Year : 2019

THAP proteins in the transcriptional control of cell proliferation

Dehaene Harmonie

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Centre Intégratif de Génomique

THAP proteins in the transcriptional control of cell proliferation

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine de l'Université de Lausanne

par

Harmonie DEHAENE

diplomée de l'École Polytechnique, Paris, France
et
Master en Biologie Médicale de l'Université de Lausanne, Suisse

Jury

Prof. Christian Hardtke, président
Prof. Winship Herr, directeur de thèse
Prof. Ivan Stamenkovic, expert
Prof. Bart Deplancke, expert

Lausanne, 2019

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Ecole Doctorale
Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président·e	Monsieur	Prof.	Christian	**Hardtke**
Directeur·rice de thèse	Monsieur	Prof.	Winship	**Herr**
Experts·es	Monsieur	Prof.	Ivan	**Stamenkovic**
	Monsieur	Prof.	Bart	**Deplancke**

le Conseil de Faculté autorise l'impression de la thèse de

Madame Harmonie Dehaene

Master ès Sciences en biologie médicale Université de Lausanne

intitulée

THAP proteins in the transcriptional control of cell proliferation

Lausanne, le 8 mars 2019

pour le Doyen
de la Faculté de biologie et de médecine

Prof. Christian Hardtke

A Aurore, mon petit soleil

Résumé vulgarisé pour le grand public

L'immense majorité des cellules de notre corps comporte la même information génétique, c'est à dire l'ensemble des instructions nécessaires à leur fonctionnement. Selon le contexte — par exemple, le type de cellule, ou encore les signaux que celle-ci reçoit — différentes unités d'information génétique peuvent être lues. Les informations peuvent donc être interprétées différemment. Les facteurs responsables de cette lecture sélective sont des protéines appelées facteurs de transcription. Dans cette thèse, je m'intéresse à une famille particulière de ces facteurs: les protéines THAP, qui sont au nombre de 12 chez l'homme. Des études précédentes ont suggéré que ces différentes protéines sont capables d'interpréter, en étroite collaboration avec une protéine partenaire appelée HCF-1, l'information génétique reliée à la prolifération cellulaire et au développement. De plus, certaines de ces protéines sont impliquées dans diverses maladies comme notamment certains cancers, une maladie neurodégénérative et un trouble rare du métabolisme de la vitamine B_{12}. Ceci rend donc la compréhension de leur mode d'action primordiale.

Ma thèse vise à mieux comprendre les mécanismes qui permettent aux protéines THAP d'interpréter l'information génétique et ainsi de réguler la prolifération cellulaire. Je montre que ces protéines sont des facteurs essentiels pour la régulation de la prolifération cellulaire, chacune possédant des fonctions et caractéristiques à la fois communes et spécifiques. En outre, je clarifie les mécanismes sous-jacents à la maladie génétique liée au trouble du métabolisme de la vitamine B_{12}. De façon générale, une meilleure compréhension des mécanismes normaux et pathogéniques liés à ces protéines est un premier pas vers une meilleure prise en charge des maladies associées.

Résumé

Les protéines THAP sont des facteurs de transcription caractérisées par leur domaine THAP, un domaine de liaison à l'ADN à doigt de zinc très bien conservé dans les espèces animales. Un nombre croissant d'études soulignent l'importance de ces protéines dans la régulation de la transcription et de la prolifération cellulaire, et leur étroite collaboration avec HCF-1, un co-régulateur de transcription. L'émergence de ces protéines ainsi que leur association avec diverses maladies humaines — comme différents cancers, la maladie neurodégénérative "dystonia 6" ou un trouble du métabolisme de la cobalamine — rend leur étude particulièrement importante.

Dans ce travail de thèse, j'étudie comment les protéines THAP régulent la transcription des gènes et la proliferation cellulaire. J'utilise une approche pluridisciplinaire combinant bioinformatique, biochimie et approches moléculaires et cellulaires, ainsi que des techniques génétiques et génomiques de pointe. Je caractérise pour commencer les différentes proteines THAP dans leur ensemble avant de concentrer progressivement mon analyse sur un nombre réduit de protéines THAP.

Tout d'abord, grâce à des outils bioinformatiques et aux banques de données disponibles, je révèle comment ces protéines ont évolué au sein dans les espèces animales, et caractérise les niveaux d'expression des gènes *THAP* chez l'homme et la souris. Ensuite, je montre que les différentes protéines THAP humaines possèdent des capacités distinctes d'homo- et d'hétérodimérisation, ainsi que de liaison avec leur partenaire supposé HCF-1. Ceci suggère l'existence d'un mécanisme complexe de régulation dans lequel les différentes protéines THAP ont des fonctions à la fois communes et particulières. Focalisant ensuite mon étude sur les deux protéines THAP7 et THAP11, j'utilise des lignées cellulaires créées sur mesure pour démontrer que ces deux protéines régulent la prolifération cellulaire. Enfin, je concentre mes efforts sur l'unique protéine THAP11 et montre qu'elle se lie à l'ADN pour réguler des gènes impliqués dans le dévelopement, la prolifération cellulaire et la transcription. En outre, je clarifie les mécanismes moléculaires à la base du trouble cobalaminique associé à une mutation du gène *THAP11*.

Les résultats obtenus dans cette thèse caractérisent les facteurs de transcription THAP comme d'importants régulateurs de la prolifération cellulaire, chacun possédant des fonctions communes et spécifiques. Une meilleure compréhension des mécanismes normaux et pathogéniques sous-jacents aux fonctions de ces protéines est un pré-requis pour une meilleure prise en charge des maladies associées.

Abstract

THAP proteins are animal-specific transcription factors, which share the THAP domain, an evolutionary conserved zinc-finger DNA-binding domain. Growing evidence implicates THAP proteins as broad transcriptional and proliferation regulators, working hand in hand with the HCF-1 transcription co-regulator. The emergence of THAP proteins and their association with diverse human diseases — such as several cancers, the dystonia 6 neurodegenerative disorder or a subgroup of cobalamin disorder — make them ripe for detailed exploration.

In this thesis work, I clarify how THAP proteins can regulate gene transcription and subsequent cell proliferation. I use a multidisciplinary strategy combining bioinformatics, biochemistry, molecular and cellular approaches, as well as state-of-the-art genetic and genomic techniques. I start by studying the whole set of THAP proteins using bioinformatics tools and gradually concentrate my analysis to smaller subsets of human THAP proteins for wet-lab experiments.

To begin with, using bioinformatics tools and available data, I shed light on the evolution of THAP proteins among animal species, and unravel the pattern of expression of human and mouse *THAP* genes. Then, I show that human THAP proteins possess differing potentials for homo- and heterodimer formation, and for binding to their putative HCF-1 partner. This suggests possibilities for an intricate THAP-protein regulatory network in which THAP proteins exhibit both shared and specific functions. Further focusing the scope of my analysis on the THAP7 and THAP11 proteins, I demonstrate that both proteins regulate cell proliferation using custom engineered THAP7 and THAP11 cell lines. Finally, I concentrate my efforts on THAP11 and show that it associates to DNA to regulate genes involved in development, cell proliferation and transcription. Particularly, I clarify the molecular mechanisms underlying the THAP11 cobalamin disorder-associated mutation.

The results developed in this thesis implicate the THAP transcriptional factors as important regulators of development and cell proliferation, each of them likely exhibiting shared and specific functions. Sheding light on the normal and pathogenic mechanisms of the THAP proteins is therefore a first step towards managing their associated diseases.

Acknowledgements

These five years of doctoral studies would have not been the same without the support of many different people, that I want to thank at the end of this doctoral journey.

To begin with, I thank Prof. Winship Herr, my PhD advisor, for his guidance, advice and help during these years. You also helped me to improve my English grammar: I will probably never forget the difference between "remember" and "remind", and I now use "however" properly! Also, many thanks to the members of my thesis committee for their valuable suggestions along the different steps of my doctoral studies: Prof. Bart Deplancke, Prof. Ivan Stamenkovic and Prof. Christian Hardtke. I was also very lucky to have been mentored by Prof. Liliane Michalik: Liliane, merci pour ton soutien, tes conseils et ta bienveillance ! Also, I would like to thank all the Herr lab members I had the chance to work with for the nice atmosphere and the stimulating scientific environment. More generally, a big thanks to all the past and present 4th floor colleagues as well as to the CIG members. I would like to particularly thank several people. Tout d'abord, Fabienne, le pilier du labo, un immense merci pour ta disponibilité, ta bonne humeur permanente et tous les conseils que tu m'as donnés. Philippe, notre expert en culture cellulaire, un merci tout spécial pour l'aide inestimable que tu m'as apportée. Travailler avec toi est un vrai plaisir, tout comme te remercier avec des gâteaux et chocolats. Un grand merci également à Maykel pour les différentes expériences que tu as réalisées pour moi sur la fin de ma thèse. Viviane, notre spécialiste en bioinformatique, merci pour tout le travail que tu as réalisé sur mes données, malgré le timing plus que serré. Merci pour ton efficacité, ta franchise et tout ce que j'ai appris en travaillant avec toi. In addition, thanks to the two platforms that have provided me technical support in my experiments: the UNIL Genomic Technology Facility for the sequencing of the samples and the EPFL Flow Cytometry Core Facility for the cell sorting of the mutant cells. Nico, François, les deux clowns du 4ème, merci avoir apporté un brin (voir un peu plus) de folie à l'étage par vos coups pendables et délires parfois douteux. Pascal, merci pour les précieux conseils de PCR que tu m'as donnés, sinon je pense que j'y serais encore. Corinne, merci pour ta bienveillance et tout le boulot administratif que tu fais pour nous tous. Et comme certaines collègues se transforment en véritables amies, un merci tout particulier à Anne-Sophie et Nathalie. Nath, la bonne fée indispensable au fonctionnement du 4ème étage et de tout le CIG, merci pour

ton immense gentillesse, tes gâteaux, nos discussions, les ravitaillements chocolat dans ton bureau, et nos séances de gymstick qui nous empêchent de marcher le lendemain. Anne-So, merci pour ton soutien, au labo comme en dehors, pour ta bonne humeur et toutes tes gentilles attentions, comme les tablettes de chocolat qui m'ont été précieuses pour boucler la fin de l'écriture. In addition, I was extremely lucky to have become, during the course of my studies, part of the Eprouvette team. Merci à toute l'équipe d'animateurs et tout particulièrement à l'équipe permanente: Patricia, Delphine, Mathilde, Séverine et Timothée, pour tout ce que j'ai appris en travaillant à l'Eprouvette et pour la confiance que vous m'accordez.

A thesis is not restricted to the lab and these 5 years of doctoral studies would not have been possible without the priceless support of my friends and family. Merci à mes amis, qu'ils soient en Suisse ou à l'autre bout du monde, pour m'avoir toujours soutenue. Mathieu et Laetitia, merci pour votre présence, même de loin. Anne-Laure, Mathieu, Guliz, Evrim et Nadja, merci pour votre soutien moral, vos encouragements et nos afterworks. Amaia, Santiago, Navina, Loïc, thanks for all the good times we had since the beginning of the master, the laughs and the crazy cheese diners. Martin, merci pour nos rires à en avoir mal au ventre, nos bêtises et notre complicité, et d'avoir toujours été present (et patient) quand j'avais besoin de râler. Les copains de "Plastik", merci pour vos encouragements malgré la distance, les filets de citron, les coeur coeur love et nos jours de l'An. Tout particulièrement, Thomas et Clotilde pour toutes nos bonnes bouffes et soirées ensemble ; et Lucile et Cécilia, merci les filles pour votre soutien sans faille, encore plus ces derniers mois, les "pump it up" réguliers pour me motiver, les rigolades pour me détendre, les bisounoursitudes pour m'encourager. Un grand merci à toute ma belle famille, et tout spécialement à Ghislaine et Stanislas, mes beaux parents, aux 4 grands-parents de mon mari (Mèrèse, Ivan, Bernadette et Philippe) et à Sophie, ma cousine d'adoption, pour m'avoir également toujours soutenue et encouragée. Un immense merci à ma famille, qui, malgré la distance, ont toujours été mes premiers fans, même s'ils ne comprenaient pas forcément ce que je trafiquais avec mes cellules. Mes parents, Agnès et Bruno, pour avoir toujours cru en moi et m'avoir poussé à me dépasser et à ne jamais rien lacher. Merci d'avoir toujours été là pour m'épauler avec amour. Mes frères et soeurs, Bérénice, Hubert, Philomène et Titus (en ordre chronologique, et non de taille !...), ainsi que les valeurs ajoutées Dom, Émilie et Fabio, merci pour votre soutien et votre amour, nos délires et blagues pourries qui ne font rire que nous, les répliques de *La cité de la peur*, les innombrables messages whatsapp sur nos 22 groupes différents, les apéros et nos vacances à la Grande-Motte. A mes grand-mères, Zaïme et Line, pour leur soutien et leur question récurrente: "Alors, ta thèse, ça avance ?" certes légèrement angoissante à force, mais qui montre surtout leur intérêt et amour. Enfin, mes derniers remerciements vont à mes deux plus grands soutiens, ma petite famille à moi: mon mari Guillaume et ma fille Aurore. Guillaume, merci de m'avoir supportée, dans tous les sens du terme, dans cette aventure. Merci pour les bons petits plats et gâteaux au chocolat réconfortants, pour m'avoir écoutée, rassurée, encouragée, ou secouée quand il y

en avait besoin. Merci (et bravo !) d'avoir intégralement relu mon manuscrit de thèse pour chasser les fautes, et merci de t'être cassé la tête à résoudre mes problèmes de LaTeX. Merci pour ton amour et ta gentillesse, j'ai énormément de chance de t'avoir à mes côtés. Aurore, mon petit soleil, mon joli clown: tu ne comprends pas encore tout, mais merci d'avoir apporté tant d'amour dans ma vie, et d'avoir la capacité de me faire oublier tout le reste quand je suis avec toi. Tu as été à la fois ma plus belle distraction et la meilleure des motivations pour finir cette thèse.

Table of contents

List of Figures

List of Tables

Chapter 1

Introduction

This study explores some of the mechanisms of cell-proliferation regulation via the control of gene expression.

1.1 General principles

To better understand this research thesis, I review of some general principles of cell proliferation, gene expression and genome organisation. The goal is not to be exhaustive but to illuminate some key points that will be important for understanding the present study.

1.1.1 Cell proliferation in mammalian cells

Cell proliferation is a fundamental process by which cells reproduce themselves by growing and then dividing into two daugther cells. This process is key in development by which multicellular organisms are generated from a single cell. In addition, in adult organisms, cell proliferation is also essential as probably every organ has the ability to regenerate (abeit to very different extents), either a normal state (e.g. the skin) or upon damage (e.g. the liver).

Cells reproduce themselves by a series of highly conserved and coordinated events called the cell cycle, which has been divided into four phases in eukaryotic cells (Figure 1.1). In the presence of growth signals, resting cells can pass a "restriction point" late in the G1 phase, thereby becoming committed to progress through the cell cycle. DNA duplication takes place in S phase (S for Synthesis), which typically requires 10 to 12 hours. In M phase, which is much shorter (about one hour), nuclear division first takes place with chromosome segregation (mitosis), followed by cell division (cytokinesis). The two "gap phases", named G1 and G2, have two majors functions. First, they allow sufficient time to synthesize the requisite proteins and organelles for cell growth in G1. Second, they ensure the satisfaction of different checkpoints along the

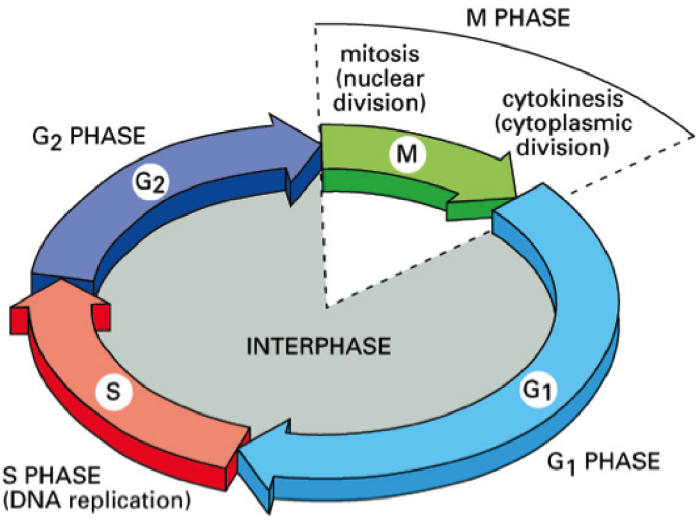


Figure 17–3. Molecular Biology of the Cell, 4th Edition.

Figure 1.1: **The eukaryotic cell cycle.** [1]

progression through the cycle, to guarantee that the previous step have been performed properly and thus that the conditions are favourable to pursue the cycle. Finally, cells in early G1 phase can either go into another round of cell cycle, or exit the cell cycle and enter a resting state, called G0, in which they can stall for hours, days or even a lifetime [1].

Cell-cycle progression is tightly regulated. Central proteins in this regulatory system are cyclins and their partners, the cyclin-dependent kinases (CDKs). While CDK concentrations stay quite constant, cyclin concentrations oscillate along the cell cycle as the result of the gene transcription and protein degradation. More generally, control of the transcription of cell-cycle-related genes provides an important layer of regulation of cell proliferation [1].

1.1.2 Regulation of transcription in eukaryotes

Most cells in our bodies possess an identical set of genetic information. Yet, different cell types exhibit different patterns of gene expression to perform their specialized functions. A given cell can also alter the expression of its genes depending on the conditions (e.g. nutrient availability, proliferation signals). To do so, eukaryotes have acquired highly sophisticated regulatory mechanisms, in particular for gene transcription, which is the first step of gene expression.

General aspects of gene-transcription regulation

Diverse classes of proteins interplay and positively or negatively impact transcription (Figure 1.2).

Sequence-specific DNA-binding transcription factors directly bind DNA regulatory elements (e.g. enhancers, promoters) in a sequence-specific manner and can activate, or repress, the transcription of the corresponding gene. They also recruit other proteins, named as transcription co-factors, that do not directly bind DNA but further contribute to the activation or repression of transcription. Often, one transcription (co-)factor can participate in either activating or repressing transcription depending on the other partners in the regulatory complex.

In the nucleus, the DNA is highly compacted and the first level of compaction is the nucleosome, which is an octamer of two of each histone protein H2A, H2B, H3 and H4, and DNA wrapped around them [2]. Such a compaction of DNA could be an obstacle for DNA transcription, which requires the binding of numerous factors to DNA, thus requiring accessibility to the DNA. Fortunately, chromatin is a highly dynamic structure: it can be remodelled by displacing the nucleosomes along the DNA, removing nucleosomes, or modifying their composition (e.g. incorporating histone variants instead of canonical ones). These modifications, mediated by chromatin-remodeling complexes, can have various transcriptional outcomes. Indeed, it results in the modification of the DNA accessibility by the transcription and regulatory machinery [3, 4]. In addition, the N-terminal tails of histones can be post-translationally modified (e.g. lysine and arginine methylation, serine and threonine phosphorylation, lysine acetylation). These modifications regulate gene transcription by locally modifying the chromatin structure (for instance, histone acetylation loosens DNA-histone interactions) and by recruiting or excluding non-histone proteins. Indeed, some proteins can "read" a so-called "histone code". These readers can in turn further influence the recruitment of additional factors — such as additional histone-modifying enzymes or chromatin remodelers — and the regulation of transcription [3, 4]. As an example, the tri-methylation of the lysine 4 in histone H3 (H3K4me3) at promoters and close to the transcription start sites is associated with high RNA polymerase II occupancy and active transcription [5, 6]. As a downstream consequence, this methylation mediates the recruitment of chromatin-remodeling machinery [6]. In contrast, unmethylated H3K4 is perceived as a signal for DNA methylation, which generally represses transcription when located at promoters and enhancers [5, 7].

Finally, scaffold proteins, which do not have a transcriptional activity *per se*, serve as binding platforms to help different effectors to bind to each other (e.g. transcription factors and chromatin modifiers) and build high-order protein complexes.

Transcription-factor dimerization as a regulatory strategy

Many eukaryotic transcription factors form dimers or higher-order oligomers. Regarding sequence-specific DNA-binding transcription factors, dimerization has two major consequences. First, it both increases DNA-binding affinity and specificity, as it simultaneously permits cooperative DNA binding and increases the length

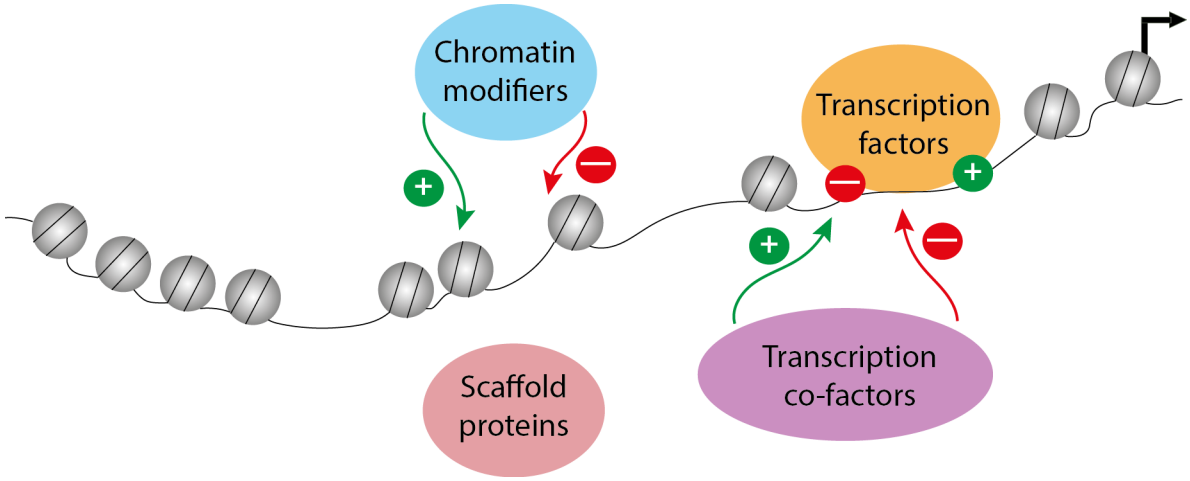


Figure 1.2: **Simplistic view of gene-transcription regulation.**

of the DNA-bound site [8]. Second, it increases the complexity of gene regulatory networks: as each monomer can potentially have several binding partners, such that multiple different dimers can be formed, each having a distinct DNA-binding specificity and transcriptional effect, and thus potentially different functions [9].

As an example, a functional E2F transcription factor is actually a heterodimer made of an E2F (E2F1-8) protein and a DP (DP1-2) protein. Each individual subunit possesses a DNA-binding domain but their DNA binding is mutually dependent. They thus regulate transcription in a synergistic manner, each subunit potentiating the E2F transcription-factor activity. Indeed, depending on the E2F partner in the E2F-DP complex, the transcriptional outcome will differ: for example, while the E2F1 member activates transcription, E2F4 represses transcription [10,11]. Another good example is the group of Max, Myc and Mad transcription factors. Max can interact with both Myc and Mad, whereas Myc and Mad are not able to bind each other. When Max is bound to Myc, the resulting Myc-Max complex activates transcription. In contrast, the Mad-Max heterodimer silences transcription [9].

1.1.3 Gene families and evolution

The study of genomes reveals that many genes belong to gene families, which are groups of genes coming from a single common ancestor gene and retaining similar sequences and often related functions [12]. This concept is used to describe genes belonging to the same genome — called paralogs. It is an efficient way of increasing the genetic diversity [13]. Several models have been proposed to explain their appearance and evolution. They all involve the concept of gene duplication within a single genome and subsequent evolution — the difference being in the way the genes (or portions of gene) are duplicated and in the way the duplicated segments evolve [14]. For instance, an intact "ancestor" gene can be duplicated, thus generating redundant copies, each of them being able to independently diverge ("divergent evolution" model) [13,14]. On one hand,

deleterious mutations can occur, resulting in non-functional (pseudo) genes. On the other hand, mutations can give rise to distinct — yet related — genes. For instance, they can diverge in expression patterns (e.g. tissue specificity) or in functions [12, 13]. As an example, the E2F transcription factors form such a family of paralogs with opposing functions: while E2F1 activates transcription, E2F4 is a transcriptional repressor [10, 11]. Of note, regarding transcription factors, the divergence of function can also be translated into a divergence of DNA-binding specificity, and thus in a divergence of the genes regulated.

This gene duplication and divergence mechanism can result in both genetic redundancy and increased complexity. While the paralogs have acquired specialized features (sub-functionalization), they often retain at least part of their original common function and thus can be, to a certain extent, redundant. This genetic redundancy can often complicate research to unravel the function of a specific protein. The resulting complexity, however, allows an increased robustness of biological processes and promotes evolution [1].

Of note, an hypothesis suggests that, at very early stages of vertebrate evolution, the whole genome successively underwent two rounds of duplication. Such duplications would have resulted in the creation of four copies of every gene. Since then, evolution has led to the divergence of the different copies as well as to the loss of some [1].

An excellent example of duplication in vertebrate evolution is the divergence and decay in the globin-gene family. Indeed, present-day globin genes result from a single ancestor globin gene. This amplification and evolution resulted in a four-chain globin molecule in modern vertebrates. This hemoglobin form is more efficient than the ancestor single-chain one to transport oxygen in larger organsisms. In addition, it also enabled the emergence of the fetal β chain in mammals, which allows the transfer of oxygen from the mother to the fetus. Also, several duplicated globin genes have lost function along the course of evolution, giving rise to pseudogenes [1].

1.2 HCF-1: a cell-cycle regulator which regulates gene transcription

HCF-1 (for Host Cell Factor 1) has been initially discovered as a key cellular protein for the lytic phase of the herpes simplex virus infection. Indeed, HCF-1 stabilizes the so-called "VP16-induced complex", a transcriptional regulatory complex formed by itself, another host protein named Oct-1, and the viral protein VP16. This complex assembles on viral immediate-early promoters to promote their transcription and initiate the lytic phase of the infection [15].

1.2.1 The HCF-1 protein

In vertebrates, HCF-1 is synthesized as a large precursor protein that undergoes proteolytic maturation. This generates non-covalently associated amino- and carboxy-terminal moieties, called HCF-1_N and HCF-1_C, respectively [15] (Figure 1.3, top). This maturation occurs at centrally-located repeats, called HCF-1_{PRO} repeats, which are highly conserved among vertebrate species. Curiously, the processing is directly performed by the OGT glycosyltransferase, which both O-GlcNAcylates and cleaves HCF-1 [16].

Additional notable features of the human HCF-1 proteins are [15] (Figure 1.3, top):

- a Kelch domain, arranged in a β-propeller structure, responsible for mediating the interaction with a large number of HCF-1 partners;
- basic and acidic regions;
- two fibronectin type 3 (Fn3) repeats involved in the association of the HCF-1_N and HCF-1_C subunits;
- a C-teminal nuclear localization signal (NLS).

Interestingly, missense mutations in the *HCFC1* gene encoding for the HCF-1 protein (located on the X chromosome) have been associated with diseases: a mental retardation syndrome (non-syndromic intellectual disability [17–20]) and neurological diseases associated with vitamin metabolism defects (X-linked cobalamin disorder [18, 21] and X-linked cobalamin disorder associated with multiple congenital abnormalities [22]).

1.2.2 HCF-1 regulates two phases of the cell cycle

It was early on suggested that HCF-1 plays a role in cell-cycle progression, as a single missense mutation in HCF-1 in a baby-hamster kidney cell line results in a temperature-sensitive cell-cycle arrest [23]. Indeed, both HCF-1 subunits ensure complementary roles in cell-cycle progression: while HCF-1_N promotes G1-to-S-phase transition, HCF-1_C allows proper mitosis and cytokinesis during M phase [24] (Figure 1.3, bottom). It has been shown that HCF-1 is required for proliferation in embryonic development and liver regeneration after 2/3 partial hepatectomy [25, 26]. The early hypothesis was that HCF-1 controls cell proliferation by regulating gene transcription [23, 27].

1.2.3 HCF-1 is a broad and versatile regulator of transcription

HCF-1 regulates — positively or negatively — transcription by serving as a "scaffold protein" as described in 1.1.2 [15]. It is an abundant chromatin-associated protein [27] that acts as a hub by connecting chromatin modifiers to selected promoter-binding transcription factors (Figure 1.3, bottom). This results in a selective

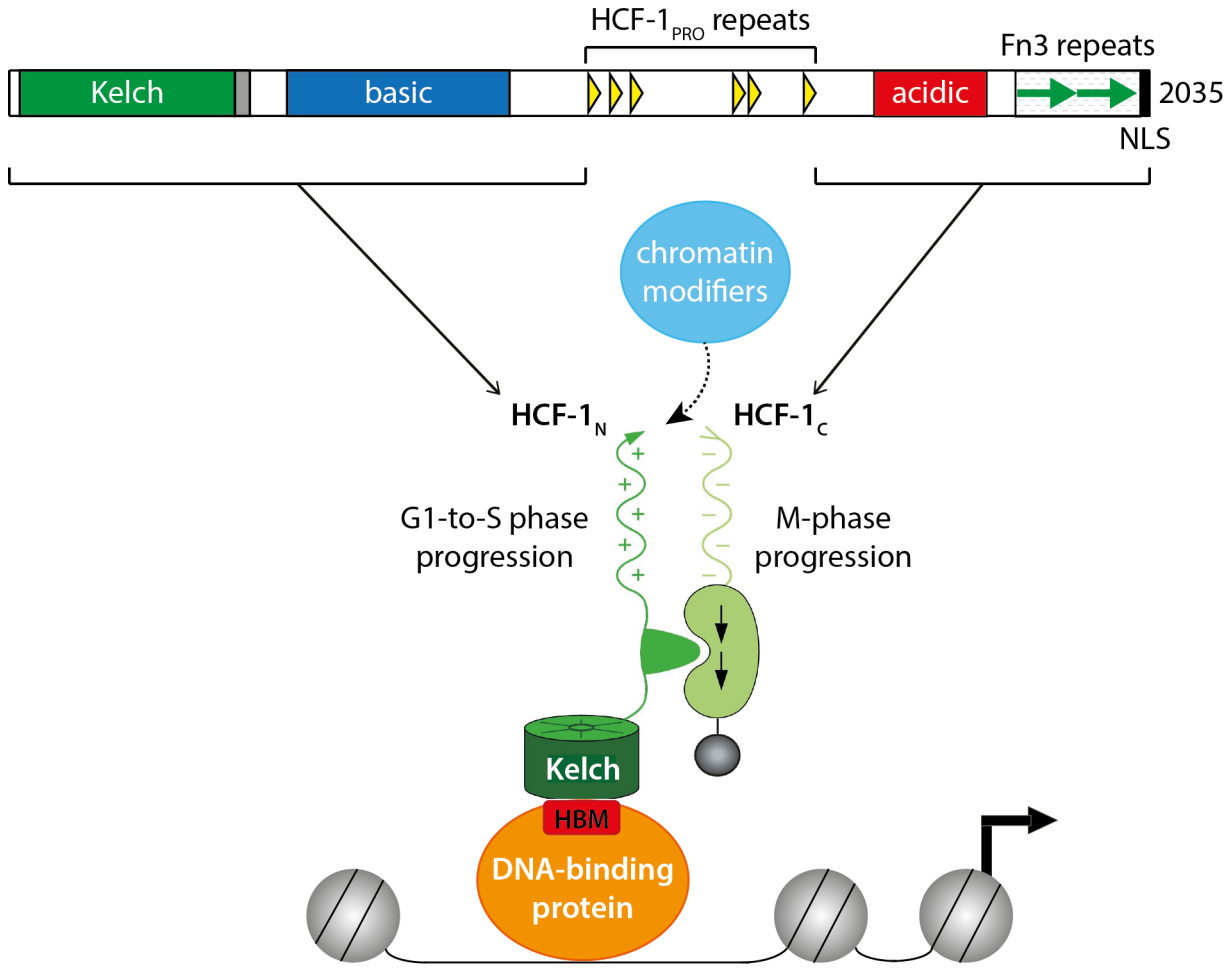


Figure 1.3: **HCF-1 is an heterodimer, each of its subunits being required for a distinct phase of the cell cycle.** Top, schematic representation of the human HCF-1 precursor protein. Bottom, HCF-1 is indirectly tethered to chromatin by interaction of its Kelch domain with HBM-bearing DNA-binding proteins, and acts as a hub by connecting chromatin modifiers to selected transcription factors.

modification of chromatin structure at specific promoters, enabling either positive or negative transcription outcomes, depending on the partners in the complex. For instance, in a cell-cycle dependent manner, HCF-1 alternatively builds activating and repressing complexes by recruiting selected histone modifiers (histone methyltransferase and histone deacetylase, respectively) to E2F-regulated promoters (E2F1 and E2F4, respectively) [28].

Of note, HCF-1 has a paralog called HCF-2 which seems to have opposing cell-cycle and transcription functions to HCF-1 (Gudkova and Herr, personal communication).

1.2.4 HCF-1 acts as a member of a chromatin complex

As mentioned, HCF-1 regulates transcription as part of transcription-regulator complexes, but does not bind DNA directly. Thus, to perform its functions, HCF-1 needs to associate with DNA-binding proteins. Of note, the precise identity of the proteins responsible for HCF-1 recruitment to chromatin remains controversial [28, 29].

Many different proteins, of diverse functions, can interact with HCF-1 (Figure 1.4). Even if it is not the sole mode of interaction, its N-terminal Kelch domain mediates a large part of HCF-1 associations [30]. The latter recognizes a so-called HCF-1 Binding Motif (HBM; consensus sequence $^{B}/_{Z}$HxY, where B and Z are aspartate/asparagine or glutamate/glutamine, respectively, and x denotes any amino-acid), which is present in a large variety of DNA-binding proteins, from transcription factors to chromatin-modifying enzymes [28, 31–37].

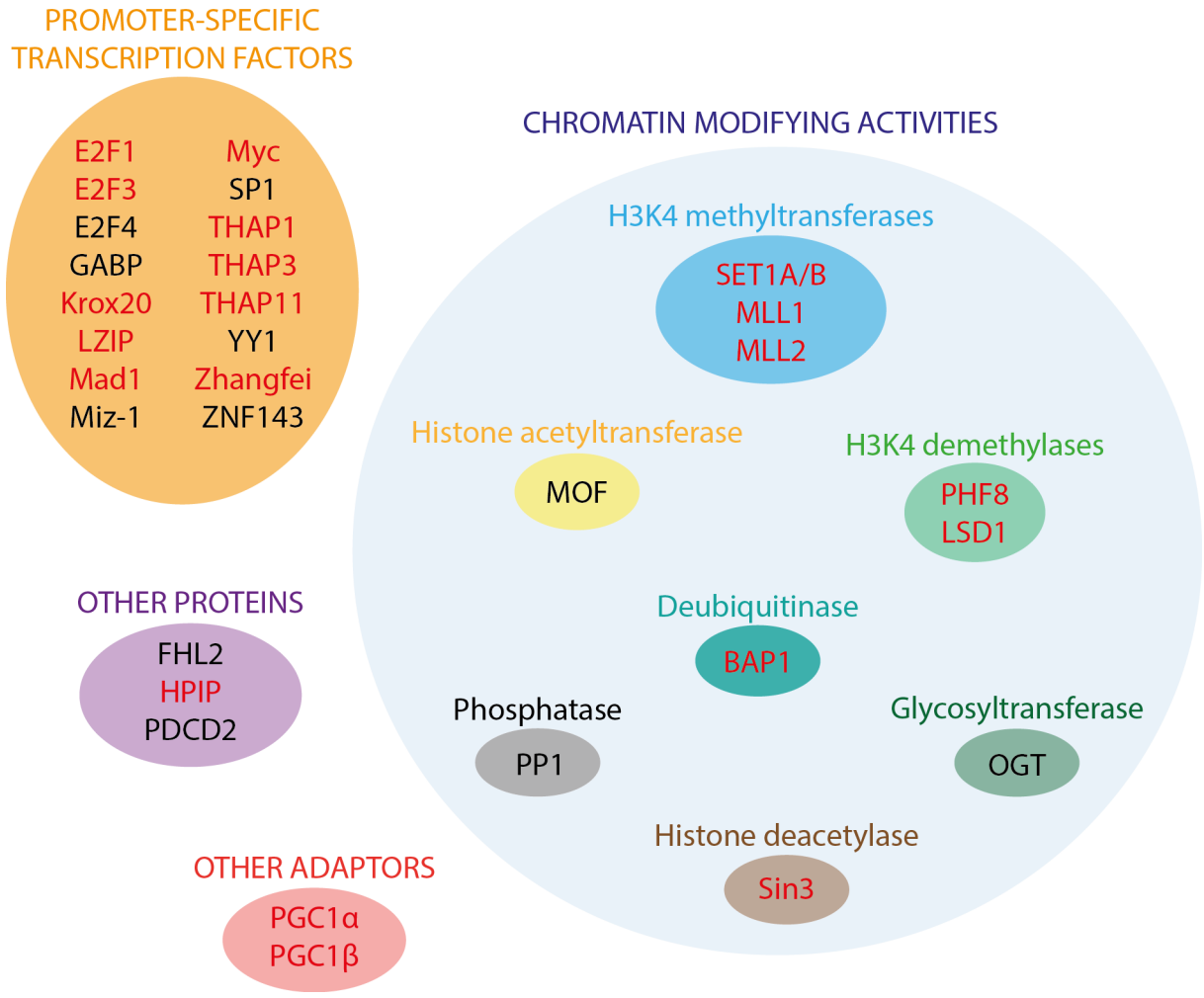


Figure 1.4: **Landscape of HCF-1-interacting proteins.** Overview of proteins known to interact with HCF-1, grouped by function. The ones bearing an HBM are in red.

In addition, most of the THAP proteins, a recently-discovered family of transcription regulators, possess an HBM sequence, and at the onset of my doctoral research three — THAP1, 3 and 11 — had been shown to interact with HCF-1 in human extracts [38, 39] (see Figure 1.4).

1.3 The THAP transcriptional regulators: a family of HCF-1 interactors

The 12 human THAP proteins are encoded by a family of gene paralogs (Figure 1.5). These proteins are defined by their N-terminal THAP (for THanatos — referring to the Greek God of Death — Associated proteins) domain, an atypical zinc-finger domain.

1.3.1 Discovery of the THAP family of proteins

This family of proteins was uncovered 15 years ago [40]. The authors noticed that the N-terminal portion of THAP1, a novel pro-apoptotic factor, shares some features with the N-terminal part of 11 other human proteins, most of them being uncharacterized at the time. They were therefore called THAP2 to THAP11, and the previously characterized p52rIPK protein was renamed as THAP0 [40].

1.3.2 Characteristics of THAP proteins

In addition to their THAP domain, the 12 human THAP proteins share several structural motifs, while some also have unique characteristics (Figure 1.5).

The THAP domain The THAP domain is a sequence-specific zinc-dependent DNA-binding domain restricted to the animal kingdom [40] (Figure 1.5, red box). It is characterized by its N-terminal location, its size of around 80-90 amino acids, and the presence of several conserved features [40, 46]:

- A "C_2CH signature" (Figure 1.6A, red-highlighted residues), which forms an atypical zinc finger. Its consensus sequence is Cys-X_{2-4}-Cys-X_{35-50}-Cys-X_2-His, where X denotes any amino-acid. These 4 residues provide the ligands for the zinc coordination and are each individually required for DNA binding of the THAP domain.
- Four key residues strictly conserved in the different THAP members (Figure 1.6A, blue-highlighted residues), defined as the THAP1 amino-acids P26, W36, F58 and P78. Similarly to the 4 zinc-finger amino-acids mentionned above, each of them is required for binding to DNA.

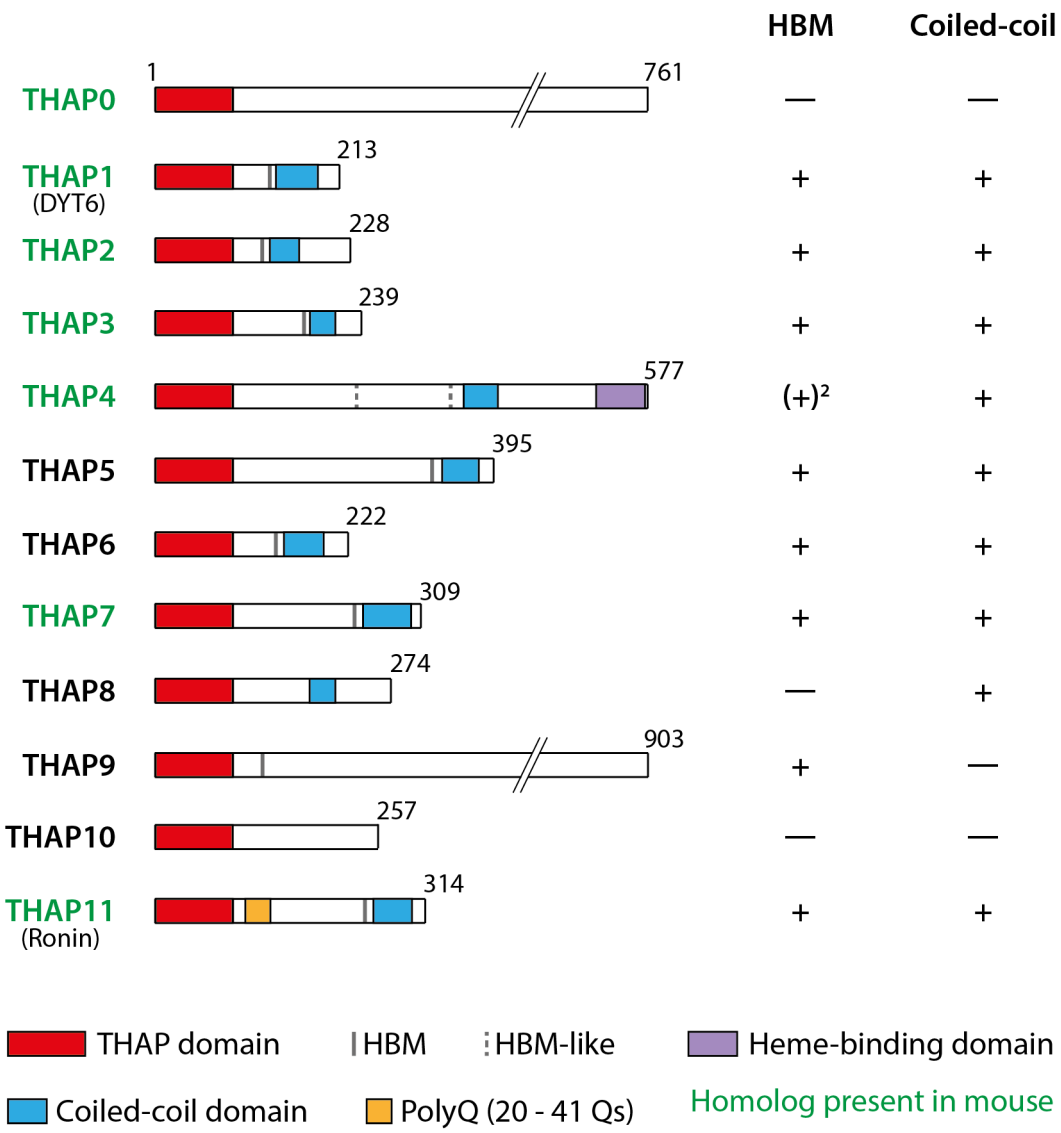


Figure 1.5: **The human family of THAP proteins.** Schematic representation of the structures of the 12 human THAP proteins.

- A C-terminal "AVPTIF box" (Figure 1.6A, green box), suggested to be necessary for the proper folding of the zinc finger.
- Several additional residues, abeit not strictly conserved, with particular physico-chemical properties.

Interestingly, the THAP domains of distinct proteins display similar three-dimentional structures (Figure 1.6B), while not recognizing the same DNA target sequence (Figure 1.6C). Thus, each THAP protein appears to have its own DNA-binding specificity [41–45].

Similarities between the THAP domain and the DNA-binding domain of the *Drosophila melanogaster* P-element transposase have been pointed out early on [40]. It has led to the hypothesis that the THAP

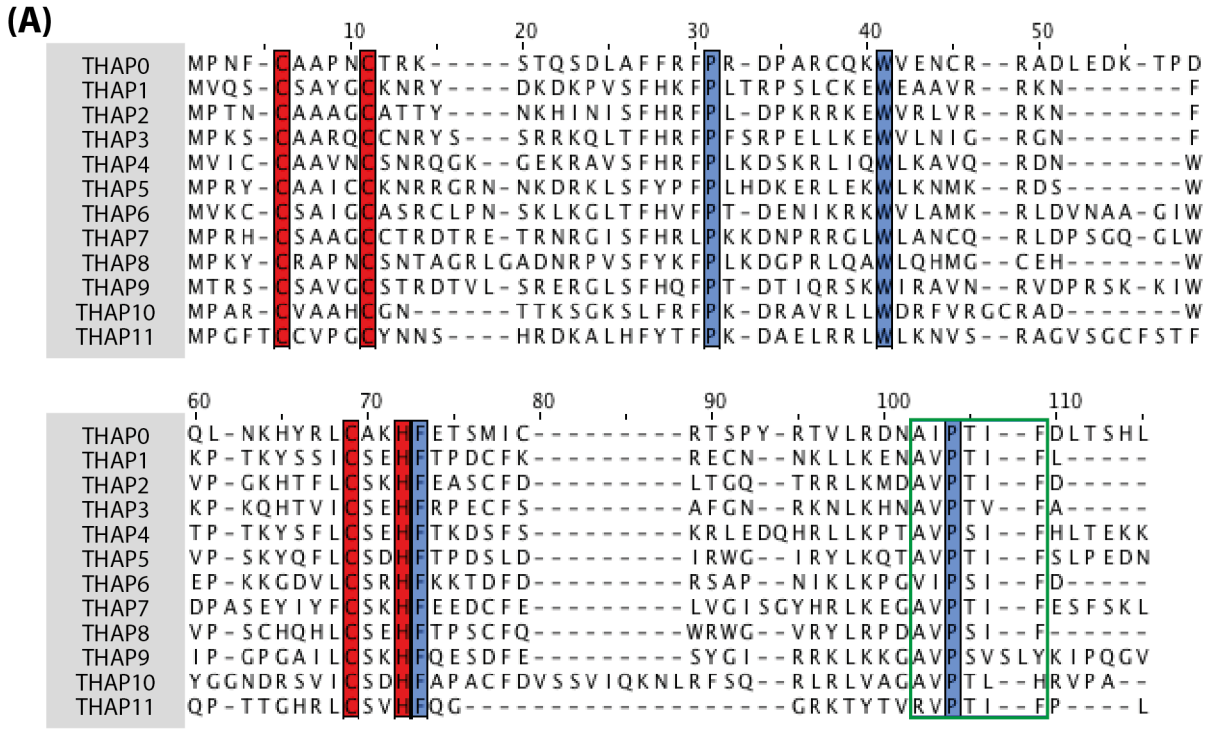

(A)
THAP0
THAP1
THAP2
THAP3
THAP4
THAP5
THAP6
THAP7
THAP8
THAP9
THAP10
THAP11
(B)
THAP1
N
Arg42
Lys46
Thr48
Lys24
C
THAP11 (Ronin)
L4
L3
L1
N
P78
F66
W87
P28
L2
THAP1
THAP2
(C)
THAP1
THAP5
THAP9
THAP11
(Ronin)
Bits
bits
5′

Figure 1.6: **The THAP domain. (A)** Sequence alignment of the 12 human THAP proteins, with the conserved features highlighted: red, "C_2CH signature" of the zinc finger; blue, strictly conserved residues; green box, "AVPTIF box"). **(B)** Cristal structure of the THAP domain of human THAP1 (left), human THAP11 (middle), and comparaison of the structures of the THAP domains of human THAP1 and THAP2 displayed in the same orientations (right) [41, 42]. **(C)** DNA-binding motifs identified for some human THAP proteins, by SELEX approach (THAP1 and THAP9), CASTing (Cyclic Amplification and Selection of Targets) (THAP5) or ChIP-seq (THAP11) [43–45].

domain originated from the domestication of P-transposable elements [47], domestication being the change from a parasitic element into a beneficial host gene [48]. Curiously, THAP9 was shown to be related to the *Drosophila melanogaster* P-element transposase and suggested to retain the DNA-transposase activity [40,49]. Nevertheless, these two proteins are overall too different from each other to be considered as orthologs.

The coiled-coil domain The coiled-coil motif is a common structural domain involved in protein homo- and heterodimerization. It is found on highly diverse classes of proteins and predictions have suggested that around 2 to 3% of all protein residues are involved in a coiled-coil domain. In particular, many transcription factors possess a coiled-coil domain, which is interesting regarding the role of protein dimerization in transcription regulation (as seen in 1.1.2) [50].

Most of THAP proteins display such a motif. Indeed, 9 out of the 12 human THAP proteins have it (Figure 1.5, blue box), which is of potential interest regarding their role as transcription factors (for more details on the identification of the coiled-coil domain in THAP proteins, see section 3.3). And yet, so far, only THAP0 [51], THAP1 [52,53] and THAP11 [54,55] have been shown to homodimerize.

The HBM sequence Several of these proteins additionally display an HBM sequence for potential HCF-1 association: 8 have a consensus HBM and THAP4 has 2 identical HBM-like sequences (Figure 1.5, grey and dashed-grey lines, respectively). On the other hand, THAP0, 8 and 10 do not possess any evident HBM sequence [38]. Of note, when both a coiled-coil domain and an HBM sequence are present in a THAP protein, they are always positioned in the same orientation — the HBM being N-terminal to the coiled-coil — and close to each other.

Heme-binding domain THAP4 has been shown to have a C-terminal heme-binding domain [56] (Figure 1.5, purple box) apparently able to catalyze the detoxification of peroxynitrite [57], connecting THAP4 with metabolism.

Poly-glutamine tract THAP11 displays a poly-glutamine tract typically of 29 amino-acids (Figure 1.5, orange box), but varying from 20 to 41 glutamine residues in different individuals. To date, it is not clear

whether this polyQ tract could cause any disease [58, 59].

Conservation in mouse Several, but not all THAP proteins have an ortholog in mouse. In the mouse genome, there are only 7 *Thap* genes out of the 12 human ones [46] (Figure 1.5, green THAP names).

1.3.3 Importance of THAP proteins

At the time of their discovery, the THAP proteins were hypothesized to be DNA-binding proteins and to have a role in gene transcription. But, it has so far only been shown explicitely for a subset of them: THAP1 [38,60], THAP3 [38], THAP5 [45], THAP7 [61,62], THAP10 [63] and THAP11 [54]. Their mechanisms of action and outcome on gene transcription are diverse, and THAP-protein and context dependent.

In the past decade, advances have shed some light on the cellular roles of several THAP members. Importantly, they have often been involved in cell proliferation — both cell growth and cell-cycle progression — and in development:

- THAP0 is a positive regulator of the Protein Kinase R (PKR), a cell growth suppressor [64];
- THAP1 positively regulates cell-cycle progression in endothelial cells by activating transcription of many pRb/E2F cell-cycle target genes, including the *RRM1* promoter, whose gene product, a subunit of ribonucleotide-diphosphate reductase, is essential for S-phase DNA synthesis [60]. In addition, THAP1 is an essential factor for mouse embryonic stem cells (ESCs): whereas not required for maintenance of the pluripotency, THAP1 loss promotes ESCs cell death and severely impairs cell proliferation by blocking cells into S phase. It is also necessary for silencing pluripotency genes and subsequent neural differentiation [65].
- THAP5 has been proposed to be a negative regulator of cell-cycle progression at the level of the G2-to-M-phase transition [66];
- THAP7 has been suggested to control cell proliferation by interaction with the Histone Nuclear Factor P (HiNF-P), a major activator of histone H4 gene transcription [67];
- THAP10 inhibits cell proliferation and promotes cell differentiation of acute myeloid leukemia cells bearing the t(8;21) translocation [63];
- THAP11 has been extensively demonstrated to regulate cell proliferation, but whether it promotes or represses cell proliferation has remained controversial [39, 54, 68, 69]. Nevertheless, THAP11 has been repeatedly shown to bind numerous cell-growth (and metabolism) and cell-cycle-related genes promoters [29,39,68–71]. In addition, numerous studies have highlighted its essential role in pluripotency [54,72,73]

and in early vertebrate development of various organs: the retina [74], the heart [71], the brain [75, 76] and during hematopoiesis [77].

Of note, as their name indicates, several THAP proteins have also been implicated in apotosis: THAP0 [51, 78], THAP1 [79] and THAP5 [45, 66].

Finally, the importance of THAP proteins in humans is highlighted by their role in diverse human diseases. First, THAP1 (also called DYT6) is associated with primary monogenic torsion dystonia, a neurodegenerative movement disorder characterized by involuntary muscle contractions. Dystonia 6, the second most frequent dystonia, is due to dominant mutations lying throughout the *THAP1* gene, with a large majority of mutations falling inside the THAP domain [80–82]. In addition, *THAP5*-gene expression has been linked with heart disease [66]. Also, as for HCF-1, THAP11 has been implicated in cobalamin disorder, an inborn vitamin deficiency associated with neurodevelopmental abnormalities [75, 83]. Finally, many THAP genes have been associated with cancer: THAP3 [84], THAP4 [85], THAP5 [45], THAP10 [63, 86] and THAP11 [39, 68, 69, 87, 88].

1.3.4 THAP proteins and HCF-1

As mentioned above (Figure 1.5), 9 out of the 12 human THAP proteins have an HBM or HBM-like sequence, through which they could potentially interact with HCF-1. Mazars and colleagues showed, in a yeast two-hybrid assay, that the HCF-1 Kelch domain binds to the human THAP proteins that bear an HBM or HBM-like sequence but not to the three THAP members that are devoid of an HBM sequence [38]. As mentioned above, the association of HCF-1 with THAP1 [38], THAP3 [38] and THAP11 [39] has also been shown in human cell extracts.

Importantly, HCF-1 has been implicated in THAP-mediated functions:

- HCF-1 is critical for *RRM1* promoter activation by THAP1 [38]. Also, dystonia-associated mutations have been uncovered in the HBM of THAP1, suggesting the importance of binding with HCF-1 for THAP1 activities. Also, it has been shown than HCF-1 is present at more than 90% of THAP1-associated promoters, this recruitment being THAP1-dependent — and more precisely, THAP1-HBM dependent [89].

- Likewise, THAP3 interaction with HCF-1 has also been shown to be required for THAP3-mediated transcription regulation [38].

- HCF-1 has been extensively demonstrated to be a critical THAP11 partner for regulation of transcription and cell proliferation, with THAP11 and HCF-1 binding and regulation of transcription to their

common promoters being mutually dependent [29, 39, 43, 54, 70, 71, 75].

Consequently, HCF-1 seems to be an important co-factor in THAP-protein activities — an importance that warrants further investigation.

1.4 Genetic methods: creating custom cell lines to investigate specific proteins

Genetics is often used to unravel the functions of genes and their associated encoded proteins. This process is called reverse-genetics as it starts from the genotype to uncover the associated phenotype. To determine the function of a specific gene, the most common approaches aim to study mutants (cells or organisms) that either lack this gene, express an altered version of it, or otherwise over-express the gene of interest [1].

These past years, many technological advances allowed the development of easy and efficient techniques to create such custom genetic tools and to decipher the function of genes.

1.4.1 The CRISPR/Cas9 technique of genome editing

Engineering specific mutations (either loss-of-function mutations, less drastic functional mutants, or disease-associated mutations) in the laboratory gives the opportunity to assess their effect on the function of proteins, then giving insights into their role and mechanism of action. The challenge is to alter the endogenous version of the gene of interest, in a precise and efficient way, without affecting the rest of the genome. The so-called CRISPR/Cas9 system, which arose this past decade and has become very popular, provides an extremely satisfying solution.

Discovery

Clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated sequences (Cas) were initially discovered as part of the adaptive immune system of microbes. It allows bacteria and archaea to defend against viruses and plasmids. By using single-strand RNA molecules, Cas nucleases are targeted to foreign DNA in a sequence-specific manner. This results in the creation of a double-strand break (DSB) at specific locations of the foreign DNA and its subsequent degradation [90, 91].

Hijacking the CRISPR/Cas immune system for precise genome editing

This immune system has been diverted to precisely and efficiently edit genomes. Here, a human codon-optimized Cas9 endonuclease is used in combination with a single-strand chimeric RNA in which the two

RNA molecules traditionnally used by the Cas nuclease are fused. This latter RNA molecule, named guide RNA (gRNA), will target the Cas9 endonuclease to a specific locus by direct base-pairing interaction of its guide sequence of 20 nucleotides with the targeted DNA locus (Figure 1.7A). Thus, this system can be designed to target virtually any location in a genome by adapting the 20-nucleotide guide sequence inside the gRNA [90,92]. The single requirement is that the target sequence must be immediately followed, in the host, by a so-called protospacer adjacent motif (PAM), necessary for Cas endonuclease activity (Figure 1.7A). The PAM requirement differs depending on the Cas endonuclease ; in the case of Cas9, it is 5'- NGG [90].

Following cleavage of the DNA at the target location, DNA-damage repair pathways are activated [90]:

- By default, error-prone non-homologous end joining (NHEJ) takes place to repair the DSB, resulting in the formation of indels. This can typically be used to create loss-of-function mutations, as indels can lead to frameshifts and premature stop codons (Figure 1.7 B, left).
- Alternatively, a repair template can be used to activate the homology-directed repair (HDR) pathway. Although occuring less frequently than NHEJ, HDR can be exploited to introduce precise modifications of the targeted DNA sequence. This requires providing the system with a repair template for the recombination, bearing the desired changes of the sequence (Figure 1.7B, right).

CRISPR/Cas9 procedure for cell lines in culture

First, a gRNA that fits ones goals needs to be designed. It consists in identifying the 20-nucleotides guide sequence which will direct the Cas9 to the locus of interest (Figure 1.7 A, blue sequence). Again, this sequence must be followed, on the genomic locus, by a 5'- NGG PAM (red line). Of note, Cas9 will cleave the target sequence around 3 base pairs upstream of the PAM (red triangle). The guide sequence and a scaffold RNA sequence (blue and red sequences, respectively) create the gRNA. This gRNA needs to be delivered to cells together with a Cas9-expression construct and, if necessary, a DNA-repair template. One easy way to do that is to create an expression plasmid encoding both the sgRNA and the Cas9 endonuclease, which can be delivered to cells by transfection [90].

In case the objective is to create specific modifications of the locus, a repair template is also necessary for HDR. An easy option is to use single-stranded DNA oligonucleotides (ssODNs), with homology arms of 40 to 80 nucleotides of each side, which are transfected into cells together with the gRNA-Cas9 expression plasmid. In that case, the gRNA sequence needs to be chosen so that the desired modification (*i.e.*, mutation) of the locus is in close proximity of the DSB generated by Cas9 (around 10 nucleotides maximum for a good efficiency) [90].

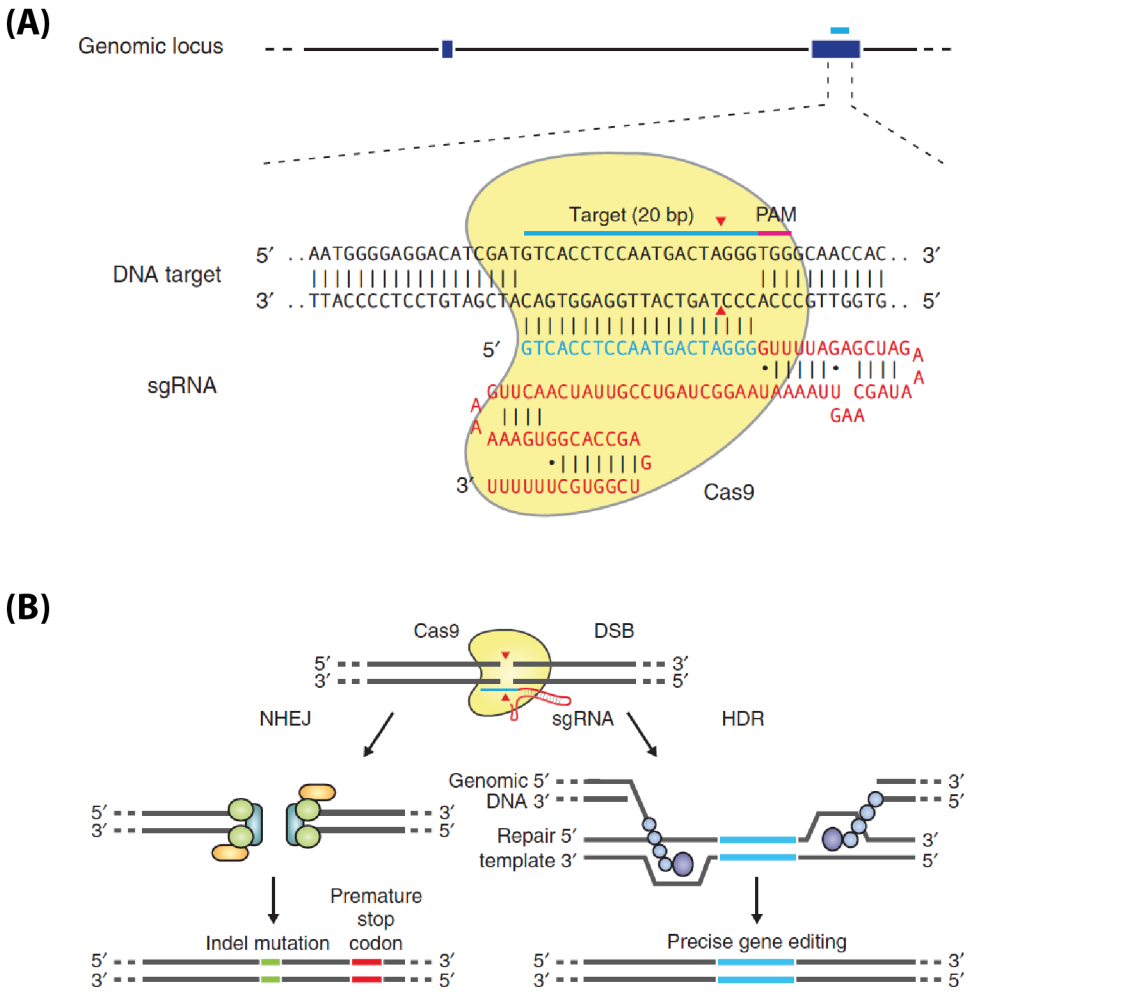


Figure 1.7: **Genome engineering using the CRISPR/Cas9 system. (A)** The gRNA is composed of the 20-nucleotides guide sequence (blue) fused with a scaffold (red). The guide sequence pairs with the DNA target (bleu bar), which is immediatly followed by a 5'- NGG PAM. It subsequently mediates the cleavage by Cas9 approximatively 3 base pairs upstream of the PAM (red triangle). **(B)** DSBs can be either repaired by NHEJ, resulting in indels (left), or, providing a repair template, by HDR, allowing precise gene editing (right). [90]

Following transfection, the cells can be clonally selected and eventually tested for the presence of indels (if no ssODN) or for the presence of the desired mutation if a ssODN has been used.

Remarks

The design of the guide RNA sequence is particularly critical for the success of the technique. Among other factors, one should try to minimize the off-target effects of the endonuclease. For this, an online CRISPR Design tool (http://tools.genome-engineering.org) has been created which greatly helps with this problem. Given an input sequence around the targeted site, it identifies suitable guide sequences (upstream of a 5'-NGG PAM) minimizing the off-target activities, and provides a list of predicted off-target effects [90].

Interestingly, the CRISPR/Cas9 system has been further used in a more general way, as a DNA targeting platform, expanding the scope of this technology. As Cas9 DNA-docking ability does not depend on it endonuclease activity, it has been possible to generate, by a single point mutation, a deactivated Cas9 enzyme (dCas9). Together with a gRNA, this dCas9 can still be targeted to DNA, but without cleaving it. If directed to a promoter region, binding of dCas9 can repress the transcription of the associated gene, by preventing RNA polymerase binding or elongation, or the recruitment of necessary transcription factors. Alternatively, dCas9 can be fused with any desired protein to generate a chimera protein, bearing the dCas9 site-specific docking activity together with the function of the protein of interest. This could be exploited to modulate gene expression in a more complex way, by fusing dCas9 with transcriptional activators, or repressors, or any chromatin-remodelling complex, and subsequently targeting the resulting factor to a specific promoter region with a gRNA. Thus, the CRISPR/Cas9 system can also be exploited to finely manipulate the expression of virtually any endogenous gene [93, 94].

Finally, the CRISPR/Cas9 genome editing allows multiplexing. Indeed, one can target mutiple genomic locations simulatenously by introducing several distinct gRNAs together. This enables high-throughput experiments with library-based approaches to perform screens [94].

1.4.2 The Flp-In™ T-Rex™ system to easily introduce an inducible gene construct in the genome of targeted cells

In addition to altering the endogenous version of a particular gene of interest (GOI), it can be of interest to introduce an ectopic gene contruct into cells. For instance, this can be used to complement the effect of a particular mutation with the wild-type (WT) allele. Alternatively, it can allow the expression of non-native genes, such as chimeras, or tagged genes (e.g. with GFP). Traditionally, this has been done by stable transfection, providing the cells with a DNA plasmid and selecting the ones having successfully integrated the foreign DNA into their genome. This integration event being very rare, however, the efficacy of this technique is extremely low. In addition, the site of insertion and the number of gene copies inserted are not controlled with this technique. The randomness of the insertion site is particularly problematic as the insertion can disrupt a gene sequence, or regulatory regions of a gene.

Principles of the Flp-In™ system

The Flp-In™ system provides an interesting solution to allow integration and expression of a unique copy of the GOI at a controlled genomic location [95]. It takes advantage of the *Saccharomyces cerevisiae*-DNA recombination system. It uses the Flp recombinase to induce DNA recombination at specific sites called FRT

(for Flp Recombination Target) sites present in both the host cell genome and the GOI expression vector. These sites are binding and cleavage sites for the recombinase. The Flp recombinase induces intermolecular DNA homologous recombination, which does not require DNA synthesis, thereby limiting the introduction of mutations.

Procedure for generating stable cell lines by Flp-mediated recombination

To create a Flp-In™ stable cell line, one has to go through the following steps (Figure 1.8) [95]:

- First, a host cell line containing an FRT site integrated into the genome, named as Flp-In™ host cell line, has to be generated. This can be done in the laboratory for virtually any cell line, by stable transfection of the appropriate plasmid. Alternatively, several Flp-In™ host cell lines have been generated by Invitrogen for various mammalian species (HEK293 and Jurkat human cells, 3T3 mouse cells, CV-1 monkey cells, CHO and BHK hamster cells), bearing a single FRT site into their genome at a controlled location. Using these cell lines is therefore a appreciable gain of time as it tremendously shorten and simplifies the creation of the Flp-In™ stable cell lines.

- Second, aforementioned host cells are co-transfected with a Flp recombinase-encoding plasmid (named as pOG44) together with an expression vector (called pcDNA™5/FRT) containing the GOI flanked with an FRT site. Consequently, the expressed Flp recombinase mediates recombination between the FRT sites present in the pcDNA™5/FRT and in the host cell genome. This results into stable integration of the GOI into the genome of the host cell.

- Third, cells with a stable integration at the FRT site can be selected using hygromycin resistance. Indeed, the pcDNA™5/FRT vector contains a degenerated hygromycin-resistance gene which is recapitulated by a proper insertion event.

The previously-described system uses the pcDNA™5/FRT vector, in which the GOI's expression is controlled by the human CMV promoter. Consequently, once integrated into host cells, the GOI is constitutively expressed. In some instances, however, it can be of interest to turn on and off the expression of the GOI in a reversible fashion. For this purpose, the Flp-In™ system has been further developed into a Flp-In™ T-Rex™ system, in which the expression of the GOI is under the positive control of tetracycline.

Adding a layer of complexity: tetracycline-regulated expression of the gene of interest

The Flp-In™ T-Rex™ system allows the generation of stable cells with a tetracycline-inducible GOI, using the Tn10-encoded tetracycline (Tet)-resistance operon from *E.Coli* [96]. Unlike previously, the GOI will not

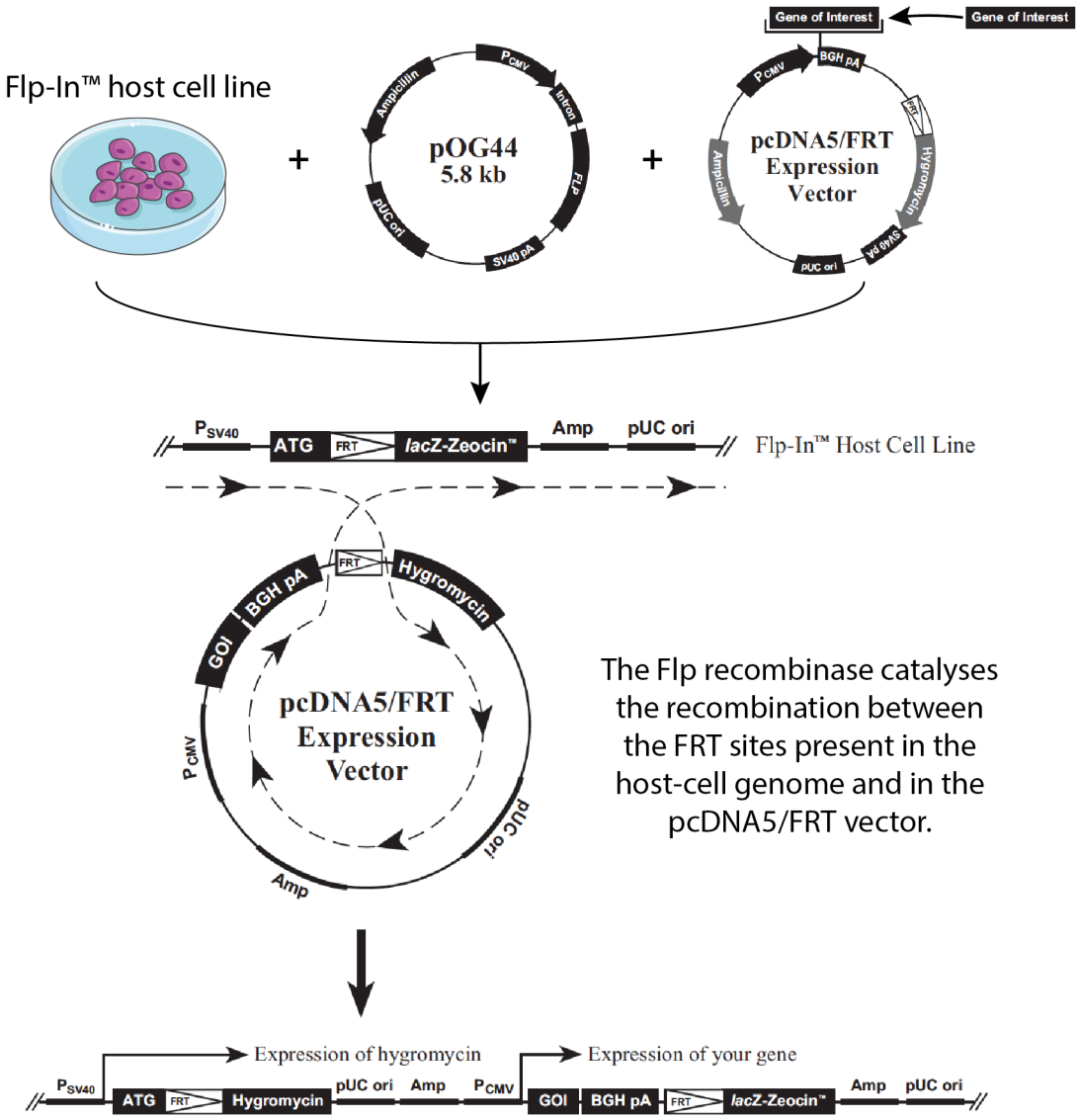


Figure 1.8: **Generation of stable cell lines using the Flp-In™ system.** The pcDNA™5/FRT expression vector containing the gene of interest (GOI) is co-transfected with the Flp recombinase-encoding vector (pOG44) into the Flp-In™ host cell line (top). The Flp recombinase mediates the recombination between the two FRT sites (middle), resulting in the integration of the expression construct into the genome of the host cell and expression of the GOI (bottom) [95].

be constitutively expressed, but rather constitutively repressed, due to the presence of 2 tandem copies of the *tet* operator ($TetO_2$) inserted into its CMV promoter, and the constitutive expression of Tet repressor by the host cell. In the absence of tetracycline, the Tet repressor expressed by the host cell homodimerizes and is recruited to the $TetO_2$ sites, thereby repressing the expression of the GOI (Figure 1.9, top). Upon treatment of the cells with tetracycline, tetracycline binds to the Tet homodimers (Figure 1.9, middle) and induces a conformation change that results into dissociation from the $TetO_2$ sites and removal of the transcriptional

repression (Figure 1.9, bottom). In summary, in the absence of tetracycline, the GOI is transcriptionnally repressed whereas, upon tetracycline treatment, it is expressed.

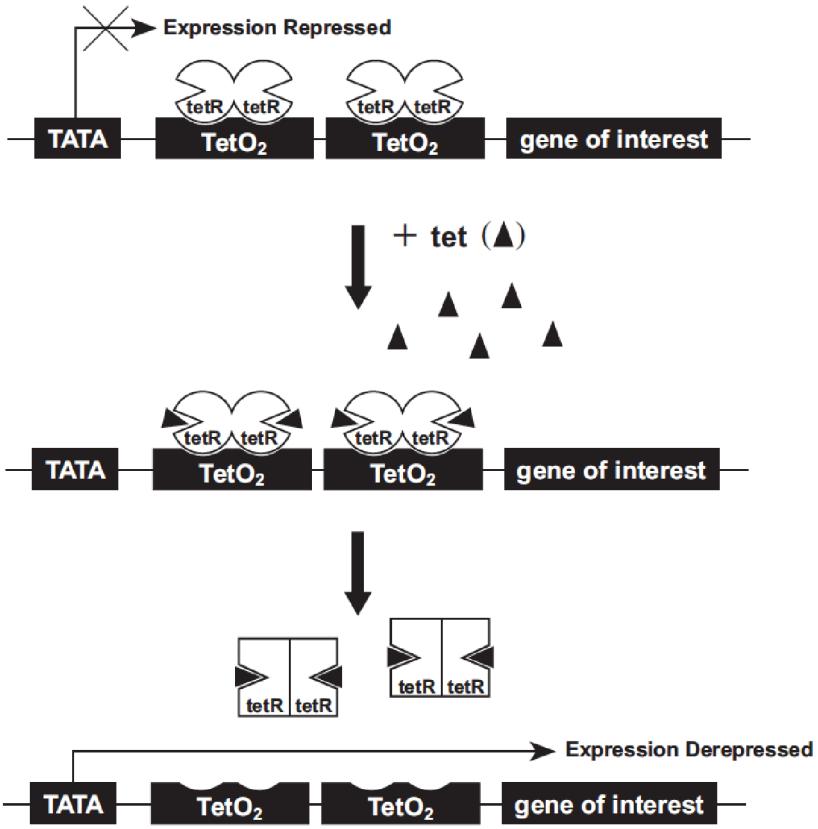


Figure 1.9: **Principle of tetracycline-mediated regulation.** The Tet repressor (tetR) constitutively expressed from the host cell homodimerizes and binds to $TetO_2$ sequences in the GOI promoter, resulting in repression of its expression (top). When added, tetracycline (tet) binds to the tetR homodimers (middle). This causes a conformational shift in the tetR homodimers and eventually their release from the $TetO_2$ sequences, and subsequent derepression of the GOI (bottom) [96].

The following points are changed compared to the previously-described Flp-In™ system:

- The host cell line needs to be further modified to stably encode the Tet repressor. As before, this can be made in virtually any cell line of interest, but the Flp-In™ T-Rex™ HEK-293 cell line has already been developed by Invitrogen.
- The GOI needs to be cloned into the pcDNA™5/FRT/TO vector, which bears the two $TetO_2$ sites in the promoter controlling the expression of the GOI.

The rest of the procedure stays the same than previously described for the creation of Flp-In™ stable cells.

Additional remarks

Because tetracycline antibiotics inhibit protein synthesis in both human and bacterial cells, it has the potential to kill the treated cells. But, the concentrations used to induce the expression of the GOI are usually too low to be toxic of the cells. Nevertheless, one should keep in mind the potential toxic effect of the treatment with tetracycline, especially if used at high concentrations. Classically, the tetracycline antibiotic used is doxycycline as it is more stable.

Of note, most cells are kept in culture supplementing their medium with fetal serum. This latter serum, however, often contains some tetracycline because it has been generated from animals fed with tetracycline-containing diet. Consequently, the use of such serum in cell culture would lead to a basal GOI expression, instead of a clean repression [96]. In case a cleaner repression of the GOI is needed (for instance, if it produces a toxic compound), one should consider of using tetracycline-depleted serum that is commercially available.

1.5 Thesis content

By reviewing the literature at the onset of my doctoral studies, I noticed that the THAP family of proteins had been barely studied as a whole, which I felt would provide an interesting research opportunity. Indeed, I felt that a global analysis of these proteins as a family, and not as separate entities, had chances to reveal interesting features. In addition, growing evidence implicates THAP proteins as broad regulators of gene transcription and resulting cell proliferation, often working in collaboration with HCF-1. Nevertheless, even at this writing, the importance of the THAP family is still emerging and most human THAP proteins remain understudied. In particular, tools to study them are often lacking.

Thus, I aimed to establish experimental tools and to use them to probe the role of human THAP proteins in the transcriptional control of cell proliferation alongside HCF-1, both by examining these proteins as a complete set, and then by more carefully investigating some specific members.

The goals of this doctoral thesis are the following:

- Consider the THAP proteins as a family, thereby unraveling some of their shared and specific features.

  The objective of Chapters 3 and 4 is to use bioinformatics and available data to respectively study how *THAP* genes have evolved during evolution and to assay their expression. Also, in Chapter 5, I describe biochemistry analyses performed on selected THAP proteins to examine how they interact, both with themselves and with HCF-1.

- Create genetic tools to study specific THAP proteins.

As mentioned above, very few tools are available so far to study the human THAP proteins, which restrains the discoveries about their funtions. Chapter 6 describes how I have engineered cell lines using recent genetic technologies, to question specific THAP proteins: THAP7 and THAP11.

- Probe the role of THAP7 and THAP11 in cell proliferation and gene transcription.

  In Chapters 7 and 8, I describe the phenotypic analyses performed on the aforementioned cell lines. Chapter 7 focuses on impacts on cell proliferation, and Chapter 8 explains how I have used genomic approaches to probe the role of THAP7 and THAP11 in regulation of transcription.

This study sheds light on how THAP proteins, in collaboration with HCF-1, can regulate gene transcription and subsequently cell proliferation. In addition, the established experiment tools can be beneficial for other researchers to explore the function of THAP proteins.

Chapter 2

Materials and methods

Here, I describe the experimental procedures of the expriments presented in this thesis.

2.1 Plasmids and site-directed mutagenesis

Plasmids WT *THAP* open reading frames were purchased from OriGene and GeneCopoeia and then cloned into the pcDNA^{TM}5/FRT/TO vector (Invitrogen V6520-20) with an T7-Flag tag or an HA tag at their C-termini. The pOG44 plasmid was from Invitrogen (V6005-20). The HCF-1-encoding plasmids are made from the pCGN vector and encode N-terminally HA and c-myc-tagged HCF-1 versions. They come from the Herr laboratory and have been deposited in the Addgene plasmid repository: pCGN-HCF-1$_{FL}$, pCGN-HCF-1$_{N1011}$ (HA-HCF-1$_N$) and pCGN-HCF-1$_{C600}$ (HA-HCF-1$_C$).

Site-directed mutagenesis Partially overlapping primer pairs were designed as recommended [97]. A polymerase chain reaction (PCR) was performed with 100 ng of the template plasmid, 0.2 µM of each forward and reverse primer, 200 µM of dNTPs and 2.5 units of Pfu turbo Taq polymerase (Agilent 600250-52) in 1X Pfu buffer supplemented with Quick Solution (Agilent 200516-51). The mix was incubated in the thermocycler for the PCR, with an extension time of 1 minute per kilobase (kb) of template plasmid. At the end of the PCR reaction, the mix was incubated with 20 units of DpnI enzyme (New England Biolabs R0176S) to digest the (methylated) template plasmid. The mutagenesis mix was then used to transform competent bacteria under carbenicillin selection (100 µg/µL, Carl Roth 6344.1). DNA was extracted from the bacteria and sequenced to assess the presence of the desired mutation in the plasmid.

2.2 Maintenance of cells in culture

Adherent cells were routinely maintained in DMEM medium (DMEM + 4.5g/L D-Glucose, L-Glutamine, Pyruvate; Gibco 419666) supplemented with 10% of heat-inactivated fetal calf serum (FCS, BioConcept 2-01F30) at 37 °C and 5% CO_2 in humidified atmosphere. For some experiments, the latter serum was replaced by tetracycline-free calf serum (FCS Tetracycline Free, BioConcept 2-01F28). In some proliferation assays, cells are cultivated at 39.5 °C.

HeLa-S cells (S for spinner) were grown in suspension. They were cultivated in JMEM medium (JMEM powder — MEM Joklik Modification with L-Glutamine without $CaCl_2$ — BioConcept 1-29P02-W) supplemented with 10% of heat-inactivated fetal calf serum and 1% of Penicillin-Streptomycin mix (10 000 units/mL Penicillin, 10000 µg/mL Streptomycin, Gibco 15140) at 37 °C and 5% CO_2 in humidified atmosphere, under constant agitation.

2.3 Cell transfection for biochemistry analyses

For THAP-protein interactions analyses as described in Chapter 5, cells were seeded in 10 cm dishes at 4×10^6 cells per dish 18 hours before transfection to allow their attachment. For the transfection, 6 µg of total DNA (3 µg of each plasmid in case of two plasmids) was resuspended into 250 µL of serum-free Opti-MEM medium (Gibco 31985). Separately, 10 µL of Lipofectamine® 2000 (Invitrogen 11668-019) was diluted into 250 µL of serum-free Opti-MEM medium. After 5 minutes incubation at room temperature, the Lipofectamine®-medium mix was added drop by drop to the DNA-medium mix. After incubating for 20 minutes at room temperature, the DNA-Lipofectamine® mix was added drop by drop onto the cells. The cells were then further incubated for 28 hours before harvesting for analysis.

2.4 Co-immunoprecipitation

Cells were on-plate lyzed in 500 µL per plate of 0.5% NP40 lysis buffer (10 mM Tris pH 8.0, 150 mM NaCl, 5 mM $MgCl_2$, 0.5% NP40, supplemented with one tablet of complete EDTA-free Protease Inhibitor Cocktail (Roche 04693132001) per 50 mL). From the lysate, 40 µL was transferred into a fresh tube for the input sample, mixed with 10 µL of 5X Laemmli buffer (250 mM Tris pH 6.8, 500 mM βmercaptoethanol, 50% glycerol, 10% SDS, with bromophenol blue) and heated at 95 °C for 5 minutes. The rest of the lysate was incubated overnight at 4 °C with 30 µL of monoclonal anti-HA (Sigma A2095) or monoclonal anti-Flag (Sigma A2220) agarose beads. After 4 washes of the beads with the 0.5% NP40 lysis buffer, the beads were resuspended in 20 µL of 5X Laemmli buffer, heated at 95 °C for 5 minutes and the supernatant was used for

western blotting (IP sample).

2.5 Western blotting

The samples were separated by SDS-PAGE before being transferred onto a nitrocellulose membrane. For samples after co-immunoprecipitation, an equal volume of input and IP samples were loaded (5 µL, then ratio input over IP is 1/30). Membranes were blocked for 1 hour in 100% blocking buffer (LI-COR Bioscences 927-40000). Membranes were incubated overnight at 4 °C with the primary antibody diluted 1/1000 in a mix of 50% PBS 0.5% Tween 20 and 50% blocking buffer. They were then washed with PBS 0.1% Tween 20 before being incubated with the appropriate secondary antibody during 1 hour at room temperature. Blots were finally visualized with the Odyssey® infra-red imaging system (LI-COR).

2.6 CRISPR/Cas9 mutagenesis

Design Using the online CRISPR Design tool (http://tools.genome-engineering.org), I selected for each mutation a suitable gRNA fullfilling the following criteria:

- being immediatly followed by a 5'- NGG PAM sequence;
- being 20-nucleotides long if it already has a G at its 5', or 21-nucleotides long adding an extra G nucleotide at its 5' if it does not have one;
- minimizing the off-target activities, meaning having a high score on the CRISPR Design tool output;
- close enough to the mutation site, more precisely the cutting site, which is 3 nucleotides upstream to the 5'- NGG PAM sequence, should be not farer than 11 nucleotides away from the mutation site;
- the PAM sequence is disrupted by the mutation or, alternatively, can be disrupted by a silent mutation.

The different gRNAs used in this thesis are listed on the Table 2.1. Please note that the target gRNA can be designed either on the sense or the antisense strand and that both forward and reverse sequences were synthesized to create a double strand gRNA. Before synthesis, I added upstream to the forward sequence a 5'- CACC sequence and a 5'- AAAC upstream of the reverse one. The resulting forward and reverse gRNA sequences, consequently partially overlapping, were synthesized by Microsynth. The extra nucleotides subsequently enable for cloning of the annealed gRNA sequences into the pSpCas9(BB)-2A-GFP backbone plasmid (Addgene plasmid 48138) using the BsbI restriction enzyme. This gRNA-expression plasmid, designed by

	gRNA			
THAP7$_{null}$	5′-	CACC G	CGCCGCCGGCTGCTGCACAC	-3′
	3′-	C	GCGGCGGCCGACGACGTGTG	CAAA -5′
THAP7$_{HBM}$	5′-	CACC G	CCAGAATGAACACAGCTACC	-3′
	3′-	C	GGTCTTACTTGTGTCGATGG	CAAA -5′
THAP7$_{\Delta CC}$	5′-	CACC	GCAGCGCCTTACTCTGGAAG	-3′
	3′-		CGTCGCGGAATGAGACCTTC	CAAA -5′
THAP11$_{null}$	5′-	CACC G	CAACAACTCGCACCGGGACA	-3′
	3′-	C	GTTGTTGAGCGTGGCCCTGT	CAAA -5′
THAP11$_{HBM}$	5′-	CACC G	CTGACGACAAGGAGTACGAA	-3′
	3′-	C	GACTGCTGTTCCTCATGCTT	CAAA -5′
THAP11$_{\Delta CC}$	5′-	CACC G	CTTGTCGTCAGGCACCACGG	-3′
	3′-	C	GAACAGCAGTCCGTGGTGCC	CAAA -5′
THAP11$_{F80L}$	5′-	CACC	GCTCATTGACGCCGCGCAGC	-3′
	3′-		CGAGTAACTGCGGCGCGTCG	CAAA -5′

Table 2.1: **List of gRNA used for CRISPR/Cas9 mutagenesis.** Forward and reverse sequences designed to create the double strand gRNA that was subsequently cloned into the pSpCas9(BB)-2A-GFP backbone plasmid. Red, core guide sequence; Blue, 5'- G added if not already present in the 20-nucleotides guide sequence; Green, extra nucleotides added for BsbI-mediated cloning into pSpCas9(BB)-2A-GFP.

Ran and colleagues [90], encodes the invariant gRNA scaffold and cloning sites for insertion of the guide sequence, together with the Cas9 nuclease and a GFP cassette for selection purposes.

The repair templates were designed as followed: from the mutation sites, 80 nucleotides upstream and downstream were taken to create an approximately 160-nucleotides long (depending on the number of nucleotides mutated). Table 2.2 lists the different single-stranded DNA oligonucleotides (ssODNs) used, which were synthesized by IDT.

Mutagenesis Cells were co-transfected with the pSpCas9(gRNA)-2A-GFP plasmid encoding for the gRNA, the Cas9 nuclease and GFP together with the appropriate ssODN repair template. For this, low-passage HEK-293 cells were seeded in 6 cm plates at 0.4×10^6 cells per plate, and incubated for 24 hours. The following day, cells were transfected as follows: 500 ng of pSpCas9(gRNA)-2A-GFP plasmid and 1 µL of the appropriate ssODN (10 µM) were diluted into 250 µL of serum-free Opti-MEM medium (Gibco 31985). Separately, 4 µL of Lipofectamine® 2000 (Invitrogen 11668-019) were diluted into 100 µL of serum-free Opti-MEM medium. After 5 minutes of incubation at room temperature, the Lipofectamine®-medium mix was

	ssODN
THAP7$_{null}$	5′- TGCCCGGAGAGCCGCTTGCGACTTAACTCCCGCCTCTTTCCCAGATG CCGCGTCACTGCTCCGCCGCCGGCTGCTGCACATGATAGACGCGCGAG ACGCGCAACCGCGGCATCTCCTTCCACAGGTCAGCGCGCGTGCGCCGC GGGCTCACGTGCGCATGCGCTAG
THAP7$_{HBM}$	5′- CTCTCGAACCACGGCCAGTCTCCCCCTCAGCGTATATGCTGCGCCTG CCCCCACCCGCCGGAGCCTACATCCAGAATGAAGCCAGCGCCCAAGTG GGCAGCGCCTTACTCTGGAAGCGGCGAGCCGAGGCAGCCCTTGATGCC CTTGACAAGGCCCAGCGCCAGCTGCAGGC
THAP7$_{\Delta CC}$	5′- ATATGCTGCGCCTGCCCCCACCCGCCGGAGCCTACATCCAGAATGAA CACAGCTACCAGGTGGGCAGCGCCTTACTCTGGTAGTAGCGAGCCGAG GCAGCCCTTGATGCCCTTGACAAGGCCCAGCGCCAGCTGCAGGCCTGC AAGCGGCGGGAGCAGCGGCTGC
THAP11$_{null}$	5′- TGGGCCGGGCCGGGCCGCGCGGCGCAGCCATGCCTGGCTTTACGT GCTGCGTGCCAGGCTGCTACAACAACTCGCACCGGTAGTAAGCGCTGC ACTTCTACACGTTTCCAAAGGACGCTGAGTTGCGGCGCCTCTGGCTCA AGAACGTGTCGCGTGCCGGCGTCAG
THAP11$_{HBM}$	5′- GGCTGGAGGCTGCCGAGTGCCCTATGGGCCCCCAGTTGGTGGTGGT AGGGGAAGAGGGCTTCCCTGATACTGGCTCCGACGCTTCGGCCTCCTT GTCGTCAGGCACCACGGAGGAGGAGCTCCTGCGCAAGCTGAATGAGC AGCGGGACATCCTGGCTCTGATGGAAG
THAP11$_{\Delta CC}$	5′- AGTTGGTGGTGGTAGGGGAAGAGGGCTTCCCTGATACTGGCTCCGA CCATTCGTACTCCTTGTCGTCAGGCACCACGGAGTAGTAGCTCCTGCGC AAGCTGAATGAGCAGCGGGACATCCTGGCTCTGATGGAAGTGAAGATG AAAGAGATGAAAGGCAGCATT
THAP11$_{F80L}$	5′- CCCACCACAGGCCACCGTCTCTGCAGCGTTCACTTCCAGGGCGGC CGCAAGACCTACACGGTACGCGTCCCCACCATCTTGCCGCTGCGCGGC GTCAATGAGCGCAAAGTAGCGCGCAGACCCGCTGGGGCCGCGGCCGC CCCGCCGCAGGCAGCAGCAGC

Table 2.2: **List of ssODN repair templates used for CRISPR/Cas9 mutagenesis.** The mutated residues are depicted in red in each sequence.

added drop by drop to the DNA-medium mix. After incubating for 20 minutes at room temperature, the DNA-Lipofectamine® mix was added drop by drop onto the cells and the cells were further incubated for 3 to 4 days.

Cells were then sorted using the FACS Aria II instrument to select for the transfected cells only, meaning the GFP-expressing ones. At the exit of the cell sorter, cells were immediatly diluted at one cell per well in 3 96-well plates, to allow single-cell cloning. In addition, the remaining GFP-positive cells were collected and

grown altogether for 2 to 3 days before being manually plated in 3 96-well plates at an average density of 0.5 cell per well. Also, the same procedure was done in parallel for GFP-negative cells as a control.

Screening Typically 3 weeks after the cell sorting and the single-cell seeding, the media of some wells starts to turn yellow, indicating that the cell in this well has survived and is reaching confluency. Each well was then splitted into two: two third of its content was used for DNA extraction and subsequent screening, while the rest was maintained in culture. DNA was extracted as follows: cells were resuspended in 15 µL of basic solution (25 mM NaOH, 0.2 mM EDTA, pH 12.0) and incubated in a thermocycler, first during 15 minutes at 68 °C, second during 30 minutes at 98 °C and third cooled down at 4 °C. Finally, 12 µL of acid solution (40 mM Tris, pH 5.0) was added and the extracted DNA was quantified.

The PCR was performed by mixing approximately 100 to 200 ng of the extracted DNA, with 0.4 µM of each forward and reverse primers, 200 µM of dNTPs and 0.4 units of high fidelity Q5 DNA polymerase (New England Biolabs M0491L) in 1X Q5 buffer. The PCR was done with appropriate annealing temperatures and extension times, which were optimized for each primer pair. Table 2.3 lists the pairs of primers used for these PCRs with their corresponding PCR conditions.

Following the PCR, half of the reaction mix was used for the enzymatic digestion with the appropriate restriction enzyme. For this, half of the PCR mix was incubated with 4 units of the appropriate enzyme, in 1X reaction buffer (which depends on the enzyme). Table 2.4 lists the different enzymes used for each reaction, together with the incubation conditions (buffer, temperature and duration). Following the digestion, PCR products and corresponding digestion products were loaded side by side on a 2% agarose gel. When a clone displays the expected digestion pattern for the mutation, PCR products were extracted from the gel by excising the gel portion and incubating it overnight in 40 µL of 1M Tris, pH 6.8. The following day, the DNA released in the Tris buffer was submitted to sequencing using one of the forward or reverse primer to confirm the presence of the mutation.

2.7 Stable cell lines

Transfection Flp-In™ T-Rex™ HEK-293 cells (Invitrogen R78007) were seeded in 10 cm dishes at 3×10^6 cells per dish 18 hours before transfection to allow their attachment. Cells were then transfected with the THAP-T7-Flag pcDNA™5/FRT/TO plasmid construct together with the pOG44 plasmid containing the gene encoding for the Flp recombinase. Here, 3 µg of the pOG44 plasmid plus 0.313 µg of THAP-T7-Flag pcDNA™5/FRT/TO plasmid were transfected as described previously for the biochemistry analyses and the cells were further incubated for 24 hours. Cells were then trypsinized and split to 3 new 10 cm dishes at

	PCR primers	PCR conditions
THAP7$_{null}$	fwd 5′- TCAAGAGAATCGGCTGGGAC rvs 5′- CGAGGCGAGCAACAACTAGC	59.8 °C 6 seconds 2.5% DMSO
THAP7$_{HBM}$	fwd 5′- GCCTTAGCAGCCCCTTTTCAG rvs 5′- TCAGTCTCAACCGCAGCCG	57 °C 6 seconds
THAP7$_{\Delta CC}$	fwd 5′- CGGAGCCTACATCCAGAATGAAC rvs 5′- TCAGTCTCAACCGCAGCCG	58.6 °C 6 seconds
THAP11$_{null}$	fwd 5′- CGCAGCCATGCCTGGCTTTACG rvs 5′- CCGCCCTGGAAGTGAACGCTGC	53 °C 7 seconds
THAP11$_{HBM}$	fwd 5′- CGGAGTTACAGGCTGCTACC rvs 5′- CTTCTCACGCAGTTCTTCGC	55 °C 6 seconds
THAP11$_{\Delta CC}$	fwd 5′- GCGTCGGAGTTACAGGCTG rvs 5′- CCTTCTCACGCAGTTCTTCG	53.2 °C 6 seconds
THAP11$_{F80L}$	fwd 5′- CGTTTCCAAAGGACGCTGAGTTGC rvs 5′- GCTGTTGCTGCTGCTGCCTGCG	70.6 °C 6 seconds

Table 2.3: **List of the PCR primer pairs used to screen the cell clones obtained after CRISPR/Cas9 mutagenesis, and their specific working conditions.** fwd, rvs: forward and reverse primers, respectively.

the following dilutions: 1/4, 1/10 and 1/20. Cells were incubated for 24 hours to allow their attachment.

Antibiotic selection and generation of single-cell clones Cells were treated with the hygromycin-B antibiotic to select for those having integrated the THAP gene construct. For this, the medium was removed from the plate and replaced with fresh medium supplemented with hygromycin-B at 100 µg/ml (LabForce HygroGold ant-hg-1). Cells were further incubated at 37 °C and the growth of colonies was monitored regularly. During this time, medium was replaced twice a week by fresh medium supplemented with 100 µg/ml of hygromycin. When the colonies are well formed, which typically took 10 to 12 days after the first addition of hygromycin, 6 colonies were randomly picked from the plates to be further expanded and eventually tested as described in the results. Cells were maintained under hygromycin selection at 100 µg/ml when cultured outside experimental procedures, while antibiotic selection was removed during the course of experiments.

	Enzyme	Buffer	Temperature	Duration
THAP7$_{null}$	NlaIII	CutSmart buffer	37°C	O/N
THAP7$_{HBM}$	AluI	CutSmart buffer	37°C	1 hour
THAP7$_{\Delta CC}$	AciI	CutSmart buffer	37°C	1 hour
THAP11$_{null}$	AgeI	CutSmart buffer	37°C	1 hour
THAP11$_{HBM}$	CviQI	NEBuffer 3.1	25°C	1 hour
THAP11$_{\Delta CC}$	SacI	CutSmart buffer	37°C	1 hour
THAP11$_{F80L}$	FauI	CutSmart buffer	55°C	1 hour

Table 2.4: **Restriction enzymes used to screen the cell clones obtained after CRISPR/Cas9 mutagenesis, and their specific working conditions.** O/N, overnight.

Doxycycline stimulation To stimulate the expression of the ectopic *THAP* construct, cells were treated with doxycycline (doxycycline hyclate Sigma D9891-1G) at 1 µg/mL final, except otherwise specified, for 24 hours or longer. The doxycycline having been resuspended using dimethyl sulfoxide (DMSO) as solvent, the same amount of DMSO was used to treat cells in parallel for the unstimulated-cell control.

2.8 Cell proliferation assays

Cells were seeded (day 0) in wells of 6-well plates at 25000 cells per well and incubated for 24 hours at 37 °C to allow their attachment. For each condition, cells were plated in duplicate. After 24 hours (day 1), if appropriate, cells were moved to temperature (39.5 °C) or treatment was performed (doxycycline, or DMSO as control). Cells of two different well per condition were then counted on day 1, and every day from day 4 to day 8. Resuts were dispayed as the ratio between the mean cell number of the two replicates at the time point (Nt), and the cell number initially seeded on day 0 (No).

2.9 Cell synchronization

For this experiment, I used Hela-S cells (S for spinner) which grow in suspension.

Double thymidine block Cells were prepared as 0.3×10^6 cells/ml in 400 mL of medium and immediatly blocked with 2 mM of thymidine (Sigma T1895) during 12 hours. After an 11-hour release in normal medium, cells were blocked again for 12 hours with 2 mM of thymidine. They were subsequently released in normal medium to let them progress synchronously through the cell cycle.

Flow cytometry for DNA-content analysis Time points were collected every two hours during 24 hours. Here, 0.5×10^6 cells were taken off from the pool of cycling cells and immediatly fixed by resuspending them in 200 µL of ice-cold PBS 1X and adding drop by drop 800 µL of 100% ice-cold ethanol. Cells were then stored at -20 °C for at least 12 hours before being analysed. The DNA of the fixed cells was stained with propidium iodide (PI, Sigma P4864-10mL) by resuspending the cells in 250 µL of PI-staining buffer (50 µg/mL of PI, 50 µg/mL of RNAse A (DNAse-free RNAse A, Roche 1119915), diluted in water) and incubating them for 20 minutes at 4 °C in the dark. As control, cells from asynchronous cells as well as unstained cells were also prepared. Cells were finally analyzed using the PE-A channel of a LSR II flow cytometer (BD Biosciences). The results were visualized using the FACSDiva Version 6.1.1 software.

2.10 High throughput RNA sequencing (RNA-seq)

Preparation of cells For the analysis of stable cells, 0.25×10^6 cells per well were seeded in 6-well plates and treated 24 hours later with 1 µg/mL of doxycycline, or the same amout of DMSO, as control. Cells were further incubated for 36 hours at 37 °C before being analyzed.

For the analysis of the mutant cells, 0.15×10^6 cells per well were seeded in 6-well plates and incubated at 37 °C for 24 hours. If necessary, some plates were switched to 39.5 °C. Cells were further incubated for 48 hours before being analyzed.

RNA extraction and sequencing Total RNA was extracted using the Qiagen RNeasy kit (Qiagen 74104) according to the manufacturer protocol. The RNA was eventually eluted in water and quantified. The samples were then entrusted to the UNIL Genomic Technology Facility (GTF) for quality control with a fragment analyser. Ribosomal RNA was removed using the Ribo-zero rRNA removal kit (Illumina MRZH11124), then libraries were prepared using the truSEQ stranded RNA LT kit (Illumina) and 50-nucleotides single-read high-throughput sequencing was performed. The results were subsequently analyzed by Dr. Viviane Praz, biotechnician in our laboratory.

2.11 Chromatin immunoprecipitation followed by high-throughput sequencing (ChiP-seq)

Chromatin preparation Cells were expanded at 37 °C and approximately 150 millions of cells were used per IP. Cells were on-plate crosslinked with 1% of formaldehyde (Sigma F1635-500ml) during precisely 8 minutes, then 0.125 M of Glycine (Axonlab A1067.5000) was added to terminate the crosslinking reaction. Cells were washed twice with cold PBS 1X and lyzed for 10 min on ice in 0.5% NP40 lysis buffer (5 mM PIPES pH 8.0, 85 mM KCl, 0.5% NP40, supplemented with one tablet of complete EDTA-free Protease Inhibitor Cocktail (Roche 04693132001) per 50 mL; 950 µL per 10 million of cells). The nuclei were recovered by high-speed centrifugation (5 minutes at 3200 xg, at 4 °C), resuspended in nuclei lysis buffer (NLB, 50 mM Tris-HCl pH 8.1, 10 mM EDTA pH 8.0, 1% SDS, supplemented with one tablet of complete EDTA-free Protease Inhibitor Cocktail) and incubated for 20 minutes at 4 °C.

Chromatin was sonicated using a Bioruptor Pico (Diagenode) to obtain fragments of around 200 bp. Sonicated chromatin was clarified by centrifugation and subsequently 1:2 diluted in 2X IP buffer (33.4 mM Tris pH 8.1, 167 mM NaCl, 167 mM LiCl, 2.4 mM EDTA pH 8.0, 2.2% Triton X-100, 0.02% SDS, supplemented with one tablet of complete EDTA-free Protease Inhibitor Cocktail) before being snap frozen in liquid nitrogen. A 50 µL aliquot was kept to analyze the chromatin quality and concentration. The frozen sonicated chromatin was stored at −80 °C.

Chromatin immunoprecipitation The day prior to the chromatin immunoprecipitation, protein G agarose beads (Roche 1243233) were washed with NLB:IP buffer (a mix of equal amounts of NLB and 2X IP buffer) and further incubated overnight under rotation at 4 °C) in NLB:IP buffer supplemented with 100 µg/mL of Bovine Serum Albumine (BSA, Sigma A8022-100).

The sonicated chromatin was thawed. A 60 µL aliquot was kept for the input sample, while, for each IP, 9 µg of chromatin (in a total volume of 1200 µL of NLB:IP buffer) was overnight incubated under rotation at 4 °C with 2 µg of anti-THAP11 antibody. Samples were further incubated under rotation at room temperature with 60 µL of the above washed and BSA-blocked protein G agarose beads. Immunoprecipitated samples were washed twice with IPWB1 buffer (IP wash buffer 1: 20 mM Tris pH 8.1, 50 mM NaCl, 2 mM EDTA pH 8.0, 1% Triton X-100 and 0.1% SDS). Each IP was performed 5 times in parallel as described and subsequently pooled to form a single IP sample at this step, after the washes with IPWB1 buffer. Pooled IP samples were then washed once with IPWB2 buffer (IP wash buffer 2: 10 mM Tris pH 8.1, 250 mM LiCl, 1 mM EDTA pH 8.0, 1% NP40 and 1% Na-deoxycholate) and finally twice with TE buffer (10 mM Tris pH 8.1, 1 mM EDTA pH 8.0). Two elutions with IPEB buffer (elution buffer: 100 mM $NaHCO_3$, 1% SDS) were

sequencially done (5 minutes at 37 °C under agitation) and pooled.

The input sample was thawed and supplemented with 190 µg of IPEB buffer. Eluates (IP) and input samples were decrosslinked by overnight incubation at 65 °C in presence of 20 µg/mL of RNAse A (DNAse-free RNAse A, Roche 1119915) and 300 mM of NaCl. The samples were further incubated during 90 minutes at 45 °C with 350 µg/mL of proteinase K (Promega V3021). Samples were finally purified using the "Nucleospin Gel and PCR clean-up" kit (Macherey-Nagel 740609) using the NTB buffer for SDS-containing samples. Samples were eluted with 50 µL of the pre-warmed (72 °C) buffer NE (from the kit), and subsequently re-eluted using the previous eluate. DNA yield was quantifed using a Qubit spectrophotometer.

Library preparation and sequencing For each IP sample, two separate libraries were prepared and sequenced, using the same immunoprecipitated-DNA material. For this, 5 ng of purified DNA was used to prepare paired-end sequencing libraries using the "MicroPlex Library Preparation" kit (Diagenode C05010014) following the manufacturer instructions. Here, 8 PCR cycles were done for DNA amplification and the DNA fragments were not size selected. Then, libraries were purified using AMPure® XP magnetic beads (Diagenode) and entrusted to the UNIL Genomic Technology Facility (GTF) for 100-nucleotide paired-end high-throughput sequencing (Illumina, HiSeq 2100) with 3 samples per line (multiplexing). The results were subsequently analyzed by Dr. Viviane Praz.

2.12 Antibodies

The antibodies used were: anti-HA (rat, Roche 10768600), anti-Flag (rabbit, Cell Signalling 2368S), anti-HCF-1 (rabbit, H12, [98]), anti-OGT (rabbit, Santa Cruz Biotechnology sc-32921), anti-tubulin (mouse, Sigma T0198) and anti-THAP11 (sheep, R & D Systems AF5727).

2.13 Bioinformatics

Analysis of THAP proteins The list of THAP proteins in the selected animal species was done by combining the list from Clouaire et al. [46] and the Pfam database [99]. Duplicates, but with different annotations, were identified by performing numerous Protein Blast analyses [100] and subsequently removed. Each putative THAP protein was finally manually verified so it indeed contains a correct N-terminal THAP domain as described in section 1.3.2.

The presence of an HBM sequence in the THAP proteins was assessed using Protein Patter Find [101], providing a fasta file with the THAP protein sequences as a query, and looking at consensus HBM sequences

($^{B}/_{Z}$HxY, where B and Z are aspartate/asparagine or glutamate/glutamine, respectively, and x denotes any amino-acid) using the following search pattern: "[DNEQ]H.Y".

The presence of a coiled-coil domain in the THAP proteins was assessed with two separate tools: COILS [102] and PairCoil2 [103]. Fasta sequences of the THAP proteins were used as queries, with default parameters, except for the p-score cut-off which was set to 0.05. Both tools gave identical results in terms of presence or absence of a coiled-coil domain, even though precise coiled-coil domain boundaries may slightly differ.

Evolutionary trees were done using the Mobyle portal [104] to perform multiple alignment (Muscle alignment), by providing an input fasta file containing the sequences to be aligned. This was done either on the full-length THAP proteins, or only on their THAP domains. The alignment was refined using the "protein boostrap distance phylogeny (alignment)" option (random number seed = 3; random number seed for multiple dataset = 7). The output of the alignment was subsequently uploaded on the iTOL (interactive Tree Of Life) online tool [105, 106] for visualization.

RNA-seq analysis The high-throughput RNA sequencing data were analyzed by Dr. Viviane Praz. Single reads were mapped onto the Hg19 human genome annotation using STAR (Spliced Transcripts Alignment to a Reference, [107]) and read counts and normalized RPKM (Reads Per Kilobase of transcript per Million mapped reads) were calculated using RSEM [108, 109]. Genes with an RPKM value below 1.2 in all the experiments were considered as not expressed and discarded of further analyses. Differential analyses were performed with DESeq2 [110] (fold-change cutoff = 0.5, adjusted pvalue = 0.05). Resulting gene sets were submitted to Gene Ontology (GO) enrichment analysis [111,112], summarized and visualized using REVIGO [113]. PCAs were done with the prcomp function of the R software, with no scaling or normalization.

ChIP-seq analysis The data obtained from the high-throughput DNA sequencing following the chromatin immunoprecipitation were analyzed by Dr. Viviane Praz. Paired-end reads were mapped onto the Hg19 human genome annotation using STAR (Spliced Transcripts Alignment to a Reference) [107]. Each fragment end was mapped individually allowing a maximum of 10 multiple genomic matches. Only the equivalent multiple matches (in terms of match length and mismatch numbers) were kept. The sequencing pairs were then reassembled with in-house scripts.

Peaks were detected using the Model-based Analysis of ChIP-Seq (MACS2) tool [114] (format = BEDPE, qvalue cutoff = 1.00e-03, duplicates allowed = 3, Broad region calling = off) and tested using a custom method developed by Dr. Viviane Praz, which identified enriched bins with a so-called "Origami method" [115]. Only MACS-identified peaks intersecting with the Origami enriched bins were kept. Peaks were further classified into 3 categories: (i) peaks present in both WT and THAP11$_{F80L}$ samples; (ii) peaks present only in the

WT sample; (iii) peaks present only in the THAP11$_{F80L}$ sample. Intersecting peaks in the different samples were represented by a Venn diagram using the Meta-Chart website (https://www.meta-chart.com). A peak was said close to a TSS if at least one nucleotide of its underlying DNA region locates within +/- 250 bp of a Pol2 transcription start site.

Peak scores were calculated as follows: (i) peak fragments of the IP sample were summed on the whole peak region, and normalized to total fragments in the sample and peak width (bringing all peaks width at a standard 1 kb); (ii) the same procedure was done for the input sample; (iii) for each sample, the peak score was calculated as the $\log_2$ of the normalized IP counts (done in (i)) minus the input normalized counts (done in (ii)). For calculation purposes, one pseudocount was added to each of the IP and input normalized counts.

Sequence comparison with known motifs was done on regions expanding 500 bp on each side of the peak middle using CentriMo [116].

The motifs under the THAP11 peaks were further analyzed to identify subtle differences in the consensus sequences. First, the peaks were redefined by taking the peak summits (meaning, the highest position of the peaks) as a reference (instead of the peak centers) and extending the positions 250 bp on each side. Second, the ZNF143 consensus motif from Hocomoco (reference ZN143_HUMAN.H11MO.0.A, here 22-bp long) was screened over the whole genome using the PWMTools web interface [117] to extract all the genomic positions containing such motif, as well as the correponding 22-bp DNA sequence. Third, sequences intersecting between the list of peaks and the genomic screen described (the latter having produced a list of more than 9 millions of sequences) were extracted. If more than one motif was identified for a given peak, only the one closest to the peak center was considered. Fourth, a consensus sequence was generated using the list of pre-aligned 22-bp long sequences and a logo was made. The number of THAP11-associated motifs per peak was defined by counting, for each peak, the number of motifs in a region expanding 1000 bp on each side of the peak maximum. The scores were defined as the difference of the IP and input log2 counts, scaled by total tags in sample and peak width.

Peaks close to TSS (+/- 5 kb) were visualized with the UCSC genome browser [118] in which wiggle files prepared by Dr. Viviane Praz were uploaded, all tracks being set with the same vertical viewing range (1 to 600).

Chapter 3

Evolution of THAP proteins in animal species

When Roussigne and colleagues [40] reported the discovery of the THAP domain, they found *THAP* genes in several animal species, but not in plants, yeast, fungi or bacteria. They therefore postulated that the THAP domain is restricted to animals. Here, I use bioinformatics tools to dig inside the landscape of THAP proteins in animal — and also potentially non-animal — species.

3.1 The THAP family within animal species

Here, I digged into the landscape of THAP proteins in selected animal model species.

3.1.1 THAP proteins in animal and non-animal species

Using the list provided by Clouaire and colleagues [46] and the Pfam database [99], I have counted all the THAP proteins in selected animal model species (Figure 3.1): *Homo sapiens* (human, 12), *Mus musculus* (mouse, 7), *Gallus gallus* (chicken, 7), *Xenopus tropicalis* (frog, 27), *Danio rerio* (zebrafish, 45), *Drosophila melanogaster* (fruit fly, 12), *Caenorhabditis elegans* (round worm, 8). I only considered as putative THAP proteins those having indeed an N-terminal THAP domain, with the appropriate characteristics (see section 1.3.2). In addition, I removed as much as possible redundant annotated sequences that are very likely to belong to the same protein. I observed that, while in 5 out of the 7 species, the number of THAP proteins remains in the same order of magnitude, this number has exploded in the zebrafish (almost 4-times more THAP proteins compared to human) and, even to a lesser extent, in *Xenopus tropicalis*. In addition, the Pfam

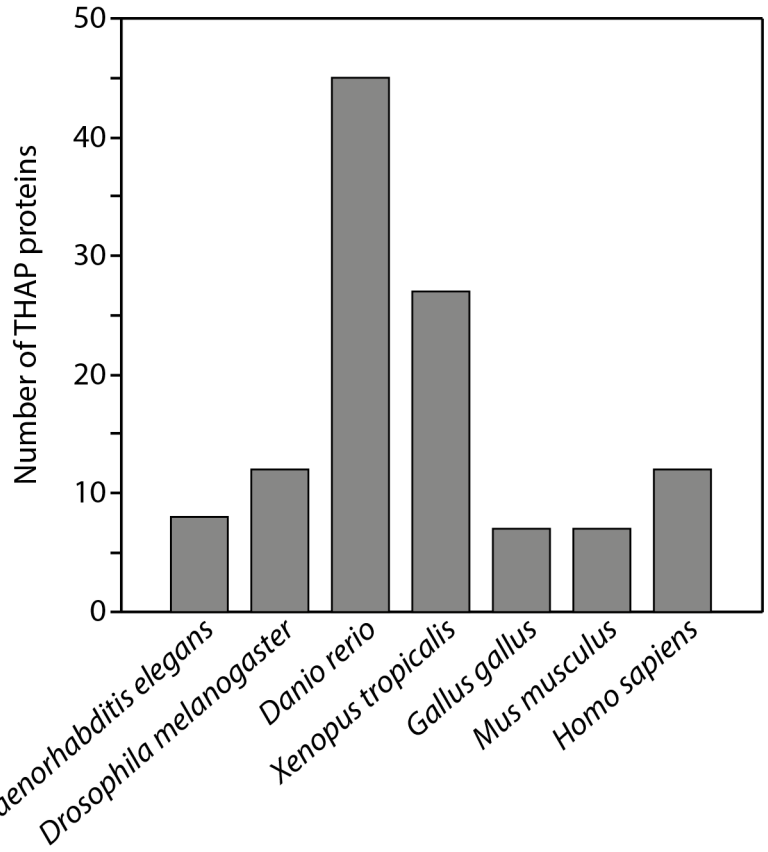


Figure 3.1: **Number of THAP proteins in animal species.**

database does not list any non-animal THAP proteins. Also, Protein Blast [100] on a a list of 122 animals, 99 protists and 48 plants, using the human THAP proteins as queries, did not reveal any non-annotated THAP proteins in either animal or non-animal species. This also confirms the fact that the THAP domain is animal specific.

3.1.2 Orthologs of human THAP proteins

Using the list of THAP proteins obtained and the gene annotations available, I investigated the presence of orthologs of the human THAP proteins in the 6 other selected species.

First, the accuracy of each already-annotated ortholog was verified by performing Protein Blast on the human-protein database using the annotated ortholog as a query, and on the selected-species protein database using the human THAP protein as a query. This confirmed all of the orthologous relationships already annotated. In addition, when there was not any ortholog for one particular THAP protein in a species, I assessed whether it is due to a lack of annotation or to a real absence of ortholog. For this, I ran Protein Blast on the selected-species protein database using the human THAP protein as a query. But, this analysis

did not reveal any additional orthologs than the ones already annotated in the databases.

I exploited a complementary method to verify differently the orthologous relationships and identify new ones. For this, I used the Mobyle portal (Mobyle@Pasteur) [104] and the iTOL (interactive Tree Of Life) online tool [105,106] to build phylogenic trees. Each tree was made by mixing the human THAP proteins and the THAP proteins from a single other species (data not shown). Again, it confirmed the already-annotated orthologous relationships without revealing any new ones.

Table 3.1 summarizes the presence (green plus) or absence (red minus) of human THAP orthologs in the selected animal species. To begin with, no ortholog of human THAP proteins has been detected in *Drosophila melanogaster* and *Caenorhabditis elegans*. In addition, some human THAP proteins have been well conserved from zebrafish to human, such as THAP0, 1, 4 and 7. In contrast, other human THAP proteins, such as THAP8 and THAP10, seem to have appeared very recently in evolution. Some conservation patterns are less easy to interprete, such as THAP2 and THAP6, because they appear in distant species but in not proximal species on the evolutionary tree. An hypothesis is that the THAP orthologs would have appeared independently in different species after speciation; it is, however, very unlikely due to the poor statistical chances that it happens. Alternatively, the *THAP* gene may have been present in anscestor species, but subsequently lost in some species after speciation, due to the divergence and decay phenomenon explained in section 1.1.3. To test this hypothesis, one could look for related pseudogenes in the genome of the considered species. Also, I noted that THAP11 is conserved in every species, but in the chicken (*Gallus gallus*). It is well known that the *Gallus gallus* genome is poorly assembled. Consequently, the presence or absence of ortholgs using genome sequences is poorly reliable in this species. Thankfully, another bird has been sequenced and assembled with much better quality: the Tibetan ground tit (*Pseudopodoces humilis*, a bird from the Tibetan plateau north of the Himalayas) [119]. I thus assessed the presence of human THAP orthologs in this species. As in *Gallus gallus*, I found orthologs for human THAP0, 1, 4, 5, 7 and 9, and not of human THAP2, 3, 6, 8 and 10. But, unlike in *Gallus gallus*, I also found an ortholog of human THAP11. This suggests that there is probably an ortholog of human THAP11 in birds, and that its absence in the *Gallus gallus* genome is due to the poor genome assembly at this species. For the rest of this study, I therefore consider the THAP11 protein sequence from *Pseudopodoces humilis* as the avian model (Table 3.1, purple plus).

3.1.3 Relationships between human THAP proteins

I then analyzed the relationships between the 12 different human THAP paralogs.

First, while the THAP proteins form a family of structurally-related proteins, they are quite diverse. Even the THAP domain, which gives them their identity, is less than 50% identical among each other.

	Homo sapiens	*Mus musculus*	*Gallus gallus*	*Xenopus tropicalis*	*Danio rerio*
THAP0	+	+	+	+	+
THAP1	+	+	+	+	+
THAP2	+	+	—	+	—
THAP3	+	+	—	—	+
THAP4	+	+	+	+	+
THAP5	+	—	+	+	—
THAP6	+	—	—	—	+
THAP7	+	+	+	+	+
THAP8	+	—	—	—	—
THAP9	+	—	+	+	+
THAP10	+	—	—	—	—
THAP11	+	+	+	+	+

Table 3.1: **Orthologs of human THAP proteins.** Green, presence of an ortholog; red, no ortholog; purple: probable ortholog found in another avian species (see text for details).

This divergence can explain why the different THAP domains do not recognize the same DNA sequences. Considering the full-length proteins, this percentage drops to around 30-35% for most pairs of human THAP proteins. This diversity can also be noted just when comparing the features of each THAP member, as seen in Figure 1.5. This suggests that, despite sharing some features and perhaps some functions, each THAP protein may have its own specifity.

Second, I produced an evolutionary tree as decribed above, to uncover the relationships between the different proteins of the human THAP family. Figure 3.2 shows the 12 human THAP proteins, sorted by sequence similarity of their THAP domain. It suggests the existence of protein couples, such as THAP2 and 10, THAP0 and 11, THAP1 and 3 or THAP5 and 8. Curiously, three of these four pairs associate an HBM-containing protein with a non-HBM protein. This observation raises the intringing hypothesis of duplications events and subsequent divergence which could have created such couples, one potentially interacting with HCF-1, while the second not.

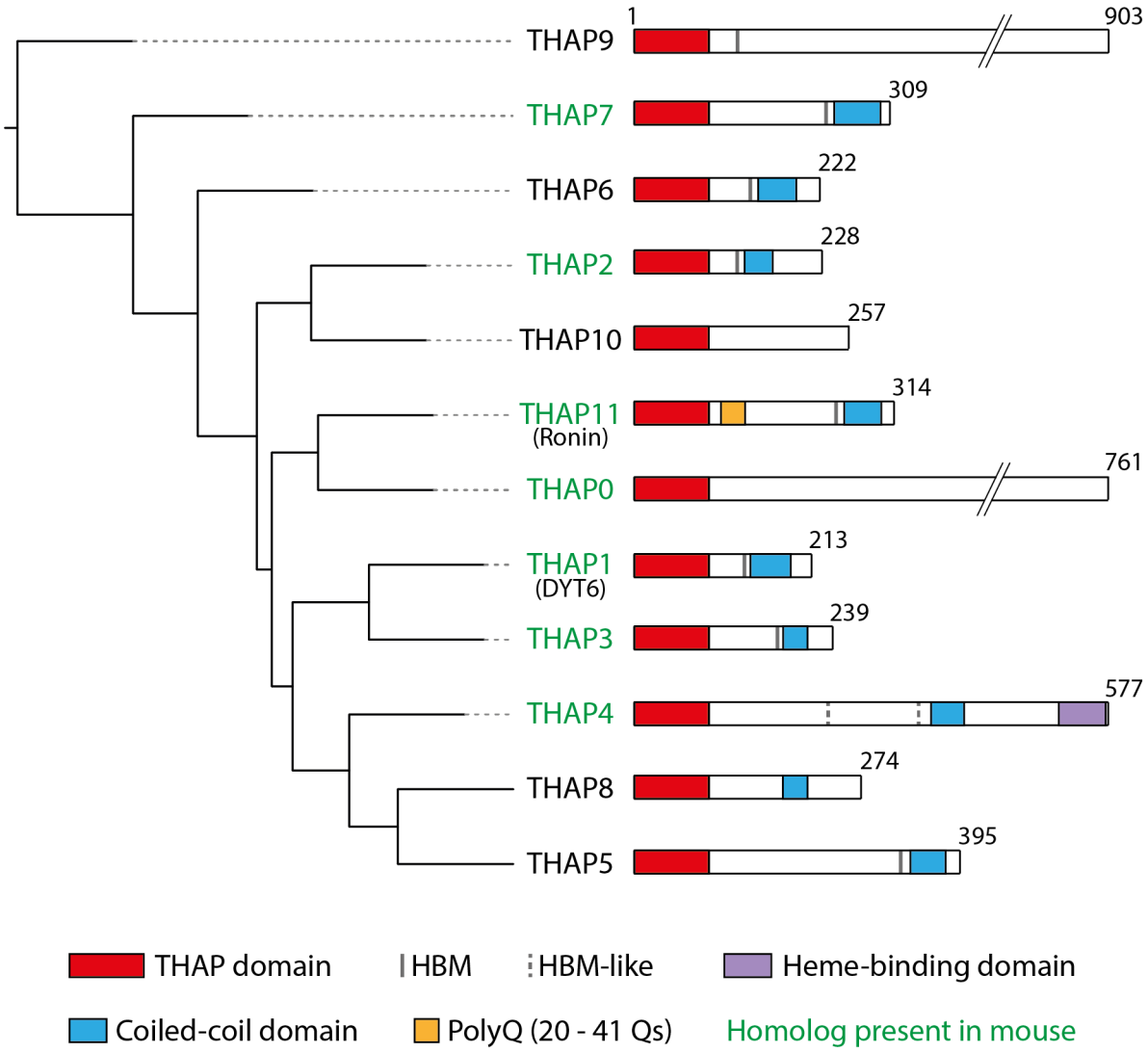


Figure 3.2: **Human THAP proteins.** The 12 human THAP proteins are sorted by sequence similarity of their THAP domain.

3.2 The HBM sequence in animal THAP proteins

I have previously explained that 9 of the 12 human THAP proteins display an HBM (*i.e.*, HCF-1 Binding Motif), and that HCF-1 seems to be an important co-factor in THAP-mediated activities (see 1.3.4). Also, HCF-1 is conserved in metazoans (Gonzalez, Praz and Herr, personal communication). Thus, I globally investigated the presence of the HBM sequence among animal THAP proteins.

3.2.1 Conservation of the HBM sequence during the evolution of the THAP proteins

The presence of the HBM sequence in each animal THAP protein listed in Figure 3.1 was explored using Protein Pattern Find [101]. Only consensus HBM sequences ($^{B}/_{Z}$HxY, where B and Z are aspartate/asparagine or glutamate/glutamine, respectively, and x denotes any amino acid) were considered and the percentage of HBM-displaying THAP proteins in each species has been calculated (Figure 3.3). This percentage is pretty low in *Caenorhabditis elegans* (13%, 1 out of 8) and progressively increases in *Drosophila melanogaster* and *Danio rerio*. Then, from *Xenopus tropicalis* to *Homo sapiens*, this percentage remains stable and high (approximately 70%). This suggests that the interaction between HCF-1 and THAP proteins has changed during evolution and is strenghtened in species closer to human.

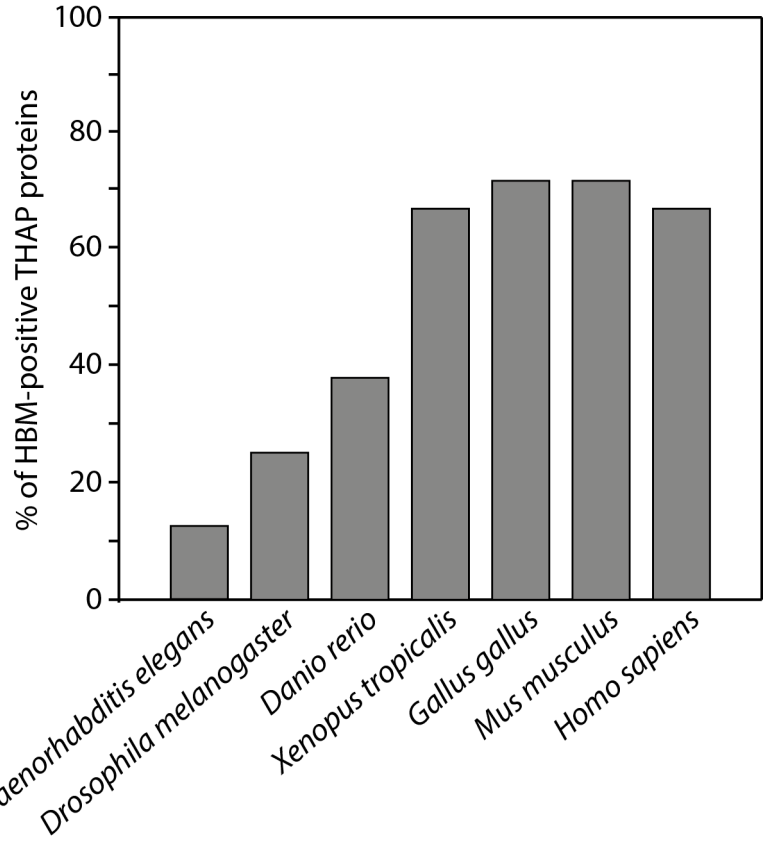


Figure 3.3: **Conservation of the HBM sequence during THAP-protein evolution.** The percentage of HBM-positive THAP proteins is displayed for each animal model species.

3.2.2 Conservation of the HBM sequence among orthologs of human THAP proteins

I then had a closer look at the conservation of the HBM sequence among the orthologs of human THAP proteins, as seen on Table 3.2. I observed that, when there are orthologs of human THAP proteins, the HBM presence or absence is well conserved. Indeed, when the HBM is present in the human proteins, it remains present at least until *Xenopus tropicalis*, and most of the time until *Danio rerio*. This is also true for the 2 HBM-like sequences from human THAP4. Also, the HBM consensus sequence allows flexibility. I noted that, while in some cases the HBM sequence is strictly conserved (e.g. THAP1 or 7), in other cases the 4 amino-acid sequence has changed during evolution, yet still respecting the consensus sequence (e.g. THAP2 where the first amino-acid position is a glutamic acid in *Homo sapiens*, and a aspartic acid in *Xenopus tropicalis*). In addition, the only human HBM-negative THAP proteins having orthologs in the chosen model species (THAP0) remains HBM negative in these species. More particularly, it is curious to realize that, among the 3 HBM-negative THAP proteins, a single one (THAP0) has orthologs in the chosen model species. It may indicate that the HBM sequence is a selected element in evolution.

To conclude, the HBM sequence appears to be a critical element in THAP proteins that is very well conserved during evolution, likely being the indication of a functional importance.

3.3 The coiled-coil domain of THAP proteins

The ability to form dimers is often important for the function of transcription factors, as mentionned in section 1.1.2. Thus, the coiled-coil dimerization domain may have a particular functional relevance in THAP proteins.

3.3.1 Presence of the coiled-coil domain in THAP proteins

Gervais and colleagues have published that all the 12 human THAP proteins display a coiled-coil domain [42]. I have used two independent bioinformatics tools to predict the presence and boundaries of coiled-coil domains in proteins: COILS [102] and PairCoil2 [103]. Both prediction tools gave the same results: contrary to what was suggested [42], not all human THAP proteins display a coiled-coil domain, as no such domain was identified in THAP0, 9 and 10 (Figures 1.5 and 3.2, blue box). This suggests that the distinct THAP proteins might not have the same ability to homo- or heterodimerize. In addition, the human family of THAP proteins displays all the possibility of HBM/coiled-coil containing/lacking proteins. Thus, the HBM and coiled-coil characteristics of each THAP protein may be particularly critical for its specific action.

	Homo sapiens	*Mus musculus*	*Gallus gallus*	*Xenopus tropicalis*	*Danio rerio*
THAP0	•	•	•	•	•
THAP1	DHNY	DHNY	DHNY	DHNY	•
THAP2	EHSY	EHSY	—	DHNY	—
THAP3	DHSY	DHSY	—	—	DHTY
THAP4	*LHSY (x2)*	*LHSY (x2)*	*LHSY (x2)*	*LHSY (x2)*	•
THAP5	EHSY	—	EDSY	EHSY	—
THAP6	EHSY	—	—	—	DHSY
THAP7	EHSY	EHSY	EHSY	EHSY	EHSY
THAP8	•	—	—	—	—
THAP9	DHNY	—	DHTY	DHLY + DHTY	•
THAP10	•	—	—	—	—
THAP11	DHSY	DHSY	DHSY	DHSY	DHSY

Table 3.2: **The HBM sequence in human THAP-protein orthologs.** Green text, HBM sequence (consensus sequence $^{B}/_{Z}$HxY, where B and Z are aspartate/asparagine or glutamate/glutamine, respectively, and x denotes any amino acid); green dot, no HBM in this ortholog; red minus, no ortholog. Italic, HBM-like sequence. Please note that the THAP11 protein sequence from *Pseudopodoces humilis* was considered as the avian (*Gallus gallus*) model (see text for details).

Contrary to what I demonstrated above for the HBM sequence, no particular conservation pattern of the coiled-coil domain was observed when looking at the different orthologs of human THAP proteins (not shown). To have a more global idea of the importance of the coiled-coil domain in THAP proteins, I plotted, for each species I worked with, the percentage of coiled-coil domain-positive THAP proteins (Figure 3.4). It shows that this percentage is slightly higher in *Gallus gallus*, *Mus musculus* and *Homo sapiens* compared to other species. In *Danio rerio* and *Xenopus tropicalis*, the percentage drops. This can be explained, at least, by the higher number of THAP proteins in these species and an expansion of non coiled-coil orthologs. To conclude, the coiled-coil domain seems to be more conserved in species closer to humans, suggesting a selection pressure in THAP proteins to possess such a domain.

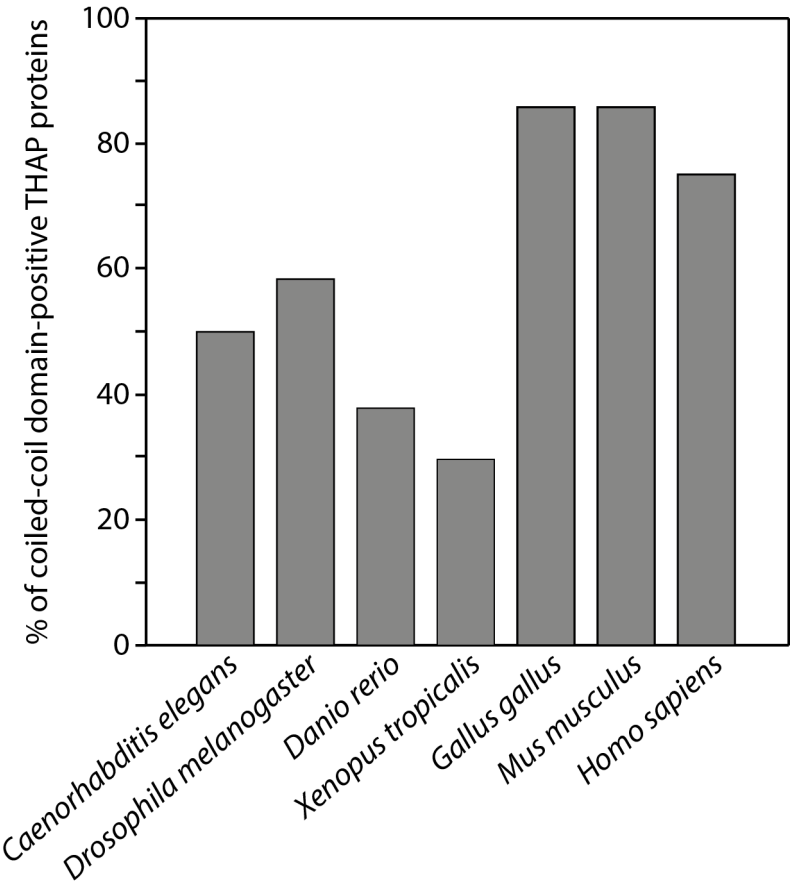


Figure 3.4: **Frequency of coiled-coil domains in THAP proteins.** The percentage of coiled-coil domain-positive THAP proteins is displayed for each animal model species.

3.3.2 Coiled-coil domains and HBM sequences

Curiously, when both an HBM sequence and a coiled-coil domain are present in the same THAP protein, they are in close proximity, the HBM being always just N-terminal (Figure 1.5). To investigate further the possibility of a link between these two domains, I assessed the distance between these two features in HBM and coiled-coil-positive THAP proteins, in all selected species (not shown). I realized that, in frog, chicken, mouse and human, the relative position of the domains remains the same and the distance between them is around 10 amino acids, even often less. But, in zebrafish, fly and worm, the distance slighlty increases. This suggests that the HBM sequence and the coiled-coil domain might be functionally linked to each other, at least in tetrapod species.

3.4 Discussion

In this chapter, I used bioinformatic tools to study features of the THAP proteins.

First, I confirmed that the THAP domain is animal specific, and listed and verified all the THAP proteins in 7 selected animal model species. One challenging problem was that, even though human and mouse genomes are extremely well annotated, it is not as well done for other species. Consequently, in *Gallus gallus*, *Xenopus tropicalis*, *Danio rerio*, *Drosophila melanogaster* and *Caenorhabditis elegans*, I obtained many "uncharacterized" as well as "predicted" proteins. This complicated the analysis as some sequences might be partial, or redundant with others, while corresponding to the same protein. Conversely, sequences can be uncorrectly annotated. These drawbacks can mistakenly "create" or "miss" proteins. I tried as much as possible to remove this bias, by carefully analyzing the sequences. In particular, I extensively performed Protein Blast analyses to identify sequences that are likely to belong to the same protein, to remove as much as possible artificial redundancy. I also verified the orthologous relationships already annotated with the human THAP proteins and tried to uncover new ones. In particular, THAP11 was surprisingly absent from the *Gallus gallus* annotations, but I showed that THAP11 is likely to be present in birds. The search for additional non-annotated orthologous relationships using either Protein Blast searches or phylogenic trees was unsuccessful. Nevertheless, these methods do not exclude the existence of yet-unnanotated orthologs of human THAP proteins in other species. To have a more complete view of the presence or absence of orthologs, one should also look at RNA-seq data and assess the presence of transcripts corresponding to human *THAP* proteins.

Finally, I had a closer look at the human THAP proteins and realized that, despite sharing the THAP domain and some additional features, the human THAP proteins sequences are pretty diverse. This can suggest two non-exclusive hypotheses:

- Despite their weak sequence similarity, the THAP proteins could still have similar functional domains and 3-dimentional structures, as we know it is already the case for the THAP domain, as seen in section 1.3.2. Thus, even being different at the amino-acid level, the 12 human THAP protein are likely to be structurally and functionaly related.

- Alternatively, this divergence may have led to THAP proteins having their own identity and specificity. As an example, the THAP domains of distinct human THAP proteins have similar 3-dimentional structures but recognize different DNA sequences (section 1.3.2). More generally, the sequence divergence of THAP proteins probably has led to the appearance of specific functions and characteristics for particular THAP members (*i.e.*, sub-functionalization). For instance, each THAP protein might have the ability to interact with a different set of co-regulators, of which some might be shared with other THAPs. Similarly, as distinct THAP proteins do not recognize the same DNA sequence, it is likely that each one regulates a different set of genes, resulting in different regulatory outcomes.

To conclude, the human family of THAP proteins displays similarities and differences, which can lead to both shared and specialized partners, functions and outcomes.

I also had a closer look at the HBM sequence, which may be of particular relevance for THAP proteins. Indeed, the HBM is present in most of human THAP members, and HCF-1 has been suggested to be a critical partner for THAP-protein functions (section 1.3.4). By probing the presence of the HBM motif in each of the listed animal THAP protein, I demonstrated that the percentage of HBM-containing THAP proteins has changed during evolution and is higher in species closest to the human one. Also, I showed that, among the orthologs of human THAP proteins, the presence or absence of the HBM sequence is relatively well conserved. Thus, it is likely that the HBM is under positive selection pressure in THAP proteins, suggesting its functional importance in this protein family. Together with the fact that both THAP and HCF-1 proteins are restricted to metazoans, it raises the interesting possibility of their co-evolution.

Furthermore, I investigated the presence and conservation of the coiled-coil domain in THAP proteins, as dimerization is of particular relevance in transcription factors (see section 1.1.2). I demonstrated that, contrary to what has been published, some THAP proteins do not bear any evident coiled-coil domain. This observation suggests that each THAP protein is likely to have a distinct capacity in dimerization, and also different protein partners. The coiled-coil domain does not show any particular pattern of conservation among human THAP orthologs, contrary to what has been shown with the HBM sequence. Nevertheless, looking overall at all THAPs suggests that this domain is under positive selection pressure in THAP proteins as well. Consequently, dimerization of THAP proteins might have a particular relevance for their functions. I also identified a possible connection between the HBM sequence and the coiled-coil domain. Indeed, when both are present in a single THAP protein, they are always close to each other, with the HBM being N-terminal of the coiled-coil domain. This curious feature is conserved from human to frog, while less obvious in more distant animals. This raises the intringing possibility of a functional link between these two domains and further, between HCF-1 binding and protein dimerization. Thus, the THAP coiled-coil domains appear to have an importance not only by themselves, but also in the context of an HCF-1 interaction.

It should be noted that this bioinformatics analysis was done at the beginning of my doctoral studies, nearly 5 years ago. As mentionned above, the sequencing annotations of some species were far from complete. But this is constantly improving and genome annotations are in permanent evolution. Consequently, some annotations used in my analyses may have changed since then, and consequently the list of proteins, especially for species other than human and mouse, may have also changed.

This *in-silico* analysis has revealed interesting features of THAP proteins, and raised attracting hypotheses on their functions and characteristics. Biochemical and functional experiments are the next step to verify these different hypotheses and uncover more precisely the relevance of the different domains, in particular the

HBM sequence and the coiled-coil domain. Indeed, it would be of particular interest to investigate further the functional role of these domains, which so far remains elusive, in future experiments. For instance, biochemical experiments would unravel whether the different THAP proteins bind to HCF-1, as well as the importance of the HBM sequence in the potential interaction. In addition, one could probe the potential of THAP proteins to form (homo- and hetero-) dimers and the contribution of the coiled-coil domain to the possible dimerization. Also, I suggested that each human THAP protein, although a member of a structurally-related protein family, may have a particular combination of characteristics which makes it unique. To test this idea, one could imagine to create THAP chimeras by combining portions of separate THAP proteins.

Another important aspect of proteins is to know their pattern of expression. Indeed, this can give insight into their role. To assess the levels of proteins, one can either use publicly-available data, or experimentally probe the protein or mRNA levels.

Chapter 4

Expression of *THAP* genes

Here, I explain how I probed the expression of human *THAP* genes in normal tissues and in cultured cells, as well as murine *Thap* genes in the regenerative liver.

4.1 Expression of human *THAP* genes in normal tissues

Here, I used available data to probe the expression of the different human *THAP* genes in normal tissues. A former group from our department performed extensive RNA sequencing in different species for 6 organs: brain, cerebellum, heart, kidney, liver and testis [120]. I took advantage of their data and analyzed the normalized expression levels of the *THAP* genes in human, expressed in RPKM (Reads Per Kilobase of transcript per Million mapped reads). In their analysis, only the genes displaying 1:1 orthology relationship for each pair of selected primates species were considered. Consequently, the expression values of 9 out of the 12 human *THAP* genes were available (*THAP1*, *2*, *4*, *6* to *11*), while *THAP0*, *3* and *5* were not. For brain, cerebellum, heart and kidney, one female and at least one male were analyzed; for liver and testis, two males (but not female — naturally, for testis) were used.

First, I compared the expression values of each human *THAP* gene available between the two sexes. The normalized expression values are similar between males and females in brain, heart and kidney, whereas in the cerebellum, males and female seem to express differently some *THAP* genes (not shown).

I then plotted the mean of the normalized expression values from the male individuals for each available human *THAP* gene and for the 6 organs. Choosing to concentrate on male individuals allowed me to probe more organs and to determine variation between individuals, as several males were used but only one female. As shown in Figure 4.1, some *THAP* genes are much more expressed than others, such as *THAP4*, *7* and *11*. In addition, while some genes display a homogenous pattern of expression in the different organs (e.g.

THAP11), some are differentially expressed according to the tissue (e.g. *THAP1*).

In conclusion, the differing expression patterns of human *THAP* genes suggest that each THAP member possesses its unique regulatory mechanism, which can be ubiquitous or tissue specific.

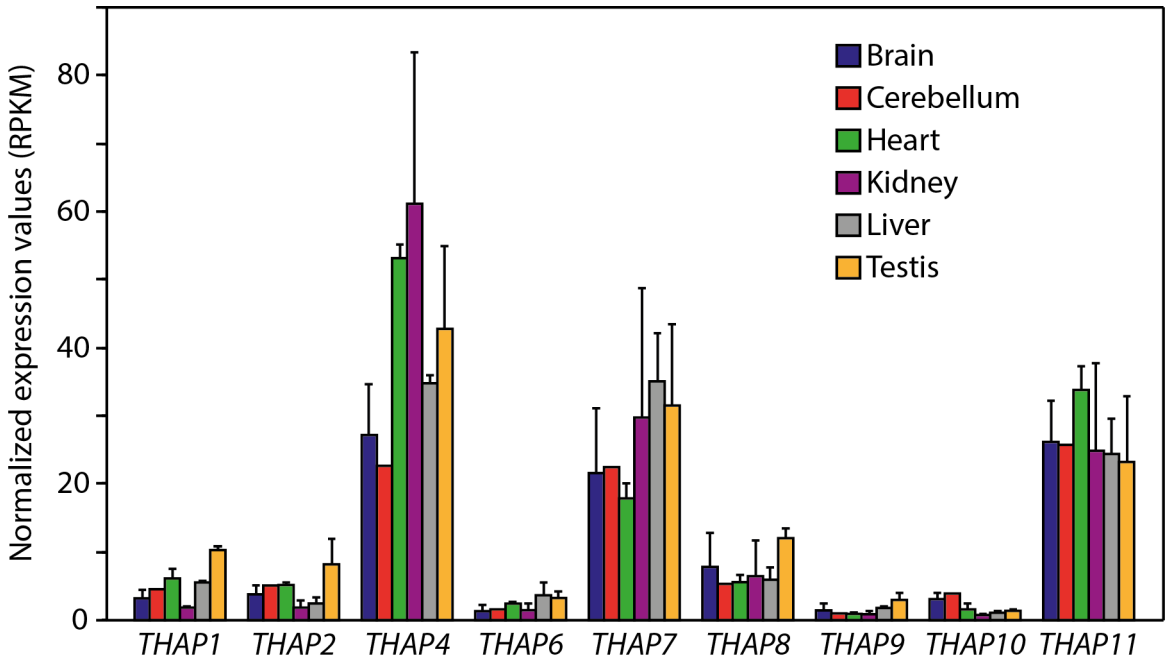


Figure 4.1: **Expression levels of *THAP* genes in human organs.** Mean +/- standard deviation of normalized RPKM expression values from males individuals. No standard deviation is shown for the cerebellum as only one male individual was used. Note that *THAP0*, *3* and *5* are missing from the histogram.

4.2 In cultured cells, most human *THAP* genes are not differently expressed along the cell cycle

Tissue-culture cells can be easily synchronized along their cell cycle using two rounds of thymidine block [121]. At the end of the second block, cells have accumulated at the G1-to-S phase transition. When released into normal medium, they progress through the cell cycle in a synchronous manner. As a result, one can obtain homogenous populations of cells at specific phases of the cell cycle.

4.2.1 The human THAP11 protein in synchronized cells

I was interested to probe the variation of THAP-protein levels during the cell cycle. The study of proteins, however, is restricted by the availability of antibodies. Thus, I was only able to probe the levels of the THAP11 protein, as it is the only THAP member for which I had a working antibody.

I performed double-thymidine block on HeLa-S cells grown in suspension. From the release in normal medium, I did a time course in which I analyzed every two hours the DNA content of the cells to assess their cell-cycle status. Figure 4.2 A shows the flow cytometry profiles of cells, for which their DNA was stained with propidium iodide. The P2 (left) and P4 (right) gates delimitate the diploid (2n) and tetraploid (4n) cells, respectively. The P3 (middle) gate delimitates the cells in between which are synthesizing DNA. Consequently, the P2, P3 and P4 gates contain the cells in G1, S and G2/M phases, respectively. The DNA profile of non-synchronized cells (top left) exhibits 48% of cells in G1 (P2), 17% of cells in S phase (P3) and 34% of cells in G2/M phases (P4). The other profiles show that cells are efficiently synchronized at G1-to-S transition prior release (86% of cells in P2), and then in S phase at 4 hours after release (67% of cells in P3), in G2/M phase at 8-10 hours (74% and 71% of cells in P4, respectively), in G1-early ($G1_E$) after 12 hours (82% of cells in P2) and in G1-late ($G1_E$) at 16 hours post relase (87% of cells in P2) [28, 121]. From 20 hours post release, the signal becomes blurred, indicating that the cells are losing their synchronization.

I also probed for the levels of cell-cycle protein markers: phospho-CDK1, cyclin A and cyclin B1. Their respective levels confirmed the identified correspondance between the time points and the cell-cycle phases (not shown).

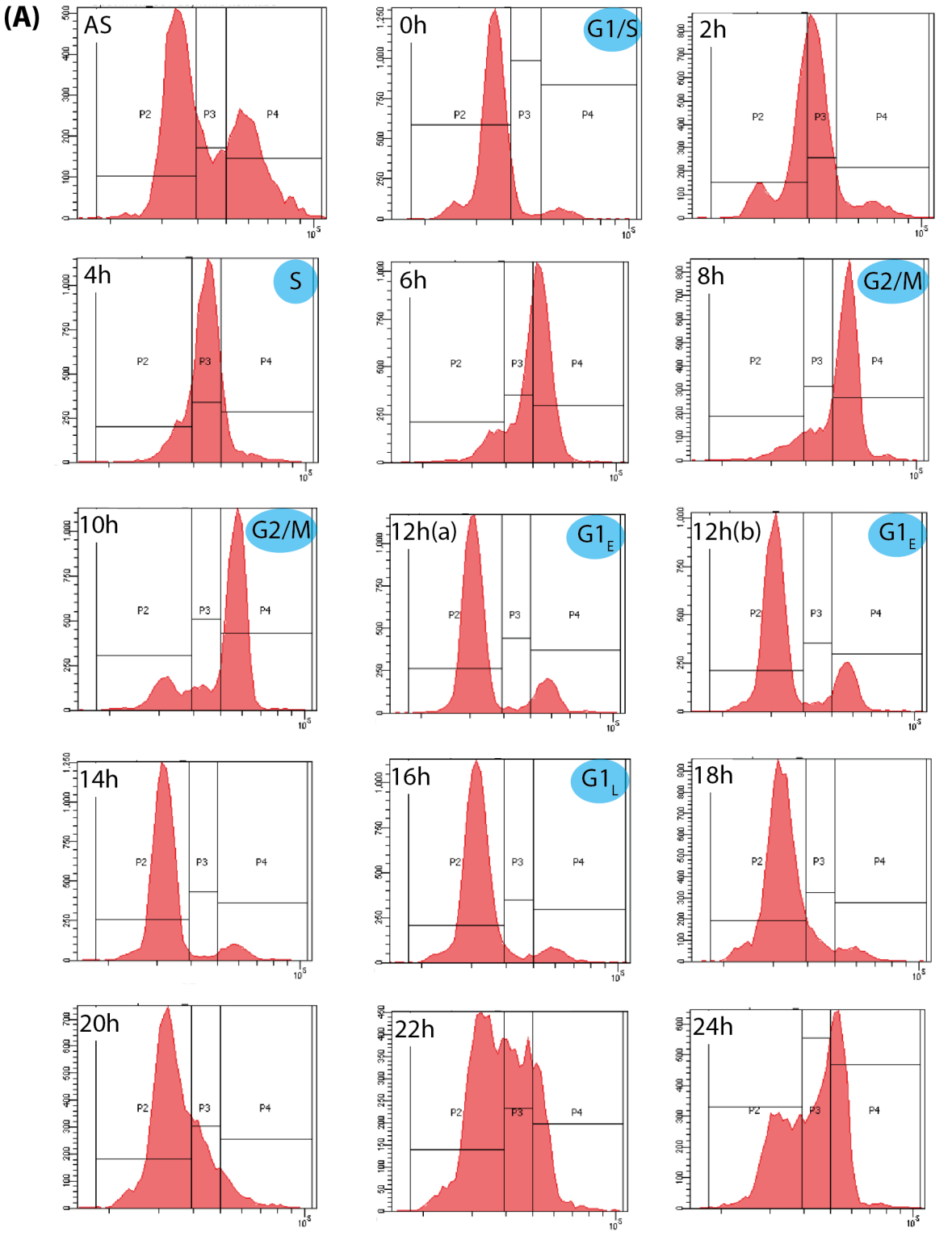
At each time point post thymidine release, I probed for THAP11 protein. Figure 4.2 B shows the result. The THAP11 protein levels were constant all along the course of the cell cycle, suggesting that its synthesis is not cell-cycle regulated.

4.2.2 Human *THAP* genes in synchronized cells

To examine the expression of the *THAP*-gene family more broadly, I studied the *THAP* mRNA levels during the course of the cell cycle, keeping in mind that mRNA and protein levels do not necessarily correlate.

Double-thymidine block was performed on HeLa-S cells grown in suspension as before. At different time points post release in normal medium, RNA was extracted from the cells and subsequently submitted to high-throughput sequencing (RNA-seq). This particular experiment was performed by Maykel Lopes, research technician, and analyzed by Dr. Viviane Praz, a lab bioinformatician. Figure 4.3 shows the mRNA levels of the corresponding *THAP* genes determined by the RNA-seq analysis, expressed in $\log_2$ of RPMK values. Thus, a 1 unit change on the $\log_2$ RPKM Y-axis scale represents a 2-fold mRNA level change.

Among the 12 *THAP* genes, only THAP8 is not expressed in HeLa-S cells, as shown by the mRNA-level curves below the cut-off (green bar). *THAP2* and *THAP3* are the only *THAP* genes that exhibit a notable variation of mRNA levels during the progression through the cell cycle: *THAP2* mRNA level progressively decreases while cells are progressing along the cell cycle, while *THAP3* mRNA level increases by almost 2.8


(A)
AS
0h
G1/S
2h
4h
S
6h
8h
G2/M
10h
G2/M
12h(a)
G1E
12h(b)
G1E
14h
16h
G1L
18h
20h
22h
24h
P2
P3
P4

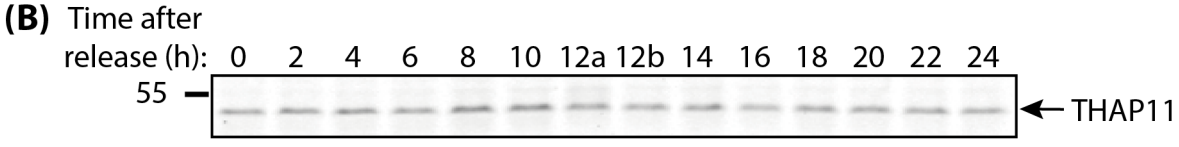

(B) Time after release (h): 0 2 4 6 8 10 12a 12b 14 16 18 20 22 24
55
THAP11

Figure 4.2: **Levels of human THAP11 protein along the cell cycle of cultured cells.** HeLa-S cells grown in suspension were synchronized at G1-to-S phase by double thymidine block and analyzed at different time points after release in normal medium. **(A)** Flow cytometry plots showing the DNA content of cells. P2, G1-phase cells; P3, S-phase cells; P4, G2 and M-phase cells (see text for details). AS, asynchronous cells. Blue circles, identification of the corresponding cell-cycle phases. **(B)** At each time point, the same amounts of whole cell lysates were probed for THAP11 protein. The 12-hour time point has been done in duplicate (12a and 12b).

fold ($\log_2$ RPKM change of 1.5) between 2 and 10 hours post release. The 9 other *THAP* genes, which are expressed, are not markedly changing along the cell cycle.

In conclusion, among the human *THAP* genes expressed in HeLa-S cells, all but *THAP2* and *THAP3* seem quite constant along the cell cycle. In contrast, *THAP2* and *THAP3* show a notable change in mRNA levels during the course of the cell cycle.

4.3 Expression of murine *Thap* genes during liver regeneration

Our laboratory has extensive experience with performing liver regeneration experiments [122,123]. Two-third partial hepatectomy in the mouse induces a highly coordinated regeneration in which remaining hepatocytes synchronously enter and progress through the cell cycle (Figure 4.4 A). This model is thus an excellent tool to study, *in vivo*, the expression of genes along the cell cycle. Two-third partial hepatectomy was performed on mice [122, 123]. At different time points post surgery, mice were sacrified, RNA extracted from the livers and submitted to high-throughput sequencing. In parallel, livers were used for ChIP-seq (Chromatin ImmunoPrecipitation followed by high-throughput sequencing) against the Pol2 polymerase (Pol2), to correlate the expression of genes with the presence of the polymerase at their promoter. This experiment was part of Dr. Leonor Rib's PhD thesis [123], who extracted for my purpose the expression levels of the murine *Thap* genes and the presence of Pol2 at their promoters or in their gene bodies. Figure 4.4 B shows that the 7 murine *Thap* genes have a remarkable homogenous expression during the progression through the cell cycle (yellow lines, mRNA levels). Similarly, the abundance of Pol2 at the gene promoter (brown lines, Pol2 promoter) and in the gene body (pink lines, Pol2 transcription unit) does not really vary either.

Thus, most of these results are consistent with the previous observation made in synchronized HeLa-S cells. Indeed, mouse *Thap0*, *1*, *4*, *7* and *11* mRNA levels are not affected by the progession along the cell cycle, similarly to human genes. By contrast, mouse *Thap2* and *3* mRNA levels are not changed during the cell cycle, while the corresponding human genes exhibit a decrease and an increase, respectively.

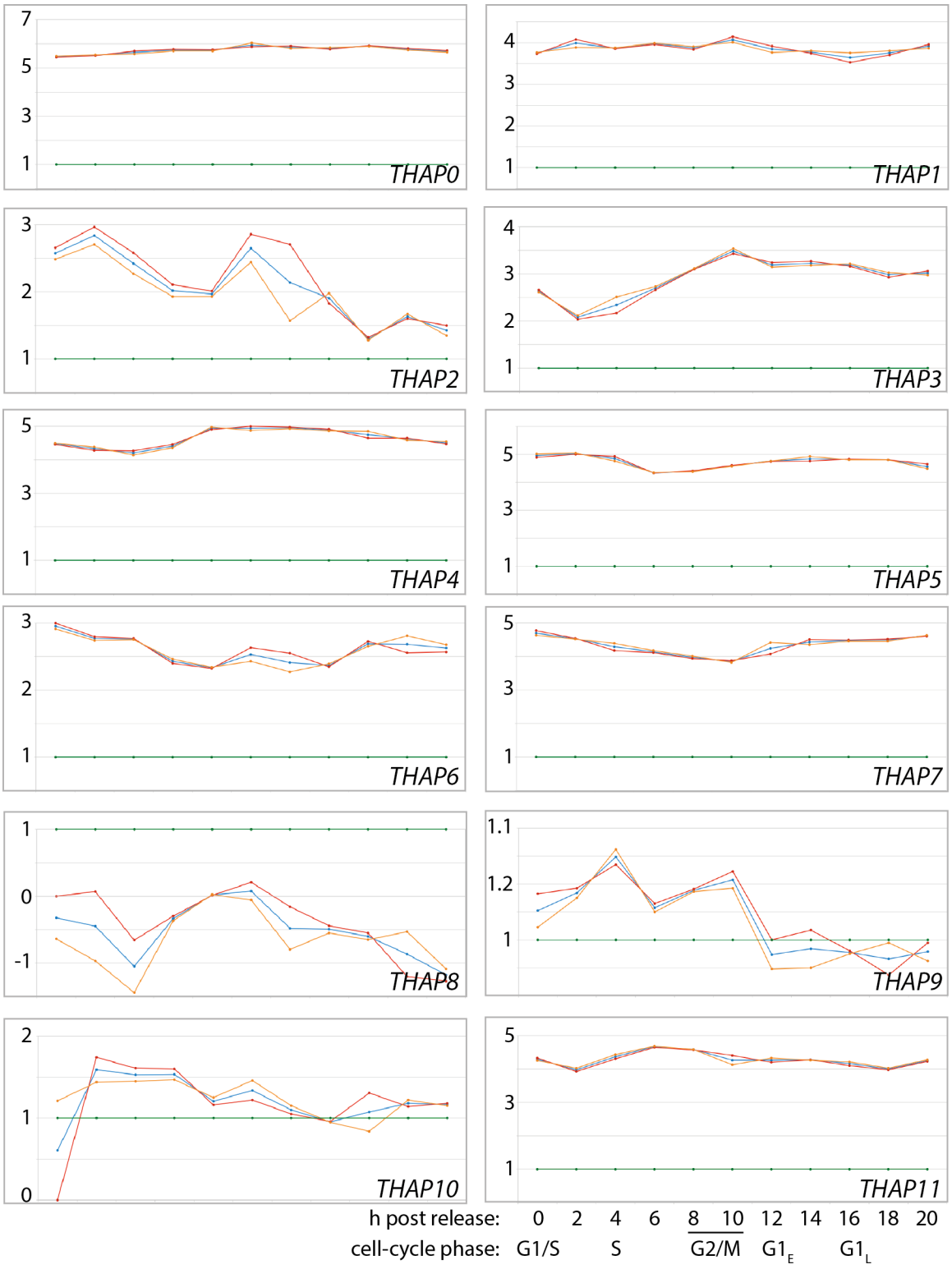


Figure 4.3: **Expression levels of human *THAP* genes along the cell cycle of cultured cells.** HeLa-S cells grown in suspension were synchronized at G1-to-S phase by double thymidine block and analyzed at different time points after release into normal medium; the time points and the associated cell-cycle phases are indicated below the *THAP11* gene results. $\log_2$ of RPKM expression values (Y axis) are displayed for each *THAP* gene. Red, replicate 1; orange, replicate 2; blue, mean of the 2 replicates; green, mRNA-level cut-off.

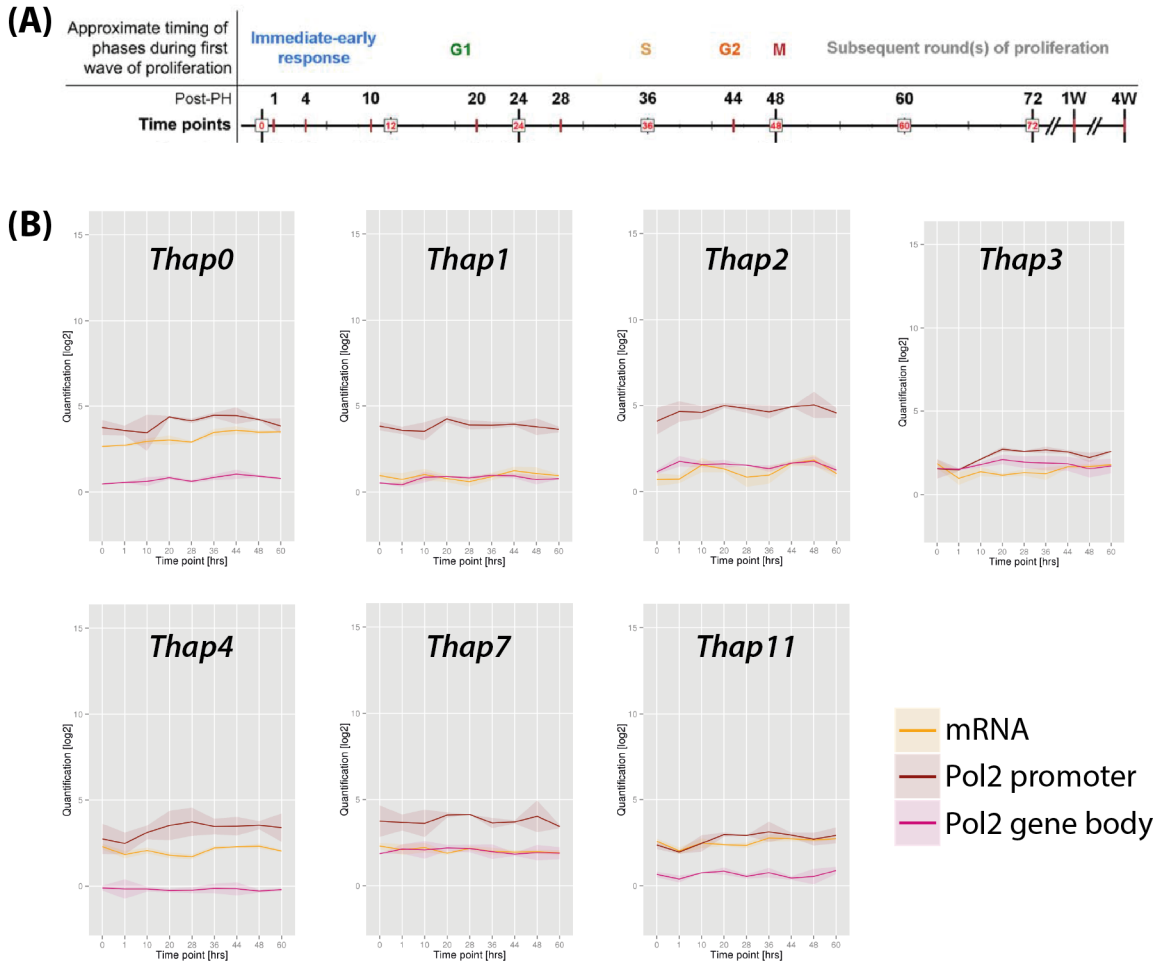


Figure 4.4: **Transcription profiles of murine *Thap* genes during liver regeneration.** Partial hepatectomy was performed and the mice were sacrified at different time points post surgery, RNA extracted from the liver and submitted to high-throughput sequencing. **(A)** Schematic showing the cell-cycle progression following partial hepatectomy [122]. **(B)** Transcript levels and Pol2 occupancy are displayed for each *Thap* gene. Orange, mRNA; brown, Pol2 at promoter; pink, Pol2 in the gene body.

4.4 Discussion

In this chapter, I have combined data of diverse sorts to probe the expression of the *THAP* genes: publicly-available data, experimental results from colleagues, and my own experiments.

First, I took advantage of gene expression analyses that have been already published [120] and for which the data are available online . This study assessed the gene expression levels in different human organs, in several males and one female. From this, I was able to extract the levels of expression of the human *THAP* genes. I observed that each *THAP* gene has a specific expression pattern, some being much more expressed than others, and some having a tissue-specific expression pattern while other showing a more ubiquitous pattern. These results suggest that at least some human THAP proteins may have tissue-specific functions.

This result meets with the observations of the preceding chapter (Chapter 3) in which each THAP protein was suggested to be unique, despite belonging to the same family. Indeed, I previously hypothesized that each THAP protein would have specialized partners, functions and outcomes. This latter idea is strengthened by the different patterns of expression of the diverse THAP members.

In addition, I probed the expression levels of *THAP* genes during the cell cycle. Indeed, as THAP proteins have been suggested to be transcription factors implicated in the control of different phases of the cell cycle, a possible level of regulation would be the modulation of their expression levels during cell-cycle progression. To test this hypothesis, I used both synchonized human cultured cells and regenerating mouse livers. The analysis of human *THAP*-gene expression shows that all but *THAP2* and *THAP3* genes remain relatively constant along the cell cycle. By contrast, *THAP2* and *THAP3* mRNA levels change during the cell cycle: they decrease, and increase, respectively, while the cell cycle progresses. The analysis of THAP11 protein levels is constistent with what has been shown at the mRNA level as the level of THAP11 protein in synchronized cultured cells remains remarkably constant.

Also, the analysis of murine *Thap* mRNA levels during liver regeneration following partial hepatectomy show both similar and different observations when compared to the human *THAP* genes. In a mouse liver regeneration experiment, all 7 *Thap* mRNA levels remains relatively constant. The results are thus different from what has been observed for human *THAP2* and *THAP3* genes.

Overall, the expression of most *THAP* genes do not appear to be particularly regulated along the cell cycle. Two particular *THAP* genes, however, seem to have their expression affected during the cell-cycle course in humans, but not in mice. The discrepancy between these two *THAP* genes and the rest of the *THAP* family, as well as the discrepancy between human and mouse *THAP2* and *THAP3* genes, may be explained differently for the two *THAP* genes. First, regarding *THAP2*, it has been shown to be downregulated in the mouse placenta and developing brain following late-gestational hypoxia [124]. When growing HeLa-S cells, we try as much as possible to keep them in suspension by cultivating them under constant agitation. Nevertheless, these cells tend to aggregate. Then, in the course of an experiment in which cells are growing, the increasing number of cells potentiates the tendency of HeLa-S cells to aggregate. When aggregated, the cells are likely to have less locally-available oxygen, and thus to suffer from mild hypoxia. As this phenomenon would tend to increase with the concentration of cells, it would thus increase while the experiment progresses. This hypothesis would not only explain why *THAP2* mRNA level was increasing in the cell-synchronization experiment, but also account for the discrepancy between human and murine *THAP2* mRNA levels in the two different experiment. Second, with regard to *THAP3*, a recent study reported *THAP3* mRNA upregulation in human oligodendrogliomas and categorized this gene, together with others, as a gene candidate favoring oligodendroglioma growth [84]. This hypothesis is coherent with the increase of *THAP3* mRNA levels

observed during the progression along the cell cycle. This does not explain, however, the fact that *THAP3* mRNA levels are not decreasing at later time points in the synchonization experiment (from 16 hours post release in normal media) to go back to the initial levels observed at time point 0 in order to start a new cell cycle. Regarding the discrepancy between human and mouse *THAP3* mRNA-level variation, one could argue a species-specific mechanism. Alternatively, *THAP3* gene expression may indeed be affected by the double-thymidie block. This hypothesis why *THAP3* mRNA level was changing during the first hours post release in normal media, and then remaining constant. It would also explain why this phenomenon was not observed in the mouse-liver regeneration experiment.

Thus, most cell-cycle specific action of THAP proteins is likely to be achieved by modulation of their binding to promoters, rather than by modulating their levels. Nevertheless, except for THAP11, I cannot exclude the possibility of a post-transcriptional regulation, which would modulate THAP-protein levels. To test these possibilities, one should probe the binding of THAP proteins at promoters along the cell cycle. ChIP-seq experiments against THAP proteins would need to be performed, either in synchronized human cells, or in post-partial hepatectomy mouse livers.

4.5 Selection of THAP proteins for further study

Having largely considered the whole set of THAP proteins for the analyses done so far, I will know focus on a smaller subset of THAP proteins. Indeed, while it is possible to investigate many proteins simultaneously using bioinformatics tools, and in some high-throughput experiments, it would be highly inefficient to try to deal with the complete set of human THAP proteins together for future experiments. I thus decided to select a restricted number of THAP proteins to study further. The criteria I used to make a choice are the following:

- To have the possibility to study endogenous proteins *in vivo*, I selected abundantly expressed ones. Even though gene expression and protein levels are not systematically correlated, I nevertheless took into account the expression levels of human *THAP* genes displayed in Figure 4.1, where the higher-expressed ones are THAP4, 7 and 11.

- As I have a particular interest in the interplay between THAP proteins and HCF-1, I selected HBM-bearing THAP proteins, but including an HBM-negative THAP protein to compare and contrast. The HBM-negative protein selected was THAP8, as it is higher expressed than the 2 other non-HBM THAP proteins (Figures 3.2 and 4.1). This led me to also include THAP5, which appears to be the closest related THAP member to THAP8 (Figure 3.2).

- It is essential to have the possibility to do validation studies in living organisms. Thus, I focused my attention on THAP proteins having an ortholog in mouse (Figure 3.2).
- Also, it is of particular interest to work on proteins involved in human diseases, especially when specific associated mutations have been uncovered. For this reason, I believe it would be extremely exciting to study THAP1 and THAP11.

To summarize, I concentrated on the following six THAP proteins for deeper investigations: THAP1, 4, 5, 7, 8 and 11.

Chapter 5

Biochemistry analysis of selected THAP proteins

In this chapter, I describe how I performed biochemistry analyses to assess the ability of THAP proteins to bind to themselves, to other members of the THAP family, or to HCF-1.

5.1 Dimerization of THAP proteins

As mentioned, most of the THAP proteins possess a putative coiled-coil dimerization domain (section 1.3.2), and thus could potentially interact with other proteins possessing a coiled-coil domain. Also, dimerization often being an important feature of transcription factors (section 1.1.2), it is of particular interest to study whether THAP proteins can form dimers. Here, I describe the potential interaction of THAP proteins, both with themselves as homodimers, and with each other as heterodimers.

5.1.1 THAP proteins can form homodimers

I first focused on the ability of THAP proteins to form homodimers. To start with, I verified the previously-described THAP11-homodimer formation [54, 55] by co-transfecting HEK-293 cells together with HA- and Flag-tagged THAP11 constructs (Figure 5.1 A) and immunoprecipitating the HA-tagged protein. As shown in Figure 5.1 B, THAP11-Flag is recovered in the THAP11-HA immunoprecipitate (compare lane 6 with lane 8). Thus, the THAP11 protein can form homodimers.

In addition, I previously suggested that the HBM sequence and the coiled-coil domain might be somehow functionally linked to one another (see section 3.3.2). I thus asked whether THAP11 homodimerization de-

pends on the HBM sequence, meaning whether this homodimerization would be sensitive to HBM disruption. I thus repeated the previous experiment using a THAP11 mutant construct in which the HBM sequence has been disrupted by point mutation: DASA in THAP11$_{HBMmut}$ instead of DHSY in THAP11$_{WT}$ (Figure 5.1 A). Comparing lane 8 with lane 10 in Figure 5.1 B shows that THAP11$_{HBMmut}$-Flag interacts as efficiently as THAP11$_{WT}$-Flag with THAP11$_{WT}$-HA. This demonstrates that THAP11 homodimerizes independently of the HBM sequence.

I performed similar experiments to assess the homodimerization ability of THAP7 and THAP8. I co-transfected cells with both HA- and Flag-tagged THAP7 or THAP8 expressing constructs (Figure 5.1 A) and immunoprecipitated the HA version. Figure 5.1 C shows that both THAP7 and THAP8 have the ability to form homodimers (lanes 8 and 10, respectively). In addition, THAP7 homodimers seem to be stronger than THAP8 ones, as the THAP8-Flag protein is recovered less efficiently than the THAP7-Flag one when immunoprecipitating for THAP8-HA or THAP7-HA proteins, respectively (compare lanes 7-8 with lanes 9-10).

5.1.2 Selected pairs of THAP proteins form heterodimers

I then tested whether THAP proteins can form heterodimers. More precisely, I assessed whether THAP11 can bind other THAP proteins by co-transfecting THAP11-HA together with Flag-tagged THAP constructs, and immunoprecipitating THAP11-HA (Figure 5.2 A). Figure 5.2 B shows that, while THAP8 is recovered in the THAP11 immunprecipitate, THAP7 is not (lanes 10 and 8, respectively). A control experiment showed that THAP8 was not recovered in the absence of HA-tagged THAP11 protein (data not shown). In addition, THAP4 is not recovered by THAP11 co-immunoprecipitation (Figure 5.2 C, lane 6). Thus, THAP11 can form heterodimers with THAP8, but not with THAP4 and THAP7.

5.1.3 Conclusions

I have shown here that the three THAP proteins tested (THAP7, 8 and 11) form homodimers. Notably, the disruption of the HBM sequence does not impair THAP11 homodimerization. In addition, one forms heterodimers with THAP11 (THAP8), whereas the others tested do not (THAP4 and 7). In conclusion, the distinct THAP proteins interestingly possess different potentials for homo- and heterodimer formation.

5.2 Interactions between THAP proteins and HCF-1

Assessing the HBM presence among THAP proteins in evolution suggested that there is a selective pressure towards its presence and maintenance in THAP proteins, likely being evidence of a functional relevance

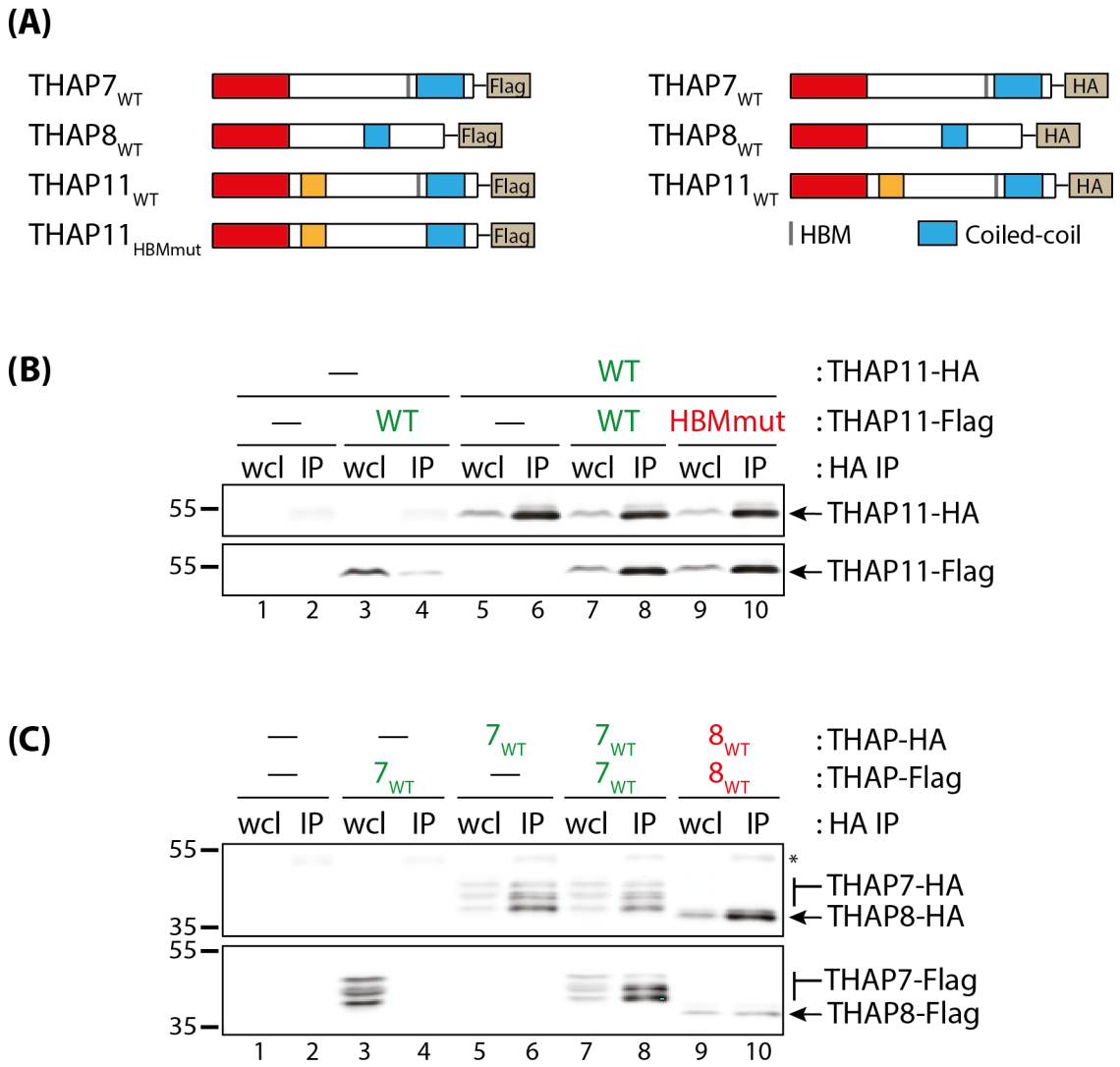


Figure 5.1: **Formation of homodimers within the THAP family.** HEK-293 cells were co-transfected with Flag- and HA-tagged THAP constructs **(A)** and whole cell lysates were then subjected to HA immunoprecipitation and analyzed by immunoblot **(B and C)**. **(B)** THAP11 forms homodimers independently of its HBM sequence. **(C)** THAP7 and THAP8 form homodimers. HBM-positive THAP proteins are depicted in green whereas HBM-negative ones are in red. * non-specific band (antibody heavy chain). wlc, whole cell lysate; IP, immunoprecipitate.

(section 3.2.1). I thus investigated the ability of several human THAP proteins to interact with HCF-1 in human-cell extracts. To present the results, I have grouped the THAP proteins by similarity of their HCF-1 interaction. First, I present two HBM-containing THAP proteins that indeed interact with HCF-1. Second, I show two THAP proteins that do not bind HCF-1 and yet contain an HBM sequence. Third, I introduce an HBM-lacking THAP protein that nevertheless interacts with HCF-1.

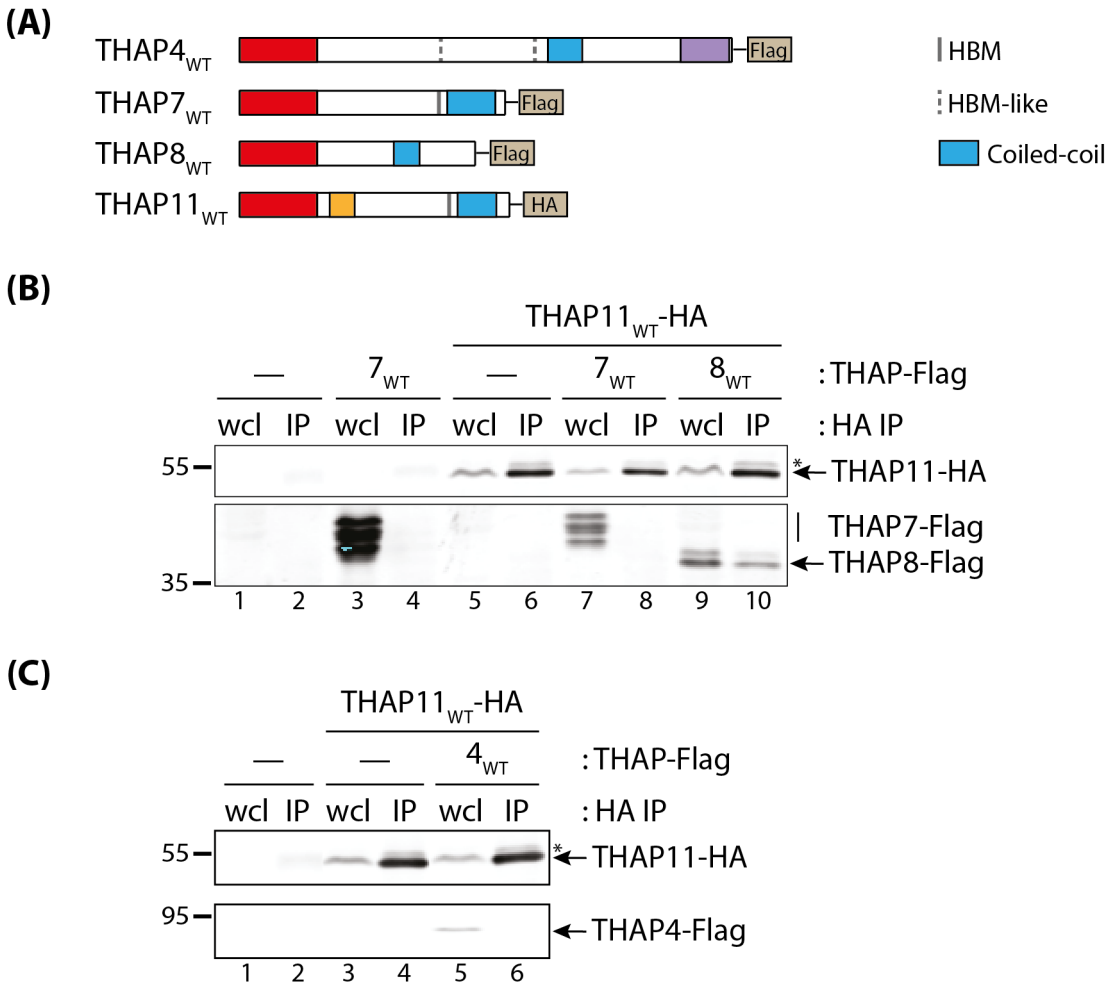


Figure 5.2: **Selected pairs of THAP proteins form heterodimers.** HEK-293 cells were co-transfected with THAP11-HA and a Flag-tagged THAP construct **(A)** and whole cell lysates were then subjected to HA immunoprecipitation and analyzed by immunoblot **(B and C)**. **(B)** THAP8, but not THAP7, heterodimerizes with THAP11. **(C)** THAP4 does not form heterodimers with THAP11. * non-specific band (antibody heavy chain). wlc, whole cell lysate; IP, immunoprecipitate.

5.2.1 HBM-containing THAP7 and THAP11 bind HCF-1

As previously mentioned, the interaction of THAP11 with HCF-1 has been documented in human cell extracts [39]. I thus intended to reproduce the result.

First, I probed the interaction of endogenous THAP11 with exogenous HCF-1 fragments. For this, human HEK-293 cells were transfected with one of the following HCF-1 constructs (Figure 5.3 A) before being immunprecipitated:

- HCF-1_{FL} representing the wild-type and full-length protein;
- HCF-1_{N} representing the N-terminal subunit;
- HCF-1_{C} representing the C-terminal part.

Figure 5.3 B shows that THAP11 is recovered in the HCF-1_{FL} and HCF-1_N immunoprecipitates (lower band, lanes 8 and 6, respectively), but not in the control and HCF-1_C samples (lanes 2 and 4). As a positive control, the presence of the OGT enzyme was probed, as OGT associates with the HCF-1_{PRO} repeats, thus with the full-length HCF-1, and not with its processed forms [16]. Indeed, OGT is co-immunoprecipitated by HCF-1_{FL} (lane 8) but not present in the control, HCF-1_C or HCF-1_N immunoprecipitates (lanes 2, 4 and 6 , respectively). This experiment thus demonstrates that exogenous HCF-1 interacts with endogenous THAP11, and that this interaction is mediated by the HCF-1 N-terminal subunit.

As THAP11 possesses an HBM sequence, it is likely that the described interaction is mediated by the HCF-1 Kelch domain and the THAP11 HBM sequence. To test this idea, I used an HCF-1 Kelch mutant bearing the P134S mutation, known to disrupt the HCF-1 Kelch:HBM interaction, and exogenous Flag-tagged THAP11 constructs, either WT or with a disrupted HBM sequence (DASA instead of DHSY) (Figure 5.3 A). Comparing lanes 2 and 4 in Figure 5.3 C shows that THAP11 efficiently binds HCF-1_N. Lanes 6 and 10 show that using a THAP11 HBM mutant or an HCF-1 P134S mutant strongly decreases the interaction, which is essentially abolished by the combination of both mutants (lane 12). Consequently, THAP11 interacts with HCF-1 in a Kelch:HBM-mediated manner.

Similarly, I have assessed the interaction of THAP7 with HCF-1, and the effect of the HBM and P134S mutations. Figure 5.4 demonstrates that THAP7 interacts with the N-terminal HCF-1 subunit (compare lanes 2 and 4), but that this interaction is impaired by the use of HBM and P134S mutants (lanes 6 and 10, respectively), the effect being even more pronounced when the two mutants are used together (lane 12). Thus, THAP7 binds to HCF-1 via their respective HBM and Kelch domains.

5.2.2 THAP4 and THAP5 do not interact with HCF-1, although containing an HBM or HBM-like sequence

A similar experiment was performed to determine whether THAP4, which has two identical, yet non-canonical, HBM sequences, can interact with HCF-1_N. Indeed, THAP4 HBM-like sequences have a leucine instead of a glutamate, glutamine, aspartate or asparagine in the first position of the HBM motif (Table 3.2). Figure 5.5 shows that THAP4_{WT} and its different HBM mutants (Figure 5.5 A) are unable to bind HCF-1_N nor the HCF-1_{P134S} Kelch mutant.

For technical reasons, THAP5 interaction with HCF-1 has been investigated in a slightly different manner. In this experiment, THAP5 was transfected alone into HEK-293 cells and subsequently immunoprecipitated, to assess the recovery of the endogenous HCF-1 protein. HCF-1 was barely recovered (Figure 5.6, compare lanes 2 and 4), suggesting that THAP5 possesses a very weak interaction, if any, with endogenous HCF-1.

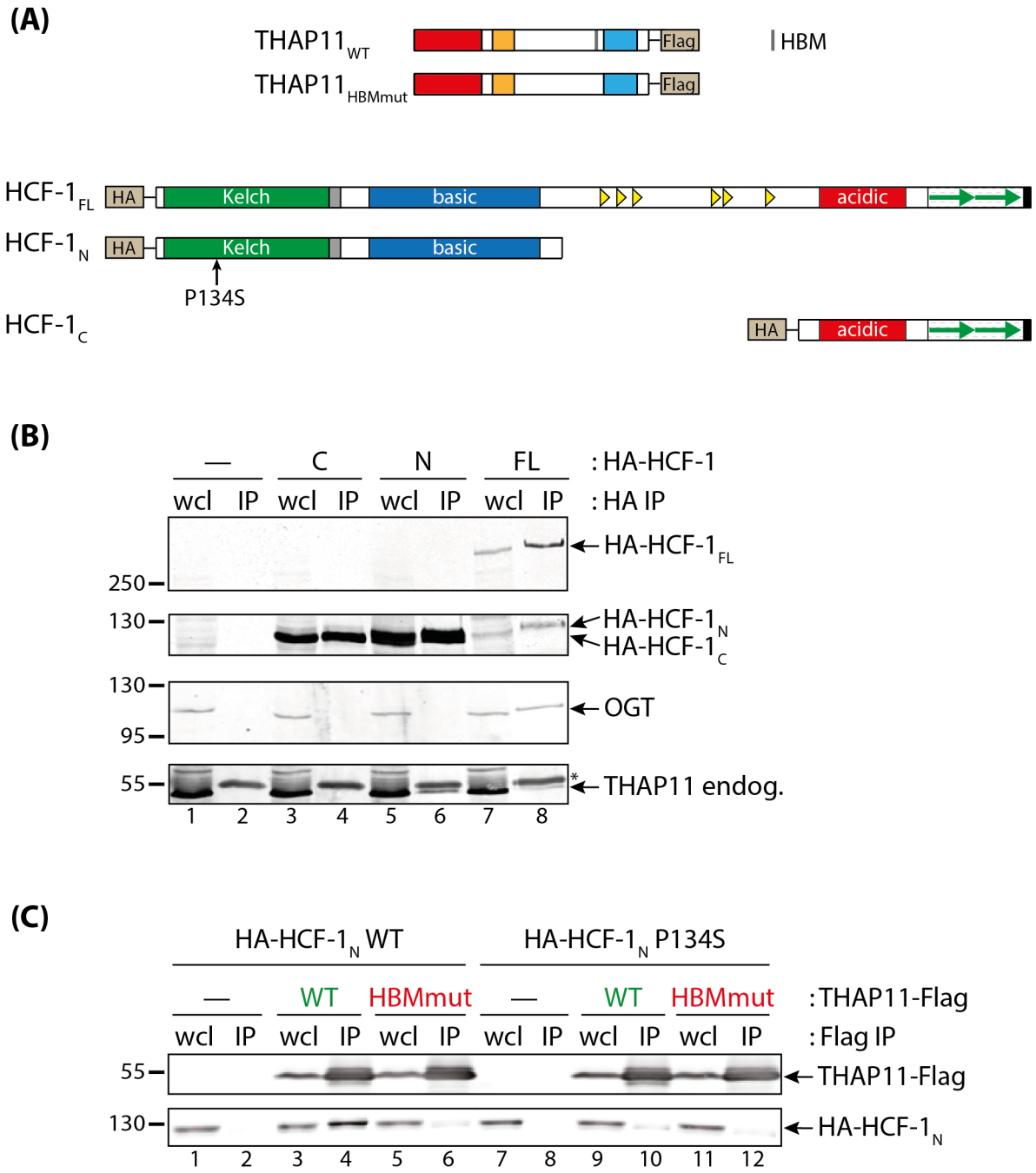


Figure 5.3: **THAP11 binds HCF-1 in a Kelch:HBM-dependent manner.** HEK-293 cells were transfected with different HA-HCF-1 constructs **(A)** and whole cell lysates were then subjected to HA immunoprecipitation and analyzed by immunoblot **(B)**. Alternatively, HEK-293 cells were co-transfected with THAP11-Flag and HA-HCF-1 constructs **(A)**, and whole cell lysates were then subjected to Flag immunoprecipitation and analyzed by immunoblot **(C)**. HBM-positive THAP proteins are depicted in green whereas HBM-negative ones are in red. * non-specific band (antibody heavy chain). wlc, whole cell lysate; IP, immunoprecipitate.

Curiously, THAP5 has an identical HBM to THAP7 (Table 3.2), the latter interacting very efficiently with co-transfected HCF-1 in an HBM-dependent manner, as shown in Figure 5.4. I ruled out the possibility

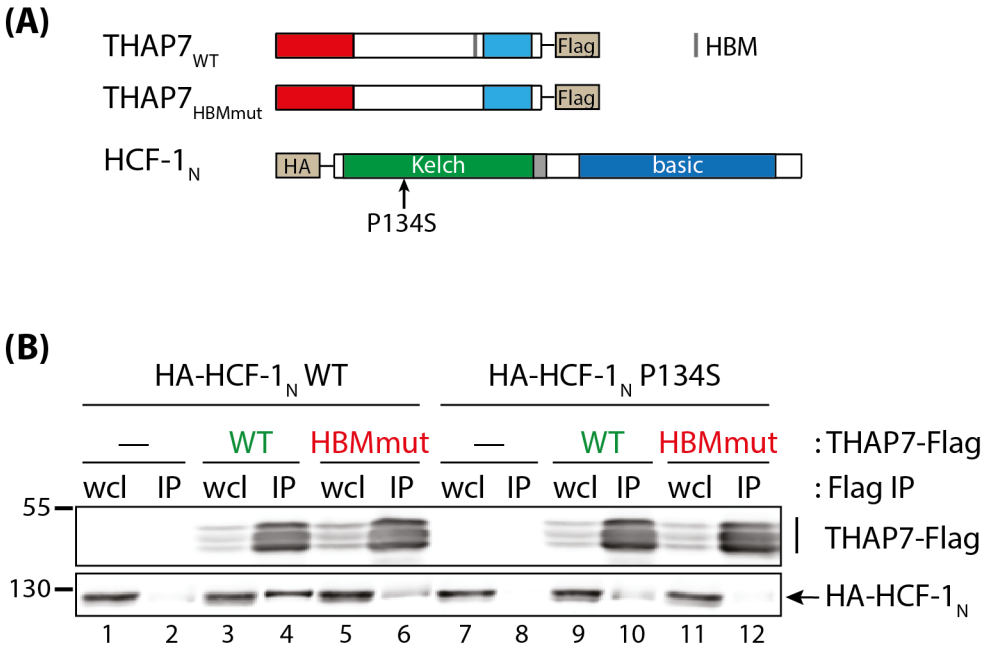


Figure 5.4: **THAP7 interacts with HCF-1 via their respective HBM and Kelch domains.** HEK-293 cells were co-transfected with THAP7-Flag and HA-HCF-1 constructs **(A)**, and whole cell lysates were then subjected to Flag immunoprecipitation and analyzed by immunoblot **(B)**. HBM-positive THAP proteins are depicted in green whereas HBM-negative ones are in red. wlc, whole cell lysate; IP, immunoprecipitate.

that this discrepancy is due to the two different experimental systems used for THAP5 and THAP7, by reproducing with THAP7 the same experimental approach of interaction with endogenous HCF-1 as for THAP5. Comparing Figures 5.6 A and B shows that THAP5 and THAP7, yet displaying identical HBM sequences, have a very different ability to bind HCF-1. Thus, more than the HBM sequence influences the interaction of THAP proteins with HCF-1.

5.2.3 An HBM-lacking THAP protein binds HCF-1: THAP8

Finally. I assessed whether THAP8, which does not display any HBM sequence, interacts with HCF-1. The same experimental set-up described above for THAP11-HCF-1 interaction revealed that, surprisingly, THAP8 interacts with HCF-1_N (data not shown).

To verify if THAP8 indeed interacts with the Kelch domain of HCF-1_N, I co-transfected a THAP8-Flag construct with an HA-tagged HCF-1 Kelch domain-only construct (HCF-1_{Kelch}, Figure 5.7 A). Figure 5.7 B shows that $THAP8_{WT}$ co-immunoprecipitates with HCF-1_{Kelch} WT (compare lanes 2 and 4). To further investigate this HBM-independent binding, I introduced an HBM into the THAP8 sequence by point mutations to determine if it strenghtens, or modifies, the interaction. The HBM was introduced at a similar location to where it is located in HBM-positive THAP proteins, meaning just N-terminal of the coiled-coil domain (Figure 5.7 A). I also used the P134S HCF-1 Kelch mutant to determine if it disrupts the

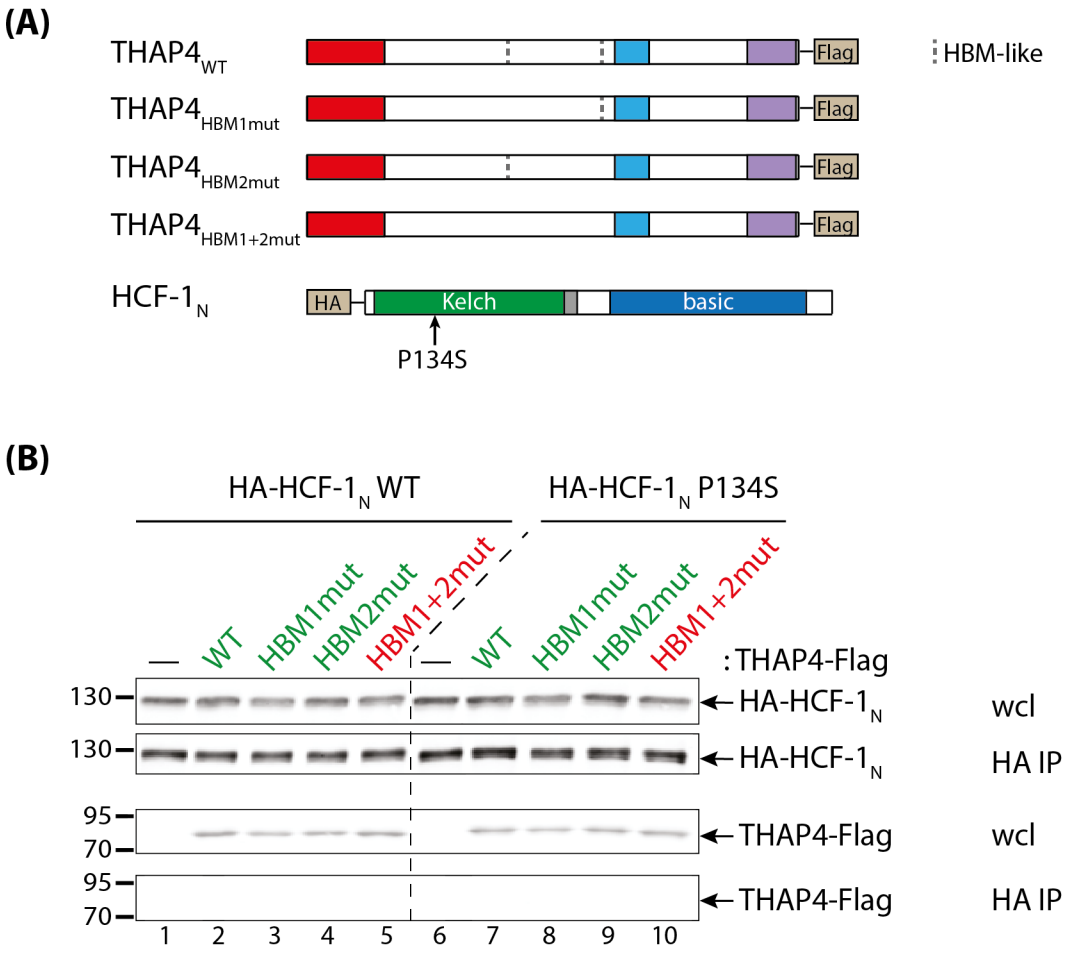


Figure 5.5: **THAP4 does not interact with HCF-1.** HEK-293 cells were co-transfected with THAP4-Flag and HA-HCF-1 constructs **(A)**, and whole cell lysates were then subjected to HA immunoprecipitation and analyzed by immunoblot **(B)**. HBM-positive THAP proteins are depicted in green whereas HBM-negative ones are in red. wlc, whole cell lysate; IP, immunoprecipitate.

interaction (Figure 5.7 A). None of these mutants, however, shows a substantial effect on THAP8-HCF-1 interaction(compare lanes 5, 6 and 7 with lane 4 in Figure 5.7 B). I thus conclude that THAP8 interacts with the Kelch domain of HCF-1 via an unknown mechanim that does not involve the well-known Kelch:HBM mode of interaction.

5.2.4 Conclusions

Here, I show that the THAP proteins have different abilities to bind HCF-1, which surprisingly do not correlate with the presence or absence of an HBM sequence:

- THAP7 and THAP11 interact with HCF-1_N in a Kelch:HBM-mediated manner. This was expected as both THAP proteins display an HBM sequence.

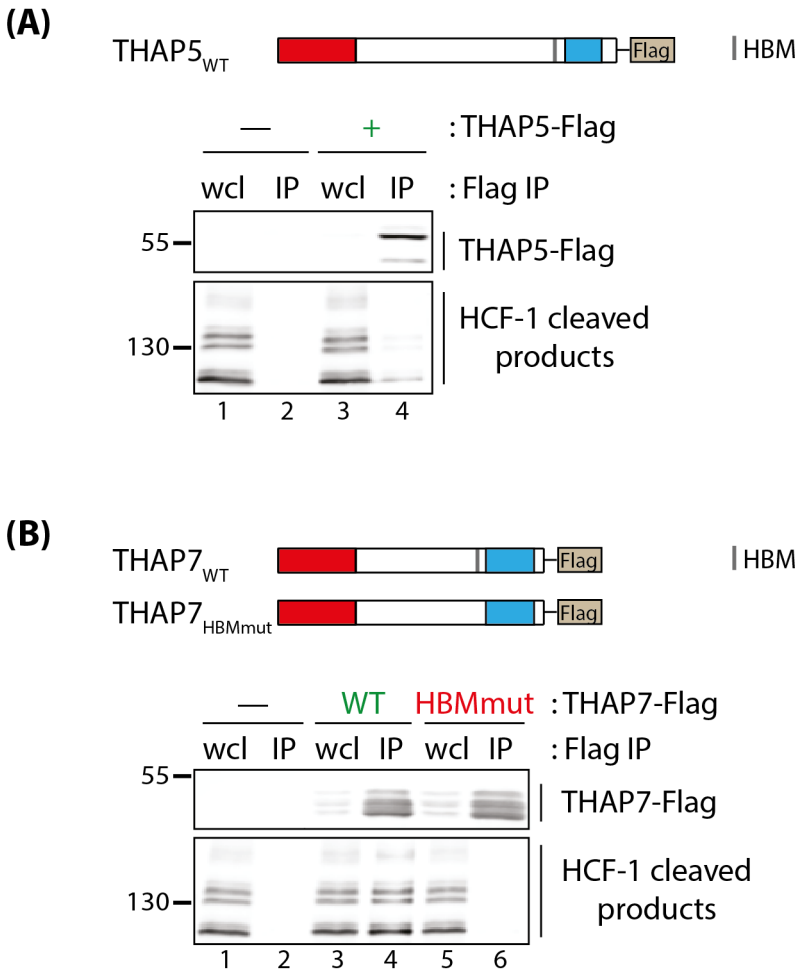


Figure 5.6: **THAP5 barely binds to endogenous HCF-1.** HEK-293 cells were co-transfected with THAP5-Flag **(A)** or THAP7-Flag **(B)** constructs, and whole cell lysates were then subjected to Flag immunoprecipitation and analyzed by immunoblot. HBM-positive THAP proteins are depicted in green whereas HBM-negative ones are in red. wlc, whole cell lysate; IP, immunoprecipitate.

- THAP4 contains two identical HBM-like sequences. Nevertheless, it does not bind HCF-1.
- THAP5 has an HBM sequence identical to THAP7, the latter efficiently interacting with HCF-1. But, THAP5 displays a very weak interaction, if any, with HCF-1.
- Conversely, THAP8, which lacks an HBM sequence, binds to HCF-1_{Kelch}. The interaction mechanism is unknown, but is independent from the well-known Kelch:HBM interaction.

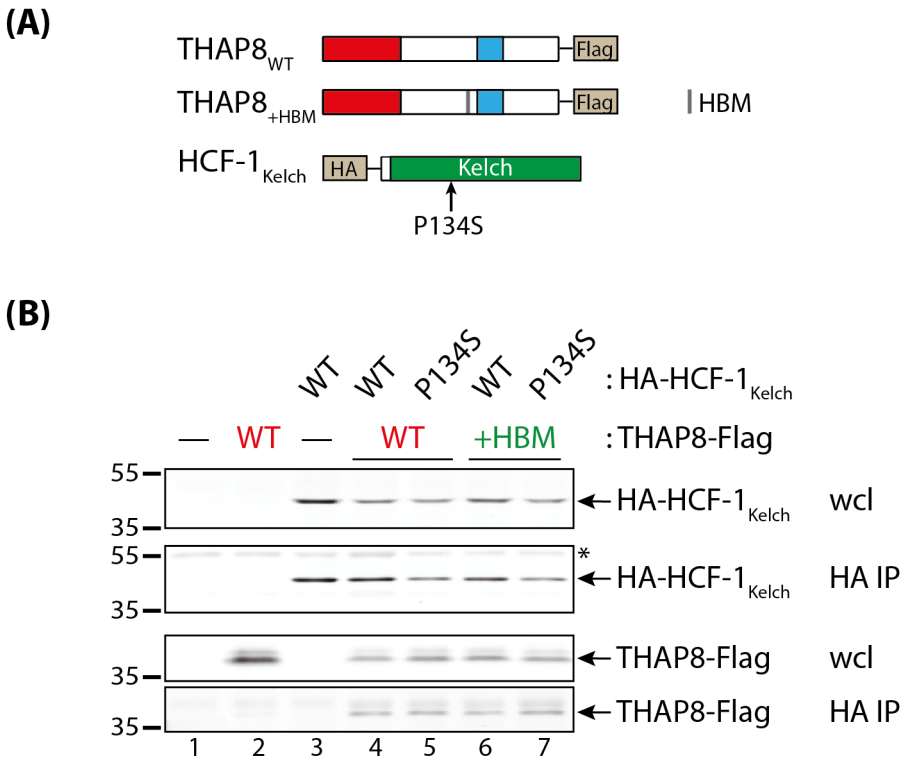


Figure 5.7: **THAP8 does interacts with HCF-1.** HEK-293 cells were co-transfected with THAP8-Flag and HA-HCF-1 constructs **(A)**, and whole cell lysates were then subjected to HA immunoprecipitation and analyzed by immunoblot **(B)**. HBM-positive THAP proteins are depicted in green whereas HBM-negative ones are in red. * non-specific band (antibody heavy chain). wlc, whole cell lysate; IP, immunoprecipitate.

5.3 THAP7 is phosphorylated

5.3.1 THAP7 immunoblotting reveals its phosphorylation

Figure 5.4 has shown that THAP7 appears as a series of bands on immunoblotting. First, I ruled out the possibility that this pattern is due to the anti-Flag antibody used. Indeed, three different antibodies (the anti-Flag, and two anti-THAP7) reveal ectopically-expressed THAP7 as a series of bands on immunoblot (data not shown). Actually, the precise number of bands remains difficult to establish with certainty as they are very close to each other: there are at least 3 bands with the middle one, or/and the lower one possibly being composed of two different bands, thus resulting in 4 or even 5 bands in total. This result is not specific to ectopically-expressed THAP7 as endogenous THAP7, although barely visible with available antibodies, also appears with the same pattern of bands (data not shown).

I then investigated whether these different bands are due to different forms of THAP7 protein, in particular different post-translational modifications. PhosphoSitePlus®, a comprehensive database of post-translational modifications (mainly from mass-spectometry experiments) [125], reveals 7 phosphorylation

sites as well as one acetylation site (Figure 5.8 A, green and pink residues, respectively). I then investigated whether THAP7 phosphorylation is responsible for the multi-band pattern. For this, a THAP7-Flag construct was transfected into HEK-293 cells, immunoprecipitated and then treated with the CIP alkaline phosphatase (Calf Intestinal Phosphatase). Comparing lanes 3 and 5 in Figure 5.8 B shows that the phosphatase treatment of THAP7 leaves only the lower THAP7 band, making the slower migrating species disappear. As a negative control, an inactivated phosphatase does not impact the pattern of THAP7 bands (Figure 5.8 B, lane 4). The treatment of the whole cell lysate with the CIP phosphatase similarly impacts the pattern of THAP7 bands, although in a less dramatic way (Figure 5.8 B, lanes 1 and 2). This result is not surprinsing as the phosphatase treatment is less efficient on non-purified material. To conclude, I have shown that the different bands of THAP7 appearing on immunoblots are due to phosphorylated forms of the protein, but these experiments did not address the functional significance of this interaction.

5.3.2 Impact of THAP7 phosphorylation on its interactions

Thus, I was interested in the possible consequences of the phosphorylation of THAP7, particularly on the previously shown interactions.

First, I have shown that THAP7 homodimerizes (Figure 5.1 C). Comparing lanes 7 and 8 shows that the lower bands of THAP7-Flag are better recovered after co-immunprecipitation of THAP7-HA. This suggests that the different forms of THAP7 have distinct capacities to form homodimers, thus the phosphorylation of THAP7 modulates its ability to dimerize.

Also, I studied whether THAP7 phosphorylation might impact its interaction with HCF-1. For this, I co-transfected HEK-293 cells with HA-tagged HCF-1_N and THAP7-Flag constructs. But, in contrast to Figure 5.4, I performed an HA-immunoprecipitation to pull down HCF-1. Figure 5.9 shows that the different THAP7 bands are not equally co-immunoprecipitated with HCF-1, the middle band being recovered more effectively (compare wcl and IP rows in lane 5). Thus, the interaction of THAP7 with HCF-1 is likely to be modulated by THAP7 phosphorylation.

5.4 Discussion

In this chapter, I have described some elements of how THAP proteins interact with each other and with HCF-1. Table 5.1 summarizes the different interactions presented here.

First, I have probed the ability of some THAP members to form homodimers and heterodimers with THAP11. This has revealed that distinct THAP proteins have differing capacities for homo- and heterodimerization:

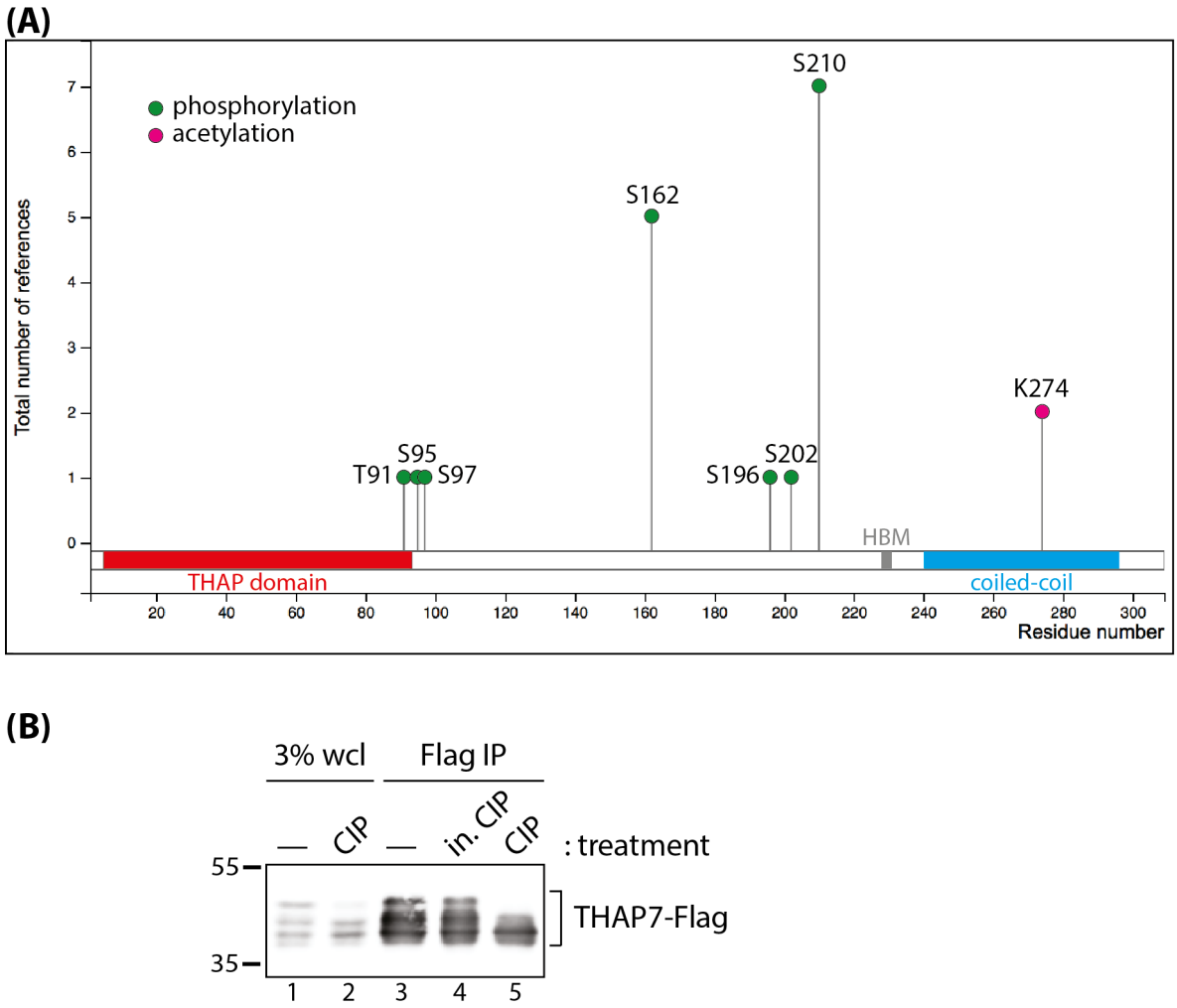


Figure 5.8: **THAP7 undergoes phosphorylation. (A)** THAP7 post-translational modifications, modified from PhosphoSitePlus®. **(B)** HEK-293 cells were transfected with THAP7-Flag and either directly treated with the CIP phosphatase (lane 2) or subjected to Flag immunoprecipitation before being treated with the CIP phosphatase (lanes 4 and 5). in. CIP, heat-inactivated CIP. wlc, whole cell lysate; IP, immunoprecipitate.

- THAP7, 8 and 11 form homodimers. This confirms the previously-published THAP11 homodimerization [54, 55] and adds two new THAP proteins to the list of the ones forming homodimers, already composed of THAP0 [51], THAP1 [52, 53] and THAP11 [54, 55].

- THAP8 heterodimerizes with THAP11, while THAP4 and 7 do not. This latter finding was unexpected as Dejosez and colleagues [54] showed an interaction between the mouse THAP7 and THAP11 (Ronin) orthologs. Surprisingly, this paper was suggesting an interaction between THAP7 and Ronin via the N-terminal half of Ronin, which does not contain the coiled-coil domain.

Thus, each THAP protein has its own ability to form dimers among the THAP family. This result is of particular interest regarding their role in transcription. Indeed, it suggests possibilities for an intricate

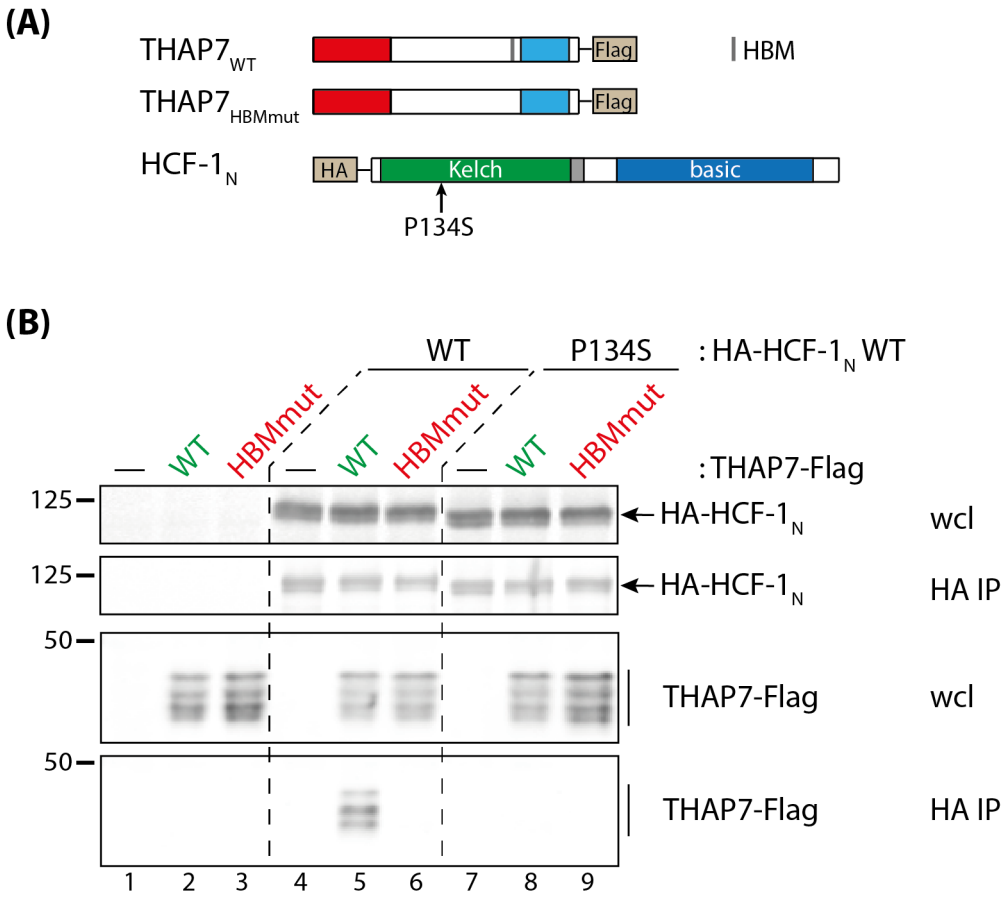


Figure 5.9: **THAP7 phosphorylation modulates its interaction with HCF-1.** HEK-293 cells were co-transfected with THAP7-Flag and HA-HCF-1 constructs **(A)**, and whole cell lysates were then subjected to HA immunoprecipitation and analyzed by immunoblot **(B)**. HBM-positive THAP proteins are depicted in green whereas HBM-negative ones are in red. wlc, whole cell lysate; IP, immunoprecipitate.

THAP-protein regulatory network in which their dimerization can modulate their DNA-binding affinity and specificity, further diversifying their gene targets and functions. Additional experiments would be useful to better understand how THAP proteins bind each other. Now, 5 out of 12 THAP members have been shown to form homodimers; also, to my knowledge, no study so far has reported that a specific THAP protein does not homodimerize. This tends to suggest that homodimerization is an ability shared by most THAP proteins, but additional experiments are required to confirm this hypothesis. In addition, it is likely that the THAP proteins dimerize through their coiled-coil motif, but it remains to be formally investigated by creating THAP-mutant proteins lacking this domain or alternatively, bearing point mutations that disrupt it.

Second, I have assessed the differing abilities of selected THAP proteins to interact with HCF-1 in human cell extracts. Mazars and colleagues [38] have previously demonstrated, in a yeast two-hybrid assay, that all HBM-containing THAP proteins bind HCF-1, but not the HBM-negative ones. I indeed showed a

	THAP4	THAP5	THAP7	THAP8	THAP11
THAP4	·				
THAP5	·	·			
THAP7	·	·	+		
THAP8	·	·	·	+	
THAP11	—	·	—	+	+
HCF-1	—	(—)	+	+	+

Table 5.1: **Summary of THAP dimerization and interaction with HCF-1.** Green plus, interaction; red minus, no interaction; (-), very weak, if any, interaction; grey dot, not tested.

Kelch:HBM-mediated interaction between HCF-1 and THAP7 or 11. But, I do not reproduce their results for THAP4, 5 and 8. Regarding THAP4 and THAP5, it is possible that the interaction is not strong enough to be detected in human cell extracts, particularly for THAP4, as it displays a non-canonical form of the HBM. On the other hand, it makes more sense to assess the interaction of proteins in their native environment, such as human cells, rather than in an artificial system such as the yeast two-hybrid one. Thus, in my hands, some but not all THAP members bind to HCF-1, and it does not correlate with the fact that they display, or not, an HBM sequence. Together with the fact that THAP5 and THAP7 bind to a very different extend to HCF-1, yet having identical HBM sequences, suggests that the HBM sequence is not the only determinant for THAP-protein binding to HCF-1. Several additional experiments would be needed to unravel in more detail how THAP proteins bind to HCF-1:

- My results suggest that the non-canonical HBM sequences of THAP4 are not sufficient to interact with HCF-1. Thus, it would be interesting to mutate these HBM-like sequences into canonical ones to probe whether it creates an interaction with HCF-1.

- To understand the discrepancy between THAP5 and THAP7, one could create swap mutants between THAP5 and THAP7 HBM-adjacent sequences. It would allow one to unravel the non-HBM determinants of binding with HCF-1.

- I showed the interaction between the HCF-1 Kelch domain and THAP8, but how they interact remains to be clarified. Binding assays between THAP8 deletion mutants and HCF-1_{Kelch} mutants available in the Herr laboratory would reveal which portions of each protein are responsible for the interaction.

In addition to the results presented here, I also tried to probe the interaction between THAP1 and HCF-1.

But, while I used the exact same procedure as for the other THAP proteins, I repeatedly had problems with excessive background in the negative controls. Consequently, I decided to focus on the THAP protein that were giving me interpretable results, and to remove THAP1 from the list of studied THAP proteins.

I previously suggested a putative interconnexion between the HBM sequence and the coiled-coil domain, in other words between the interaction with HCF-1 and THAP dimerization. I showed here that the ability of THAP11 to homodimerize is HBMindependent. This result does not mean, however, that the dimerization of THAP proteins and their interaction with HCF-1 are independent. Notably, I have shown that the HBM and the coiled-coil domain are in close proximity in human proteins (for instance, there are 9 amino-acids between these 2 motifs in THAP11). Consequently, the dimerization of THAP proteins through their coiled-coil domain and their binding to HCF-1 via their HBM are likely to influence each other, at least simply because of sterical reasons. Indeed, THAP dimerization and HCF-1-interaction are likely to be two non-simultaneous events. Thus, it would be interesting to probe whether THAP proteins can dimerize in the context of an HCF-1 interaction, and vice versa, whether THAP proteins can interact with HCF-1 once they have formed an hetero- or homodimer. For instance, one can imagine to probe whether THAP proteins could simultanenously dimerize and interact with HCF-1, by performing sequencial co-immunoprecipitation. If this hypothesis of non-simultaneous binding is correct, it would be interesting to understand in which contexts THAP proteins choose to dimerize, or interact with HCF-1, respectively.

Finally, I demonstrated that THAP7 undergoes phosphorylation and that this phosphorylation appears to modulate THAP7's ability to form dimers with THAP11 and to bind HCF-1. Thus, the phosphorylation of THAP7 is likely to have a regulatory role for its interactions and functions. So far, 7 THAP7 residues — 6 serine and one threonine — have been identified as phosphorylated in mass-spectrometry experiments (PhosphoSitePlus®, [125]). Interestingly, none of these residues lies within the HBM sequence or the coiled-coil domain, which are responsible — or suspected to be — for the THAP:HCF-1 and THAP:THAP interactions. Point mutations of these residues, either in alanine or in a phosphomimetic residue (e.g. aspartate instead of a putative phospho-serine), would reveal which ones are responsible for the mobility shifts observed and allow one to attribute the different immunoblot bands to different THAP7 phosphorylated versions.

I should point out the fact that the biochemical analyses presented in this chapter are dealing with ectopically-expressed proteins in human cells, which is quite an artificial system. Results should thus be confirmed using endogenous proteins. Working with endogenous proteins necessitates sensitive-enough antibodies to detect the endogenous version of the proteins — the latter being often expressed at a much lower level than the exogenous one. Also, the protein needs to be actually present in the working system (here, human cell lines). I tested several commercial antibodies against THAP proteins in a collection of human cell lines (HeLa-S, U2OS, HEK293, DLD-1, IMR90). Although most of them recognized the exogenous THAP

proteins, only THAP1 and THAP11 antibodies were able to detect the endogenous proteins. Gene expression data of HEK293 and HeLa-S cells (not shown and Figure 4.3, respectively) available in the Herr laboratory suggested that, abeit different levels of expression, *THAP* genes are expressed in these cells, abeit at different levels. Gene expression and protein level, however, do not necessarily correlate. Consequently, I cannot completely rule out the possibility that the corresponding proteins are not expressed, explaining the absence of detection by the antibody.

During the course of this study, I ordered the generation of a peptide-based THAP7 custom antibody from a biotech company (Covalab). The antibody was generated using chicken as a host, because the human *THAP7* sequence was too well conserved in other available species (rabbit, mouse, rat) to lead to an efficient immunization. Also, peptides were chosen carefully to avoid any cross-reactivity with other THAP proteins (e.g. outside of the THAP domain). Unfortunately, whereas the resulting purified antibody gave excellent results when tested in ELISA against the peptides, it was very disappointing when tested in western blotting. Indeed, it was barely able to recognize the ectopically-overexpressed THAP7 protein in western blotting, and generated a large number of non-specific bands, making it unusable for immunoblotting. Also, when tested in chromatin immunoprecipitation, it did not allow for any enrichment. In summary, THAP11 is the only studied THAP member for which I have a sufficiently-good antibody, thus the only one for which I can study the endogeous protein.

After having removed THAP1 from the list of THAP proteins that I am focusing on, the subset of studied THAP proteins was reduced to 5 members. To concentrate my efforts, I decided to focus on only two THAP proteins to study in more depth their cellular roles, particularly in cell proliferation and gene transcription. I chose the following two THAP members, in the context of a study centered on the interplay between THAPs and HCF-1:

- THAP11 as its already-suggested HCF-1-dependent involvement in transcription and proliferation is extremely exciting, while requiring further investigation. Also, I confirmed that THAP11 interacts with HCF-1, using the well characterized Kelch:HBM mode of interaction. Importantly, I have a good antibody for it, thus I have the possibility to work on the endogenous protein, which is extremely valuable. Furthermore, a human disease-associated mutation have been identified in THAP11 (p.F80L), which makes this protein even more interesting.

- THAP7 as its role in transcription was previously suggested, and as I clearly demonstrated its Kelch and HBM-dependent interaction with HCF-1.

Chapter 6

Creation of custom cell lines to investigate the cellular roles of THAP7 and THAP11

In this chapter, I describe how I exploited recent genetic technologies to engineer specific mutant THAP7 and THAP11 cell lines. First, I used CRISPR/Cas9 genome-editing technology to perform reverse genetics and alter the *THAP7* or *THAP11* endogenous genes. Second, I created cells stably synthesizing ectopic THAP7 or THAP11 proteins.

6.1 CRISPR/Cas9 engineered mutant cells

I used the CRISPR/Cas9 genome-editing system to engineer specific point mutations in the endogenous *THAP7* and *THAP11* genes. The mutations I introduced were of two types: functional (reverse-genetic strategy) and human-disease associated. I should mention here that I have benefited from priceless help from Philippe Lhôte, cell-culture technician, in the creation and screening of these cell lines.

6.1.1 Point mutations in a human host cell line

As described in the introduction (section 1.4), the CRISPR/Cas9 system allows one to precisely modify the genome of specific cells. When properly designed, the gRNA and the repair template can target the Cas9-mediated cleavage at the desired genomic location and introduce specific DNA modifications, respectively. This technology can be applied on virtually any cell type. Because of the following reasons, I decided to use

human embryonic kidney HEK-293 cells as parental cells:

- At the time at which I intented to use this technology, the CRISPR/Cas9 technique had already been well established in the Herr laboratory by Laura Sposito, a PhD student, in HEK-293 cells;
- HEK-293 cells are easy to maintain in culture and synthesize high levels of proteins;
- HEK-293 cells are easy to transfect, which is of great importance as cells need to be transfected with the different components of the CRISPR/Cas9 system;
- Even if HEK-293 cells are not genetically normal, they have been transformed from primary cells, as opposed to cell lines derived from tumors, and may be genetically more reliable.

The details about the CRISPR/Cas9 mutagenesis protocol are explained in Chapter 2. Briefly, cells transfected with the CRISPR/Cas9 components and a repair template were clonally selected. Each clone was then analyzed by PCR and subsequent digestion with a restriction enzyme. Indeed, each mutagenesis strategy has been designed so the mutation created, or disrupted, a specific restriction site.

Figure 6.1 summarizes the 7 different mutants that I set out to obtain. For each of THAP7 and THAP11, I intended to create the following functional mutants:

- THAP$_{null}$ mutants. One very classical approach of reverse genetics is to assess the effect of the gene loss. Thus, I aimed to create cells in which the *THAP* gene has been disrupted. For this, my strategy was to introduce two stop codons in a row near the very beginning of the *THAP* coding sequence. This should result in the synthesis of an extremely truncated THAP protein with only few amino acids left, likely being non-functional.
- THAP$_{HBM}$ mutants. I have a particular interest in the role of HCF-1 in THAP-mediated activities, and I demonstrated that THAP7 and THAP11 proteins interact with HCF-1 via their HBM sequence. Thus, I wished to investigate what is happening when a given THAP protein has lost its ability to bind HCF-1, by precisely mutating its HBM sequence — as previously done in ectopically-expressed proteins (Chapter 5).
- THAP$_{\Delta CC}$ mutants devoid of the coiled-coil domain. The ability of THAP proteins to form dimers probably being critical for their functions and the coiled-coil domain likely being responsible for their dimerization, I had an interest in probing the effects of the lack of this domain in THAP proteins. As in THAP proteins, the coiled-coil domain is the most C-terminal feature (Figure 1.5), the truncation can be achieved by introducing two stop codons in a row at the beginning of the coiled-coil domain.

This would result in a truncated THAP protein devoid of its coiled-coil domain, but still retaining its HBM sequence.

Also, a disease-associated mutation has been identified in THAP11 (p.F80L), resulting in a cobalamin disorder [75, 83], but so far its mechanism of action remains unclear. It would thus be extremely interesting to introduce this point mutation into cells, to create a cellular model and study the altered mechanisms of this mutation.

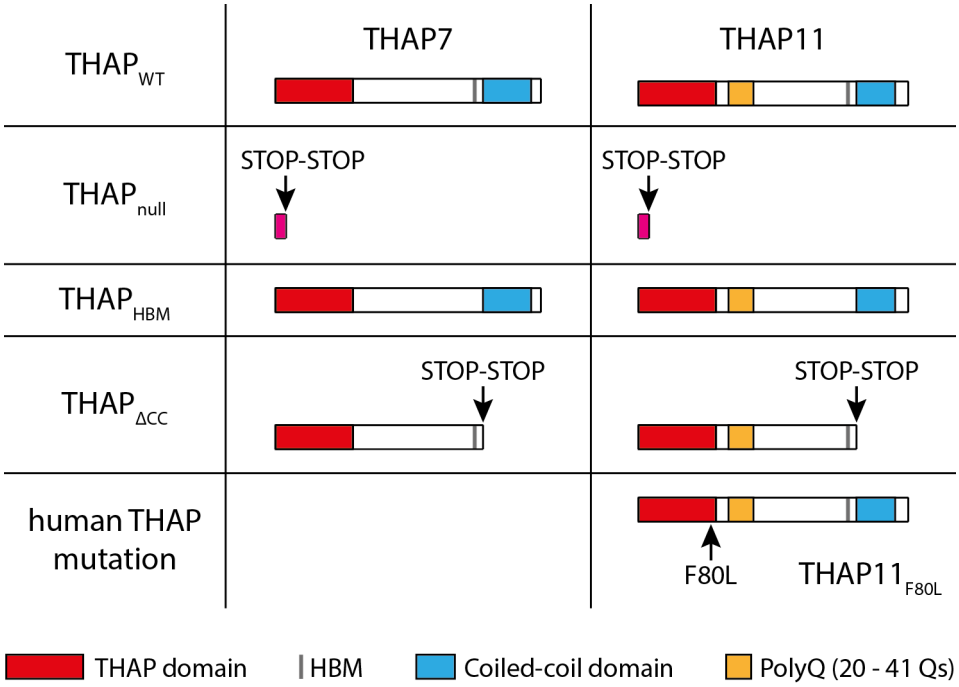


Figure 6.1: **THAP7 and THAP11 point mutants.** HEK-293 cells are subjected to CRISPR/Cas9-mediated genome editing in the aim of obtaining these 7 different mutants.

To rule out clone-specific observations and off-targets effects, I tried to obtain for each mutant, if possible, two independent clones from two different transfection experiments.

6.1.2 Generation of THAP7$_{null}$ and THAP11$_{null}$ cell lines

To create THAP7$_{null}$ and THAP11$_{null}$ mutants as described, I had to create two successive stop codons by point mutagenesis at the very beginning of the *THAP* coding sequence. To increase the efficiency of the mutagenesis, I minimized the number of nucleotides to be mutated. Also, the mutations should be around 10 nucleotides from the cleavage site, the latter being determined by a PAM sequence on the target gene. Considering these restrictions, the selected position of the stop codons would result in the generation of truncated THAP7 and THAP11 proteins of 12 and 17 amino-acids, respectively. These truncated THAP proteins would only retain the two first cysteines of the THAP domain "C_2CH signature". Consequently,

the THAP zinc fingers would be incomplete and the resulting truncated THAP proteins would be likely non-functional.

THAP7$_{null}$ mutant cell lines

The introduction of the two selected stop codons described above required the mutation of 4 distinct nucleotides (Figure 6.2 A). As the mutations created a NlaIII restriction site that is absent in the WT sequence (Figure 6.2 B), the cell clones were screened by the ability of the NlaIII enzyme to digest the DNA at the site of the mutations.

The first mutagenesis attempt gave us 59 clones to test, from which we were able to identify one clone with the desired homozygous mutations (# 29). A second independent attempt provided us with 101 clones, from which two had the desired mutations in a homozygous state (# 32 and # 44). Figures 6.2 B and C show the pattern of NlaIII digestion of these 3 clones compared to WT, and their sequencing chromatograms, respectively. Because it comes from a separate transfection, clone # 29 is independent from the clones # 32 and # 44.

THAP11$_{null}$ mutant cell lines

Despite having screened 139 clones from two separate transfection experiments, we were not able to find a single cell clone with the desired 4-nucleotide mutation as depicted in Figure 6.3. We were thus unable to obtain a cell line devoid of the THAP11 protein.

6.1.3 THAP7 and THAP11 HBM-mutant cell lines

As seen in Chapter 5, mutating the second and last positions of the 4 amino-acid HBM sequence into alanine residues is sufficient to disrupt the HBM sequence and prevent THAP7 and THAP11 from interacting with HCF-1. T thus designed CRISPR/Cas9 mutagenesis strategies to create such mutations in HEK-293 cells.

THAP7$_{HBM}$ mutant cell lines

Mutation of the THAP7 EHSY HBM sequence into EASA disrupts an AluI restriction site (Figures 6.4 A and B). From the 80 clones tested, 2 have the homozygous desired mutation (clones # 34 and # 76). In addition, clone # 50 is unclear: its sequencing chromatogram — which has been confirmed by a second sequencing — suggested the presence of the mutations, but still displays residual signals for the WT nucleotides (C, A, T and A). Thus, the homo- or heterozygous status of this clone remains uncertain. Curiously, the clone # 42 is heterozygous at the two first nucleotide positions, while being homozygous at the two last ones (Figures 6.4

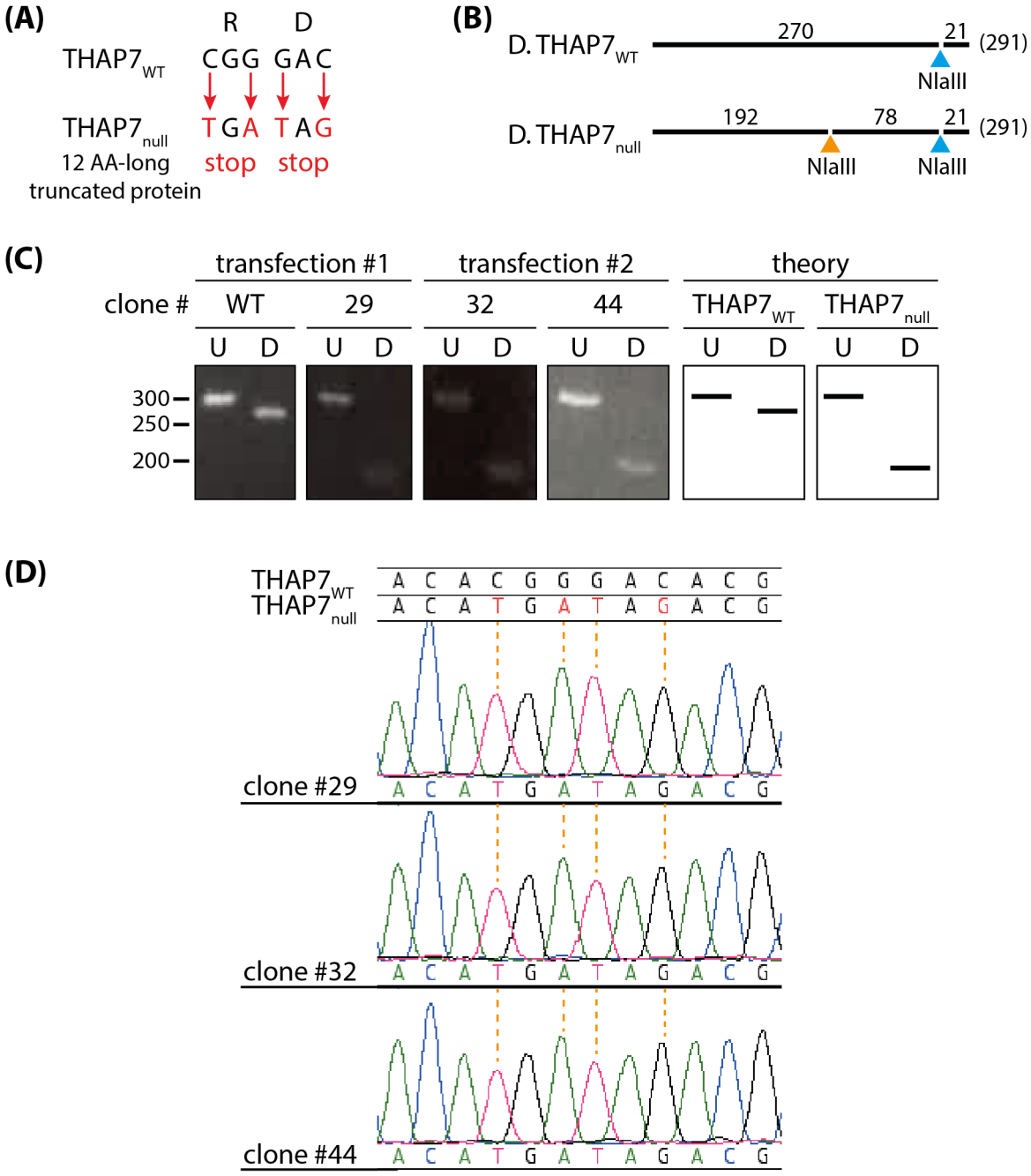


Figure 6.2: **THAP7$_{null}$ mutant cells.** Two successive stop codons were created by mutating 4 nucleotides at the beginning of the *THAP7* coding sequence **(A)**, and cell clones were subsequently screened with the NlaIII enzyme, as the mutation creates one restriction site (orange triangle) **(B)**. **(C)** NlaIII digestion patterns of WT and mutant cell clones, compared to the theoretical ones. **(D)** Sequencing chromatograms of the mutant cell clones, compared to the WT and the expected mutant sequences, the mutated residues deing depicted in red. U, undigested PCR fragment; D, NlaIII digested fragment.

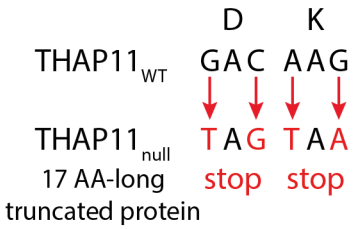


Figure 6.3: **Strategy for THAP11$_{null}$ mutagenesis.** Four mutations were designed to create two successive stop codons at the beginning of the *THAP11* coding sequence.

C and D). As clones # 42 and # 50 are odd, I did not include them in future analyses and instead focused on the clear homozygous mutant clones # 34 and # 76.

THAP11$_{HBM}$ mutant cell lines

Figures 6.5 A and B show that the suppression of the THAP11 HBM sequence by mutation of 4 nucleotides disrupts a CviQI restriction site originally present in the WT sequence. Unfortunately, we did not identify any THAP11$_{HBM}$ mutant cells on the first attempt of mutagenesis, in which we screened 87 clones. In a second attempt, we suprisingly obtained only 12 clones. While we still did not obtain any homozygous mutant clone, we found an heterozygous THAP11$_{HBM}$ mutant (clone # 1, Figures 6.5 C and D). These cells, however, failed to survive as, despite the great attention we provided them, they ultimately died. These results suggest that the THAP11 HBM mutation is detrimental to HEK-293 cell growth even in a heterozygous state.

6.1.4 Coiled-coil truncated THAP7 and THAP11 mutant cell lines

Thanks to the fact that the coiled-coil domain is the most C-terminal feature of THAP proteins, introducing stop codons at the very beginning of this domain will result in a truncated protein devoid of the coiled-coil domain, but still retaining its other features — particularly, its THAP domain and HBM sequence. Here, I introduced two successive stop codons to completely delete the coiled-coil domain, leaving only one amino-acid from the original coiled-coil domain of THAP7, and not any for THAP11. Consequently, the HBM is left intact and located 9 amino-acids from the C-terminus of both truncated proteins.

THAP7$_{\Delta CC}$ cell lines

Figure 6.6 A shows the nucleotides mutated to create two stop codons at the very beginning of the THAP7 coiled-coil domain. The resulting cell clones were tested using the AciI enzyme, as the mutations disrupt an AciI restriction site originally present in the WT sequence (Figure 6.6 B). From the 79 clones tested, we obtained 6 homozygous for the desired 3-nucleotide mutation: clones # 7, # 22, # 27, # 36, # 40 and # 58 (Figures 6.6 C and D). Clones # 22 and # 58, however, had trouble proliferating, making them difficult to

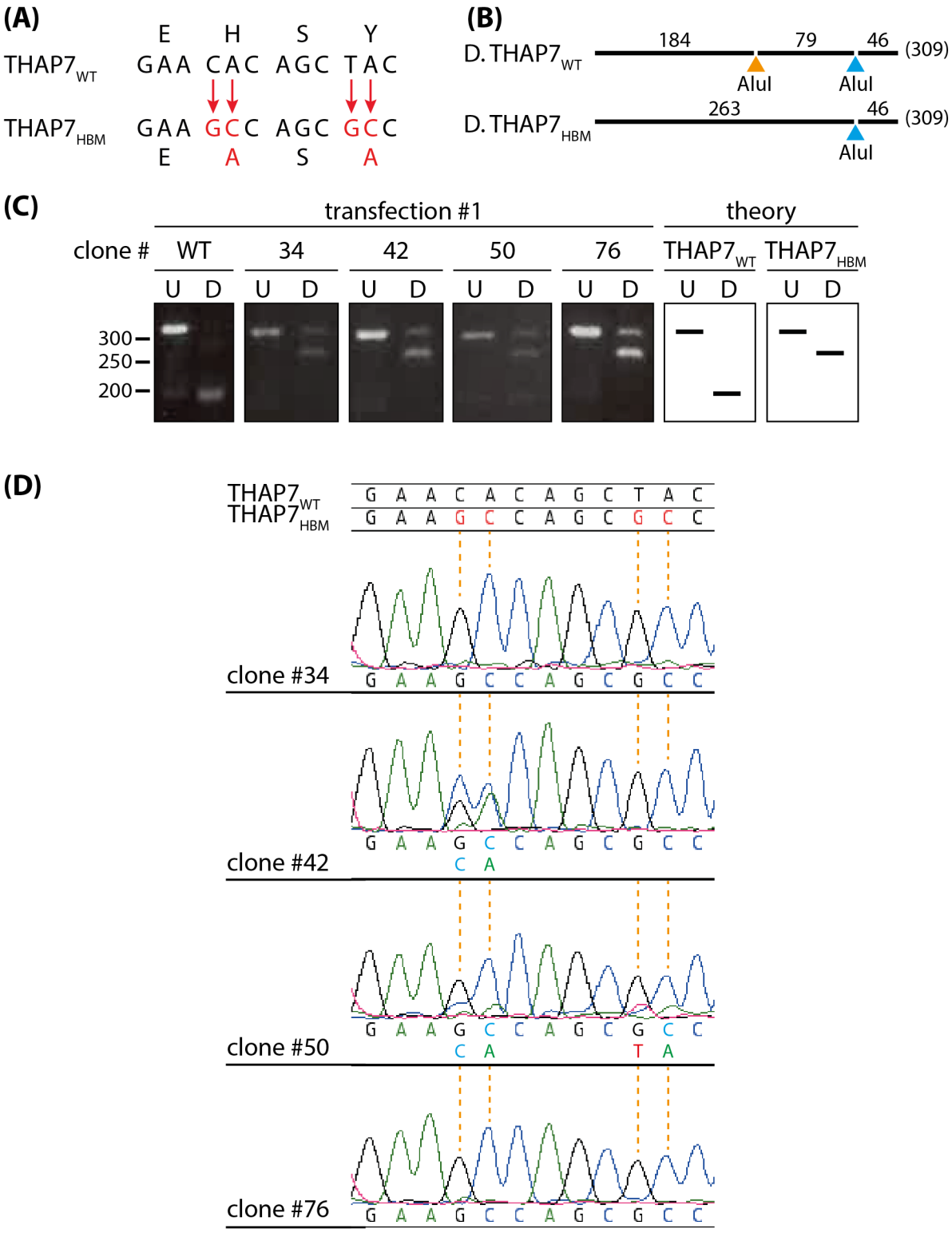


Figure 6.4: **THAP7$_{HBM}$ mutant cells.** The THAP7 EHSY HBM sequence was changed into EASA by mutating 4 nucleotides **(A)**, and cell clones were subsequently screened with the AluI enzyme, as the mutation disrupts one restriction site (orange triangle) **(B)**. **(C)** AluI digestion patterns of WT and mutant cell clones, compared to the theoretical ones. **(D)** Sequencing chromatograms of the mutant cell clones, compared to the WT and the expected mutant sequences, the mutated residues being depicted in red. U, undigested PCR fragment; D, AluI digested fragment.

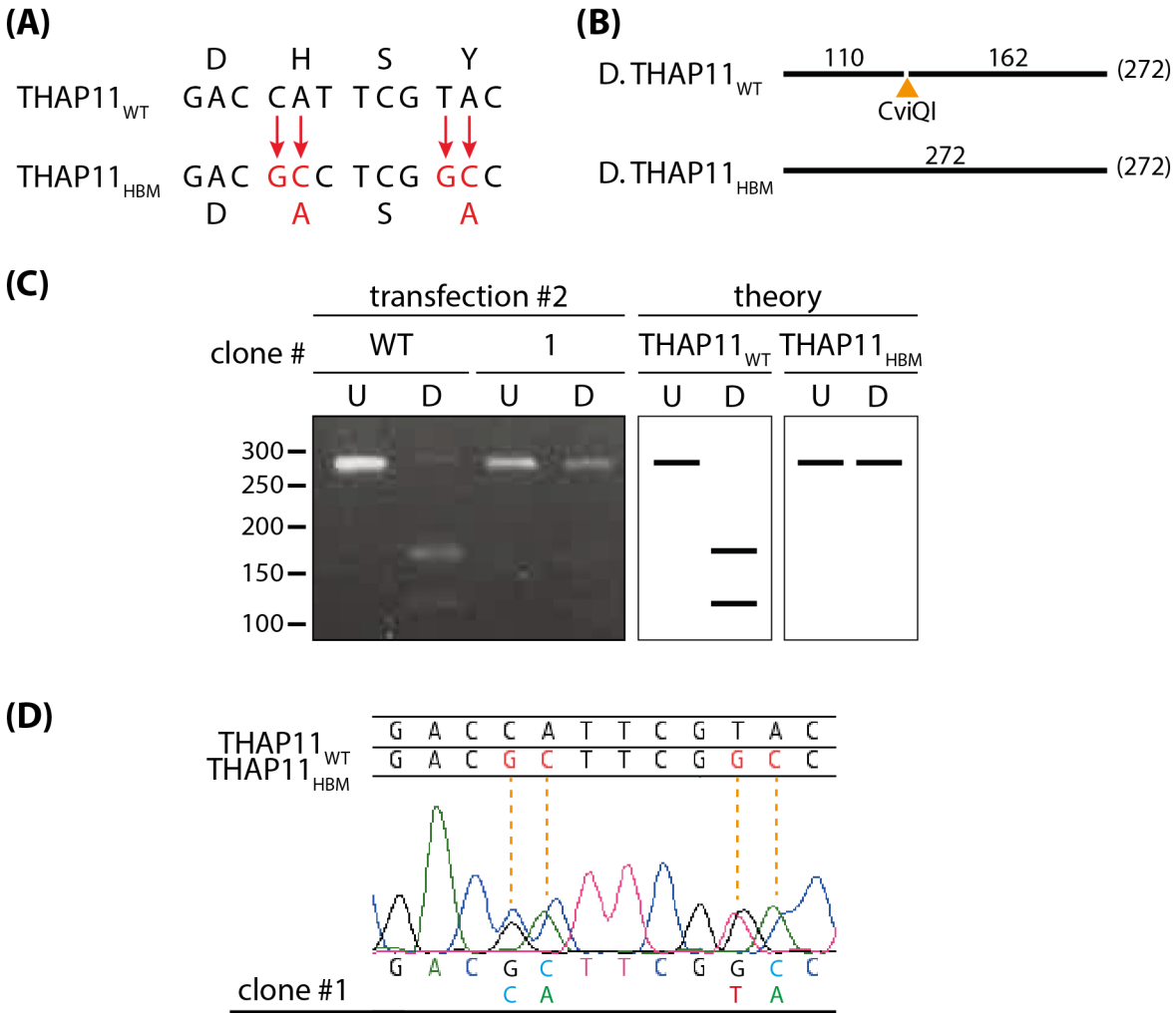


Figure 6.5: **THAP11$_{HBM}$ mutant cells.** The THAP11 DHSY HBM sequence was changed into DASA by mutating 4 nucleotides **(A)**, and cell clones were subsequently screened with the CviQI enzyme, as the mutation disrupts one restriction site (orange triangle) **(B)**. **(C)** CviQI digestion patterns of WT and the mutant cell clone, compared to the theoretical ones. **(D)** Sequencing chromatograms of the mutant cell clone, compared to the WT and the expected mutant sequences, the mutated residues being depicted in red. U, undigested PCR fragment; D, CviQI digested fragment.

keep in culture. Cells were nevertheless expanded sufficiently to be frozen, but they were consequently not analyzed further as I concentrated my efforts on the 4 other clones obtained.

THAP11$_{\Delta CC}$ cell lines

For construction of the THAP11$_{\Delta CC}$ cell line, a first experiment gave us 91 single-cell clones to test. We were not able, however, to find any clone bearing the desired 2-nucleotide mutations depicted in Figure 6.7. The repetition of the experiment resulted in only 30 clones and unfortunately, none of them was positive for the mutations as well. Thus, we were unable to obtain any THAP11$_{\Delta CC}$ cell clone, indicating that loss of the THAP11 coiled-coil domain is deleterious to HEK-293-cell growth and/or viability. .

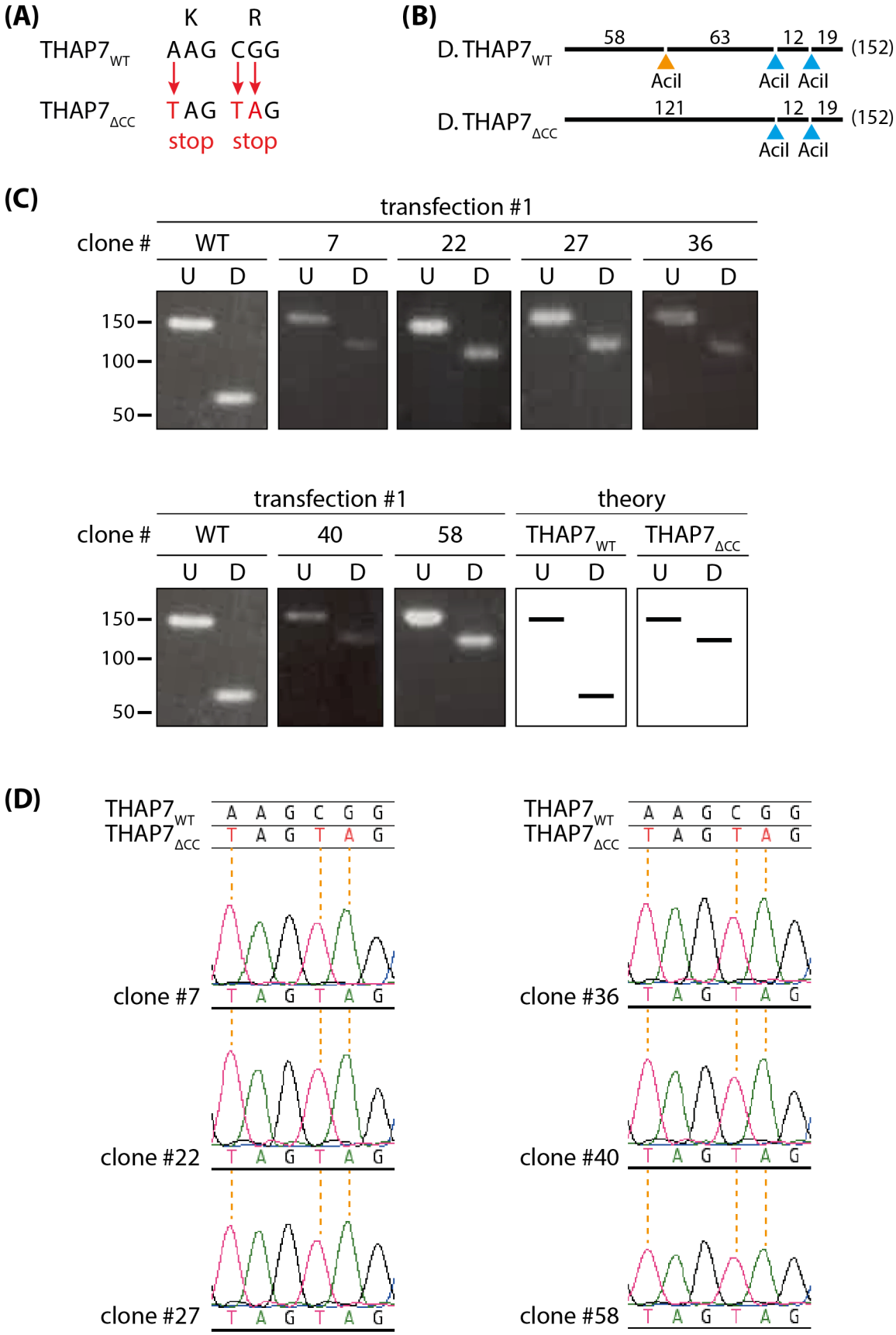

(A)
K R
THAP7WT AAG CGG
THAP7ΔCC TAG TAG
stop stop
(B)
D. THAP7WT
58 63 12 19 (152)
AciI AciI AciI
D. THAP7ΔCC
121 12 19 (152)
AciI AciI
(C)
transfection #1
clone # WT 7 22 27 36
U D U D U D U D U D
150
100
50
transfection #1 theory
clone # WT 40 58 THAP7WT THAP7ΔCC
U D U D U D U D U D
150
100
50
(D)
THAP7WT A A G C G G
THAP7ΔCC T A G T A G
clone #7
clone #22
clone #27
clone #36
clone #40
clone #58

Figure 6.6: **THAP7$_{\Delta CC}$ mutant cells.** Two successive stop codons were created by mutating 3 nucleotides at the very beginning of the THAP7 coiled-coil domain **(A)**, and cell clones were subsequently screened with the AciI enzyme, as the mutation disrupts one restriction site (orange triangle) **(B)**. **(C)** AciI digestion patterns of WT and mutant cell clones, compared to the theoretical ones. **(D)** Sequencing chromatograms of the mutant cell clones, compared to the WT and the expected mutant sequences, the mutated residues being depicted in red. U, undigested PCR fragment; D, AciI digested fragment.

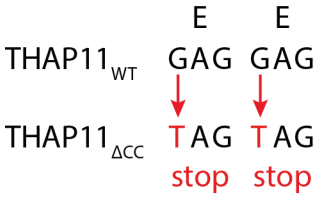


Figure 6.7: **Strategy for THAP11$_{\Delta CC}$ mutagenesis.** Two nucleotides were mutated to create two successive stop codons at the beginning of the THAP11 coiled-coil domain.

6.1.5 Creation of a cell line bearing a human THAP11 disease-associated mutation

I also aimed to create a cell line bearing the recently uncovered THAP11 disease-associated p.F80L mutation (THAP11$_{F80L}$, [75]).

The F80L THAP11 mutation associated with cobalamin disorder

Cobalamin disorders refer to a group of disorders of intracellular cobalamin (vitamin B_{12}) metabolism associated with neurodevelopmental abnormalities. An X-linked subgroup, named *cblX*, is due to a variety of recessive HCF-1 mutations [126]. HCF-1, together with THAP11, transcriptionaly regulate cobalamin metabolism by regulating the expression of *MMACHC*, whose gene product is a key enzyme in the cobalamin metabolic pathway. In *cblX* patients, however, mutant HCF-1 fails to activate *MMACHC* expression, resulting in a pronounced decrease of *MMACHC* mRNA and protein levels, likely being the cause of the cobalamin deficiency [75,126]. Recently, Quintana and colleagues [75] have reported a patient with *cblX*-like phenotypic characteristics, but without any MMACHC or HCF-1 mutation. Instead, this patient had an homozygous missense mutation in the *THAP11* gene (n.C240G) resulting in the change of a phenylalanine into a leucine at the end of its THAP domain (p.F80L) [75]. This phenylalanine amino acid is highly conserved across vertebrate species [75] and also among the different human THAP proteins. Indeed, it is the last position of the THAP domain "AVPTIF box" which has been suggested to be necessary for the proper folding of the zinc finger (Figure 6.1, see also section 1.3.2 and Figure 1.6 A, green box). Thus, the mutation likely affects the folding and then the stability of the THAP11 protein. Similarly to *cblX* patients with HCF-1 mutations, the THAP11$_{F80L}$ patient exhibited a drastic reduction in *MMACHC* expression [75]. To further

clarify the pathology of the THAP11$_{F80L}$ mutation, I translated it into a cellular model by precisely mutating the endogenous *THAP11* gene in HEK-293 cells.

Translating a human disease into a cell-culture model: the THAP11$_{F80L}$ cell line

I exploited once more the CRISPR/Cas9 genome-editing strategy to introduce the p.F80L mutation depicted in Figure 6.8 A in HEK-293 cells. In a first transfection experiment, we were able to obtain 88 clones to test with the FauI enzyme, as the corresponding n.C240G mutation disrupts a FauI restriction site (Figure 6.8 B). From these 88 clones, sequencing chromatograms revealed that clone # 8 is homozygous for the mutation, while clone # 20 is heterozygous (Figures 6.8 C and D). We repeated the experiment to obtain a second independent homozygous mutant, but unfortunately, none of the 128 tested clones bore the desired mutation.

Impact of the F80L mutation on the THAP11 protein

To assess the effect of the p.F80L mutation on the THAP11 protein, a western blot was performed using lysates of WT and THAP11$_{F80L}$ mutant cells. With the help of Maykel Lopes, equal amount of proteins from the two lysates were separated and visualized with the anti-THAP11 antibody. Figure 6.9 A reveals that the band corresponding to the WT protein is barely visible in the THAP11$_{F80L}$ mutant cells compared to the WT cells (lanes 2 and 1, respectively). It is unlikely that the F80L mutation causes a change in the migration pattern of the THAP11 protein, as there is not any new band appearing in the THAP11$_{F80L}$ cell lysate. Consequently, this observation demonstrates that the THAP11$_{F80L}$ cells have lower THAP11 protein levels compared to WT cells. To more accurately estimate the amount of THAP11$_{F80L}$ mutant protein compared to the THAP11$_{WT}$ one, Maykel Lopes sequentially diluted the lysate of the WT cells by 2 fold. Figure 6.9 B indicates, however, a non-linear esponse of the detection assay. Thus, the THAP11 protein level is reduced in the THAP11$_{F80L}$ cells compared to WT ones, but the assay did not allow me to quantify this decrease.

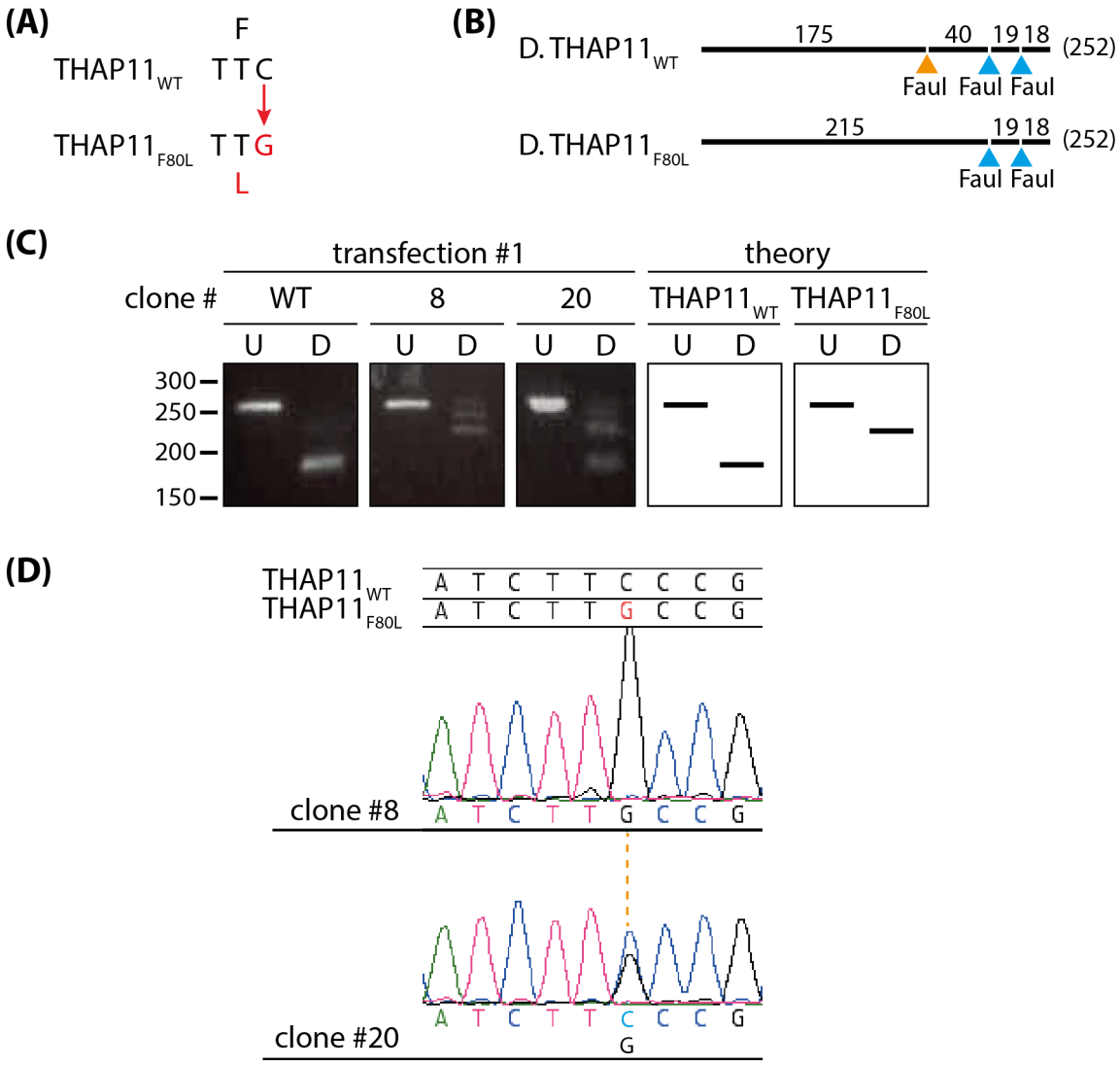


Figure 6.8: **THAP11$_{F80L}$ mutant cells.** The cell clones bearing the n.C240G, p.F80L mutation **(A)** were screened with the FauI enzyme, as the mutation disrupts one restriction site (orange triangle) **(B)**. **(C)** FauI digestion patterns of WT and mutant cell clones, compared to the theoretical ones. **(D)** Sequencing chromatograms of the mutant cell clones, compared to the WT and the expected mutant sequences, the mutated residues being depicted in red. U, undigested PCR fragment; D, FauI digested fragment.

6.1.6 Discussion of the generation of CRISPR/Cas9-designed mutant THAP7 and THAP11 HEK-293-based cell lines

Table 6.1 summarizes the different homozygous and heterozygous cell clones obtained.

I was able to obtain the 3 THAP7 functional mutant cells that I set out to obtain. I obtained 3 THAP7$_{null}$ cell clones from 2 separate transfection experiments. Thus, THAP7$_{null}$ clones # 29 can be considered as independent from clones # 32 and # 44. I also obtained 2 homozygous THAP7$_{HBM}$ mutant cell clones — as well as two others that are likely heterozygous. These two homozygous clones come from a single

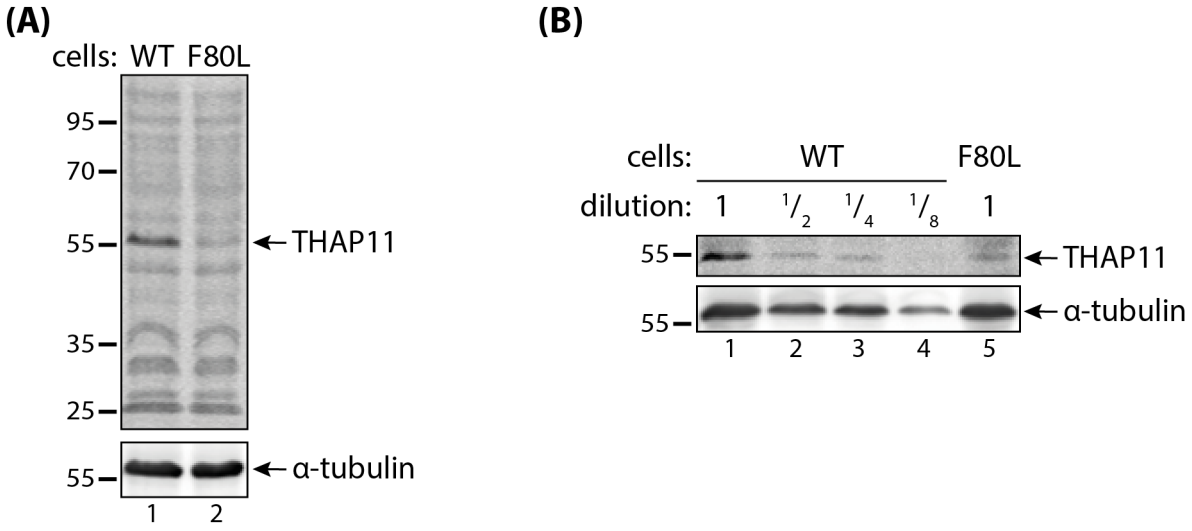


Figure 6.9: **Comparison of THAP11$_{WT}$ and THAP11$_{F80L}$ protein levels.** Whole cell lysates of the WT and THAP11$_{F80L}$, clone #8 cell lines were analyzed by immunoblot. **(A)** Equal amount of proteins were loaded and the whole gel picture is depicted. **(B)** Serial dilution of the WT cell lysate were performed, the dilution ratio being expressed relative to the undiluted lysates. α-tubulin was used as loading control.

transfection experiment, so they are not necessarily independent. Nevertheless, cells were cultured for only 3 days between the transfection and single-cell clonal selection was carried out immediatly after. Consequently, there are very few chances that the different clones originally come from the same transfected cell and thus they are likely to be independent too, even though it cannot be formally concluded as such. Finally, I got 6 different homozygous THAP7$_{\Delta CC}$ cell clones. Again, these different clones have been generated from the same transfection experiment and thus cannot be formally considered as independent — even though they are likely to be so. Among these 6 clones, 2 of them have critically impaired cell survival and/or proliferation, making them difficult to keep in culture. As I was lucky enough to get 4 additional homozygous clones, I decided to not deal with these two impaired clones and to focus on the 4 remaining ones. Nevertheless, it would have been extremely interesting to understand the reasons of the different behavior of these two specific cell clones. These cells may have off-target effects in cell survival and/or proliferation-related genes. Comparing the genomic sequence of these cell clones with the genomes of the other THAP7$_{\Delta CC}$ clones and of the WT cells would possibly unravel any underlying off-target mutated genes. To not lose focus of this PhD-thesis project, I did not do pursue this question, but still kept them frozen for possible future investigation by others.

Regarding THAP11, Table 6.1 immediately reveals that very few mutagenesis experiments were successful. No single THAP11$_{null}$ or THAP11$_{\Delta CC}$ clones, and only an heterozygous THAP11$_{HBM}$ one, were obtained, although the experiment was conducted twice for the two latter mutations. A tempting hypothesis to explain these results is that the homozygous alteration of THAP11 (the loss of either the entire protein, its coiled-

	THAP7		THAP11	
$THAP_{null}$	#29	#32, #44		
$THAP_{HBM}$	#34, #76 #42, #50		#1†	
$THAP_{\Delta CC}$	#7, #22†, #27 #36, #40, #58†			
human THAP mutation			#8, #20 F80L	

Table 6.1: **Summary of THAP7 and THAP11 mutant cell clones obtained by CRISPR/Cas9 genome editing.** Independent clones coming from separate transfection experiments are placed in separated columns. Red, homozygous mutants; blue, heterozygous mutants. †, non-viable cell clones (see text for details).

coil domain or its HCF-1 interaction) is lethal for the cells, which would explain the failure of obtaining such mutant clones. Notably, the complete loss of mouse Thap11 (Ronin) is embryonically lethal [54]. In addition, the heterozygous $THAP11_{HBM}$ mutant cells were extremely unhappy in culture and ultimately died even before we were able to freeze some, despite all the good care we provided them. This result raises the intringing hypothesis that the $THAP11_{HBM}$ mutant protein is dominant negative, impairing the WT protein's functions. To test this, I could imagine to assess the effect of the stable synthesis of an ectopic $THAP11_{HBM}$ mutant protein in WT cells. Finally, I succeeded in getting two $THAP11_{F80L}$ cell clones, one homozygous and one heterozygous. Unfortunately, I was not able to obtain any additional independent homozygous $THAP11_{F80L}$ mutant clones in a second transfection experiment. Consequently, the results obtained further with the homozygous $THAP11_{F80L}$ clones will have to be interpreted cautiously, as I will not be able to rule out any clone-specific effects. From now on and for the rest of this study, these homo- and heterozygous mutant cells will be referred as $THAP11^{F80L/F80L}$ and $THAP11^{F80L/+}$ cells, respectively. Please note that this designation is not meant to mean that the cells are diploid of the locus.

Interestingly, I demonstrated that $THAP11^{F80L/F80L}$ cells express significantly lower THAP11 protein levels compared to WT cells. This difference is not due to an altered recognition of the mutant protein by the antibody, as this antibody recognizes the last 85 amino acids of the THAP11 (amino-acids 226 to 314) while the mutation is at position 80 (p.F80L). In addition, the western blotting has been performed in denaturating conditions, thus a different conformation of the mutant protein would neither affect its recognition by the antibody. Thus, the difference observed truly reveals a difference in protein levels in the different cell lines, which can either result from a decrease in protein synthesis (transcription and/or translation) or from an increase in protein degradation. Considering the fact that the mutated amino-acid is

part of the THAP-domain "AVPTIF box" suggested to be necessary for the proper folding of the zinc finger, the F80L mutation is likely to impair the folding of the THAP11 protein, thereby leading to an increased degradation of the mutant THAP11$_{F80L}$ protein. This result is in accordance with a published abstract [83] (the corresponding paper being in the writing process) also suggesting that the THAP11$_{F80L}$ corresponding protein is unstable. I cannot rule out the possibility, however, that the THAP11$_{F80L}$ mutant protein is simply synthesized less. An hypothesis here is that, as THAP11 is a transcription factor, it may activate its own transcription — the F80L mutation preventing it from doing so. I will demonstrate in a future chapter, however, that the THAP11 p.F80L mutation does not impact THAP11 binding to its own promoter neither the *THAP11* mRNA levels in THAP11$^{F80L/F80L}$ cells (Chapter 8).

THAP7$_{null}$ and THAP7$_{\Delta CC}$ mutants were engineered by creating stop codons at the beginning of the *THAP7* coding sequence or the coiled-coil domain, respectively. I cannot, however, exclude the possibility of a read-through of the stop codons, even though I inserted two successive — and different, in the case of THAP7$_{null}$ mutant — ones to avoid this phenomenon. The effectiveness of the stop codons should thus be verified in western blotting by the absence of the full-length protein for THAP7$_{null}$ cells, and by the generation of a truncated THAP7 protein in THAP7$_{\Delta CC}$ cells. Unfortunately, the lack of a good THAP7 antibody able to detect the endogenous protein prevents me from doing such a control.

Interestingly, the ploidy of HEK-293 cells is still debated, from near-tetraploid to hypo- and hyper-triploidy [127]. Consequently, obtaining the aforementioned mutations in a homozygous state would mean to have mutated the 3 or 4 gene copies in these cells. Alternatively, observing an heterozygous clone such as THAP11$_{HBM}$ clone # 1 or THAP11$^{F80L/+}$ one, where the peaks of the WT and mutated nucleotides are of similar height on the sequencing chromatograms (Figures 6.5 D and 6.8 D) suggests that equivalent number of gene copies are WT and mutated. Relative to HEK-293 ploidy, it likely means that the cells are tetraploid (or alternatively, diploid) at the *THAP11* locus. The HEK-293 cell ploidy could also explain what was observed for THAP7$_{HBM}$ clone # 50, where the peaks of the WT nucleotides were much lower than the mutant ones on the sequencing chromatogram (Figure 6.4 D). Indeed, one can imagine that a single *THAP7* gene copy remains WT while the other ones — 2 or 3, depending the ploidy at this specific locus — are mutated, which would explain a much higher signal of the mutated residues in the sequencing. Finally, the fact that the CRISPR/Cas9 genome editing enabled me to create, in a single mutagenesis step, homozygous mutants at triploid or tetraploid locus highlights the power and efficiency of this technique.

6.2 Stable cell lines containing an inducible *THAP11* or *THAP7* gene construct

Complementary to the mutations of the endogenous *THAP7* and *THAP11* genes described above, I created stable cell lines containing an ectopic *THAP*-gene expression construct. In parallel, I suspected some of the CRISPR/Cas9-mediated mutations to be potentially lethal for the cells. I thus envisioned the use of THAP-inducible cells lines as host cells, so I could benefit from the presence of the exogenous THAP protein to keep the cells alive, and suppress its synthesis to perform experiments. Consequently, the generation of cell lines with inducible *THAP* constructs would both enable me to probe the effect of the forced synthesis of the THAP proteins of interest and increase the chances of success of the CRISPR/Cas9 mutagenesis at the same time.

6.2.1 Integration of ectopic and inducible gene constructs into human host cells

To generate such inducible cell lines, I took advantage of the Flp-In™ T-Rex™ technology described in the introduction (section 1.4). I decided to use HEK-293-based host cells as it would allow me to compare the stable cells with the CRISPR/Cas9 mutant ones, the latter being generated in HEK-293 cells.

In addition to the generation of cells containing inducible $THAP11_{WT}$ or $THAP7_{WT}$ constructs — to probe the effects of THAP11 or THAP7 forced synthesis — I also created cells with an inducible $THAP11_{HBM}$ mutant construct to determine whether the $THAP11_{HBM}$ mutant protein has a dominant-negative effect. Indeed, the only $THAP11_{HBM}$ mutant cell clone that I obtained with CRISPR/Cas9 was heterozygous and died almost immediately. This result has led to the hypothesis that the $THAP11_{HBM}$ mutant protein has a dominant deleterious effect on cells. To test this hypothesis, I thus suggested to assess the effects of the introduction of a $THAP11_{HBM}$ version into WT cells containing the endogenous $THAP11_{WT}$ protein.

The protocol to generate Flp-In™ T-Rex™ stable cells is detailed in Chapter 2. Briefly, cells were transfected together with the Flp recombinase-encoding vector (pOG44) and the pcDNA™5/FRT/TO containing a $THAP11_{WT}$, $THAP11_{HBM}$ or $THAP7_{WT}$-Flag construct. After hygromycin selection of the cells having the appropriate stable integration, cells were clonally selected before being subsequently tested. The three different *THAP* constructs were Flag-tagged to be easily detected on immunoblots.

As explained, I had initially intended to mutate the endogenous *THAP7* or *THAP11* gene with the CRISPR/Cas9 system in the stable $THAP7_{WT}$ or $THAP11_{WT}$ cells, respectively, and to benefit in parallel from the presence of the WT exogenous construct. Thus, the latter construct has to be Cas9 resistant so that the CRISPR/Cas9 systemb will selectively affect the endogenous *THAP* gene. This can be achieved by silently

mutating the PAM sequence of the exogenous *THAP* construct used in the design of the CRISPR/Cas9 mutagenesis. These stable cell lines, however, were finally not used to generate the CRISPR/Cas9-mediated mutant cells, for reasons that I will explain later. Nevertheless, the Flp-In™ T-Rex™ THAP7$_{WT}$ and THAP11$_{WT}$ cells were engineered before this decision, thus with Cas9-resistant *THAP7*$_{WT}$ and *THAP11*$_{WT}$ constructs.

6.2.2 Creation of Flp-In™ T-Rex™ THAP11$_{WT}$ cells

I started by creating the stable cells containing the *THAP11*$_{WT}$ gene construct. After hygromycin selection of the transfected cells — to select for the cells having properly integrated the construct at the FRT site — single-cell cloning was carried out. From all the clones obtained, 6 were randomly selected for further testing.

To begin with, I probed the level of the ectopic THAP11$_{WT}$-Flag protein under stimulation with a tetracycline antibiotic, here doxycycline. Figure 6.10 A shows that the six cell clones exhibit a high THAP11$_{WT}$-Flag level after 3 days of treatment with doxycycline. Also, in every clone, the level of ectopic THAP11$_{WT}$-Flag protein is much higher than the level of the endogenous THAP11 protein. Interestingly, the level of endogenous THAP11 protein does not seem to be affected by the forced synthesis of the ectopic THAP11$_{WT}$-Flag counterpart (compare lanes 5 and 6 with lane 1, in which the confluency of the cells were similar).

As the six clones behaved similarly, two were randomly chosen for further characterization: clones # 13 and # 14. To assess whether the expression of the THAP11$_{WT}$-Flag construct is properly repressed in the absence of doxycycline, cells were treated for 28 hours with, or without, doxycycline. Normal FBS, used to supplement the cell medium, is likely to contain some tetracycline, which would then lead to a basal THAP11$_{WT}$-Flag level, even in the absence of doxycycline treatment. I thus cultivated cells either in medium supplemented with normal FBS (normal medium) or switched them two days prior to the experiment into medium supplemented with tetracycline-free FBS (tet-free medium). In the absence of doxycycline, THAP11$_{WT}$-Flag is expressed at a very low, but still detectable level in clone # 13, while significantly expressed in clone # 14 (Figure 6.10 B, lower panel, compare lanes 3 and 5 and lanes 7 and 9, respectively). Importantly, the use of tet-free medium does not suppress the basal THAP11$_{WT}$-Flag synthesis in the absence of doxycycline stimulation (lane 3 versus 5, and lane 7 versus 9). To rule out the possibility that the cells were not treated for long enough in tet-free medium to have completely washed tetracycline out of the cells, and thus still contain some tetracycline at the time of the experiment, I repeated the previous experiment, but cultivating the cells for an entire week in tet-free medium before analysis. But again, the tet-free medium does not enable the suppression of the basal THAP11$_{WT}$-Flag synthesis in unstimulated cells (not shown). To conclude, the tet-free medium does not have the expected effect and I was not able to totally avoid a basal expression of the

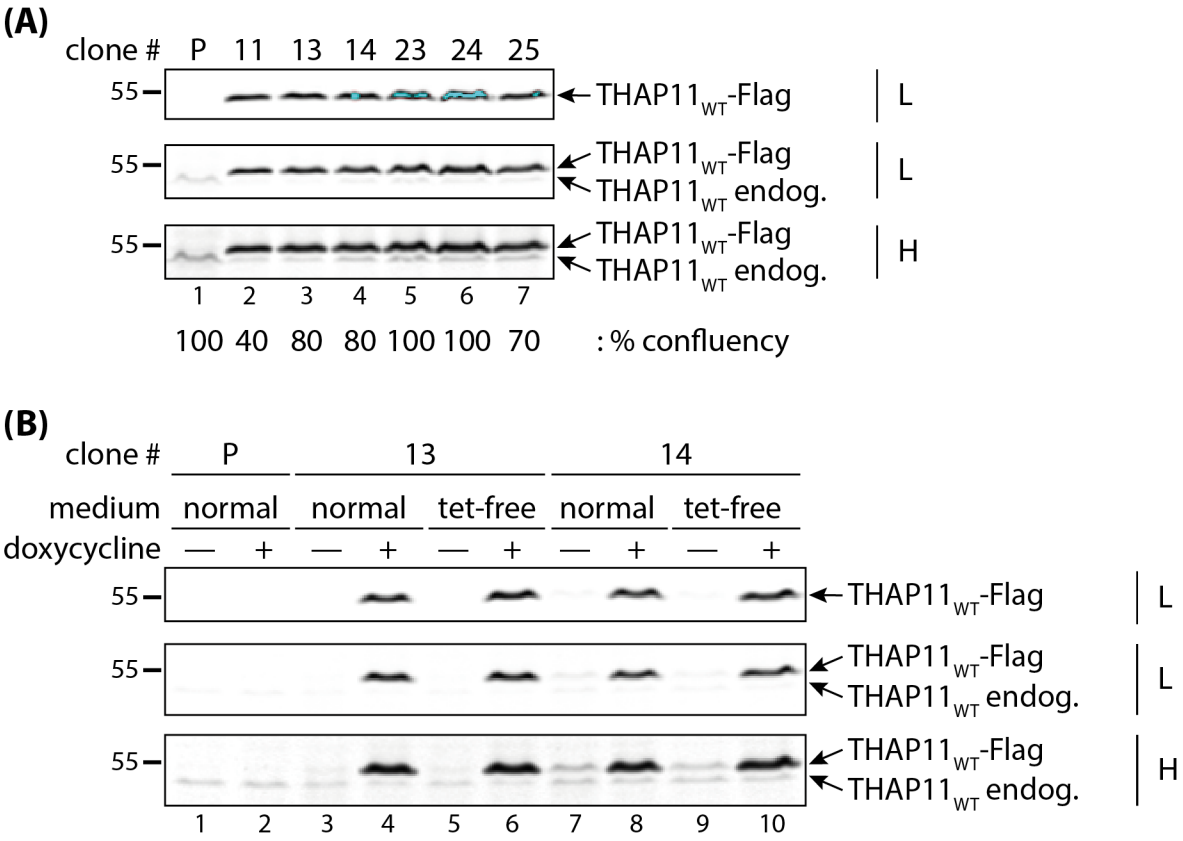


Figure 6.10: **Flp-In™ T-Rex™ THAP11$_{WT}$ cells. (A)** The 6 selected cell clones were treated during 3 days with 1 µg/ml of doxycycline and whole cell lysates were analyzed by immunoblot. The cell confluency at the time of the cell lysis is depicted. **(B)** Clones # 13 and # 14 cultured in normal or tet-free medium were treated during 28 hours with 1 µg/ml of doxycycline (+) or with DMSO as control (—), and whole cell lysates were analyzed by immunoblot. Top pannel, anti-Flag antibody; bottom 2 pannels, anti-THAP11 antibody. P, parental cells. L, low exposure; H, high exposure.

ectopic *THAP11$_{WT}$-Flag* construct. Curiously, the basal level of THAP11$_{WT}$-Flag synthesis in the absence of doxycycline varies in the differing clones. As it is lower in clone # 13 than in clone # 14, I worked with clone # 13 in future experiments. In addition, I only used normal medium, as the use of tet-free one has no evident effect.

I was also interested to determine whether I could manipulate the inducible system more finely than a simple on-off expression. For this, I probed whether using different doses of doxycycline would allow different levels of THAP11$_{WT}$-Flag synthesis. The standard doxycycline concentration to induce ectopic-construct expression is 1 µg/ml. Doxycycline titration with concentrations ranging from 0.01 to 10 µg/ml did not allow any modulation in THAP11$_{WT}$-Flag synthesis level, which is very high in all conditions (Figure 6.11 A). Thus, doxycycline is in large excess at the standard 1 µg/ml concentration, where the system is already saturated. I repeated the experiment with doxycline concentrations a thousand times lower, ranging from 0.01 to 10 ng/ml. Figure 6.11 B shows that low doxycycline concentrations induce lower THAP11$_{WT}$-Flag

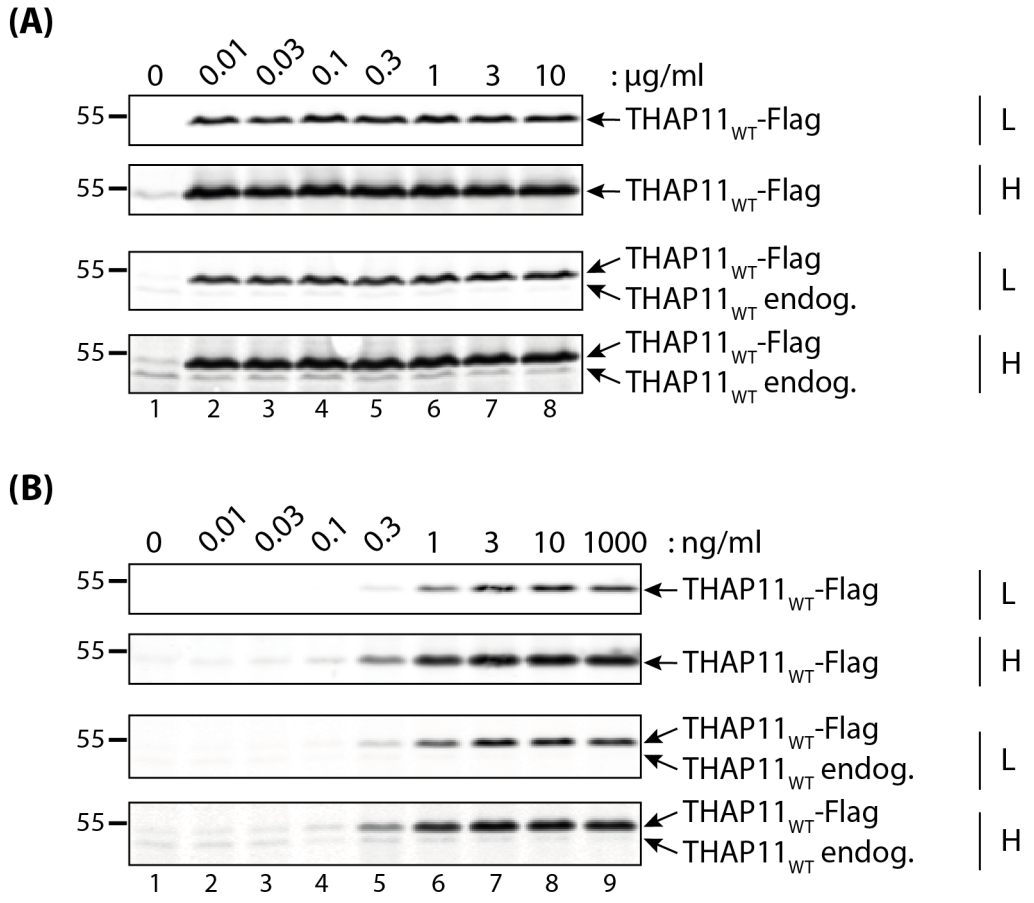


Figure 6.11: **Doxycycline dose response in Flp-In™ T-Rex™ THAP11$_{WT}$ cells.** Cells were treated with different doxycycline concentrations for 28 hours and whole cell lysates were analyzed by immunoblot. Doxycycline concentrations ranged from 0.01 to 10 µg/ml **(A)** or from 0.01 to 10 ng/ml **(B)**. Top 2 pannels, anti-Flag antibody; bottom 2 pannels, anti-THAP11 antibody. L, low exposure; H, high exposure.

synthesis, and thus the level of ectopic THAP11$_{WT}$-Flag protein can indeed be manipulated using different doxycycline concentrations. Notably, the use of doxycycline at 0.3 ng/ml (lane 5) allows for an interesting intermediate between the leaky and the very high THAP11$_{WT}$-Flag synthesis (lane 9). As a remark, I observed that doxycycline at 10 µg/ml starts to be toxic for the cells, as cell death was notably higher at this concentration than at lower ones. Indeed, doxycycline being an antibiotic, its use is potentially toxic for the cells. While the 10 µg/ml concentration appears to be toxic, I observed no obvious difference in cell viability at the lower concentrations, in both short-term and long-term experiments (see below), compared to the control DMSO treatment. Thus, except for the 10 µg/ml concentration, the other ones are not toxic — particularly, the 1 µg/ml used routinely.

So far, I only performed short term doxycycline treatments (3 days or 28 hours). I thus assessed the effect of a long-term treatment with doxycycline, on both the ectopic THAP11$_{WT}$-Flag and the endogenous THAP11 proteins. Figure 6.12 demonstrates that cells treated for 5 weeks, twice a week, with doxycycline

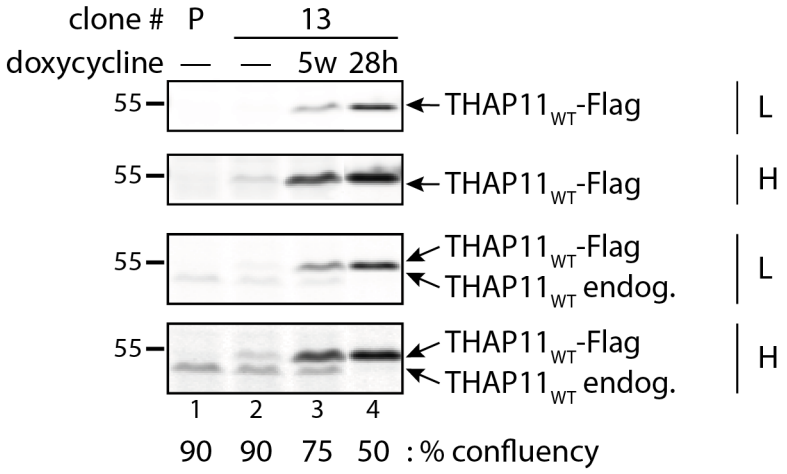


Figure 6.12: **Effect of prolonged doxycycline treatment on Flp-In™ T-Rex™ THAP11$_{WT}$ cells.** Cells were treated with 1 µg/ml of doxycycline for 28 hours (28h) or during 5 weeks (5w), twice a week, or treated with DMSO as control (—) and whole cell lysates were analyzed by immunoblot. The cell confluency at the time of the cell lysis is depicted. Top 2 pannels, anti-Flag antibody; bottom 2 pannels, anti-THAP11 antibody. P, parental cells. L, low exposure; H, high exposure.

express a lower level of THAP11$_{WT}$-Flag protein compared to cells treated for only 28 hours with the same concentration of doxycycline (compare lanes 3 and 4, taking into account the respective confluency of cells depicted below the figure). The prolonged treatement, however, does not affect the endogenous THAP11 protein, as seen by comparing lane 2 with lane 3. Please note that the endogenous THAP11 protein is barely detected in the 28-hours sample (lane 4), but this is likely due to the low confluency of cells, thus a low amount of total material.

Finally, it is interesting to note that the THAP11$_{WT}$-Flag protein level in the absence of doxycycline is generally comparable — or slightly lower depending the experiment and the cell clone used — to the level of endogenous THAP11 (e.g. Figure 6.11 A and B, lanes 1 or Figure 6.12, lane 2).

6.2.3 Creation of Flp-In™ T-Rex™ THAP11$_{HBM}$ cells

Similarly, cells containing a *THAP11$_{HBM}$-Flag* construct, in which the HBM has been disrupted, were generated. Interestingly, I did not have more trouble in obtaining these cells than in obtaining the Flp-In™ T-Rex™ THAP11$_{WT}$ ones. After hygromycin selection and single-cell cloning, 6 clones were randomly selected and tested for the synthesis of the THAP11$_{HBM}$-Flag under stimulation, or not, with doxycycline. Figure 6.13 A reveals that the six clones strongly express THAP11$_{HBM}$-Flag after 28 hours of treatment with doxycycline, the leaky synthesis of the ectopic construct differing between the clones. All of the 6 cell clones survived well in culture and did not display any obvious defect, suggesting that the presence of the THAP11$_{HBM}$-Flag protein does not severely impact the survival of these cells. For future experiments,

clone # 21 was used (a random selection). Interestingly, the level of the endogenous THAP11 protein is not affected by the synthesis of the THAP11$_{HBM}$-Flag mutant protein, the latter being lowly (no doxycyline) or strongly (with doxycycline) expressed (Figure 6.13 B for clone # 21, and not shown for the 5 other clones).

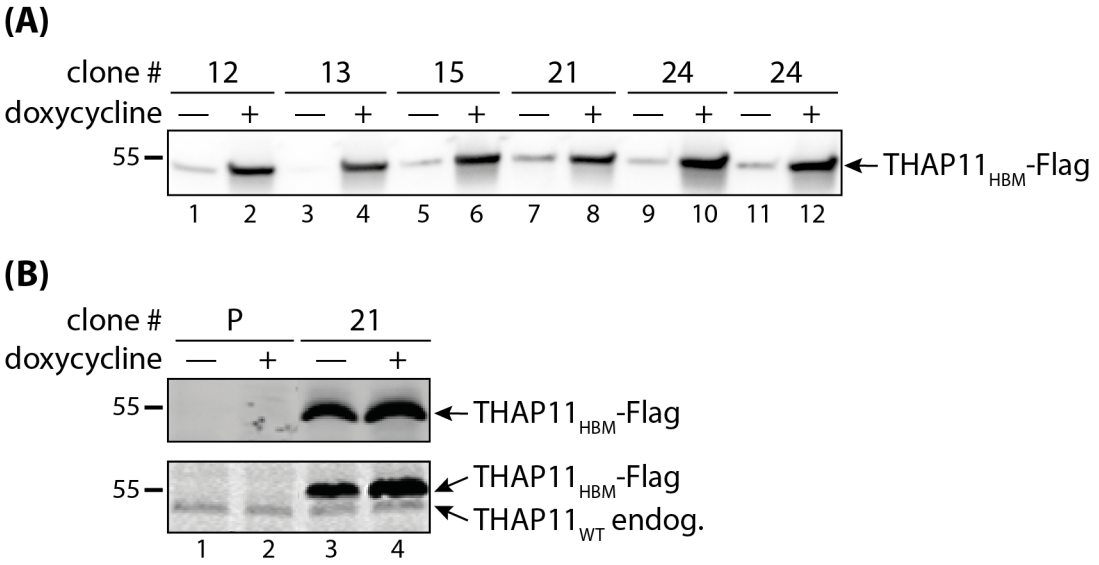


Figure 6.13: **Flp-In™ T-Rex™ THAP11$_{HBM}$ cells. (A)** The 6 selected cell clones were treated during 28 hours with 1 µg/ml of doxycycline (+), or with DMSO as control (—), and whole cell lysates were analyzed by immunoblot with an anti-Flag antibody. **(B)** Cells from clone # 21 or parental cells were treated for 6 days with 1 µg/ml of doxycycline, or with DMSO as control (—), and whole cell lysates were analyzed by immunoblot. Top pannel, anti-Flag antibody; bottom 2 pannel, anti-THAP11 antibody. Please note that this immunoblot comes from the experiment described in Figure 7.5. P, parental cells. L, low exposure; H, high exposure.

6.2.4 Creation of Flp-In™ T-Rex™ THAP7$_{WT}$ cells

With the same approach, Flp-In™ T-Rex™ THAP7$_{WT}$ cells were created, and 6 clones were randomly selected. Figure 6.14 shows that the 6 clones express high levels of the THAP7$_{WT}$-Flag protein after 3 days of doxycycline treatment. Due to the lack of a THAP7 antibody, I was unfortunately not able to probe the effect of the ectopic THAP7$_{WT}$-Flag protein on the endogenous THAP7 one. By random selection, future studies were done using clone # 14.

6.2.5 Discussion of the generation of stable cell lines containing an inducible *THAP11* or *THAP7* gene construct

Using the Flp-In™ T-Rex™ system, I was able to engineer stable cell lines synthesizing, under the control of tetracycline (here, doxycycline), one of the following Flag-tagged ectopic THAP proteins: THAP11$_{WT}$, THAP11$_{HBM}$ or THAP7$_{WT}$. The generation of such cells worked very well, and I obtained for each type

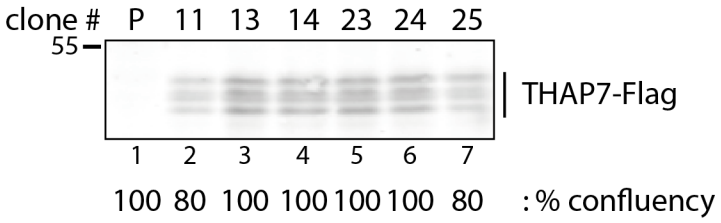


Figure 6.14: **Flp-In™ T-Rex™ THAP7$_{WT}$ cells.** The 6 selected cell clones were treated during 3 days with 1 µg/ml of doxycycline and whole cell lysates were analyzed by immunoblot. P, parental cells.

many cell clones, among which the 6 randomly tested expressed high levels of the ectopic THAP protein when treated with doxycycline. In the absence of doxycycline, the cells nevertheless possess low levels of the ectopic proteins, which vary depending on the cell clone. This low level being likely due to the presence of tetracycline in the serum used to supplement the cell culture medium, I tested the use of tetracycline-free serum. This did not, however, suppress the basal expression of the ectopic *THAP* constructs in the absence of doxycycline. Thus, the inducible aspect of these cells does not work perfectly, as I am not able to cleanly suppress the expression of the ectopic contruct in the absence of doxycycline treatment. Because of this leaky expression, these cells were not used as hosts for CRISPR/Cas9-mediated mutagenesis, as the key point was to be able to completely suppress the synthesis of the ectopic *THAP* construct. I can nevertheless make the best of this drawback by exploiting the leaky synthesis, particularly as this level is comparable to the one of the endogenous THAP protein (at least for THAP11; THAP7 was not tested for this due to the absence of a THAP7 antibody). These cells can thus be used in two ways: with no doxycycline, they express the ectopic THAP protein, at a low level similar to the endogenous THAP one, while upon treatment with doxycycline, the ectopic THAP is strongly expressed. Also, I showed that the use of different doxycycline concentrations enables me to modulate the expression level of the ectopic construct. Thus, the leaky expression and the use of different doxycycline concentration can be exploited to obtain intermediate levels of synthesis of the ectopic THAP protein (between very low to very high synthesis), which might be less artificial than the high forced synthesis.

Curiously, I observed that the THAP11$_{WT}$-Flag level is lower after a prolonged doxycycline treatment than after a short treatment (1 to 3 days). I can hypothesize that a compensatory mechanism takes place to reduce the expression of the ectopic THAP construct when it is overexpressed for a certain length of time. Alternatively, it can be the sign that some of the cells may not have the inserted THAP construct — either they have lost it, or they never had it — and that the treatment with doxycycline somehow positively favors these cells at the expense of the Flp-In™ T-Rex™ THAP11$_{WT}$ stable cells.

Interestingly, using Flp-In™ T-Rex™ THAP11$_{WT}$ cells, I demonstrated that the endogenous THAP11 protein is not affected by the presence of the exogenous THAP11-Flag one, the latter being WT or HBM

mutant, highly or lowly expressed (*i.e.*, treated, or not, with doxycycline), briefly or continuously expressed (*i.e.*, treated for a short, or a long, period with doxycycline). Thus, the THAP11 protein may not regulate the expression of its own gene.

Finally, I succeeded in creating Flp-In™ T-Rex™ THAP11$_{HBM}$ cells, and I did not encounter more difficulties with the *THAP11*$_{HBM}$ construct than with the WT one. Contrary to what I suggested earlier, the THAP11$_{HBM}$ protein might not have such a deleterious effet on cells, the WT version of the protein being present. This cell line needs to be further tested to probe whether it behaves differently than the Flp-In™ T-Rex™ THAP11$_{WT}$ one.

6.3 Conclusion

In this chapter, I have created a collection of cellular models to investigate the role of THAP7 and THAP11 proteins. THAP7 functional mutants and a THAP11 disease-associated mutant were created in the endogenous genes of HEK-293 cells: THAP7$_{null}$, THAP7$_{HBM}$, THAP7$_{\Delta CC}$ and THAP11$_{F80L}$. In addition, I obtained cell lines stably expressing THAP11$_{WT}$, THAP11$_{HBM}$ or THAP7$_{WT}$ at a low basal state, and at a high level upon treatment with doxycycline. I exploited these cell lines to probe the effect of THAP7 and THAP11 proteins on cell proliferation and gene transcription, and to decipher their mechanism of action.

Chapter 7

Cell proliferation roles of THAP7 and THAP11

In the following chapter, I describe how I used the previously generated cell lines to probe the role of THAP7 and THAP11 in cell proliferation. With the help of Philippe Lhôte, I performed proliferation experiments in which we followed the cells for 8 days.

7.1 THAP7$_{null}$ and THAP7$_{\Delta CC}$ mutations retard cell proliferation

To test cell proliferation rates, HEK-293 WT and THAP7$_{null}$ cells from clones # 29, # 32 and # 44 were seeded at the same density on day 0 and cell proliferation was followed during a week from day 1 by counting cells from two duplicate plates every 24 hours (except on days 2 and 3). Figure 7.1 displays for each time point the mean of the cell count between the two replicates (Nt), relative to the initial cell number on day 0 (No). All of these three THAP7$_{null}$ cell lines (orange, grey and yellow lines), which are homozygous, have a delay in cell proliferation compared to WT cells (blue line). The three mutant cells are able to proliferate, but at a lower rate than the WT ones. Thus, the disruption of the *THAP7* gene impairs HEK-293-cell proliferation.

Growth curves using THAP7$_{HBM}$ and THAP7$_{\Delta CC}$ cells were done similarly. Figure 7.2 shows that the two homozygous THAP7$_{HBM}$ mutant clones have different behaviors regarding cell proliferation: whereas clone # 76 cells (grey line) grow similarly to the WT cells (blue line), cells from clone # 34 (orange line) display a marked impairment in cell proliferation. Indeed, almost half as many clone # 34 cells were counted on day 8 compared to WT or clone # 76 cells. Due to the discrepancy between the two mutant cell clones,

it is difficult to draw any conclusions regarding the impact of the HBM mutation in THAP7.

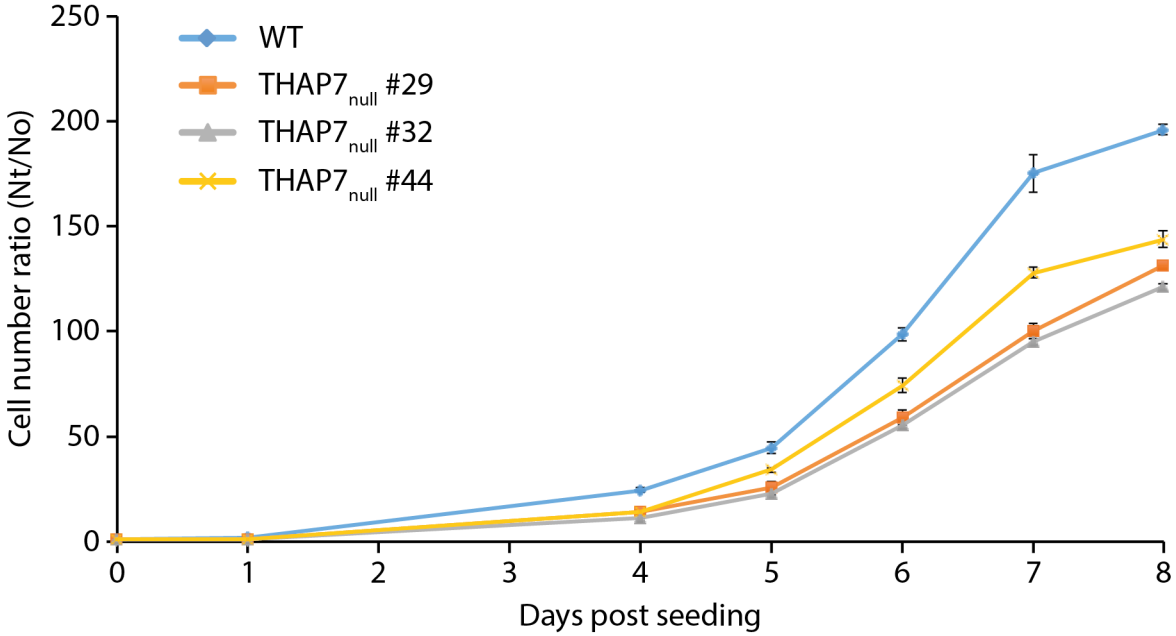


Figure 7.1: **Proliferation assays of HEK-293 WT and THAP7$_{null}$ cells.** Cells were seeded at the same density on day 0, and 2 plates of each cell line were counted every 24 hours from day 1 (except on days 2 and 3). The ratio of the mean cell count between the two replicates (Nt) and the initial cell number (No) is displayed, +/- the standard deviation.

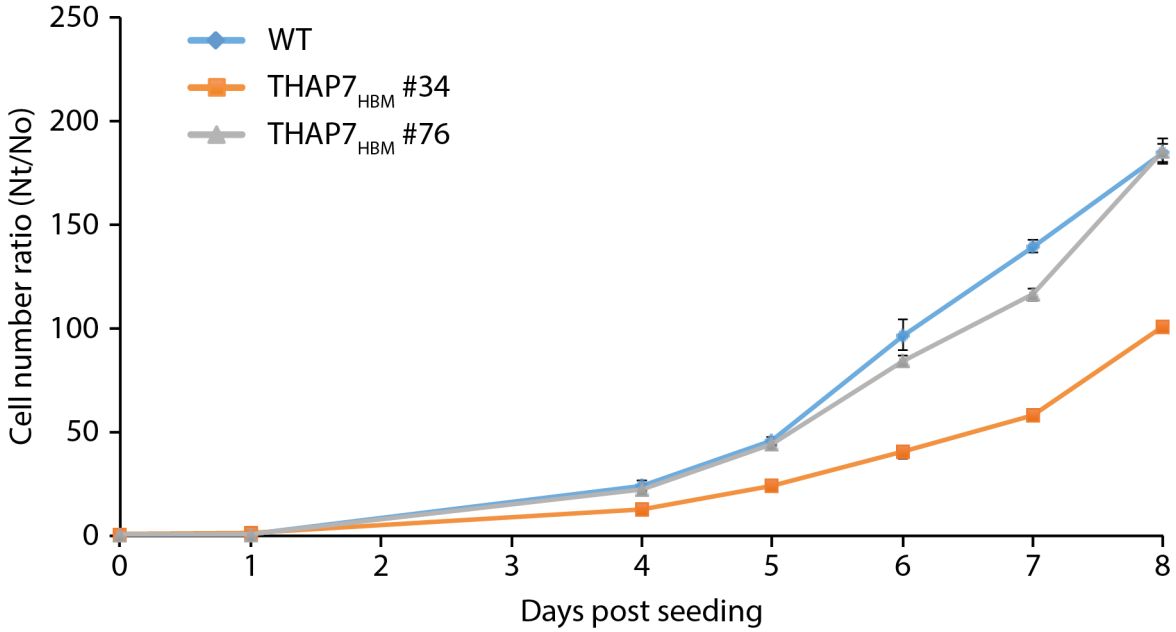


Figure 7.2: **Proliferation assays of HEK-293 WT and THAP7$_{HBM}$ cells.** Cells were seeded at the same density on day 0, and 2 plates of each cell line were counted every 24 hours from day 1 (except on days 2 and 3). The ratio of the mean cell count between the two replicates (Nt) and the initial cell number (No) is displayed, +/- the standard deviation.

Regarding THAP7$_{\Delta CC}$ cells, Figure 7.3 reveals that the 4 homozygous clones (orange, grey, yellow and green lines) consistently exhibit a slow down in cell proliferation compared to WT cells (blue line). Remarkably, the absence of the coiled-coil domain in THAP7 has the same effect as the total loss of the protein, as it slows down cell proliferation to the same extent than the THAP7$_{null}$ mutation (compare with Figure 7.1).

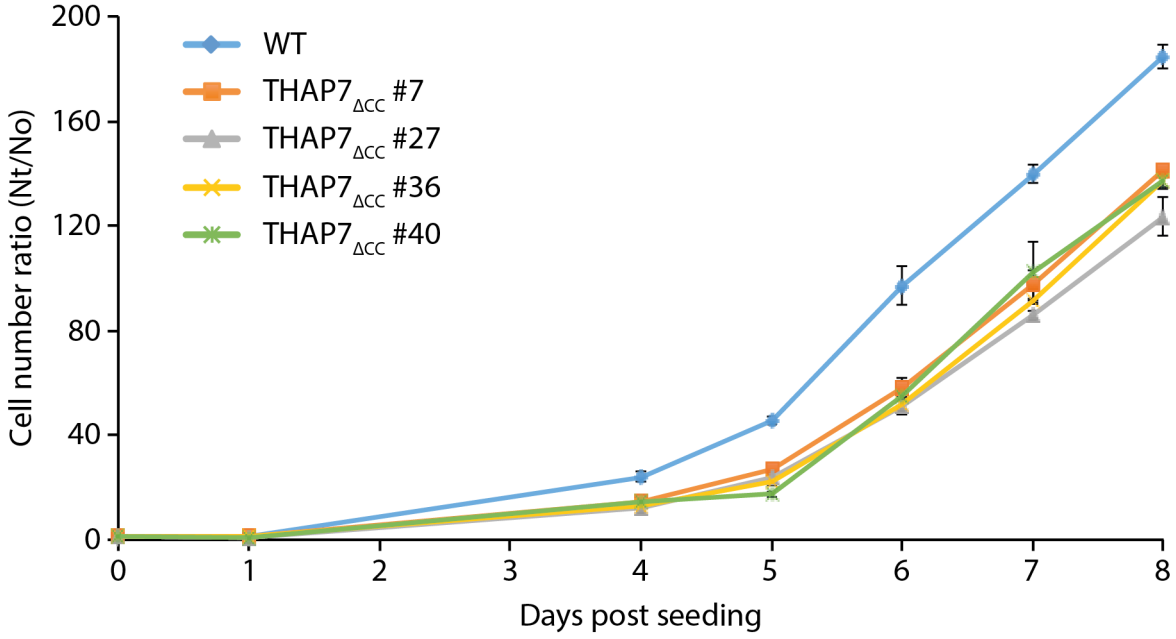


Figure 7.3: **Proliferation assays of HEK-293 WT and THAP7$_{CC}$ cells.** Cells were seeded at the same density on day 0, and 2 plates of each cell line were counted every 24 hours from day 1 (except on days 2 and 3). The ratio of the mean cell count between the two replicates (Nt) and the initial cell number (No) is displayed, +/- the standard deviation.

In conclusion, both the total loss of THAP7 protein and the truncation of its coiled-coil domain retard cell proliferation, while effect of the THAP7 HBM mutation cannot be interpreted.

7.2 Cell proliferation is moderately affected by the forced synthesis of the THAP7$_{WT}$ protein

In the same manner, I assessed the effect of the reinforced synthesis of the THAP7$_{WT}$ protein. For this, I used the Flp-In^{TM} T-Rex^{TM} THAP7$_{WT}$ cells created in Chapter 6, which can express very low or high levels of the ectopic THAP7$_{WT}$-Flag protein, depending on the treatment. I then compared the proliferation of cells with no (parental cells), very low levels (DMSO-treated) or high levels (doxycycline-treated) of THAP7$_{WT}$-Flag protein (Figure 7.4 A). The same number of cells were plated for each condition, and treated 24 hours later (day 1) with doxycycline (DOX, dashed lines), or DMSO (solid lines) as control. Comparing parental cells treated with DMSO and doxycyline (Figure 7.4 B, solid and dashed blue lines, respectively) demonstrates

that doxycycline treatment by itself does not affect cell proliferation. The leaky synthesis of THAP7$_{WT}$-Flag (DMSO-treated cells, solid orange line) slightly slows down cell proliferation compared to parental cells, the effect being more pronounced when THAP7$_{WT}$-Flag is highly expressed by stimulation with doxycycline (dashed orange line). Consequently, the ectopic synthesis of THAP7 moderatly slows down cell proliferation, the effect being THAP7$_{WT}$-Flag-dose dependent.

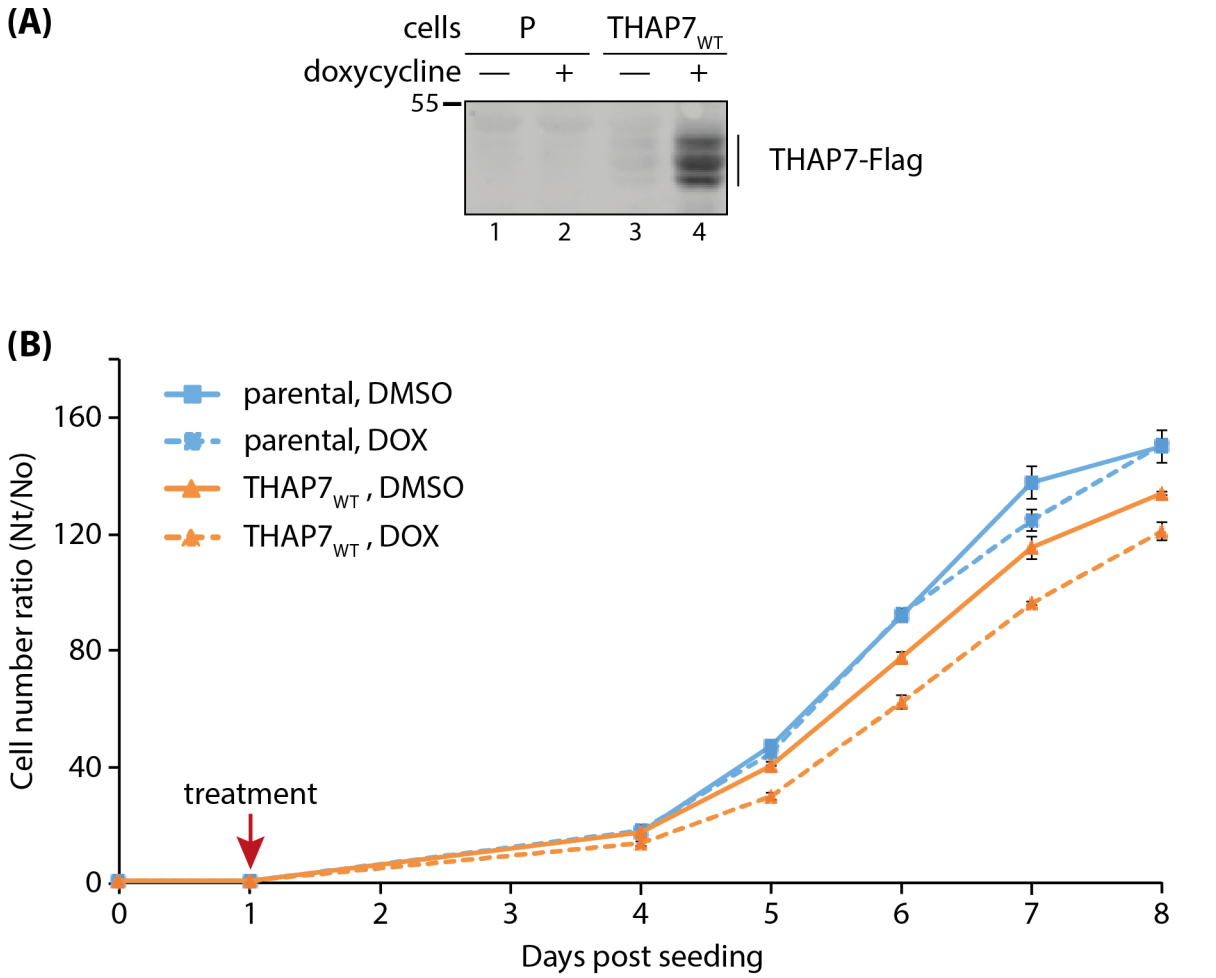


Figure 7.4: **Proliferation assays of Flp-In™ T-Rex™ THAP7$_{WT}$ cells.** Cells were seeded at the same density on day 0, and treated 24 hours later with 1 µg/ml of doxycycline (DOX, dashed line), or DMSO as control (solid line). From day 1, 2 plates of each cell line were counted every 24 hours (except on days 2 and 3). **(A)** Equal amounts of proteins from whole cell lysates of day 7 cells were analyzed by immunoblot. P, parental cells. **(B)** The ratio of the mean cell count between the two replicates (Nt) and the initial cell number (No) is displayed, +/- the standard deviation.

7.3 The forced synthesis of an ectopic THAP11 protein impairs cell proliferation

Similarly, I probed the effect of the reinforced synthesis of THAP11 on cell proliferation, using the Flp-In™ T-Rex™ THAP11$_{WT}$ cells (Chapter 6). While the leaky synthesis of THAP11$_{WT}$-Flag in the absence of doxycycline (DMSO treatment) is at a level comparable to the endogenous THAP11 protein one, using doxycycline allows to strongly induce THAP11$_{WT}$-Flag resulting in a large excess of the ectopic protein (Figure 7.5 A, lanes 3 and 4, respectively). The low synthesis of THAP11$_{WT}$-Flag (DMSO treatment) slows down cell proliferation when compared to parental cells (Figure 7.5 B, compare the solid orange line with the blue lines). In addition, the impairment in cell proliferation is more pronounced when the synthesis of THAP11$_{WT}$-Flag is reinforced by stimulation with doxycycline (dashed orange line).

In addition, the absence of an HBM sequence in the ectopic THAP11 protein (THAP11$_{HBM}$ cells) also slows down cell proliferation compared to parental cells (Figure 7.5 C, compare the solid green line with the blue lines). Again, the increased synthesis of the ectopic THAP11$_{HBM}$-Flag protein due to doxycycline treatment (Figure 7.5 A, lanes 5 and 6) further slows down cell proliferation (compare the solid and dashed green lines). Curiously, the ectopic THAP11$_{HBM}$-Flag protein is here expressed in large excess compared to the endogenous protein, even in the absence of doxycycline stimulation.

To conclude, the forced synthesis of both THAP11$_{WT}$ and THAP11$_{HBM}$ impairs cell proliferation, the decrease of proliferation rate being correlated with their respective levels.

7.4 The THAP11 cobalamin-disorder mutation impairs cell proliferation

Similarly, cell proliferation was followed for the cells bearing the cobalamin-disorder associated mutation. Growth curves were done for both the homozygous THAP11$^{F80L/F80L}$ and heterozygous THAP11$^{F80L/+}$ cells. As the stability of the THAP11 protein is affected by the p.F80L mutation (see section 6.1.5), the resulting mutant protein may be particularly sensitive to temperature change. Indeed, an increase of temperature may then destabilize even more the THAP11$_{F80L}$ mutant protein. To test this hypothesis, proliferation assays were performed at both 37 °C and 39.5 °C. For this, cells were seeded on day 0 at the same density and incubated at 37 °C. Then, 24 hours later (day 1), half of the cells were transferred at 39.5 °C for the following 7 days. Figure 7.6 shows that temperature has an effect on WT cells: as expected, cells grown at 39.5 °C proliferate slower than the ones grown at 37 °C (dashed and solid blue lines, respectively). In

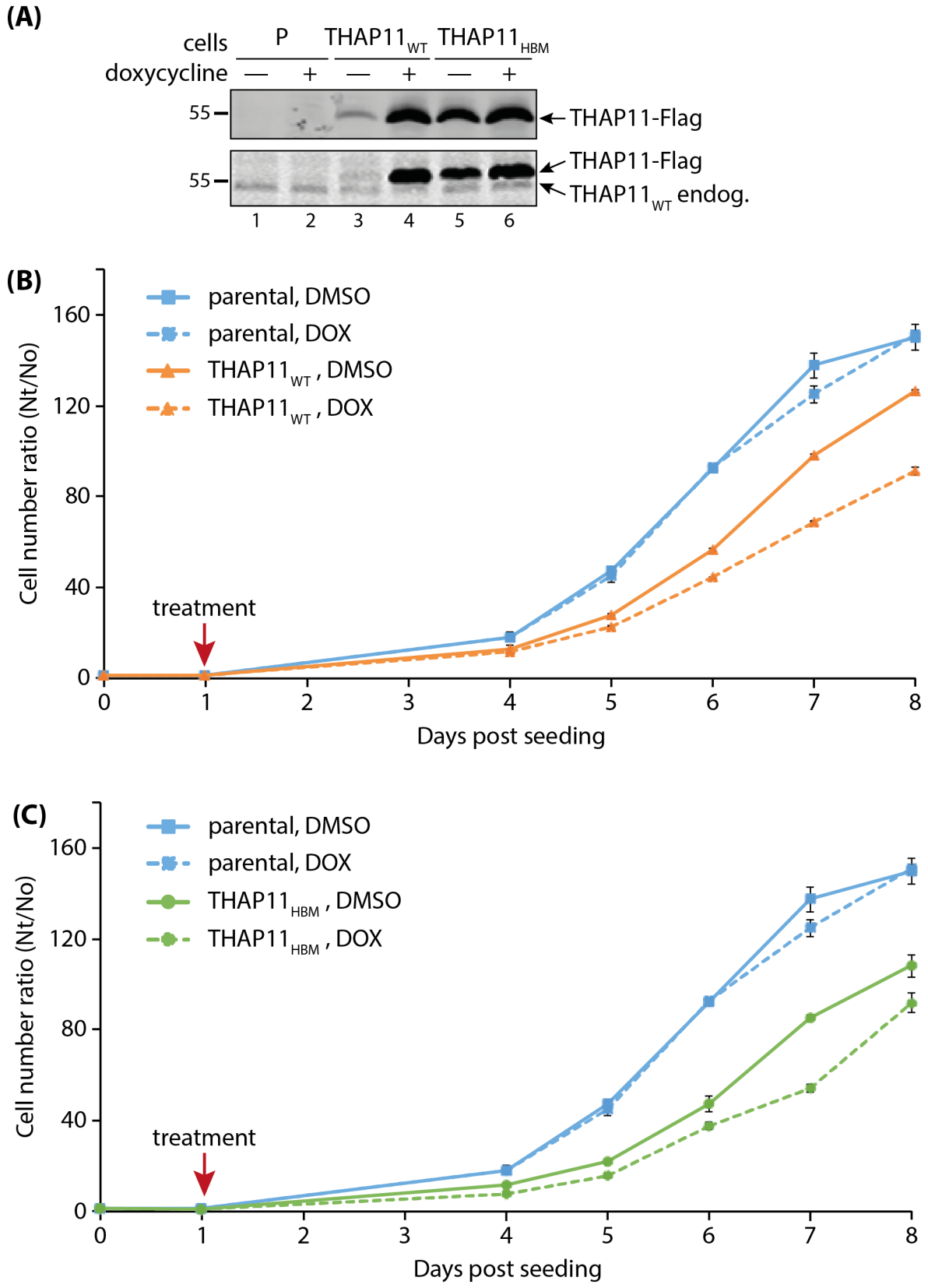

(A)
cells
P
THAP11WT
THAP11HBM
doxycycline
—
+
—
+
—
+
55
THAP11-Flag
55
THAP11-Flag
THAP11WT endog.
1
2
3
4
5
6
(B)
parental, DMSO
parental, DOX
THAP11WT , DMSO
THAP11WT , DOX
Cell number ratio (Nt/No)
160
120
80
40
0
treatment
0
1
2
3
4
5
6
7
8
Days post seeding
(C)
parental, DMSO
parental, DOX
THAP11HBM , DMSO
THAP11HBM , DOX
Cell number ratio (Nt/No)
160
120
80
40
0
treatment
0
1
2
3
4
5
6
7
8
Days post seeding

Figure 7.5: **Proliferation assays of Flp-In™ T-Rex™ THAP11$_{WT}$ and THAP11$_{HBM}$cells.** Cells were seeded at the same density on day 0, and treated 24 hours later with 1 µg/ml of doxycycline (DOX, dashed line), or DMSO as control (solid line). From day 1, 2 plates of each cell line were counted every 24 hours (except on days 2 and 3). **(A)** Equal amounts of proteins from whole cell lysates of day 7 cells were analyzed by immunoblot. P, parental cells. **(B)** and **(C)** The ratio of the mean cell count between the two replicates (Nt) and the initial cell number (No) is displayed, +/- the standard deviation.

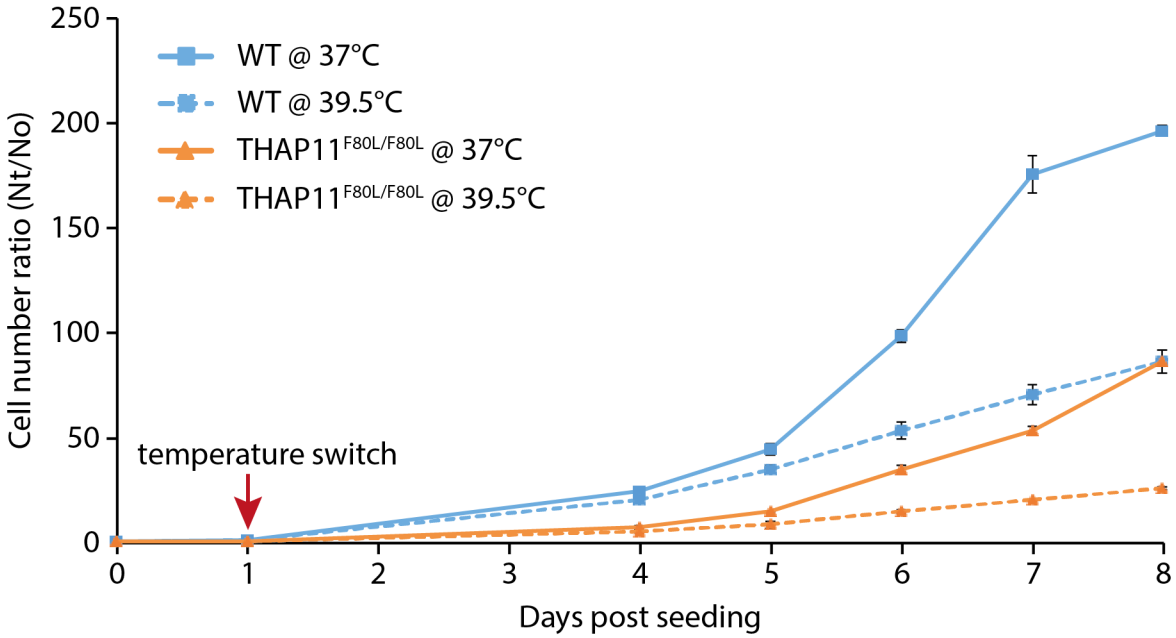


Figure 7.6: **Proliferation assays of HEK-293 WT and homozygous THAP11$^{F80L/F80L}$ cells.** Cells were seeded at the same density on day 0, and half of the cells were transferred at 39.5 °C on the following day (dashed lines) while the rest of the cells stayed at 37 °C (solid lines). From day 1, 2 plates of each cell line were counted every 24 hours (except on days 2 and 3). The ratio of the mean cell count between the two replicates (Nt) and the initial cell number (No) is displayed, +/- the standard deviation.

addition, homozygous THAP11$^{F80L/F80L}$ cells display a marked impairment in cell proliferation when compared to WT cells, which is worsened by the increase of temperature. At 37 °C, we counted less than half THAP11$^{F80L/F80L}$ cells compared to WT cells (compare solid orange and blue lanes, respectively), and almost four times less when the cells have been switched at 39.5 °C (compare dashed orange and blue lanes, respectively).

Regarding the heterozygous THAP11$^{F80L/+}$ cells, Figure 7.7 shows that they have an intermediate phenotype. When grown at 37 °C, THAP11$^{F80L/+}$ cells have a smaller cell count in days 6 and 7 compared to WT cells, but finally catch up on day 8 (compare solid orange and blue lines, respectively). In contrast, heterozygous mutant cells cultivated at 39.5 °C retain their defect in cell proliferation until day 8, where there were more than three times less THAP11$^{F80L/+}$ cells than WT cells (dashed orange and blue lanes, respectively).

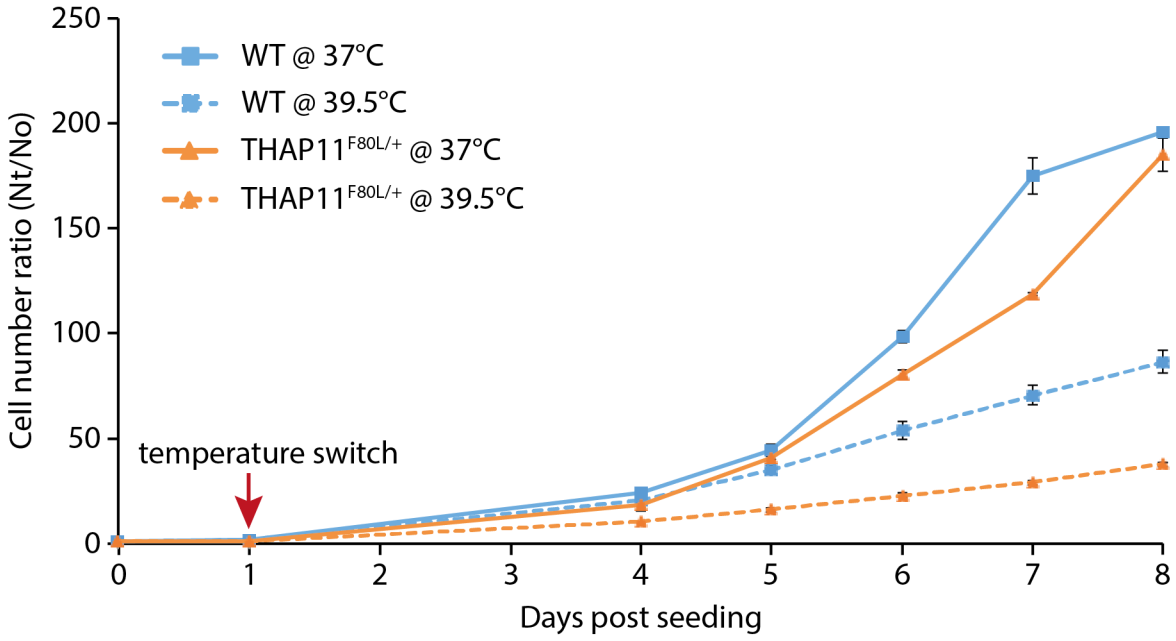


Figure 7.7: **Proliferation assays of HEK-293 WT and heterozygous THAP11$^{F80L/+}$ cells.** Cells were seeded at the same density on day 0, and half of the cells were transferred at 39.5 °C on the following day (dashed lines) while the rest of the cells stayed at 37 °C (solid lines). From day 1, 2 plates of each cell line were counted every 24 hours (except on days 2 and 3). The ratio of the mean cell count between the two replicates (Nt) and the initial cell number (No) is displayed, +/- the standard deviation.

Thus, the p.F80L mutation in THAP11 impairs cell proliferation at both temperatures, the effect being more pronounced when cells are grown at 39.5 °C. In addition, cells bearing the homozygous mutation have a more pronounced phenotype than the heterozygous ones.

7.5 Discussion

In this chapter, I have probed, with the help of Philippe Lhôte, the role of THAP7 and THAP11 proteins in cell proliferation.

I demonstrated that THAP7 is an important factor for HEK-293-cell proliferation. First, its absence in the CRISPR/Cas9 THAP7$_{null}$ cells slows down cell proliferation. Second, the lack of its coiled-coil domain in THAP7$_{\Delta CC}$ cells also markedly reduces the rate of proliferation, and its absence has a similar effect. Thus, its ability to dimerize is likely to be necessary to control cell proliferation. The role of its interaction with HCF-1, however, remains uncertain, as the two different THAP7$_{HBM}$ homozygous clones behave very differently. To conclude, THAP7 appears to be important, though not essential — as cells carry on proliferating even when it is missing — for HEK-293-cell proliferation, and its mechanism of action likely involves its dimerization. In Chapter 5, I showed that THAP7 homodimerizes, but does not form heterodimers with THAP11. Also,

homodimerization is often important for transcription-factor activies, as explained in section 1.1.2. It is thus likely that THAP7 homodimerizes to regulate gene expression and subsequent cell proliferation.

In addition, the increased THAP7 synthesis in the Flp-In™ T-Rex™ $THAP7_{WT}$ cells also slows down cell proliferation, albeit its forced synthesis has a weaker effect than its absence. Interestingly, in the latter stable cells, the reduction of cell proliferation is proportional to the level of the THAP7 forced synthesis. Thus, a precise level of THAP7 protein appears to be critical for its effect on cell proliferation, as any manipulation of its level impairs cell proliferation. Importantly, stable increased synthesis of a protein into cells is a very artificial experiment, which can be quite difficult to interprete. For instance, an hypothesis would be that THAP7 negatively regulates its own synthesis, directly or indirectly. It would explain the similar effect of the overexpression and knock-out experiments, as the forced synthesis of THAP7 would then result in a negative feedback to decrease THAP7 (endogenous) levels. Although mRNA and protein levels do not necessarily correlate, this hypothesis could be tested by probing, on western blotting, the level of the endogenous THAP7 protein in Flp-In™ T-Rex™ $THAP7_{WT}$ cells — stimulated or not with doxycycline. Alternatively, one can imagine doing THAP7 ChIP-seq and to assess whether THAP7 protein is present on the *THAP7* regulatory regions, which would be an indication of its possible direct auto-regulation. These experiments are, however, not possible here as I do not possess any good THAP7 antibody.

Otherwise, increasing THAP7 protein concentration could result in diluting out the critical THAP7 partners (e.g. HCF-1). One can imagine that having too many THAP7 molecules results in a whole set of THAP7 proteins in excess and not bound by their usual partners, which would thus be less or non functional. Thus THAP7 in excess would retain the ability to bind DNA and thus compete with other THAP7 molecules that are bound by their usual partners. More generally, the increased synthesis of any protein has the potential to result in an impaired function of the protein, by a dominant-negative mechanism. It would also explain why the phenotype is dose-dependent, as increasing the synthesis of the ectopic protein would enhance the dilution of THAP7 partners. This highlights how cautious the interpretation of an forced-synthesis experiment should be. Overall, these different results tend to show that THAP7 is a positive regulator of cell proliferation in HEK-293 cells. This conclusion, however, is in contradition with a previous study suggesting that THAP7 would repress cell proliferation as it inhibits Histone Nuclear Factor P (HiNF-P)-mediated histone H4 transcription, necessary for G1-to-S phase transition and S-phase progression [67]

For THAP11, the impossibility to obtain any $THAP11_{null}$ cell clones (Chapter 6) suggested that THAP11 is an essential protein for cell survival and/or proliferation, at least in HEK-293 cells. The results obtained here with the $THAP11^{F80L/F80L}$ and $THAP11^{F80L/+}$ mutant cells point in the same direction. In Chapter 6, I demonstrated that the homozygous p.F80L mutation results in a marked decrease in protein level, likely being the consequence of protein destabilization and subsequent degradation. In these $THAP11^{F80L/F80L}$ mutant

cells, the homozygous mutation severely hampers cell proliferation, which suggests that THAP11 positively regulates cell proliferation. This result is in accordance with clinical manifestations of the X-linked cobalamin disorder which include neurodevelopment defects (developmental delay, microcephaly, brain malformations) [126]. Indeed, the mutation in THAP11$_{F80L}$ is then likely to prevent THAP11 from positively regulating cell proliferation and subsequent development, leading to the aforementioned developmental impairments. In agreement with this, several recent studies have suggested that THAP11 (or Ronin, in mice), is involved not only in cell proliferation, but more largely in early vertebrate development of several organs, by controling the proliferation and differentiation fates of the cells [71,72,74–77,83]. In addition, heterozygous THAP11$^{F80L/+}$ mutant cells initially display a delay in cell proliferation compared to the WT cells, before catching up in the later time points. This result suggests that the cells are able to adapt and to compensate for the partial loss of the THAP11$_{WT}$ protein. This explains why the THAP11$_{F80L}$ mutation causing the *cblX*-like disease has only been identified as an homozygous state, as the heterozygous one would not probably cause any symptoms.

Interestingly, increasing the cell-growth temperature worsened the effect of the p.F80L mutation on cell proliferation. Indeed, the homozygous THAP11$^{F80L/F80L}$ mutant cells barely proliferate when they are switched to 39.5 °C. The mutant protein is likely to be even more destabilized at elevated temperature, which would explain the more severe phenotype 39.5 °C. In addition, heterozygous cells for the cobalamin-disorder mutation, which are not severely affected at 37 °C, exhibit a defect in cell proliferation at 39.5 °C almost as pronounced as the one of the homozygous mutant cells. Consequently, at this temperature, the WT THAP11 version seems unable to compensate for the heterozygous mutation. Thus, THAP11$_{F80L}$ mutation probably has an incomplete penetrance at 37 °C, while a complete penetrance at 39.5 °C. An hypothesis would be that the mutated THAP11$_{F80L}$ protein is not completely lost and probably still retains some function at 37 °C, while being more destabilized and likely non functional at 39.5 °C. More precisely, the decrease of THAP11 protein level in heterozygous cells (likely due to a degradation of the mutant THAP11 protein version) probably does not lead to a strong-enough reduction of total (WT and mutant) THAP11 protein level. When grown at 39.5 °C, the THAP11$_{F80L}$ mutant protein is likely to be even more unstable, thus even more degradaded. The resulting more severe decrease in THAP11 level may then be important enough to impact THAP11 function. It would thus be interesting to compare the THAP11 protein level in both the homozygous THAP11$^{F80L/F80L}$ and heterozygous THAP11$^{F80L/+}$ mutant cells at 37 °C and 39.5 °C. Also, it likely means that a single functional allele of THAP11 is not sufficient for THAP11 functions. Clinically speaking, the heterozygous mutation should then not cause any defect as our body is normally maintained at 37 °C. Also, THAP11 seems to mainly impact development, and to have minor effects in adult organisms. Nevertheless, one could theoretically imagine that a fetus heterozygous for the THAP11$_{F80L}$ mutation could

have developmental impairments in case of a prenatal (in-utero) exposure to fever, the latter happening when the mother suffers from fever.

In addition, I showed that the reinforced synthesis of THAP11$_{WT}$ also hampers cell proliferation in HEK-293 cells, the magnitude of the effect being proportional to the level of synthesis. Similarly to THAP7, I thus demonstrated that the increased synthesis of THAP11 as a similar effect, although weaker, than its (probable) loss of function due to the p.F80L mutation. Similarly to what I explained above, THAP11 enhanced synthesis is likely to result in a dominant-negative effect. HCF-1 has been shown to be a critical THAP11 partner for regulation of transcription and cell proliferation [29, 39, 43, 54, 70, 71, 75], thus THAP11 increased level could dilute out the HCF-1 molecules, leaving numerous THAP11 molecules unbound to HCF-1 and consequently non functional. The latter unbound THAP11 protein are unlikely to bind DNA by themselves, as THAP11 and HCF-1 binding to their common promoters has been suggested to be mutually dependent. THAP11, however, has the ability to form homodimers (Chapter 5, Figure 5.1 B), which is likely to be important for its DNA binding and subsequent regulation of transcription. Thus, an excess of THAP11 protein may result in formation of THAP11 homodimers bound by a single HCF-1 molecule, instead of two, which may not be sufficient for their synergistic transcription regulation.

Furthermore, I demonstrated that the forced synthesis of an HBM-depleted THAP11 protein — thus unable to bind HCF-1 (as shown in Chapter 5, Figure 5.1 B) — slows down cell proliferation with a dose-dependent effect. The overexpressed THAP11$_{HBM}$ protein being in such large excess compared to the endogenous THAP11$_{WT}$ one, two sorts of THAP11 dimers can be formed : either endogenous THAP11$_{WT}$ with ectopic THAP11$_{HBM}$ mutant heterodimers, or ectopic THAP11$_{HBM}$ mutant homodimers. The latter being unable to bind any HCF-1 molecule, they are likely not bound to DNA and non functional. The former heterodimer, however, would have the ability, through the endogenous THAP11$_{WT}$ moiety, to recruit a single HCF-1 protein and to bind DNA. Nevertheless, the recruitment of a single HCF-1 molecule to the THAP11 dimer may not be sufficient for their synergistic regulation of transcription. Consequently, the forced synthesis of the HBM mutant THAP11 protein is likely to also have a dominant negative effect by saturating the endogenous WT protein. This result can be linked to the fact that the heterozygous CRISPR/Cas9-mediated THAP11$_{HBM}$ mutant cell line obtained in Chapter 6 ultimately died. Indeed, the death of these cells suggested a possible dominant-negative effect of the HBM mutant protein and motivated the creation of stable cells expressing the THAP11$_{HBM}$ protein. It is suprising, however, that the effect in the mutant cells, likely expressing similar levels of THAP11$_{WT}$ and THAP11$_{HBM}$ protein, would be stronger than the one observed in the stable cells, in which the THAP11$_{HBM}$ protein is expressed at a much higher level than the WT version, even without doxycycline stimulation.

Most of the time, the experiments we performed show very similar results for different clones of the same

mutation. This is the case, for example, for THAP7$_{null}$ and THAP7$_{\Delta CC}$ clones. Thus, the results obtained there are very strong and I can be very confident. Regarding THAP11$^{F80L/F80L}$ cells, I had only one homozygous clone which theoritically does not give as much confidence in the results obtained. I demonstrated, however, that the heterozygous cells have an intermediate behavior between the WT and THAP11$^{F80L/F80L}$ homozygous cells. It then gives quite some confidence in the results obtained with the single clone. For the stable cells, however, we used a single mutant for each type of cell to perform the proliferation experiments. The results obtained should then be considered cautiously, even though the different clones of each cell type behave similarly when routinely maintained in culture. Unfortunately, in the case of THAP7$_{HBM}$ mutant cells, the two clones we used behave very differently, making any interpretation difficult. It is consequently necessary to obtain additional THAP7$_{HBM}$ cell clones to conclude on the role of the HCF-1 interaction on THAP7-mediated proliferation effect. In addition, it is likely that the two clones have different off-target effects due to the CRISPR/Cas9 mutagenesis. It would be interesting to understand the underlying differences which would explain their different behaviors. For this, one could sequence and compare their genome to identify specific off-target mutations on each clone.

Here, I note that although most of my custom cell lines have exhibited an impaired cell proliferation compared to WT cells, none of the mutation or over-synthesis experiments resulted in a complete suppression of cell proliferation, or even in cell death. This was actually expected, as the process of the cell line creation had necessarily selected cells able to proliferate to a certain extent. Indeed, both the CRISPR/Cas9 mutagenesis and the creation of stable cells involved a step of clonal selection, in which cells have to survive and proliferate at least to a certain extent, to be able to be selected and tested. It then automatically gets rid of the non-proliferating and dying cells. It is thus important to keep this in mind as it can be responsible from a bias in the cell line clones selected and the results I obtained using them.

Overall, these different results demonstrate that THAP7 and THAP11 are, at least in HEK-293 cells, positive regulators of cell proliferation that need to be expressed at a precise level. Their respective mechanisms of action, however, remain largely unclear. For instance, previous studies suggested that both THAP proteins regulate G1-to-S phase transition and S-phase progression during the cell cycle [29,67,74]. To probe this, one could compare the cell-cycle progression of WT and CRISPR/Cas9 mutant cells — particularly, THAP7$_{null}$ and THAP11$^{F80L/F80L}$ cells — using flow cytometry as described in Chapter 4 for synchronized cells. In addition, it would be interesting to assess the proliferative role of THAP7 and THAP11 in other cell types, to verify that the latter conclusions are not specific to HEK-293 cells.

In the rest of this study, however, I more globally examined the transcriptional role of THAP7 and THAP11, which is likely to underly their proliferation effects.

Chapter 8

Role of THAP11 on gene transcription

Using the aforementioned stable and mutant cell lines, genomic analyses were performed to decipher how THAP7 and THAP11 regulate gene transcription and subsequent cell proliferation. Here, I described the analysis of the data related to the THAP11 protein.

In this chapter, I thus describe the state-of-the-art high-throughput RNA sequencing (RNA-seq) performed on the THAP11$^{F80L/F80L}$ and THAP11$^{F80L/+}$ mutant cells as well as the Flp-In^{TM} T-Rex^{TM} THAP11$_{WT}$ and THAP11$_{HBM}$ stable cells. Also, I present the chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) executed on the WT and THAP11$^{F80L/F80L}$ mutant cells. Importantly, I was lucky enough to benefit from the valuable help from Philippe Lhôte and Maykel Lopes, two lab technicians, for the experimental part. The subsequent sequencings were executed by the Lausanne Genomic Technologies Facility (LGTF). Also, the computational data analysis was performed by Dr. Viviane Praz, our lab bioinformatician.

8.1 Transcriptomic analysis of THAP11 stable cells supports the involvment of THAP11 in cell proliferation and development

To begin with, we performed a large scale transcriptomic analysis on the Flp-In^{TM} T-Rex^{TM} cell lines expressing THAP11$_{WT}$ and THAP11$_{HBM}$. A single clone per cell type (the same as used in Chapter 3) was used, but technical duplicates were done for each of them. Parental Flp-In^{TM} T-Rex^{TM} cells were used in parallel as a reference. Cells were grown and treated for 36 hours at 37 °C with DMSO (all cells) or doxycycline (all but Flp-In^{TM} T-Rex^{TM} THAP11$_{HBM}$ cells). Total RNA was then extracted, the ribosomal RNA removed and the resulting RNA was subjected to single-read high-throughput sequencing.

Dr. Viviane Praz started by summarizing the normalized RNA-seq data using a principal component analysis (PCA). Indeed, PCA allows a simplification of the complexity of high-dimensional data, which is the case for large-scale experiments such as RNA-seq. By transforming the data, PCA summarizes the large number of correlated variables into a smaller number of new, uncorrelated variables. The latter are named principal components (PC1, PC2, ...). The first principal component accounts for the largest possible variability within the data, and each of the following principal components accounts for the largest possible part of the remaining variability. As a result, PCA allows one to identify patterns within the data and to highlight similarities and differences. In addition, it allows a graphical representation of such high-dimensional data, in other words it enables to visually see how the data look like.

The PCA plot in Figure 8.1 shows the mean of the 2 technical replicates (square) with the dashed-line circle representing the two standard errors for each principal component. The first principal component (PC1), represented on the X-axis and accounting for 28% of the variance, mainly segregates the different cell lines. Curiously, the THAP11$_{WT}$ and THAP11$_{HBM}$-expressing cells segregate on opposite sides of the parental cells: while the Flp-In^{TM} T-Rex^{TM} THAP11$_{WT}$ cells are on the right of the parental ones, the Flp-In^{TM} T-Rex^{TM} THAP11$_{HBM}$ are on the left.

The second principal component (PC2), on the Y axis and responsible for 12% of the remaining variance, mainly differentiates between the treatments (DMSO versus doxycycline). Curiously, the treatment seems to have opposite effects regarding the PC2 component for the parental and THAP11$_{WT}$-expressing cells. Indeed, doxycycline treatment induces a marked displacement along the PC2 (Y) axis towards to top in the parental cells, while this displacement is very small and towards the bottom for Flp-In^{TM} T-Rex^{TM} THAP11$_{WT}$ cells. Interestingly, the standard-error circle is elongated along the Y-axis for each sample, suggesting that the two replicates segregate along the Y axis. Thus, the PC2 component may also account for the difference between the replicates, meaning, the technical and sequencing variability.

Dr. Viviane Praz also did several differential analyses between the different cell lines and conditions, followed by Gene Ontology (GO) analyses. First, when analyzing the effect of the treatment on the parental cells, only one gene, *ZACN*, is differentially expressed between parental cells treated with DMSO or doxycycline. This gene is also upregulated in the doxycycline-treated Flp-In^{TM} T-Rex^{TM} THAP11$_{WT}$ sample compared to the DMSO-treated one. Thus, doxycycline treatment at the concentration used is likely to have a negligible adverse effect besides the activation of the ectopic protein, as it is limited to the upregulation of the sole *ZACN* gene.

Second, we compared the gene expression of parental and Flp-In^{TM} T-Rex^{TM} THAP11$_{WT}$ cells, in both treatment conditions. When treated with DMSO, 374 genes were upregulated and 564 were downregulated in the Flp-In^{TM} T-Rex^{TM} THAP11$_{WT}$ cells compared to the parental ones. Upon doxycycline treatment, 358

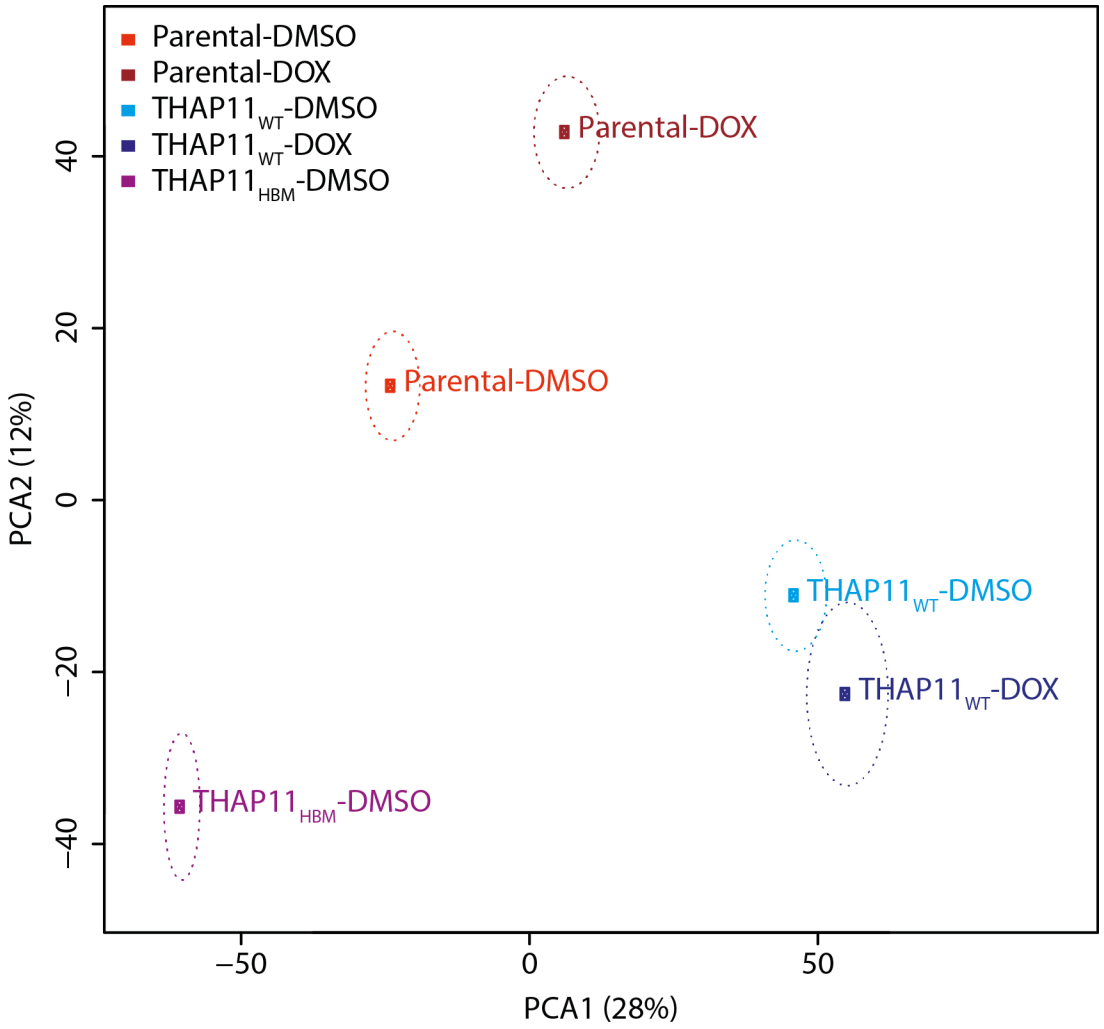


Figure 8.1: **Principal component analysis of RNA-seq data from THAP11 stable cell lines.** PCA1 and 2 resulting from RNA-seq data of Flp-In™ T-Rex™ parental (red), THAP11$_{WT}$ (blue) and THAP11$_{HBM}$ cells (purple), treated during 36 hours with 1 µg/ml of doxycycline (DOX, dark colors) or with DMSO (light colors) as control. The squares represent the mean of the 2 technical replicates with the dashed-line circle representing the standard error.

genes are upregulated and 1043 are downregulated in the Flp-In™ T-Rex™ THAP11$_{WT}$ cells compared to the parental ones. These numbers immediately suggest that the treatment with doxycycline strenghtens the effects of THAP11$_{WT}$ ectopic synthesis, which is what has already been observed when assaying the proliferation of such cells. Indeed, I observed that Flp-In™ T-Rex™ THAP11$_{WT}$ cells proliferate slower than parental ones, the effect being more pronounced upon treatment with doxycycline (see Chapter 7).

Dr. Viviane Praz submitted the list of differentially expressed genes to Gene Ontology (GO) analysis. The results for cells treated with DMSO or doxycycline were very similar, the ones obtained with doxycycline-treated cells being even stronger (Supplemental Tables S1, S2, S3 and S4). This was actually expected as the doxycycline treatment increases the ectopic THAP11$_{WT}$-Flag synthesis and reinforces the phenotypic effects

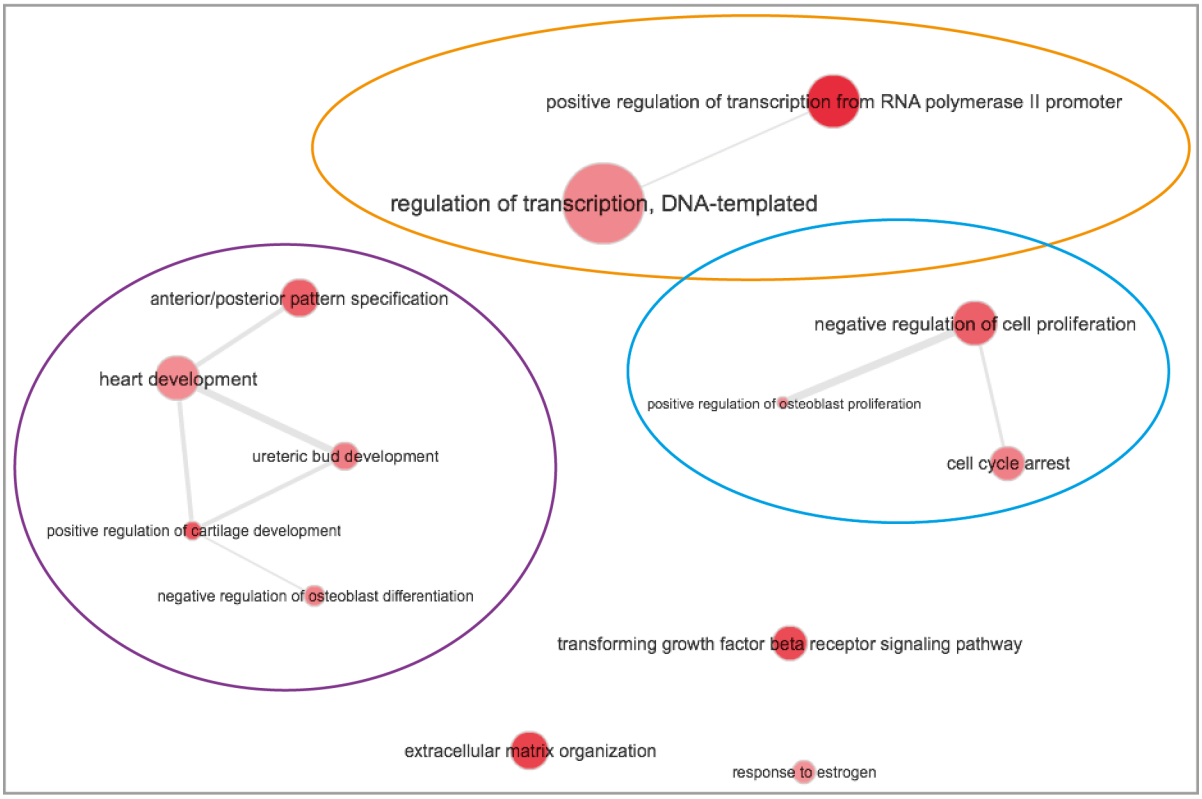
as seen in Chapter 7. Upregulated genes in the Flp-In™ T-Rex™ THAP11$_{WT}$ cells reveal pathways mainly related to development (purple), cell proliferation (blue), and also in transcription (orange) (Figure 8.2 A for doxycycline-treated cells, showing a visual summary of GO terms associated to downregulated genes, and Supplemental Tables S1 and S3 for the full lists of GO terms obtained for DMSO and doxycycline-treated cells, respectively). Alternatively, genes downregulated in the Flp-In™ T-Rex™ THAP11$_{WT}$ cells are associated to mitochondria (GO cellular composents: mitochondrion, mitochondrial matrix, mitochondrial inner membrane; green dots) (Supplemental Tables S2 and S4 for the full lists of GO terms obtained for DMSO and doxycycline-treated cells, respectively). These analyses again support the idea that THAP11 is involved in gene transcription, cell proliferation and development. They also make sense with what has been observed in the proliferation assays, meaning that THAP11 is involved in cell proliferation, with an increased phenotype when it is more highly overexpressed (doxycycline treatment) (Chapter 7).

Of note, many fewer genes are differentially expressed between Flp-In™ T-Rex™ THAP11$_{WT}$ cells treated with DMSO and doxycycline (27 upregulated and 66 downregulated) and the GO analyses did not identify any relevant pathway affected (Supplemental Table S5). Together with the proliferation experiments (Chapter 7), the latter results strenghten the idea that most of the effect due to the ectopic THAP11 protein is already triggered by its leaky synthesis. The reinforced THAP11 synthesis due to doxycycline treatment only enhances the phenotype slightly.

Third, we analyzed the gene expression of the Flp-In™ T-Rex™ THAP11$_{HBM}$ cells to probe whether the HBM sequence is necessary for the role of THAP11 on gene transcription. Comparing parental and Flp-In™ T-Rex™ THAP11$_{HBM}$ cells treated with DMSO, a small number of genes were differentially expressed. Indeed, only 68 genes were upregulated and 167 downregulated upon forced synthesis of the ectopic THAP11$_{HBM}$ protein. In addition, GO analyses did not reveal any significant pathway affected (Supplemental Table S6). Thus, gene expression does not seem to be severely affected by the forced synthesis of the THAP11$_{HBM}$ ectopic protein. In addition, 479 genes were upregulated and 577 downregulated in the Flp-In™ T-Rex™ THAP11$_{WT}$ cells compared to the Flp-In™ T-Rex™ THAP11$_{HBM}$ ones (DMSO treatment). When submitting these lists of genes to GO analysis, it revealed an association with mitochondria for downregulated genes (green dots) and pathways linked to development, proliferation and transcription for upregulated genes (purple, blue and orange, respectively; Figure 8.2 B for a visual summary of GO terms associated to downregulated genes, and Supplemental Tables S7 and S8 for the full lists of GO terms). The results are actually the same than what was obtained when comparing the parental and THAP11$_{WT}$-expressing cells (see above). Consequently, the THAP11 HBM sequence appears to be necessary for THAP11 to alter gene transcription.

To conclude, transcriptomic analyses using THAP11 stable cells demonstrated that THAP11 is implicated,

(A) GO biological processes, genes upregulated in the Flp-In T-Rex THAP11$_{WT}$ cells compared to parental cells (doxycycline treatment)



(B) GO biological processes, genes upregulated in the Flp-In T-Rex THAP11$_{WT}$ cells compared to Flp-In T-Rex THAP11$_{HBM}$ cells (DMSO treatment)

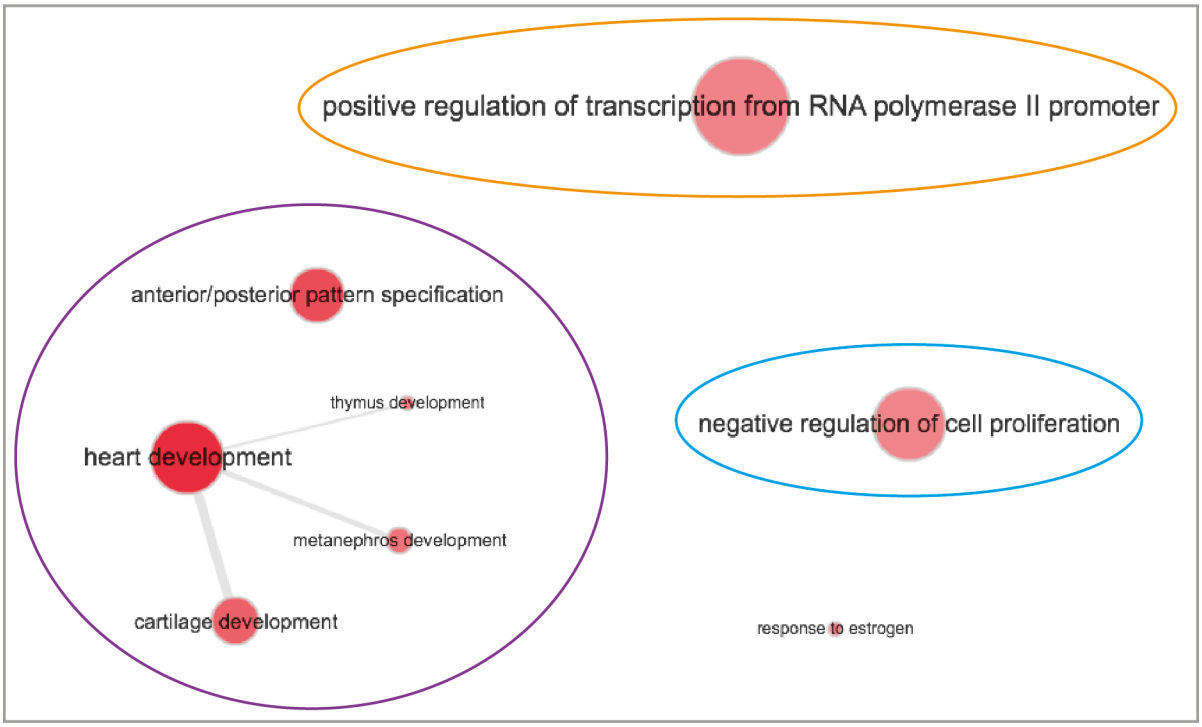


Figure 8.2: **Visual summary of Gene-Ontology analyses of RNA-seq data from THAP11 stable cell lines.** List of upregulated and downregulated genes in the interesting sample comparisons were separately submitted for Gene-Ontology analysis and visualized using REVIGO and the interactive graph tool. GO biological processes from **(A)** genes upregulated in the Flp-In™ T-Rex™ $THAP11_{WT}$ cells compared to WT cells, treated with doxycycline and **(B)** genes upregulated in the Flp-In™ T-Rex™ $THAP11_{WT}$ cells compared to Flp-In™ T-Rex™ $THAP11_{HBM}$ cells, treated with DMSO. The size of each GO term is proportional to its associated p value resulting from the GO analysis output.

in an HBM-dependent manner, in the regulation of transcription, development and cell proliferation. In addition, its function appears to be tightly linked to mitochondria.

8.2 Genomic analyses of the cobalamin-disorder associated THAP11 mutation

Previous analyses of the $THAP11^{F80L/F80L}$ and $THAP11^{F80L/+}$ cell lines showed a severe proliferation impairment of these cells compared to WT ones. To decipher the mechanism underlying the THAP11 cobalamin-disorder mutation, we used two complementary approaches on the mutant cells:

- a genome-wide analysis of THAP11 DNA association in WT and mutant cells, using ChIP-seq against the THAP11 protein in the homozygous $THAP11^{F80L/F80L}$ cells;
- a transcriptomic analysis using RNA-seq, to compare the gene expression profiles between WT, $THAP11^{F80L/F80L}$ and $THAP11^{F80L/+}$ cells.

8.2.1 THAP11 DNA association is selectively disrupted at a subset of DNA sites by the p.F80L mutation

As explained, the THAP11 p.F80L mutation affects the last amino acid of the "AVPTIF" box of the THAP domain, which has been suggested to be involved in the proper folding of the THAP domain. The THAP domain being responsible for DNA association, the resulting mutant protein could have an impaired ability to bind to DNA. Also, previously published studies [75, 83], as well as my own experiments described above, suggest that the resulting mutant protein is likely less stable than the WT protein.

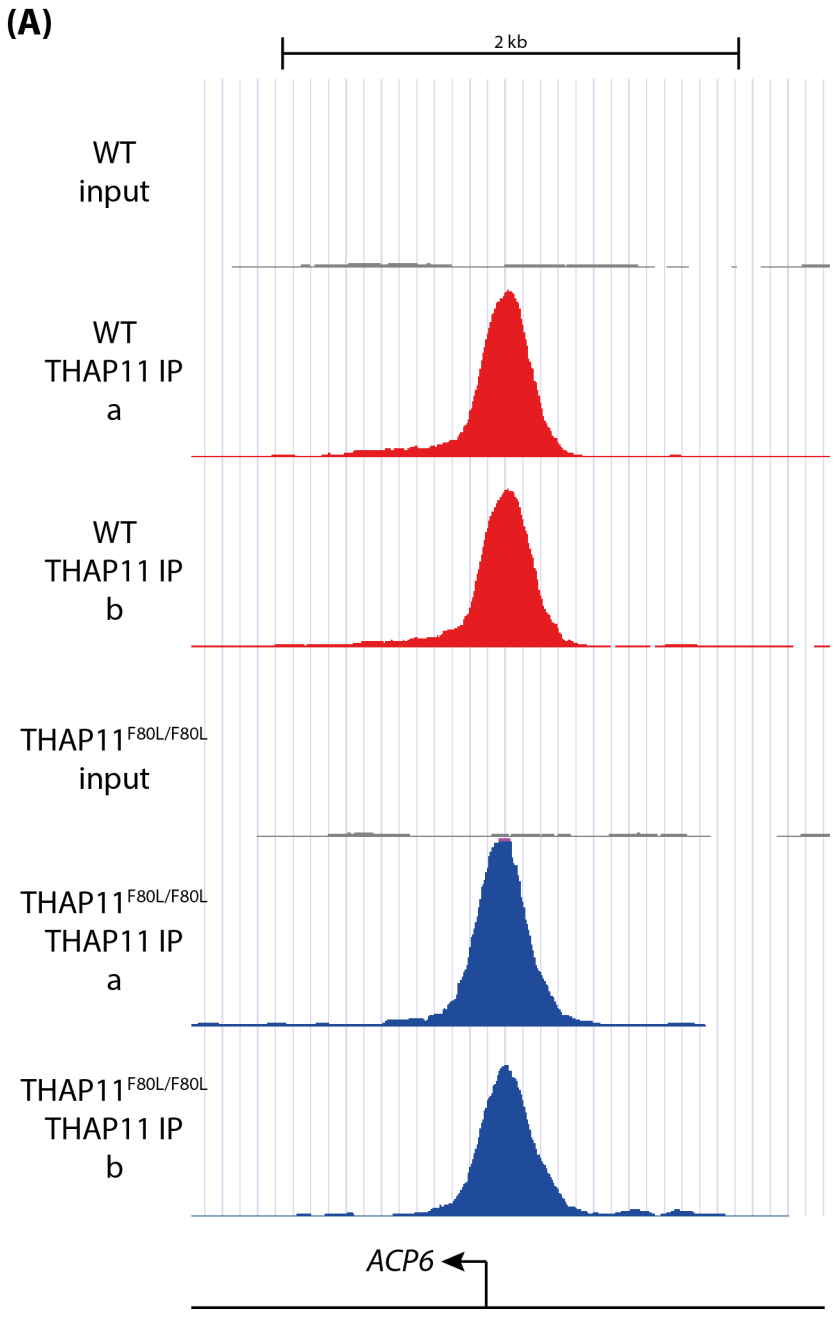
To probe the effect of the THAP11 p.F80L mutation on THAP11 DNA binding, Maykel Lopes performed a chromatin immunoprecipitation against the THAP11 protein, followed by high-throughput sequencing, in both the WT and the homozygous $THAP11^{F80L/F80L}$ cells. Crosslinked chromatin from these cells was sonicated and immunoprecipitated by an antibody directed to the C-terminus of THAP11, thus

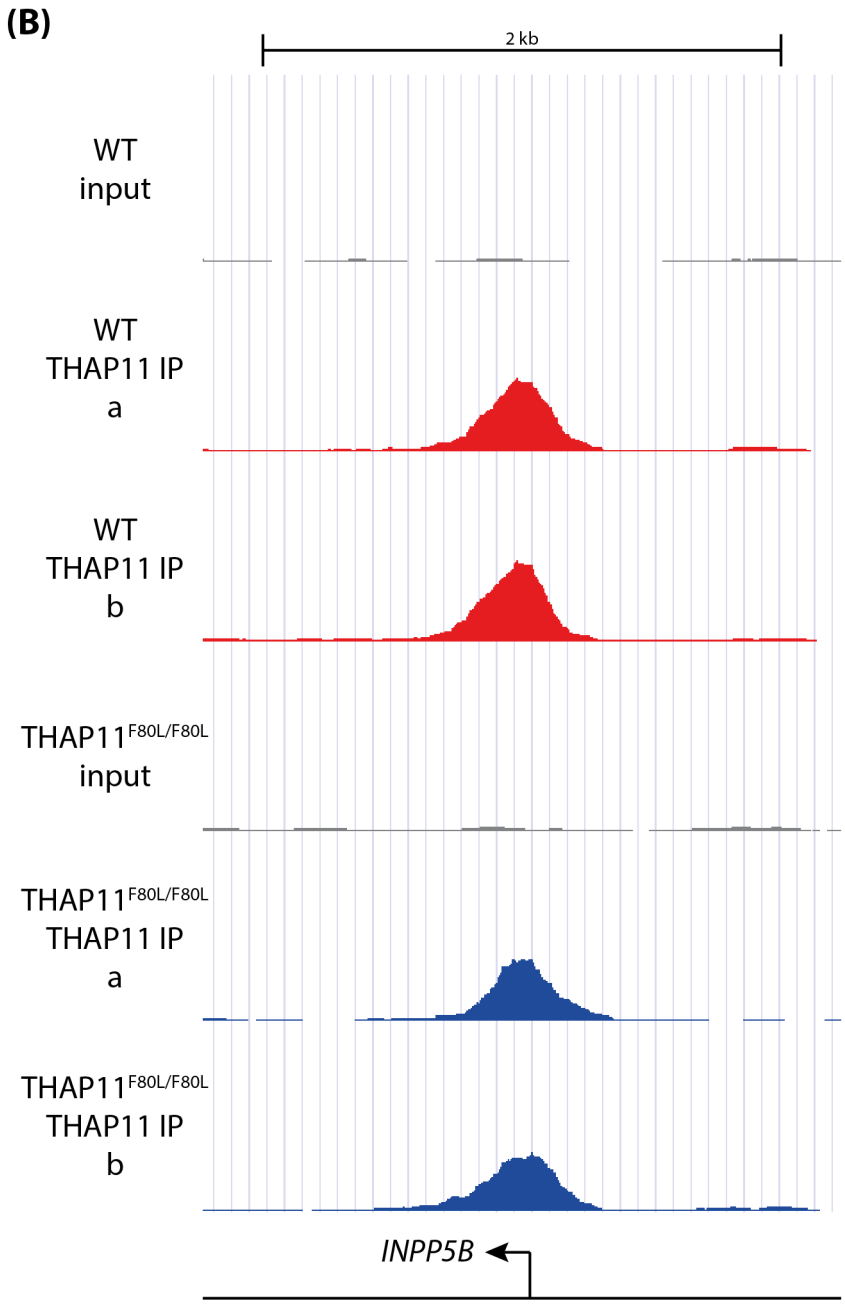
not competing with its DNA binding. The recovered chromatin and the input — the total, meaning non-immunoprecipitated, chromatin — were decrosslinked and submitted to paired-end sequencing. The data were analyzed by Dr. Viviane Praz. After mapping the reads onto the human genome, peaks were detected in each sample. Each of THAP11 ChIP, in the two different cell lines, was sequenced in duplicate (two different libraries were prepared from the same immunoprecipitated chromatin and sequenced). Comparing the intersection of peaks between the replicate samples, we observed that 72 to 90% of the peaks in one of the two replicate are found in the other one, depending the relative comparison made (data not shown). Also, Figure 8.3 shows the two pairs of sequencing duplicates for 3 peak examples, and demonstrates how similar the data are inside each replicate pair. Consequently, the sequencing data are extremely robusts. For future visualization of peaks, a single replicate per sample is shown (replicate (a) for each sample). Also, for each cell line, the sum of the peaks found in the duplicates will be considered. Naturally, when a peak is present in both replicates, it is only counted once. If needed to choose between the peak data coming from one or the other replicate, sample (a) was used. When later quantifiying the peaks, the mean between the replicates was considered.

Figures 8.4 A and C show that 2341 peaks were detected in the WT cells, and only 1473 in the THAP11$^{F80L/F80L}$ cells, with about half of the peaks being located close to an annotated transcription start site (TSS; 250 bp on each side).

These peaks were classified into three categories: (i) peaks present in both the WT and the THAP11$^{F80L/F80L}$ samples ("common"); (ii) peaks present only in the WT sample ("WT specific"); (iii) peaks present exclusively in the THAP11$^{F80L/F80L}$ sample ("mutant specific"). Figures 8.4 B and C summarize the total number of peaks identified in each of the 3 categories, and the number and percentage of these peaks that are located close to a TSS (+/- 250 bp). It shows that more than half of common peaks are located close to a TSS, while only a third of WT-specific ones are. Interestingly, only 3 peaks close to a TSS were identified as specific to the THAP11$^{F80L/F80L}$ sample. Visualization of these peaks using the UCSC genome browser showed that the 3 mutant-specific TSS-associated peaks are actually very low peaks that can be considered as background (Figure 8.5). Thus, the THAP11$^{F80L/F80L}$ mutant protein remains bound to only a subset of its promoter targets, and does not exhibit any evident *de novo* promoter binding compared to the WT protein. For the rest of the study, I thus focus only on the common and WT-specific peak categories.

The effect of the THAP11 p.F80L mutation is DNA-site specific. For instance, THAP11 is located at the *MMACHC* promoter in WT cells, but absent in the THAP11$^{F80L/F80L}$ mutant cells (Figure 8.6 A). As a reminder, the associated MMACHC protein is an essential enzyme in the cobalamin pathway, and 90% of cobalamin disorders result from mutations in the *MMACHC* gene. The loss of *MMACHC*-promoter


(A)
2 kb
WT
input
WT
THAP11 IP
a
WT
THAP11 IP
b
THAP11F80L/F80L
input
THAP11F80L/F80L
THAP11 IP
a
THAP11F80L/F80L
THAP11 IP
b
ACP6


(B)
2 kb
WT
input
WT
THAP11 IP
a
WT
THAP11 IP
b
THAP11F80L/F80L
input
THAP11F80L/F80L
THAP11 IP
a
THAP11F80L/F80L
THAP11 IP
b
INPP5B

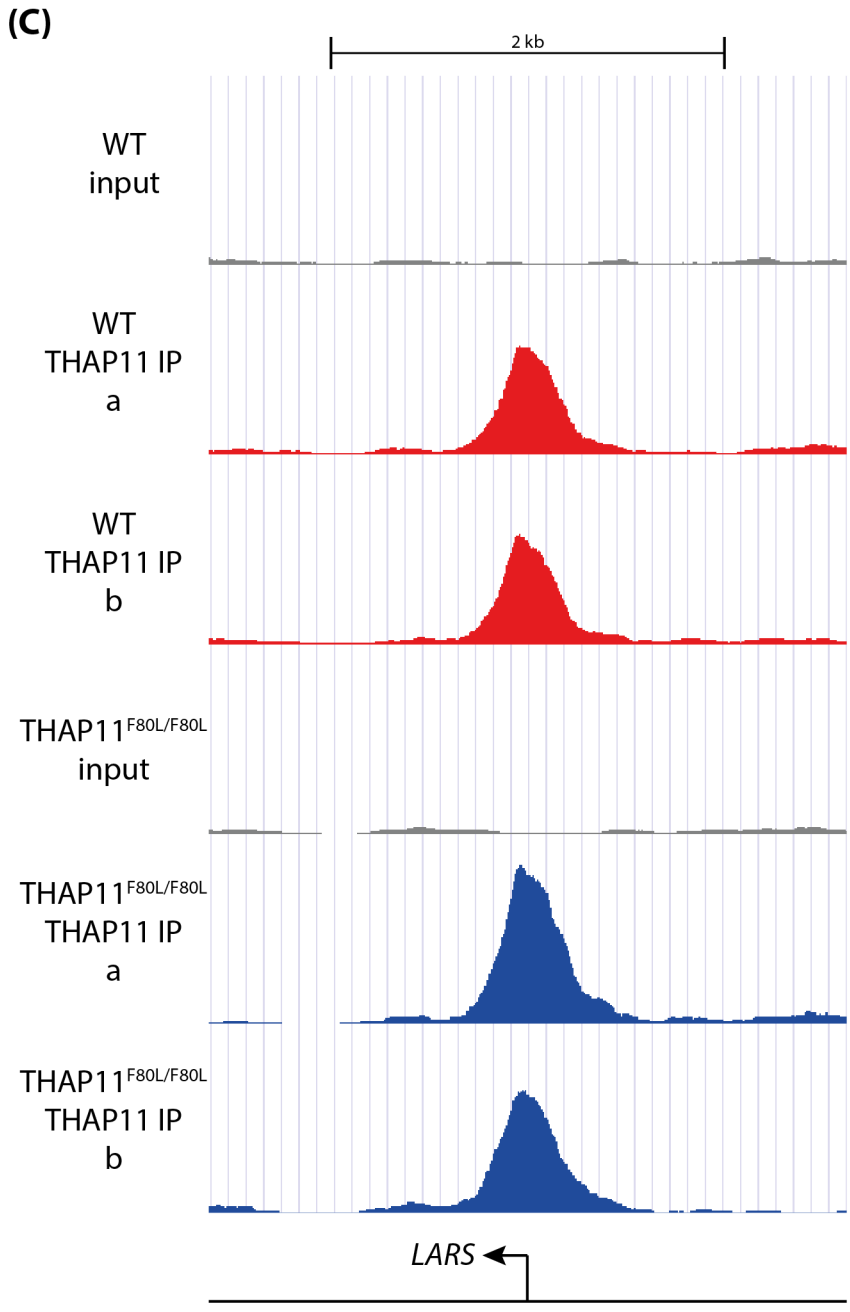


Figure 8.3: **Visualization of the sequencing duplicates at three ChIP-seq peaks.** Peaks were visualized using the UCSC genome browser, all tracks being set with the same vertical viewing range (1 to 600).

(A)

	All peaks	Peaks close to a TSS
WT cells	2341	1114 (48 %)
THAP11$^{F80L/F80L}$ cells	1473	737 (50 %)

(B)

	All peaks	Peaks close to a TSS
common	1261	734 (58 %)
WT specific	1080	380 (35 %)
mutant specific	212	3 (1 %)

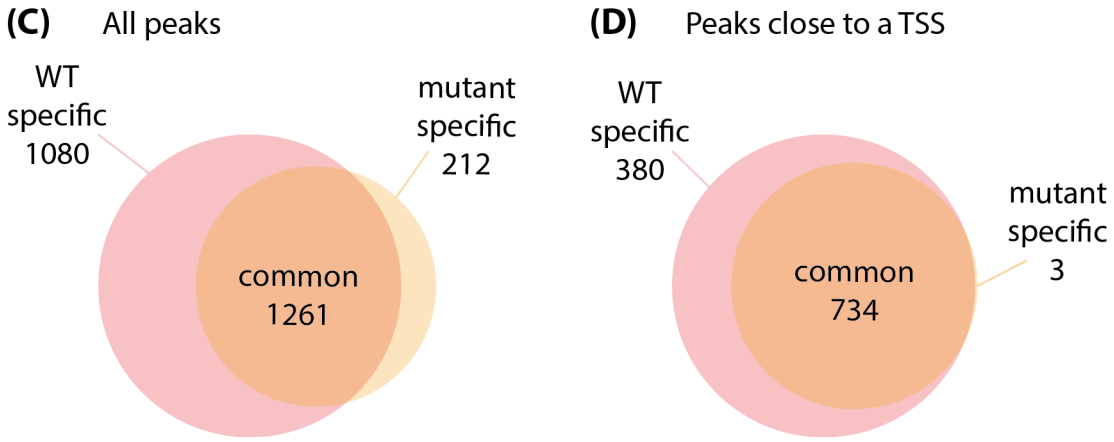
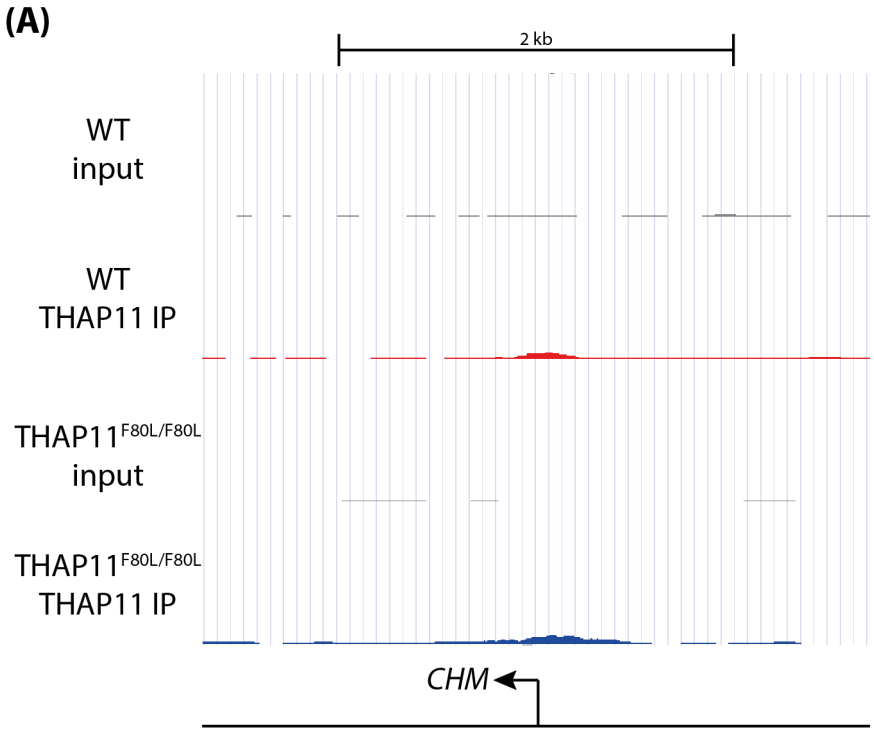
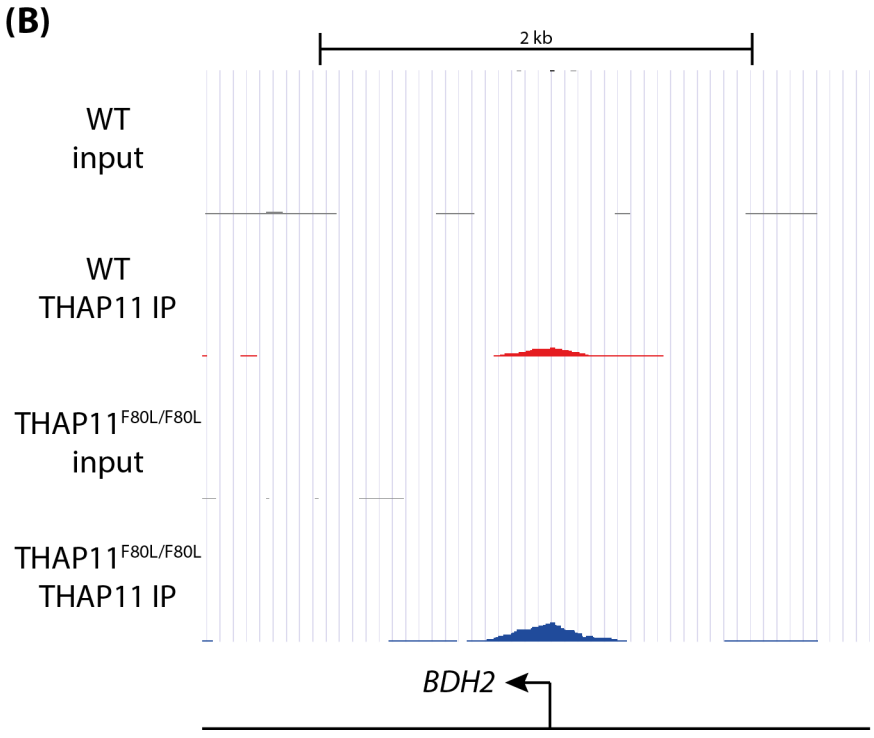


Figure 8.4: **Number of identified peaks after THAP11 ChIP-seq in WT and THAP11$^{F80L/F80L}$ cells.** Total number of peaks, number and percentage of peaks close to a TSS (250 kb on each side) in each cell type **(A)** or in each of the three peak categories **(B)**. Venn diagrams visually showing the overlap of all peaks **(C)** or TSS-close peaks **(D)** between the two different cell lines. Red, WT cells; light orange, THAP11$^{F80L/F80L}$ cells.


(A)
2 kb
WT
input
WT
THAP11 IP
THAP11F80L/F80L
input
THAP11F80L/F80L
THAP11 IP
CHM


(B)
2 kb
WT
input
WT
THAP11 IP
THAP11F80L/F80L
input
THAP11F80L/F80L
THAP11 IP
BDH2

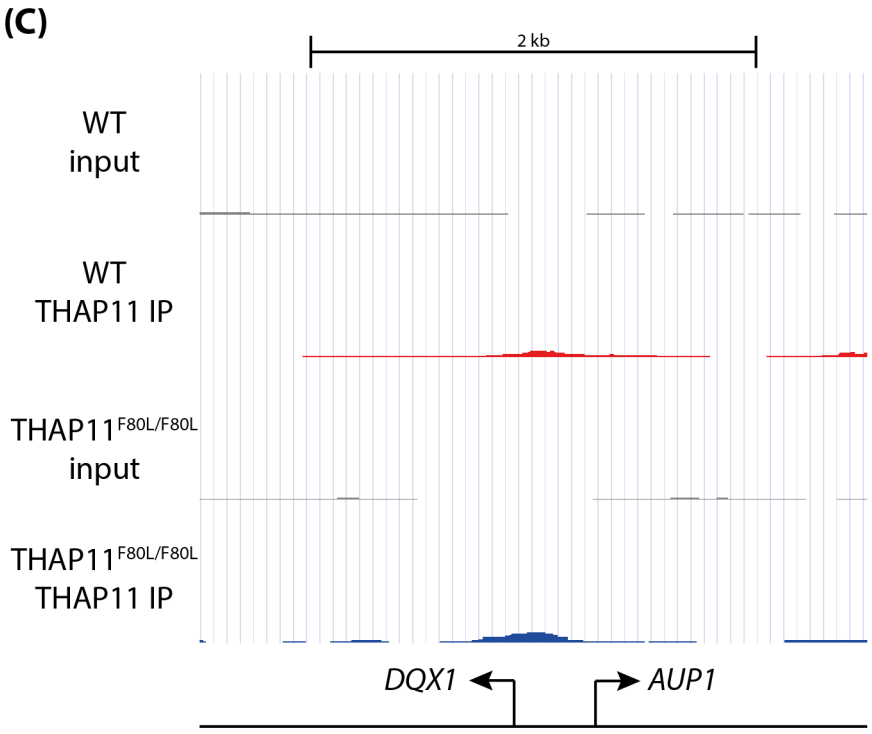


Figure 8.5: **Visualization of the 3 TSS-associated mutant-specific peaks.** Peaks were visualized using the UCSC genome browser, all tracks being set with the same vertical viewing range (1 to 600).

binding by THAP11 in THAP11$^{F80L/F80L}$ cells is thus particularly interesting regarding the pathogenicity of the THAP11 p.F80L mutation [75]. Examining a larger segment of the *MMACHC* genomic region reveals additional THAP11-binding sites, that are differentially affected by the p.F80L mutation. THAP11 binding at the *TOE1 / MUTYH* bidirectional promoter (left peak) is also lost, while retained at the *TMEM69 / GPBP1L1* one (right peak) (Figure 8.6 B).

Thus, we considered the peaks falling into two simplistic categories, depending on whether the peak is retained or not in the THAP11$^{F80L/F80L}$ sample: common versus WT specific peaks. Importantly, the common-peak classification does not discriminate between the peaks that are relatively similar between the two samples, and the ones that, although present in both samples, exhibit a notable change in the peak level between the samples. In a more developed analysis, additional peak categories may be created to take into account the more subtle changes between the WT and mutant samples.

In conclusion, the THAP11 p.F80L mutation results in a selective disruption of THAP11 DNA binding at specific genomic locations. While some DNA sites remain unaffected by the mutation, other sites suffer from a partial to complete loss of THAP11 DNA binding. Also, the THAP11 p.F80L mutation does not create *de-novo* THAP11 promoter-binding sites.

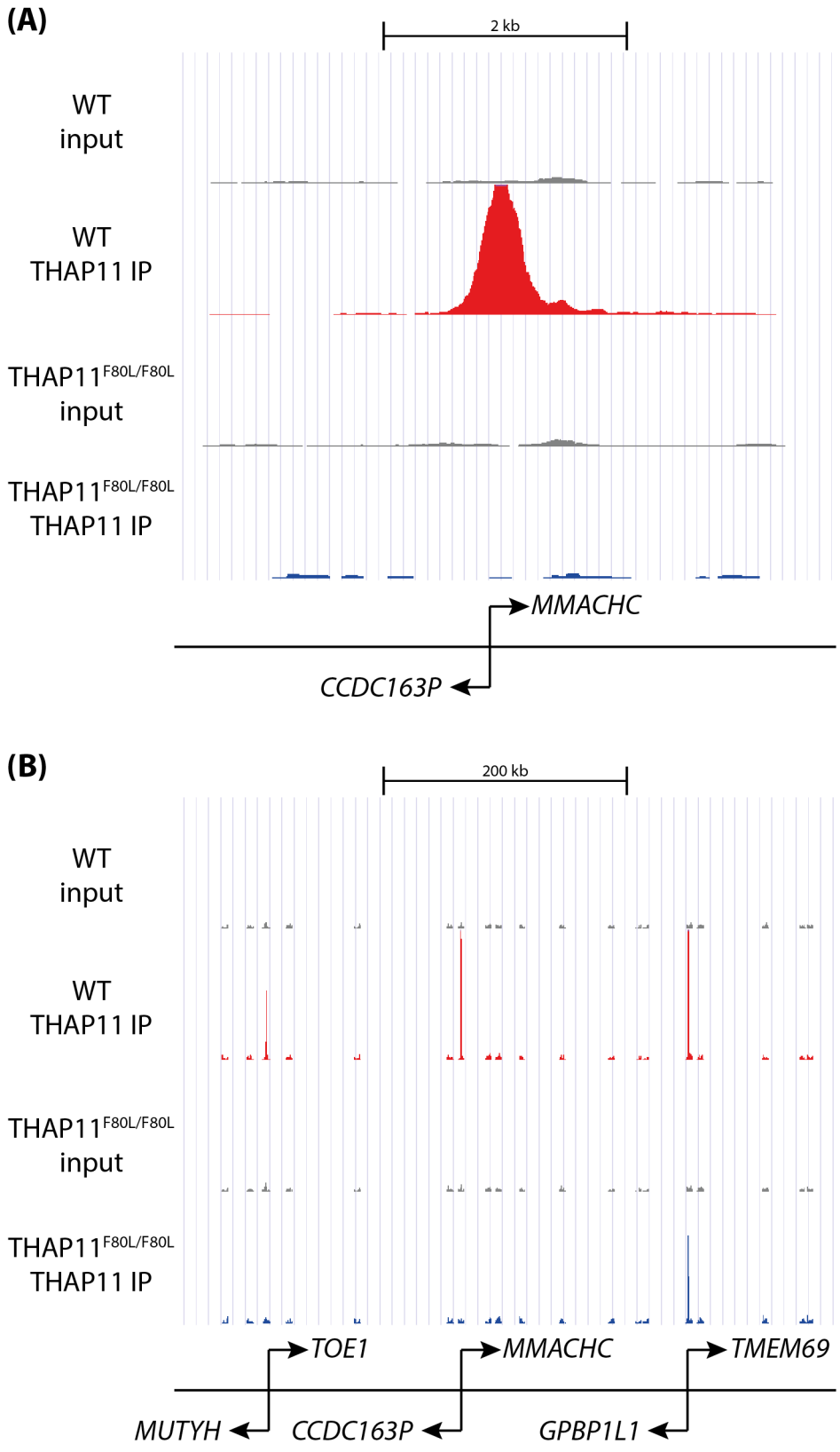


Figure 8.6: **THAP11$_{F80L}$ mutant protein selectively dissociate from a subset of DNA sites.** *MMACHC*-promoter region **(A)** and 100-time zoom out **(B)**. Peaks were visualized using the UCSC genome browser, all tracks being set with the same vertical viewing range (1 to 600).

8.2.2 Why is the THAP11$_{F80L}$ mutant protein not bound to some specific DNA sites?

It is intriguing that some genomic locations are dramatically affected, whereas others are not. To understand differences that may exist between the DNA sites that remain bound by the THAP11$_{F80L}$ mutant protein (common peaks) and the ones lost upon THAP11 p.F80L mutation (WT-specific peaks), Dr. Viviane Praz performed different motifs analyses of the DNA sequence below the peaks.

To begin, we looked for recurrent motifs in the DNA sequences underlying the THAP11 peaks, separately for each peak category. We used CentriMo [116], a tool for sequence comparison with known motifs, to search for motifs within 1000 bp on each side of the peak middle. Interestingly, the results were extremely similar for both sets of peaks. For both peak categories, CentriMo revealed several enriched motifs that are actually simply different versions of the same motif. Figure 8.7 shows the top three motif hits found for each peak category with their respective E-value (A), and the probability of each motif relative to the peak center (B). The two first ones, named as ETS1 and SMARCC2 and 21-bp long, are just the reverse complement of each other (Figure 8.7 C, compare the two top motifs). The third one, called ZNF143 and 15-bp long, is a 5' truncation of the two former motifs (Figure 8.7 C, compare the first and last motifs). Further down in the motif list output, alternate versions of the ZNF143 motif come up, being 15 to 22-bp long. All of the motifs listed in the CentriMo output actually end up being the same ones, or truncated versions, or reverse complements, of the aforementioned ETS1 and SMARCC2 motifs. In conclusion, the CentriMo analysis revealed that a single motif is associated to the THAP11 peaks, irrespective of the THAP11 p.F80L mutation. The motif identified here, called as ETS1, SMARCC2 or ZNF143 in the CentriMo motif database, is also the same as identified as the THAP11 motif or Ronin-binding motif [43, 70]. For the remainder of this thesis, it will be referred to as the THAP11-associated motif (TAM). These results led to think that the difference between the common and WT-specific peak sets is probably more subtle than a complete difference in the underlying DNA sequence.

To assess whether this difference is due to subtle variations among TAMs, Dr. Viviane Praz created a pipeline to generate, separately for each peak category, a consensus sequence (for more details, see the explained procedure in the Chapter 2). Figure 8.8 B and C display the consensus sequences obtained for each of the common and WT-specific peak category, respectively. The consensus sequence is extremely similar, but weaker for the WT-specific peaks. A part of the weakness of the consensus sequence of WT-specific peaks might be explained by the smaller number of motif-containing sequences used to build the consensus of the WT-specific peak category, compared to the number for the common category (561 versus 1159 sequence, respectively). This, however, is not likely to account for the whole difference of consensus strength. Notably,

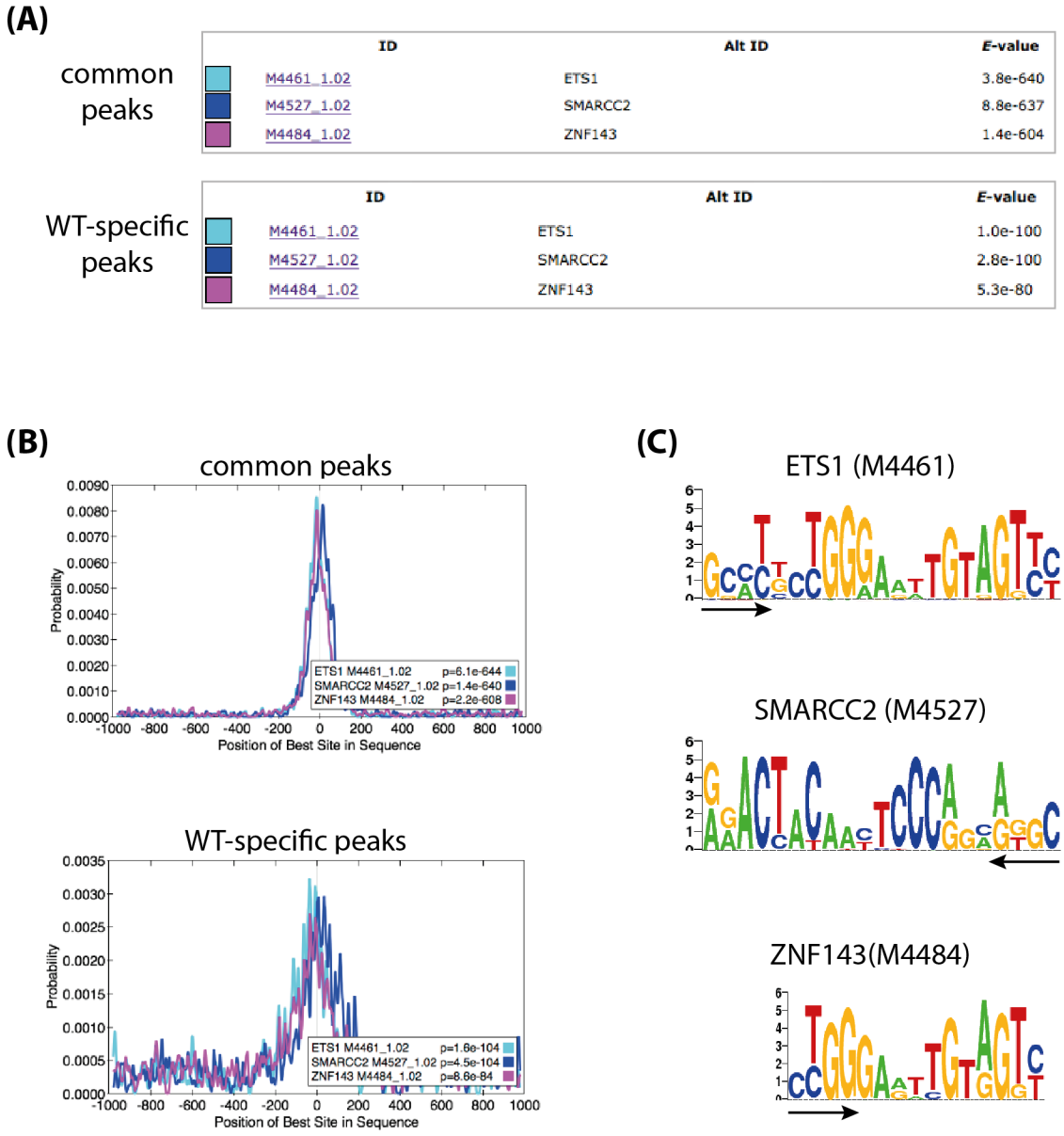


Figure 8.7: **Motif analysis of DNA sequences underlying each category of peaks using Centrimo.** **(A)** Top three hits found for each peak category with their respective E-values. **(B)** Motif-probability graph showing the probability of finding the given motif at a specific distance of the peak center. **(C)** Motif logo of each of the 3 hits provided by Centrimo, with their respective ID.

if the guanine nucleotides are taken as reference, the thymidine in position 17 is severely diminished in the WT-specific peak consensus sequence compared to the common-peak one. More generally, the 3 thymidines in position 15, 17 and 20, as well as the 2 adenines in positions 12 and 18 have a diminished strength in the WT-specific peak consensus sequence compared to the common-peak one, if taking the neighboring guanine as references. Other less important nucleotides also show a diminution in the WT-specific peak motif. Overall, the two consensus motifs are relatively similar regarding their key features (such as the guanine triplet in position 9 to 11 and the 2 guanines in position 16 and 19), but secondary nucleotide positions appear to be weaker in the motif generated with the sequences underlying WT-specific peaks compared to common ones. Actually, the weakness of the TAM under WT-specific peaks was already suggested by the lower E-values obtained in the CentriMo output for the WT-specific set compared to the common one (Figure 8.7 A). Thus, the fact that the mutant THAP11$_{F80L}$ protein still binds, or not, to DNA locations likely depends on the strength of the underlying motif: if the DNA motif is strong enough, the mutant THAP11$_{F80L}$ protein will still be able to bind this genomic location, while it will not bind if the motif is weaker. This would explain the selective disappearance of binding sites bearing a weaker motif.

In the previous analysis, we only considered a single motif per peak — the one closest to the peak summit — whereas many THAP11-bound DNA sites actually exhibit several TAMs. We also counted all the TAMs found within 1000 bp on each side of the peak maximum. Figure 8.9 shows for each of the two groups of peaks the number and percentage of peaks with (one or more) and without a motif (A) as well as the distribution of motif number per peak (B). The results demonstrate that most common peaks (92%) have at least one TAM (dark-green bar). In constrast, almost half of WT-specific peaks (48%) do not have any evident TAM, as shown on Figure 8.9 B (light-red bars). Regarding the peaks having at least one TAM, there is not much difference between the common and WT-specific peak categories regarding the distribution of the motif number per peak, as shown by the peak distribution (Figure 8.9 B, compare the dark-red and red bars for motif numbers from 1). Interestingly, for both categories, a large proportion of peaks exhibits between 2 to 4 motifs per peak.

These different analyses strongly suggest that with the p.F80L mutation THAP11 retains some DNA-binding ability. The resulting mutant protein does not have any *de-novo* binding activity but rather remains bound to its strongest DNA locations. These are the ones that have at least one TAM, the latter being the strongest possible. Thus, the criterion that determines whether the mutant THAP11$_{F80L}$ protein is still able or not to bind a specific DNA location is its apparent affinity for the associated DNA sequence.

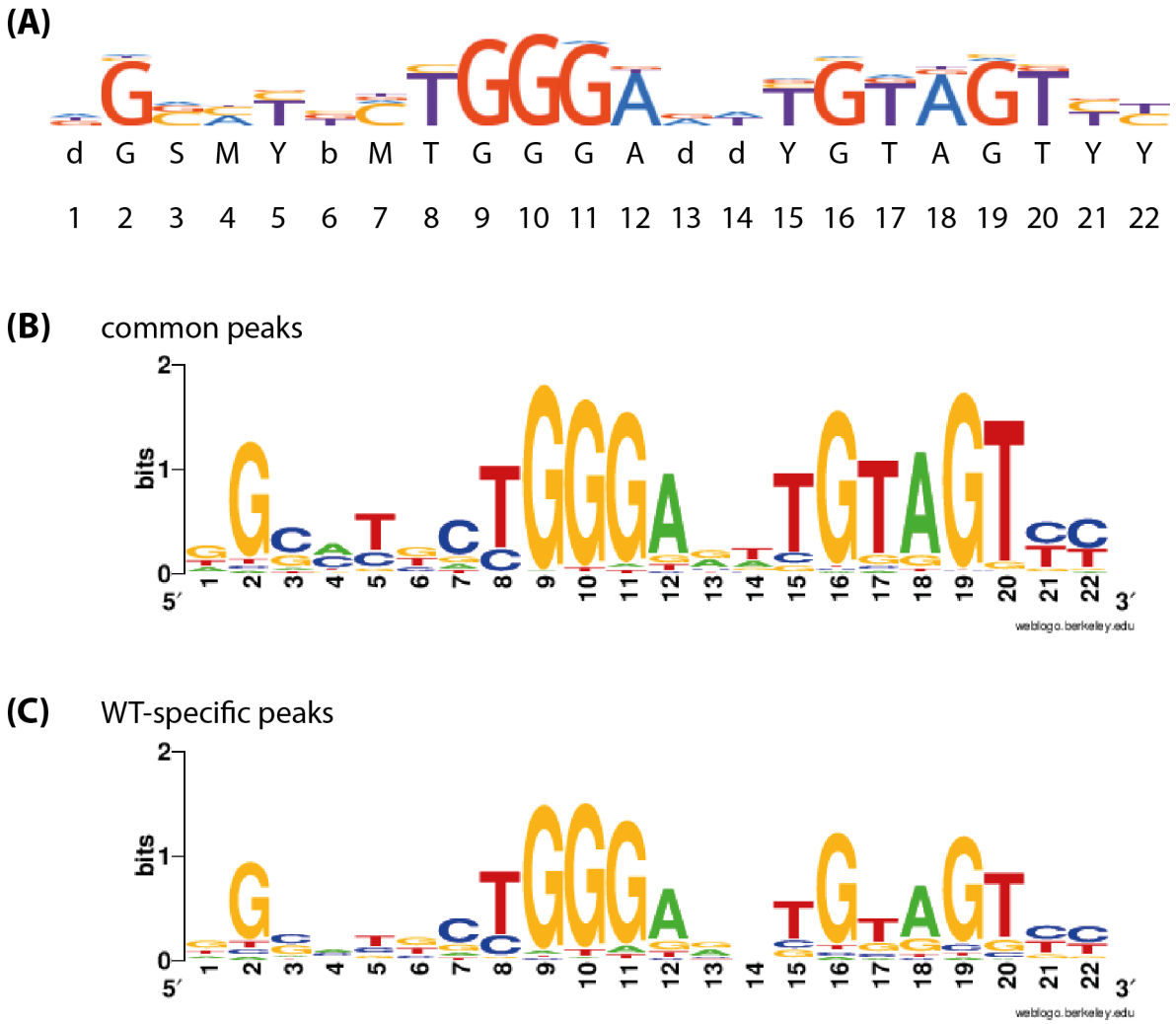


Figure 8.8: **TAM consensus sequence in common and WT-specific peak categories. (A)** 22-bp long ZNF143 motif logo from Hocomoco (ZN143_HUMAN.H11MO.0.A) used as a reference, the consensus sequence being depicted below. **(B)** and **(C)** Logo of consensus sequences obtained for common and WT-specific peaks, respectively.

8.2.3 The THAP11 F80L mutation affects a specific subset of THAP11-bound promoters

THAP11 regulates transcription, and has been previously shown to be tightly associated to TSSs [43, 70]. I already mentioned that one half of THAP11 binding sites lies close to an annotated TSS, both in WT and THAP11$^{F80L/F80L}$ mutant cells (Figures 8.4 A). This finding is even better illustrated by the genomic distribution charts in Figure 8.10. Indeed, half of both WT and THAP11$_{F80L}$ protein DNA association happens in a 500 bp-window (+/- 250bp) around an annotated TSS (TSS_PROX). The association of THAP11 to TSSs is even more prominent than what the pie charts visually suggests at first sight, as the TSS_PROX category is much more restrictive than the other ones that encompass much larger regions of the genome.

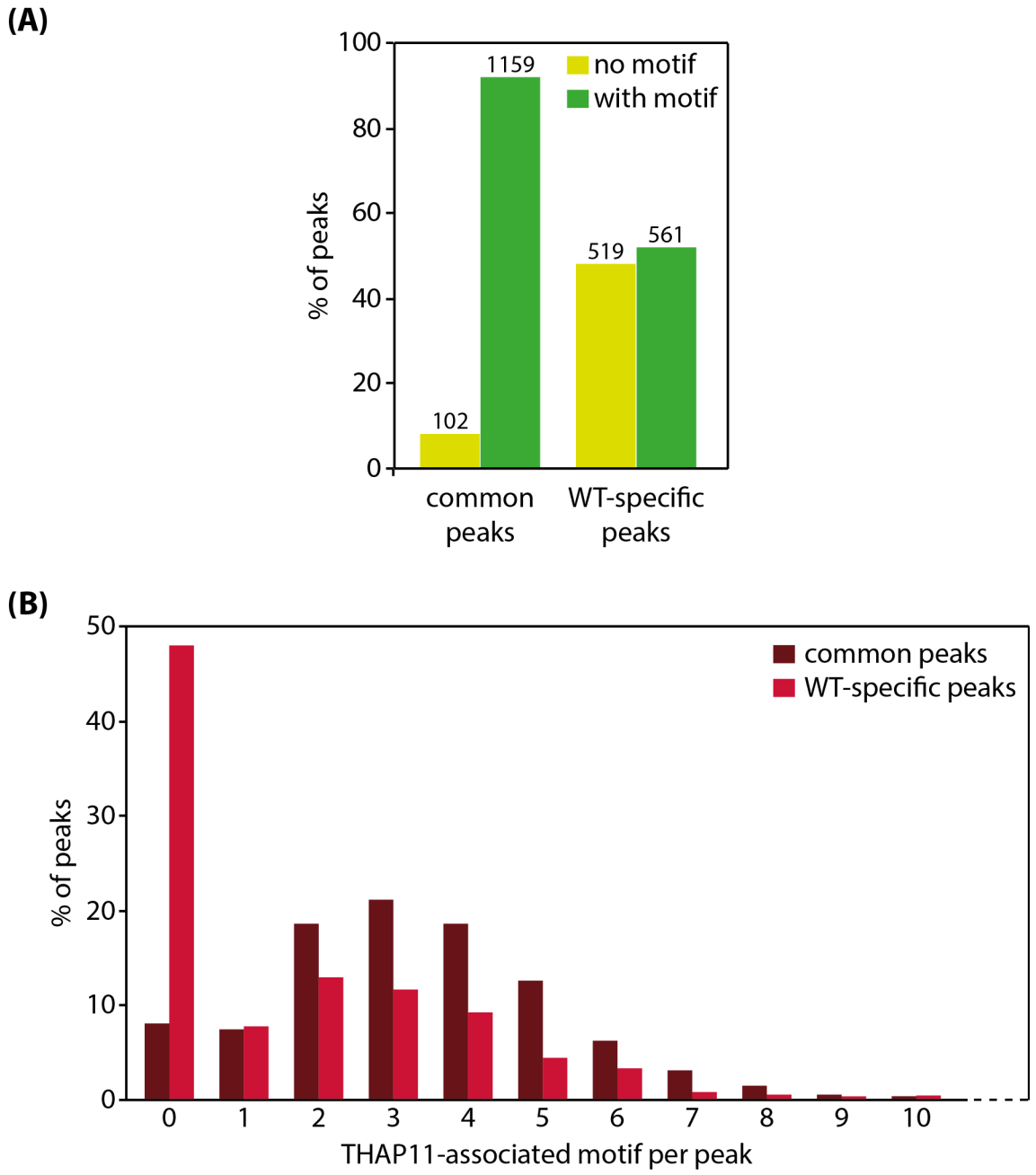


Figure 8.9: **Percentage of peaks with a THAP11-associated motif.** The number and percentage of peaks with a TAM was calculated separately for each peak category. **(A)** Percentage of peaks with (regardless of the motif number), and without, a motif, the raw numbers being written above the bars. **(B)** Distribution of the motif number per peak. Only peaks with a number of motifs up to 10 were shown, but some rare peaks have a higher TAM number, up to 18.

For instance, a remarkably higher proportion of THAP11 peaks is located in the TSS_PROX category, which is 500-bp wide, compared to the 1500-bp wide TSS_DIST category.

As both the WT and the p.F80L mutant THAP11 proteins associate preferentially to TSSs, Dr. Viviane Praz prepared a composite plot showing the THAP11 occupancy in a window of +/- 1 kb around TSSs. For this, peaks in the TSS_PROX category were considered and pilled up using the TSS as a reference. Figure 8.11 shows that both WT and mutant THAP11 proteins preferentially associate 50 bp upstream to TSSs, on average. This observation is in agreement with the average 77-bp distance to TSSs that have been already proposed [70]. The THAP11$^{F80L/F80L}$-cell curve (dark-blue line) is less high than the WT curve (dark-red line) due to the fewer number of peaks in the mutant sample and to the fact that the WT-specific peaks tend to be lower (see Figure 8.13 later).

In conclusion, I demonstrate that THAP11 prefentially associates just upstream to TSSs, this feature being unaffected by the THAP11 p.F80L mutation. As I am particularly interested in the role of THAP11 in gene transcription, the rest of the study focuses on the TSS-related peaks — meaning, THAP11 peaks for which at least one bp is located within a -250 to +250 bp window of an annotated TSS (TSS_PROX category).

So far, I have shown that the THAP11 p.F80L mutation induces a selective THAP11 DNA dissociation at genomic sites for which THAP11 has a lower apparent affinity. This selective modification of THAP11 DNA-binding pattern thus likely affects gene transcription as both WT and mutant THAP11 proteins are associated with TSSs. Indeed, it has been previously shown that fibroblasts from the THAP11$^{F80L/F80L}$ patient exhibit a different transcriptome compared to fibroblasts from healthy patients. We thus analyzed the genes for which THAP11 promoter binding has been altered by the p.F80L mutation. For each THAP11 TSS-associated peak as defined above, scores were calculated by normalizing the tag counts in the ChIP sample by the tag counts in the input sample and the peak width. When the peak is present in both replicates — which is the case for most of the peaks — the mean score between the two replicates is considered. The dot plot in Figure 8.12 shows the distribution of TSS-associated common (dark-red dots) and WT-specific (light-red dots) peaks according to their WT score (X axis) and the fold change between the WT and THAP11$^{F80L/F80L}$ samples (Y axis, a positive fold change meaning that the score is higher in the WT sample compared to the mutant one).

This plot revealed an impressive segregation between common (dark-red dots) and WT-specific (light-red dots) peaks. First, the red cloud of WT-specific peaks is above the dark-red one of the common peaks. This confirms the previously-made segregation of peaks between the common and WT-specific categories: the WT-specific peaks are the ones present in the WT sample but disappearing in the mutant one, thus logically have an higher fold change between the two samples. Not surprinsingly, the two distributions are not exactly

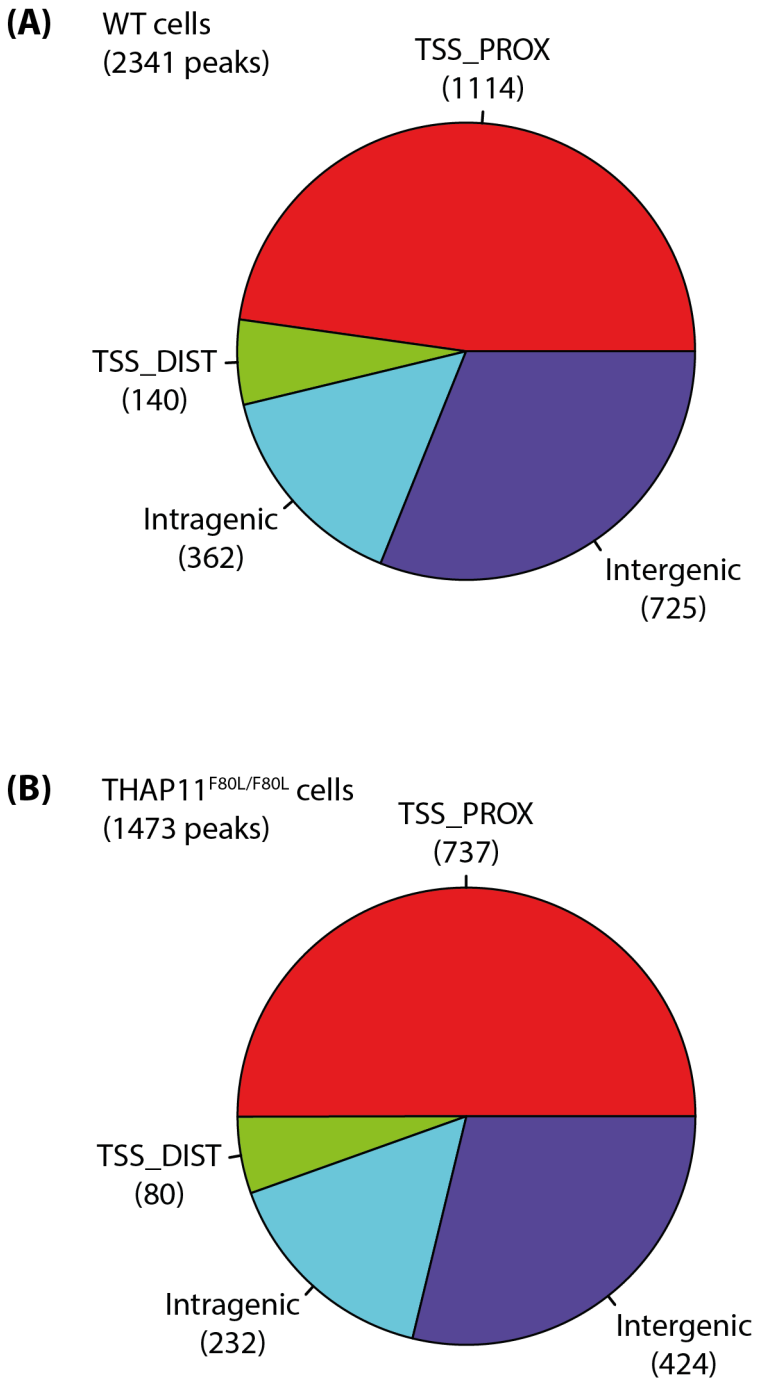


Figure 8.10: **Genomic-distribution chart of THAP11 DNA association in WT and THAP11$^{F80L/F80L}$ cells.** The proportion and number of peaks in each of the genomic category are displayed for WT **(A)** and THAP11$^{F80L/F80L}$ cells **(B)**. TSS_PROX, peaks for which at least one bp is located within a -250 to +250 bp window of an annotated TSS; TSS_DIST, peaks which at least one bp is located within a -1 to +1 kp window of an annotated TSS, but not included in the TSS_PROX category; intergenic, peaks found in repeats and between genes; intragenic, peaks found in exons and introns.

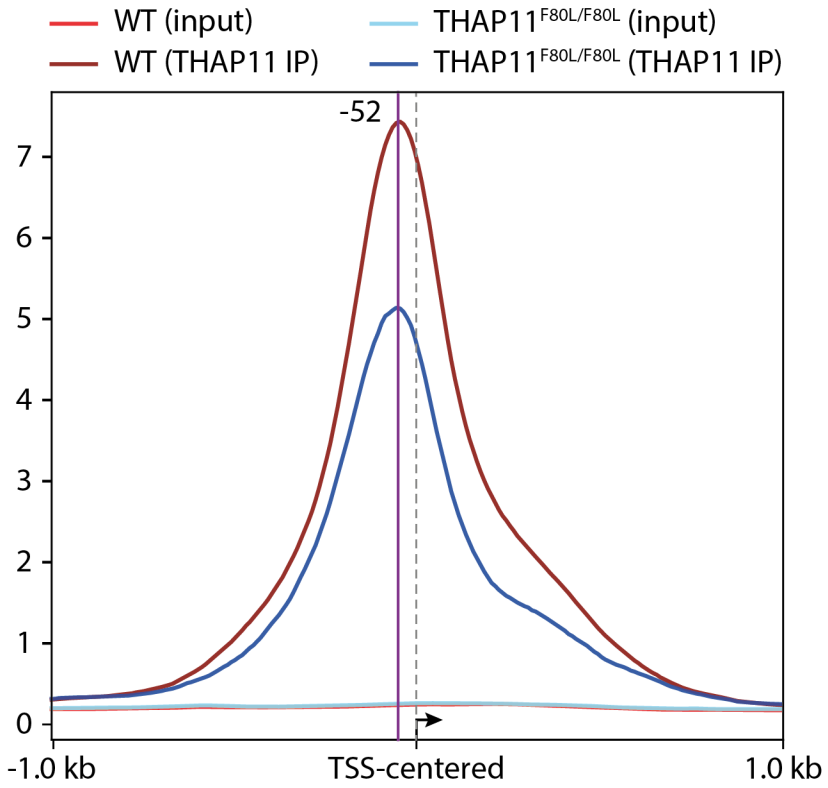


Figure 8.11: **THAP11 occupancy around transcription start sites in WT and THAP11$^{F80L/F80L}$ cells.** Composite plot showing the THAP11 occupancy (Y axis) in a +/- 1kb window around TSSs for each WT **(A)** and THAP11$^{F80L/F80L}$ cells **(B)**. Only peaks categorized as TSS_PROX were considered (+/- 250 bp window). Dashed-grey line, TSS position; solid-purple line, average position of peaks.

on the top of each other horizontally fashion, but rather diagonally, as the fold change increases with the WT score. Of note, the dot labelled as (a) represents a peak categorized in the common category, but bearing a $\log_2$ fold-change of 2.5. It may thus seem mislabelled. It has, however, a very low WT score, suggesting that it may be simply background. Indeed, the peak comes from the repeated *RNA5S* genes genomic region, in which input and IP samples diplay the same signal strength. In addition, WT and THAP11$^{F80L/F80L}$ samples are similar at this genomic region. Consequently, this peak is not mislabelled, it is simply not a real peak and has been identified as such due to the nature of its associated genomic sequence.

Second, the dot plot shows that the peaks in the dark-red cloud (common peaks) are shifted to the right of the plot, thus have overall higher WT scores than the light-red cloud ones (WT-specific peaks). This is even better shown by the distribution of WT scores for each peak category, displayed in Figure 8.13. This means that the peaks that remain bound by the THAP11$_{F80L}$ mutant protein are the ones with a higher score, thus the strongest peaks. It nicely fits the model that I am so far elaborating in which the THAP11$_{F80L}$ mutant protein remains only bound to the genomic sites for which it has the higher affinity.

Interestingly, the top hit regarding the fold change (labelled as (b) on 8.12) is the peak associated to

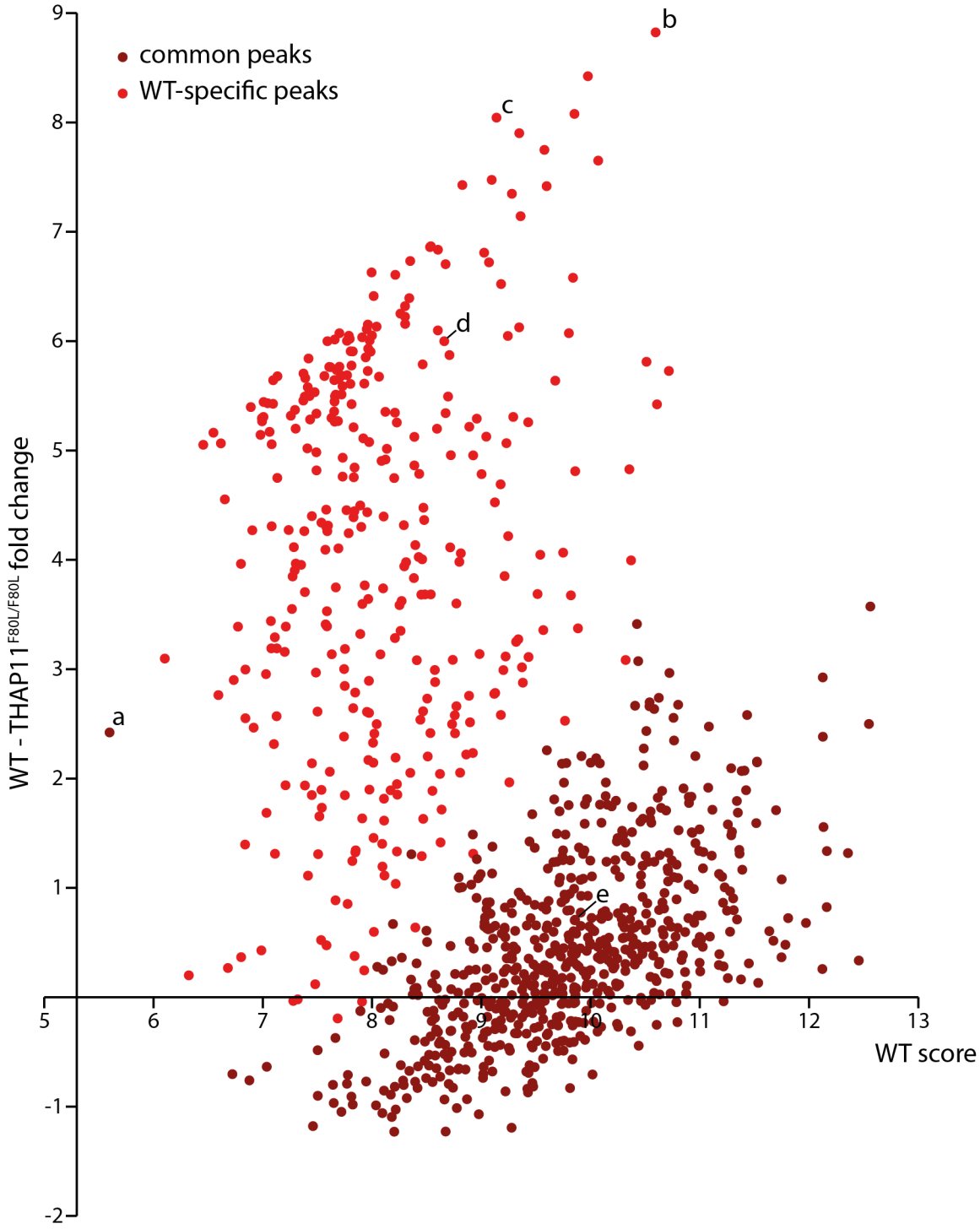


Figure 8.12: **Distribution of peaks according to their score in the WT sample and the fold change between the WT and THAP11$^{F80L/F80L}$ sample scores.** Each dot represents a peak colored according to its category: dark-red dots, common peaks; light red dots: WT-specific peaks. The X axis represents the score of the peak in the WT sample (the mean between the two replicates, if applicable). The Y axis represents the fold change between the WT and THAP11$^{F80L/F80L}$ samples, meaning the difference between the respective ($\log_2$) scores. Specific peaks of interest discussed in the text are labelled with letters.

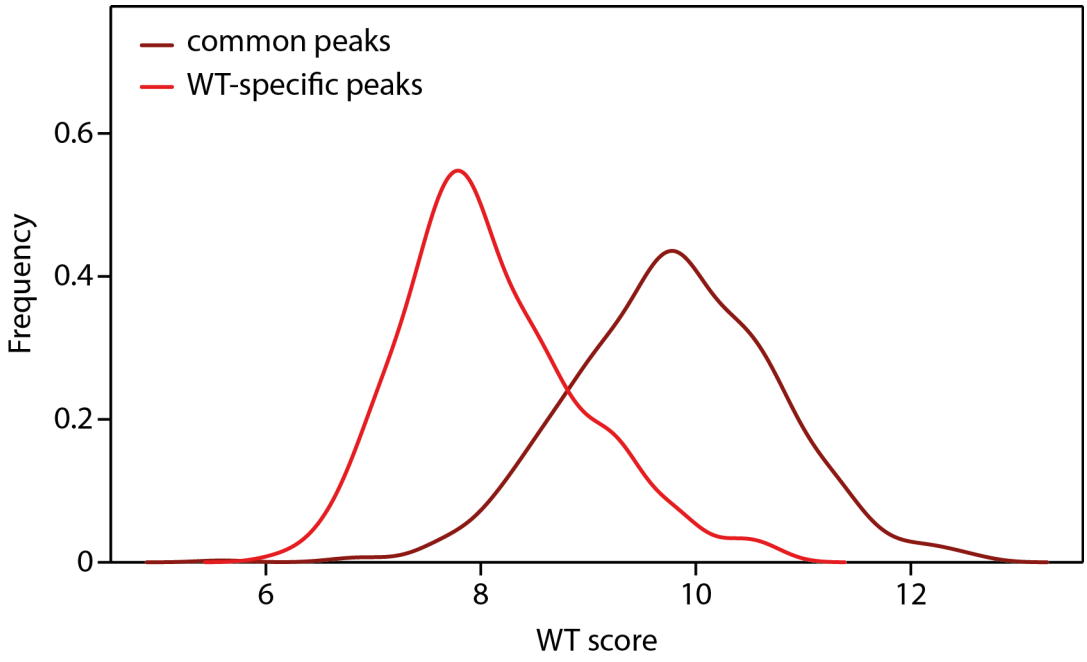


Figure 8.13: **Distribution of WT scores for each peak category.**

the *DPF1* gene. Its associated protein is a neurospecific transcription factor essential for proper neuronal development. It belongs to the nBAF neuron-specific chromatin remodeling complex which regulates the transition from proliferating neural progenitors to differentiated neurons [128]. It is particularly intringuing as the cobalamin disorder is associated with neurodevelopmental abnormalities. Furthermore, the *MMACHC* gene promoter is among the top peak hits with the higher fold change between the WT and mutant samples (Figure 8.12, peak labelled as (c)). Indeed, THAP11 DNA binding is completely lost upon the THAP11 p.F80L mutation, as seen in Figure 8.6 A. Thus, *MMACHC* gene expression is likely to be affected by the loss of THAP11 binding. More generally, the selective modification of THAP11 promoter-binding pattern is thus likely to affect the expression of various genes.

8.2.4 Overview of differential gene expression due to the THAP11$_{F80L}$ mutation

We then submitted the WT and mutant cells to a large-scale transcriptomic analysis by RNA extraction and sequencing to analyze the change in gene expression due to the THAP11 p.F80L mutation. Thanks to the ease of the RNAseq procedure, we performed, as for the proliferation assays, the experiment at both 37 °C and 39.5 °C, for the WT and both the homozygous THAP11$^{F80L/F80L}$ and the heterozygous THAP11$^{F80L/+}$ cell lines. Indeed, I previously showed that the heterozygous cells exhibit an intermediate behavior between the WT and the homozygous mutant cells, and thus it is interesting to examine these cells in parallel. Also, even if there is no temperature-specific effect in the proliferation impairment, I still demonstrated that increasing

the temperature reinforces the proliferation phenotype, and thus may facilitate the interpretation of the results. Cells were grown for 3 days in duplicate for each condition, either at 37 °C, or at 39.5 °C for the two last days. RNA was extracted and subsequently sequenced to probe the whole transcriptomic changes between the WT and the mutant cells.

As done previously, Dr. Viviane Praz summarized the normalized data using a PCA, as shown in Figure 8.14. First, there is a general temperature effect. Indeed, each cell line is displaced in the same direction (and magnitude) by the increase of temperature. Again, no temperature-specific effect was observed for the mutation. This PCA plot also discriminates the samples by their genotype. Curiously, the heterozygous mutation implies a deplacement similar to the temperature shift. In contrast, the homozygous mutation displaces the samples on the PCA in a perpendicular direction of the PCA plot. This demonstrates that the heterozygous THAP11$^{F80L/+}$ cells are more similar to the WT cells, compared to the homozygous mutant ones.

Then, Dr. Viviane Praz performed several differential analyses on the RNA-seq data, to understand which genes are differentially expressed between the different samples. Here, only data from cells grown at 37 °C were used as I have just shown that there is no temperature-specific effect. For each comparison, she subsequently submitted the list of upregulated and downregulated genes for a Gene Ontology (GO) analysis. In the homozygous THAP11$^{F80L/F80L}$ cells, 346 genes were upregulated and 669 downregulated, compared to the WT cells. When sujected to GO analysis, upregulated genes in the homozygous mutant cells reveal pathways largely related to transcription, DNA binding and chromatin (Supplemental Table S9, orange dots), and a pathway related to neuronal development (dendritic spine development, purple dot). In addition, genes downregulated in the homozygous mutant cells are associated to mitochondria (Supplemental Table S10, green dots), and also reveal transcription and development-associated GO terms (orange and purple dots, respectively). Conversely, many fewer genes were differentially expressed when THAP11 is mutated in an heterozygous fashion: only 171 genes were upregulated and 142 downregulated, in the heterozygous mutant cells compared to the WT ones. When these lists of differentially expressed genes were submitted for GO analysis, no particular pathway stood out for genes upregulated in the heterozygous mutant cells (Supplemental Tables S11), while some development and proliferation-associated GO terms are revealed by downregulated genes (S12, purple and blue dots, respectively). Interestingly, the comparison of the homozygous THAP11$^{F80L/F80L}$ cells with the heterozygous ones leads to results similar to the comparison with the WT cells. Indeed, 479 genes were upregulated and 925 downregulated, in the homozygous mutant cells compared to the heterozygous THAP11$^{F80L/+}$ cells. GO analysis revealed, as for the comparison with the WT cells, pathways related to transcription (upregulated genes, orange dots) and to development (downregulated genes, purple dots) as well as association with mitochondria (downregulated genes, green

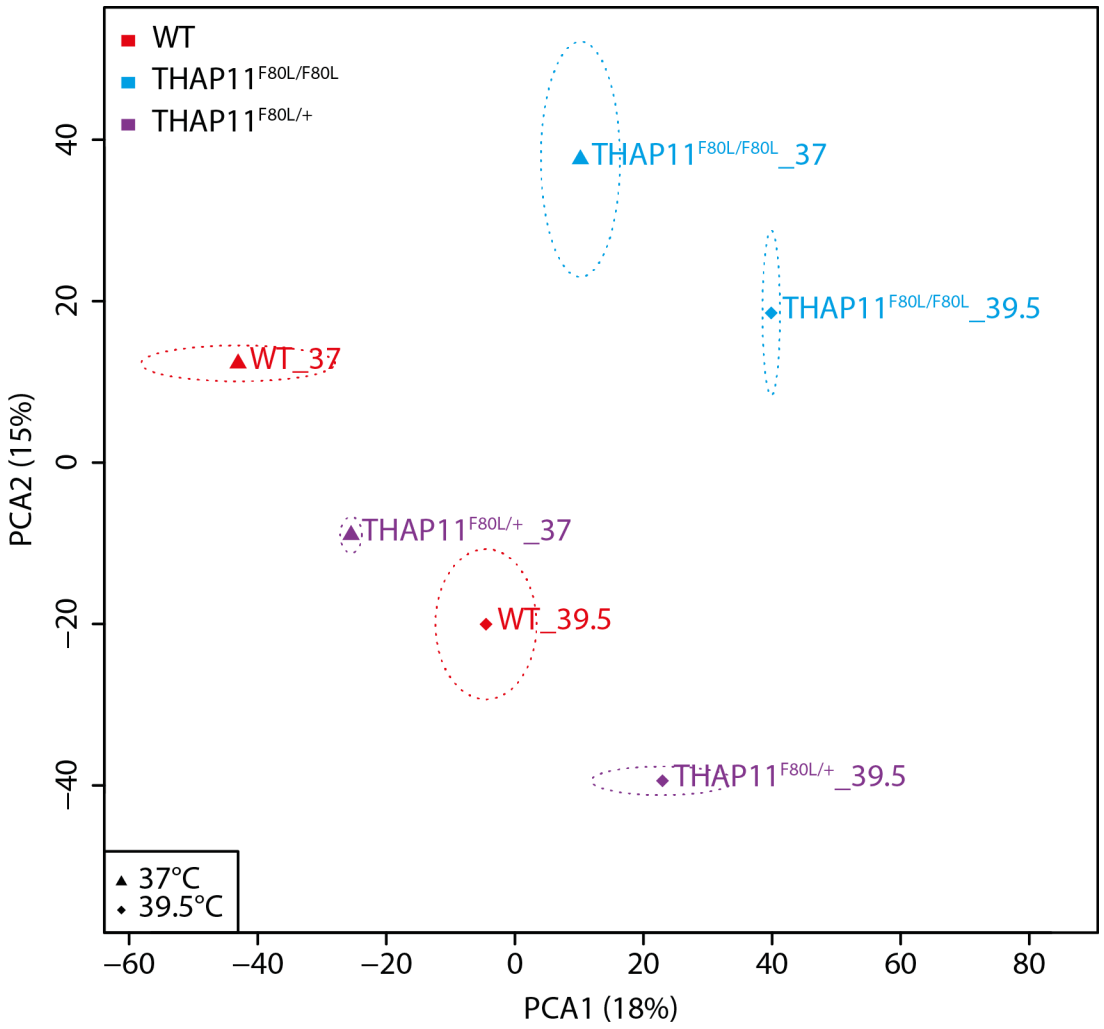


Figure 8.14: **Principal component analysis of RNA-seq data from WT, THAP11$^{F80L/F80L}$ and THAP11$^{F80L/+}$ cell lines.** PCA1 and 2 resulting from RNA-seq data of WT (red), THAP11$^{F80L/F80L}$ (homozygous, blue) and THAP11$^{F80L/+}$ (heterozygous, purple) cells, cultivated for 3 days at 37 °C or for 1 day at 37 °C and the 2 following days at 39.5 °C. The triangles and squares represent the mean of the 2 technical replicates with the dashed-line circle representing the standard error. Triangles represent cells grown at 37 °C while diamonds represent cells grown at 39.5 °C.

dots) (Supplemental Tables S13 and S14, respectively). The two latter comparisons between the homozygous and heterozygous clones confirm again the fact that the heterozygous THAP11$^{F80L/+}$ cells are more similar to WT cells than to homozygous mutant cells, as was already seen in the proliferation assays (at least at 37 °C) and with the PCA plot.

I looked closer at the top differentially expressed genes (in terms of p value) between the WT and the homozygous THAP11$^{F80L/F80L}$ cells at 37 °C (Figure 8.15). The first hit is *HIST1H1D*, which encodes for the linker histone H1.3 involved in the condensation of chromatin into higher-order structures. Second and third are the *TMOD2* (tropomodulin 2) and the *MMACHC* genes, respectively. Indeed, *HIST1H1D* is more than 10-time less expressed, *TMOD2* is completely absent (almost 6-time lower), and *MMACHC* is more than 4-times less expressed in the homozygous mutant cells compared to the WT cells (Figure 8.15 A, compare the solid blue and red bars; please note that the data are expressed in $\log_2$ of normalized expression values, so a change of 1 unit means a 2-fold change in the expression of the gene). As mentioned above, THAP11 DNA binding is lost upon p.F80L mutation at the *MMACHC* promoter (Figure 8.6). Similarly, the THAP11$_{F80L}$ mutant protein also dissociates from the *TMOD2* promoter in the homozygous mutant cells (Figure 8.15 B and Figure 8.12, peak labelled as (d)). Thus, the selective dissociation of THAP11 from a subset of promoters due to the mutation results in a severe impairment in the expression of the corresponding genes. In contrast, no THAP11 occupancy was observed at the *HIST1H1D* promoter. While a portion of affected genes are likely to be primary targets of THAP11 affected by the mutation, others may conversely represent downstream effectors. Also, the mRNA levels of the three top hits are similar in cells grown at 37 °C and 39.5 °C (Figure 8.15, compare solid and hatched bars). This suggests again that there is no temperature-specific effect on these cells: the increase of temperature does not reinforce the disease mechanism, but rather adds an unrelated constraint to the cells. It thereby explains the behavior of homozygous THAP11$^{F80L/F80L}$ cells at elevated temperature that we have seen both on proliferation assays and on the previous PCA plot. Interestingly, heterozygous cells exhibit similar expression levels of these three genes at both temperatures (solid and hatched-purple bars). It thereby suggests once again that the heterozygous cells are not much affected by the heterozygous presence of the mutation. Furthermore, Quintana and colleagues have also previously demonstrated that both *TMOD2* and *MMACHC* mRNA levels are dramatically reduced in THAP11$^{F80L/F80L}$ patient-derived fibroblasts compared to healthy-patients fibroblasts [75]. As in their study, we did not identify any genes involved in cobalamin metabolism other than *MMACHC*, being differentially expressed between the HEK-293 WT and homozygous THAP11$^{F80L/F80L}$ cells. Thus, the dysregulation of *MMACHC* gene expression resulting from the loss of promoter binding is likely to account for a sizeable part of the disease phenotype, which is tightly linked to the cobalamin pathway.

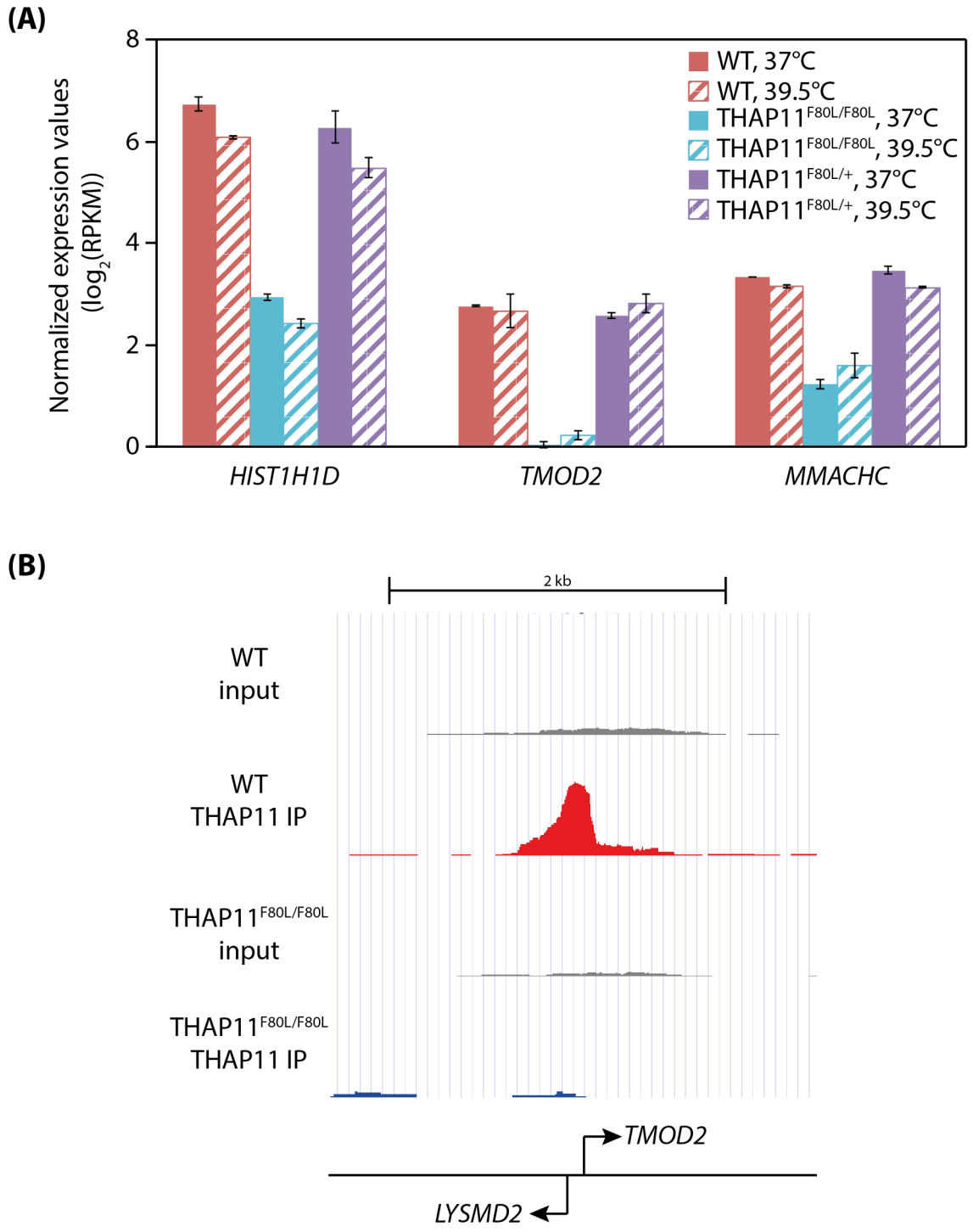


Figure 8.15: **Top-three differentially expressed genes between WT and THAP11$^{F80L/F80L}$ mutant cells.** The genes differentially expressed between the WT cells and THAP11$^{F80L/F80L}$ mutant cells were ranked by their p value and the top three genes were selected. **(A)** Normalized $\log_2$(RPKM) expression values of the three top gene hits, in WT, THAP11$^{F80L/F80L}$ and THAP11$^{F80L/+}$ cells, at both temperatures, shown as the mean +/- standard deviation of the two technical replicates. **(B)** Visualization of THAP11 occupancy at the TMOD2 promoter in WT cells and THAP11$^{F80L/F80L}$ mutant cells using the UCSC genome browser, all tracks being set with the same vertical viewing range (1 to 600).

To summarize, the results presented here support the idea that the THAP11 p.F80L mutation leads to a loss of THAP11 binding to specific promoters, which in turn alters the expression of the corresponding genes and subsequent downstream targets.

8.2.5 Its THAP11 gene expression affected by the p.F80L mutation?

I previously demonstrated that the level of THAP11 protein is diminished in THAP11$^{F80L/F80L}$ cells compared to WT cells (see Chapter 6). Thus, I wondered whether this would be the result of a decrease of *THAP11* gene transcription. Indeed, one could imagine that the WT THAP11 protein may regulate the expression of its own gene, and that the p.F80L mutation could disrupt its binding to its own promoter, leading to an impaired *THAP11* gene expression. The RNA-seq and ChIP-seq experiments described above allow to test this hypothesis.

As shown on Figure 8.16 A, THAP11 protein is indeed bound to its own promoter. In addition, there are two very strong TAMs under this peak, which are likely to be responsible for THAP11 recruitement (black triangles). THAP11 thus has the potential to regulate the expression of its own gene. The mutant THAP11$_{F80L}$ protein, however, still binds the THAP11 promoter, even though the peak is slightly lower in the THAP11$^{F80L/F80L}$ sample compared to the WT one (Figure 8.16 A, and peak labelled (e) in Figure 8.12). In addition, *THAP11* gene expression is not affected by the THAP11 p.F80L mutation, as seen by the similar mRNA levels in the WT and mutant cells (Figure 8.16 B).

To summarize, although THAP11 has the potential to regulate the expression of its own, the THAP11 p.F80L mutation does not affect *THAP11* gene expression. Consequently, the diminished level of mutant THAP11 protein in the THAP11$^{F80L/F80L}$ cells is not due to a decrease in THAP11 mRNA levels.

8.3 Discussion

In this chapter, I discussed high-throughput analyses performed to understand how THAP11 controls gene transcription and cell proliferation. I was very lucky to benefit from help for these analyses, both on the experimental side and on the computational side.

First, we compared gene expression between parental cells and stable cells expressing an ectopic THAP11 protein, either WT or HBM mutant. In agreement with previous studies [29, 39, 54, 68–77], we revealed that THAP11 is involved in gene transcription, development and cell proliferation. For instance, heart development is a developmental pathway that particularly stands out in our GO analysis, and Fujita and colleagues have suggested that Ronin (the mouse Thap11) governs early heart development [71]. In addition, THAP11 is associated with genes involved in mitochondrial function, which is in agreement with a previous

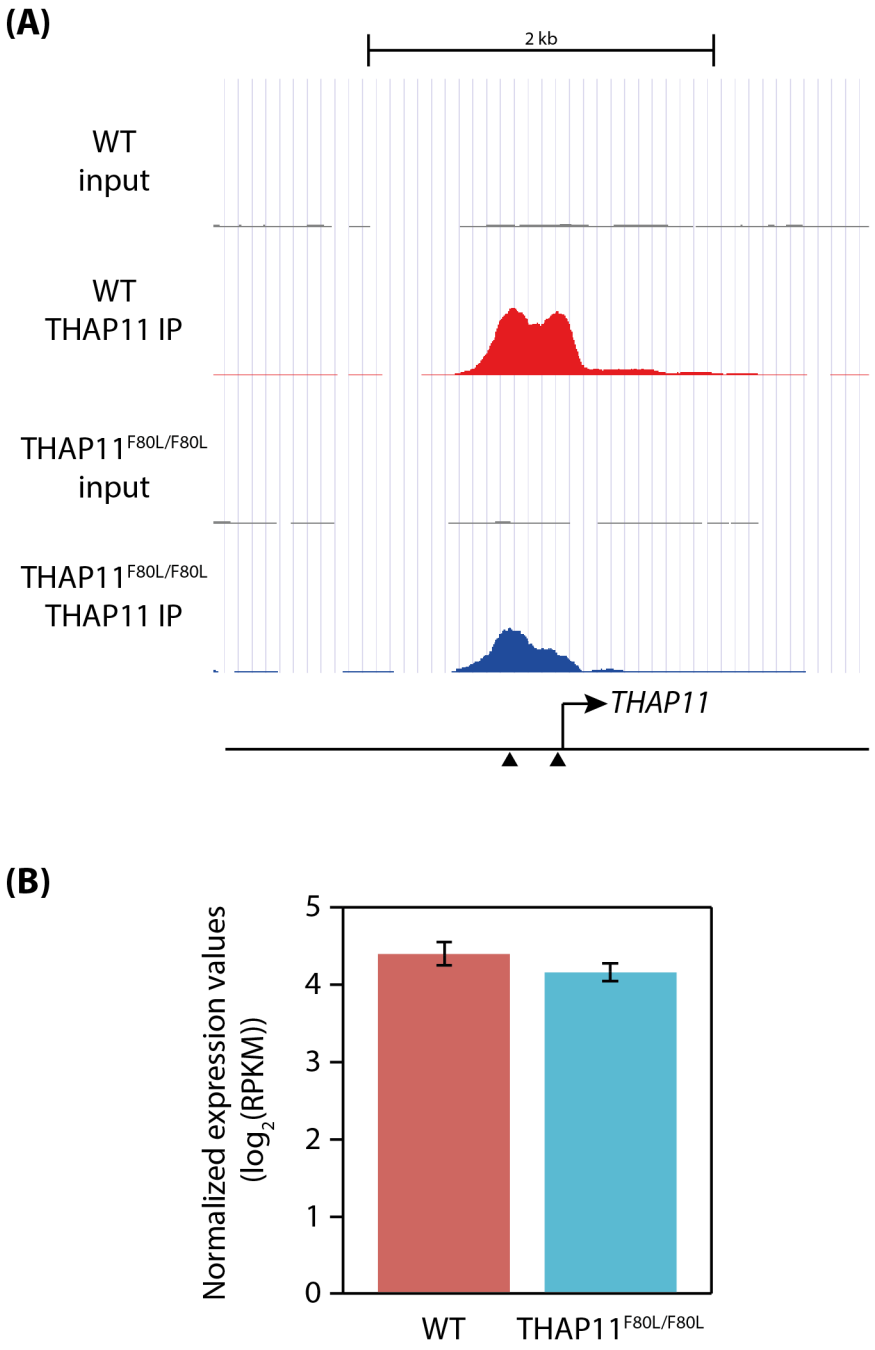


Figure 8.16: **THAP11 binding to its own promoter and effect of the THAP11$_{F80L}$ mutation.** **(A)** Visualization of THAP11 recruitment to its own promoter, in WT and THAP11$^{F80L/F80L}$ mutant cells, using the UCSC genome browser, all tracks being set with the same vertical viewing range (1 to 600). The two black triangles represent the two strong TAMs. **(B)** Normalized $\log_2$(RPKM) expression values of the *THAP11* gene, in WT and THAP11$^{F80L/+}$ mutant cells, shown as the mean +/- standard deviation of the two technical replicates.

study showing that Ronin is essential for mitochondrial function in the developing retina [74]. Interestingly, we showed that THAP11 functions are HBM dependent, as mutation of the HBM sequence in the ectopic THAP11 construct severely attenuates the gene expression changes implied by THAP11 ectopic synthesis. This is also in accordance with previous publications [29, 39, 43, 54, 70, 71, 75].

Probing the gene-expression changes between WT and THAP11$^{F80L/F80L}$ mutant cells also revealed an association to mitochondria, and pathways linked to transcription and development. Thus, the two experiments, ectopically expressing or mutating THAP11, respectively, are in agreement by revealing the same regulated pathways. Not surprisingly, the effect in the THAP11$^{F80L/F80L}$ mutant cells is milder and pathways are less standing out compared to the ectopic-synthesis experiment. Indeed, this is expected because the mutation has a more subtle effect than the forced expression, as it affects only a subset of THAP11-bound promoters. Notably, proliferation-associated pathways do not stand out in the GO analyses of mutant cells, while showing up in the stable-cell experiments and that proliferation assays demontrated a strong impact of the mutation on cell proliferation (Chapter 7). In addition, the direction of gene-expression changes in the two set ups are the same: upregulated genes in the stable or mutant cells point towards transcription and development, while downregulated genes point towards mitochondria, when compared to the WT or parental cells. This observation meets with the proliferation experiments in which THAP11 ectopic synthesis had a similar impact on cell proliferation than the p.F80L mutation (Chapter 7). This strenghtens the previous hypothesis which suggested that THAP11 increased level has a dominant-negative effect.

As mentioned, both THAP11 stable and mutant cells revealed GO cellular components associated to mitochondria. This is particularly interesting as the cobalamin disorders have been associated with alteration of mitochondria [129,130]. Also, cells bearing different HCF-1 *cblX*-causing mutations exhibit a morphological change of their mitochondria (Laura Sposito, personal communication).

Second, I focused on the human THAP11 *cblX*-associated mutation and we performed a genome-wide analysis of THAP11 DNA association, in both WT and THAP11$^{F80L/F80L}$ cells. It revealed a selective dissociation of the mutant THAP11 protein from a subset of DNA sites. On the other hand, the THAP11 p.F80L mutation does not enable THAP11 to bind additional promoters compared to the WT protein. We thus investigated the reason behind the selective DNA dissociation, meaning the difference between the DNA locations that remain bound by the mutant protein, compared to the ones that are still bound despite the mutation. While both categories of peaks reveal the same underlying motif — referred as THAP11-associated motif (TAM) — it is stronger for the peaks that remain bound despite the THAP11 p.F80L mutation compared to the ones that disappear. Also, most of peaks still present in THAP11$^{F80L/F80L}$ cells have at least an underlying TAM, and they exhibit a high score in the WT sample. By contrast, the disappearing peaks have lower WT scores, and only half of them have a TAM. These conclusions led me to

build the following model: upon p.F80L mutation, THAP11 protein selectively dissociates from its weakest DNA sites, which are the ones that have a weaker, if any, TAM. This results in a selective dissociation of THAP11 from specific promoters, the associated genes likely being affected.

Indeed, a transcriptomic analysis comparing the gene expression profiles of WT and THAP11$^{F80L/F80L}$ cells revealed, as already mentioned, differentially expressed genes associated with transcription and mitochondria. The third most affected gene is actually *MMACHC*, the associated protein being a key enzyme in the cobalamin pathway. Most of cobalamin-disorder forms are due to mutations in the *MMACHC* gene. It has already been suggested that THAP11, together with HCF-1, regulates *MMACHC* transcription [75]. I have demonstrated here that the mutant THAP11$_{F80L}$ protein dissociates from the *MMACHC* promoter, resulting in a marked decrease in gene transcription. This specific change in DNA binding and subsequent gene transcription is likely to be responsible of a large part of the mutation pathogenicity. It is thus not surprising that THAP11 and HCF-1 mutations fall into the same group of cobalamin disorders. As it has been suggested that they would inter-dependently regulate *MMACHC* expression, the dissociation of any of the two from the *MMACHC* promoter would result in a similar phenotype, accounting mainly for the dramatic decrease in *MMACHC* gene expression. Nevertheless, I suspect that other affected genes — primarily downstream targets — also play a role in the pathogenicity and the specificity of the *cblX* cobalamin-disorder type. For instance, the *cblX* group of cobalamin-disorder is distinct from *cblC* one, which is due to mutations in the *MMACHC* gene. Biochemical characteristics of *cblC* and *cblX* are similar, though milder in *cblX* patients, but the two disorders are clinically distinct, *cblX* exhibiting a more severe neurologic phenotype [126].

The gene ontology analyses did not identify any cobalamin-related pathway. This is actually expected: even though *MMACHC* dysregulation is likely to account for the majority of the mutation pathogenicity, it is the sole cobalamin-related gene affected by the mutation. The MMACHC protein is an enzyme and thus will not have any downstream transcriptional effect. Thus, it is not surprising that the cobalamin pathway was not identified by the ontology analyses as it concerns a single gene. Curiously, however, GO analyses revealed the KEGG pathway associated to lupus (systemic lupus erythematosus), which is an auto-immune disease. The putative association between lupus and THAP11 is thus worth being investigated, as the cause of this disease remains unknown, the cure non-existant and the treatments not very effective.

Third, I also investigated whether the THAP11 p.F80L mutation affects the expression of its own gene. I showed that THAP11 has the potential to regulate *THAP11* gene expression because it binds its own promoter. This binding is not greatly affected in the presence of the p.F80L mutation. Also, in agreement with what has been published [75], I showed that *THAP11* gene expression is not affected in THAP11$^{F80L/F80L}$ mutant cells. Thus, the decreased level of THAP11 mutant protein observed in the THAP11$^{F80L/F80L}$ cells (Figure 6.9) in not a consequence of a lack of THAP11-promoter self recruitement and subsequent decrease

in gene expression. Alternatively, it can be due to a decrease in mRNA translation or to an increased degradation of the mutant protein. As THAP11 p.F80L is suspected to affect the folding of its THAP domain, a degradation mechanism is more likely responsible for the lower level of THAP11 mutant protein.

Fourth, the analysis of the heterozygous THAP11$^{F80L/+}$ mutant cells is contradictory with the proliferation results (Chapter 7). Indeed, I showed that despite a small delay in cell proliferation at specific time points, heterozygous cells grown at 37 °C end up behaving similarly to WT cells. On the other hand, when grown at 39.5 °C, heterozygous cells display a sizeable impairment in cell proliferation. Transcriptomic analyses demonstrated that, overall, these cells have a transcriptome more similar to WT cells than to homozygous mutant cells, both at 37 °C and at 39.5 °C. Particularly, they do not exhibit any change in *MMACHC* expression. The discrepancy may come from the fact that RNA was extracted from cells only 2 days after the temperature switch. Indeed, when looking at Figure 7.7, the impairment of cell proliferation in heterozygous THAP11$^{F80L/+}$ mutant cells compared to WT ones observed at day 3 (2 days post temperature switch), when grown at 39.5 °C, is less pronounced that at the end time points. Thus, an hypothesis is that, at 39.5 °C, heterozygous cells are initially able to compensate for the hemizygous loss of the WT THAP11 protein, but eventually fail to compensate and exhibit defects. To probe this hypothesis, it would be interesting to repeat the transcriptomic analyses at later times after the temperature switch. It is unlikely, however, that the compensation would be mediated by THAP11 auto-regulation to increase the expression of its own gene, as RNA-seq data of heterozygous cells show no variation in the *THAP11* mRNA levels in the heterozygous cells grown at 37 °C or 39.5 °C. Whatever the case, the heterozygous THAP11 p.F80L mutation does not causes any significant defect when grown at 37 °C. It would be interesting to know whether the siblings of the THAP11$^{F80L/F80L}$ *cblX*-like patient are asymptomatic carriers of the mutation in an heterozygous state. These data, however, are unavalaible as the parents of the patient declined any genetic testing on themselves and their four other healthy children. Thus, it is not possible to know whether the p.F80L mutation has appeared spontaneously, in an homozygous fashion, in the patient, or whether it has been transmitted from the parents to their son. Due to the low likelihood of a spontaneous homozygous mutation, it is more probable that the parents are heterozygous carriers of the mutation. This variant must be extremely rare, as it was not found in public databases.

Overall, these results lead me to propose a model for the mechanism and pathogenicity of the THAP11 p.F80L mutation associated with cobalamin disorder. On a whole, I suggest that the effects of the THAP11 p.F80L mutation are not due to a change in THAP11 DNA-binding ability, but rather simply due to a decrease level of the protein, which then becomes a limiting factor. I propose that, upon mutation, the resulting THAP11$_{F80L}$ protein is unstable — probably because of folding issues of the THAP domain — leading to its increased degradation. As the protein level is much lower, it is limited for binding to DNA

and probably in too small quantity to bind to all the promoter sites normally occupied by the WT protein. Consequently, the limited amount of THAP11$_{F80L}$ mutant protein will preferentially bind to the DNA sites for which it has the strongest affinity, which are the ones that display at least one underlying TAM and the strongest one. Then, a subset of THAP11-regulated genes is lacking the THAP11-promoter binding and are thus misregulated. Among them, *MMACHC* is strongly downregulated in the mutant cells, which probably accounts for a large part of the pathogenicity. As HCF-1 and THAP11 co-regulate *MMACHC* gene expression, it is thus not surprising that mutations in both protein resulting in DNA-binding impairments lead to extremely similar diseases that are classified together (*cblX*). Other misregulated genes are likely to also play some role in the disease pathogenicity and to account for the specificity of the *cblX* disease group.

Thus, a decrease in THAP11 protein level such as the one observed in THAP11$^{F80L/F80L}$ cells has dramatic consequences. Indeed, it induces broad changes in the transcriptional program. As already shown in Figure 4.1, THAP11 is evenly expressed in the different human tissues. In addition, the Human Protein Atlas [131] references that, despite differential levels of mRNA, THAP11 protein is present at similar levels in the different human tissues. Thus, THAP11-mediated transcription in the different organs is likely to trigger similar transcriptional programs. Particularly, it probably triggers *MMACHC* expression in all the tissues, which is reassuring as the latter enzyme is necessary for cell metabolism.

It is important here to point out again that this analysis has been done on a single clone per cell type. Thus, the results, even if they are in agreement with the disease pathogenicity and previous analysis of the patients fibroblasts [75], will need to be confirmed with additional independent cell clones.

Future studies on THAP11$^{F80L/F80L}$ cells will work to correlate ChIP-seq and RNA-seq data to unravel the primary gene targets, besides *MMACHC*, affected by the THAP11 p.F80L mutation. It would also clarify downstream target genes and pathways. This will probably help to highlight the mechanisms underlying the mutation pathogenicity, particularly the difference between the *cblX* and the *cblC* groups, and the neurological involvement in the *cblX* type. Indeed, cobalamin-metabolism impairment causes neurological disorders by itself, as vitamin B_{12} is particularly important for the normal functioning of the brain and the nervous system. The *cblX* disease group, however, exhibit a more severe neurological trouble compared to the *cblC* one caused by the sole *MMACHC* mutation, suggesting that additional neurological mechanisms take place in the THAP11$^{F80L/F80L}$ patient and other *cblX* ones.

Chapter 9

Concluding thoughts

When starting my PhD work, I reviewed the literature regarding HCF-1, which is the main research topic of the Herr laboratory. Doing so, I encountered the THAP11 protein presented as an important co-factor for HCF-1. Digging into the subject, I realized that THAP11 belongs to a family of related proteins, some of them having already been suggested to interact and collaborate with HCF-1 at the onset of my work. First, I was extremely interested by these proteins and their interplay with HCF-1. While some of them were largely understudied at that time, some have already been proved to bear exciting properties, such as their transcriptional role in cell proliferation. Second, I rapidly realized that a global analysis of these different proteins as a family, and not as separate entities, was lacking from the published reports. I thus felt that the study of these proteins as an entire family may provide an interesting research opportunity. In particular, studying a family of related factors enables me to compare and to contrast their respective properties. Third, as their importance was still emerging, tools to study them were often lacking. Nevertheless, this last point did not discourage me given my enthusiasm of studying this family of proteins. Thus, I decided to study the THAP family of proteins and to try to understand how they regulate transcription and subsequently cell proliferation.

Thanks to the power of bioinformatics, publicly-available bio-databases and recent state-of-the-art high-throughput technologies, I was able to study in parallel all the different human THAP proteins, and to shed light on their structure, their evolution among animals and their pattern of expression in mice and human (Chapters 3 and 4). I showed that, despite sharing similarities such as their defining THAP domain, the THAP proteins are also notably diverse regarding their sequence and structure. Also, examining human and mouse gene expression revealed that each *THAP* gene has a specific expression pattern. Furthermore, I demonstrated that the HBM sequence and coiled-coil domain in THAP proteins are under positive selective pressure, indicating the importance of their interaction with HCF-1 and with themselves. Subsequently,

biochemistry analysis of selected THAP proteins (Chapter 5) showed that each of them possesses differing capacities for self binding — homo- and heterodimerization — and for interaction with HCF-1. These various results led to the conclusion that the different THAP proteins display similarities and differences, which can lead to both shared and specialized partners, functions and outcomes. At a certain point in my research, I realized that my analysis was limited by the available experimental tools, such as antibodies or engineered cells. While the generation of an antibody directed to the THAP7 protein was not successful, I managed to engineer several custom cell lines — with site-specific genomic mutations or stably transfected — to study further the two THAP7 and THAP11 proteins (Chapter 6). The phenotypic study of these different cell lines demonstrated that both THAP7 and THAP11 are required for proper cell proliferation (Chapter 7). Finally, I worked further on THAP11 and showed that it associates with DNA to regulate genes involved in development, cell proliferation and transcription (Chapter 8). Doing so, I also clarified the molecular mechanisms underlying the THAP11 cobalamin disorder-associated mutation. I proposed a model in which the THAP11 p.F80L mutation induces the destabilization of the THAP11 protein leading to its subsequent degradation. This results in a pronounced decrease of THAP11 protein level, THAP11 then probably becoming a limiting factor for DNA binding. I thus further hypothesize that THAP becomes limiting for binding to gene promoters and consequently associates with DNA sites for which it has the highest affinity, causing a selective dissociation of the mutant THAP11 protein from a subset of promoter sites, leading to a misregulation of the corresponding genes. The *MMACHC* gene being the sole misregulated gene linked to the cobalamin pathway, it likely accounts for a large part of the mutation pathogenicity (Chapter 8). In summary, the results presented in this thesis implicate the THAP transcriptional factors as important regulators of development, and cell proliferation and metabolism that happen to be associated to a variety of diseases.

I have already extensively discussed the described results at the end of each chapter and suggested many follow-up research directions. Here, I rather illuminate some more general considerations related to my thesis work. In particular, my thesis makes me dive into the field of human genetics, which was relatively new for me. Subsequent to the completion of this work, I realize how powerful this field has become, notably thanks to the development of state-of-the-art techniques in the past decade.

First, DNA sequencing has been a revolution in the field of human biology, but not exactly in the way it was anticipated. Before the completion of the human-genome sequencing, people had tremendous hopes on what it would bring. Particularly, it was often thought that the human genome would be quite easy to read and that it would solve many things in and of itself, particularly diseases such as cancer. It became rapidly clear, however, that knowing the sequence of the human genome was not going to give the magical result. Indeed, the human genome is far more complex to decipher than anticipated and remains, even over

15 years after the first completion of the Human Genome Project, in good part a mystery. Consequently, people placing a lot of hope in the knowledge of the human-genome sequence were disappointed.

Nevertheless, the (near) completion of the human-genome sequence was a gigantic step forward in the research in biology and medicine. It nevertheless provides the (near-complete) sequence of the human genome, which, even if not as easy to read as initially expected, is a bottomless source of information and research. In addition, it has been the starting point of the development of state-of-the-art high-throughput sequencing technologies. The cost to sequence the human genome has been reduced by more that 2000-fold in a decade, while the speed of sequencing increased considerably: it took almost 15 years to get the first sequence of the human genome, while nowadays it is possible to fully sequence the genome of an individual within a day. This revolution had enormous impacts on, among other fields, the identification, research and management of rare genetic diseases. Indeed, the democratization of high-throughput sequencing techniques allowed to understand disease genetics and this revolution has benefited most to the field of rare genetic diseases. The identification of the genetic trait responsible for a particular disease not only allows to decipher the molecular mechanisms but also has the potential to lead to its better clinical management. Furthermore, diseases are an extremely valuable source of knowledge: they exhibit the malfunctions of the human body and thus are a window onto how it works. Thus, integrating rare genetic diseases — which in an of themselves are individually much more numerous than common ones — susbtantially enlarges the disease window and provides many more opportunities to understand the human body.

Also, DNA sequencing has revolutionized the way of doing genetics as it does not only enable me to study the genome of cells *per se*. Indeed, it also enables the study of the transcriptome (RNA-seq, thanks to the reverse transcription of RNA into DNA). Furthermore, it allows the analysis of numerous aspects of the chromatin, such as the proteins bound to it, directly or not (ChIP-seq and derived techniques, DNA adenine methyltransferase identification DamID), its conformation and organization (chromosome-conformation capture techniques(3C)-based methods or chromatin accessibility assays such as MNase-seq or DNase-seq) or its composition (bisulfite sequencing for DNA methylation, modification-specific histone ChIPseq for histone composition and modifications). Consequently, DNA sequencing is not only a poweful tool for genetic studies, but also for epigenetics and gene-expression concerns.

Second, I used in my thesis research the state-of-the art CRISPR/Cas9 genome-editing technology. It allowed me to create mutant cell lines, both functional mutants to decipher the mechanism of action of my proteins of interest as well as disease-associated mutants. Here, I used a very simple application of the CRISPR/Cas9 system, but it is much more powerful and can be used to create whole organisms with the desired mutation. As a research point of view, this enables the recapitulation of a genetic disease in a genetically homogenous context, having mutant samples (cells, tissues, animals) carrying the exact genetic

modification as well as control ("healthy") samples for comparison as opposed to "real life" where each patient exhibits additional variability that is not linked to the disease. Thus, this technique is an incredibly powerful research tool. Furthermore, in a therapeutic point of view, the CRISPR/Cas9 genome-editing technology has the potential to precisely correct the genetic alteration responsible for a given genetic disease. CRISPR/Cas9-mediated gene therapy is still at its infancy and different problems still need to be solved, such as accuracy (off-target effects, raising safety issues), efficiency, and delivery methods [132, 133]. Nevertheless, this field is rapidly progressing and an increasing number of pre-clinical and clinical gene-editing trials are currently in progress [133]. Thus, one can have the optimistic hope that within a couple of decades, CRISPR/Cas9-mediated gene therapy will allow the correction of the genetic cause underlying some diseases.

To conclude, I feel lucky to have been able to use these two incredibly powerful techniques that are DNA sequencing and CRISPR/Cas9-mediated genome editing, and it taught me a lot. To finish with, I just want to raise some considerations about the development of these amazing technologies, particularly ethically speaking. Indeed, the development and democratization of these two techniques has raised important ethical issues that researchers, politicians and even lay people are struggling with.

Regarding DNA sequencing, the possibility of sequencing, at low cost, the genome of any individual raises questions about the use of these sequencing data. An example is the interpretations that can be inferred with such data, for instance at the scale of a single individuals. Nowadays, more and more companies such as 23andMe® or Pathway Genomics® are bein established, providing a DNA-sequencing service to anybody willing to pay for it, advertising the possibility to decipher the customer's origins or his/her genetic features, such as the explanation of his/her personality traits, or even the prediction of disease susceptibility. This last service related to diseases is, in my opinion, the most problematic one as the information is released to the customer without much background explanation, while the customer can be far from educated for properly understanding such results. This can result in dramatic consequences for the customer and generating aberrant behaviors, for example in the case he/she would understand that he/she would end up having Alzheimer's disease while his/her DNA simply contains alleles associated to an increased susceptibility to develop this disease. In addition, this also raises the question of availability and privacy of such data. Indeed, they could be for instance unfairly exploited by insurance companies (disease-associated data). Also, pharaceutical companies nowadays pay a fortune to buy such bio-databases, such as what happened recently (summer 2018) when 23andMe® sold almost its entire set of data to the english pharmaceutical company GSK (GlaxoSmithKline). I thus believe that DNA sequencing is an extremely valuable and powerful tool, but its democratization now urgently requires the creation of regulations and ethical guidelines.

As far as the CRISPR/Cas9 technology, I would like to highlight that manipulation of the genome, particularly the human one, is not without facing ethical issues. Indeed, mad scientists could envision to

not only repair the genome in case of genetic disorders, but also to "improve" it. Thus, the extent and circumstances to which the genome can be manipulated urgently needs to be defined by ethical advisory commitees.

As a conclusion, I have shared in these few pages my amazement regarding the revolution that has transformed (and is still transforming) the field of human genetics, and also my personal feeling that ethical guidelines as well as regulations are urgently needed.

Appendix

Genes upregulated in the Flp-In T-Rex THAP11$_{WT}$ cells compared to parental cells (DMSO treatment)

KEGG pathways terms	
Pvalue	*Ontology description*
8.19E-07	TGF-beta signaling pathway
6.92E-05	Inflammatory bowel disease (IBD)
Total genes found in KEGG pathways : 17 (over 374)	

GO biological process terms		
Pvalue		*Ontology description*
2.83E-10	●	anterior/posterior pattern specification
3.82E-10	●	embryonic skeletal system morphogenesis
1.21E-06	●	positive regulation of transcription from RNA polymerase II promoter
1.45E-06	●	transcription, DNA-templated
3.39E-06	●	cartilage development
5.09E-05	●	negative regulation of cell proliferation
5.73E-05	●	metanephros development
Total genes found in GO biological process : 93 (over 374)		

GO molecular function terms		
Pvalue		*Ontology description*
5.24E-08	●	sequence-specific DNA binding
1.32E-07	●	transcription factor activity, sequence-specific DNA binding
7.81E-07	●	RNA polymerase II core promoter proximal region sequence-specific DNA binding
Total genes found in GO molecular function : 55 (over 374)		

GO cellular component terms	
Pvalue	*Ontology description*
7.25E-06	extracellular region
4.92E-05	intracellular
Total genes found in GO cellular component : 86 (over 374)	

Supplemental Table S1: **GO terms associated with upregulated genes in the Flp-In™ T-Rex™ THAP11$_{WT}$ cells compared to parental cells (DMSO treatment).** GO terms related to development, proliferation and transcription are highlighted by purple, blue and orange dots, respectively.

Genes downregulated in the Flp-In T-Rex $THAP11_{WT}$ cells compared to parental cells (DMSO treatment)

KEGG pathways terms

Total genes found in KEGG pathways : 0 (over 564)

GO biological process terms

Total genes found in GO biological process : 0 (over 564)

GO molecular function terms

Pvalue	*Ontology description*
3.09E-08	protein binding
8.35E-06	G-protein coupled receptor activity
2.32E-05	RNA binding
2.37E-05	protein homodimerization activity

Total genes found in GO molecular function : 289 (over 564)

GO cellular component terms

Pvalue	*Ontology description*
1.48E-07	mitochondrion
3.36E-06	mitochondrial matrix
6.80E-06	cytoplasm
2.44E-05	nucleus
2.71E-05	cytosol
5.63E-05	nucleoplasm
6.26E-05	myelin sheath
8.17E-05	cytosolic small ribosomal subunit

Total genes found in GO cellular component : 354 (over 564)

Supplemental Table S2: **GO terms associated with downregulated genes in the Flp-In™ T-Rex™ $THAP11_{WT}$ cells compared to parental cells (DMSO treatment).** GO terms related to mitochondria are highlighted by green dots.

Genes upregulated in the Flp-In T-Rex THAP11$_{WT}$ cells compared to parental cells (DOX treatment) 1/2

KEGG pathways terms	
Pvalue	*Ontology description*
5.07E-13	TGF-beta signaling pathway
1.50E-07	Signaling pathways regulating pluripotency of stem cells
8.20E-07	Cell adhesion molecules (CAMs)
1.12E-05	Hippo signaling pathway
1.47E-05	Inflammatory bowel disease (IBD)
1.54E-05	ECM-receptor interaction
9.10E-05	Pathways in cancer
1.00E-04	Hypertrophic cardiomyopathy (HCM)
Total genes found in KEGG pathways : 62 (over 538)	

GO biological process terms	
Pvalue	*Ontology description*
1.50E-08	positive regulation of transcription from RNA polymerase II promoter
9.14E-08	extracellular matrix organization
1.79E-07	transforming growth factor beta receptor signaling pathway
5.78E-07	positive regulation of cartilage development
1.09E-06	BMP signaling pathway
1.23E-06	negative regulation of cell proliferation
1.42E-06	anterior/posterior pattern specification
2.21E-06	embryonic skeletal system morphogenesis
6.62E-06	SMAD protein signal transduction
1.10E-05	positive regulation of pathway-restricted SMAD protein phosphorylation
2.00E-05	ureteric bud development
2.27E-05	cell cycle arrest
2.60E-05	metanephros development
2.82E-05	negative regulation of osteoblast differentiation
3.90E-05	branching involved in ureteric bud morphogenesis
4.12E-05	regulation of transcription, DNA-templated
5.89E-05	cellular response to BMP stimulus
6.30E-05	heart development
6.98E-05	transcription, DNA-templated
7.70E-05	positive regulation of osteoblast proliferation
9.42E-05	positive regulation of gene expression
1.00E-04	response to estrogen
Total genes found in GO biological process : 167 (over 538)	

GO molecular function terms	
Pvalue	*Ontology description*
1.69E-07	transcription factor activity, sequence-specific DNA binding
6.59E-07	protein binding
4.32E-06	zinc ion binding
6.10E-06	calcium ion binding
1.39E-05	heparin binding
2.12E-05	transforming growth factor beta binding
2.69E-05	poly(A) RNA binding
4.44E-05	RNA polymerase II core promoter proximal region sequence-specific DNA binding
8.36E-05	sequence-specific DNA binding
Total genes found in GO molecular function : 325 (over 538)	

Genes upregulated in the Flp-In T-Rex THAP11$_{WT}$ cells compared to parental cells (DOX treatment)
2/2

GO cellular component terms	
Pvalue	*Ontology description*
1.19E-07	cell surface
3.07E-07	plasma membrane
1.24E-06	extracellular exosome
1.54E-06	integral component of membrane
2.30E-06	extracellular region
3.02E-06	extracellular space
5.95E-05	cytoplasm
7.89E-05	extracellular matrix
8.49E-05	lysosomal membrane
Total genes found in GO cellular component : 383 (over 538)	

Supplemental Table S3: **GO terms associated with upregulated genes in the Flp-In™ T-Rex™ THAP11$_{WT}$ cells compared to parental cells (doxycycline treatment).** GO terms related to development, proliferation and transcription are highlighted by purple, blue and orange dots, respectively.

Genes downregulated in the Flp-In T-Rex THAP11$_{WT}$ cells compared to parental cells (DOX treatment)

KEGG pathways terms	
Pvalue	*Ontology description*
9.53E-08	Olfactory transduction
3.68E-06	Parkinson:s disease
1.85E-05	Oxidative phosphorylation
9.59E-05	Huntington:s disease
Total genes found in KEGG pathways : 23 (over 1043)	

GO biological process terms	
Pvalue	*Ontology description*
5.69E-07	G-protein coupled receptor signaling pathway
1.44E-05	DNA repair
1.73E-05	translation
6.55E-05	mitochondrial respiratory chain complex I assembly
Total genes found in GO biological process : 78 (over 1043)	

GO molecular function terms	
Pvalue	*Ontology description*
2.40E-12	protein binding
1.62E-10	G-protein coupled receptor activity
8.92E-06	metal ion binding
9.91E-06	endodeoxyribonuclease activity
Total genes found in GO molecular function : 563 (over 1043)	

GO cellular component terms	
Pvalue	*Ontology description*
4.47E-16	mitochondrion
9.86E-14	mitochondrial inner membrane
1.49E-09	mitochondrial matrix
2.44E-07	nucleus
1.54E-06	cytosol
1.79E-05	nucleoplasm
2.38E-05	cytoplasm
7.21E-05	plasma membrane
Total genes found in GO cellular component : 704 (over 1043)	

Supplemental Table S4: **GO terms associated with downregulated genes in the Flp-In™ T-Rex™ THAP11$_{WT}$ cells compared to parental cells (doxycycline treatment).** GO terms related to mitochondria are highlighted by green dots.

(A) Genes upregulated in the Flp-In T-Rex THAP11$_{WT}$ cells treated with DOX compared to treated with DMSO

KEGG pathways terms
Total genes found in KEGG pathways : 0 (over 27)

GO biological process terms
Total genes found in GO biological process : 0 (over 27)

GO molecular function terms
Total genes found in GO molecular function : 0 (over 27)

GO cellular component terms
Total genes found in GO cellular component : 0 (over 27)

(B) Genes downregulated in the Flp-In T-Rex THAP11$_{WT}$ cells treated with DOX compared to treated with DMSO

KEGG pathways terms
Total genes found in KEGG pathways : 0 (over 66)

GO biological process terms
Total genes found in GO biological process : 0 (over 66)

GO molecular function terms
Total genes found in GO molecular function : 0 (over 66)

GO cellular component terms	
Pvalue	*Ontology description*
7.53E-06	integral component of peroxisomal membrane
Total genes found in GO cellular component : 3 (over 66)	

Supplemental Table S5: **GO terms associated with differentially expressed genes between DMSO and doxycycline treatment of Flp-In™ T-Rex™ THAP11$_{WT}$ cells.** Genes upregulated **(A)** and downregulated **(B)** in the Flp-In™ T-Rex™ THAP11$_{WT}$ cells treated with doxycycline compared to treated with DMSO were submitted to Gene Ontology analysis.

(A) Genes upregulated in the Flp-In T-Rex THAP11$_{HBM}$ cells compared to parental cells (DMSO treatment)

KEGG pathways terms
Total genes found in KEGG pathways : 0 (over 68)

GO biological process terms
Total genes found in GO biological process : 0 (over 68)

GO molecular function terms
Total genes found in GO molecular function : 0 (over 68)

GO cellular component terms
Total genes found in GO cellular component : 0 (over 68)

(B) Genes downregulated in the Flp-In T-Rex THAP11$_{HBM}$ cells compared to parental cells (DMSO treatment)

KEGG pathways terms
Total genes found in KEGG pathways : 0 (over 167)

GO biological process terms	
Pvalue	*Ontology description*
5.54E-05	behavioral response to pain
Total genes found in GO biological process : 3 (over 167)	

GO molecular function terms
Total genes found in GO molecular function : 0 (over 167)

GO cellular component terms	
Pvalue	*Ontology description*
1.45E-05	proteinaceous extracellular matrix
Total genes found in GO cellular component : 10 (over 167)	

Supplemental Table S6: **GO terms associated with differentially expressed genes in the Flp-In™ T-Rex™ THAP11$_{HBM}$ cells compared to parental cells (DMSO treatment).** Genes upregulated **(A)** and downregulated **(B)** in the Flp-In™ T-Rex™ THAP11$_{HBM}$ cells compared to parental cells (DMSO treatment) were submitted to Gene Ontology analysis.

Genes upregulated in the Flp-In T-Rex THAP11$_{WT}$ cells
compared to Flp-In T-Rex THAP11$_{HBM}$ cells (DMSO treatment)

KEGG pathways terms	
Pvalue	*Ontology description*
9.82E-07	TGF-beta signaling pathway
Total genes found in KEGG pathways : 11 (over 479)	

GO biological process terms		
Pvalue		*Ontology description*
3.35E-08	●	heart development
4.65E-07	●	anterior/posterior pattern specification
2.27E-06	●	cartilage development
1.14E-05	●	embryonic skeletal system morphogenesis
1.36E-05	●	metanephros development
3.96E-05	●	thymus development
3.99E-05	●	angiogenesis
4.42E-05	●	positive regulation of transcription from RNA polymerase II promoter
4.54E-05		response to estrogen
4.77E-05	●	negative regulation of cell proliferation
7.62E-05	●	positive regulation of fibroblast proliferation
Total genes found in GO biological process : 81 (over 479)		

GO molecular function terms		
Pvalue		*Ontology description*
2.21E-09		protein binding
3.22E-06		calcium ion binding
1.39E-05	●	sequence-specific DNA binding
1.67E-05	●	transcription factor activity, sequence-specific DNA binding
1.76E-05		actin binding
2.87E-05		collagen binding
Total genes found in GO molecular function : 280 (over 479)		

GO cellular component terms	
Pvalue	*Ontology description*
7.72E-07	integral component of membrane
2.10E-05	extracellular exosome
Total genes found in GO cellular component : 200 (over 479)	

Supplemental Table S7: **GO terms associated with upregulated genes in the Flp-In™ T-Rex™ THAP11$_{WT}$ cells compared to Flp-In™ T-Rex™ THAP11$_{HBM}$ cells (DMSO treatment).** GO terms related to development, proliferation and transcription are highlighted by purple, blue and orange dots, respectively.

Genes downregulated in the Flp-In T-Rex THAP11$_{WT}$ cells compared to Flp-In T-Rex THAP11$_{HBM}$ cells (DMSO treatment)

KEGG pathways terms	
Pvalue	*Ontology description*
7.60E-08	Ribosome
1.54E-07	Oxidative phosphorylation
1.16E-06	Huntington:s disease
1.91E-06	Parkinson:s disease
1.43E-05	Alzheimer:s disease
Total genes found in KEGG pathways : 38 (over 577)	

GO biological process terms	
Pvalue	*Ontology description*
6.77E-10	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
1.02E-09	SRP-dependent cotranslational protein targeting to membrane
1.20E-08	viral transcription
1.66E-08	translation
1.12E-06	translational initiation
3.28E-06	rRNA processing
2.97E-05	DNA repair
4.61E-05	G-protein coupled receptor signaling pathway
Total genes found in GO biological process : 54 (over 577)	

GO molecular function terms	
Pvalue	*Ontology description*
1.29E-15	protein binding
1.39E-07	unfolded protein binding
1.92E-07	structural constituent of ribosome
2.72E-07	RNA binding
6.78E-07	G-protein coupled receptor activity
1.62E-06	poly(A) RNA binding
3.24E-06	protein homodimerization activity
Total genes found in GO molecular function : 341 (over 577)	

GO cellular component terms	
Pvalue	*Ontology description*
3.48E-12	cytosol
2.71E-11	● mitochondrion
3.19E-11	● mitochondrial inner membrane
1.71E-10	nucleoplasm
5.31E-10	● mitochondrial matrix
4.13E-09	cytoplasm
8.07E-09	cytosolic small ribosomal subunit
2.01E-08	ribosome
1.87E-07	nucleus
9.27E-07	myelin sheath
2.40E-05	small ribosomal subunit
9.83E-05	integral component of membrane
Total genes found in GO cellular component : 447 (over 577)	

Supplemental Table S8: **GO terms associated with downregulated genes in the Flp-In™ T-Rex™ THAP11$_{WT}$ cells compared to Flp-In™ T-Rex™ THAP11$_{HMB}$ cells (DMSO treatment).** GO terms related to mitochondria are highlighted by green dots.

Genes upregulated in the THAP11$^{F80L/F80L}$ cells compared to WT cells (37°C)

KEGG pathways terms

Total genes found in KEGG pathways : 0 (over 346)

GO biological process terms

Pvalue	*Ontology description*
5.42E-20	regulation of transcription, DNA-templated
2.72E-19	transcription, DNA-templated
2.38E-04	regulation of G-protein coupled receptor protein signaling pathway
6.14E-04	dendritic spine development

Total genes found in GO biological process : 97 (over 346)

GO molecular function terms

Pvalue	*Ontology description*
3.05E-16	nucleic acid binding
1.91E-14	metal ion binding
3.67E-12	transcription factor activity, sequence-specific DNA binding
2.95E-10	DNA binding
6.58E-06	protein binding
1.31E-04	RNA polymerase II regulatory region sequence-specific DNA binding
8.72E-04	cytokine receptor activity

Total genes found in GO molecular function : 219 (over 346)

GO cellular component terms

Pvalue	*Ontology description*
2.46E-11	intracellular
5.14E-11	nucleus
2.05E-04	phagocytic vesicle membrane

Total genes found in GO cellular component : 147 (over 346)

Supplemental Table S9: **GO terms associated with upregulated genes in the THAP11$^{F80L/F80L}$ (homoyzygous) cells compared to WT cells** (37 °C)**.** GO terms related to development and transcription are highlighted by purple and orange dots, respectively.

Genes downregulated in the THAP11$^{F80L/F80L}$ cells compared to WT cells (37°C)

KEGG pathways terms	
Pvalue	*Ontology description*
2.57E-24	Systemic lupus erythematosus
1.69E-23	Alcoholism
1.42E-06	Viral carcinogenesis
3.97E-04	Olfactory transduction
Total genes found in KEGG pathways : 45 (over 669)	

GO biological process terms	
Pvalue	*Ontology description*
8.47E-07	chromatin silencing
1.95E-05	nucleosome assembly
2.58E-04	endoderm formation
3.63E-04	nitric oxide biosynthetic process
4.60E-04	actin filament organization
5.17E-04	G-protein coupled receptor signaling pathway
5.21E-04	actin cytoskeleton organization
8.47E-04	acyl-CoA metabolic process
8.58E-04	immune response
8.69E-04	long-chain fatty-acyl-CoA biosynthetic process
Total genes found in GO biological process : 65 (over 669)	

GO molecular function terms	
Pvalue	*Ontology description*
7.51E-07	protein binding
1.74E-05	G-protein coupled receptor activity
3.32E-05	DNA binding
2.65E-04	protein C-terminus binding
3.62E-04	catalytic activity
6.54E-04	four-way junction DNA binding
Total genes found in GO molecular function : 358 (over 669)	

GO cellular component terms	
Pvalue	*Ontology description*
9.20E-12	nucleosome
2.56E-10	mitochondrion
1.87E-07	mitochondrial inner membrane
1.31E-06	extracellular exosome
2.30E-06	nucleus
7.48E-06	nuclear nucleosome
1.64E-05	mitochondrial matrix
1.50E-04	protein complex
2.13E-04	cytosol
2.38E-04	cytoplasm
6.34E-04	plasma membrane
7.13E-04	microtubule organizing center
8.42E-04	microtubule
Total genes found in GO cellular component : 478 (over 669)	

Supplemental Table S10: **GO terms associated with downregulated genes in the THAP11$^{F80L/F80L}$ (homozygous) cells compared to WT cells** (37 °C)**.** GO terms related to development, transcription and mitochondria are highlighted by purple, orange and green dots, respectively.

Genes upregulated in the THAP11$^{F80L/+}$ cells compared to WT cells (37°C)

KEGG pathways terms
Total genes found in KEGG pathways : 0 (over 171)

GO biological process terms	
Pvalue	*Ontology description*
4.07E-05	intrinsic apoptotic signaling pathway by p53 class mediator
8.40E-05	DNA damage response, signal transduction by p53 resulting in cell cycle arrest
1.13E-04	protein transport
1.28E-04	positive regulation of extrinsic apoptotic signaling pathway in absence of ligand
Total genes found in GO biological process : 20 (over 171)	

GO molecular function terms	
Pvalue	*Ontology description*
6.63E-04	protein binding
6.63E-04	tumor necrosis factor-activated receptor activity
Total genes found in GO molecular function : 84 (over 171)	

GO cellular component terms
Total genes found in GO cellular component : 0 (over 171)

Supplemental Table S11: **GO terms associated with upregulated genes in the THAP11$^{F80L/+}$ (heterozygous) cells compared to WT cells** (37 °C)**.**

Genes downregulated in the THAP11$^{F80L/+}$ cells compared to WT cells (37°C)

KEGG pathways terms	
Pvalue	*Ontology description*
1.24E-05	Alcoholism
2.24E-05	Systemic lupus erythematosus
3.21E-05	Viral carcinogenesis
2.99E-04	Inflammatory mediator regulation of TRP channels
5.64E-04	Pathways in cancer
5.71E-04	Axon guidance
6.09E-04	Glioma
8.34E-04	HTLV-I infection
8.43E-04	Melanoma
Total genes found in KEGG pathways : 25 (over 142)	

GO biological process terms	
Pvalue	*Ontology description*
1.01E-05	response to progesterone
7.79E-05	hair follicle development
1.39E-04	wound healing
4.38E-04	positive regulation of cell proliferation
4.39E-04	inactivation of MAPK activity
4.93E-04	negative regulation of axon extension involved in axon guidance
6.84E-04	response to estradiol
6.86E-04	response to drug
7.20E-04	positive regulation of endothelial cell proliferation
8.28E-04	regulation of cell differentiation
Total genes found in GO biological process : 29 (over 142)	

GO molecular function terms
Total genes found in GO molecular function : 0 (over 142)

GO cellular component terms	
Pvalue	*Ontology description*
3.25E-05	nucleosome
2.84E-04	extracellular exosome
4.90E-04	intracellular
5.59E-04	nucleus
Total genes found in GO cellular component : 73 (over 142)	

Supplemental Table S12: **GO terms associated with downregulated genes in the THAP11$^{F80L/+}$ (heterozygous) cells compared to WT cells** (37 °C)**.** GO terms related to development and proliferation are highlighted by purple and blue dots, respectively.

Genes upregulated in THAP11$^{F80L/F80L}$ cells compared to THAP11$^{F80L/+}$ cells (37°C)

KEGG pathways terms	
Total genes found in KEGG pathways : 0 (over 479)	

GO biological process terms		
Pvalue		*Ontology description*
3.73E-21	●	regulation of transcription, DNA-templated
1.61E-19	●	transcription, DNA-templated
4.96E-05	●	negative regulation of transcription from RNA polymerase II promoter
7.55E-05		G-protein coupled receptor signaling pathway
7.27E-04		negative regulation of endothelial cell migration
Total genes found in GO biological process : 133 (over 479)		

GO molecular function terms		
Pvalue		*Ontology description*
6.81E-18		metal ion binding
4.25E-15	●	nucleic acid binding
1.20E-13	●	DNA binding
5.15E-09	●	transcription factor activity, sequence-specific DNA binding
1.14E-06		protein binding
2.30E-04		zinc ion binding
2.98E-04		cytokine receptor activity
Total genes found in GO molecular function : 318 (over 479)		

GO cellular component terms	
Pvalue	*Ontology description*
1.30E-15	nucleus
6.65E-11	intracellular
Total genes found in GO cellular component : 197 (over 479)	

Supplemental Table S13: **GO terms associated with upregulated genes in the THAP11$^{F80L/F80L}$ (homozygous) cells compared to THAP11$^{F80L/+}$ (heterozygous) cells** (37 °C)**.** GO terms related to transcription are highlighted by orange dots.

Genes downregulated in the THAP11$^{F80L/F80L}$ cells compared to THAP11$^{F80L/+}$ cells (37°C)

KEGG pathways terms	
Pvalue	*Ontology description*
7.59E-09	Alcoholism
2.84E-08	Systemic lupus erythematosus
7.67E-07	Olfactory transduction
4.78E-06	Viral carcinogenesis
8.04E-04	Arginine biosynthesis
9.43E-04	Pyrimidine metabolism
Total genes found in KEGG pathways : 54 (over 925)	

GO biological process terms	
Pvalue	*Ontology description*
1.84E-07	G-protein coupled receptor signaling pathway
3.81E-04	cellular amino acid metabolic process
5.25E-04	phosphorylation
6.90E-04	response to drug
7.03E-04	intrinsic apoptotic signaling pathway
8.04E-04	2-oxoglutarate metabolic process
8.44E-04	intrinsic apoptotic signaling pathway in response to DNA damage by p53
8.65E-04	urogenital system development
9.24E-04	metabolic process
Total genes found in GO biological process : 85 (over 925)	

GO molecular function terms	
Pvalue	*Ontology description*
4.30E-10	G-protein coupled receptor activity
4.63E-08	protein binding
8.01E-06	hydrolase activity
4.56E-05	catalytic activity
3.62E-04	isomerase activity
Total genes found in GO molecular function : 443 (over 925)	

GO cellular component terms	
Pvalue	*Ontology description*
9.61E-12	mitochondrion
1.10E-07	cytoplasm
4.00E-06	cytosol
1.04E-05	mitochondrial matrix
6.16E-05	plasma membrane
7.06E-05	extracellular exosome
1.03E-04	nucleus
1.14E-04	mitochondrial inner membrane
1.21E-04	integral component of plasma membrane
3.84E-04	neuronal cell body
6.72E-04	microtubule
6.92E-04	peroxisome
8.88E-04	intracellular
Total genes found in GO cellular component : 647 (over 925)	

Supplemental Table S14: **GO terms associated with downregulated genes in THAP11$^{F80L/F80L}$ (homozygous) cells compared to THAP11$^{F80L/+}$ (heterozygous) cells** (37 °C)**.** GO terms related to development and mitochondria are highlighted by purple and green dots, respectively.

Bibliography

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell, Fifth Edition.* 5th ed., 2008.

[2] K. Luger, A. W. Mader, R. K. Richmond, D. F. Sargent, and T. J. Richmond, "Crystal structure of the nucleosome resolution core particle at 2 . 8 A," *Nature*, vol. 389, pp. 251–260, 1997.

[3] T. Kouzarides, "Chromatin Modifications and Their Function," *Cell*, vol. 128, no. 4, pp. 693–705, 2007.

[4] B. Li, M. Carey, and J. L. Workman, "The Role of Chromatin during Transcription," *Cell*, vol. 128, no. 4, pp. 707–719, 2007.

[5] V. W. Zhou, A. Goren, and B. E. Bernstein, "Charting histone modifications and the functional organization of mammalian genomes," *Nature Reviews Genetics*, vol. 12, no. 1, pp. 7–18, 2011.

[6] A. J. Ruthenburg, C. D. Allis, and J. Wysocka, "Methylation of Lysine 4 on Histone H3: Intricacy of Writing and Reading a Single Epigenetic Mark," *Molecular Cell*, vol. 25, no. 1, pp. 15–30, 2007.

[7] E. Shen, H. Shulha, Z. Weng, and S. Akbarian, "Regulation of histone H3K4 methylation in brain development and disease," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1652, 2014.

[8] N. J. Marianayagam, M. Sunde, and J. M. Matthews, "The power of two: Protein dimerization in biology," *Trends in Biochemical Sciences*, vol. 29, no. 11, pp. 618–625, 2004.

[9] G. D. Amoutzias, D. L. Robertson, Y. Van de Peer, and S. G. Oliver, "Choose your partners: dimerization in eukaryotic transcription factors," *Trends in Biochemical Sciences*, vol. 33, no. 5, pp. 220–229, 2008.

[10] N. Dyson, "The regulation of E2F by pRB-family proteins.," *Genes & development*, vol. 12, pp. 2245–62, aug 1998.

[11] J. M. Trimarchi and J. A. Lees, "Sibling rivalry in the E2F family," *Nature Reviews Molecular Cell Biology*, vol. 3, no. 1, pp. 11–20, 2002.

[12] J. P. Demuth and M. W. Hahn, "The life and death of gene families," *BioEssays*, vol. 31, no. 1, pp. 29–39, 2009.

[13] T. Ohta, "Evolution of gene families.," *Gene*, vol. 259, no. 1-2, pp. 45–52, 2000.

[14] M. Nei and A. P. Rooney, "Concerted and birth-and-death evolution of multigene families.," *Annual review of genetics*, vol. 39, no. 11, pp. 121–52, 2005.

[15] J. Wysocka and W. Herr, "The herpes simplex virus VP16-induced complex: The makings of a regulatory switch," *Trends in Biochemical Sciences*, vol. 28, no. 6, pp. 294–304, 2003.

[16] F. Capotosti, S. Guernier, F. Lammers, P. Waridel, Y. Cai, J. Jin, J. W. Conaway, R. C. Conaway, and W. Herr, "O-GlcNAc transferase catalyzes site-specific proteolysis of HCF-1," *Cell*, vol. 144, no. 3, pp. 376–388, 2011.

[17] L. Huang, L. A. Jolly, S. Willis-Owen, A. Gardner, R. Kumar, E. Douglas, C. Shoubridge, D. Wieczorek, A. Tzschach, M. Cohen, A. Hackett, M. Field, G. Froyen, H. Hu, S. A. Haas, H. H. Ropers, V. M. Kalscheuer, M. A. Corbett, and J. Gecz, "A noncoding, regulatory mutation implicates HCFC1 in nonsyndromic intellectual disability," *American Journal of Human Genetics*, vol. 91, no. 4, pp. 694–702, 2012.

[18] L. A. Jolly, L. S. Nguyen, D. Domingo, Y. Sun, S. Barry, M. Hancarova, P. Plevova, M. Vlckova, M. Havlovicova, V. M. Kalscheuer, C. Graziano, T. Pippucci, E. Bonora, Z. Sedlacek, and J. Gecz, "HCFC1 loss-of-function mutations disrupt neuronal and neural progenitor cells of the developing brain," *Human Molecular Genetics*, vol. 24, no. 12, pp. 3335–3347, 2015.

[19] P. S. Tarpey, R. Smith, E. Pleasance, A. Whibley, S. Edkins, C. Hardy, S. O'Meara, C. Latimer, E. Dicks, A. Menzies, P. Stephens, M. Blow, C. Greenman, Y. Xue, C. Tyler-Smith, D. Thompson, K. Gray, J. Andrews, S. Barthorpe, G. Buck, J. Cole, R. Dunmore, D. Jones, M. Maddison, T. Mironenko, R. Turner, K. Turrell, J. Varian, S. West, S. Widaa, P. Wray, J. Teague, A. Butler, A. Jenkinson, M. Jia, D. Richardson, R. Shepherd, R. Wooster, M. I. Tejada, F. Martinez, G. Carvill, R. Goliath, A. P. De Brouwer, H. Van Bokhoven, H. Van Esch, J. Chelly, M. Raynaud, H. H. Ropers, F. E. Abidi, A. K. Srivastava, J. Cox, Y. Luo, U. Mallya, J. Moon, J. Parnau, S. Mohammed, J. L. Tolmie, C. Shoubridge, M. Corbett, A. Gardner, E. Haan, S. Rujirabanjerd, M. Shaw, L. Vandeleur, T. Fullston, D. F. Easton, J. Boyle, M. Partington, A. Hackett, M. Field, C. Skinner, R. E. Stevenson, M. Bobrow, G. Turner, C. E. Schwartz, J. Gecz, F. L. Raymond, P. A. Futreal, and M. R. Stratton, "A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation," *Nature Genetics*, vol. 41, no. 5, pp. 535–543, 2009.

[20] C. Koufaris, A. Alexandrou, G. Tanteles, V. Anastasiadou, and C. Sismani, "A novel HCFC1 variant in male siblings with intellectual disability and microcephaly in the absence of cobalamin disorder," *Biomedical Reports*, pp. 215–218, 2015.

[21] H. C. Yu, J. L. Sloan, G. Scharer, A. Brebner, A. M. Quintana, N. P. Achilly, I. Manoli, C. R. Coughlin, E. A. Geiger, U. Schneck, D. Watkins, T. Suormala, J. L. Van Hove, B. Fowler, M. R. Baumgartner, D. S. Rosenblatt, C. P. Venditti, and T. H. Shaikh, "An X-linked cobalamin disorder caused by mutations in transcriptional coregulator HCFC1," *American Journal of Human Genetics*, vol. 93, no. 3, pp. 506–514, 2013.

[22] M. Gérard, G. Morin, A. Bourillon, C. Colson, S. Mathieu, D. Rabier, T. Billette de Villemeur, H. Ogier de Baulny, and J. F. Benoist, "Multiple congenital anomalies in two boys with mutation in HCFC1 and cobalamin disorder," *European Journal of Medical Genetics*, vol. 58, no. 3, pp. 148–153, 2015.

[23] H. Goto, S. Motomura, A. C. Wilson, R. N. Freiman, Y. Nakabeppu, K. Fukushima, M. Fujishima, W. Herr, and T. Nishimoto, "A single-point mutation in HCF causes temperature-sensitive cell-cycle arrest and disrupts VP16 function," *Genes and Development*, vol. 11, no. 6, pp. 726–737, 1997.

[24] E. Julien and W. Herr, "Proteolytic processing is necessary to separate and ensure proper cell growth and cytokinesis functions of HCF-1," *EMBO Journal*, vol. 22, no. 10, pp. 2360–2369, 2003.

[25] S. Minocha, T. L. Sung, D. Villeneuve, F. Lammers, and W. Herr, "Compensatory embryonic response to allele-specific inactivation of the murine X-linked gene Hcfc1," *Developmental Biology*, vol. 412, no. 1, pp. 1–17, 2016.

[26] S. Minocha, S. Bessonnard, T. L. Sung, C. Moret, D. B. Constam, and W. Herr, "Epiblast-specific loss of HCF-1 leads to failure in anterior-posterior axis specification," *Developmental Biology*, vol. 418, no. 1, pp. 75–88, 2016.

[27] J. Wysocka, P. T. Reilly, and W. Herr, "Loss of HCF-1-chromatin association precedes temperature-induced growth arrest of tsBN67 cells.," *Molecular and cellular biology*, vol. 21, pp. 3820–9, jun 2001.

[28] S. Tyagi, A. L. Chabes, J. Wysocka, and W. Herr, "E2F Activation of S Phase Promoters via Association with HCF-1 and the MLL Family of Histone H3K4 Methyltransferases," *Molecular Cell*, vol. 27, no. 1, pp. 107–119, 2007.

[29] J. B. Parker, H. Yin, A. Vinckevicius, and D. Chakravarti, "Host cell factor-1 recruitment to E2F-bound and cell-cycle-control genes is mediated by THAP11 and ZNF143.," *Cell reports*, vol. 9, pp. 967–82, nov 2014.

[30] A. C. Wilson, R. N. Freiman, H. Goto, T. Nishimoto, and W. Herr, "VP16 targets an amino-terminal domain of HCF involved in cell cycle progression.," *Molecular and cellular biology*, vol. 17, pp. 6139–46, oct 1997.

[31] R. N. Freiman and W. Herr, "Viral mimicry: Common mode of association with HCF by VP16 and the cellular protein LZIP," *Genes and Development*, vol. 11, no. 23, pp. 3122–3127, 1997.

[32] R. L. Luciano and A. C. Wilson, "HCF-1 Functions as a Coactivator for the Zinc Finger Protein Krox20," *Journal of Biological Chemistry*, vol. 278, no. 51, pp. 51116–51124, 2003.

[33] J. Wysocka, M. P. Myers, C. D. Laherty, R. N. Eisenman, and W. Herr, "re- peats (Wilson et al. 1993a, 1995; Kristie et al. 1995). These cleavages result in stable N- (HCF-1," *Genes & Development*, pp. 896–911, 2003.

[34] L. R. Thomas, A. M. Foshage, A. M. Weissmiller, T. M. Popay, B. C. Grieb, S. J. Qualls, V. Ng, B. Carboneau, S. Lorey, C. M. Eischen, and W. P. Tansey, "Interaction of MYC with host cell factor-1 is mediated by the evolutionarily conserved Myc box IV motif," *Oncogene*, vol. 35, p. 3613, nov 2015.

[35] A. Yokoyama, Z. Wang, J. Wysocka, M. Sanyal, D. J. Aufiero, I. Kitabayashi, W. Herr, and M. L. Cleary, "Leukemia Proto-Oncoprotein MLL Forms a SET1-Like Histone Methyltransferase Complex with Menin To Regulate Hox Gene Expression," *Molecular and Cellular Biology*, vol. 24, no. 13, pp. 5639–5649, 2004.

[36] Y. J. Machida, Y. Machida, A. A. Vashisht, J. A. Wohlschlegel, and A. Dutta, "The deubiquitinating enzyme BAP1 regulates cell growth via interaction with HCF-1," *Journal of Biological Chemistry*, vol. 284, no. 49, pp. 34179–34188, 2009.

[37] S. Misaghi, S. Ottosen, A. Izrael-Tomasevic, D. Arnott, M. Lamkanfi, J. Lee, J. Liu, K. O'Rourke, V. M. Dixit, and A. C. Wilson, "Association of C-Terminal Ubiquitin Hydrolase BRCA1-Associated Protein 1 with Cell Cycle Regulator Host Cell Factor 1," *Molecular and Cellular Biology*, vol. 29, no. 8, pp. 2181–2192, 2009.

[38] R. Mazars, A. Gonzalez-de Peredo, C. Cayrol, A. C. Lavigne, J. L. Vogel, N. Ortega, C. Lacroix, V. Gautier, G. Huet, A. Ray, B. Monsarrat, T. M. Kristie, and J. P. Girard, "The THAP-Zinc finger protein THAP1 associates with coactivator HCF-1 and O-GlcNAc transferase a link between DYT6 and DYT3 dystonias," *Journal of Biological Chemistry*, vol. 285, no. 18, pp. 13364–13371, 2010.

[39] J. B. Parker, S. Palchaudhuri, H. Yin, J. Wei, and D. Chakravarti, "A Transcriptional Regulatory Role of the THAP11-HCF-1 Complex in Colon Cancer Cell Function," *Molecular and Cellular Biology*, vol. 32, no. 9, pp. 1654–1670, 2012.

[40] M. Roussigne, S. Kossida, A.-C. Lavigne, T. Clouaire, V. Ecochard, A. Glories, F. Amalric, and J.-P. Girard, "The THAP domain: a novel protein motif with similarity to the DNA-binding domain of P element transposase.," *Trends in biochemical sciences*, vol. 28, pp. 66–9, feb 2003.

[41] D. Bessière, C. Lacroix, S. Campagne, V. Ecochard, V. Guillet, L. Mourey, F. Lopez, J. Czaplicki, P. Demange, A. Milon, J. P. Girard, and V. Gervais, "Structure-function analysis of the THAP zinc finger of THAP1, a large C2CH DNA-binding module linked to Rb/E2F pathways," *Journal of Biological Chemistry*, vol. 283, no. 7, pp. 4352–4363, 2008.

[42] V. Gervais, S. Campagne, J. Durand, I. Muller, and A. Milon, "NMR studies of a new family of DNA binding proteins: The THAP proteins," *Journal of Biomolecular NMR*, vol. 56, no. 1, pp. 3–15, 2013.

[43] M. Dejosez, S. S. Levine, G. M. Frampton, W. A. Whyte, S. A. Stratton, M. C. Barton, P. H. Gunaratne, R. A. Young, and T. P. Zwaka, "Ronin/Hcf-1 binds to a hyperconserved enhancer element and regulates genes involved in the growth of embryonic stem cells," *Genes and Development*, vol. 24, no. 14, pp. 1479–1484, 2010.

[44] A. Sabogal, A. Y. Lyubimov, J. E. Corn, J. M. Berger, and D. C. Rio, "THAP proteins target specific DNA sites through bipartite recognition of adjacent major and minor grooves," *Nature Structural and Molecular Biology*, vol. 17, no. 1, pp. 117–124, 2010.

[45] M. P. Balakrishnan, L. Cilenti, C. Ambivero, Y. Goto, M. Takata, J. Turkson, X. S. Li, and A. S. Zervos, "THAP5 is a DNA-binding transcriptional repressor that is regulated in melanoma cells during DNA damage-induced cell death," *Biochemical and Biophysical Research Communications*, vol. 404, no. 1, pp. 195–200, 2011.

[46] T. Clouaire, M. Roussigne, V. Ecochard, C. Mathe, F. Amalric, and J.-P. Girard, "The THAP domain of THAP1 is a large C2CH module with zinc-dependent sequence-specific DNA-binding activity.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 19, pp. 6907–12, 2005.

[47] H. Quesneville, D. Nouaud, and D. Anxolabehere, "Recurrent Recruitment of the THAP DNA-Binding Domain and Molecular Domestication of the P-Transposable Element," *Molecular Biology and Evolution*, vol. 22, pp. 741–746, mar 2005.

[48] S. E. Hammer, S. Strehl, and S. Hagemann, "Homologs of Drosophila P Transposons Were Mobile in Zebrafish but Have Been Domesticated in a Common Ancestor of Chicken and Human," *Molecular Biology and Evolution*, vol. 22, pp. 833–844, apr 2005.

[49] S. Majumdar, A. Singh, and D. C. Rio, "The human THAP9 gene encodes an active P-element DNA transposase," *Science*, vol. 339, no. 6118, pp. 446–448, 2013.

[50] P. Burkhard, J. Stetefeld, and S. V. Strelkov, "Coiled coils: A highly versatile protein folding motif," *Trends in Cell Biology*, vol. 11, no. 2, pp. 82–88, 2001.

[51] Y. Lin, A. Khokhlatchev, D. Figeys, and J. Avruch, "Death-associated protein 4 binds MST1 and augments MST1-induced apoptosis," *Journal of Biological Chemistry*, vol. 277, no. 50, pp. 47991–48001, 2002.

[52] C. Sengel, S. Gavarini, N. Sharma, L. J. Ozelius, and D. C. Bragg, "Dimerization of the DYT6 dystonia protein, THAP1, requires residues within the coiled-coil domain," *Journal of Neurochemistry*, vol. 118, no. 6, pp. 1087–1100, 2011.

[53] A. Richter, R. Hollstein, E. Hebert, F. Vulinovic, J. Eckhold, A. Osmanovic, R. Depping, F. J. Kaiser, and K. Lohmann, "In-depth Characterization of the Homodimerization Domain of the Transcription Factor THAP1 and Dystonia-Causing Mutations Therein," *Journal of Molecular Neuroscience*, vol. 62, no. 1, pp. 11–16, 2017.

[54] M. Dejosez, J. S. Krumenacker, L. J. Zitur, M. Passeri, L. F. Chu, Z. Songyang, J. A. Thomson, and T. P. Zwaka, "Ronin Is Essential for Embryogenesis and the Pluripotency of Mouse Embryonic Stem Cells," *Cell*, vol. 133, no. 7, pp. 1162–1174, 2008.

[55] C. D. Cukier, L. Maveyraud, O. Saurel, V. Guillet, A. Milon, and V. Gervais, "The C-terminal region of the transcriptional regulator THAP11 forms a parallel coiled-coil domain involved in protein dimerization," *Journal of Structural Biology*, vol. 194, no. 3, pp. 337–346, 2016.

[56] C. M. Bianchetti, C. A. Bingman, and G. N. Phillips, "Structure of the C-terminal heme-binding domain of THAP domain containing protein 4 from Homo sapiens," *Proteins: Structure, Function and Bioinformatics*, vol. 79, no. 4, pp. 1337–1341, 2011.

[57] G. De Simone, A. di Masi, F. Polticelli, and P. Ascenzi, "Human nitrobindin: the first example of an all-β-barrel ferric heme-protein that catalyzes peroxynitrite detoxification.," *FEBS open bio*, vol. 8, pp. 2002–2010, dec 2018.

[58] N. Pandey, U. Mittal, A. K. Srivastava, and M. Mukerji, "SMARCA2 and THAP11: Potential candidates for polyglutamine disorders as evidenced from polymorphism and protein-folding simulation studies," *Journal of Human Genetics*, vol. 49, no. 11, pp. 596–602, 2004.

[59] R. H. Yin, Y. Li, F. Yang, Y. Q. Zhan, M. Yu, C. H. Ge, W. X. Xu, L. J. Tang, X. H. Wang, B. Chen, Y. Yang, J. J. Li, C. Y. Li, and X. M. Yang, "Expansion of the polyQ repeats in THAP11 forms intranuclear aggregation and causes cell G0/G1 arrest," *Cell Biology International*, vol. 38, no. 6, pp. 757–767, 2014.

[60] C. Cayrol, C. Lacroix, C. Mathe, V. Ecochard, M. Ceribelli, E. Loreau, V. Lazar, P. Dessen, R. Mantovani, L. Aguilar, and J. P. Girard, "The THAP-zinc finger protein THAP1 regulates endothelial cell proliferation through modulation of pRB/E2F cell-cycle target genes," *Blood*, vol. 109, no. 2, pp. 584–594, 2007.

[61] T. Macfarlan, S. Kutney, B. Altman, R. Montross, J. Yu, and D. Chakravarti, "Human THAP7 is a chromatin-associated, histone tail-binding protein that represses transcription via recruitment of HDAC3 and nuclear hormone receptor corepressor," *Journal of Biological Chemistry*, vol. 280, no. 8, pp. 7346–7358, 2005.

[62] T. Macfarlan, J. B. Parker, K. Nagata, and D. Chakravarti, "Thanatos-associated protein 7 associates with template activating factor-Ibeta and inhibits histone acetylation to repress transcription.," *Molecular endocrinology (Baltimore, Md.)*, vol. 20, no. 2, pp. 335–47, 2006.

[63] Y. Li, Q. Ning, J. Shi, Y. Chen, M. Jiang, L. Gao, W. Huang, Y. Jing, S. Huang, A. Liu, Z. Hu, D. Liu, L. Wang, C. Nervi, Y. Dai, M. Q. Zhang, and L. Yu, "A novel epigenetic AML1-ETO/THAP10/miR-383 mini-circuitry contributes to t(8;21) leukaemogenesis," *EMBO Molecular Medicine*, vol. 9, no. 7, pp. 933–949, 2017.

[64] M. Gale, C. M. Blakely, D. A. Hopkins, W. Mark, M. Wambach, P. R. Romano, G. Michael, M. W. Melville, and M. G. Katze, "Regulation of Interferon-Induced Protein Kinase PKR : Modulation of P58IPK Inhibitory Function by a Novel Protein , P52rIPK.," *Mol Cell Biol.*, 1998.

[65] F. Aguilo, Z. Zakirova, K. Nolan, R. Wagner, R. Sharma, M. Hogan, C. Wei, Y. Sun, M. J. Walsh, K. Kelley, W. Zhang, L. J. Ozelius, P. Gonzalez-Alegre, T. P. Zwaka, and M. E. Ehrlich, "THAP1: Role in Mouse Embryonic Stem Cell Survival and Differentiation," *Stem Cell Reports*, vol. 9, no. 1, pp. 92–107, 2017.

[66] M. P. Balakrishnan, L. Cilenti, Z. Mashak, P. Popat, E. S. Alnemri, and A. S. Zervos, "THAP5 is a human cardiac-specific inhibitor of cell cycle that is cleaved by the proapoptotic Omi/HtrA2 protease during cell death.," *American journal of physiology. Heart and circulatory physiology*, vol. 297, pp. H643–53, 2009.

[67] A. Miele, R. Medina, A. J. Van Wijnen, G. S. Stein, and J. L. Stein, "The interactome of the histone gene regulatory factor HiNF-P suggests novel cell cycle related roles in transcriptional control and RNA processing," *Journal of Cellular Biochemistry*, vol. 102, no. 1, pp. 136–148, 2007.

[68] C. Y. Zhu, C. Y. Li, Y. Li, Y. Q. Zhan, Y. H. Li, C. W. Xu, W. X. Xu, H. B. Sun, and X. M. Yang, "Cell growth suppression by thanatos-associated protein 11(THAP11) is mediated by transcriptional downregulation of c-Myc," *Cell Death and Differentiation*, vol. 16, no. 3, pp. 395–405, 2009.

[69] S. Nakamura, D. Yokota, L. Tan, Y. Nagata, T. Takemura, I. Hirano, K. Shigeno, K. Shibata, S. Fujisawa, and K. Ohnishi, "Down-regulation of Thanatos-associated protein 11 by BCR-ABL promotes CML cell proliferation through c-Myc expression," *International Journal of Cancer*, vol. 130, no. 5, pp. 1046–1059, 2012.

[70] J. Michaud, V. Praz, N. J. Faresse, C. K. JnBaptiste, S. Tyagi, F. Schütz, and W. Herr, "HCFC1 is a common component of active human CpG-island promoters and coincides with ZNF143, THAP11, YY1, and GABP transcription factor occupancy," *Genome Research*, vol. 23, no. 6, pp. 907–916, 2013.

[71] J. Fujita, P. Freire, C. Coarfa, A. L. Benham, P. Gunaratne, M. D. Schneider, M. Dejosez, and T. P. Zwaka, "Ronin Governs Early Heart Development by Controlling Core Gene Expression Programs," *Cell Reports*, vol. 21, no. 6, pp. 1562–1573, 2017.

[72] J. Durruthy-Durruthy, M. Wossidlo, S. Pai, Y. Takahashi, G. Kang, L. Omberg, B. Chen, H. Nakauchi, R. Reijo Pera, and V. Sebastiano, "Spatiotemporal Reconstruction of the Human Blastocyst by Single-Cell Gene-Expression Analysis Informs Induction of Naive Pluripotency," *Developmental Cell*, vol. 38, no. 1, pp. 100–115, 2016.

[73] B. A. Seifert, M. Dejosez, and T. P. Zwaka, "Ronin influences the DNA damage response in pluripotent stem cells," *Stem Cell Research*, vol. 23, pp. 98–104, 2017.

[74] R. A. Poché, M. Zhang, E. M. Rueda, X. Tong, M. L. McElwee, L. Wong, C. W. Hsu, M. Dejosez, A. R. Burns, D. A. Fox, J. F. Martin, T. P. Zwaka, and M. E. Dickinson, "RONIN Is an Essential Transcriptional Regulator of Genes Required for Mitochondrial Function in the Developing Retina," *Cell Reports*, vol. 14, no. 7, pp. 1684–1697, 2016.

[75] A. M. Quintana, H. C. Yu, A. Brebner, M. Pupavac, E. A. Geiger, A. Watson, V. L. Castro, W. Cheung, S. H. Chen, D. Watkins, T. Pastinen, F. Skovby, B. Appel, D. S. Rosenblatt, and T. H. Shaikh, "Mutations in THAP11 cause an inborn error of cobalamin metabolism and developmental abnormalities," *Human molecular genetics*, vol. 26, no. 15, pp. 2838–2849, 2017.

[76] A. Achilleos, X. Tong, and R. A. Poché, "A New Role of Ronin (Thap11) in the Neural Crest and Craniofacial Development in the Mouse," *The FASEB Journal*, vol. 31, no. 1_supplement, pp. 387.2–387.2, 2017.

[77] X. Z. Kong, R. H. Yin, H. M. Ning, W. W. Zheng, X. M. Dong, Y. Yang, F. F. Xu, J. J. Li, Y. Q. Zhan, M. Yu, C. H. Ge, J. H. Zhang, H. Chen, C. Y. Li, and X. M. Yang, "Effects of THAP11 on erythroid differentiation and megakaryocytic differentiation of K562 cells," *PLoS ONE*, vol. 9, no. 3, 2014.

[78] A. Kimchi, "DAP genes: Novel apoptotic genes isolated by a functional approach to gene cloning," *Biochimica et Biophysica Acta - Reviews on Cancer*, vol. 1377, no. 2, pp. 13–33, 1998.

[79] M. Roussigne, C. Cayrol, T. Clouaire, F. Amalric, and J. P. Girard, "THAP1 is a nuclear proapoptotic factor that links prostate-apoptosis-response-4 (Par-4) to PML nuclear bodies," *Oncogene*, vol. 22, no. 16, pp. 2432–2442, 2003.

[80] L. J. Ozelius and S. B. Bressman, "Genetic and clinical features of primary torsion dystonia," *Neurobiology of Disease*, vol. 42, no. 2, pp. 127–135, 2011.

[81] D. C. Bragg, I. A. Armata, F. C. Nery, X. O. Breakefield, and N. Sharma, "Molecular pathways in dystonia," *Neurobiology of Disease*, vol. 42, no. 2, pp. 136–147, 2011.

[82] M. S. LeDoux, J. Xiao, M. Rudzińska, R. W. Bastian, Z. K. Wszolek, J. A. Van Gerpen, A. Puschmann, D. Momčilović, S. R. Vemula, and Y. Zhao, "Genotype-phenotype correlations in THAP1 dystonia: Molecular foundations and description of new cases," *Parkinsonism and Related Disorders*, vol. 18, no. 5, pp. 414–425, 2012.

[83] A. Achilleos, X. Tong, and R. A. Poché, "Ronin (Thap11) is Implicated in a New Cobalamin Deficiency Syndrome Impacting the Central Nervous System," *The FASEB Journal*, vol. 31, no. 1_supplement, pp. 746.2–746.2, 2017.

[84] J. Gladitz, B. Klink, and M. Seifert, "Network-based analysis of oligodendrogliomas predicts novel cancer gene candidates within the region of the 1p/19q co-deletion.," *Acta neuropathologica communications*, vol. 6, no. 1, p. 49, 2018.

[85] F. Abate, A. C. da Silva-Almeida, S. Zairis, J. Robles-Valero, L. Couronne, H. Khiabanian, S. A. Quinn, M.-Y. Kim, M. A. Laginestra, C. Kim, D. Fiore, G. Bhagat, M. A. Piris, E. Campo, I. S. Lossos, O. A. Bernard, G. Inghirami, S. Pileri, X. R. Bustelo, R. Rabadan, A. A. Ferrando, and T. Palomero, "Activating mutations and translocations in the guanine exchange factor VAV1 in peripheral T-cell lymphomas," *Proceedings of the National Academy of Sciences*, vol. 114, no. 4, pp. 764–769, 2017.

[86] E. de Souza Santos, S. A. de Bessa, M. M. Netto, and M. A. Nagai, "Silencing of LRRC49 and THAP10 genes by bidirectional promoter hypermethylation is a frequent event in breast cancer," *International Journal of Oncology*, vol. 33, no. 1, pp. 25–31, 2008.

[87] R. A. Johnson, K. D. Wright, H. Poppleton, K. M. Mohankumar, D. Finkelstein, S. B. Pounds, V. Rand, S. E. Leary, E. White, C. Eden, T. Hogg, P. Northcott, S. MacK, G. Neale, Y. D. Wang, B. Coyle, J. Atkinson, M. Dewire, T. A. Kranenburg, Y. Gillespie, J. C. Allen, T. Merchant, F. A. Boop, R. A. Sanford, A. Gajjar, D. W. Ellison, M. D. Taylor, R. G. Grundy, and R. J. Gilbertson, "Cross-species genomics matches driver mutations and cell compartments to model ependymoma," *Nature*, vol. 466, no. 7306, pp. 632–636, 2010.

[88] W. X. Lian, R. H. Yin, X. Z. Kong, T. Zhang, X. H. Huang, W. W. Zheng, Y. Yang, Y. Q. Zhan, W. X. Xu, M. Yu, C. H. Ge, J. T. Guo, C. Y. Li, and X. M. Yang, "THAP11, a novel binding protein of PCBP1, negatively regulates CD44 alternative splicing and cell invasion in a human hepatoma cell line," *FEBS Letters*, vol. 586, no. 10, pp. 1431–1438, 2012.

[89] R. Hollstein, B. Reiz, L. Kötter, A. Richter, S. Schaake, K. Lohmann, and F. J. Kaiser, "Dystonia-causing mutations in the transcription factor THAP1 disrupt HCFC1 cofactor recruitment and alter gene expression," *Human molecular genetics*, vol. 26, no. 15, pp. 2975–2983, 2017.

[90] F. A. Ran, P. D. Hsu, J. Wright, V. Agarwala, D. A. Scott, and F. Zhang, "Genome engineering using the CRISPR-Cas9 system," *Nature Protocols*, vol. 8, no. 11, pp. 2281–2308, 2013.

[91] P. Horvath and R. Barrangou, "CRISPR/Cas, the immune system of bacteria and archaea.," *Science (New York, N.Y.)*, vol. 327, pp. 167–70, jan 2010.

[92] S. W. Cho, S. Kim, J. M. Kim, and J. S. Kim, "Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease," *Nature Biotechnology*, vol. 31, no. 3, pp. 230–232, 2013.

[93] T. R. Sampson and D. S. Weiss, "Exploiting CRISPR/Cas systems for biotechnology," *BioEssays*, vol. 36, no. 1, pp. 34–38, 2014.

[94] P. Mali, K. M. Esvelt, and G. M. Church, "Cas9 as a versatile tool for engineering biology," *Nature Methods*, vol. 10, no. 10, pp. 957–963, 2013.

[95] Life Technologies Corporation, "Flp-In™ System For Generating Stable Mammalian Expression Cell Lines by Flp Recombinase-Mediated Integration," *User guide*, vol. E, no. November, p. 40, 2010.

[96] Life Technologies Corporation, "Flp-In™ T-REx™ Core Kit For Generating Stable, Inducible Mammalian Expression Cell Lines by Flp Recombinase-Mediated Integration," *User guide*, no. 25, pp. 1–9, 2012.

[97] L. Zheng, U. Baumann, and J. L. Reymond, "An efficient one-step site-directed and site-saturation mutagenesis protocol.," *Nucleic acids research*, vol. 32, no. 14, 2004.

[98] A. C. Wilson, K. LaMarco, M. G. Peterson, and W. Herr, "The VP16 accessory protein HCF is a family of polypeptides processed from a large precursor protein," *Cell*, vol. 74, no. 1, pp. 115–125, 1993.

[99] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate, and M. Punta, "Pfam: the protein families database.," *Nucleic acids research*, vol. 42, pp. D222–30, jan 2014.

[100] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool.," *Journal of molecular biology*, vol. 215, pp. 403–10, oct 1990.

[101] P. Stothard, "The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences.," *BioTechniques*, vol. 28, pp. 1102, 1104, jun 2000.

[102] A. Lupas, M. Van Dyke, and J. Stock, "Predicting coiled coils from protein sequences.," *Science (New York, N.Y.)*, vol. 252, pp. 1162–4, may 1991.

[103] A. V. McDonnell, T. Jiang, A. E. Keating, and B. Berger, "Paircoil2: improved prediction of coiled coils from sequence.," *Bioinformatics (Oxford, England)*, vol. 22, pp. 356–8, feb 2006.

[104] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput.," *Nucleic acids research*, vol. 32, no. 5, pp. 1792–7, 2004.

[105] I. Letunic and P. Bork, "Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation.," *Bioinformatics (Oxford, England)*, vol. 23, pp. 127–8, jan 2007.

[106] I. Letunic and P. Bork, "Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy.," *Nucleic acids research*, vol. 39, pp. W475–8, jul 2011.

[107] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "STAR: ultrafast universal RNA-seq aligner.," *Bioinformatics (Oxford, England)*, vol. 29, pp. 15–21, jan 2013.

[108] B. Li and C. N. Dewey, "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.," *BMC bioinformatics*, vol. 12, p. 323, aug 2011.

[109] B. Li, V. Ruotti, R. M. Stewart, J. A. Thomson, and C. N. Dewey, "RNA-Seq gene expression estimation with read mapping uncertainty.," *Bioinformatics (Oxford, England)*, vol. 26, pp. 493–500, feb 2010.

[110] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, vol. 15, p. 550, dec 2014.

[111] The Gene Ontology Consortium, "Expansion of the Gene Ontology knowledgebase and resources.," *Nucleic acids research*, vol. 45, no. D1, pp. D331–D338, 2017.

[112] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.," *Nature genetics*, vol. 25, pp. 25–9, may 2000.

[113] F. Supek, M. Bošnjak, N. Škunca, and T. Šmuc, "REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms," *PLoS ONE*, vol. 6, p. e21800, jul 2011.

[114] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu, "Model-based analysis of ChIP-Seq (MACS).," *Genome biology*, vol. 9, no. 9, p. R137, 2008.

[115] M. Renaud, V. Praz, E. Vieu, L. Florens, M. P. Washburn, P. L'Hôte, and N. Hernandez, "Gene duplication and neofunctionalization: POLR3G and POLR3GL.," *Genome research*, vol. 24, pp. 37–51, jan 2014.

[116] T. L. Bailey and P. Machanick, "Inferring direct DNA binding from ChIP-seq," *Nucleic Acids Research*, vol. 40, no. 17, p. e128, 2012.

[117] G. Ambrosini, R. Groux, and P. Bucher, "PWMScan: a fast tool for scanning entire genomes with a position-specific weight matrix," *Bioinformatics*, vol. 34, no. 14, pp. 2483–2484, 2018.

[118] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and a. D. Haussler, "The Human Genome Browser at UCSC," *Genome Research*, vol. 12, pp. 996–1006, may 2002.

[119] Y. Qu, H. Zhao, N. Han, G. Zhou, G. Song, B. Gao, S. Tian, J. Zhang, R. Zhang, X. Meng, Y. Zhang, Y. Zhang, X. Zhu, W. Wang, D. Lambert, P. G. Ericson, S. Subramanian, C. Yeung, H. Zhu, Z. Jiang, R. Li, and F. Lei, "Ground tit genome reveals avian adaptation to living at high altitudes in the Tibetan plateau," *Nature Communications*, vol. 4, no. May, pp. 1–9, 2013.

[120] D. Brawand, M. Soumillon, A. Necsulea, P. Julien, G. Csárdi, P. Harrigan, M. Weier, A. Liechti, A. Aximu-Petri, M. Kircher, F. W. Albert, U. Zeller, P. Khaitovich, F. Grützner, S. Bergmann, R. Nielsen, S. Pääbo, and H. Kaessmann, "The evolution of gene expression levels in mammalian organs," *Nature*, vol. 478, no. 7369, pp. 343–348, 2011.

[121] G. Chen and X. Deng, "Cell Synchronization by Double Thymidine Block.," *Bio-protocol*, vol. 8, sep 2018.

[122] S. Minocha, D. Villeneuve, L. Rib, C. Moret, N. Guex, and W. Herr, "Segregated hepatocyte proliferation and metabolic states within the regenerating mouse liver," *Hepatology Communications*, vol. 1, no. 9, pp. 871–885, 2017.

[123] L. Rib, D. Villeneuve, S. Minocha, V. Praz, N. Hernandez, and N. Guex, "Cycles of gene expression and genome response during mammalian tissue regeneration," *bioRxiv Genomics*, pp. 1–19, 2018.

[124] R. Trollmann, H. Rehrauer, C. Schneider, G. Krischke, N. Huemmler, S. Keller, W. Rascher, and M. Gassmann, "Late-gestational systemic hypoxia leads to a similar early gene response in mouse placenta and developing brain," *AJP: Regulatory, Integrative and Comparative Physiology*, vol. 299, no. 6, pp. R1489–R1499, 2010.

[125] P. V. Hornbeck, B. Zhang, B. Murray, J. M. Kornhauser, V. Latham, and E. Skrzypek, "PhosphoSitePlus, 2014: mutations, PTMs and recalibrations.," *Nucleic acids research*, vol. 43, pp. D512–20, jan 2015.

[126] J. L. Sloan, N. Carrillo, D. Adams, and C. P. Venditti, *Disorders of Intracellular Cobalamin Metabolism*. 2018.

[127] A. A. Stepanenko and V. V. Dmitrenko, "HEK293 in cell biology and cancer research: phenotype, karyotype, tumorigenicity, and stress-induced genome-phenotype evolution.," *Gene*, vol. 569, pp. 182–90, sep 2015.

[128] G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T. I. Stein, R. Nudel, I. Lieder, Y. Mazor, S. Kaplan, D. Dahary, D. Warshawsky, Y. Guan-Golan, A. Kohn, N. Rappaport, M. Safran, and D. Lancet, "The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses.," *Current protocols in bioinformatics*, vol. 54, pp. 1.30.1–1.30.33, 2016.

[129] S. Krahenbuhl, D. B. Ray, S. P. Stabler, R. H. Allen, and E. P. Brass, "Increased hepatic mitochondrial capacity in rats with hydroxy-cobalamin[c-lactam]-induced methylmalonic aciduria.," *The Journal of clinical investigation*, vol. 86, pp. 2054–61, dec 1990.

[130] E. P. Frenkel, A. Mukherjee, C. R. Hackenbrock, and P. A. Srere, "Biochemical and ultrastructural hepatic changes during vitamin B12 deficiency in animals and man.," *The Journal of biological chemistry*, vol. 251, pp. 2147–54, apr 1976.

[131] M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, and F. Pontén, "Proteomics. Tissue-based map of the human proteome.," *Science (New York, N.Y.)*, vol. 347, p. 1260419, jan 2015.

[132] H. Mollanoori and S. Teimourian, "Therapeutic applications of CRISPR/Cas9 system in gene therapy.," *Biotechnology letters*, vol. 40, pp. 907–914, jun 2018.

[133] M. L. Maeder and C. A. Gersbach, "Genome-editing Technologies for Gene and Cell Therapy.," *Molecular therapy : the journal of the American Society of Gene Therapy*, vol. 24, pp. 430–46, mar 2016.