Aus dem Institut für Humangenetik

Theoretische Medizin und Biowissenschaften

der Medizinischen Fakultät

der Universität des Saarlandes, Homburg/Saar

Analyses of human endogenous retrovirus-encoded proteins with potential relevance for human biology and diseases

Dissertation zur Erlangung des Grades eines Doktors der Naturwissenschaften

der Medizinischen Fakultät

der UNIVERSITÄT DES SAARLANDES

2020

vorgelegt von

Giuseppe Rigogliuso

geboren am 30.01.89 in Lanusei, Italien

Tag der Promotion: 27.11.2020
Dekan: Prof. Dr. M.D. Menger
Berichterstatter: Prof. J. Mayer
   Prof. U. Walldorf

John Laroche: *...the point is, what's so wonderful is that every one of these flowers has a specific relationship with the insect that pollinates it. There's a certain orchid look exactly like a certain insect so the insect is drawn to this flower, its double, its soul mate, and wants nothing more than to make love to it. And after the insect flies off, spots another soul-mate flower and makes love to it, thus pollinating it. And neither the flower nor the insect will ever understand the significance of their lovemaking. I mean, how could they know that because of their little dance the world lives? But it does. By simply doing what they're designed to do, something large and magnificent happens. In this sense they show us how to live, how the only barometer you have is your heart. How, when you spot your flower, you can't let anything get in your way.*

From *Adaptation* by Charlie Kaufman

Contents

1. Summary

The human genome harbors numerous sequences derived from ancient retroviral infections, so-called human endogenous retroviruses (HERVs). The HERVs of the HERV-K(HML-2) group, in short HML-2, encode various retroviral proteins that, if expressed, may affect cell biology. HML-2 transcription is upregulated in different disease contexts including presence of HML-2 encoded proteins. Some HML-2 proteins have been implicated in disease development, although definitive associations with diseases were not reported yet.

This work focused on HML-2 protease (Pro) and HML-2 integrase (IN), two HML-2 encoded proteins that received little attention so far, although their enzymatic activities can, in principle, have detrimental effects in the cell.

In the first part of this work we investigated HML-2 Pro. Besides the retroviral proteins that constitute the virion, some retroviral aspartyl proteases, among them HIV-1 protease, are known to process cellular proteins. We wondered whether HML-2 Pro proteolytic activity can process cellular proteins as well. By cleaving cellular proteins, HML-2 Pro may impair cellular processes, thus have a role in disease development. The major aim for this part was to identify cellular proteins processed by HML-2 Pro. We also performed initial experiments documenting HML-2 Pro activity against cellular proteins in a cellular environment and provided insights into cellular processes that could be potentially influenced by HML-2 Pro activity.

HML-2 Pro was successfully purified and reaction conditions for assaying its activity *in vitro* were optimized. Purified HML-2 Pro was incubated with cellular proteins derived from HeLa cells. Employing terminal amine isotope labeling of substrates (TAILS) we then identified cellular proteins cleaved by HML-2 Pro. TAILS experiment and subsequent analyses of raw data produced a list of 872 human proteins as putative cellular substrates of HML-2 Pro. Identified proteins were profiled using suited bioinformatic tools. Processed proteins could be assigned to different cellular compartments and various, often disease-relevant cellular processes. We verified through additional experiments processing of selected candidate proteins by HML-2 Pro *in vitro* and *in vivo*. Cleavage by HML-2 Pro was confirmed for 9 out of 14 selected candidate proteins *in vitro*. Further verification experiments demonstrated *in vivo* cleavage of cellular proteins by HML-2 Pro in a cellular environment. Sizes of processing products observed for some of the tested proteins coincided with product sizes predicted by TAILS, thus corroborating TAILS results. We documented cell death and activation of apoptotic processes during HML-2 Pro overexpression, hence providing initial experimental evidence of cellular processes being influenced by HML-2 Pro activity. Finally, we obtained preliminary evidence of

functional, enzymatically active endogenous HML-2 Pro being present in some tumor cell lines known to express HML-2 proteins.

Our results suggest that hundreds of cellular proteins are potential substrates of HML-2 Pro. It is therefore conceivable that upregulated HML-2 transcription might lead to increased expression of active HML-2 Pro that might subsequently affect various cellular processes by degrading cellular proteins. HML-2 Pro is present in cells of some tumor types where its proteolytic activity could play a role. HML-2 Pro thus deserves further attention because cellular processes impaired by its activity potentially contribute to human diseases.

In the second part of this work we investigated HML-2 IN, an enzyme that, if expressed, might locate to the cell nucleus and exhibit catalytic activities inducing DNA damage and formation of DNA double-strand breaks (DSBs). DSBs are a severe type of DNA damage that may contribute to genome instability and tumor development. We aimed at evaluating whether HML-2 IN causes DNA damage resulting in formation of DSBs.

Following transient expression of HA-tagged wild-type and mutant HML-2 IN in HeLa cells, we performed immunofluorescence assays for monitoring HML-2 IN cellular localization through HA-tag detection as well as formation of DSBs through analysis of 53BP1 foci.
Transiently expressed HML-2 IN localized to nuclei but we did not observe significant differences regarding 53BP1 foci when comparing cells expressing wild-type HML-2 IN and controls. Although our results point towards HML-2 IN not causing DSBs when expressed in HeLa cells the experimental approach that we established can be considered a valuable starting point for further, more comprehensive investigations.

Zusammenfassung

Analysen Humaner Endogener Retrovirus-kodierter Proteine hinsichtlich deren Relevanz für die menschliche Biologie und Erkrankungen

Das menschliche Genom enthält zahlreiche Sequenzen, die von ehemaligen retroviralen Infektionen stammen, sogenannte humane endogene Retroviren (HERVs). Die HERVs der HERV-K (HML-2) Gruppe, kurz HML-2, kodieren verschiedene retrovirale Proteine, die, wenn sie exprimiert werden, die Zellbiologie beeinflussen können. Die HML-2 Transkription wird in verschiedenen Krankheitskontexten hochreguliert, einschließlich der Expression HML-2-kodierter Proteinen. Einige HML-2 Proteine sind eventuell an der Entwicklung bestimmter Krankheiten beteiligt, obwohl definitive Assoziationen mit solchen Krankheiten noch nicht beschrieben sind.

Diese Arbeit konzentrierte sich auf HML-2 Protease (Pro) und HML-2 Integrase (IN), zwei HML-2 kodierte Proteine, die bisher wenig Beachtung fanden, obwohl ihre enzymatischen Aktivitäten im Prinzip schädliche Auswirkungen auf die Zelle haben können.

Im ersten Teil dieser Arbeit haben wir HML-2 Pro untersucht. Es ist bekannt, dass neben den retroviralen Virion-bildenden Proteinen einige retrovirale Aspartylproteasen, darunter die HIV-1-Protease, auch noch zelluläre Proteine spalten können. Es bestand die Frage, ob die proteolytische Aktivität von HML-2 Pro auch zelluläre Proteine spalten kann. Durch die Spaltung von zellulären Proteinen kann HML-2 Pro zelluläre Prozesse beeinträchtigen und somit eine Rolle bei der Krankheitsentwicklung spielen. Das Hauptziel dieses Teils der Arbeit war die Identifizierung von zellulären Proteinen, die von HML-2 Pro prozessiert wurden. Wir führten auch erste Experimente durch, die die HML-2 Pro Aktivität gegen zelluläre Proteine in einer zellulären Umgebung dokumentierten, und lieferten Einblicke in zelluläre Prozesse, die möglicherweise durch die HML-2 Pro Aktivität beeinflusst werden könnten.

HML-2 Pro wurde erfolgreich gereinigt und die Reaktionsbedingungen zum Testen seiner Aktivität *in vitro* wurden optimiert. Gereinigtes HML-2 Pro wurde mit zellulären nativen Proteinen aus HeLa-Zellen inkubiert. Unter Verwendung der terminalen Aminisotopenmarkierung von Substraten (TAILS) identifizierten wir zelluläre Proteine, die durch HML-2 Pro prozessiert wurden. Das TAILS-Experiment und die anschließende Analyse der Rohdaten ergaben eine Liste von 872 humanen Proteinen als mutmaßliche zelluläre Substrate von HML-2 Pro. Die dentifizierten Proteine wurden unter Verwendung geeigneter bioinformatischer Werkzeuge charakterisiert. Die identifizierten Proteine konnten verschiedenen Zellkompartimenten und verschiedenen, oft krankheitsrelevanten Zellprozessen zugeordnet werden. Wir verifizierten ferner durch zusätzliche Experimente die

Prozessierung ausgewählter Kandidatenproteine durch HML-2 Pro *in vitro* und *in vivo*. Eine solche Prozessierung durch HML-2 Pro wurde für 9 von 14 ausgewählten Kandidatenproteinen *in vitro* bestätigt. Weitere Verifikationsexperimente zeigten *in vivo* die Prozessierung von zellulären Proteinen durch HML-2 Pro in einer zellulären Umgebung. Die Größen der für einige der getesteten Proteine beobachteten Prozessierungsprodukte stimmten hierbei mit den von TAILS vorhergesagten Produktgrößen überein, was die TAILS-Ergebnisse bestätigte. Wir dokumentierten weiter den Zelltod und die Aktivierung apoptotischer Prozesse während der Überexpression von HML-2 Pro und lieferten damit erste experimentelle Beweise, dass zelluläre Prozesse durch die HML-2 Pro-Aktivität beeinflusst werden. Schließlich erhielten wir vorläufige Beweise dafür, dass funktionelles, enzymatisch aktives endogenes HML-2 Pro in einigen Tumorzelllinien, von denen HML-2 Proteinexpression bekannt ist, ebenso exprimiert wird.

Unsere Ergebnisse legen nahe, dass Hunderte von zellulären Proteinen potenzielle Substrate von HML-2 Pro sind. Es ist daher denkbar, dass hochregulierte HML-2 Transkription zu einer erhöhten Expression von aktivem HML-2 Pro führt, die anschließend verschiedene zelluläre Prozesse durch Abbau zellulärer Proteine beeinflussen könnte. HML-2 Pro ist in Zellen einiger Tumortypen vorhanden, in denen seine proteolytische Aktivität eine Rolle spielen könnte. HML-2 Pro verdient daher weitere Aufmerksamkeit, da zelluläre Prozesse, die durch seine Aktivität beeinträchtigt werden, möglicherweise zu menschlichen Krankheiten beitragen.

Im zweiten Teil dieser Arbeit untersuchten wir HML-2 IN, ein Enzym, das bei Expression im Zellkern lokalisieren und katalytische Aktivitäten aufweisen könnte, die eine DNA-Schädigung und die Bildung von DNA-Doppelstrangbrüchen (DSBs) induzieren. DSBs sind eine schwerwiegende Art von DNA-Schäden, die zur Instabilität des Genoms und zur Tumorentwicklung beitragen können. Wir wollten untersuchen, ob HML-2 IN DNA-Schäden verursacht, die zur Bildung von DSBs führen.

Nach der Expression von HA-markiertem Wildtyp- und mutiertem HML-2 IN in HeLa-Zellen führten wir Immunfloreszenzexperimente zur Untersuchung der HML-2 IN-Zelllokalisation durch HA-Tag-Detektion sowie zur Analyse der Bildung von DSBs durch Analyse von 53BP1-Foci durch.
Transient exprimiertes HML-2 IN lokalisiert im Zellkern, jedoch beobachteten wir keine signifikanten Unterschiede in Bezug auf 53BP1-Foci im Vergleich von Zellen, die Wildtyp-HML-2 IN oder Kontrollen exprimierten. Obwohl unsere Ergebnisse darauf hindeuten, dass HML-2 IN bei Expression in HeLa-Zellen keine DSBs verursacht, kann der von uns etablierte experimentelle Ansatz als wertvoller Ausgangspunkt für weitere, umfassendere Untersuchungen angesehen werden.

2. Introduction

2.1. Retroviruses and endogenous retroviruses

Retroviruses, classified in the family *Retroviridae*, represent a vast group of enveloped RNA viruses with the ability to reverse transcribe their genome into a linear double-stranded DNA (dsDNA) that is subsequently integrated into the genome of their host cell. Because of such replication strategy, retroviruses can exist as exogenous virus particles or as endogenous DNA sequences permanently integrated in the host genome, the latter called proviruses [211]. Retroviruses that target somatic cells can be transmitted to new individuals through typical horizontal transmission of virus particles. Vertical transmission of virus particles can happen from mother to fetus, but if germ cells or gametes are infected, another way of vertical transmission is possible: if a provirus is integrated in a reproductive cell that subsequently forms a zygote, thus originating a new organism, a copy of that provirus will be carried by each cell of the newborn. Starting from that individual, the provirus will be transmitted in a mendelian fashion as a host allele, having a chance to be fixed in the entire progeny. Through this mechanism of vertical transmission, during the evolution of vertebrates, in the course of many million years, thousands of retrovirus elements have been integrated and fixed in the genome of species. Such retroviral elements are called endogenous retroviruses (ERVs) [137].

2.2 Human endogenous retroviruses (HERVs)

Originally identified in mice and chickens, ERVs were found in the genome of all vertebrates studied so far, including humans [102]. The first human endogenous retrovirus (HERV) was discovered in 1981 [133]. Since then, more than 700.000 retroviral elements have been identified in the human genome, corresponding to almost the 8% of the entire human genome [75,137]. Surprisingly, this is more than the percentage of human genome that encodes proteins, which corresponds to approximately 1.5% [84]. In some species, for example, in koala or mice, an active process of germ line colonization by novel ERV is still ongoing. In humans, it seems that the process of HERV endogenization has virtually ceased [128]. However, as will be described later, this does not mean that HERVs no longer influence human biology.

2.2.1 Proviral structure

HERVs share with exogenous retroviruses the typical proviral structure [128] (Figure 1). Their sequence consists of ~10 kb in full-length proviruses, although it can be shorter in defective proviruses (see later). The proviral genome is enclosed at the 5' and 3' ends by sequence regions called long terminal repeats (LTRs). LTRs are composed of regions U3, U5 and R. Such regions contain sequence elements that have a regulatory function in the retroviral replication cycle, in particular during reverse transcription, integration and transcription. In proximity of the 5' LTR is located the primer binding

site (PBS), which is the binding site for a tRNA that serves as primer during initiation of reverse transcription. In the portion of the proviral genome flanking the PBS, there is a coding-region containing the viral genes: *gag*, *pro*, *pol* and *env*. *gag* encodes the Gag (group-specific antigen) polyprotein that generates three structural proteins that compose the virion: matrix (MA), capsid (CA) and nucleocapsid (NC). *pro* encodes the viral protease (Pro). *pol* encodes the reverse transcriptase (RT), RNaseH and integrase (IN). *env* encodes the surface glycoprotein (SU) and transmembrane protein (TM), which are included in the lipidic envelope that surround the virion. Apart from these canonical proteins, some HERVs are characterized by a more complex genome that in some cases encode additional proteins. For instance, some HERV groups contain motifs for a deoxyuridine triphosphatase (dUTPase) [135]. Such dUTPase motifs has been found in variable genome positions among different groups, in particular in an N-terminal region of *pro* or in a C-terminal region of *pol* [135]. Other accessory proteins can derive from alternative splicing of the *env* gene (see chapter 2.3.1).

Looking at the genomic structure of the multitude of retrovirus-derived elements dispersed in the human genome, it became clear that, after integration, the majority of HERVs have been subjected to mutations and recombination processes. Such processes have reduced most of HERVs to defective proviruses with lower or totally missing coding capacity [206]. One representative case of defective HERVs are the so-called "solo LTRs", which consist of just an LTR generated by homologous recombination between the two proviral LTRs. Of note, solo LTRs are present at approximately 10-fold higher numbers in the human genome compared to HERV proviral sequences [128]. It is conceivable that evolutionary processes favored the proliferation of defective proviruses as a consequence of their reduced impact on genome stability. Although, several HERVs still retain intact genes and ability to encode functional proteins.
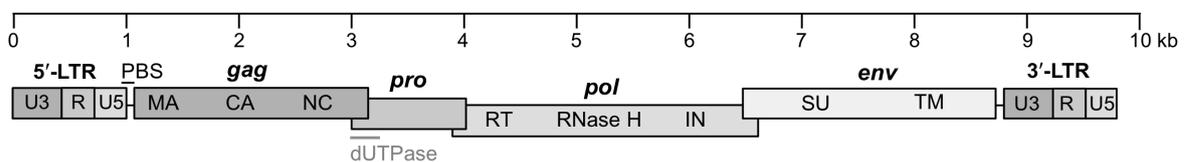
Figure 1: Representative structure of a HERV provirus. A full-length provirus consist of nearly 10 kb composed of two long terminal repeats (LTRs) that flank the internal portion of the viral genes *gag*, *pro*, *pol* and *env*. LTRs can be divided into the domains U3, R and U5. Reverse transcription starts from the primer binding site (PBS), located between the 5' LTR and *gag*. Viral genes encode proteins composed of different domains: MA = matrix, CA = capsid, NC = nucleocapsid, RT = reverse transcriptase, IN = integrase, SU = surface protein, TM = transmembrane protein. Several proviruses encode a protein domain for dUTPase, in some cases located in an N-terminal region of *pro*.

2.2.2 Classification and nomenclature of HERVs

Various criteria have been employed for classification and nomenclature of HERVs. For example, according to sequence similarities in the *pol* gene, HERVs have been broadly grouped in three main classes that basically reflect their similarities with exogenous retroviruses [70,88,206]: Class I elements

are more closely related to exogenous *gamma-* and *epsilonretroviruses*, Class II to *alpha-* and *betaretroviruses,* and Class III to *spumaviruses*. Another HERV classification was based on the tRNA that binds to the PBS to initiate reverse transcription [38]. For instance, HERV-K indicates a human ERV (or ERV lineage) that uses a lysin (K) tRNA, whereas HERV-W indicates an ERV that uses a tryptophan (W) tRNA. Some specific HERV loci were named using arbitrary designations that have evolved according to internal laboratory criteria based on, for instance, HERVs neighboring genes (e.g. HERV-ADP [126]), clone number (e.g. HERV-S71 [218]) or place of identification (e.g. HERV-K(HML-2.HOM) [134]). Very often in the literature a specific HERV locus was reported in different studies with different names, increasing confusion. Only recently a system has been proposed that would unify the nomenclature of ERVs to the level of individual proviral loci [71].

2.2.3 Impact of HERVs on human biology

After their discovery, retrovirus-related elements have been considered for a relatively long time as "useless" junk sequences, without a functional role in the cell. Nowadays, it is known that through different mechanisms HERVs can influence cellular processes, inducing both positive or negative effects [113].

Beneficial roles of HERVs are the consequence of adaptation processes, which provided novel functions to specific HERVs that became advantageous for the host. Several HERVs, including those that are relatively degraded in terms of coding capacity, can have important functions as cellular gene regulators. HERV LTRs harbor promoter and enhancer elements, binding sites for transcription factors, and a polyadenylation signal. These transcription-related elements provided fundamental functions for the integrated provirus, but it was demonstrated that they can contribute as functional elements to cellular genes [37]. For instance, LTRs can serve as alternative gene promoter, bringing to variable expression of cellular genes in different tissues or conditions [13]. There are also rarer examples of LTRs that replaced the original promoter of a gene, serving as unique promoter [167]. One representative case is the transcription of the amylase gene in the human salivary gland, which depends on the activity of a HERV-E LTR [177].

Another way HERVs can influence the cellular processes is through the production of retroviral proteins. The most remarkable example of HERV-derived proteins with a known cellular function is represented by syncytins, that are proteins encoded by cellular genes originated from the *env* gene of HERVs [115]. These proteins have been exapted on multiple occasions and independently in diverse mammalian species to play a role in placentation [115]. Syncytin-1, encoded by a proviral locus of the HERV-W group, has fusogenic properties that appear to be fundamental during the formation of the syncytial trophoblast layer during placenta development [131]. Curiously, it has been speculated also an involvement of syncytin-1 in multiple sclerosis [173], suggesting potential deleterious effects of this HERV-derived protein. Although nothing has been concretely established so far concerning the

relationship between syncytin-1 and multiple sclerosis, this example can be used to remark that, domesticated or not, HERVs could also have detrimental roles [113]. In fact, negative effects of apparently harmless HERVs could be suddenly triggered by various mechanisms. For instance, it is known that most HERVs are typically silenced in somatic cells by epigenetic mechanisms including, but not limited to, DNA methylation [128]. CpG methylation was demonstrated to have a role on the transcriptional activity of many HERVs [116]. As a consequence, it is conceivable that conditions of epigenetic reprogramming or altered epigenetic controls, with associated loss of DNA methylation, could reactivate suppressed HERVs. In support of this notion is the observation of increased levels of HERV expression after treatment of cells with demethylating agents [197]. HERV overexpression or HERV expression in situations when they should not be expressed could have detrimental consequences for the cell. As insertional elements, HERVs could have the potential to cause deleterious mutations, for example, disrupting the integrity of a gene or inducing activation of oncogenes, when integrated in their proximity. Also, HERV-derived proteins could potentially interact with cellular components and impair cellular processes, leading to alteration of cell functions. These phenomena could, for example, take place during tumorigenesis, when the genome methylation level is remarkably altered.

Some HERV groups could be more relevant than other in the disease context. In the following, attention will be focused on HERV-K(HML-2), a group of HERVs that has been investigated for its potential roles in tumor and other diseases. The remaining part of this introduction will describe general characteristics of HERV-K(HML-2), then consider specific aspects related to the two proteins studied here.

2.3 HERV-K(HML-2) proviruses

HERV-K is the only group of HERVs having human-specific members [54]. HERV-K group is currently divided into 11 subgroups, created on the base of similarities in the RT sequence [4,199]. The HERV-K subgrups are indicated with the acronim HML, for human MMTV-like, because they are closely related to the betaretrovirus mouse mammary mumor virus (MMTV). The most recently integrated and best preserved HERV-K proviruses belong to the HERV-K(HML-2) subgroup [139], hereafter referred to as HML-2. Because of several peculiar features, HML-2 has attracted a considerable interest in the scientific community. In 1986, the complete nucleotide sequence of HERV-K10 was published [148] (later classified as member of the HML-2 group), describing for the first time a HERV with intact ORFs for retroviral proteins. Because of its peculiar integrity, HERV-K10 has been used as a reference in the study of HML-2 proviruses. Various research groups started to investigate HERV-K10 and other HML-2 elements to elucidate properties, activities and eventually define potential functions in the cell. In a recent analysis, 944 HML-2 solo LTR elements and 91 full-

length HML-2 proviruses have been catalogued, reporting a number of exceptionally well-preserved proviruses [199]. As a consequence of relatively recent integrations, the HML-2 group also includes human-specific proviruses, some of which are polymorphic in the human population [11,81,208], i.e., specific proviral loci that are not fixed in the population and can be found in various allelic forms, including cases of full-length proviruses, solo LTRs, ERVs tandem repeats, or no viral sequence in the concerned genome position at all. One study that analyzed more than 2500 sequenced human genomes identified 36 polymorphic HML-2 proviruses, with insertion frequencies ranging from <0.0005 to >0.75 [219]. Taking into account that many unidentified HML-2 proviral loci may be polymorphic, it is conceivable that the number of HML-2 loci known at the moment could be an underestimation of the actual number in the entire human population [199].

2.3.1 Genomic organization

The genomic organization of HML-2 proviruses resemble that of betaretroviruses, with the four major genes (*gag*, *pro*, *pol*, *env*) having different, overlapping ORFs (Figure 2 A) [78]. Based on sequence properties, HML-2 proviruses can be clustered in different groups. Depending on the absence or presence of a 292-bp deletion in the junction between *pol* and *env* genes HML-2 proviruses can be classified as type 1 or type 2 [123], respectively. Another HML-2 variant, harboring a 96-bp insertion in a central region of the *gag* gene has been categorized in a group named HERV-K(OLD), with additional sequence features of HERV-K(OLD) proviruses suggesting a longer-time presence in the human genome [160]. Based on phylogenetic analysis of the LTR sequences, HML-2 can be also classified in three subgroups: LTR5Hs, LTR5A or LTR5B [64], with the former group including the most recently acquired proviruses, whereas the last two groups including evolutionarily older proviruses [30].

2.3.2 Proviral transcripts and proteins

For transcription of HML-2 proviruses, the 5' LTR provides the promoter recognized by the cellular RNA polymerase II. Transcription starts in the U5 region and results in full-length viral RNA molecules and subgenomic-sized RNA molecules that are spliced by the host cell spliceosome. Two donor sites (SD) and two acceptor site (SA) are present for splicing of the HML-2 transcripts (see Figure 2 A). All mRNA species (see Figure 2 B) are modified with a 5´CAP and a 3´poly-A tail, and exported to the cytoplasm, where they serve as mRNAs for translation of viral proteins. Full-length viral RNA transcripts are also packaged into newly generated virions as genomic RNA [78]. Typical for retroviruses, some HML-2 proteins are produced as polyproteins that need to be processed in order to release the single proteins. Gag, Pro and Pol are translated from a full-length, unspliced viral mRNA. Their translation is initiated at the starting AUG codon of the Gag ORF. Pro and Pol (the second and third ORFs) are produced as a Gag-Pro and Gag-Pro-Pol polyprotein. Since ORFs are arranged in

different reading frames, a mechanism of ribosomal frameshift is necessary to produce the different products. Gag is translated when the termination codon at the end of Gag ORF is recognized. Occasionally, the Gag termination codon is not reached due to slippage of the ribosome upstream of the Gag ORF stop codon. In this case, translation continues in a different reading frame that produces Gag-Pro fusion protein, instead of Gag alone. A similar frameshift mechanism can occasionally happen upstream of the Pro stop codon, generating Gag-Pro-Pol [174]. All protein domains embedded in the polyproteins mentioned above, are released by the proteolytic activity of Pro, which releases itself from the Gag-Pro polyprotein. The canonical proteins MA, CA and NC are released through cleavage of Gag polyprotein by Pro [185]. Similar to MMTV, at least one additional protein, designated p15, is encoded between the MA and CA subdomains of HML-2 Gag. The precise functions of this protein are still poorly understood [66]. Almost all HML-2 loci contain dUTPase motifs within the N-terminal region of *pro* [135]. dUTPase and Pro are released through cleavage of Gag-Pro polyprotein by Pro. Mature RT, RnaseH and IN are released through cleavage of Gag-Pro-Pol polyprotein by Pro. A spliced form of the full-length mRNA encodes the Env protein, whose typical domains SP, SU and TM are released by cellular proteases [78].

In addition to transcripts for translation of Gag-Pro-Pol and Env, HML-2 type 2 proviruses can generate an additional mRNA via alternative splicing of the *env* gene. This transcript encodes an accessory protein, named Rec, that appears to be a peculiarity of the HML-2 group [136]. Rec displays similarities to proteins encoded by complex exogenous retroviruses, in particular Rev/Rex proteins of HIV-1 and human T-lymphotropic virus (HTLV) [130], as well as Rem protein of MMTV [83]. In the life cycle of those exogenous retroviruses, that accessory protein facilitates shuttling of unspliced viral transcripts from the nucleus to the cytoplasm. Rec likely was a functional equivalent of those proteins in the exogenous precursor of HML-2 [129]. Due to the 292-bp deletion, type 1 proviruses have lost a SD and are incapable of encoding Env or Rec. Of note, an alternative SD site located upstream of the 292-bp deletion is instead used to splice an mRNA that encodes a 9-kDa protein named Np9 [7,77]. The role of Np9 in HML-2 biology is not known, but both Np9 and Rec may be involved in cell tumorigenesis, as will be described later.
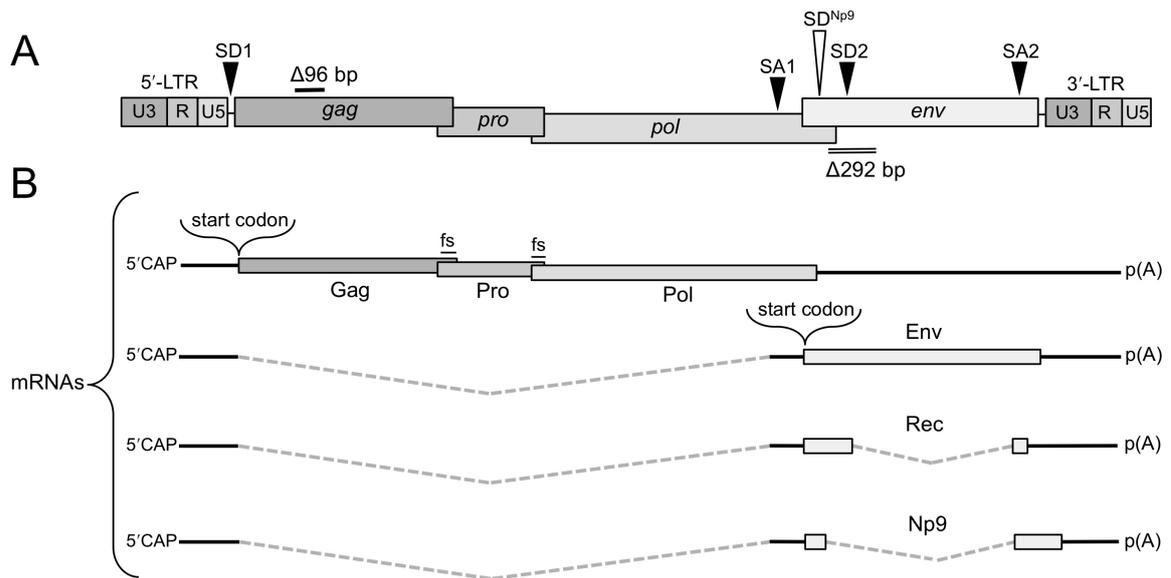
Figure 2: Proviral structure and transcripts of HERV-K(HML-2). A) General structure of an HML-2 full-length provirus with ORFs of the retroviral genes *gag*, *pro*, *pol*, *env* and flanking LTRs. The 96-bp deletion (Δ96 bp) in the *gag* ORF, which differentiates HML-2 from the older HERV-K (OLD) proviruses and the 292-bp deletion (Δ292 bp) at the *pol-env* junction (dotted line), which differentiate type 1 and 2 proviruses (see text), are shown. The position of splice donor (SD) and splice acceptor (SA) sites are shown. **B)** Splice variants of mRNA of the HML-2 proviruses of types 1 and 2. HML-2 transcription starts from the R region. All transcripts are modified with a 5' CAP and a poly-A tail (pA). Introns are represented with dotted lines. A full-length transcript includes three ORFs encoding proteins Gag, Pro and Pol. In this transcript there is only a starting codon at the 5' end of Gag, thus Pro and Pol are translated through ribosomal frameshifts (fs). A second mRNA, where a region is spliced between SD1 and SA1 encodes the protein Env. A third mRNA encodes the Rec protein. That mRNA is the product of an alternative splicing of the *env* mRNA, between SD2 and SA2, and is produced only in proviruses of type 2, which do not have the 292-bp deletion. In type 1 proviruses, SD2 is missing due to the 292-bp deletion but an alternative splice donor site (SD[NP9]) is used immediately upstream of the 292-bp deletion. As a result, the alternative splicing of the *env* mRNA in type 1 proviruses generates a transcript that encodes for Np9 protein.

2.3.3 Transcription of HML-2 in healthy and disease-derived tissues

Although HML-2 transcription can be suppressed by epigenetic mechanisms [116], several studies identified HML-2 transcripts in human tissues. An analysis conducted more than 10 years ago in our laboratory detected 23 transcriptionally active HML-2 proviruses [58]. Additional transcribed proviruses were identified subsequently [15,142]. Of note, HML-2 transcription in healthy tissues was reported by various research groups. In the placenta, an organ where decreased epigenetic regulation has been observed, HML-2 expression appears particularly active [152]. Also, testes and other steroid hormone-regulated tissues appear to be privileged for HML-2 expression [3]. Expression profiles of other healthy tissues, including brain, heart, peripheral blood mononuclear cells, lung, liver and breast, indicate low-level transcription of several HML-2 proviruses [58]. Detection of HERV transcripts in healthy tissues sparks speculation about beneficial roles of HML-2 products [113], although, the few HERVs that seem to be under strong positive selective pressure, do not belong to the HML-2 group [21]. On the contrary, upregulation of HML-2 proviruses has been strongly associated with several disease conditions, notably tumors. Elevated HML-2 transcription was reported for certain types of

germ cell tumor (GCT) including seminoma [76,179], teratocarcinoma [15] and ovarian cancer [215]. HML-2 transcripts have been detected in prostate cancer [72]. HML-2 is upregulated also in melanoma [183]. Concerning melanoma, one specific HML-2 locus that has been found to be transcribed only in melanoma-derived samples but not in melanocytes, might represent a marker for melanoma [183]. HML-2 *env* transcripts are also present in human breast cancer, with significantly higher amounts of transcripts than in normal breast tissue [214]. Expression of HML-2 *env* transcripts was 5–10-fold higher in breast cancer cell lines treated with estradiol and progesterone than in cells without treatment [214]. Moreover, HML-2 RNA was found at very high titers in plasma of patients with lymphomas [42].

Apart from tumor diseases, HML-2 proviruses may play a role in other pathological conditions for which HML-2 upregulation was observed. For example, HML-2 transcripts are increased in brain samples from patients with bipolar-disorder and schizophrenia in comparison with control samples [61]. HML-2 was reported to be upregulated also in neurons of patients with amyotrophic lateral sclerosis (ALS) [51], although such evidence could not be corroborated in subsequent analyses [65,138]. Significant upregulation of HML-2 mRNA levels was observed for rheumatoid arthritis compared to inflammatory and healthy controls [62]. Moreover, HML-2 RNA sequences were found in plasma samples from HIV-1 positive patients but rarely detected in plasma of patients with hepatitis C or control subjects [41].

2.3.4 Expression of HML-2 proteins in disease-derived tissues

In human cancers and cell lines, not only HML-2 transcripts but also HML-2 proteins have been identified. Representative examples will be described in the following. HML-2 Env, Rec and Np9 proteins were detected in primary melanoma and melanoma derived cell lines [29]. Several studies demonstrated the oncogenic potential of such HML-2 proteins. Stable expression of HML-2 Env in non-transformed breast epithelial MCF10A cell line induced tumorigenic processes, resulting in altered cell morphology and increased motility [119]. In HEK293T cells, 24 h post-transfection (hpt) with expression vector encoding HML-2 Env, many cellular genes showed modified expression levels. Among the identified upregulated genes, there was a strong enrichment of transcription factors associated with oncogenesis [119]. Similar conclusions come from investigations of HML-2 Rec and Np9. Both proteins have been reported to interact physically and functionally with nuclear promyelocytic zinc finger (PLZF) protein, a tumor suppressor that has been implicated in leukemogenesis in humans and spermatogenesis in mice [48]. Also, Rat-1 fibroblasts stably expressing HML-2 Rec have undergone cell transformation when injected in immunodeficient mice [19]. Analogous experiments, with injection of cells overexpressing Np9 into mice, showed much faster tumor growth than controls without Np9 [33]. Np9 protein is highly expressed also in leukemia patients

where it may be essential for growth of myeloid and lymphoblastic leukemia cells [33]. When Np9 was expressed *in vitro* or *in vivo* proliferation of leukemia cells was significantly promoted [33]. Moreover, crucial cellular pathways like ERK, AKT, Notch1 and β-catenin, involved in the survival of leukemia stem cells, were activated by Np9 [33].

HML-2 Env protein was found also in brain tissue from patients with ALS, where it could have a role in ALS pathophysiology [138,224]. Expression of Env protein in human neuronal cell cultures caused neurotoxicity and neuronal death [121]. Expression of Env protein in transgenic mouse neurons led to motor dysfunction, DNA damage, morphological and functional changes [121]. Thus, HML-2 could be involved in some neurologic diseases [35], although the contribution of HML-2 elements in neurodegeneration is speculative at the moment and requires further investigations [138].

A novel HML-2 protein derived from Env, called Env-SP, was identified in our laboratory [168]. Env-SP is the signal peptide of HML-2 type 2 Env that, instead of being degraded, is relatively stable and localizes to various cell compartments [168]. It has been hypothesized that, in case of expression, Env-SP may interfere with important cellular processes, potentially being involved in immune evasion of GCT cells or tumorigenesis [168].

As Env-SP, HML-2 dUTPase is another protein that it could be associated with detrimental effects if expressed, although its expression in human cells has not been investigated in detail yet. For example, it was demonstrated that dUTPase has a pro-inflammatory potential, triggering secretion of cytokines and activation of NF-κB and other immune response factors [5]. In this context, HML-2 dUTPase has been proposed as potential contributor to psoriasis, a chronic inflammatory immune disease of the skin [5].

Concerning HML-2 Gag, antibodies against that protein were detected in patients with GCTs, especially seminoma [179]. A larger amount of HML-2 Gag protein is produced in GCT tissues, where HML-2 protease-cleaved Gag was also detected, indicating presence of active HML-2 Pro in GCT tissue [179]. Immature and properly processed HML-2 Gag and Env proteins were observed also in plasma of patients with lymphoma [42].
Moreover, retrovirus-like particles derived from expression of HML-2 were observed in human teratocarcinoma many years ago [20,170]. Such viral particles were identified more recently also in melanoma [28,144], in the plasma of lymphoma patients [42] and in the context of multiple sclerosis [109]. It is known that Gag has the ability to induce the budding of virus-like particles from the cell, even when expressed in the absence of the other virus-encoded proteins. However, isolated retroviral particles harbor reverse transcriptase activity and were shown to contain mature forms of Gag and Env proteins [144], suggesting that apart from Gag, also other functional HML-2 components were

expressed and packaged into viral particles. Up to date, there is no study that demonstrated infectivity of HML-2 encoded particles budding from cells. However, taking into account that some HML-2 elements are polymorphic, it is conceivable that in some individuals HML-2 proviral loci capable of producing infectious particles could exist [10]. It is also known that reversion of a few nucleotide mutations could render a defective provirus replication-competent, as demonstrated by two independent studies [50,118].

In summary, expression of HML-2 proviruses is known to be up-regulated in several disease-derived tissues. In those tissues, not only transcripts but also proteins, and retrovirus-like particles originating from HML-2 proviruses have been observed. HML-2 products may represent initiating factors and/or promoters of pathological conditions. Most evidence associates HML-2 products with tumorigenesis. Actual functional relevance of HML-2 products in disease contexts has not been confirmed with certainty so far. Further specific analyses are required.

2.4 HERV-K(HML-2) encoded protease and integrase

Some HML-2 proteins were not studied in greater detail so far, although they may have crucial roles in human health and disease as well. HML-2 encoded protease (Pro) and integrase (IN) were investigated in the work presented for their potential impact on cellular biology. In the following, typical features of retroviral Pro and IN will be summarized, followed by a description of current knowledge on HML-2 Pro and IN. As will be noted, currently available evidence regarding the impact of retroviral Pro and IN on cell biology derives from studies of exogenous retroviruses, in particular HIV-1.

2.4.1 Retroviral proteases

Retroviral proteases belong to a specific class within the family of aspartic proteases [220] and are functional as homodimers. A monomer is formed by the duplication of four core structural elements: a β-hairpin; a wide loop containing a highly conserved DTG (Asp-Thr-Gly) catalytic triad; a C-terminal α-helix; a second β-hairpin, also called "flap", that is fundamental for the binding and release of substrates [220]. Dimerization of two Pro monomers in retroviral proteases creates a cleft between the two subunits, where substrate is bound and where the catalytic site is located [52]. The two aspartic acids from the DTG triad of each Pro monomer cooperate in presence of a water molecule to perform substrate lysis. Cleavage of substrates does not happen at random amino acid positions. Cleavage sites are recognized with a certain specificity that varies between retroviral proteases. For example, each retroviral protease can cleave with higher specificity cleavage sites in the polyproteins of its own virus [220]. When the substrate is accommodated in the Pro catalytic site, the efficiency of the cleavage derives from more or less favorable interactions between side chains of substrate and corresponding

subsites located in the cleft of Pro. By convention, the peptide bond that is cleaved is referred to as the scissile bond, with flanking amino acids in N-terminal direction referred to as P1, P2, P3, etc. Amino acids flanking in C-terminal direction are referred to as P1′, P2′, P3′, etc. Accordingly, the scissile bond is located between P1 and P1′ [200]. The properties of amino acids in P1 and P1′ positions appear to be major determinants of Pro specificity. A study involving proteases from 10 different retroviruses reported that approximately 80% of all known Pro cleavage sites can be classified into two groups: type 1 sites have an aromatic residue in P1 (favouring phenylalanine or tyrosine) and prolin in the P1′ position; while type 2 sites favor leucine over phenylalanine or tyrosine in P1 and have a preference for valin, leucin or alanin in P1′ [154]. Amino acids with side chains branched at the β-carbon are excluded from the P1 position of both types.

In several retroviruses, Pro is translated as part of a Gag-Pro polyprotein from which Pro is released by self-processing. The precise mechanism of the initial steps that release the Pro domain from the Gag-Pro precursor was studied for HIV-1 Pro [154]. It appears that through dimerization of two Gag-Pro precursors the Pro domain embedded in the precursor can cause intra- or intermolecular release of free Pro [154]. The binding properties of the Gag domain enable interactions of precursor proteins and formation of viral particles through a process called particle assembly. Independent of the cellular compartment in which the assembly takes place (in the cell membrane for some retroviruses and in the cytoplasm for others) activation of Pro is known to typically occur during or after the budding of new retroviral particles from the host cellular membrane. During this phase, Pro activity induces a characteristic change of morphology in the virion that is called maturation [200]. However, several studies of retroviruses from different genera documented Pro activity also in the cytoplasm, thus not just during maturation of retroviral particles. For example, premature activation of Pro has been observed during acute HIV-1 infection, resulting in accumulation of processed HIV-1 Gag that was excluded from the assembly pathway [96]. In some HIV strains, up to 50% of HIV-1 Pro activity is not associated with virions due to low packaging efficiency [96]. Intracellular activation of HIV-Pro and intracellular processing of the HIV-1 Gag-Pol polyprotein was also observed when overexpressing Gag-Pol in cells [97]. For MMTV, that is phylogenetically closely related to HERV-K(HML-2), intracellular processing of MMTV Gag was also confirmed, with evident accumulation of unprocessed Gag polyprotein in presence of inactive Pro [223].

2.4.2 Cellular proteins as substrates of retroviral proteases

In addition to catalyzing maturation of viral proteins, retroviral proteases may also process functionally relevant cellular proteins. Cleavage of such proteins may have severe consequences for cell biology, as demonstrated by several pubblications. A number of proteins associated with the cytoskeleton can be cleaved by HIV-1 Pro. Among those, cleavage of actin, vimentin, spectrin, desmin, myosin, tropomyosin, glial fibrillary acidic protein [187–189] and microtubule associated protein-2 [213] was demonstrated more than 20 years ago *in vitro*. Cleavage of actin and vimentin was also observed *in vivo* [79,204]. In the case of vimentin, the protein cleavage was accompanied by altered cytoskeletal architecture and nuclear morphology [79]. Also, it was observed that fibroblasts microinjected with HIV-1 Pro assume a rounded form, detached from the surface of the cell culture flask, or were dying [79]. Using the same approach, it was reported cleavage of various proteins comprising the focal adhesion plaque [190], a multi-protein structure necessary for interaction between cell and extracellular matrix, but also involved in signal transduction, cell proliferation and regulation of gene expression. Processing of vimentin was also reported for proteases of bovine leukemia virus, Mason–Pfizer monkey virus, and myeloblastosis-associated virus [194].

Wagner et al. (2015) [212] reported cleavage of serine-threonine kinases RIPK1 and RIPK2, both involved in innate immune response, by HIV-1 Pro, followed by reduced NF-κB activation [212]. NF-κb is a protein complex that controls transcription, cytokine production and cell survival, whose regulation is critical during cellular stress. Interestingly, Rivière et al. (1991) [165] observed processing of the precursor of NF-κB during HIV-1 acute infection, with an increased level of active NF-κB in the nucleus [165]. Other important cellular factors processed by HIV-1 Pro are pro-interleukin 1β [204] and serine-threonine kinases NDR1 and NDR2 [49].

Among the restriction factors that are part of the cellular antiviral defense, cleavage of APOBEC3 by the proteases of murine retrovirus and feline immunodeficiency virus was observed [222]. Proteases from several members of the family *Retroviridae*, including MMTV, are able to cleave translation initiation factor eIF4GI, severely inhibiting cap-dependent translation [1]. Translation initiation factor eIF3d is also known to be targeted by HIV-1 Pro [87].

Of further note, a recent investigation reported more than 120 cellular substrates processed by HIV-1 Pro *in vitro*, including proteins having functions with biological relevance [82]. In addition, several lines of evidence suggest that retroviral proteases may process cellular factors that can induce cell death. Nie et al. (2002) [146] demonstrated that HIV-1 Pro directly cleaves and activates procaspase 8 in T-cells, followed by induction of the molecular events leading to apoptosis. Rumlová et al. (2014) [169] observed HIV-1 Pro localizing to mitochondria and demonstrated that HIV-1 Pro mediates cleavage of mitochondrial proteins Tom22, VDAC and ANT, leading to release of apoptosis-inducing factors. Blanco et al. (2003) [18] expressed HIV-1 Pro in yeast, observing drastic alteration of

membrane integrity followed by cell lysis. Similarly, HIV-1 Pro expression in mammalian COS7 cells increased plasma membrane permeability and cell lysis by necrotic processes [18].

Altogether, quite a number of host cellular proteins are susceptible to proteolysis by retroviral proteases. Proteins identified are diverse, locate to various cellular compartments and participate in various, often important cellular processes. Among those processes are, for instance, fundamental processes performed by the cytoskeleton (intracellular transport, chromosome segregation, cell signaling etc.), but also cell adhesion, innate immune response, cell signaling, translation initiation, apoptosis. For several proteins, e.g. actin, vimentin, RIPK1, RIPK2, NF-κB, NDR1, NDR2, eIF4GI, processing by retroviral protease was documented through experiments conducted *in vivo*, demonstrating that observed proteolytic events can take place in the cellular environment. There is evidence of detrimental effects on cell biology as a consequence of proteolytic activity of retroviral proteases, e.g, perturbation of the cytoskeletal organization along with alteration of cell morphology, altered levels of cellular regulatory factors, translation inhibition, induction of apoptosis, and cell lysis.

2.4.3 HML-2 protease

Sequence analysis and biochemical characterization of HML-2 Pro confirmed pronounced similarities with retroviral aspartic proteases [205]. Typical retroviral protease motifs are present in the HML-2 Pro ORF, including the DTG motif, the flap motif and the GRDLL motif (corresponding to the C-terminal α-helix) [207]. As in Betaretroviruses, the HML-2 Pro gene is located in a separate ORF [148]. Expression of the full-length HML-2 Pro ORF in bacteria resulted in self-processing of the translation product with accumulation of a ~18 kDa mature enzyme composed of 156 amino acids. The HML-2 Pro core domain, corresponding to the 106 N-terminal amino acids of the mature enzyme, was defined on the basis of sequence similarity with HIV-1 Pro. That region is 28% identical with HIV-1 Pro on the amino acid level [205]. No similarity with HIV-1 Pro was observed for the C-terminal 50 amino acids, although such a C-terminal extension is characteristic for proteases of betaretroviruses. HML-2 Pro is translated as a Gag-Pro polyprotein that self-processes and further processes Gag subdomains [143]. The cleavage sites of HML-2 Pro within HML-2 Gag were recently mapped by two different research groups [66,111]. For all cleavage sites identified, but one site, the amino acid in P1 was hydrophobic and unbranched at the β-carbon, whereas proline, tyrosine and valine were found in P1′. The HML-2 Pro N-terminal self-processing site was also identified [205]. Self-processing was demonstrated to occur at the aa sequence KAAY-WASQ, thus with a tyrosine in P1 and a Tryptophan in P1′. The C-terminal cleavage site has not been precisely mapped as yet, although initial studies subjecting a purified HML-2 Pro to prolonged incubation observed some limited cleavage in the region corresponding to the C-terminus of the Pro ORF [149,205]. In a recent publication, a profile of HML-2 Pro cleavage site specificity was obtained from analysis of 95 cleavage sites identified through

incubation of a tryptic peptide library with purified HML-2 Pro. That study revealed for HML-2 Pro a preference for aromatic amino acids in P1 (thus suggesting a preference for retroviral Pro cleavage sites of type 1, see above), although a considerable proportion (21%) of cleavage sites had glycine in P1. Aromatic and aliphatic amino acid residues were found for P1′, and acidic residues (aspartic acid or glutamic acid) were found in 32% of cleavage sites for P2 [17].

Concerning cleavage specificity, it was also demonstrated that HML-2 Pro is able to process HIV-1 Gag at the authentic MA/CA cleavage site of HIV-1 Pro [205]. That finding led to the hypothesis that HML-2 Pro may complement processing of HIV-1 poplyprotein during HIV-1 infection [205], a finding of potential relevance in view of the fact that HML-2 Pro was found insensitive to various HIV-1 Pro inhibitors used in treatment of HIV-infected patients [149].

Several HML-2 proviruses harboring an intact Pro ORF were identified in the human genome. Apart from the abovementioned HERV-K10, another example relevant for experiments performed in the present studies is a provirus discovered by our research group and originally termed HERV-K(HML-2.HOM) [134], subsequently reported by other studies with alternative designations HERV-K108, HERV-K(C7) or ERVK-6, the latter being the official locus name [64]. Located in human chromosome 7p22.1, HERV-K(HML-2.HOM) harbors full-length ORFs for all retroviral proteins, with an inactivating mutation only within the reverse transcriptase coding region. The proviral allele originally investigated was present as a tandem arrangement consisting of two proviruses and three LTRs, with the central LTR shared by both proviruses [161]. Extensive analysis identified allelic variants of HERV-K(HML-2.HOM), for example, alleles consisting of only a single provirus or a solo LTR, alleles with mutations within Gag or Pro domains and alleles with an intact reverse transcriptase [161,199]. Other HML-2 loci able to encode an intact Pro are known [118], and the majority of those loci also retain an intact *gag* gene, that is necessary for translation of Pro (see chapter 2.4.1). Potentially relevant variations in the Pro sequence have been found for different HML-2 proviruses that may affect interactions with substrates [207].

There is evidence of transcription of Pro-encoding HML-2 loci in various human tissues (for instance, see [58,183]). Moreover, considerable levels of HML-2 Gag protein are detectable in GCT tissues, where products of Gag processing, expected if processed by active HML-2 Pro, were identified [179]. In GCT cell lines, HML-2-encoded retroviral particles budding from the cellular membrane can also be identified [20,170]. Of note, HML-2-encoded retroviral particles with a morphology that suggested a maturation process due to Pro activity have been observed [16]. Thus, in those tissues and derived cell lines there is strong evidence of presence of enzymatically active HML-2 Pro. Moreover, bacterial expression of constructs harboring HML-2 Gag-Pro ORFs showed self-processing of Pro from the Gag-Pro precursor and processing of Gag subdomains [134,143], confirming functionality of HML-2 Pro.

2.4.4 Retroviral integrases

As their name suggests, retroviral integrases (INs) are enzymes that catalyze the integration of the retroviral dsDNA genome into the genome of the host cell. INs are commonly composed of three conserved domains: The N-terminal domain (NTD), the catalytic core domain (CCD), and the C-terminal domain (CTD). The NTD contains a highly conserved HHCC motif, composed of two histidine (H) and cysteine (C) residues that, via binding of Zn^{2+}, fold the NTD in a compact three-helical bundle [120]. The CCD domain contains a DDE motif, composed of two highly conserved aspartic acid (D) residues and a glutamic acid (E) residue that together constitute a catalytic triad. The two aspartic acids coordinate a cationic cofactor, usually Mg^{2+}, necessary for IN catalytic activity. The CTD domain contains conserved tryptophan residues and is involved functionally in the interaction with DNA, stabilizing the protein-DNA complex [120].

IN is usually produced as part of the Gag-Pro-Pol polypeptide precursor, then released by Pro activity. In the cytoplasm, IN proteins assemble on viral dsDNA ends forming a stable nucleoprotein complex referred to as the intasome [120]. IN is catalytically active as a multimeric protein. The number of integrase molecules required to form the intasome differs between viral genera, for example, four subunits are typical for spumaviruses, whereas lentiviral IN comprises up to sixteen subunits [53]. Betaretroviruses employ eight IN subunits that create an octameric architecture, specifically a tetramer-of-dimers [53]. Apart from IN and viral DNA, the intasome also includes host cellular proteins that have various functions, for example, identification of a suitable integration site or protection from proteolysis [67].Two of the IN molecules enclosed in the intasome carry out the integration process by virtue of two distinct reactions: 3′ processing and strand transfer [46]. The first reaction occurs in the cytoplasm during migration of the intasome toward the cellular nucleus. 3′ processing results in the removal of two (or three) nucleotides from each 3′ end of the retroviral dsDNA. Such cleavage is specific and occurs downstream of a conserved CA dinucleotide present in retroviral LTRs. A nucleophilic attack by the OH group of a water molecule is involved, although other molecules may also provide the nucleophile necessary for this process [193]. After 3′ processing, the retroviral dsDNA presents recessed 3′-OH ends exposing highly reactive hydroxyl groups. The strand transfer takes place when the intasome, having entered the cell nucleus, approaches and interacts with the host DNA. At that stage, IN coordinates a nucleophilic attack by the retroviral free 3′ hydroxyl groups to a phosphodiester bond of the host DNA. This reaction occurs at both ends of the viral DNA molecule with an offset of some base pairs between the two points of insertion. As a consequence of integrase-mediated joining, the two ends of viral DNA are ligated to the target DNA, however portions of single-stranded DNA (ssDNA) remain between host DNA and provirus. After the strand transfer, the cellular DNA repair machinery is necessary to restore genome integrity, thus complete the integration [120].

3′ processing and strand transfer have been modelled *in vitro* using short oligonucleotide substrates designed to mimic the retroviral DNA LTRs [100]. Experiments *in vitro* demonstrated that IN can also

exhibit two other activities: 1) disintegration, which may be considered as the strand transfer reaction in reverse [34]; 2) nonspecific alcoholysis, that takes place through nucleophilic attack of the DNA phosphodiester bonds by the OH group from a variety of molecules, including small alcohol molecules [99–101,193]. Through the latter activity, IN can potentially induce nicking of any DNA target sequence at almost any internal site.

2.4.5 Retroviral integrases as cause of DNA damage

Disruption of genome integrity by new provirus integrations, temporary formation of ssDNA gaps during the integration process and DNA lesions (potentially) induced by nonspecific alcoholysis are processes by which an IN could potentially induce genomic instability of the host cell. Each time that the stable structure of a dsDNA is altered and an ssDNA region becomes exposed, there is a hotspot for the formation of a DNA double-strand break (DSB), a serious damage in a cell genome [176]. DSBs impact cellular processes drastically, for instance, inducing transcriptional activation of genes, post-transcriptional modifications, genetic recombination, cell-cycle arrest, and in some cases, programmed cell death [44,103]. DSBs are typically repaired by the cellular repair system in order to prevent cell health, but gene translocations, rearrangements, amplifications, and deletions arising during repair and misrepair of DSBs may contribute to cell transformation and tumor development [210].

There is evidence to believe that retroviral IN can induce DNA damage in the cell genome, potentially resulting in formation of DSBs. It was observed that cells respond to retroviral DNA integration by activating the repair machinery in a manner similar to the DSB response [176], thus suggesting formation of DSBs due to presence of IN activity. A decrease in retroviral infectivity has been observed in cells lacking DSB repair enzymes [44], suggesting that during retroviral integration there is formation of DSBs that need to be repaired for efficient completion of the retroviral DNA integration. Although the biological role of IN nonspecific alcoholysis (see above) is unknown, it is conceivable that, if occurring in cells, such uncontrolled activity may cause DNA damage in the cell genome. For instance, in a previous study, expression of HIV-1 IN in yeast cells led to a lethal phenotype that was not induced in presence of inactive HIV-1 IN mutated in the DDE motif [151]. Considering that there were no HIV-1 LTRs present in yeast, it has been hypothesized that cell death in that case was due to nonspecific alcoholysis, causing cleavage of genomic DNA and cell death. It appears that IN activity can also interfere with DNA recombination both *in vitro* and *in vivo* [151], thus providing further support to the notion that retroviral IN may be a source of genomic instability.

2.4.6 HML-2 integrase

An HML-2 IN domain was originally identified in the HERV-K10 provirus by sequence similarity with Rous sarcoma virus (RSV) IN [105]. Specifically, HML-2 IN is encoded by the 3´ region of the *pol* gene, where a sequence corresponding to a 288 aa protein, was demonstrated to suffice for IN functionality [105]. Protein modelling indicated for HML-2 IN the typical domain architecture of betaretroviral IN [25]. Several HML-2 loci potentially able to encode a full-length IN have been identified in the human genome. Apart from proviruses HERV-K10 and HERV-K(HML-2.HOM) mentioned above, at least another 20 HERV-K(HML-2) proviruses containing a full-length IN were identified [25]. Among HML-2 proviruses with full-length IN, 9 displayed intact active sites motifs and no known inactivating mutations [25]. Those 9 loci have been shown to be transcriptionally active [15,58]. Functionality of HML-2 IN was demonstrated through a series of experiments conducted *in vitro,* using short substrates mimicking retroviral LTR ends. In those experiments, HML-2 IN could efficiently perform 3' processing and strand transfer on HML-2 LTR substrates, but it was also active on LTR substrates derived from HIV-1 and RSV [105]. The relaxed substrate specificity of HML-2 IN appeared to be a peculiar characteristic of this enzyme not typical for other retroviral INs. For example, HIV-1 IN and RSV IN could act only on their own viral LTRs [105]. It is conceivable that HML-2 IN can also perform nonspecific alcoholysis, although never investigated so far. In view of the numerous human loci containing an intact HML-2 IN, and in view of the HML-2 IN activities confirmed *in vitro*, it is conceivable that, if expressed, HML-2 IN could have the potential to induce DNA damage in the cell genome.

2.5 Aims of the study

Although most of HERVs have been severely mutated during evolution, the HML-2 group includes proviruses that still encode intact proteins. Some HML-2 proteins exhibit activities that could potentially damage the cell, with consequential activation of pathological processes. HML-2 upregulation and HML-2 proteins have been documented in various human diseases whose causes are not fully understood. HML-2 proteins may play a role in the development of those diseases. Thus, investigation into HML-2 proteins could be the key to clarify the biological mechanisms behind some human conditions.

In the present study, we focused on HML-2 protease (Pro) and HML-2 integrase (IN), two enzymes that have not been extensively investigated concerning biological and clinical significance.

Regarding HML-2 Pro, it was previously confirmed through experiments *in vitro* and *in vivo* that the enzyme is functional against HML-2 Gag, a protein specifically processed by HML-2 Pro in the context of the retroviral replication cycle. Processing products of HML-2 Gag protein were observed in germ cell tumors and derived cell lines, indicating presence of an endogenous active HML-2 Pro in those cells. Considering evidence of active HML-2 Pro in human cells, it was interesting to know whether, apart from targeting its specific retroviral substrates, HML-2 Pro activity could also target cellular proteins. As for the latter point, there is well documented evidence for some other retroviral proteases, among them HIV-1 Pro. Through cleavage of cellular proteins, HML-2 Pro could potentially affect cellular processes leading to pathological cell states. The ability of HML-2 Pro to process cellular proteins was not investigated so far.

Particular investigations of HML-2 Pro comprised the main part of the study. The major aims were to establish whether HML-2 Pro is able to process cellular proteins and to identify cellular proteins specifically cleaved by this enzyme. We also aimed at producing initial indications of how cells respond to HML-2 Pro expression, thus providing initial insights into a functional impact of HML-2 Pro on cell biology and possible relevance for human diseases. To achieve major aims, we accomplished various sub-aims. We aimed at purifying HML-2 Pro by applying a procedure suited to preserve enzymatic functionality. We also aimed at purifying a mutated, enzymatically inactive HML-2 Pro. Employing a specifically designed fluorescence *in vitro* assay, we aimed at characterizing the enzymatic activity of HML-2 Pro biochemically and optimizing reaction conditions for subsequent experiments. Incubating purified HML-2 Pro with native human proteins *in vitro* and subsequently subjecting reactions to a sophisticated proteomics technique called terminal amine isotope labeling of substrates (TAILS), we aimed at identifying cellular proteins cleaved by HML-2 Pro *in vitro* and producing a list of human proteins that could be considered as substrates of HML-2 Pro. We also aimed

at validating selected cellular substrates of HML-2 Pro in more specific experiments suited to verify processing of cellular proteins *in vitro* and *in vivo*. We aimed at gathering indications of cellular processes and molecular pathways that might become affected by HML-2 Pro activity. We therefore compiled for cellular substrates of HML-2 Pro information on cellular localization, cellular processes, and involvement in diseases, employing suited bioinformatics tools.

Regarding HML-2 IN, it was previously demonstrated through experiments *in vitro* that HML-2 IN can interact with DNA and perform catalytic reactions that disrupt DNA integrity. If expressed in cells, HML-2 IN might interact with the host genomic DNA and induce DNA damage leading to DNA double-strand breaks (DSBs). As an endogenous potential genotoxic agent, HML-2 IN could represent a contributor to genomic instability, a hallmark of various human diseases including cancer. The notion that HML-2 IN might be a cause of genomic instability is intriguing, but there is no experimental evidence in support of this hypothesis. There is even lack of knowledge as to whether HML-2 IN localizes in the cell nucleus, which is prerequisite for speculations on an impact of HML-2 IN at level of the cell genome.

The second part of our study therefore concerned HML-2 IN. We aimed at assessing experimentally whether HLM-2 IN causes DNA damage. In particular, we performed immunofluorescence experiments targeting transiently expressed HA-tagged HML-2 IN and a biomarker for DSBs thereby evaluating whether HML-2 IN could localize to the cell nucleus and induce DSBs in the cell genome.

3. Materials and methods

3.1 Materials

3.1.1 Chemicals

Chemical	Manufacturer
1,4-Dithiothreitol (DTT)	Carl Roth (Karlsruhe, Germany)
5-Brom-4-chlor-3-indoxyl-β-Dgalactopyranosid (X-Gal)	AppliChem (Darmstadt, Germany)
Acetic Acid Glacial ≥ 99.7%	VWR (Radnor, USA)
Acrylamid/Bisacrylamid (30%) (37,5:1)	Bio-Rad (Hercules, USA)
Agar, Bacto	BD (Franklin Lakes, USA)
Agarose, Ultra Pure	Invitrogen/Life Technologies (Carlsbad, USA)
Ammonium Peroxodisulfate (APS)	Carl Roth (Karlsruhe, Germany)
Ampicillin-Natriumsalt	Carl Roth (Karlsruhe, Germany)
Anthranilyl-HIV Protease Substrate Trifluoroacetate salt	Bachem (Bubendorf, Switzerland)
Bis(2-hydroxyethyl)amino-tris(hydroxymethyl) methan (Bis-Tris)	Carl Roth (Karlsruhe, Germany)
Bovine Serum Albumin (BSA)	Sigma-Aldrich (St. Louis, USA)
Brilliant Blue R250	Carl Roth (Karlsruhe, Germany)
Bromophenol Blue	Carl Roth (Karlsruhe, Germany)
cOmplete EDTA-free protease inhibitor cocktail	Sigma-Aldrich (St. Louis, USA)
Deoxynucleoside Triphosphate Set (dNTPs)	Roche (Basel, Schweiz)
Dimethyl sulfoxide (DMSO)	Carl Roth (Karlsruhe, Germany)
Dulbecco's Phosphate Buffered Saline (DPBS)	Gibco/Life Technologies (Carlsbad, USA)
Ethidium Bromide	Carl Roth (Karlsruhe, Germany)
Ethylene Glycol	Carl Roth (Karlsruhe, Germany)
Ethylenediaminetetraacetic acid (EDTA)	Carl Roth (Karlsruhe, Germany)
Formaldehyde Solution 37%	Carl Roth (Karlsruhe, Germany)
Glycerin	Zentrales Chemikalienlager, UdS, Saarbrücken
Indinavir Sulfate (orb321983)	Biozol (Eching, Germany)
Isopropyl-1-thio-β-D-galactopyranoside (IPTG)	Carbolution Chemicals GmbH (St. Ingbert, Germany)
L-^{35}S-methionine (370 MBq, 10mCi/ml)	Hartmann Analytic (Braunschweig, Germany)
Kanamycin	Sigma-Aldrich (St. Louis, USA)
Methanol	Zentrales Chemikalienlager, UdS, Saarbrücken
N,N,N',N'-Tetramethylethylendiamin (TEMED)	Sigma-Aldrich (St. Louis, USA)
NuPAGE™ 4x LDS Sample Buffer	Thermo Fisher (Waltham, USA)
NuPAGE™ Antioxidant	Thermo Fisher (Waltham, USA)
NuPAGE™ MES SDS Running Buffer	Thermo Fisher (Waltham, USA)
NuPAGE™ MOPS SDS Running Buffer	Thermo Fisher (Waltham, USA)
NuPAGE™ Transfer Buffer	Thermo Fisher (Waltham, USA)
Opti-MEM™ Reduced Serum Medium (31985062)	Gibco/Life Technologies (Carlsbad, USA)
Pancaspase Inhibitor Q-VD-Oph hydrate	Sigma-Aldrich (St. Louis, USA)

Pepstatin A	Calbiochem Merck/Millipore (Burlington, USA)
Pepstatin A−Agarose	Sigma-Aldrich (St. Louis, USA)
Piperazine-N,N′-bis(2-ethanesulfonic acid) (PIPES)	Carl Roth (Karlsruhe, Germany)
Poly-L-Lysine Hydrobromide	Sigma-Aldrich (St. Louis, USA)
Ponceau S P3504	Sigma-Aldrich (St. Louis, USA)
Potassium Sodium Tartrate Tetrahydrate	Carl Roth (Karlsruhe, Germany)
Powdered Milk Blotting Grade	Carl Roth (Karlsruhe, Germany)
Rotiphorese NF-Urea	Carl Roth (Karlsruhe, Germany)
Sodium Chloride (NaCl)	VWR (Radnor, USA)
Sodium Deoxycholate	Sigma-Aldrich (St. Louis, USA)
Sodium Dodecyl Sulfate (SDS)	Carl Roth (Karlsruhe, Germany)
Sodium Hydroxide (NaOH)	Zentrales Chemikalienlager, UdS, Saarbrücken
Tris(hydroxymethyl)-Aminomethan (Tris)	Carl Roth (Karlsruhe, Germany)
Triton X-100	Carl Roth (Karlsruhe, Germany)
Trypan Blue Solution	Sigma-Aldrich (St. Louis, USA)
Trypton, Bacto	BD (Franklin Lakes, USA)
Tween®-20	Carl Roth (Karlsruhe, Germany)
Vectashield® Mounting Medium with DAPI	Vector Laboratories (Burlingame, USA)
Yeast Extract, Bacto	BD (Franklin Lakes, USA)

3.1.2 Buffers

Buffers for agarose gel electrophoresis

1x TAE Running Buffer

40 mM	Tris-HCl, pH 8.0
1 mM	EDTA
20 mM	Glacial acetic acid

10x DNA Loading Buffer

10% [v/v]	10x TAE Running Buffer
70% [v/v]	Glycerol
20% [v/v]	20 mM EDTA, pH 8.0
0.2 g	Bromophenol blue

Buffers for protein recovery from inclusion bodies

5x TE Buffer

| 0,1M | Tris-HCl pH 7.5 |
| 5 mM | EDTA |

Solving Buffer

0,1M	Tris-HCl pH 7.5
8 M	Urea
1 mM	DTT

Dialysis Buffer

20 mM	PIPES, pH 6.5
1 M	NaCl
1 mM	DTT

Buffers for protease purification by affinity chromatography

Buffer A

50 mM	PIPES, pH 6.5
1 M	NaCl
1 mM	EDTA
1 mM	NaK tartrate
10% [v/v]	Glycerol

Buffer B

0.1 M	Tris-HCl, pH 8.0
1 mM	NaK tartrate
10% [v/v]	Glycerol
5% [v/v]	Ethylene glycol

Buffers for Western blot

10x TBS (Tris-Buffered Saline)
500 mM	Tris-HCl, pH 7.5
1.5 M	NaCl

TBS-T
1x	TBS
0.05% [v/v]	Tween 20

Blocking Buffer
1x	TBS
5% [w/v]	Nonfat dry milk

Buffers for protein staining

Coomassie Staining Solution
0.2% [w/v]	Coomassie Brilliant Blue R250
50% [v/v]	Methanol
10% [v/v]	Acetic acid

Coomassie Destaining Solution
50% [v/v]	Methanol
10% [v/v]	Acetic acid

Ponceau S Solution
0.1% [w/v]	Ponceau S
10% [v/v]	Acetic acid

Buffer for lysis of mammalian cells under denaturing conditions

RIPA Buffer
50 mM	Tris-HCl, pH 8.0
150mM	NaCl
1% [v/v]	NP-40
0.5% [w/v]	Sodium deoxycholate
0.1% [w/v]	SDS

3.1.3 Enzymes

Enzyme	Buffer	Manufacturer
*Bam*HI-HF® (20000 U/ml)	CutSmart® Buffer	NEB (Ipswich, USA)
*Bgl*II (10000 U/ml)	NEBuffer™ 3.1	NEB (Ipswich, USA)
*Eco*RI-HF® (20000 U/ml)	CutSmart® Buffer	NEB (Ipswich, USA)
*Hind*III (10000 U/ml)	NEBuffer™ 2.1	NEB (Ipswich, USA)
*Msc*I (10000 U/ml)	CutSmart® Buffer	NEB (Ipswich, USA)
*Nhe*I-HF® (20000 U/ml)	CutSmart® Buffer	NEB (Ipswich, USA)
*Not*I-HF® (20000 U/ml)	CutSmart® Buffer	NEB (Ipswich, USA)
Phusion® High-Fidelity DNA Polymerase (2000 U/ml)	5x Phusion® HF Buffer	NEB (Ipswich, USA)
*Pst*I (10000 U/ml)	NEBuffer™ 3.1	NEB (Ipswich, USA)
rAPid Alkaline Phosphatase (1000 U/ml)	rAPid Alkaline Phosphatase Buffer (10x)	Merck/Sigma-Aldrich (St. Louis, USA)
T4 DNA Ligase (400000 U/ml)	T4 DNA Ligase Buffer (10x)	NEB (Ipswich, USA)
T4 Polynucleotide Kinase (10000 U/ml)	T4 DNA Ligase Buffer (10x)	NEB (Ipswich, USA)
Taq DNA Polymerase (5000 U/ml)	10x PCR Buffer with $MgCl_2$	Merck/Sigma-Aldrich (St. Louis, USA)
Trypsin-EDTA		Gibco/Life Technologies (Carlsbad, USA)
*Xho*I (10000 U/ml)	NEBuffer™ 3.1	NEB (Ipswich, USA)

3.1.4 Molecular biology kits

Kit name	Manufacturer
FuGENE® HD Transfection Reagent	Promega (Madison, USA)
NucleoBond® PC 100	Macherey-Nagel (Düren, Germany)
NucleoSpin® Gel and PCR Clean-up	Macherey-Nagel (Düren, Germany)
peqGOLD Plasmid Miniprep Kit I	Peqlab (Erlangen, Germany)
pGEM®-T Easy Vector System	Promega (Madison, USA)
SignalFire™ Elite ECL Reagent	Cell Signaling Technology (Zweigniederlassung, Germany)
TNT® T7 Quick for PCR DNA	Promega (Madison, USA)

3.1.5 Antibodies

Primary antibodies *

Antibody	Host / Clonality	Dilution	Manufacturer
α-53BP1 Alexa Fluor 488 (NB100-304AF488)	Rabbit / Polyclonal	IF 1:800	Novus Biologicals USA (Littleton, USA)
α-53BP1 (NB100-304)	Rabbit / Polyclonal	IF 1:5000	Dr. Nadine Schuler (Klinik für Strahlentherapie und Radioonkologie. Homburg, Germany)
α-HML-2 Gag n. 2548	Rabbit / Polyclonal	WB 1:800	Mueller-Lantzsch et al. 1993 [143]
α-GFP	Rabbit / Polyclonal	WB 1:1000	Prof. Dr. Schlenstedt Gabriel (Homburg, Germany)
α-HA (clone 3F10)	Rat / Monoclonal	WB 1:5000 IF 1:800	Prof. Dr. Friedrich Grässer (Institute of Virology, University of Saarland. Homburg, Germany)
α-Histone H3 (D1H2)	Rabbit / Monoclonal	WB 1:2000	Cell Signaling Technology (Zweigniederlassung, Germany)
α-HML-2 Pro n. 9367	Rabbit / Polyclonal	WB 1:1000 IF 1:1000	Schommer et al. (1996) [185]
α-Ki67/MKI67	Mouse / Monoclonal	IF 1:1000	Novus Biologicals USA (Littleton, USA)
Apoptosis Western Blot Cocktail			Abcam (Cambridge, UK)
250x primary antibodies cocktail containing:		WB 1:250	
α-Actin	Rabbit / Monoclonal		
α-Pro/p17-caspase-3	Rabbit / Monoclonal		
α-Cleaved PARP1	Mouse / Monoclonal		

* Antibody dilutions employed are reported for Western blot (WB) and/or immunofluorescence (IF)

Antibody	Host / Clonality	Dilution	Manufacturer
α-Mouse IgG (A9044)	Rabbit/Polyclonal	WB 1:5000	Merck/Sigma-Aldrich (St. Louis, USA)
α-Rat IgG (A5795)	Rabbit / Polyclonal	WB 1:5000	Prof. Friedrich Grässer (Institute of Virology, University of Saarland)
α-Rabbit IgG Alexa Fluor 488	Goat / Polyclonal	IF 1:1000	Invitrogen/Fisher Scientific (Waltham, USA)
α-Rat IgG Alexa Fluor 594	Goat / Polyclonal	IF 1:1000	Invitrogen/Fisher Scientific (Waltham, USA)
α-Mouse IgG Alexa Fluor 647	Goat / Polyclonal	IF 1:1000	Invitrogen/Fisher Scientific (Waltham, USA)
Apoptosis Western Blot Cocktail 100x secondary antibodies cocktail containing:		WB 1:100	Abcam (Cambridge, UK)
α-Mouse IgG	Goat/Polyclonal		
α-Rabbit IgG	Rabbit/Monoclonal		

* Antibody dilutions employed are reported for Western blot (WB) or immunofluorescence (IF)

3.1.6 Molecular standards for agarose and polyacrylamide gels

Lengths of DNA fragments in agarose gel were estimated employing the 2-Log DNA Ladder (0.1-10.0 kb) (N3200S, New England Biolabs, Ipswich, USA). 5-10 µl per lane of 2-Log DNA Ladder were loaded onto at least one lane per agarose gel. Molecular weight of proteins in SDS-PAGE and Western Blot was estimated employing the Precision Plus Protein Dual Color Standards (1610374, Bio-rad, Hercules, USA) (Figure 3). 5-10 µL of protein standard per lane were loaded onto at least one lane per polyacrylamide gel.
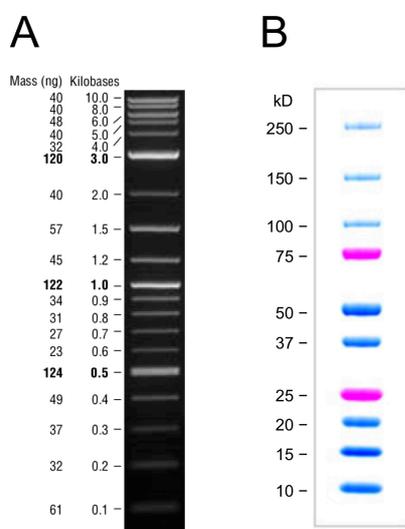
Figure 3: Molecular weight standards for agarose and polyacrylamide gel electrophoresis. A) 2-Log DNA Ladder (0.1-10.0 kb) (N3200S, New England Biolabs, Ipswich, USA). **B)** Precision Plus Protein Dual Color Standards (1610374, Bio-rad, Hercules, USA).

3.1.7 Culture media and cell lines

Prokaryotic cell culture media

LB-Medium		LB-Agar for plates
1% [w/v]	Trypton	1.5% [w/v] Agar
0.5% [w/v]	Yeast extract	in LB-Medium
170 mM	NaCl	
pH 7.0		

Prokaryotic cell culture media were subjected to autoclave sterilization. LB-Agar was cooled down to approximately 50°C and poured in sterile Petri dishes immediately after sterilization of the LB-Agar solution. LB-Medium and LB-Agar plates were stored at 4°C.

Antibiotic standard concentrations

Ampicillin	Kanamycin
100 μg/ml	50 μg/ml

LB-Medium and LB-Agar plates supplemented with ampicillin at standard concentration are referred to in the text as LB_{amp} and $LB\text{-}Agar_{amp}$ plate, respectively.

Mammalian cell culture media

Medium	Manufacturer
Dulbecco's Modified Eagle's Medium (DMEM) (41966029)	Gibco/Life Technologies (Carlsbad, USA)
RPMI-1640 (21875034)	Gibco/Life Technologies (Carlsbad, USA)
Opti-MEM™ Reduced Serum Medium (31985062)	Gibco/Life Technologies (Carlsbad, USA)

DMEM and RPMI media were supplemented with:

10% [v/v] Heat-Inactivated Fetal Bovine Serum (FBS) (S0615 0114G)	Merck/Millipore (Burlington, USA)
1% [v/v] Penicillin-Streptomycin (10,000 U/mL) (15140122)	Gibco/Life Technologies (Carlsbad, USA)

Prokaryotic cells

Cell line	Source
NEB® 5-alpha Competent *E. coli* (High Efficiency) (C2987H)	NEB (Ipswich, USA)
BL21 (DE3) Competent *E. coli* (C2527H)	NEB (Ipswich, USA)

Eukaryotic cells

Cell line	Tissue (human)	Tumor type	Medium	Usual split ratio for culturing
HEK293T	Embryonic kidney		DMEM	1:30
HeLa (DSMZ)	Cervix	Cervical cancer	DMEM	1:10
MeWo	Skin, derived from lymph node-metastasis	Melanoma	RPMI	1:10
SK-MEL-28	Skin, derived from axillary lymph node-metastasis	Melanoma	RPMI	1:5
Tera-1	Testis, derived from lung-metastasis	Embryonal carcinoma	RPMI	1:5

Human cell lines were cultured starting from cell culture stocks available in-house.

3.1.8 Plasmids

Plasmid	Copy Number	Antibiotic resistance	General purpose	Manufacturer
pET11d	Low	Ampicillin	Protein expression in prokaryotic cells	Agilent (Santa Clara, USA)
phCMV	High	Ampicillin	Protein expression in eukaryotic cells	In-house, originally contributed by Dr. Alessia Ruggieri
pEGFP-C1	High	Kanamycin, Neomycin	Expression of EGFP protein or EGFP-fused proteins in eukaryotic cells	Clontech Laboratories (Otsu, Japan)
pcDNA3.1	High	Ampicillin, Neomycin	Protein expression in eukaryotic cells	Invitrogen/Fisher Scientific (Waltham, USA)
pSG5	High	Ampicillin	Protein expression in eukaryotic cells	Agilent Technologies (Santa Clara, USA)
pGEM-T Easy (A1360)	High	Ampicillin	Sub-cloning of DNA constructs	Promega (Madison, USA)

3.1.9 DNA-oligonucleotides

Oligonucleotides were synthesized by Eurofins MWG Genomics GmbH (Ebersberg, Germany). Purification method was High Purity Salt Free (HPSF), a cartridge purification technology based on liquid chromatography. The usual synthesis scale was 0.01 µmol or 0.05 µmol for oligonucleotides with lengths less or greater than 50 bases, respectively.

Upon reception, primers were resuspended in H_2O_{dd} to produce a 100 µM stock solution, based on information provided by the manufacturer. For each primer, a working solution of 25 µM was prepared for setting up PCR reactions.

Table 1: Primers used for generation of DNA constructs for protease and integrase expression [*]

Primers for protease studies			
Primer name	**Orientation**	**Sequence (5´-3´direction)**	**Remarks**
HOMProFOR	forward	AGCTAGCGACTATAAAGGCGAAATTCAA	*Nhe*I restriction site
HOMProREV	reverse	AGGATCCTTAGGGCATGGTGATTTCCGCACC	*Bam*HI restriction site
Pro3415F	forward	AGGATCCGCCG*CCACC*ATGGGAAGCACTGATCCAACAGG	*Bam*HI restriction site; Kozak sequence
Pro3946R	reverse	AGGATCCGGTTTAGGAGGCTCTACAGTGG	*Bam*HI restriction site
phCMV-GFPproFOR	forward	CGCGGATCCGGAAGCACTGATCCAACAGG	*Bam*HI restriction site
phCMV-GFPproREV	reverse	CGCGGATCCGGTTTAGGAGGCTCTACAGTGG	*Bam*HI restriction site
Pro-mutDTG-FOR	forward	GGTAAACACTGGAGCAGATGT	Introduction of a G->A substitution
Pro-mutDTG-REV	reverse	AACCCTTCAAACTGTTTTCCTTGA	
Pro-mutCGA-FOR	forward	TCTGTGGGGTAAAGATTTATTAC	Introduction of a CG->AA substitution
Pro-mutCGA-REV	reverse	TTAAGAGGAATTGAAGTAATCATTG	
Primers for integrase studies			
Primer name	**Orientation**	**Sequence (5´-3´direction)**	**Remarks**
HOM-HA-IN-FOR	forward	GGAT*CCACC*ATGGCATATCCT**TATGATGTTCCTGATTATGCT**GCAGGGCCTTTGACTAAAGC	*Bam*HI restriction site; Kozak sequence, HA-tag
HOM-HA-IN-REV	reverse	GGATCCTAGGTGCTTTTCTTTGCATCTC	*Bam*HI restriction site
HOM-INdelDNGfor	forward	CCAGGATATTGTAGTAAAGCTTTC	
HOM-INdelDNGrev	reverse	AGTTTTGATTTTTTCTGGAACTCC	

[*] Names and sequences of primers employed for generation of plasmid constructs for prokaryotic expression, eukaryotic expression and site-directed mutagenesis of HML-2 protease and HML-2 integrase. Restriction sites in oligonucleotide sequences are underlined, Kozak sequence is in italics and HA-tag is in bold. Nucleotide substitutions are double underlined.

Table 2: Primers for amplification of DNA templates employed for generation of proteins through *in vitro* transcription/translation system [*]

Primer Name	Orientation	Sequence (5´-3´direction)
T7gesgag-2	forward	GGATCCTAATACGACTCACTATAGGAACAGA*CCACC*ATGGGGCAAACTAAAAGT
gesgagREV	reverse	TAGGCAGGGGTCCATATAC
MAP2K2-FOR	forward	GGATCCTAATACGACTCACTATAGGGAACAG*CCACC*ATGCTGGCCCGGAGGAAG
MAP2K2-REV	reverse	TTA**AGCGTAATCTGGAACATCGTATGGGTA**CACGGCGGTGCGCG
C15ORF57-FOR	forward	GGATCCTAATACGACTCACTATAGGGAACAG*CCACC*ATGAAAATGTTTGAGAGCGCTG
C15orf57-REV	reverse	TTA**AGCGTAATCTGGAACATCGTATGGGTA**CTGTTCTGCTGCTGCTGG

31

HSP90AB1-FOR	forward	GGATCC*TAATACGACTCACTATAGGG*AACAG*CCACC*ATGCCTGAGGAAGTGCACC
HSP90AB1-REV	reverse	TTA**AGCGTAATCTGGAACATCGTATGGGTA**ATCGACTTCTTCCATGCGAG
FSCN1-FOR	forward	GGATCC*TAATACGACTCACTATAGGG*AACAG*CCACC*ATGACCGCCAACGGCACA
FSCN1-REV	reverse	TTA**AGCGTAATCTGGAACATCGTATGGGTA**GTACTCCCAGAGCGAGGC
CALR-FOR	forward	GGATCC*TAATACGACTCACTATAGGG*AACAG*CCACC*ATGCTGCTATCCGTGCCG
CALR-REV	reverse	TTA**AGCGTAATCTGGAACATCGTATGGGTA**CAGCTCGTCCTTGGCCT
PSMC4-FOR	forward	GGATCC*TAATACGACTCACTATAGGG*AACAG*CCACC*ATGGAGGAGATAGGCATCTTGG
PSMC4-REV	reverse	TTA**AGCGTAATCTGGAACATCGTATGGGTA**CTTGTAAAACTCATGCTCCTGC
DDX3X-FOR	forward	GGATCC*TAATACGACTCACTATAGGG*AACAG*CCACC*ATGAGTCATGTGGCAGTGGA
DDX3X-REV	reverse	TTA**AGCGTAATCTGGAACATCGTATGGGTA**GTTACCCCACCAGTCAACC
S100A4-FOR	forward	GGATCC*TAATACGACTCACTATAGGG*AACAG*CCACC*ATGGCGTGCCCTCTGG
S100A4-REV	reverse	TTA**AGCGTAATCTGGAACATCGTATGGGTA**TTTCTTCCTGGGCTGCTTATC
ENO1-FOR	forward	GGATCC*TAATACGACTCACTATAGGG*AACAG*CCACC*ATGTCTATTCTCAAGATCCATGCC
ENO1-REV	reverse	TTA**AGCGTAATCTGGAACATCGTATGGGTA**CTTGGCCAAGGGGTTTCTG
PDIA3-FOR	forward	GGATCC*TAATACGACTCACTATAGGG*AACAG*CCACC*ATGCGCCTCCGCCG
PDIA3-REV	reverse	TTA**AGCGTAATCTGGAACATCGTATGGGTA**GAGATCCTCCTGTGCCTTCT
RNASEH2B-FOR	forward	GGATCC*TAATACGACTCACTATAGGG*AACAG*CCACC*ATGGCCGCTGGCGTG
RNASEH2B-REV	reverse	TTA**AGCGTAATCTGGAACATCGTATGGGTA**GGAAGGACAAACTGAAACATGTAAA
TUBA1A-FOR	forward	GGATCC*TAATACGACTCACTATAGGG*AACAG*CCACC*ATGCGTGAGTGCATCTCCA
TUBA1A-REV	reverse	TTA**AGCGTAATCTGGAACATCGTATGGGTA**GTATTCCTCTCCTTCTTCCTCAC
CIAPIN1-FOR	forward	GGATCC*TAATACGACTCACTATAGGG*AACAG*CCACC*ATGGCAGATTTTGGGATCTCTG
CIAPIN1-REV	reverse	TTA**AGCGTAATCTGGAACATCGTATGGGTA**GGCATCATGAAGATTGCTATCAC
RANBP1-FOR	forward	GGATCC*TAATACGACTCACTATAGGG*AACAG*CCACC*ATGAGGACCATGATTTCCACTG
RANBP1-REV	reverse	TTA**AGCGTAATCTGGAACATCGTATGGGTA**TTGCTTCTCCTCAGCATCCT
HSP90AA1-FOR	forward	GGATCC*TAATACGACTCACTATAGGG*AACAG*CCACC*ATGCCTGAGGAAACCCAGA
HSP90AA1-REV	reverse	TTA**AGCGTAATCTGGAACATCGTATGGGTA**GTCTACTTCTTCCATGCGTGAT

* Target sequence-specific sequence portions are underlined, Kozak sequence is in italics, T7 primer is underlined with a dashed line, HA-tag is in bolb. All reverse primers include a stop codon (TTA or TAG) at the 5´end (thus at the 3' end of the final PCR-product).

Table 3: Primers used for sequencing of plasmids generated in this study

Primer name	Orientation	Sequence (5´-3´direction)
CMV fwd	forward	CGCAAATGGGCGGTAGGCGTG
EGFP-C F	forward	CATGGTCCTGCTGGAGTTCGTG
pEGFP-C.rev	reverse	TGCATTCATTTTATGTTTCAGG
SP6	reverse	GATTTAGGTGACACTATAG
T7	forward	TAATACGACTCACTATAGGG

3.1.10 Devices

Instrument	Manufacturer
Agarose gel electrophoresis chamber	Renner GmbH (Dannstadt, Germany)
Analytical balance (Basic)	Sartorius (Göttingen, Deutschland)
Analytical precision balance (Europe 60)	Gibertini Elettronica (Novate Milanese, Italy)
Autoclave (VE-150)	Systec (Wettenberg, Germany)
Automatic cell counter (Luna)	Logos Biosystems (Annandale, USA)
Biophotometer (6131)	Eppendorf (Hamburg, Germany)
Blotting system (XCell II™ Blot module)	Invitrogen/Life Technologies (Carlsbad, USA)
Blue light transilluminator (Flu-0-Blu)	Biozym Scientific (Hessisch Oldendorf, Germany)
CCD-camera (DP71)	Olympus (Tokyo, Japan)
CCD-camera (LU105M)	Lumenera (Ottawa, Kanada)
Centrifuge (Rotofix 32)	Hettich (Tuttlingen, Germany)
Centrifuge (Universal 1200-01 D-7200)	Hettich (Tuttlingen, Germany)
Dounce homogenizer	Fortuna (Germany)
Electrophoresis system (PerfectBlueTM Mini ExW)	Peqlab/Avantor (Erlangen, Germany)
Electrophoresis system (XCell SureLock™ Mini-Cells)	Invitrogen/Life Technologies (Carlsbad, USA)
Flow cytometer (BD FACSCanto™ II)	BD Biosciences (San Jose, USA)
Freezing container (CoolCell)	BioCision (Mill Valley, USA)
Gel documentation system (DeVision Dbox)	DC Science Tec (Hohengandern, Germany)
Horizontal mixer (RM-5)	CAT (Ballrechten-Dottingen, Germany)
Ice machine (F75L)	Migel (Milano, Italy)
Imaging system (Chemidoc Touch)	Bio-Rad (Hercules, USA)
Incubation shaker (Multitron II)	Infors (Basel, Switzerland)
Incubator (Galaxy 170S CO_2)	Eppendorf (Hamburg, Germany)
Incubator (Jouan)	Thermo Fisher Scientific (Waltham, USA)
Inverted light microscope (Axiovert 25)	Carl Zeiss (Jena, Germany)
Liquid nitrogen tank (Biosafe-MDβ)	Cryotherm (Denver, USA)
Magnetic stirrer (IKAMAG RCT)	IKA (Staufen, Germany)
Microcentrifuge (MIKRO 120)	Hettich (Tuttlingen, Germany)
Micropipettes (Pipetman) (10-20-200-1000 µl)	Gilson (Middleton, USA)
Microplate reader (Spark 10M)	Tecan (Männedorf, Switzerland)
Microscope (AX70 system)	Olympus (Tokyo, Japan)
Mixing block (MB-102)	Bioer (Hangzhou, China)
Molecular imager (Typhoon 9410)	GE Healthcare (Uppsala, Sweden)
pH meter (Seven Easy)	Mettler-Toledo (Gießen, Germany)
Pipette controller (accu-jet pro)	Brand (Wertheim, Deutschland)
Power supply (Standard Power Pack p25)	Biometra/Analytikjena (Göttingen, Germany)
Refrigerated centrifuge (2-16K)	Sigma (Osterode am Harz, Germany)
Refrigerated microcentrifuge (5417R)	Eppendorf (Hamburg, Germany)
Shaker (KS 125 Basic)	IKA (Staufen, Germany)
Shaker (VORTEX 3)	IKA (Staufen, Germany)

Sonicator (Ultrasonics Sonifier S-250A)	Branson (Danbury, USA)
Spectrophotometer (Infinite® m200)	Tecan (Männedorf, Switzerland)
Spectrophotometer (NanoDrop™ 2000)	Thermo Fisher Scientific (Waltham, USA)
Thermoblock (TCR 100)	Carl Roth (Karlsruhe, Germany)
Thermocycler (TGradient)	Biometra/Analytikjena (Göttingen, Germany)
Ultracentrifuge (Avanti J-26 XP)	Beckman Coulter (Brea, USA)
Ultracentrifuge rotor (JLA-8.100)	Beckman Coulter (Brea, USA)
UV light transilluminator	Wealtec (Sparks, USA)
Water heater bath (Thermomix 5 BU)	B. Braun (Melsungen, Germany)
Water purification system (Milliq Integral 15)	Millipore (Bedford, USA)

3.1.11 Consumables

Consumables	**Manufacturer**
12-well cell culture plates (665180)	Greiner BioOne (Frickenhausen, Germany)
8-well cell culture chambers (94.6140.802)	Sarstedt (Nümbrecht, Germany)
96-well microplates (655087)	Greiner BioOne (Frickenhausen, Germany)
Amicon® Ultra-15 centrifugal filter unit (UFC9003)	Merck/Millipore (Burlington, USA)
Cell culture flasks 175 cm² (661175)	Greiner BioOne (Frickenhausen, Germany)
Cell culture flasks 75 cm² (658175)	Greiner BioOne (Frickenhausen, Germany)
Dialysis tubing (68100)	Thermo Fisher Scientific (Waltham, USA)
FACS tubes (5 ml, 75x12 mm, PS) (55.1579)	Sarstedt (Nümbrecht, Germany)
Falcon tubes (15 ml - 50 ml)	Greiner BioOne (Frickenhausen, Germany)
Gel casting cassettes (NC2010)	Novex/Invitrogen (Winston-Salem, USA)
Parafilm M Bemis	VWR (Radnor, USA)
PCR tubes (0.2 ml)	VWR (Radnor, USA)
Pipette tips (10 μl)	Sorenson Biosciences (Salt Lake City, USA)
Pipette tips (20 μl - 200 μl - 1000 μl)	VWR (Radnor, USA)
Polypropylene disposal bags	Carl Roth (Karlsruhe, Germany)
PVDF membrane (Amersham Hybond-P)	GE Healthcare (Uppsala, Sweden)
Serological pipettes (5 ml - 10 ml - 25 ml)	Greiner BioOne (Frickenhausen, Germany)
Syringe filters (Minisart 0.4 μm - 0.2μm)	Sartorius (Göttingen, Deutschland)
Syringes (Injekt 20 ml)	B. Braun (Melsungen, Germany)
Tubes (301003)	Glasgerätebau Ochs (Bovenden, Germany)
Whatman paper	GE Healthcare (Uppsala, Sweden)

3.1.12 Software, online-tools and databases

- *CellSens Imaging Software* (Olympus, Tokyo, Japan): analysis and acquisition of images via Olympus AX70 system
- *DeVision G* V2.0 (DC Science Tec, Hohengandern, Germany): documentation of agarose gels
- *Geneious Prime* (Biomatters Ltd., Auckland, New Zealand): analysis and processing of DNA and protein data
- *Graphic Converter 10* (Lemke Software GmbH, Peine, Germany): image processing
- *Image Lab 5.2.1* (2014) (Bio-Rad, Hercules, USA): analysis and processing of images captured with Chemidoc; quantification of band volume intensities
- *Mendeley Desktop V1.19* (Mendeley Ltd., London, UK): reference manager
- *NCBI (National Center for Biotechnology Information)* (Bethesda, Maryland, USA): database providing biomedical and genomic information
- *Office 2016 for Mac* (Microsoft, Redmond, USA): texting, data analysis, slides
- *Venn diagrams tool* (http://bioinformatics.psb.ugent.be/webtools/Venn/): intersection of data sets, Venn diagrams
- *PANTHER (Protein ANalysis THrough Evolutionary Relationships)* classification system (http://pantherdb.org/about.jsp): classification of proteins and their genes; identification of cellular localization and function of gene products
- *Catalogue Of Somatic Mutations In Cancer* (COSMIC) *database* (https://cancer.sanger.ac.uk/cosmic): identification of cancer diseases related to a certain gene/protein
- *Online Mendelian Inheritance in Man (OMIM) database* (https://www.omim.org): identification of disease phenotypes related to a certain gene/protein
- *Uniprot database* (https://www.uniprot.org): recovery of protein sequences and protein functional information

3.2 Methods

3.2.1 Manipulation of nucleic acids

3.2.1.1 Polymerase chain reaction

Polymerase chain reaction (PCR) [175] is an *in vitro* enzymatic assay that allows for amplification of specific DNA fragments. The key component of this methodology is a DNA polymerase that is resistant to heat inactivation. The DNA polymerase derived from the bacterium *Thermus aquaticus* (*Taq*) is commonly employed as a standard for applications that involve the PCR. *Taq* polymerase has a temperature optimum of 70-80°C and an activity half-life of 45-50 min at 95°C [117]. The PCR reaction takes place in a defined buffer that usually includes $MgCl_2$, necessary as cofactor of the *Taq* DNA polymerase. General PCR buffers also include Tris-HCl, KCl, glycerol and additional variable components, with a final buffer pH of approximately 8.0. PCR buffers are usually provided by manufacturers as concentrated mixtures that include all components necessary for optimal polymerase reaction conditions. In addition to the DNA polymerase and its proper buffer, other elements to be included in a PCR reaction are as follows: 1) a DNA template; 2) primers, short oligonucleotides complementary to the target DNA region; 3) deoxynucleoside triphosphates (dNTPs), the molecular blocks for building new DNA strands. The thermal cycler is the device that, via programmed temperature shifts, permits cycling of the three main PCR steps: 1) denaturation of the dsDNA template generating ssDNA; 2) annealing of the primers to each of the ssDNA templates; 3) synthesis and extension of a nascent DNA strand (complementary to the DNA template) by addition of free dNTPs to the 3´-terminus of the primers and subsequently synthesized DNA-strand. Concentration of reaction components, duration and temperature of each step of the procedure, are usually recommended by manufacturers. Some parameters can be optimized towards best possible results. For instance, it may be necessary to evaluate the melting temperature of primers, purity of the DNA template and the duration of the extension step, the latter depending on the length of the amplified DNA fragment. Primers can vary in length and they can be designed so that specific sequence features, such as restriction sites, promoters or nucleotide substitutions necessary for downstream applications, are added in the final DNA product. A characteristic of certain thermostable polymerases, including *Taq* DNA polymerase, is the addition of a 3'-adenine overhang at the 3' ends of the amplified DNA fragment, that becomes useful in some cloning systems (see chapter 3.2.1.6).

In our work, we performed PCRs employing *Taq* DNA Polymerase (Sigma-Aldrich) for amplification of specific DNA constructs to be cloned into expression vectors and for generation of PCR products to be used in an *in vitro* transcription/translation system. *Taq* DNA Polymerase was supplied with a standard 10x PCR buffer including $MgCl_2$ at 15 mM. dNTPs were prepared as a stock solution (dNTP mix) containing dATP, dCTP, dGTP, and dTTP at a concentration of 2.5 mM each. Typical composition and cycling parameters of PCR reactions performed in our lab with *Taq* DNA Polymerase are given in Table 4 and Table 5, respectively.

Table 4: Composition of PCR reactions with *Taq* DNA polymerase

Component	Final concentration
10x PCR Buffer	1x
Template DNA	~ 0.1 ng/µl
dNTP mix	100 µM (each dNTP)
Forward primer	0.25 µM
Reverse primer	0.25 µM
Taq DNA Polymerase	0.05 unit/µl
H_2O_{dd} to a final volume of 50 µl	

Table 5: PCR cycling parameters for PCR reactions with *Taq* DNA polymerase

PCR step	Temperature °C	Duration
Initial denaturation step	94°C	3 min
Around 30 cycles of:		
Denaturation of DNA Template	94°C	1 min
Primer annealing	56°C	1 min
Primer extension	72°C	1 min per 1000 bp
Final extension	72°C	10 min
Hold	8°C	∞

As will be described in more detail in chapter 3.2.1.9, we also employed the PCR method for site-directed mutagenesis. In such a procedure, pGEM-T Easy vector harboring the DNA fragment to be mutated was amplified and subsequently re-ligated. Considering that the *Taq* polymerase has a relatively high error rate during DNA synthesis and is usually not suited for amplification of longer PCR amplicons, we used Phusion® High-Fidelity DNA Polymerase (NEB), a highly accurate thermostable DNA polymerase that is suitable for amplification of long PCR-amplicons including a low error rate. The 10x Phusion® HF Buffer (supplied by the manufacturer) included $MgCl_2$ at 7.5 mM. The dNTP mix was the same as for PCRs with *Taq* DNA polymerase. Reaction conditions used for PCRs with Phusion® High-Fidelity DNA Polymerase and cycling parameters are given in Table 6 and Table 7, respectively.

Table 6: Composition of PCR reactions with Phusion DNA polymerase

Component	Final concentration
5x Phusion® HF Buffer	1x
Template DNA	~ 0.3 ng/µl
dNTPs mix	100 µM (each dNTP)
Forward primer	0.5 µM
Reverse primer	0.5 µM
Phusion® High-Fidelity DNA Polymerase	1 unit/50 µl
H_2O_{dd} to a final volume of 50 µl	

Table 7: PCR cycling parameters used for PCR reactions with Phusion DNA polymerase

PCR step	Temperature °C	Duration
Initial denaturation step	98°C	30 sec
Around 30 cycles of:		
Denaturation of DNA Template	98°C	10 sec
Primers annealing	55°C	20 sec
Primers extension	72°C	80 sec
Final extension step	72°C	10 min
Hold	8°C	∞

3.2.1.2 Agarose gel electrophoresis

In agarose gel electrophoresis, application of an electric field causes migration of charged biomolecules through an agarose matrix, separating those biomolecules by size. This method is mostly employed for separation of nucleic acids, that migrate towards the positively charged anode because of negatively charged phosphates in the sugar phosphate backbone. Fluorescent agents intercalating with nucleic acids are commonly used to visualize nucleic acid fragments subsequent to gel electrophoresis. Composition of buffers employed for agarose gel electrophoresis is reported in chapter 3.1.2. In our experiments, DNA samples were mixed with 10x DNA Loading Buffer (1x final concentration) before loading them onto an agarose gel containing 1-2% [w/v] agarose in 1x TAE Running Buffer. DNA was separated by applying an electrical field in standard electrophoresis chambers filled with 1x TAE Running Buffer. DNA samples were usually separated at 80 V for 1.5 h. 2-Log DNA Ladder (NEB) was used as size standard for DNA fragments (see chapter 3.1.6). For documentation purposes, DNA fragments were detected by ethidium bromide present in the agarose gel at a final concentration of 0.625 µg/ml, and visualized under UV light (365 nm) subsequent to electrophoresis. For experiments requiring preserved integrity of nucleic acids, agarose gels lacked ethidium bromide and were instead soaked, after electrophoresis, in 1x GelStar® Stain (Lonza) in 1x TAE for 30 min, and the DNA was then visualized and documented using a blue light (470 nm) transilluminator.

3.2.1.3 PCR clean-up and DNA extraction from agarose gels

NucleoSpin® Gel and PCR Clean-up Kit (Macherey Nagel) was employed for DNA purifications from enzymatic reactions as well as for purifications of DNA fragments from agarose gels. Buffers were supplied with the kit and the kit was used according to the manufacturer's recommendations. In this procedure the sample is mixed with a binding buffer and, in case of a cut-out gel band, is heated to dissolve the agarose. The DNA is bound to a silica membrane of a centrifugable column in presence of chaotropic salt. Contaminations are removed by washing steps in presence of ethanol. The purified DNA is then eluted under low salt conditions in a mildly alkaline elution buffer.

3.2.1.4 Nucleic acids concentration measurements

Nucleic acids concentration was measured by spectrophotometric absorption at 260 nm (RNA/DNA absorption) using a NanoDrop^{TM} 2000 Spectrophotometer. Nucleic acids purity was determined by measuring absorption at 280 nm (proteins absorption) and calculating the ratio of absorption at 260 nm and absorption at 280 nm. A value greater than or equal to 1.8 indicated sufficient purity of DNA.

3.2.1.5 Digestion of DNA by restriction enzymes

Restriction enzymes (REs) are fundamental tools employed to cut dsDNAs into fragments. The cleavage of a dsDNA by a RE occurs at a so-called restriction site, a specific nucleotide sequence recognised by the RE. We involved REs in our cloning strategies to generate plasmids harboring desired sequence regions and as a screening tool to confirm presence and orientation of DNA inserts in plasmids generated. REs and buffers employed were from New England Biolabs (NEB). Each RE exhibits an optimal enzymatic activity at conditions that can differ between enzymes. RE reactions were usually performed as shown in Table 8. Buffers varied between REs as recommended by the manufacturer (see chapter 3.1.3). The volume of RE was typically 5% of the total reaction volume. The final concentration of RE varied depending on the enzyme's stock concentration, the latter being 10000 U/ml for some REs and 20000 U/ml for others (see chapter 3.1.3). In reactions with two different REs (double digests), the volume of each RE was typically 2.5% of the total reaction volume. All the REs employed required incubation at 37°C, as recommended by the manufacturer. Incubation time was 2-3 h when REs were used for confirmation of presence and orientation of inserts in plasmids. For preparation of DNA inserts to be subsequently cloned into expression plasmids, RE reactions were incubated overnight for higher yield of restriction fragments and a reduced amount of undigested plasmid. When recommended by the manufacturer, REs were inactivated by incubation for 10 min at 80°C.

Table 8: Typical reaction composition for DNA digestion using restriction enzymes

2 μl	10x NEB Buffer
~1 μg	DNA
1 μl	Restriction enzyme (10000 U/ml or 20000 U/ml)
H_2O_{dd} to a final volume of 20 μl	

3.2.1.6 Ligation of DNA inserts into plasmids

In combination with restriction enzymes, DNA ligation, mediated by DNA ligases, enables a workflow where a defined DNA fragment is cloned into a plasmid or transferred from one plasmid to another.

pGEM-T Easy vector system

PCR products generated for cloning purposes were cloned into pGEM-T Easy vector (Figure 4) using pGEM®-T Easy Vector Systems (Promega). pGEM-T Easy vector is a linearized vector with a single 3´-terminal thymidine overhang at both 3´ends that greatly improves ligation efficiency of 3´ A-tailed fragments generated by *Taq* DNA polymerase. Standard ligation reactions with pGEM-T Easy vector were set up as given in Table 9.

Table 9: Standard reaction for ligation of PCR product into pGEM-T Easy vector

5 µl	2x Rapid Ligation Buffer
1 µl	pGEM-T Easy vector (50 ng/µl)
~150 ng	PCR product
1 µl	T4 DNA Ligase (3000 U/ml)
H_2O_{dd} to a final volume of 10 µl.	

A reaction with a control insert DNA (supplied with the kit) and a reaction without PCR product were usually included as controls for further blue/white color screening of clones (see below). After overnight incubation at 4°C, 2 µl from each reaction were used for transformation of NEB® 5-alpha Competent *E. coli* (see 3.2.1.7). Transformed cells were plated onto LB-Agar$_{amp}$ plates containing 80 µg/ml 5-Brom-4-chlor-3-indoxyl-β-Dgalactopyranosid (X-Gal). The pGEM T-Easy vector is engineered in a way that successful cloning of an insert interrupts a sequence coding for β-galactosidase. X-Gal is a colorless compound that can be cleaved by β-galactosidase resulting in a reaction product of bright blue-color. Thus, bacteria transformed with vector harboring an insert grow into white colonies, whereas those transformed with only the vector lacking an insert result in blue colonies. Control ligations without PCR product generated blue colonies only, whereas controls with control insert DNA generated white colonies and a relatively low number of blue colonies.
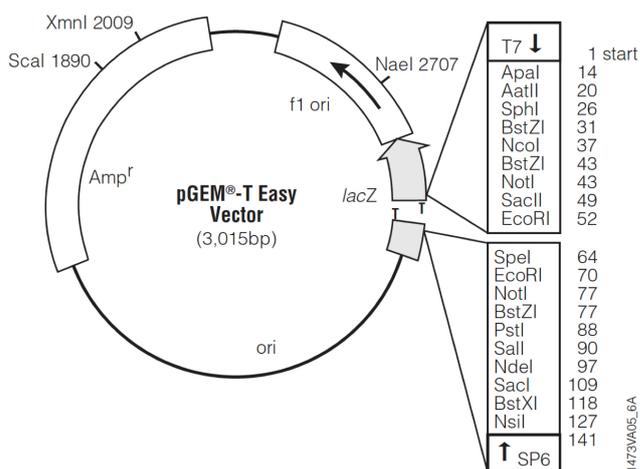
Figure 4: pGEM-T Easy vector map and selected sequence features. Figure retrieved from the pGEM®-T and pGEM®-T Easy Vector Systems Technical Manual (Rev. 6/15 TM042).

Ligation of DNA inserts into linearized expression vector

DNA inserts were excised from pGEM-T Easy plasmids by restriction enzyme digests, purified subsequent to agarose gel electrophoresis, and cloned into a linearized expression vector using T4 DNA Ligase (NEB). In order to reduce self-ligation of linearized expression vector backbone, linearized plasmid was first treated with rAPid Alkaline Phosphatase (Merck/Sigma-Aldrich) to remove 5′-phosphoryl termini in plasmid DNA strands. By doing so, ligation of linearized vector DNA was possible only via the 5′-phosphoryl termini provided by the DNA strands to be inserted into the expression vector. Dephosphorylation reactions (Table 10) were incubated for 10 min at 37°C followed by 2 min at 75°C for heat inactivation of rAPid Alkaline Phosphatase. Dephosphorylated vector DNA was directly used in ligation reactions (Table 11) in presence of to be cloned DNA fragments. In some cases, control reactions with non-dephosphorylated vector or with dephosphorylated vector, in both cases without insert, were also performed in parallel in order to evaluate background ligation of plasmid backbone. Ligation reactions were incubated overnight at 16°C, followed by 10 min incubation at 65°C in order to heat-inactivate T4 DNA Ligase. 1-5 µl of ligation mixtures were used for transformations of *E. coli* cells each.

Table 10: Reaction for dephosphorylation of linearized vector by alkaline phosphatase

0.5 µg	Linearized expression vector
2 µl	rAPid Alkaline Phosphatase Buffer (10X)
1 µl	rAPid Alkaline Phosphatase (1000 units/ml)
H_2O_{dd} to a final volume of 20 µl	

Table 11: Reaction for ligation of vector and insert DNA by DNA ligase

2 µl	10x T4 DNA Ligase Buffer
50 ng	Dephosphorylated expression vector DNA
~150 ng	Insert DNA
1 µl	T4 DNA Ligase (40000 units/ml)
H_2O_{dd} to a final volume of 20 µl	

3.2.1.7 Transformation of *E. coli* cells

Transformation is the process by which an exogenous DNA is introduced into a bacterial cell. There are various bacterial strains that can be transformed very efficiently and are suitable for molecular biology applications. In our work, NEB® 5-alpha Competent *E. coli* (C2987H, NEB), a high efficiency strain derivative of *E. coli* DH5α cells, were employed for propagation of plasmids and screening of plasmid clones; while BL21 (DE3) Competent *E. coli* (C2527I, NEB) were employed for higher-level production of HML-2 Pro.

Transformation of NEB® 5-alpha Competent *E. coli* was performed as follows. 20 μl of cells were pipetted into an Eppendorf tube and kept on ice. 1 μl containing approximately 1 ng of plasmid DNA was gently added to the cell mixture, followed by incubation on ice for 30 min. Cells were then subjected to heat shock for 30 sec at 42°C in a waterbath and incubated again on ice for 5 min. 80 μl of LB-Medium was then gently added to transformed cells and the mixture was incubated for 1 h at 37°C. During this incubation step, *E. coli* cells could recover from the heat shock and establish antibiotic resistance gained by transformed plasmids. The 100 μl mixture was then plated onto an LB-Agar plate for the purpose of subsequent screening of plasmid-harboring clones, whereas it was directly inoculated into LB-Medium for the purpose of subsequent isolation of larger amounts of plasmid DNA. In both cases, LB-Medium was supplemented with specific antibiotics at standard concentration (see chapter 3.1.7). *E. coli* cells were grown overnight in liquid culture flasks in a shaking incubator at 37° C and 200 rpm. The following day, liquid cultures were subjected to plasmid-DNA isolation, whereas *E. coli* colonies grown on LB-Agar plates (with antibiotics) were further screened for transformed clones. The protocol employed for transformation of *E. coli* BL21 (DE3) Competent cells is described in chapter 3.2.2.3.

3.2.1.8 Isolation of plasmid DNA from *E. coli* cells

The peqGOLD Plasmid Miniprep Kit I (PeqLab) was used for isolating relatively small amounts (15-25 μg) of plasmid DNA from *E. coli* cells, with that amount of plasmid DNA being sufficient for screening of plasmid constructs. The principle of the purification is based on alkaline lysis of bacteria and subsequent binding of plasmid DNA to a silica membrane in a centrifugable column. DNA binding occurs in presence of high salt concentration, proteins and other contaminants are removed through a wash step, nucleic acids are eluted with a low salt buffer. In our procedure, a single *E. coli* colony, harboring transformed plasmid, was picked from an LB-Agar plate with a plastic tip. The tip of the tip, thus *E. coli* cells, were streaked (forming a line) in a particular cell of a numbered grid of an LB-Agar plate (Master-plate). The same tip, that is remaining *E. coli* cells, was then inoculated into 5 ml of LB-Medium, with that tube labeled with the same number as the grid cell. This was repeated for at least 10 plasmid clones per plasmid construct generated. The 5 ml cultures were incubated overnight in an orbital shaker at 37°C. Then, *E. coli* cells were centrifuged for 10 min at 5000 g and cell pellets

were further processed following the kit manufacturer's recommendations. Plasmid DNAs were eluted into 100 μl of elution buffer each (provided with the kit) and stored at -20°C after measuring DNA concentration. The Master-plate, incubated overnight at 37°C and then stored at 4°C, served as a source of positive *E. coli* cells for producing larger amounts of plasmid DNA, sparing another transformation by inoculating the desired *E. coli* clone from the Master-plate directly into a liquid culture.

Isolation of larger amounts (20-100 μg) of plasmid DNA were carried out using the NucleoBond® PC 100 plasmid DNA purification kit (Macherey Nagel). The method employs a modified alkaline/SDS lysis procedure to prepare the bacterial cell pellet for plasmid purification. Nucleic acids are denatured under alkaline conditions. Potassium acetate is added to the denatured lysate, causing formation of a precipitate containing chromosomal DNA and other cellular compounds. Plasmid DNA remains in solution. Plasmid DNA is then specifically bound to an anion exchange column under low salt concentration and appropriate pH conditions, washed and eluted under high salt concentration. Plasmid DNA is then concentrated by isopropanol precipitation, centrifugation, a wash step, and finally dissolved in H_2O_{dd}. For our procedure, an *E. coli* colony picked from a Master-plate, or a just a liquid culture of transformed *E. coli* cells (see above), were inoculated and cultured overnight at 37°C, in a flask containing LB-Medium supplemented with the appropriate antibiotic. Culture volumes were 30 ml or 100 ml when *E. coli* cells harbored, respectively, high-copy or low-copy number plasmids. Plasmid DNA was eluted into 100 μl of H_2O_{dd}. Plasmid DNA concentration was determined by UV spectrophotometry and subsequently adjusted to 500 ng/μl with H_2O_{dd}.

3.2.1.9 Generation of plasmids harboring mutated variants of HML-2 Pro and HML-2 IN coding sequences

Enzymatically inactive Pro and IN variants were generated by site-directed mutagenesis. Site-directed mutagenesis was performed via PCR, employing two strategies for generation of nucleotide substitution or deletion, respectively (Figure 5). The general procedure consists of PCR amplification of the complete sequence of a plasmid harboring the sequnce region to be mutated. PCR primers used for the amplification are designed in a way that the resulting PCR product (representing the linearized plasmid) contains the desired mutation. Re-ligation of the linear PCR product creates a circular plasmid harboring the mutation of interest.

In our procedure we used Phusion® High-Fidelity DNA Polymerase with PCR conditions as indicated in chapter 3.2.1.1. Pro and IN constructs cloned into pGEM-T Easy vector served as DNA templates for PCR. We aimed at generating two Pro mutants (Pro-mutDTG and Pro-mutGRDL) and an IN mutant (IN-mut). The strategy employed for generating Pro mutants involved nucleotide substitution introduced by the forward PCR primers (Figure 5). More specifically, a single nucleotide exchange G→A was introduced in PromutDTG-FOR, the forward primer for generation of plasmids pGEM Pro

mut DTG (for pET11d) and pGEM Pro-mut DTG (for phCMV). A dinucleotide cexhange CG →AA was introduced by Pro-mutCGA-FOR, the forward primer for generation of plasmid pGEM Pro-mutCGA (for pET11d) (see Table 1 for primer sequences). The strategy for generating IN-mut introduced a nucleotide deletion. Nucleotides 5'-GACAATGGA-3' were deleted employing PCR primers complementary to the template DNA yet spacing away from each other by a nucleotide distance that excludes the sequence region to be deleted.
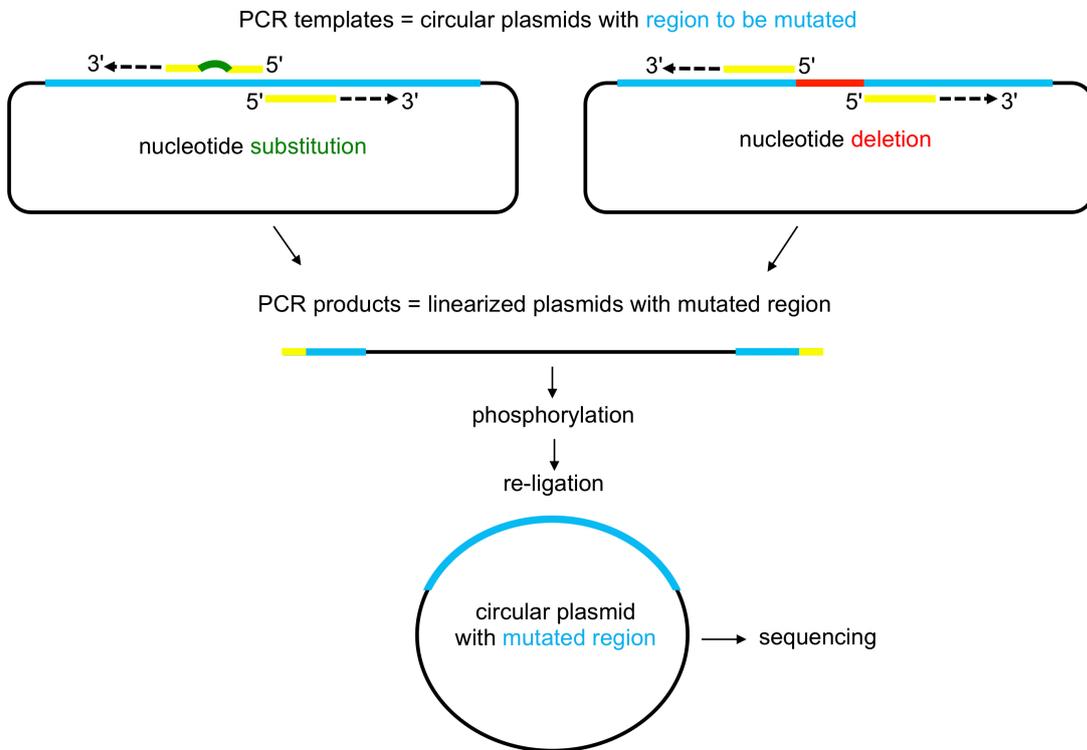
Figure 5: Schematic representation of the strategy employed for generation of plasmids harboring mutated variants of Pro and IN coding sequences. PCR site-directed mutagenesis was performed by two methods for generation of nucleotide substitution or deletion. Plasmids harboring the sequence region of interest to be mutated comprised the template DNA. PCR primers are depicted in yellow. Protein coding sequences are shown in blue. For nucleotide substitution, the forward primer contains the mutation (in green). For nucleotide deletions, primers are spaced away from each other to exclude the region to be deleted (in red). Following PCR, steps for generating circularized functional plasmids were the same for all constructs.

Resulting PCR products represented linear pGEM plasmids, including the DNA insert, yet harboring the mutated variants of the protein coding sequences of interest. Circular plasmids could be recreated by re-ligation of such linear plasmids. For this purpose, PCR products were first purified following agarose gel electrophoresis in order to remove template plasmid DNA template and other potential contaminations. Since primers used for the PCR were not phosphorylated at their 5' ends resulting PCR products were treated with T4 Polynucleotide Kinase (NEB) (T4 PNK) which catalyzes the transfer of a phosphate from ATP to the 5′-hydroxyl terminus of a polynucleotide, thus adding a 5' phosphoryl group. The phosphorylation reaction (Table 12) was set up with T4 DNA Ligase Buffer (NEB) which

contains 1 mM ATP and provides suitable reaction conditions for T4 PNK activity. Following 30 min of incubation at 37°C, T4 PNK was heat-inactivated by incubation for 20 min at 65°C.

Re-ligation of 5'-phosphorylated PCR products was therefore done by addition of 1 µl of T4 DNA Ligase (NEB) directly to the 5'-phosphorylation reaction and incubation overnight at 16°C. Finally, reactions were incubated for 10 min at 65° C in order to heat-inactivate the T4 DNA Ligase. Circular pGEM plasmids harboring desired mutants of Pro and IN were verified by sequencing (see below) before subsequent cloning steps.

Table 12: Reaction for 5'-phosphorylation of linear plasmids generated by PCR-mediated site-directed mutagenesis

2 µl	10x T4 DNA Ligase Buffer
~150 ng	PCR product
1 µl	T4 PNK (10000 units/ml)
H_2O_{dd} to a final volume of 20 µl	

3.2.1.10 Generation of plasmids for HML-2 Pro and HML-2 IN expression

This section describes plasmids employed and cloning strategies used for generation of plasmids for prokaryotic and eukaryotic expression of HML-2 Pro and for eukaryotic expression of HML-2 IN. PCR conditions used for amplification of DNA constructs with *Taq* DNA Polymerase are reported in chapter 3.2.1.1. Sequences of primers mentioned in this section are listed in Table 1.

3.2.1.10.1 Plasmid backbones

pET-11d

The pET-11d vector (Agilent) was used for expression of HML-2 Pro in *E. coli*. A description of the plasmid and a plasmid map can be found in chapter 3.2.2.3.

phCMV-G

The phCMV-G vector [95] expresses the G glycoprotein of vesicular stomatitis virus (VSV-G) under control of the CMV early promoter (Figure 6 A). We employed this plasmid as eukaryotic expression vector, replacing (via *Bam*HI digestion) the VSV-G sequence with the insert of interest, consisting of a Kozak consensus sequence followed by HML-2 Pro coding sequence. The Kozak sequence plays an important role in initiation of translation, ensuring translation from the correct initiation codon and increasing translation efficiency in eukaryotic cells [110]. A β-globin intron (enhancing expression of coding sequence of interest because more efficient export of spliced mRNA from the nucleus) [95] and a polyadenylation signal are located upstream and downstream of the insert, respectively. An ampicillin-resistance gene for antibiotic selection and propagation of plasmid-harboring bacteria is also present.

pEGFP-C1

The pEGFP-C1 vector (Clontech Laboratories, GenBank Accession: U55763) encodes a variant of wild-type GFP optimized for brighter fluorescence, therefore called enhanced GFP (EGFP). The sequence flanking EGFP contains a Kozak consensus sequence that enhances translation efficiency in eukaryotic cells. The multiple cloning site in pEGFPC1 is located between the EGFP coding sequences and the SV40 poly A (Figure 6 B). The pEGFP-C1 provides a pUC origin of replication for high copy number propagation in *E. coli*. A bacterial promoter upstream of a neomycin resistance cassette (Neo^r) expresses kanamycin resistance in *E. coli*. The sequence of interest is cloned into the MCS, in-frame with EGFP, so that it can be expressed in mammalian cells as a protein fused to the C-terminus of EGFP. Fusions to the C terminus of EGFP retain the fluorescent properties of the native protein. We employed pEGFP-C1 for expression of EGFP (in controls) and EGFP fused with HML-2 Pro in cell culture.

pcDNA3

The pcDNA3 vector (Invitrogen/Fisher Scientific) is designed for expression of proteins in mammalian cells. pcDNA3 harbors a CMV immediate-early promoter for high-level expression, a multiple cloning site flanked by the T7 and SP6 promoters, a pUC origin of replication for high copy number replication in *E. coli*, and an ampicillin-resistance gene (Figure 6 C). For proper and efficient translation, it is required that the insert contains a Kozak consensus sequence with an ATG initiation codon. We employed pcDNA3 for expression of candidate proteins in cell culture. In each construct, the ATG of the Kozak sequence was the starting codon of the cloned candidate-protein encoding sequence.

pSG5

The pSG5 vector (Agilent Technologies, Santa Clara, USA) is a high-copy plasmid that can be used for expression of proteins *in vitro* and *in vivo* (Figure 6 D). After transient transfection of cell lines, expression is driven from the SV40 promoter and terminated by a polyadenylation signal downstream of the MCS. *In vitro* transcription can be driven from the T7 promoter. The desired expression sequence can be cloned via restriction enzyme sites (E*co*RI, *Bam*HI and *Bgl*II). An ampicillin resistance gene serves for antibiotic selection in transformed bacteria. We employed pSG5 for expression of HML-2 IN in cell culture.
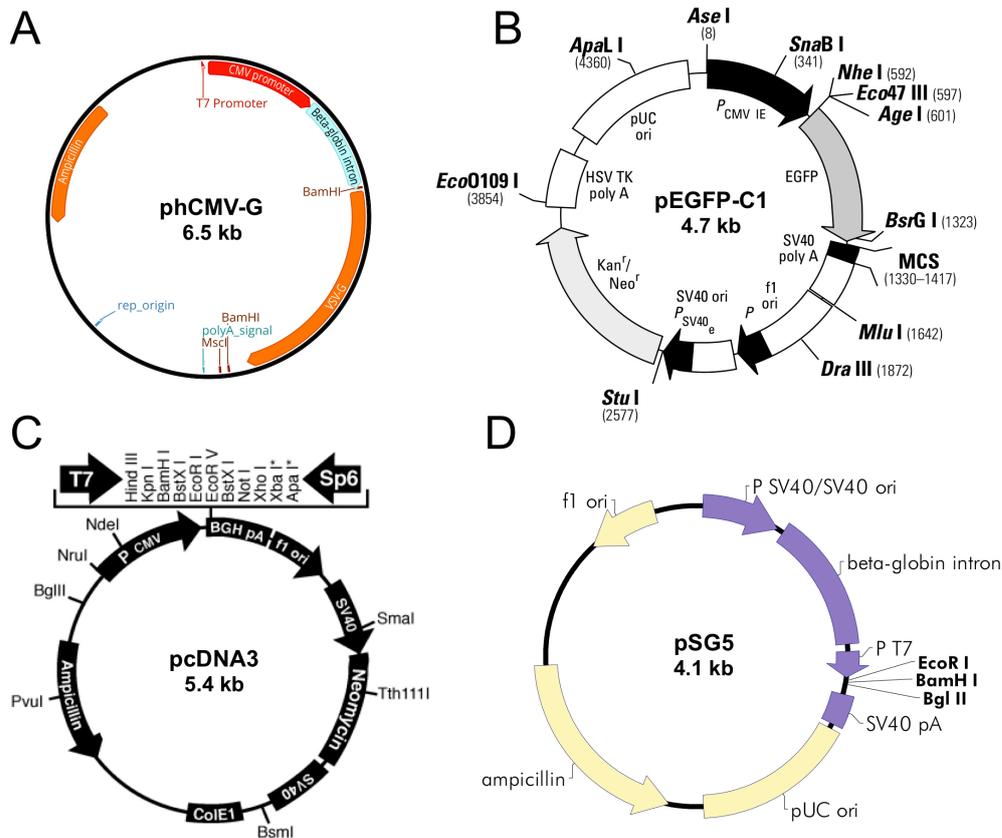
Figure 6: Plasmid maps of vectors employed for protein expression in mammalian cells. **A)** Plasmid map of phCMV-G. Only restriction sites used in our cloning strategies are depicted. The plasmid map was generated by us using the software Geneious (Biomatters). **B)** Plasmid map of pEGFP-C1 (retrieved from BD Biosciences Clontech Catalog 6084-1, Version PT3028-5). **C)** Plasmid map of pcDNA3 (retrieved from Invitrogen catalog A-150228). **D)** Plasmid map of pSG5 (retrieved from Agilent Technologies pSG5 Vector Instruction manual (Rev. C.0)).

3.2.1.10.2 Novel plasmid constructs generated

pET11d-Pro

For prokaryotic expression of HML-2 Pro, nt 3276–3768 from cloned proviral locus HERV-K(HML-2.HOM) (GenBank acc. no. AF074086.2 [134]), cloned in pBluescript II vector and available in-house, were amplified by PCR using the particular plasmid DNA as template. The forward primer (HOMProFOR) added an *Nhe*I site and the reverse primer (HOMProREV) added a stop codon and a downstream *Bam*HI site to the PCR product. The PCR product was cloned into pGEM-T Easy vector giving rise to plasmid pGEM HOM Pro (for pET11d). pGEM HOM Pro (for pET11d) was used as DNA template for generation of protease mutants by PCR site-directed mutagenesis (see 3.2.1.9). To clone the actual expression plasmid, the insert was released from pGEM HOM Pro (for pET11d) and respective mutated variants (pGEM Pro mut DTG (for pET11d), pGEM Pro-mutCGA (for pET11d)) by an *Nhe*I/*Bam*HI digest and cloned into *Nhe*I/*Bam*HI-digested pET-11d vector giving rise to plasmids pET11d Pro, pET11d Pro-mut DTG and pET11d Pro-mutCGA. For screening of clones, presence and orientation of the insert was checked by digestion with *Msc*I and subsequent visualization

of expected DNA fragments by agarose gel electrophoresis. A schematic representation of HML-2 Pro construct cloned into plasmid for Pro expression in *E. coli* is shown in Figure 7.
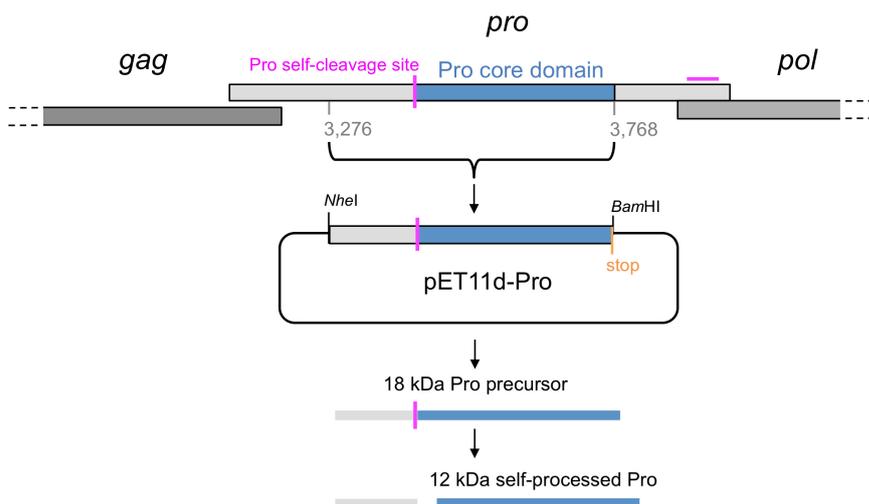
Figure 7: Schematic representation of HML-2 Pro cloned into pET11d vector for Pro expression in *E. coli*. In the upper part, the Pro ORF is depicted in the context of the HML-2 proviral genome. The region corresponding to the Pro core domain is colored in blue. A known Pro self-cleavage site in the N-terminus is indicated by a vertical violet bar. The precise location of a Pro self-cleavage site in the C-terminus is not known, thus a violet horizontal bar indicates the region where such a cleavage might occur. The Pro ORF subregion cloned into into pET-11d (see also Figure 9) is indicated with a bracket. Nucleotide numbering is as reported for the HERV-K(HML-2.HOM) provirus in GenBank acc. no. AF074086.2 [134]. The engineered stop codon (orange) and the restriction enzyme sites used for cloning are indicated. The Pro precursor expressed in *E. coli* and the products of Pro self-processing are depicted.

phCMV-Pro

For eukaryotic expression of HML-2 Pro, nt 3415-3946 from cloned proviral locus HERV-K(HML-2.HOM) (GenBank acc. no. AF074086.2 [134]), cloned in pBluescript II vector and available in-house, were amplified by PCR using the particular plasmid DNA as template. The forward primer (Pro3415F) added a *Bam*HI site, a spacer and a Kozak consensus sequence. The reverse primer (Pro3946R) added a *Bam*HI site. The PCR product was cloned into pGEM-T Easy vector giving rise to plasmid pGEM Pro (for phCMV). pGEM Pro (for phCMV) was used as DNA template for generation of protease mutants by PCR site-directed mutagenesis (see 3.2.1.9). The insert was released from pGEM Pro (for phCMV) and the respective mutated variant (pGEM Pro-mut DTG (for phCMV)) by a *Bam*HI digestion and cloned into *Bam*HI-digested phCMV-G vector, giving rise to plasmids phCMV-Pro and phCMV-Pro mut. For screening of clones, presence and orientation of the insert was checked by digestion with *Msc*I and subsequent visualization of expected DNA fragments by agarose gel electrophoresis. An HML-2 Pro construct cloned into plasmids for expression of HML-2 Pro in mammalian cells is depicted in Figure 8.

pEGFP-Pro

For eukaryotic expression of an EGFP-Pro fusion protein, Pro and Pro-mut coding sequences were amplified by PCR using pGEM Pro (for phCMV) and pGEM Pro-mut DTG (for phCMV) as DNA

templates. Forward primer (phCMV-GFPproFOR) and reverse primer (phCMV-GFPproREV) added a *Bam*HI site at the 5' and 3' end, respectively, of PCR products. PCR products were digested with *Bam*HI and cloned into *Bam*HI-digested pEGFP-C1 vector, in-frame with the EGFP ORF, giving rise to plasmids pEGFP-Pro and pEGFP-Pro-mut. For screening of clones, presence and orientation of the insert was checked by double-digestion with *Eco*RI-HF and *Msc*I (in NEB CutSmart® Buffer) and subsequent visualization of expected DNA fragments by agarose gel electrophoresis. An HML-2 Pro construct cloned into plasmids for expression of EGFP-fused Pro is depicted in Figure 8.
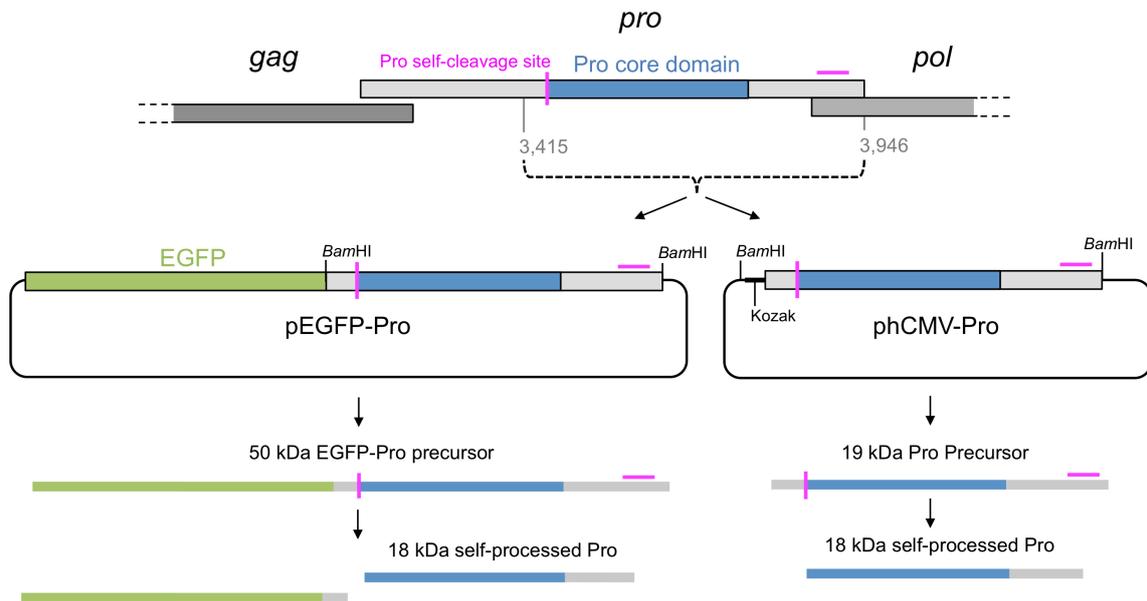
Figure 8: Schematic representation of HML-2 Pro construct cloned into different plasmids for expression of EGFP-fused Pro and Pro alone in mammalian cells. In the upper part, the Pro ORF is depicted in the context of the HML-2 proviral genome. The region corresponding to the Pro core domain is indicated in blue. A known Pro self-cleavage site in the N-terminus is indicated by a vertical violet bar. The precise location of the Pro self-cleavage site in the C-terminus is not known, a violet horizontal bar indicates the region where such a cleavage site might occur. A Pro ORF subregion (nt 3415-3946, indicated by a dotted bracket; see also Figure 9) from HERV-K(HML-2.HOM) provirus (GenBank acc. no. AF074086.2 [134]) was cloned into pEGFP and phCMV. The Kozak sequence (introduced by during PCR) and restriction enzyme sites used for cloning are indicated. The Pro precursors expressed in mammalian cells and the products of self-processed Pro are depicted.
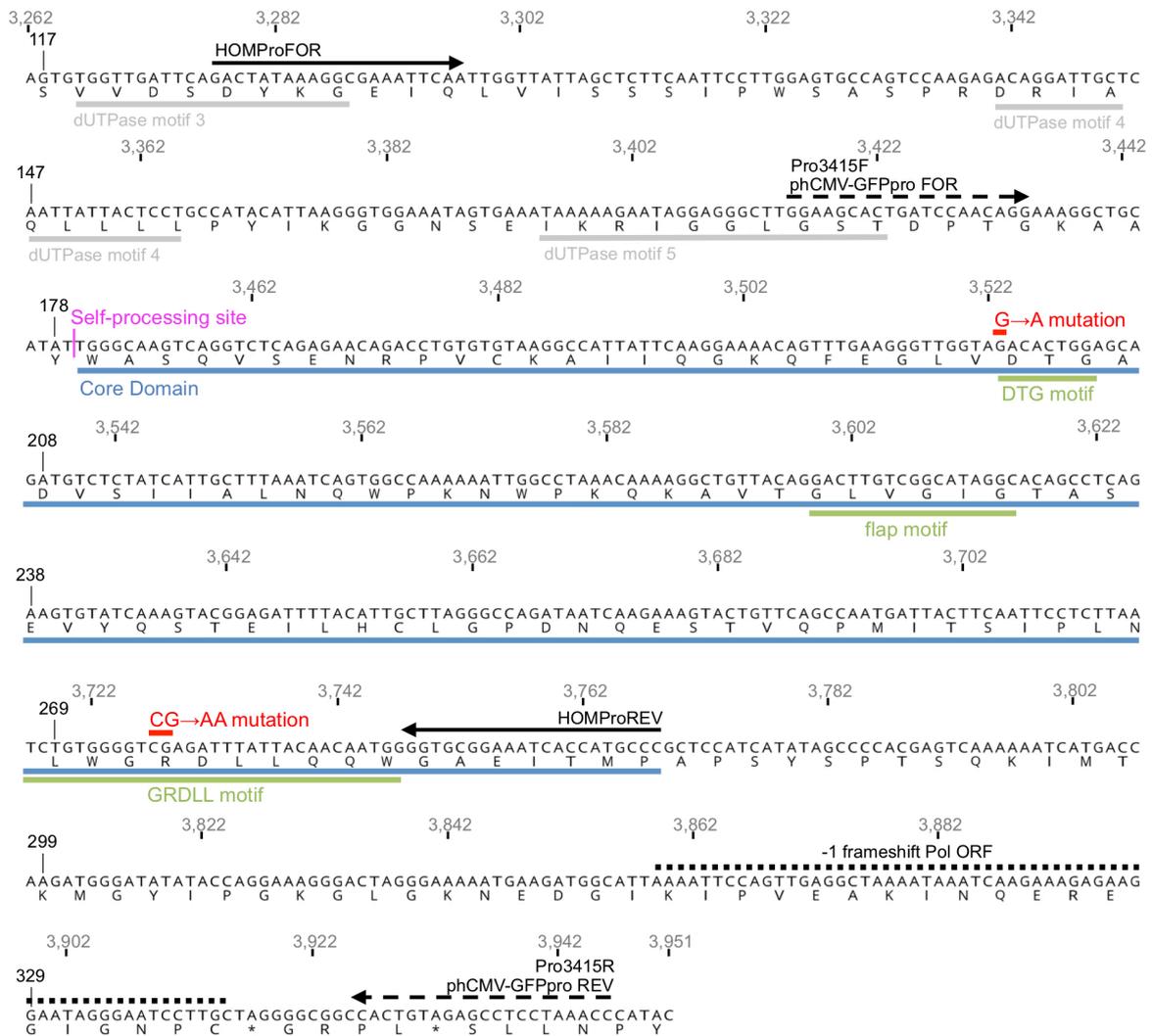
Figure 9: Pro ORF sequence used for generation of constructs for expression of HML-2 Pro. The DNA sequence, including nucleotide numbering, are based on the HERV-K(HML-2.HOM) provirus as reported in GenBank acc. no. AF074086.2 [134]. The corresponding aa sequence is given with numbering based on the Pro ORF. Forward and reverse primers used for generation of Pro constructs for Pro purification are indicated by solid line arrows. Primers for generation of Pro constructs for Pro expression in mammalian cells are indicated by dotted line arrows. Furthermore, Pro core domain (blue), a known self-processing site (violet), conserved catalytic motifs (green), dUTPase catalytic motifs 3-4 (grey), nucleotides mutated for generation of Pro mutants (red), and the frameshift region for translation of Pol ORF (dotted line) are indicated.

pcDNA3 plasmids for eukaryotic expression of HA-tagged candidate proteins

Full-length coding sequences for human proteins HSP90AA1, CIAPIN1, C15orf57, MAP2K2 and TUBA1A were present in plasmids purchased from GE Healthcare/Dharmacon. Clone identifiers of cloned coding sequences of candidate proteins are reported in Table 12. We employed those plasmids as DNA templates for PCR reactions amplifying full-length coding sequences of respective proteins. PCR primers were the same as those used for generation of DNA templates for *in vitro* translations of candidate proteins (see chapter 3.2.2.2). The forward primer added a Kozak consensus sequence. The reverse PCR primer added a human influenza hemagglutinin (HA) tag in-frame to the 3' end of the coding sequence, followed by a stop codon. The PCR product was cloned into pGEM-T Easy vector, released by a *Not*I digest and cloned into *Not*I-digested pcDNA3 vector, giving rise to plasmids

pcDNA3 HSP90AA1, pcDNA3 CIAPIN1, pcDNA3 C15orf57, pcDNA3 MAP2K2, and pcDNA3 TUBA1A. Presence and orientation of the insert was checked for each plasmid by digestion with restriction enzymes and subsequent visualization of expected DNA fragments by agarose gel electrophoresis. Plasmid pcDNA3 HSP90AA1 was double-digested with *Bgl*II and *Xho*I in NEBuffer 3.1; pcDNA3 CIAPIN1 was digested with *Hind*III; pcDNA3 C15orf57 was digested with *Bgl*II; pcDNA3 MAP2K2 was digested with *Pst*I; pcDNA3 TUBA1A was digested with *Eco*RI-HF.
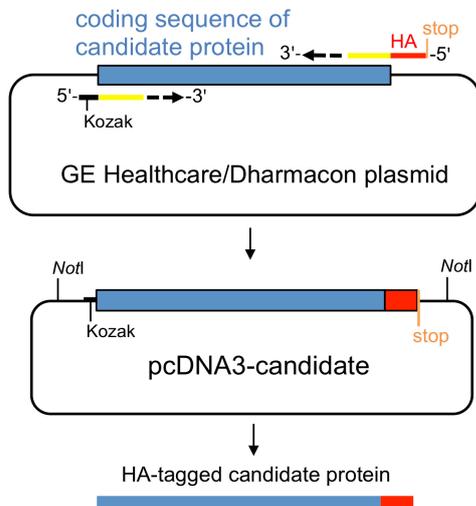
Figure 10: Schematic representation of the cloning strategy for generation of pcDNA3 plasmids for expression of HA-tagged candidate proteins in mammalian cells. Plasmids purchased from GE Healthcare/Dharmacon contained coding sequences of candidate proteins that were cloned into pcDNA3 including a Kozak sequence, an HA-tag and a stop codon. Primers used for the amplification of the coding sequence of candidate proteins are depicted. A *Not*I restriction site used for cloning into pcDNA3 is indicated. Corresponding protein expressed in mammalian cells after transfection of pcDNA3-candidate plasmids is depicted at the bottom.

pSG5-IN

For eukaryotic expression of HML-2 IN, nt 5592-6483 of cloned proviral locus HERV-K(HML-2.HOM) (GenBank acc. no. AF074086.2 [134]) were amplified by PCR. The forward primer (HOM-HA-IN-FOR) added a *Bam*HI site, a Kozak consensus sequence, a spacer and an HA-tag. The reverse primer (HOM-HA-IN-REV) added a stop codon and a downstream *Bam*HI site to the PCR product. The PCR product was cloned into pGEM-T Easy vector, giving rise to plasmid pGEM HOM-IN pSG5. pGEM HOM-IN pSG5 was used as DNA template for generation of a mutant integrase by PCR site-directed mutagenesis (see chapter 3.2.1.9). The insert was released from pGEM HOM-IN pSG5, as well as the respective mutated version (pGEM HOM-IN pSG5 Δ3aa), by a *Bam*HI digest and cloned into *Bam*HI-digested pSG5 vector, giving rise to plasmids pSG5 HOM-IN and pSG5 HOM-IN Δ3aa. For screening of clones, presence and orientation of the insert was checked by digestion with *Eco*RI-HF and subsequent visualization of expected DNA fragments by agarose gel electrophoresis.
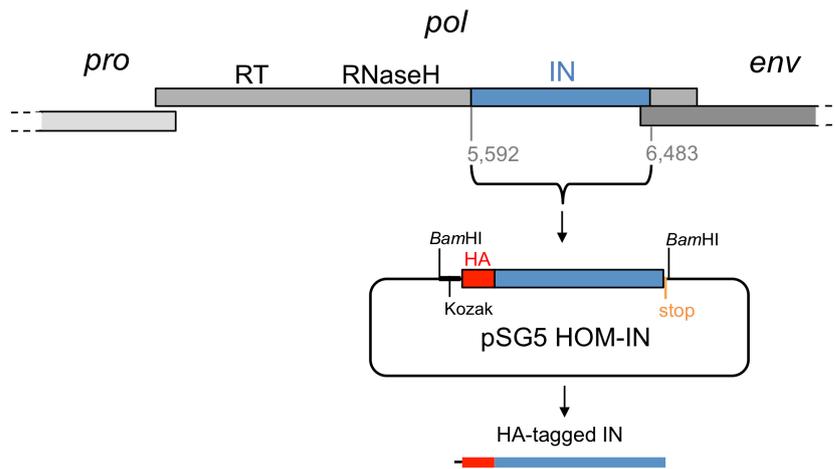
Figure 11: Schematic representation of HML-2 IN construct cloned into pSG5 vector for HML-2 IN expression in mammalian cells. In the upper part, the Pol ORF is depicted in the context of the HML-2 proviral genome. Based on previous studies concerning HERV-K10 IN [105], a subregion (nt 5592-6483, in blue) corresponding to the IN-encoding region of the HERV-K(HML-2.HOM) Pol ORF (GenBank acc. no. AF074086.2 [134]) was cloned into pSG5 expression vector. Kozak sequence (introduced by us) and restriction sites used for cloning are indicated. The HA-tagged IN protein expressed in mammalian cells after transfection of pSG5 HOM-IN is depicted at the bottom.

3.2.1.11 Verification of plasmid sequences

For each DNA construct generated in this study, the plasmid portion that included the protein-coding sequence of interest was verified by Sanger sequencing [178]. Sequencing was performed by Seq-IT GmbH (Kaiserslautern) utilizing an ABI 3730 DNA sequencer. One or two sequence reads per plasmid construct were usually generated and sufficient for sequence verification. Vector-specific sequencing primers were used, thus primers that bound vector sequence flanking the cloned portion. Sequence reads of sufficient quality were usually around 700 bp in length. For the purpose of sequence comparisons, sequence reads were aligned to reference sequences using Geneious software (Biomatters). Combinations of primers (for primer sequences see Table3) employed for sequencing of plasmid constructs were as follows:

Plasmid	Primers
pGEM-T Easy constructs	T7 + SP6
pET11d Pro	T7
pSG5 HOM-IN	T7
pEGFP-Pro	EGFP-C F + pEGFP-C.rev
pcDNA3 candidates	CMV fwd + SP6

3.2.2 Protein techniques

3.2.2.1 Protein concentration measurement

There are many techniques for measuring protein concentration. When the extinction coefficient is known for a protein of interest, a protein´s UV absorbance at 280 nm can be used to calculate the concentration of that protein. We have used that method to determine the concentration of purified HML-2 Pro, utilizing the extinction coefficient of HML-2 Pro as predicted by Geneious software (Biomatters). This method is rapid and does not require special reagents. Alternative assays are necessary to determine protein concentrations when there are interfering substances with absorbance at 280 nm and diverse proteins present in a protein solution. To determine protein concentration of cell lysates, we employed the Bio-Rad DC^{TM} Protein Assay Kit, a colorimetric assay based on the reaction of proteins with an alkaline copper tartrate solution and Folin reagent. The assay is similar to the well-documented Lowry protein assay [124] and furthermore compatible with various commonly used detergents used for cell lysis. As reported in the manual of the Bio-Rad DC^{TM} Protein Assay Kit, there are two steps which lead to color development. Proteins react with copper ions (Cu^{2+}) in an alkaline medium. Subsequently, copper-treated proteins reduce a Folin–Ciocalteu reagent by loss of 1, 2, or 3 oxygen atoms, thereby producing one or more of several possible reduced species which have a characteristic blue color with maximum absorbance at 750 nm and minimum absorbance at 405 nm. The assay was performed in 96-well microplates or 1.5 ml Eppendorf tubes. For each measurement assay, a protein standard with BSA was prepared in the same buffer as samples, as recommended by the manufacturer. Absorbance at 750 nm was measured using a Spark$^{®}$ 10M microplate reader (Tecan) or a NanoDrop^{TM} 2000 Spectrophotometer (Thermo Fisher Scientific).

3.2.2.2 Translation of proteins *in vitro* using the TNT T7 system

The TNT$^{®}$ T7 Quick for PCR DNA Kit (Promega), also referred to as TNT T7 system, is a coupled transcription/translation system that we used for *in vitro* production of HML-2 Gag protein, used as known substrate of HML-2 Pro, as well as for production of selected candidate proteins investigated for processing by HML-2 Pro. The TNT T7 system is able to produce proteins ranging in size from 10–150 kDa, starting from a DNA template generated by PCR.

The design of primers was as follows. The PCR primer portion complementary to the 5' end of the protein coding sequence was designed to have a melting temperature of approximately 58°C. Additional sequence elements were present in the PCR primers. In the forward primer, a T7 phage RNA polymerase promoter, flanked by a short spacer sequence on each side (the spacer in the 5´ terminus being a *Bam*HI restriction site), and a Kozak consensus sequence enabling efficient translation initiation, were present upstream of the protein coding-specific portion. The reverse primer added a sequence portion encoding an HA-tag and a stop codon downstream of the protein coding

sequence, effectively replacing the own stop codon of the protein coding sequence. Sequence elements included in PCR primers are shown in Figure 12.

Forward Primer

Spacer	T7 Promoter	Spacer	Kozak + start codon	cDNA-specific sequence

5´–GGATCC　TAATACGACTCACTATAGGG　AACAG　CCACC<u>ATG</u>　NNNNNNNN–3´

Reverse Primer

Stop	HA-tag	cDNA-specific sequence

5´–TTA　AGCGTAATCTGGAACATCGTATGGGTA　NNNNNNNNN–3´
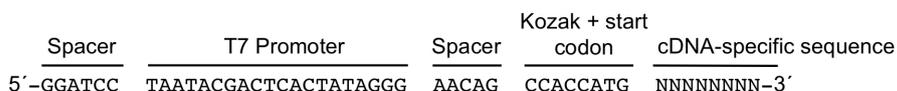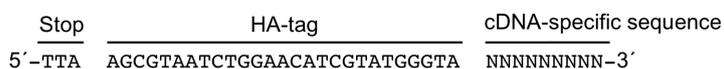
Figure 12: Sequence elements included in PCR primers for generation of DNA templates for translation of proteins *in vitro* by TNT T7 system. See the text for details.

PCR reactions were performed with *Taq* DNA polymerase employing conditions as reported in chapter 3.2.1.1. In those PCR reactions, the coding region of full-length protein was amplified from purified plasmid template DNA. The coding region of HML-2 Gag was cloned in pBluescript II vector and available in-house. Plasmids containing the coding region of candidate proteins were purchased from GE Healthcare/Dharmacon. Clone identifiers of cloned coding sequences of candidate proteins are reported in Table 13.

Table 13: Clone identifiers of cloned coding sequences of proteins investigated for processing by HML-2 Pro

Gene/protein name	Clone ID		Gene/protein name	Clone ID
C15orf57	2823236		MAP2K2	2961198
CALR	4299303		PDIA3	4712175
CIAPIN1	30389410		PSMC4	4046205
DDX3X	3617040		RANBP1	3935906
ENO1	4799584		RNASEH2B	5195426
HSP90AA1	40118488		S100A4	4247807
HSP90AB1	3621040		TUBA1A	6050536

Resulting PCR products were analyzed by agarose gel electrophoresis to verify amplification of desired PCR products and no byproducts present at considerable amounts. Desired PCR products were directly used as DNA templates for the coupled transcription/translation reaction. Reactions were set up as reported in Table 14, then incubated for 90 min at 30°C and frozen at -20°C immediately afterwards. Translation-grade L-^{35}S-methionine (370 MBq, 10 mCi/ml; Hartmann Analytic, Braunschweig, Germany) was employed in the reaction for generation of radiolabeled proteins, whereas "cold" L-methionine was included for generation of non-radiolabeled proteins.

Table 14: Reaction for generation of proteins *in vitro* by using a TNT T7 system

2.5 μl	PCR generated DNA template
22 μl	TNT® T7 PCR Quick Master Mix
0.5 μl	L-^{35}S-methionine (370 MBq, 10 mCi/ml) or L-methionine (1mM)

3.2.2.3 Prokaryotic expression and purification of HML-2 Pro

pET expression system and BL21 (DE3) competent cells

The pET expression system (Agilent) was used for prokaryotic expression of HML-2 Pro from the provirus HERV-K(HML-2.HOM). pET-11d (Figure 13) is a 5.7 kb plasmid, having as main features a T7 promoter with a lac operator (lac O), a ribosome binding site (RBS), a β-lactamase ORF for ampicillin resistance, a pBR322 origin of replication (which result in low copy number), and a lacI repressor ORF. The protein coding sequence of interest is cloned downstream of, and in-frame with, the RBS. pET-11d plasmid constructs were propagated in BL21 (DE3) Competent *E. coli* cells (NEB). BL21 (DE3) is an engineered strain of *E. coli* that carries a gene coding for the lac repressor (LacI) and a gene coding for the T7 RNA polymerase under control of a lacUV5 promoter. When pET-11d is transformed into BL21 (DE3), transcription of the gene of interest cloned in pET-11d is prevented via two mechanisms. First, LacI interacts with the lacUV5 promoter, repressing expression of T7 RNA polymerase (the gene of interest is transcribed from a T7 promoter, thus in absence of T7 RNA polymerase its transcription cannot take place). Second, LacI interacts with the lac operator located upstream of the gene of interest in the pET-11d vector. Binding of LacI prevents transcription of the gene of interest by formation of DNA structures that occlude passage of T7 RNA polymerase. Isopropyl-1-thio-β-D-galactopyranoside (IPTG) is a structural mimic of lactose that is able to bind LacI causing conformational changes that reduce its affinity for DNA [43]. Thus, IPTG is well suited for inducing T7 promoter-mediated target gene expression by unlocking the two repression mechanisms mentioned above. Moreover, IPTG is not degraded or otherwise utilized by *E. coli* cells, ensuring a stable IPTG level over hours during which the target protein will be produced, eventually comprising the majority of cellular proteins.
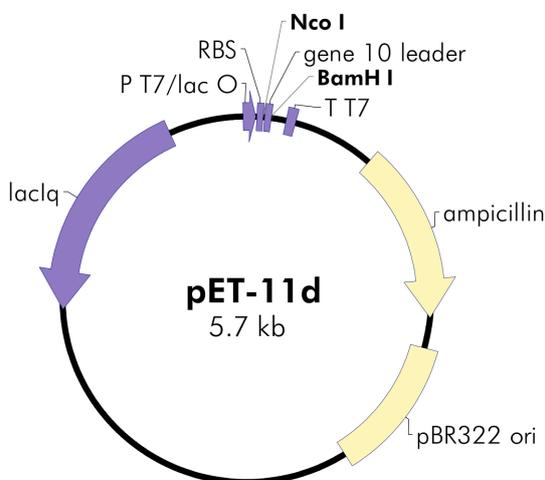
Figure 13: pET-11d plasmid map. Retrieved from Agilent pET System Vectors instruction manual (211521-12).

HML-2 Pro expression in BL21 (DE3) *E. coli*

50 μl of BL21 (DE3) *E. coli* cells (transformation efficiency 1–5 x 10^7 cfu/μg pUC19 DNA) were carefully mixed with 100 ng of pET11d Pro plasmid (see chapter 3.2.1.10.2) and incubated on ice for 30 min. The mixture was then subjected to a heat shock at 42°C for 90 sec and subsequently placed on ice for 5 min. Transformed cells were directly spread onto an agar-plate with ampicillin at standard concentration and incubated overnight at 37°C. Then, a single colony of *E. coli* cells was inoculated into 100 ml of LB$_{amp}$ and incubated overnight at 37°C. 20 ml of the overnight culture was then inoculated into 1 L LB$_{Amp}$ and incubated at 37°C until the absorbance at 600 nm had reached 0.6. Expression of HML-2 Pro was then induced by addition of IPTG at a final concentration of 0.4 mM. After 3 h at 37°C, bacterial cells were pelleted by centrifugation at 6800 g for 30 min at 4°C and the bacterial pellet was stored at -80°C.

Lysis of bacteria and protein recovery from inclusion bodies

High level expression of recombinant protein in *E. coli* often results in accumulation of expressed protein molecules in aggregates called inclusion bodies [191]. To recover the protein of interest, it was thus necessary to extract, isolate and solubilize inclusion bodies before the actual process of Pro purification by chromatography. Compositions of buffers employed for protein recovery from inclusion bodies are reported in chapter 3.1.2. The bacterial pellet derived from 1 L of cell culture, stored at -80°C until further processing, was resuspended in 50 ml of pre-cooled 5x TE Buffer and passed 3-4 times through a glass Dounce homogenizer kept constantly on ice. The lysate mixture was then subjected to sonication on ice, using a Branson Ultrasonics™ Sonifier S-250A (program: 10 x 10 sec, duty cycle 40%, output control 5). The cell lysate was centrifuged for 30 min at 3600 g and 4°C. During centrifugation, inclusion bodies mostly precipitated as part of the insoluble fraction of cell lysate. Thus, following centrifugation the pellet was subjected to subsequent processing steps, whereas the supernatant was discarded. For the solubilization step, the pellet containing inclusion bodies was washed twice with 20 ml of 5x TE Buffer each and then dissolved in 100 ml of Solving Buffer. During this process, high concentration of chaotropes, such as urea, results in disruption of protein structure making necessary a refolding process by removal of the solubilization agent. Removal of urea with accompanying refolding of proteins was achieved by dialysis. Dialysis is the separation of molecules in solution, based on their differential diffusion through a semipermeable membrane. For this purpose, we employed SnakeSkin™ Dialysis Tubing (10K MWCO, 22 mm) (Thermo Fisher Scientific) composed of regenerated cellulose. A piece of membrane of a suited size was rehydrated in H$_2$O$_{dd}$ and filled with the protein mixture solubilized in Solving Buffer. The dialysis tubing was sealed at both ends with tubing clips and floated in Dialysis Buffer using an appropriately sized container. Dialysis Buffer was kept in constant slow agitation using a magnetic stirrer. The protein mixture was dialyzed against 4 L of Dialysis Buffer at 4°C for 3 h and then against 4 L of fresh Dialysis Buffer at 4°C overnight.

HML-2 Pro purification by affinity chromatography

Chromatography is a technique for separation of specific components of a homogeneous mixture. In molecular biology, a protein mixture can be fractionated using various chromatography methods, based on physical and chemical properties of biomolecules. Affinity chromatography is based on specific interactions that can occur on protein surfaces. In general, a solution of proteins (mobile phase) is mixed with a porous solid matrix (stationary phase) covalently coupled with a ligand, for instance a substrate, an inhibitor or an antibody. The two phases are packed in a chromatography column and subjected to wash steps. Proteins specifically interacting with the ligand are retained in the matrix, all other proteins and molecules are washed out. The protein-ligand interactions are then disrupted through an elution step that induces the release of proteins bound. In this way, desired proteins can be isolated in a pure and concentrated form. Being a sensitive method, affinity chromatography requires considerable accuracy; only fine-tuned conditions guarantee a high yield and high purity of the desired protein.

We employed affinity chromatography for purification of HML-2 Pro from the HERV-K(HML-2.HOM) provirus. In our experiment, the mobile phase was a mixture of proteins recovered from inclusion bodies and derived from a lysate of bacteria overexpressing HML-2 protease (see above). The stationary phase consisted of agarose beads coupled to ligand pepstatin A, a specific and reversible aspartyl protease inhibitor. Our procedure followed a previously described protocol with minor modifications [112]. Composition of buffers employed for HML-2 Pro purification are reported in chapter 3.1.2.

The dialyzed solution prepared through the procedure for protein recovery from inclusion bodies (see previous paragraph) was centrifuged for 30 min at 6800 g and 4°C to remove precipitated proteins and then mixed at equal volumes with Buffer A. High salt concentration and low pH of Buffer A were two important parameters required for the next step, specifically interaction between pepstatin A and the mature HML-2 Pro during affinity chromatography. 5 ml of pepstatin A-agarose (Sigma) suspension (referred to as matrix in the following) were washed with 50 ml H_2O_{dd} and then 50 ml Buffer A, and subsequently added to the dialyzed protein solution previously mixed with Buffer A. The resulting matrix-protein solution was incubated overnight at 4 °C with slow agitation. The protein-matrix solution was then centrifuged for 15 min at 980 g and 4°C. The supernatant (representing the flow-through), containing most of the unbound proteins, was removed. The matrix was resuspended in 45 ml of Buffer A (Wash 1) and packed onto a chromatography column that was pre-conditioned with Buffer A. After Wash 1, another 30 ml of Buffer A were loaded onto the column in order to wash the matrix a second time (Wash 2). Proteins remaining in the column up to this point, i.e. proteins bound to pepstatin A-agarose, were eluted from the matrix with Buffer B, which has a higher pH and lower ionic strength than Buffer A (see section 3.1.2, compare pH and NaCl concentration of the two buffers). In fact, as binding of pepstatin A to HIV-1 Pro is pH-dependent and occurs at low pHs but is reduced above pH 7 [162], one can assume the same for HML-2 Pro. Six elution fractions of 5 ml each were

collected by gravity-flow. Aliquots were collected during the various purification steps in order to monitor collected aliquots by SDS-PAGE and subsequent protein visualization by Coomassie-staining of PAA-gels. Protease-containing elution fractions were pooled and concentrated using an Amicon® Ultra-15 Centrifugal Filter Unit (3000 MWCO). Protein concentration was achieved by centrifugation for 40 min at 4000 g and 4°C, to reach a final volume of approximately 2 ml. Protease concentration was determined utilizing DC^{TM} Protein assay (Bio-rad) and UV spectrophotometry using a molar absorption coefficient of 29115 M^{-1} cm^{-1} s (see also chapter 3.2.2.1). The concentrated protein solution was aliquoted and stored at -80°C.

3.2.2.4 *In vitro* enzymatic fluorescence assays for optimization of HML-2 Pro activity

Fluorophores are substances that absorb energy and re-emit part of that energy as radiation. When an exited fluorophore (donor) is in very close proximity (closer than approximately 10 nm) to a molecule (acceptor) whose absorption spectrum overlaps with the emission spectrum of the donor, a transfer of energy between the two molecules can occur, thus quenching emission from the fluorophore. Fluorescence Resonance Energy Transfer (FRET) is a technique based on this phenomenon that can be used to determine whether two molecules are within a certain distance to each other [127]. In fact, FRET is useful in assays measuring activity of a protease. In such assays, protease activity can be monitored using substrates in which a fluorophore is separated from a quencher by a short peptide sequence containing a protease-specific cleavage site. The assay that we established for monitoring and optimization of HML-2 Pro activity took advantage of the FRET technology. The molecule that we used as substrate of HML-2 Pro is 2-aminobenzoyl-Thr-Ile-Nle-p-nitro-Phe-Gln-Arg-NH_2 (BACHEM, also referred to as Anthranilyl-substrate). The Anthranilyl-substrate harbors a fluorescent 2-aminobenzoic acid (or anthranilic acid) in place of the acetyl group as donor and a p-NO_2-Phe at the P1' position as acceptor (Figure 14). Cleavage of the peptide by Pro releases the fluorescent N-terminal tripeptide from its close apposition to the quenching nitrobenzyl group, resulting in enhanced fluorescence (Figure 14).

Anthranilyl-substrate was incubated with HML-2 Pro under various reaction conditions as described in the Results section. Reactions were set up in 96-well microplates (Greiner Bio-One 655087) with a 50 µl final volume each. Fluorescence measurements were taken using a Tecan Infinite® M200 spectrophotometer, with excitation at 280 nm and emission measured at 420 nm. Reactions were monitored at 37°C by detecting the fluorescent emission at 420 nm for each reaction condition every 4 min and a total of up to 180 min. When plotting Pro activity, values of fluorescence emission at 420 nm were corrected by subtracting background fluorescence emission at 420 measured in parallel from control reactions without Pro.
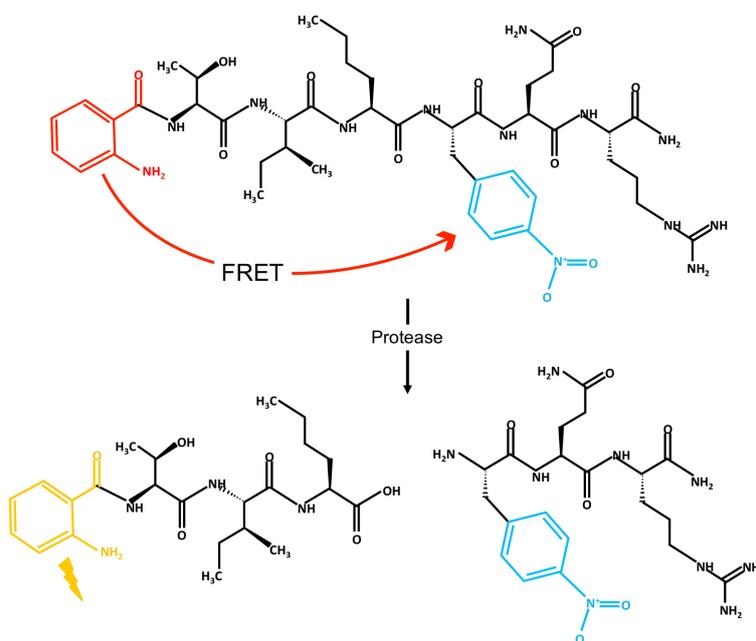
Figure 14: Molecular structure of the Anthranilyl substrate. The molecular structure of the Anthranilyl-substrate (2-aminobenzoyl-Thr-Ile-Nle-p-nitro-Phe-Gln-Arg-NH$_2$) is shown before (top) and after (bottom) cleavage by Pro. The 2-aminobenzoic acid (donor), depicted in red, is quenched by the nitrobenzyl group (acceptor) depicted in blue. For the N-terminal tripeptide produced after cleavage by Pro, the 2-aminobenzoic acid depicted in yellow is not quenched.

3.2.2.5 Separation of proteins by SDS-PAGE

Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) is a method for protein separation under denaturing conditions. Prior to loading protein samples into a polyacrylamide gel (PAA-gel), protein samples are mixed with a sample buffer that contains reducing agents and sodium dodecyl sulfate (SDS). Protein samples are then exposed to high temperature (usually 70°C for 10 min or 90°C for 5 min) in order to denature proteins. Reducing agents, such as dithiothreitol (DTT), unfold polypeptides by reducing disulfide bonds between cysteine residues. SDS is an anionic detergent that binds unfolded proteins at a ratio of approximately 1 SDS molecule per two amino acids, thus further contributing to protein denaturation. SDS introduces negative charges that mask the intrinsic protein charge. The resulting SDS-protein complexes are linearized and negatively charged, thus during electrophoresis they migrate towards the anode upon application of an electrical field. The PAA-gels are usually composed of a layer with large sized pores (stacking gel) and a layer with smaller sized pores (running gel). The stacking gel allows the concentration of proteins at the interface with the running gel, where the actual separation start. During the protein migration through the running gel, larger proteins migrate more slowly through the gel pores than shorter ones. As a consequence, the PAA-gel behaves like a molecular sieve separating proteins by molecular weight. The sample buffer usually includes a dye, such as bromophenol blue, that helps tracking the migration of the protein sample through the gel. Moreover, a protein marker with proteins of known molecular weight is loaded in at least one gel lane as a reference for protein sizes.

For our experiments we employed an XCell SureLock™ Mini-Cell Electrophoresis System (Thermo Fisher/Novex). PAA-gels were prepared as shown in Table 15. Gels were manually cast into empty gel cassettes (Thermo Fisher/Novex). The desired gel porosity was achieved by varying the percentage of acrylamid/bis-acrylamid (Bio-Rad). The discontinuous NuPAGE Bis-Tris buffer system was employed in two variants, depending on the desired protein resolution. NuPAGE® MOPS SDS Running Buffer was used to resolve proteins of medium molecular weight. NuPAGE® MES SDS Running Buffer was used to resolve proteins of relatively low molecular weight. Samples were prepared by adding 4x NuPAGE® LDS Sample Buffer and DTT at 50 mM final concentration to protein mixtures. Before loading, samples were heated at 70°C for 10 min (as recommended by the manufacturer) and then briefly centrifuged to collect condensate. Samples containing proteins generated by the TNT T7 system were heated at 65°C for 15 min to reduce the amount of extra protein bands at higher molecular weight, as suggested by the TNT T7 system manual.

For monitoring HML-2 Pro during expression in *E. coli*, 1 ml aliquots of bacterial cell culture were pelleted by centrifugation for 5 min at 6800 g. Cell pellets were resuspended in 50 µl 5x TE Buffer. 12 µl of cell suspension were mixed with 4 µl of 4x NuPAGE® LDS Sample Buffer and incubated for 5 min at 95°C. After centrifugation, supernatants were loaded onto a PAA-gel and separated by SDS-PAGE. Precision Plus Protein Dual Color Standard (Bio-rad) was used as molecular weight standard for proteins (see chapter 3.1.6). Electrophoresis was performed at 180 V constant. To maintain proteins in a reduced state during electrophoresis, a NuPAGE® Antioxidant (Thermo Fisher Scientific) was added to the running buffer of the cathode electrophoresis chamber.

Table 15: Pipetting scheme for preparation of stacking and running gel used for SDS-PAGE [*]

Components	Stacking gel	Running gel		
	4%	10%	12%	15%
3.5x Bis-Tris buffer pH 6.8	0.5	1.42	1.42	1.42
30% Acrylamid/Bis solution (37.5:1)	0.25	1.67	2	2.5
H_2O_{dd}	1.02	1.91	1.58	1.08
10% [w/v] APS	0.02	0.025	0.025	0.025
TEMED	0.01	0.007	0.007	0.007

[*]3.5x Bis-Tris Buffer composition is: 52.32g Bis(2-hydroxyethyl) amino-tris(hydroxymethyl) methan (Bis-Tris) in 200 ml H_2O_{dd}, pH 6.5-6.8 (pH is adjusted with HCl). Acrylamide and methylenebisacrylamide are crosslinked through a radical polimerization that involves ammonium peroxodisulfate (APS) as radical initiator, and tetramethylethylenediamine (TEMED) as catalyst. Volumes are given in ml.

3.2.2.6 Western blot

Following separation by SDS-PAGE, proteins can be transferred, that is, blotted, from the PAA-gel to a solid support membrane. This opens up a spectrum of possibilities in terms of protein detection and analysis not feasible using a rather fragile gel matrix.

The transfer relies on the same electrophoretic principles that drive migration of proteins during separation in SDS-PAGE but is done using an electric field oriented perpendicular to the surface of the

gel. The membrane, placed between the gel surface and the positive electrode, can vary in porosity and is usually made of nitrocellulose or polyvinylidene difluoride (PVDF). PVDF has a high protein affinity but needs to be activated with methanol in order to create a positively charged membrane surface thus allowing for interaction with SDS-complexed, thus negatively charged proteins. During blotting, proteins are immobilized on the membrane, mirroring their original position in the gel. After transfer, desired proteins can be detected on the blot membrane using, for instance, specific antibodies. Before antibody probing, a blocking step is required to reduce non-specific binding of the antibody to protein-free areas of the membrane. 5% [w/v] BSA or 5% [w/v] non-fat dry milk, both diluted in Tris-buffered saline (TBS) or phosphate-buffered saline (PBS), are common blocking solutions. Probing by indirect immunodetection involves an unlabeled primary antibody which is specific for the target protein, followed by binding of a labeled secondary antibody directed against the primary antibody. Antibodies can be diluted in blocking solution, with their concentration potentially requiring optimization to achieve best results. Moreover, incubations with immunochemical reagents are followed by wash steps, fundamental to remove unbound antibodies that could otherwise cause high background signal. Detergents, such as Tween-20, are commonly included in standard washing solutions. In a final step, depending on the label present in the secondary antibody, a variety of systems can be employed for detection of the target protein. For instance, enzyme-based chemiluminescence has been extensively developed and in particular antibodies conjugated with horseradish peroxidase (HRP) are a convenient method of choice. For this method, hydrogen peroxide induces HRP to catalyze oxidation of luminol that results in blue emission at 450 nm as a reaction product. The light emission on the blot membrane corresponds to the position of the target protein and can be detected on X-ray film or through digital imaging. The light signal can be intensified using detection reagents that include, for instance, modified phenols that increase HRP activity. So-called enhanced chemiluminescence (ECL) is crucial for detecting proteins present at relatively low-levels and, more generally, when high sensitivity and accurate quantification are important.

In our Western blot procedure, following SDS-PAGE, proteins were transferred onto Hybond 0.2 μm PVDF membrane (Amersham/GE Healthcare) using an XCell II™ Blot module and 1x NuPAGE® Transfer Buffer with 0.1% [v/v] NuPAGE antioxidant and methanol added to a final concentration of 10% [v/v] when transferring one gel, or 20% [v/v] methanol when transferring two gels in one blot module. The PVDF membrane was activated by incubation in methanol for 30 sec, 30 sec wash in H_2O_{dd} and then equilibrated in 1x NuPAGE Transfer Buffer (including methanol) for 20 min. The blot elements were assembled in the gel membrane sandwich as shown in Figure 15. Transfer of proteins was performed at 30 V constant for 1 h.

Figure 1 | Blotting Pad
+ | Blotting Pad
| Filter Paper
| Transfer Membrane
| Gel
| Filter Paper
| Blotting Pad
| Blotting Pad
Cathode Core (-)

Figure 2 | Blotting Pad
+ | Blotting Pad
| Filter Paper
| **Transfer Membrane**
| **Second Gel**
| Filter Paper
| Blotting Pad
| Filter Paper
| **Transfer Membrane**
| **First Gel**
| Filter Paper
| Blotting Pad
| Blotting Pad
Cathode Core (-)
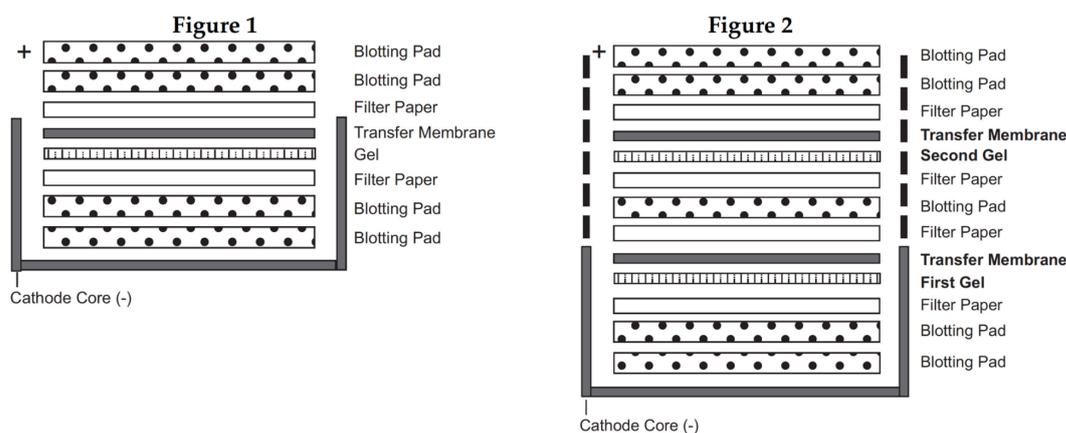
Figure 15: Assembly of blot elements in the XCell II™ Blot module. Blots were performed by transferring one (left) or two (right) gels in a blot module. Blotting pads and filter papers were soaked in transfer buffer before assembling the blot module. The gel surface was wetted with transfer buffer before positioning the pre-soaked transfer membrane on it. Figure taken from the NuPAGE® Technical Guide (Rev. date: 29 October 2010 Manual part no. IM-1001).

After transfer, the blot module was disassembled and the membrane was allowed to dry for 1 h at room temperature. The dried membrane was usually labelled in the upper right extremity using a soft graphite pencil to mark the side of the membrane containing proteins. When it was necessary to probe different parts of the membrane with different antibodies the membrane was cut accordingly using a surgical blade. Dry membrane was re-activated by incubation for 30 sec in methanol and 30 sec in H_2O_{dd}. The following steps, specifically blocking, probing and washing, were performed by putting the blot membrane in a 50 ml Falcon tube placed on a lab roller set to constant rotation. Blot membranes were blocked for 1 h in Blocking Buffer (see chapter 3.1.2 for buffer composition) and incubated overnight at 4°C with primary antibody, diluted in Blocking Buffer. Secondary antibody incubation was done using peroxidase-coupled antibodies diluted in Blocking Buffer, for 2 h at room temperature. Following primary and secondary antibody incubations, the membrane was washed three times for 10 min in 50 ml TBS-T (1x TBS, 0.05% [v/v] Tween-20) each. Signal detection was done using SignalFire™ Elite ECL Reagent (Cell Signaling Technology). 1x ECL reagent was prepared by combining equal parts of 2x Reagent A and 2x Reagent B, usually 250 µl of each, mixing, and then spreading of the solution on the membrane in a dropwise manner. The membrane was then sealed in a polypropylene envelope to facilitate equal distribution of the ECL reagent on the membrane surface. After 1 min incubation, excess ECL reagent was removed. ECL signal on membrane, still in the polypropylene envelope to prevent drying-out, was documented using a Chemidoc™ Imaging System (Bio-Rad). Subsequent image analysis utilized ImageLab 5.2.1 software (Bio-Rad).

3.2.2.7 Protein staining

For detection of proteins in PAA-gels following electrophoresis, PAA gels were briefly rinsed in H_2O_{dd} and incubated for 1 h with Coomassie Staining Solution (see chapter 3.1.2 for buffer composition) at gentle agitation. Subsequently, gels were washed thoroughly in Coomassie Destaining Solution (see

chapter 3.1.2 for buffer composition) at gentle agitation until complete removal of excess stain. Coomassie staining was also used for visualization of proteins on PVDF membranes. Since Coomassie staining is permanent, it was employed only after protein detection by ECL in the course of the Western Blot procedure. To do so, after ECL imaging, membranes were incubated for 2 min in Coomassie Staining Solution and subsequently washed for 2 min in Coomassie Destaining Solution. Membranes were left at room temperature for complete drying. Coomassie-staining of proteins on PVDF membranes was employed in some experiments to evaluate the transfer efficiency and as a loading control. Ponceau S is an alternative protein stain that is used to visualize proteins blotted onto a membrane. It is less sensitive than Coomassie-staining but with the advantage of being quick and reversible. Thus, following protein transfer, we used in certain contexts Ponceau S Solution (see chapter 3.1.2 for buffer composition) to visualize proteins in PVDF membranes in order to evaluate transfer efficiency and, when necessary, cut parts of interest of membranes for subsequent immunodetection.

3.2.2.8 Immunofluorescence microscopy

The combination of immunostaining and fluorescence microscopy provides a powerful methodology for studying the localization and distribution of specific proteins inside the cell. We employed the immunofluorescence (IF) technology to investigate potential effects of HML-2 integrase on DNA stability and in experiments addressing HML-2 protease-induced cell death. In the following, IF procedures employed in the presented work will be described in general. To perform IF, cells were seeded onto a proper surface that allowed for efficient cell adhesion and no subsequent interference with the microscopic light source. In our experiments, we seeded approximately 2×10^4 cells/well into 8-well cell culture slides with a removable chamber frame (Sarstedt). Following treatments, cells were rinsed briefly with 1x DPBS (Gibco/Life Technologies) and incubated for 15 min in 300 µl/well fixing solution (2% [v/v] formaldehyde in 1x DPBS, pH 7.4). After three wash steps with 500 µl/well 1x DPBS each, cells were permeabilized by incubation for 5 min in 300 µl/well permeabilizing solution (0.2% [v/v] Triton X-100 in 1x DPBS). During permeabilization, 8-well slides were placed on ice. Three washes in 500 µl/well 1x DPBS preceded the blocking step, done by incubation for 60 min in blocking solution (1% [w/v] BSA in 1x DPBS). Similar to the Western blot procedure, the blocking step minimized unspecific binding of antibodies. Subsequently, cells were incubated overnight at 4°C with primary antibody diluted in blocking solution. During this incubation, the 8-well slide was placed in a humidified chamber to reduce evaporation. Excess unbound antibody was removed by three washes in 500 µl/well 1x DPBS, 10 min each. Secondary antibodies were diluted in blocking solution, added to the cells and incubated for 1 h at room temperature in the humified chamber. During incubations with antibodies the 8-well slide was kept in the dark. Furthermore, handling was done in a darkened environment when antibodies were conjugated with fluorescent dyes. After removal of

secondary antibody solution and three 10 min washes in 1x DPBS, the 8-chamber frame was carefully detached from the slide and slides were mounted with Vectashield® Mounting Medium (Vector Laboratories), containing 4´,6-diamidino-2-phenylindole (DAPI) which fluoresces with an emission maximum at 461 nm (blue) when bound to double-stranded DNA.

Cells were analyzed with an Olympus AX70 fluorescence microscopy system, equipped with a filter set for visualization of DAPI, FITC, Texas-Red and infrared. Images were captured with a camera and analyzed using CellSens Imaging Software (Olympus, Tokyo, Japan). When expressing EGFP or recombinant proteins tagged with EGFP, the EGFP signal was, in certain experimental contexts, observed in living cells (without immunostaining) using an Axiovert 25 CFL inverted microscope (Carl Zeiss GmbH) equipped with reflected light fluorescence illumination.

3.2.2.9 Terminal amine isotopic labeling of substrates (TAILS)

Terminal amine isotopic labeling of substrates (TAILS) [68] is a quantitative proteomics method for labeling and isolation of N-terminal peptides. Proteolysis of proteins generates shorter peptides displaying new N-terminal ends (N-termini). Therefore, TAILS can be employed as a strategy to identify substrates of a protease of interest through the comparison between N-termini isolated from a protease-treated sample and N-termini isolated from an untreated control sample [114]. N-termini are labeled and blocked. Several processing steps result in purification and concentration of blocked N-terminal peptides. Liquid chromatography-tandem mass spectrometry (LC-MS/MS) enables the analysis and quantification of N-terminal peptides and neo-N-terminal peptides, providing their relative quantities in the treated and untreated samples. Following MS/MS, various bioinformatic tools are employed to identify actual human proteins having generated particular peptide fragments, actual cleavage positions within proteins, protein length, protein features, etc., in order to compile relevant information on substrate proteins.

In our work, TAILS was employed as a search strategy for the identification of potential substrates of HML-2 Pro. Parts of the TAILS procedure were performed in our laboratory. A protein mixture derived from lysis of HeLa cells was incubated in presence of purified HML-2 Pro at experimental conditions as described in the Results section. Incubations in the presence of 200 μM pepstatin A, effectively blocking HML-2 Pro activity, were performed as control reactions without HML-2 Pro activity. Treated samples and respective controls were sent to the laboratory of Prof. Dr. Oliver Schilling (University of Freiburg), where subsequent parts of the TAILS analysis were performed using well-established protocols [106]. Data generated by the TAILS procedure were analyzed in our laboratory using Microsoft Excel (Office 2016) and other analysis tools as described in the Results section. Datasets resulting from TAILS analyses were provided to us in tabular format. Various data filtering and data overlap functions were used for analysis of data and extrapolation of results. Lists of protein IDs were used for Gene Ontology (GO) analysis employing the PANTHER (Protein Analysis Through

Evolutionary Relationships) classification system [140]. Protein IDs were also used for retrieving respective gene names from UniProt [40]. Lists of gene names were used for intersections with the Catalogue Of Somatic Mutations In Cancer (COSMIC) [60] and the Online Mendelian Inheritance in Man (OMIM) database [2].

Preparation of cellular proteins from HeLa cells

For preparation of cellular proteins for incubation with HML-2 Pro, a total of 1.4×10^8 HeLa cells grown to near confluence in eight 160 cm^2 tissue culture flasks were washed with 1x DPBS and detached by trypsinization (see also chapter 3.2.3.1). Cells were collected in 20 ml 1x DPBS per flask, pelleted for 5 min at 250 g, resuspended in 0.5 ml of 5 mM MES pH 6.0 supplemented with protease inhibitors (cOmplete, Mini, EDTA-free, Roche) at the recommended concentration and subjected to cell lysis by three freeze-thaw cycles. Each freeze-thaw cycle was performed as follows: 15 min incubation at -80°C, 10 min at room temperature and incubation on ice until complete thawing. During the first incubation on ice, glass beads were added to the samples. During each incubation on ice, samples were mixed by vortexing for 5 sec every 5 min. Protein lysates were centrifuged at 4°C for 30 min at 16.100 g. Supernatants were pooled in a fresh tube and protein concentration was measured. HeLa protein lysate was stored in aliquots at -80°C.

Preparation of TAILS reactions

For preparation of TAILS reactions, 2 mg of HeLa total proteins were incubated with 200 nM HML-2 Pro in a 2 ml reaction volume. Reaction buffer contained 1 M NaCl, 2% [v/v] DMSO, 100 mM PIPES at pH 5.5 or pH 7. Untreated controls included 200 μM pepstatin A. To promote Pro inhibition by pepstatin A, samples were pre-incubated for 10 min at room temperature before addition of HeLa total protein lysate as the last reaction component. All reaction mixtures were incubated at 37°C for 75 min. Reactions were then frozen at -80°C. Samples were shipped on dry-ice to the laboratory of Prof. Schilling for subsequent processing steps and tandem mass-spectrometry towards the identification of cellular proteins cleaved by HML-2 Pro.

3.2.3 Manipulation of human cells

3.2.3.1 Culturing of human cell lines

We employed various adherent human cell lines in our studies (see chapter 3.1.7). All cell lines were cultured at 37°C and 5% [v/v] CO_2 in a medium supplemented with 10% [v/v] heat-inactivated fetal calf serum (FCS), 50 μg/ml penicillin and 50 μg/ml streptomycin. For HeLa and HEK293T cells, Dulbecco's Modified Eagle's Medium (DMEM) was used. For Tera-1, SK-MEL-28, and MeWo cells, RPMI 1640 medium was used. For primary cell cultures, cells were cultured in 75 cm^2 flasks and split between one to two times per week depending on the growth rate of each cell line. Cells were split

when their confluence had reached approximately 80% of the total growth area. The procedure for splitting cell cultures was as follows: removal of the growth medium from the flask, washing of the cell monolayer with sterile 1x DPBS, addition of 2 ml of 0.05% Trypsin-EDTA (Gibco/Life Technologies) and incubation for 5-10 min at 37°C, until cells detached from the bottom of the flask. Cells were then collected by addition of 8 ml of fresh medium and completely detached by careful pipetting. An aliquot of the detached cells was transferred to a new flask containing fresh culture medium. Usual dilutions used to split various cell lines are given in chapter 3.1.7. Cells were double-checked under the microscope before further culturing them.

3.2.3.2 Determination of number of cells

To estimate the number of human cells in a cell suspension, after cell detachment by Trypsin-EDTA (see above), an aliquot of cell suspension was mixed at an equal ratio with 0.4% [w/v] Trypan Blue Solution (Sigma). Trypan blue is a negatively charged dye that is not absorbed by vital cells, the functional cell membrane of which selectively excludes trypan blue. In contrast, trypan blue can pass through the damaged membrane of dead cells. Trypan blue can therefore be employed as a marker of cell vitality. Detached cells were subjected to cell counting employing a Luna™ Automated Cell Counter (Logos Biosystem) that was compatible with trypan blue staining, thus numbers of live as well as dead cells in the cell suspension could be determined. Determination of numbers of cells was used as a reference to seed desired numbers of live cells into plates.

3.2.3.3 Transient transfection of cells

For transient expression of proteins of interest in mammalian cells, we transiently transfected cells utilizing FuGENE® HD Transfection Reagent (Promega; in the following indicated as FuGENE reagent), a nonliposomal preparation consisting of a mixture of lipids and other components that is suitable for transfection of DNA into a wide variety of cell lines, that furthermore comes with relatively little cell toxicity. We employed standard transfection conditions as suggested by the manufacturer. Specifically, cells were seeded on the day before transfection. Cells were plated at cell numbers so that they were approximately 80% confluent on the day of transfection. Plasmid DNA and FuGENE reagent were added to Opti-MEM™ Reduced Serum Medium (Gibco). Total volume, amount of DNA per well and amount of FuGENE reagent per well are given in Table 16. The ratio of FuGENE reagent (µl) and DNA (µg) was 3:1. Transfection mixes consisting of plasmid DNA, FuGENE reagent, and Opti-MEM medium were incubated for 15 min at room temperature to allow for formation of FuGENE reagent:DNA complexes. Transfection mixtures were added to wells in a drop-wise manner or pipetted into cell culture medium in case of smaller volumes. Transfected cells were incubated at 37°C, 5% [v/v] CO_2 and further processed depending on the experimental context.

Table 16: Conditions established for preparation of transfection mixes

Plate	Total transfection volume (per well)	Amount of DNA (per well)	Amount of FuGENE reagent (per well)
6-well plate	100 µl	2.5 µg	7.5 µl
12-well plate	50 µl	1 µg	3 µl
8-well cell culture slide	10 µl	0.2 µg	0.6 µl

3.2.3.4 Harvesting of cells

To harvest cells (from flasks or wells), we used the following procedure. Growth medium was removed by aspiration, the cell monolayer was washed with sterile 1x DPBS; 0.05% [v/v] Trypsin-EDTA were added (100 µl/well for 12-well plate, 500 µl/well for 6-well plate, 5 ml for T-160 flask) and incubated for 5-10 min at 37°C until detachment of cells from the culture surface. Cells were then collected by addition of fresh growth medium and careful pipetting for complete cell detachment. Cell suspension was then transferred into a tube and centrifuged for 5 min at 300 g. Supernatant was removed, cell pellet was washed with 1x DPBS and centrifuged again. Cell pellets were stored at -80°C or directly used for subsequent procedures.

3.2.3.5 Fluorescence-activated cell sorting (FACS)

Fluorescence-activated cell sorting (FACS) is used to profile cells in a heterogeneous cell suspension. A flow cytometer is used to perform FACS. In a flow cytometer, employing a laser-based technology [22], individual cells can be analyzed and grouped in distinct sub-populations on the basis of physical and chemical characteristics. For such analysis, it is important that samples are prepared in a way that cells are properly separated from each other. In a flow cytometer, cells flow one by one across a laser beam that, in association with detectors, measures optical properties correlated with cellular characteristics. For instance, detecting light scattered at different angles gives indications about cell morphology, e.g., forward scatter (FSC) correlates with cell size and side scatter (SSC) is proportional to cell granularity. Fluorescence emission can be measured for gating cells that contain proteins labeled with a fluorophore. Cell count, cell cycle evaluation, apoptosis measurements, and many other applications can be achieved with this powerful methodology.

In our work, FACS was employed for evaluating cell death induced in a population of cells expressing HML-2 Pro fused with EGFP. For sample preparation, HEK293T cells were seeded and transfected in 12-well plates. After transfection, cells were collected at different time points. At each time point, cells were harvested by trypsinization and counted (see above). Each sample consisted of cells derived from two wells treated in the same way. 2×10^5 cells/sample were transferred into a fresh Eppendorf tube and washed twice with 500 µl 2% [v/v] FCS in 1x DPBS and centrifugation for 5 min at 300 g. Cell pellets were reconstituted in 150 µl 2% [w/v] paraformaldehyde in 1x DPBS, transferred to FACS tubes and stored at 4°C in the dark until measurement with a flow cytometer. We utilized in our studies a BD FACSCanto™ II flow cytometer with FACS Diva software version 6.1.3. FACS and data

analyses were performed with the help of Dr. Tina Schmidt and Dr. David Schub (Institute of Virology, University of Saarland).

3.2.3.6 Lysis of human cells

Lysis of HeLa cells under non-denaturing conditions for preparation of protein mixture for TAILS reaction, is described in 3.2.2.9. Lysis of human cells for whole protein extraction under denaturing conditions was performed as follows. After harvesting of cells from culture plates, cell pellets were frozen at -80°C and resuspended in RIPA Buffer (buffer composition in chapter 3.1.2) enriched with protease inhibitors (cOmplete, Mini, EDTA-free, Roche) at the recommended concentration, and pepstatin A at 1μg/μl. Pellets combined from two 12-well replicates were resuspended in 100 μl RIPA Buffer. Pellets from combination of two 6-well replicates were resuspended in 300 μl RIPA Buffer. Resuspended cells were incubated for 20 min on ice and mixed by vortexing for 5 sec every 10 min. Insoluble cell debris was pelleted by centrifugation at 13,000 g for 15 min at 4°C. Supernatant was transferred to a fresh Eppendorf tube and protein concentration was measured. Cell lysates were stored at -20°C.

3.2.3.7 Storage of human cells in liquid nitrogen and thawing

For long-term storage, eukaryotic cells were stored in a Biosafe®-MDβ (Cryotherm) liquid nitrogen tank. The freezing medium was composed of growth medium supplemented with 20% (v/v) FCS and 10% (v/v) dimethyl sulfoxide (DMSO). DMSO acts as a cryo-protective agent that prevents formation of ice crystals in the cell lumen thus avoiding damage of the cellular membranes. To prepare cell aliquots for cryo-storage, following cell detachment with Trypsin-EDTA (see above), cells were pelleted by centrifugation for 5 min at 300 g. The supernatant was removed and the cell pellet was resuspended very gently in freezing medium. Cells were aliquoted in 1 ml cryo-vials and frozen down using a CoolCell™ at -80°C that allowed for a constant cooling rate of approximately 1°C per minute. Frozen cells were then transferred to a Biosafe®-MDβ liquid nitrogen storage tank. Cell stocks stored in liquid nitrogen were used for starting a new cell culture of a particular cell line. To thaw cells after storage in liquid nitrogen, cell-containing cryo-vials were incubated in a water bath at 37°C until the freezing medium was completely molten, then transferred into a 75 cm^2 flask with growth medium pre-warmed to 37°C and incubated overnight at 37°C, 5% [v/v] CO_2. The following day, growth medium was replaced to remove residual DMSO and dead cells, and the surviving cells were further cultured.

4. Results

The present study focused on the biological impact of two viral proteins, protease (Pro) and integrase (IN), derived from the HERV-K(HML-2.HOM) provirus. HERV-K(HML-2.HOM) has been described for the first time in a study conducted in our laboratory and in the Department of Virology (University of Saarland, Homburg, Germany) [134]. The HERV-K(HML-2.HOM) provirus is located in chromosome 7 and represents one of the most intact HERVs identified. HERV-K(HML-2.HOM) harbors complete ORFs for all retroviral genes. The Pro and IN ORFs retained conserved amino acids within catalytic motifs. The encoded Pro was confirmed to be functional, being able to self-process and process the HERV-K(HML-2) Gag protein [134]. The IN ORF of HERV-K(HML-2.HOM) was found to be almost identical in sequence to the HERV-K10 IN, the latter known to be functional [105]. Also, the HERV-K(HML-2.HOM) provirus contains LTRs with intact sequence motifs for initiation and termination of transcription [134]. HERV-K(HML-2.HOM) transcripts were detected in malignant cell lines [183] as well as in normal human tissues [184], thus suggesting that proteins encoded by HERV-K(HML-2.HOM) might be expressed in various human cells.

In our work, the two viral proteins of interest, specifically HML-2 Pro and IN, were investigated employing different approaches. Most of our efforts were dedicated to the study of HML-2 Pro and the identification of its potential cellular substrates. The experimental approach was based on purification and further characterization of HML-2 Pro, the latter to further assess its behavior under defined reaction conditions. We then set up *in vitro* reactions with purified HML-2 Pro to perform state of the art proteomics analyses for identification of cellular substrates of HML-2 Pro. We used online protein/gene databases to identify biological properties and potential disease involvement of HML-2 Pro substrates. We verified *in vitro* and *in vivo* the processing of several selected human proteins by HML-2 Pro. Moreover, several observations prompted experiments in order to investigate an association between HML-2 Pro expression and cell death. Finally, we collected preliminary evidence of endogenous HML-2 Pro activity in selected tumor cell lines.

As for the experimental approach to investigate HML-2 IN, we set up an immunofluorescence assay for monitoring DSBs through the detection of 53BP1, a DSB marker. Following transient expression of HA-tagged wild-type and mutant HML-2 IN in cell culture, we evaluated, through the established immunofluorescence assay, whether HML-2 IN activity induces formation of DSBs in the cell genome. During the assay, visualization of transiently expressed HML-2 IN allowed us to obtain information about its cellular localization.

In the following, results will be reported in two main chapters. Chapter 4.1 is related to HML-2 Pro and chapter 4.2 is related to HML-2 IN.

4.1 HML-2 protease

4.1.1 Expression of HML-2 Pro in *E. coli* and protein purification

For studies on HML-2 Pro, we first aimed at purifying HML-2 Pro under defined conditions that preserved enzyme integrity and guaranteed a relatively high protein yield. A DNA construct harboring a subregion of the wild-type HML-2 Pro ORF was cloned into prokaryotic expression plasmid pET-11d. The cloned construct (see Materials and Methods, Figure 7) included: 1) the HML-2 Pro core domain, analogous to the HERV-K10 core domain previously described [205]; 2) an N-terminal flanking region comprised of 171 bp upstream of the Pro core domain; 3) an engineered stop codon at the aa position that was previously reported as the C-terminal end of the Pro core domain [205]. The resulting protein, referred also as HML-2 Pro precursor, was expressed in *E. coli* BL21 (DE3) cells following induction by IPTG. Since a known Pro self-processing site was present in the HML-2 Pro precursor, a portion of 59 aa residues at the N-terminal end of HML-2 Pro precursor was expected to be cleaved off in an autocatalytic manner.

HML-2 Pro protein expression was confirmed by Western blot, employing α-HML-2 Pro polyclonal antibody 9367 [185]. Aliquots of *E. coli* BL21 (DE3) cells before and after addition of IPTG were examined. HML-2 Pro precursor appeared as a protein of approximately 18 kDa that already self-processed to the 12 kDa mature form during prokaryotic expression (Figure 16).
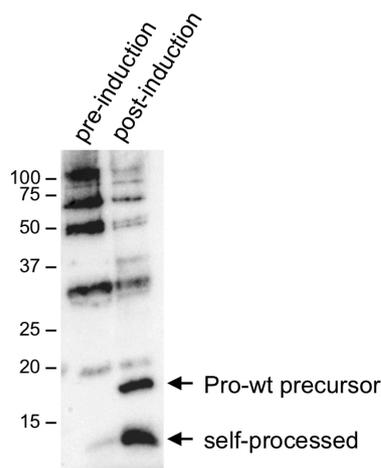
Figure 16: Successful expression of HML-2 Pro in BL21 (DE3) cells. Aliquots from cell culture of BL21 (DE3) cells harboring an expression plasmid encoding HML-2 Pro were collected before and 2 h after induction by IPTG. Cells were pelleted and resuspended in 50 µl 5x TE Buffer. 12 µl of cell suspension were mixed with 4 µl of 4x NuPAGE LDS Sample Buffer and incubated for 5 min at 95°C. After centrifugation, supernatants were loaded onto a 15% PAA-gel and separated by SDS-PAGE. HML-2 Pro was detected by Western blot using α-HML-2 Pro polyclonal antibody no. 9367 [185]. 18 kDa HML-2 Pro precursor and 12 kDa self-processed Pro are indicated.

HML-2 Pro was purified from inclusion bodies following a previously published protocol [112] with minor modifications. In our protocol, the cell lysis procedure included an additional homogenization step before sonication to improve cell lysis, and the chromatography procedure was performed by gravity, without use of a fast protein liquid chromatography system (see chapter 3.2.2.3). The

purification procedure employed pepstatin A, a specific, non-covalent inhibitor of retroviral aspartyl proteases, coupled to agarose beads as stationary phase for affinity chromatography. Monitoring of proteins during the purification steps (Figure 17) confirmed that contaminant proteins bound to the pepstatin A-agarose matrix were largely removed during the first wash step. Only one co-purifying protein of ~45 kDa was still detectable in the first elution step. That contaminant protein greatly reduced and eventually disappeared in the following elution steps. HML-2 Pro became eluted starting with the second elution step as a protein of ~12 kDa. The 18 kDa Pro precursor observed after induction in *E. coli* was not visible in eluates (Figure 17). This suggested that HML-2 Pro completely self-processed during the purification/renaturation procedure.
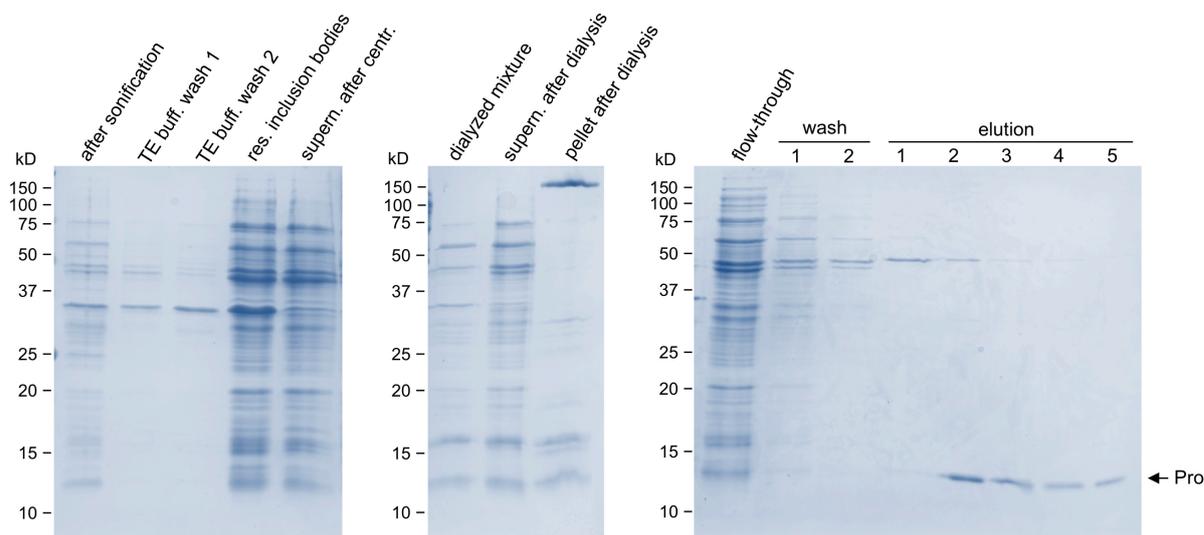
Figure 17: Purification of HML-2 Pro. To assess purification of prokaryotically expressed HML-2 Pro, samples were taken at various steps of the procedure and separated by SDS-PAGE. Proteins were visualized by Coomassie-staining of PAA-gels. Coomassie-stained gels are shown. Documented steps of the procedure are as follows: cell lysate after sonification; two washes of inclusion bodies in TE Buffer (TE buff. wash 1/2); resuspension (res.) of inclusion bodies in a urea-containing buffer; supernatant (supern.) after centrifugation of resuspended inclusion bodies; protein mixture after dialysis; supernatant and pellet after dialysis and centrifugation; flow through after incubation of dialyzed protein mixture with pepstatin A-agarose; two washes after binding; five elution fractions. Molecular masses of marker proteins are reported on the left each. Purified, self-processed HML-2 Pro migrated at ~12 kDa (see the label).

Based on Coomassie-stained proteins, eluates 3, 4, and 5 of purified HML-2 Pro showed almost complete absence of contaminant proteins (Figure 17). Thus, those eluates, with a total volume of 5 ml per eluate, were pooled together and further concentrated from a starting volume of 15 ml to a final volume of 2 ml using an Amicon® Ultra-15 Centrifugal Filter Unit. Aliquots of pooled eluates before and after concentration were subjected to SDS-PAGE for monitoring the concentration process and ensure lack of protein degradation (Figure 18). Molar absorption coefficient of self-processed HML-2 Pro, as predicted by Geneious software, was used to determine a Pro final concentration of 270 µg/ml, corresponding to a molar concentration of 23 µM. Aliquots of concentrated Pro were stored at -80°C.
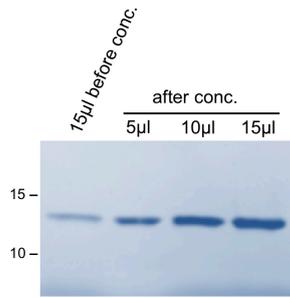
Figure 18: Concentration of purified HML-2 Pro. 15 µl of pooled eluate before concentration and 5 µl, 10 µl, and 15 µl of pooled eluate after concentration were separated by SDS-PAGE. Proteins were visualized by means of Coomassie-staining. Note that the ~12 kDa protein band corresponding to HML-Pro is enriched when compared to before concentration, and that no degradation products were visible.

Experiments to be performed with active wild-type HML-2 Pro required various controls including an inactive HML-2 Pro. For this purpose, two enzymatically inactive mutants of wild-type HML-2 Pro were generated by site-directed mutagenesis. As for the first mutant (named Pro-mutDTG), a single nucleotide mutation G→A was introduced causing an aa substitution D→N within the catalytic DTG-motif [180] (Figure 19). As for the second mutant (named Pro-mutGRDL), Wondrak et al. (1991) [221] had reported successful purification, employing pepstatin-A affinity chromatography, of an inactive HIV-1 protease mutated in the GRNLL region, specifically through an aa substitution R→K [221]. The HML-2 Pro ORF harbors an analogous GRDLL region, in which, similarly, we substituted R→K by mutating the corresponding triplet sequence CGA →AAA (Figure 19).
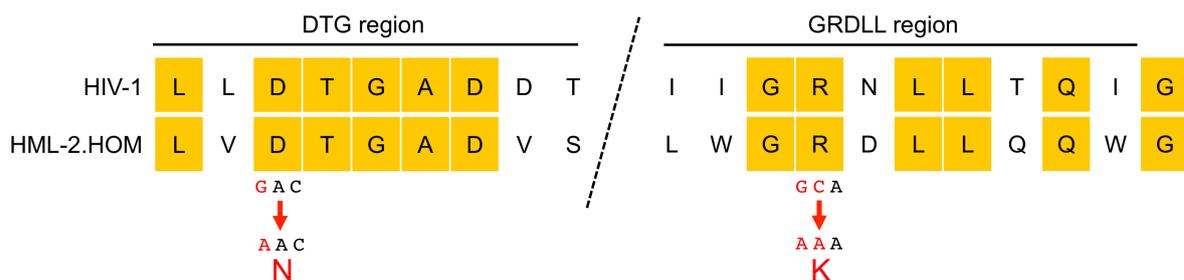
Figure 19: Amino acid sequence comparison between conserved regions of HIV-1 Pro and HERV-K(HML-2.HOM) Pro. The aa sequence of DTG and GRDLL regions of HIV-1 *pro* are compared with the corresponding regions of HERV-K(HML-2.HOM) *pro* [134]. Nucleotide exchanges and resulting aa substitutions thus generated in Pro mutants (see text) are indicated in red.

Pro-mutDTG and Pro-mutGRDLL proteins were then expressed in *E. coli* BL21 (DE3) cells, as described for wild-type HML-2 Pro. Just as for purification of wild-type HML-2 Pro, various steps of the purification of the two HML-2 Pro mutants were monitored by SDS-PAGE and Coomassie-staining (Figure 20). The two HML-2 Pro mutants could be expressed efficiently in *E. coli* BL21 (DE3) cells, with visible accumulation of the 18 kDa Pro precursor. However, neither Pro-mutDTG nor Pro-mutGRDLL could be purified when employing the protocol established for wild-type HML-2 Pro. This was most likely due to inefficient binding of the mutant proteins to pepstatin A-agarose. Strong accumulation of the 18 kDa Pro precursors but no visible processed form of 12 kDa clearly indicated

that both Pro mutants were unable to self-process. After incubation with pepstatin A-agarose, both mutants were present at considerable amounts in the flow-through and in the following wash steps, that is, they were eluted together with other unbound proteins. Apart from a ~45 kDa protein also observed during purification of wild-type HML-2 Pro (see above), no proteins were detected in eluate fractions, thus further confirming unsuccessful purification of the two HML-2 Pro mutants.
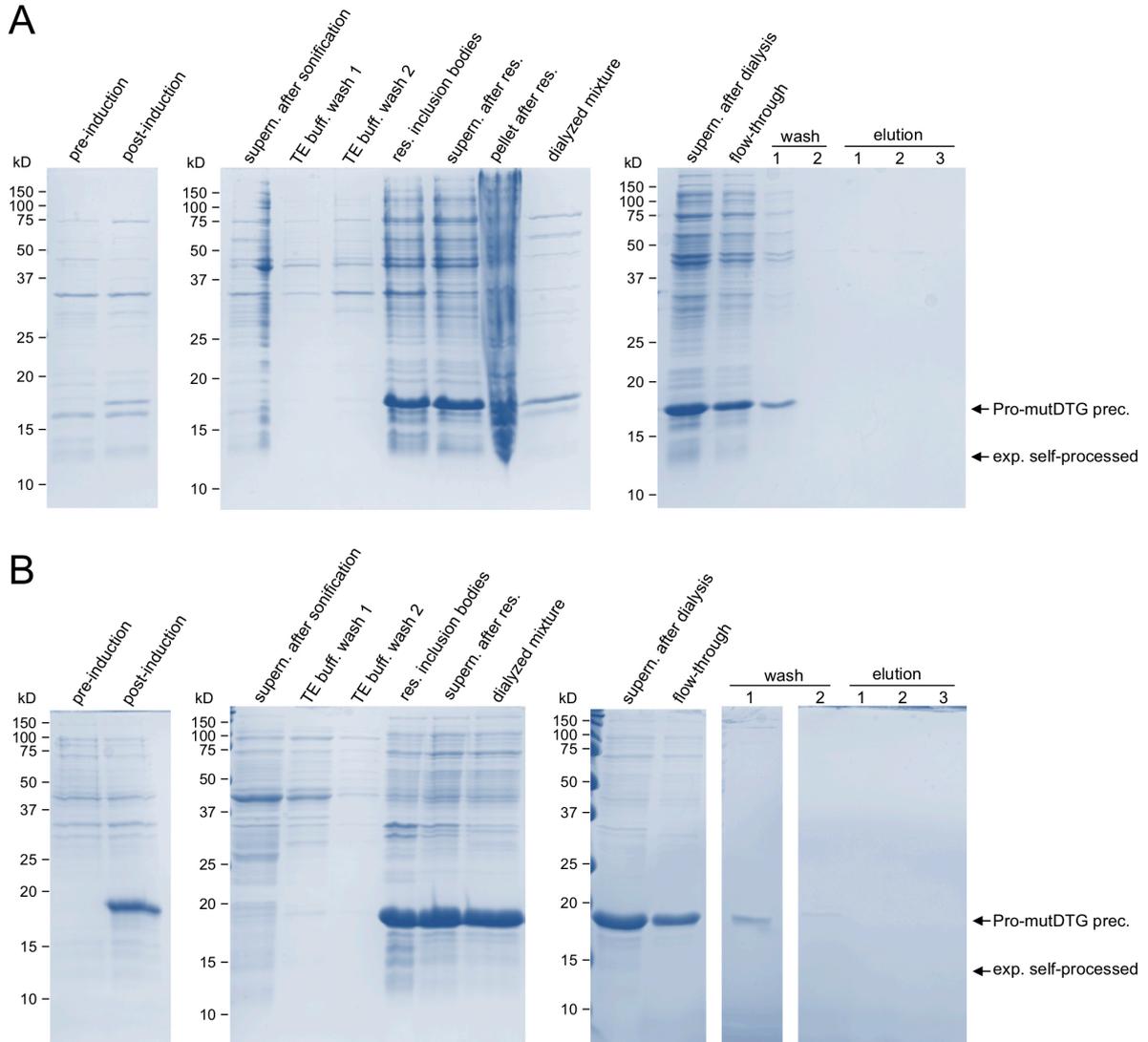
Figure 20: Ineffective purification of two HML-2 Pro mutants. As for wild-type HML-2 Pro (see Figure 17), various steps during the purification procedure of Pro-mutDTG **(A)** and Pro-mutGRDL **(B)** were monitored by SDS-PAGE and subsequent Coomassie-staining of proteins. Documented steps of the procedure are as follows: pre- and post-induction of protein expression by IPTG; supernatant (supern.) after centrifugation following sonification; two washes of inclusion bodies in TE Buffer; resuspension (res.) of inclusion bodies in urea-containing buffer; supernatant and pellet (Pro-mutDTG only) after centrifugation of resuspended inclusion bodies; protein mixture after dialysis; supernatant after centrifugation following dialysis; flow-through after incubation of bacterial lysate with pepstatin A-agarose; washes after binding and three elution fractions. Note that mutant HML-2 Pro precursors (prec.) of 18 kDa are expressed and that no such proteins, as well as no expected (exp.) self-processed HML-2 Pro of 12 kDa are detectable in the elution fractions.

4.1.2 Verification of HML-2 Pro activity after purification

During expression and purification of HML-2 Pro, we documented Pro self-processing from the 18 kDa precursor to the 12 kDa form, providing preliminary evidence of Pro enzymatic activity (see above). However, we did not have confirmation of HML-2 Pro activity after concentration and storage at -80°, that could also have affected HML-2 Pro functionality. Moreover, the results so far could not discriminate between proteolytic Pro activity *in cis* and *in trans*. We therefore aimed at confirmation of enzymatic activity of our HML-2 Pro aliquots stored at -80°, with an accompanying confirmation of proteolytic activity *in trans*. To do so, purified HML-2 Pro was combined in a reaction with *in vitro* translated HML-2 Gag, a known substrate of HML-2 Pro [185] that was expected to be efficiently and specifically cleaved by proteolytically active Pro. The reaction buffer that we employed was based on conditions reported for a previous study on HML-2 Pro [112]. The buffer was composed of 20 mM PIPES, 100 mM NaCl, 1 mM DTT, 10% [v/v] glycerol, pH 6.5. Of note, the final composition of the reaction buffer was probably slightly different due to poorly definable additional components added from the *in vitro* transcription/translation reaction producing HML-2 Gag protein. Control reactions without Pro as well as without buffer were also performed. Reactions were incubated for 1 h at 37°C, then subjected to SDS-PAGE and Western blot. HML-2 Gag was detected using α-Gag polyclonal antibody no. 2548 [143]. Proteolytic activity of purified HML-2 Pro *in trans* was evident by a reduced amount of full-length Gag protein and appearance of a protein band of ~37 kDa corresponding to Gag matrix p15 domain (MA-p15) [66] released from Gag full-length protein by HML-2 Pro activity (Figure 21). Intensity of the MA-p15 band further increased when HML-2 Pro was present in the reaction at a higher concentration. The MA-p15 domain was not produced in absence of buffer or in a control reaction without Pro (Figure 21). We therefore could confirm proteolytic activity *in trans* for purified HML-2 Pro.
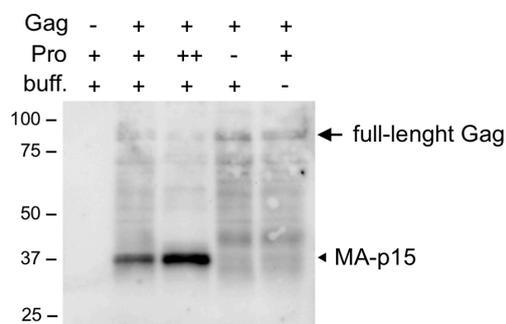
Figure 21: Confirmation of proteolytic activity *in trans* of purified wild-type HML-2 Pro. HML-2 Gag of ~78 kDa was produced employing a coupled *in vitro* transcription/translation system (also referred to as TNT T7 system, see 3.2.2.2). 1 μl of the HML-2 Gag containing TNT T7 reaction was mixed with HML-2 Pro, with the latter at final concentrations of 3 μM (+) and 9 μM (++), in a final volume of 10 μl each. Control reactions were run without Gag (-), Pro, or buffer (buff.). Reactions were incubated for 1 h at 37°C. Gag was detected by Western blot using polyclonal α-Gag antibody no. 2548 [143]. Full-length Gag and MA-p15, that is a major processing product formed by HML-2 Pro activity (see text), are indicated.

4.1.3 Characterization of HML-2 Pro activity and optimization of reaction conditions using an *in vitro* fluorescence assay

Following confirmation of purified HML-2 Pro proteolytic activity, we further characterized HML-2 Pro activity through a series of *in vitro* assays with varying reaction conditions. The primary purpose was the optimization of reaction conditions for downstream analysis, specifically identification of human proteins cleaved by HML-2 Pro *in vitro* (see chapter 4.1.4). At the same time, we wanted to validate and potentially expand previously published biochemical properties of HML-2 Pro [205]. To do so, we set up a fluorescence assay where purified HML-2 Pro was incubated with 2-aminobenzoyl-Thr-Ile-Nle-p-nitro-Phe-Gln-Arg-NH$_2$ (hereafter referred to as Anthranilyl-substrate), a substrate derived from a known cleavage site of HIV-1 Pro in HIV-1 Gag. The Anthranilyl-substrate was previously employed for HIV-1 Pro studies [45] and was expected to be suitable for HML-2 Pro activity assays because of similar specificity profiles of the two proteases [17]. The substrate, that makes use of fluorescence resonance energy transfer (FRET) (see chapter 3.2.2.4), was engineered in a way that cleavage by protease releases a fluorescent molecule from the substrate's N-terminus that is then no longer quenched by a molecule at the substrate's C-terminus, resulting in enhanced fluorescence emission when subjected to excitation at 280 nm (see Figure 14).

Reactions were set up in 96-well microplates placed on ice. Numbers of reactions per experiment were variable, depending on the number of conditions tested. For each condition, a control reaction without Pro was included. Each reaction was prepared in duplicate with a final volume of 50 μl. The Anthranilyl-substrate was usually added as the last reaction component. Microplates were then placed in a Tecan Infinite® m200 spectrophotometer set at 37°C during the entire measurement. Fluorescence measurements were done with excitation at 280 nm and emission at 420 nm, as indicated by the manufacturer (Bachem) of the Anthranilyl-substrate. Fluorescent signal intensity was monitored in each well every 4 min for a maximum of 180 min. Change in intensity of fluorescence emission at 420 nm was indicative of the amount of cleaved product, thus of HML-2 Pro activity. Fluorescence emission was plotted after subtracting background fluorescence measured for reactions without Pro.

The following description of results is divided into three sections: 1) Selection of a buffer composition suited for HML-2 Pro reactions; 2) Examination of HML-2 Pro activity at variable pH for definition of the pH optimum for proteolytic activity; 3) Establishing the concentration of pepstatin-A required for complete inhibition of HML-2 Pro.

4.1.3.1 Selection of a buffer composition suited for HML-2 Pro reactions

In order to identify a buffer that allowed for high HML-2 Pro activity and to characterize reaction parameters that we had to consider for our subsequent experiments, HML-2 Pro activity was assayed in four different buffer systems previously reported in studies on HIV-1 Pro or HML-2 Pro. Buffer compositions examined varied with regard to the type of buffering agent (PIPES, MES, or MES-TRIS), salt concentration, pH, concentration of glycerol, presence of reducing agent (DTT), and presence of EDTA.

Buffer compositions were as follows:
Buffer 1: 20 mM PIPES, 100 mM NaCl, 1 mM DTT, 10% [v/v] glycerol, pH 6.5
Buffer 2: 50 mM MES, 1 M NaCl, 20% [v/v] glycerol, 1 mM EDTA, pH 5.0
Buffer 3: 50 mM MES, 1 M NaCl, 1 mM EDTA, pH 5.0
Buffer 4: 100 mM MES-TRIS, 1.25 M NaCl, pH 6.0

Buffers 1 and 2 were previously reported in Kuhelj et al. (2001) [112] to study HERV-K Pro. Buffer 3 derived from Buffer 2, having the same composition except for lack of glycerol. Buffer 4 was previously employed [156] for characterization of HIV-1 Pro. As Anthranilyl-substrate and pepstatin A (see later) were resolved in DMSO, for which a negative effect on HIV protease was previously reported [92], we also set up reactions to evaluate the influence of DMSO on HML-2 Pro activity. Those reactions were set up in Buffer 4, that was already found to be suitable for HML-2 Pro activity in preliminary experiments (not shown). Since Anthranilyl-substrate was resolved in DMSO we could not set up meaningful reactions without DMSO. In fact, each reaction contained a minimum amount of DMSO corresponding to 2% [v/v] of the total reaction volume. Reactions with 4% [v/v] and 8% [v/v] DMSO were performed to observe effects of higher DMSO concentrations on HML-2 Pro activity. For each condition, Anthranilyl-substrate was present in reactions at a final concentration of 20 µM, and HML-2 Pro at 460 nM. Fluorescent emission at 420 nm was measured every 4 min during the course of an 80 min incubation.

Based on fluorescence emission, HML-2 Pro was found to be most active in Buffers 3 and 4 (Figure 22 A), that were both characterized by relatively high NaCl concentrations, no glycerol, and a pH below 6.5. In Buffer 2, HML-2 Pro displayed less activity than in Buffer 3. Buffers 2 and 3 differed by the presence of glycerol, thus indicating a negative effect of glycerol on HML-2 Pro activity. No change in fluorescence emission, that is, no Pro activity was observed for Buffer 1. Buffer 1 presented the lowest ionic strength, highest pH, and it contained DTT as opposed to all other buffers. Also, glycerol was included in that buffer. Thus, probably because of a combination of factors affecting Pro activity negatively, Buffer 1 was completely unsuitable for processing of the Anthranilyl-substrate by

HML-2 Pro. From this experiment we could not evaluate the potential influence of PIPES or DTT on Pro activity. However, we could observe that MES was clearly suitable and DTT not indispensable for Pro activity. Thus, based on the above described results, a buffer composed of 100 mM MES and 1 M NaCl was selected for subsequent optimization reactions.

DMSO at concentrations of 4% [v/v] and 8% [v/v] reduced Pro activity by approximately 20% and 50%, respectively, when compared with reactions containing 2% [v/v] DMSO (Figure 22 B). Thus, we confirmed a negative effect of DMSO on HML-2 Pro activity. We also considered important for further experiments to keep DMSO concentration as low as possible to reduce Pro inhibition. In subsequent experiments, DMSO concentration never exceeded 4% [v/v].
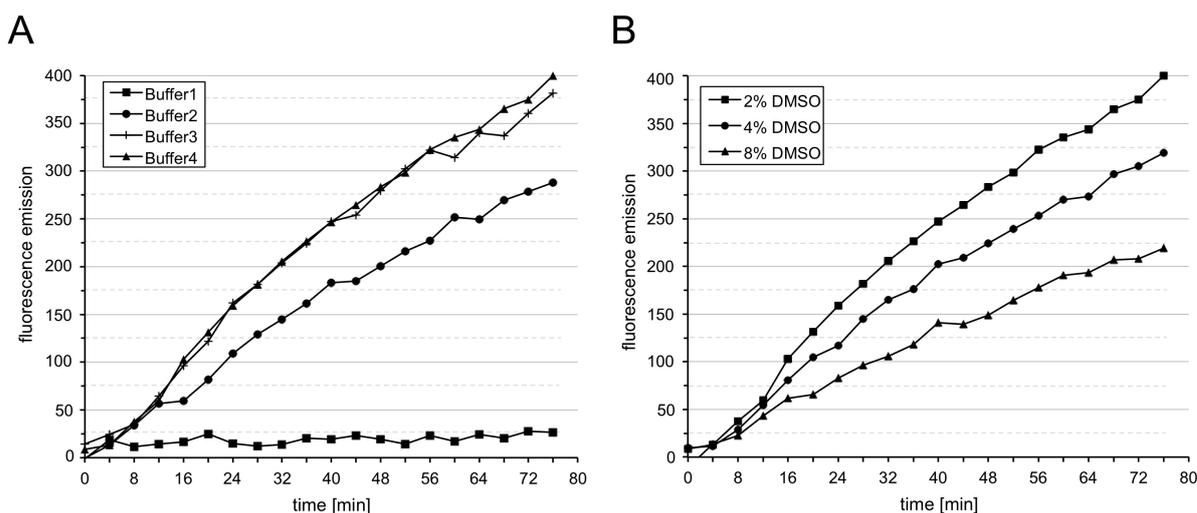
Figure 22: Monitoring HML-2 Pro activity in four different buffer systems and with increasing DMSO concentrations. 460 µM purified HML-2 Pro was incubated with 20 µM Anthranilyl-substrate at different reaction conditions. Fluorescent emission at 420 nm with background correction (Y-axis) was measured for the indicated time periods (X-axis). Although reactions were run in a single experiment, for the sake of clarity, monitoring of reactions in Buffer 4 at various DMSO concentrations are depicted separately in B. **A)** HML-2 Pro activity was monitored in four different buffer systems (see text for buffer compositions). All reactions in A contained 2% DMSO. **B)** Effect of DMSO on HML-2 Pro activity was evaluated monitoring HML-2 Pro activity in Buffer 4 with increasing DMSO concentrations. Final DMSO concentrations are indicated.

4.1.3.2 Examination of HML-2 Pro activity at variable pH and definition of pH optimum for proteolytic activity

Enzymes can be variably sensitive to pH changes. Generally, they tend to have an optimal pH at which they are catalytically most active. Therefore, pH is a fundamental parameter that should be adjusted when setting up enzymatic reactions. Based on results presented in the previous section, we confirmed that a relatively low pH was favorable for HML-2 Pro reactions, as also deduced from previously published findings [205]. In order to further optimize reaction conditions and to learn how much HML-2 Pro tolerates pH variations, we designed experimental assays where processing of the Anthranilyl-substrate by HML-2 Pro was monitored at pH ranging from pH 4.5 to pH 8. As concluded in the previous section, the reaction buffer selected was composed of 100 mM MES and 1 M NaCl. In a first

assay, processing of the Anthranilyl-substrate by HML-2 Pro at pH ranging from pH 4.5 to pH 6.5 was monitored for 130 min (Figure 23 A). HML-2 Pro was found to be active at all pH tested, with highest activity at pH 5.5. At pH 4.5, pH 5.0, pH 6.0, and pH 6.5, HML-2 Pro activity was, respectively, approximately 25%, 10%, 45%, and 70% lower compared to pH 5.5 (Figure 23 A, B).
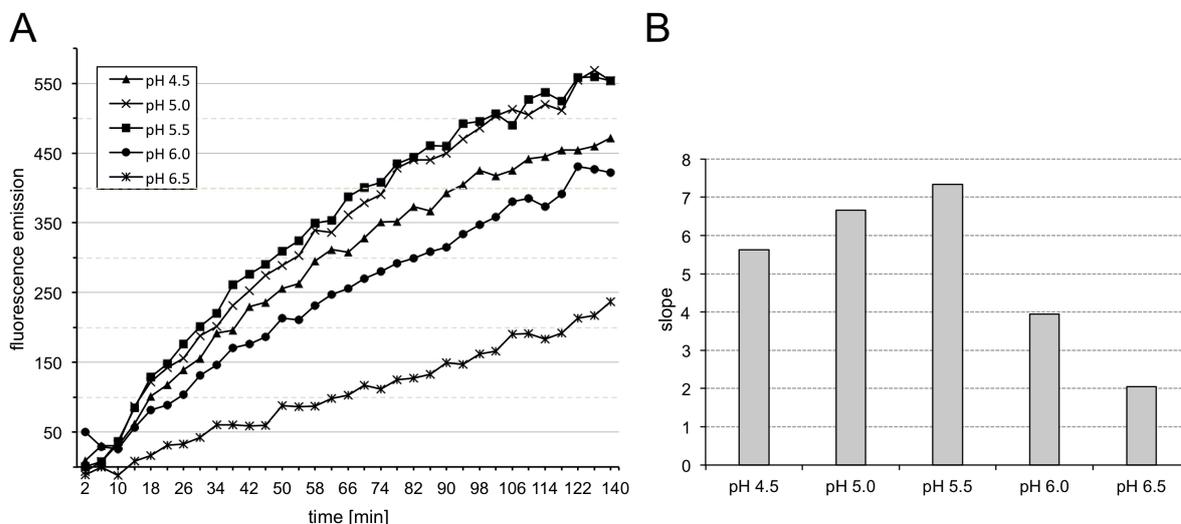
Figure 23: Determination of pH optimum for HML-2 Pro activity. A) HML-2 Pro activity was assayed in a buffer composed of 100 mM MES, 1 M NaCl, 2% [v/v] DMSO, adjusted to pH between 4.5 and 6.5 as indicated. Fluorescence emission at 420 nm with background correction is depicted on the Y-axis. Fluorescence emission was measured for the time periods indicated on the X-axis. **B)** Bar chart depicting slope of reaction curves shown in A during the first 40 min of the experiment. The bar chart allows for more accurate evaluation of Pro activity at various pHs and determination of the Pro pH optimum.

Considering that HML-2 Pro displayed proteolytic activity also at the higher pH (pH 6.5) tested in the first assay, we performed another assay with pH ranging from 5.5 to 8 to further evaluate Pro activity at neutral and basic pH. Moreover, the experiment was monitored for a longer time, specifically 180 min, to compensate for lower HML-2 Pro activity. Although HML-2 Pro activity was gradually reduced when gradually increasing pH from 5.5 to pH 8, HML-2 Pro was still active at pH 6.9, with 10% of the activity measured at pH 5.5. A retained low activity at pH 7.5, with 2% of the activity measured at pH 5.5, as well as pH 8.0, with 1% of the activity measured at pH 5.5, was also detectable (Figure 24 A).

Approximately 120 minutes into the experiment, fluorescence emission did not further increase for reactions at pH 5.5 and pH 6. For reactions at pH 6.5 and higher, fluorescence emission further increased until the end of the experiment. Very likely, in the course of the experiment the Anthranilyl-substrate became depleted at pH 5.5 and pH 6 due to higher activity of HML-2 Pro compared to pH ≥6.5.

We wondered if the lower activity of Pro at pH > 6.5 was due to the lesser suitability of the MES buffer at these pH values. To evaluate the influence of the buffering agent in the HML-2 Pro activity, another assay was performed at identical conditions except with MES (buffering pH range of 5.5-6.7 [57]) in the reaction replaced by PIPES (buffering pH range of 6.1-7.5 [57]). Monitoring HML-2 Pro activity, PIPES buffer was more suitable than MES buffer at all pHs tested. At pH 5.5 and 6.0, Pro was

approximately 25% less active in MES buffer when compared with PIPES. At pH > 6.0 the reduction in Pro activity between MES and PIPES was gradually more pronounced, approximately 55%, 70%, 80%, 90% lower at pH 6.5, 6.9, 7.5, 8.0, respectively (Figure 24 A, B, C). Our results, suggested that PIPES buffer could be more suitable than MES buffer in our system, in particular when monitoring Pro activity at pH > 6.0.
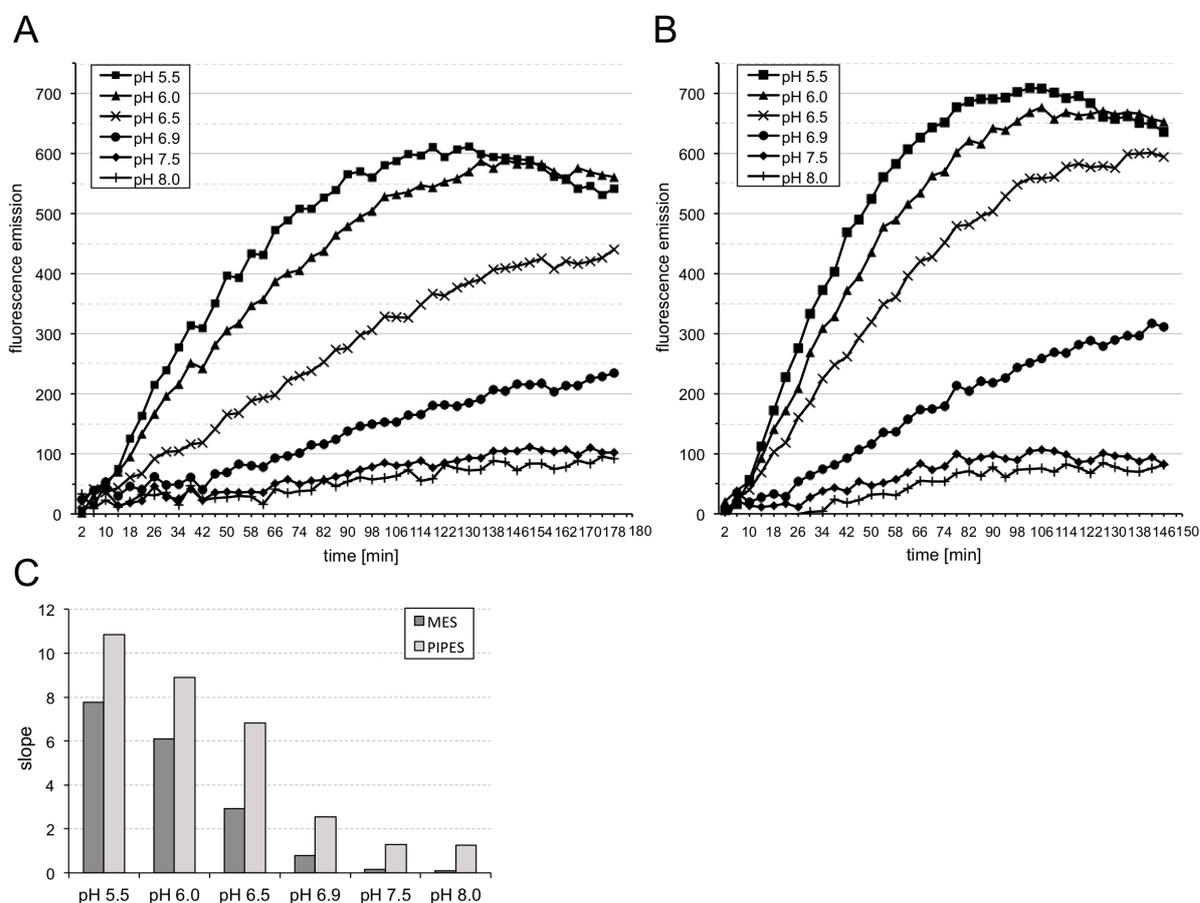
Figure 24: Monitoring HML-2 Pro activity at acidic and basic pH. HML-2 Pro activity was monitored in reaction buffers containing MES **(A)** or PIPES **(B)** at pH ranging from 5.5 to 8.0 during a longer incubation time. Time points above 150 min were considered not relevant for the comparison, thus for the sake of convenience they were not depicted in B. Fluorescence emission at 420 nm with background correction is depicted on the Y-axis. Fluorescence emission was measured for the time periods indicated on the X-axis. **C)** Bar chart depicting slope of reaction curves shown in A and B during the first 60 min of the experiments.

4.1.3.3 Monitoring HML-2 Pro saturation by Anthranilyl-substrate and inhibition by pepstatin A

As will be described in more detail in chapter 4.1.4, the experimental set-up for identification of cellular proteins that are potential HML-2 Pro substrates is essentially based on the comparison between two *in vitro* reactions. Both reactions include total cellular protein lysate plus enzymatically active HML-2 Pro in one reaction, and inactive HML-2 Pro in the other. It was originally planned to include in the reaction with inactive Pro a mutated, enzymatically inactive variant of HML-2 Pro. Since mutant HML-2 Pro could not be purified successfully by means of pepstatin A-agarose (see chapter 4.1.1) it was decided to repress HML-2 Pro activity by a specific inhibitor of aspartyl proteases. For this

purpose, we selected pepstatin A and established the concentration of pepstatin A, relative to that of HML-2 Pro, required to effectively inhibit HML-2 Pro activity in the reaction.

Concentration of Anthranilyl-substrate necessary to saturate HML-2 Pro activity

Before assaying pepstatin A inhibition, we considered it relevant to know the relative concentration of Anthranilyl-substrate necessary to saturate HML-2 Pro activity. For pepstatin A being a competitive inhibitor, the inhibitory effect was influenced by the concentration of HML-2 Pro substrate, that competes for interaction with HML-2 Pro. This aspect was relevant for subsequent incubations of HML-2 Pro with total cell lysate for which it was difficult to predict the amount of interactions of HML-2 Pro with cellular substrate proteins. Therefore, assaying pepstatin A inhibition at a saturating concentration of Anthranilyl-substrate indicated the concentration of pepstatin A required in the negative control, in order to effectively inhibit HML-2 Pro activity also under saturating interactions between Pro and cellular substrates.

We thus defined the Anthranilyl-substrate concentration at which HML-2 Pro in solution was saturated. To do so, various reactions were performed with fixed HML-2 Pro concentration (460 nM) and different Anthranilyl-substrate concentrations ranging from 5 μM to 60 μM (Figure 25). Resulting fluorescence emission-curves reached in some cases a plateau, indicating depletion of Anthranilyl-substrate. Specifically, fluorescence emission-curves of reactions with 5, 10, and 20 μM Anthranilyl-substrate reached a plateau after approximately 40, 80, and 130 min, respectively. During the course of the 140-min experiment, change in fluorescence emission did not reach a plateau in reactions with Anthranilyl-substrate at concentrations of 40 and 60 μM. Slopes of reaction curves based on the first 40 min of incubation were calculated as indicators of the reaction rate. When plotting slopes of curves relative to substrate concentrations we observed that increasing substrate concentrations up to 20 μM increased reaction rates proportionally. Reaction rates did not further increase when further increasing substrate concentrations from 40 μM to 60 μM. Thus, we concluded that in our experimental system activity of HML-2 Pro present at 460 nM became saturated by a substrate concentration of approximately 40 μM, corresponding to an enzyme:substrate ratio of 1:86.9.
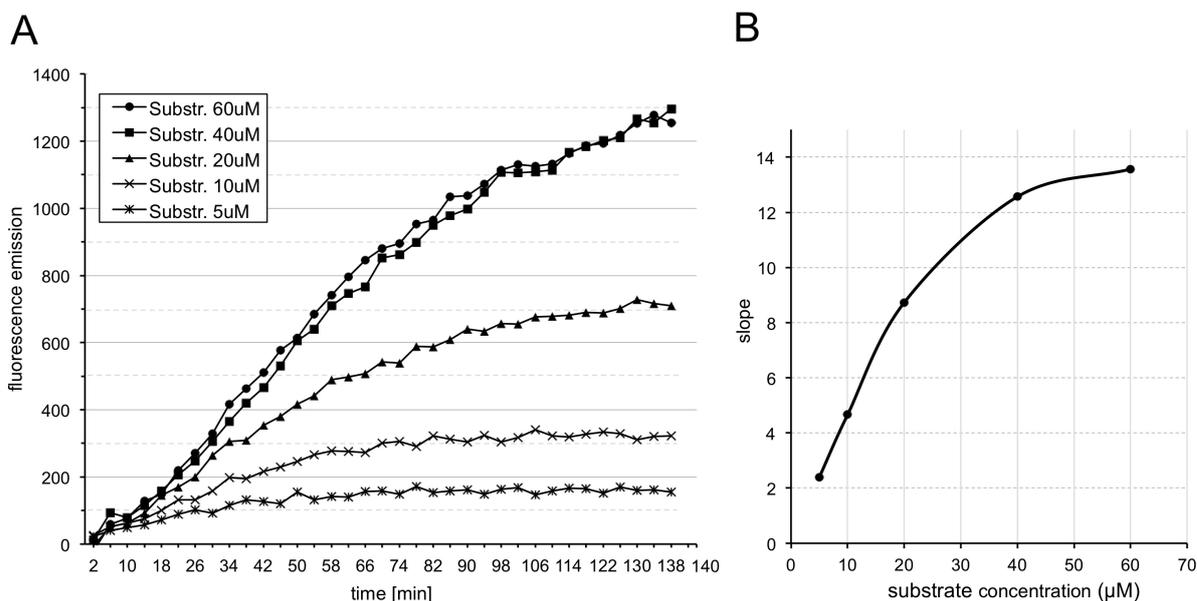
Figure 25: HML-2 Pro activity at various substrate concentrations. A fixed HML-2 Pro concentration of 460 nM was assayed in combination with Anthranilyl-substrate concentrations ranging from 5 μM to 60 μM. **A)** Fluorescence emission at 420 nm with background correction is depicted on the Y-axis. Fluorescence emission was measured for the time periods as indicated on the X-axis. **B)** Slopes of five reaction curves, based on the first 40 min of incubation, were calculated and are depicted (Y-axis) relative to substrate concentration (X-axis). The slope of the reaction curves reached a plateau at a concentration of Anthranilyl-substrate of approximately 40 μM, indicating that HML-2 Pro became saturated at that substrate concentration.

Establishing the concentration of pepstatin A required for efficient inhibition of HML-2 Pro

Inhibition of HML-2 Pro by pepstatin A was monitored at 5 different concentrations of inhibitor ranging from 1 μM to 200 μM (Figure 26). Of note, when setting up reactions, HML-2 Pro and pepstatin A were pre-incubated for 10 min at room temperature before addition of the Anthranilyl-substrate as the last reaction component. The pre-incubation had the purpose of allowing for interaction between HML-2 Pro and pepstatin A before addition of Anthranilyl-substrate, thus optimizing inhibition.

A reaction without pepstatin A was included as a reference indicating 100% HML-2 Pro activity. Other reactions with pepstatin A present were compared with the reference reaction to evaluate the reduction of Pro activity. Results showed that 1 μM, 10 μM 100 μM, and 200 μM pepstatin A reduced Pro activity to approximately 60%, 18%, 12%, and 0%, respectively (Figure 26). Without pre-incubation, inhibition in presence of 200 μM pepstatin A was considerable but not complete (still 2% Pro activity). Thus, we concluded that 200 μM pepstatin A effectively inhibited HML-2 Pro present at 460 nM when the reaction was pre-incubated for 10 min before addition of Anthranilyl-substrate.
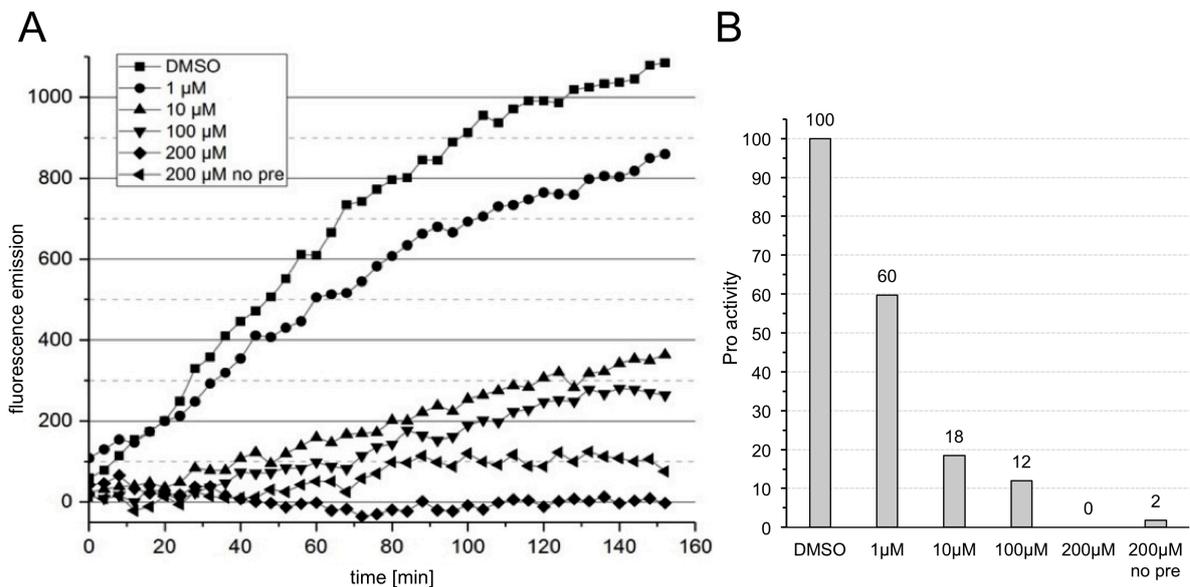
Figure 26: Effects of pepstatin A concentration on HML-2 Pro activity. A) The inhibitory effect of pepstatin A on HML-2 Pro activity was assayed at various concentrations of pepstatin A ranging from 1 µM to 200 µM as indicated. Reactions contained 460 µM purified HML-2 Pro and 40 µM Anthranilyl-substrate. When setting up reactions, HML-2 Pro was pre-incubated with pepstatin A for 10 min before addition of Anthranyl-substrate. A reaction with 200 µM pepstatin without preincubation before addition of substrate was included ("no pre"). Fluorescence emission at 420 nm with background correction is depicted on the Y-axis. Fluorescence emission was measured for the time periods indicated on the X-axis. **B)** Slopes of reaction curves, based on the first 40 min of incubation, were calculated and the HML-2 Pro activity was depicted in the bar chart as percentage of activity relative to a reaction with DMSO only, the latter defined as 100% activity.

4.1.4 Identification of human cellular proteins cleaved by HML-2 Pro

Retroviral aspartyl proteases can process not only retrovirus-encoded proteins but also cellular proteins (see Introduction, chapter 2.4.2), as shown, for instance, in a recent study [82] related to cellular substrates of HIV-1. In that study, a proteome-wide analysis was employed to identify more than 120 cellular proteins that were processed by HIV-1 *in vitro*. Cellular protein substrates identified in that study located to different cellular compartments and many of the proteins had physiologically important roles. Thus, that study as well as other previous studies (see Introduction, chapter 2.4.2) suggest that the proteolytic activity of a retroviral protease could potentially impact the cell. Up to date, no such particular proteomic analyses were conducted with regard to substrates of endogenous retroviruses-encoded aspartyl proteases.

In the following, we employed a sophisticated experimental approach that could tell whether HML-2 Pro is able to process human cellular proteins, specifically identifying proteins eventually processed and providing more detailed information on cleavage events.

Our experimental system involved *in vitro* reactions combining a native protein mixture derived from HeLa cells with purified, active HML-2 Pro. The reaction buffer was optimized with regard to buffering agent, salt concentration, and pepstatin A concentration (see chapter 4.1.3). Concerning pH, since HML-2 Pro was enzymatically active within an acidic to basic pH range (see chapter 4.1.3.2) we

examined HML-2 Pro activity at two different pH conditions. One experiment was performed at pH 5.5, defined by us as optimal pH for HML-2 Pro activity (in our system). The other experiment was performed at pH 7.0, at which HML-2 Pro showed reduced but still significant activity. Given that HML-2 Pro localization inside a cell was not investigated so far and HML-2 Pro could potentially locate and be active in various cellular compartments, reactions a two different pHs was expected to inform about the ability of HML-2 Pro to cleave substrates at pH conditions similar to those of acidic compartments, like endosomes, or to those of neutral compartments, like cytosol.

In collaboration with Prof. Dr. Oliver Schilling (University of Freiburg), we utilized a sensitive proteome-wide methodology for identification of protease substrates, called Terminal Amine Isotopic Labeling of Substrates (TAILS). The data generated by TAILS were further examined in our laboratory.

4.1.4.1 Preparation of a protein mixture from HeLa cells

A cellular protein mixture for incubation with HML-2 Pro was extracted through lysis of HeLa cells by freeze/thaw cycles in a buffer composed of 100 mM MES, 1 M NaCl, pH 6.0 (see chapter 3.2.2.9). Cell lysis conditions avoided use of detergents that are usually employed for efficient membrane disruption during cell lysis. Such detergents may have altered protein structure and interfered with downstream TAILS procedures. Preparation of HeLa proteome involved centrifugation of the cell lysate for removal of cell membranes and cellular debris. The supernatant was stored and used for reactions with purified HML-2 Pro. Detection of histone H3 in supernatants derived from HeLa lysates confirmed disruption of nuclear compartments and release of nuclear proteins (Figure 27 A).

4.1.4.2 Small-scale experiment to confirm reaction conditions for TAILS

Before setting up actual reactions for TAILS, reaction conditions were verified in a small-scale experiment that, besides established reaction components, included *in vitro*-translated HML-2 Gag protein for monitoring HML-2 Pro activity via processing of a known substrate. For this experiment, 200 μM HML-2 Pro was incubated for 1 h with 1 μl of *in vitro*-translated HML-2 Gag in presence of 20 μg of proteins of HeLa total cell lysate, at pH 5.5 and pH 7 and with pepstatin A at two different concentrations, in a final reaction volume of 20 μl. Processing of HML-2 Gag was then examined by Western blot using α-HML-2 Gag rabbit polyclonal antibody 2548 [143] (Figure 27 B). HML-2 Pro processed HML-2 Gag efficiently at pH 5.5, reducing the amount of full-length Gag and generating two distinct Gag cleavage products, one of ~37 kDa likely representing the MA-p15 domain, and the second, slightly smaller one likely representing the CA-NC domain [66]. As expected, HML-2 Pro exhibited lower activity at pH 7. Only the MA-p15 Gag cleavage product could be detected with certainty and at a lower amount in comparison with the reaction at pH 5.5. When comparing reactions with and without HeLa total cell lysate (Figure 27 B, lanes 3 and 4) presence of HeLa cellular proteins

seemed to influence HML-2 Gag processing, slightly reducing the degree of full-length Gag processing. In presence of pepstatin A at concentrations of 200 µM and 100 µM (Figure 27 B, lanes 5 and 6), the amount of full-length Gag was not significantly different from that in the control (Figure 27 B, lane 2). Also, Gag cleavage products were drastically lower than respective reactions without pepstatin A (Figure 27 B, lanes 3 and 4), thus suggesting that pepstatin A inhibited HML-2 Pro activity efficiently.

	Lane 1	2	3	4	5	6	7	8
TNT T7 Gag		+	+	+	+	+	+	+
PIPES pH 5.5	+	+	+	+	+	+		
PIPES pH 7.0							+	+
HeLa cell lysate	+	+		+	+	+	+	+
HML-2 Pro	+		+	+	+	+	+	++
pepstatin A					++	+		
gel lane	1	2	3	4	5	6	7	8

kD 100 –
75 – ← full-lenght Gag
50 –
37 – ← MA-p15 / CA-NC domain
25 –
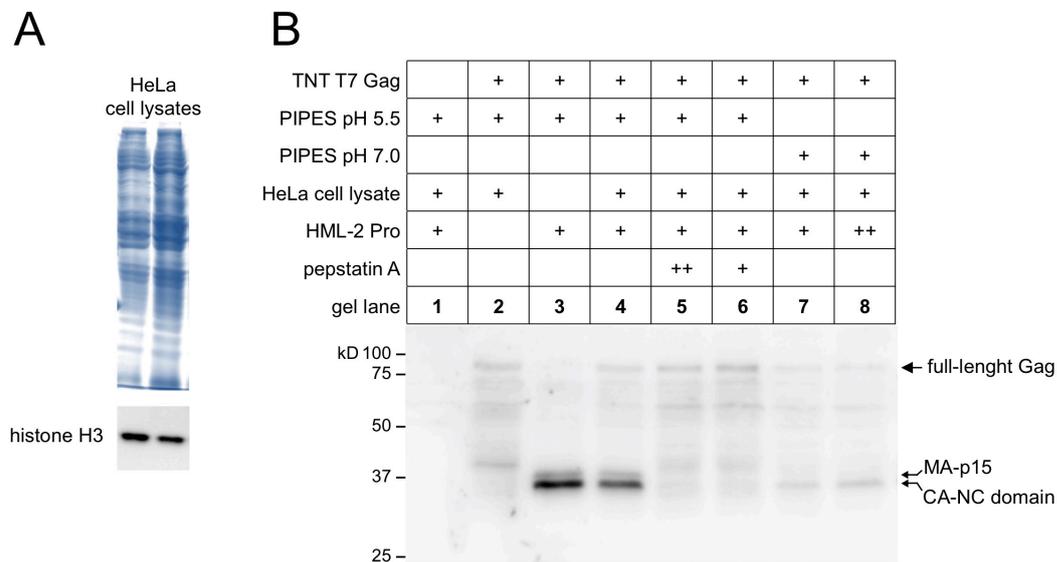
Figure 27: Verification of experimental conditions for a full-scale TAILS experiment. A) For confirmation of efficient disruption of nuclei during lysis, 15 µg total cellular proteins from a HeLa cell lysate prepared by a freeze/thaw lysis procedure (see 3.2.2.9) were subjected to SDS-PAGE and Western blot. Histone H3, that locates to the nucleus, was detected using a monoclonal α-histone H3 antibody. In the upper part, the PAA-gel was stained with Coomassie after electrophoresis for documenting presence of manifold other cellular proteins. **B)** Various reactions were set up as indicated. Reaction components were included (+) as follows: 1 µl of *in vitro* translated HML-2 Gag reaction produced in a TNT T7 system; 100 mM PIPES pH 5.5 or pH 7.0; 20 µg protein from HeLa total cell lysate; 200 nM (+) and 460 nM (++) HML-2 Pro; 100 µM (+) and 200 µM (++) final concentration of pepstatin A. Final reaction volume was 20 µl each. Full-length Gag and two processing products (MA-p15 and CA-NC domain) produced by HML-2 Pro activity (see text) are indicated.

4.1.4.3 Set up of larger-scale reactions for TAILS

The small-scale experiment confirmed that at the established conditions for TAILS reactions HML-2 Pro was enzymatically active *in trans*, specifically HML-2 Pro efficiently processed HML-2 Gag, a known Pro substrate [185]. Thus, the next step was to prepare larger-scale reactions for the actual TAILS experiment in order to examine processing of cellular proteins. The final volume of reactions for TAILS was 100-fold larger than for the small-scale experiment as the TAILS procedure required a larger amount of total protein in a larger volume. We wanted to ensure a sufficiently large reaction volume if repeated TAILS analysis would be required.

For larger-scale TAILS reactions, 2 mg of HeLa total protein were incubated with HML-2 Pro at a final concentration of 200 nM in a final reaction volume of 2 ml. Buffer conditions were as follows: 1

M NaCl, 2% [v/v] DMSO, 100 mM PIPES at pH 5.5 or pH 7. Control reactions in presence of 200 μM pepstatin A, effectively blocking HML-2 Pro activity, were also included. Control reactions with pepstatin A were pre-incubated for 10 min at room temperature before addition of HeLa total protein as the last reaction component. All reaction mixtures were thoroughly mixed and incubated at 37°C for 75 min. Two replicate reactions were prepared for each pH condition. As will be described later, because of technical issues with specific devices for TAILS in the laboratory of Prof. Schilling, two more replicates were subsequently prepared in order to repeat reactions at pH 7. Reactions were frozen at -80°C, packed and shipped, on dry ice, to the laboratory of Prof. Oliver Schilling at University of Freiburg. Reactions were then subjected to further steps of the TAILS procedure in that laboratory, specifically actual identification of cellular proteins cleaved in presence of HML-2 Pro.

4.1.4.4 TAILS procedure

The TAILS methodology employed in our study is based on an N-terminomics approach capable of distinguishing in amino acid chains between "regular" N-termini and N-termini being protease cleavage products [106]. Since TAILS was not performed in our laboratory, the basic procedure [106,108] will be described only briefly below.

After incubation of HeLa-derived protein mixture with active HML-2 Pro (designated Pro-treated sample) or inhibited HML-2 Pro (designated untreated sample), samples are denatured and reduced. Lysine amines and free (α)-amino groups of N-termini of proteins and their protease cleavage products are simultaneously blocked and labelled with stable isotopes applying a reductive dimethylation. Methyl groups containing ^{13}C (heavy) or ^{12}C (light) are introduced in Pro-treated samples and untreated samples, respectively. The labelling with stable isotopes allows the distinction between peptides originated in the Pro-treated sample and untreated sample. Pro-treated samples and respective controls are pooled together to ensure identical treatment in the following steps. The combined protein mixtures are then subjected to trypsinization generating novel internal tryptic peptides with free N-termini, whereas labeled N-termini of the original proteins remain blocked. Blocked N-terminal peptides are selectively isolated and concentrated by negative selection. To do so, hyperbranched polymers specifically reacting with free N-termini of newly generated tryptic peptides are added to samples. Dimethylated lysines, isotopically labeled N-terminal peptides and the neo-N-terminal peptides of their cleavage products remain unbound, thus they are recovered by ultrafiltration. Relative abundances of peptides and actual peptide sequences in Pro-treated and control samples are then determined by liquid chromatography-tandem mass spectrometry (LC-MS/MS) and LC-MS/MS data analysis. A combination of MS/MS measurements and stringent bioinformatic analysis assign peptides to respective human protein sequences, and also provide precise information on positions of cleavage sites within those human proteins.

Raw data output from TAILS analyses was provided to us in tabular format. Tables reported non-prime and prime peptide fragments identified, thus cleavage sites in human proteins, and various additional information. A representative dataset with TAILS results for two cleaved human proteins is exemplified in Table 17.

Table 17: Example of data output from a TAILS experiment

Protein ID 1	P07900	P36507
Protein ID 2	Q58FF7	
Protein ID 3	P08238	
Experiment	pH5.5 Rep2	pH5.5 Rep2
Non prime	RKKLSELLRY	EQQKKRLEAF
Prime	YTSASGDEMVSLKDYCTR	LTQKAKVGELKDDDFER
P1	Y	F
P1´	R	L
Name	Heat shock protein HSP 90-alpha	Dual specificity mitogen-activated protein kinase kinase 2
Alternative Name	Heat shock 86 kDa; Lipopolysaccharide-associated protein 2; Renal carcinoma antigen NY-REN-38	ERK activator kinase 2;MAPK/ERK kinase 2
Fold change (Log2)	9.966	2.036
Fc error		
Cleavage position	464	56
Protein length	732	400
Modifications	N-ter +37.08 Da, K +37.08 Da, C +57.02 Da	N-ter +37.08 Da, K +37.08 Da
Hyperscore	26	39,4
Ppm	-2,0398	-0,3741
Charge	3	3
Mass	2154.05	2135.33
GO compartment	cytoplasm; cytosol; endocytic vesicle lumen; extracellular exosome; extracellular region; ficolin-1-rich granule lumen; lysosomal lumen; melanosome; membrane; myelin sheath; nucleoplasm; nucleus; plasma membrane; ruffle membrane; secretory granule lumen;	cell-cell junction; cytoplasmic side of plasma membrane; cytosol; early endosome; endoplasmic reticulum; extracellular region; focal adhesion; Golgi apparatus; late endosome; microtubule; mitochondrion; nucleus; perinuclear region of cytoplasm; peroxisomal membrane; plasma membrane
GO process	chaperone-mediated autophagy; chaperone-mediated protein complex assembly; ciliary basal body docking; ERBB2 signaling pathway; Fc-gamma receptor signaling pathway involved in phagocytosis; G2/M transition of mitotic cell cycle; mitochondrial transport; neutrophil degranulation; positive regulation of nitric oxide biosynthetic process; positive regulation of telomerase activity; protein import into mitochondrial outer membrane; protein refolding; protein stabilization; protein unfolding; receptor-mediated endocytosis; regulation of cellular response to heat; regulation of nitric-oxide synthase activity; regulation of protein complex assembly; regulation of protein ubiquitination; response to antibiotic; response to cold; response to heat; response to unfolded protein; signal transduction; telomerase holoenzyme complex assembly; telomere maintenance via telomerase; vascular endothelial growth factor receptor signaling pathway;	activation of MAPK activity; ERK1 and ERK2 cascade; MAPK cascade; negative regulation of gene expression; peptidyl-serine autophosphorylation; positive regulation of ERK1 and ERK2 cascade; positive regulation of production of miRNAs involved in gene silencing by miRNA; positive regulation of protein serine/threonine kinase activity; positive regulation of transcription, DNA-templated; proteolysis in other organism; regulation of early endosome to late endosome transport; regulation of Golgi inheritance; regulation of stress-activated MAPK cascade
GO function	ATP binding; ATPase activity; disordered domain specific binding; DNA polymerase binding; GTPase binding; histone deacetylase binding; identical protein binding; MHC class II protein complex binding; nitric-oxide synthase regulator activity; nucleotide binding; protein homodimerization activity; protein tyrosine kinase activity; protein tyrosine kinase binding; RNA binding; TPR domain binding; unfolded protein binding;	ATP binding; MAP kinase kinase activity; metal ion binding; PDZ domain binding; protein serine/threonine kinase activator activity; protein serine/threonine kinase activity; protein serine/threonine/tyrosine kinase activity; protein tyrosine kinase activity; scaffold protein binding; signal transducer, downstream of receptor, with serine/threonine kinase activity
SP localization	Cytoplasm ECO 0000250 UniProtKB P07901	Cytoplasm ECO 0000269 PubMed 10409742

The sequence of each identified peptide fragment was automatically associated with a human protein and a respective protein ID, that is, an accession number identifying proteins unambiguously in protein sequence databases. For proteins with known isoforms, IDs of isoforms were also reported (see Table 17, Protein ID 2 and Protein ID 3). For the sake of convenience, we considered in our analysis only the first protein ID. The protein ID and the cleavage position, as defined by sequences of non-prime

and prime peptide fragments, were two fundamental pieces of information that we used for extraction of numbers of the proteolytically generated N-termini. The fold-change, derived from comparison of abundances of respective peptide fragments between protease-treated sample and untreated sample, was given as the binary logarithm ($\log_2 n$) indicating enrichment of respective peptide fragments. For each identified peptide, aa residues flanking the cleavage site and several peptide properties were also reported, such as Gene Ontology (GO) terms [8] for cellular process, function and compartment of human proteins identified by TAILS as cleaved by HML-2 Pro.

4.1.4.5 Analysis of raw data output from TAILS

TAILS results were subjected to filtering and data overlap employing various tools implemented in Microsoft Excel (Office 2016) in order to select high-confidence biologically relevant candidate substrates from all possible candidates identified from neo-N-terminal peptides. We focused mainly on protein IDs, cleavage position and fold-change. To simplify the terminology employed in the following, processing of a particular protein at a specific aa position will be designated as "cleavage event".

As for the TAILS experiment at pH 5.5, TAILS identified for the two replicates greater than 9800 and 8500 cleavage events (Figure 28). Thus, as an initial rule to collect proteolytically generated N-termini, we filtered cleavage events that were enriched at least 2-fold upon HML-2 Pro incubation. Applying such a filter, 4370 cleavage events were identified in replicate 1 and 2633 cleavage events in replicate 2 (Figure 28). Of those, 931 cleavage events were common to both replicates (Figure 29), corresponding to 548 human proteins cleaved in the same position. The number of replicated cleavage events exceeded the effective number of cleaved proteins indicating that many proteins were cleaved in more than one position. In fact, although the majority of proteins was cleaved in a single position, several proteins were cleaved at more than one position (Figure 30). In some cases, multiple cleavage events per protein were surprisingly numerous (Figure 30). Proteins with a higher number of different cleavages within the particular protein are reported in Table 18. For example, heat shock protein 90 alpha family class B member 1 (HSP90AB1) was cleaved in 30 and 50 positions in replicate 1 and replicate 2, respectively. For myosin heavy chain 9 (MYH9), 25 and 60 cleavage events were observed, for actin beta (ACTB) 38 and 32, for eukaryotic translation elongation factor 1 alpha 1 (EEF1A1) 21 and 25. In total, we identified 322 proteins common to both replicates for which there was evidence for 2-fold enriched multiple cleavages (Figure 29).

When focusing our attention on positions of cleavage events, we noticed that among cleavage events that did not occurr exactly in the same position in both replicates, many had nevertheless occured within a relatively narrow region of a particular protein in both replicates. In approximately 50% of

cases, at least one cleavage had occurred in the same 50 aa-region within a particular protein. In 25% of cases, at least one cleavage had occurred within a 5 aa-region within a particular protein. We hypothesized that HML-2 Pro is able to specifically interact with some protein regions and then process the protein within that region but not necessarily at a specific cleavage site. More specific investigations would be necessary to understand if structural or chemical characteristics of some proteins could eventually influence such a phenomenon. Such investigations were out of the aim of the study though. However, in view of this observation, when we next aimed at identifying actual proteins cleaved in both replicates (with at least 2-fold enrichment) we did not consider only those proteins necessarily cleaved in the same position within a protein. Using such a criterium, we identified 2024 and 1170 proteins in replicate 1 and 2, respectively. Those corresponded to 809 human proteins cleaved in both experiments, thus with replicated evidence of cleavage by HML-2 Pro.

As for the TAILS experiment at pH 7, two replicates were initially performed. One replicate at pH 7 could not be processed and completed properly due to technical issues with devices in the laboratory of Prof. Schilling. Two more replicates were performed later on for the sake of a more complete analysis. Thus, we will refer in the following to three replicates at pH 7. Like for the TAILS at pH 5.5, raw data from three TAILS replicates at pH 7 were subjected to a filtering procedure. Greater than 3100 cleavage events were identified for replicates 1, greater than 3500 for replicate 2 and greater than 7000 for replicate 3 (Figure 28). Raw numbers already indicated fewer cleavage events than for the TAILS experiment at pH 5.5, possibly due to lower enzymatic activity of HML-2 Pro at pH 7. For an initial estimate of cleavage events replicated in all three assays, a less stringent (1.5-fold instead of 2-fold) enrichment cut-off was applied for the first filtering step; 1074 (replicate 1), 514 (replicate 2) and 2314 (replicate 3) cleavage events were enriched greater than 1.5-fold (Figure 28). Overlapping cleavage events with 1.5-fold enrichment among the three replicates identified 11 cleavage events common to all three replicates, 58 cleavage events common to replicate 1 and 2, 182 cleavage events common to replicate 1 and 3, and 41 cleavage events common to replicate 2 and 3 (Figure 29). Thus, some degree of variability of common cleavage events among replicates was observed. Replicate 1 and 3, which showed a higher number of enriched cleavage events and the greatest overlap (182) of 1.5-fold enriched cleavage events, were used for a more stringent filtering to identify replicated cleavage events enriched greater than 2-fold. At a cut-off value of 2-fold change, 694 (replicate 1) and 1985 (replicate 3) cleavage events were generated by HML-2 Pro, with a total of 129 cleavage events common to both replicates (Figure 29).
Similar to the TAILS experiment at pH 5.5, we observed at pH 7 multiple cleavage events within several of the human proteins (Figure 30). For instance, we identified 74 proteins common to replicate 1 and 3 for which there was reported evidence of 2-fold enriched multiple cleavages. Among those proteins were 15 and 10 cleavage events in replicates 1 and 3, respectively, for HSP90AB1, 32 and 6 events for MYH6, 15 and 12 for VCP, and 23 and 11 events for HSPA8 (Table 18). Overlapping the

two replicates at pH 5.5 with replicate 1 and 3 at pH 7 revealed 52 common proteins with 2-fold enrichment cleaved in more than one position.

As for the two replicates at pH 5.5, we also identified in replicates 1 and 3 at pH 7 proteins cleaved with at least 2-fold enrichment, yet not necessarily cleaved in the same position within a protein. At a cut-off value of 2-fold change, 342 (replicate 1) and 1134 (replicate 3) different human proteins were cleaved by HML-2 Pro. Combining replicates 1 and 3, a total of 217 human proteins showed replicated evidence of cleavage by HML-2 Pro at pH 7 (Figure 29).

Lastly, we evaluated the overlap in cleaved proteins among the two TAILS replicates at pH 5.5 (pH5.5_Rep1, pH5.5_Rep2) and two TAILS replicates at pH 7 (pH7_Rep1, pH7_Rep3) (Figure 29, four-color Venn diagrams). That analysis identified 154 different human proteins cleaved by HML-2 Pro that were detected in the four TAILS experiments when applying 2-fold enrichment (Figure 29). 187 cleaved proteins were common to the four experiments when applying 1.5-fold enrichment for the two replicates at pH 7 and 2-fold enrichment for replicates at pH 5.5 (not shown).

Altogether, we considered the above findings to bear evidence of potential processing of human proteins by HML-2 Pro. HML-2 Pro was able to cleave many cellular proteins at acidic as well as neutral pH. The 809 human proteins common to the two pH 5.5 experiments plus the 217 proteins common to the two pH 7 experiments (see above), resulted in a total of 872 unique proteins (154 proteins were common to the two lists, thus 809 + 63), all of them with at least 2-fold enrichment, thus with strong evidence of processing by HML-2 Pro *in vitro*. The 872 proteins (see Appendix, Table 21) can be considered as cellular substrates of HML-2 Pro that could be potentially processed by HML-2 Pro also *in vivo*.
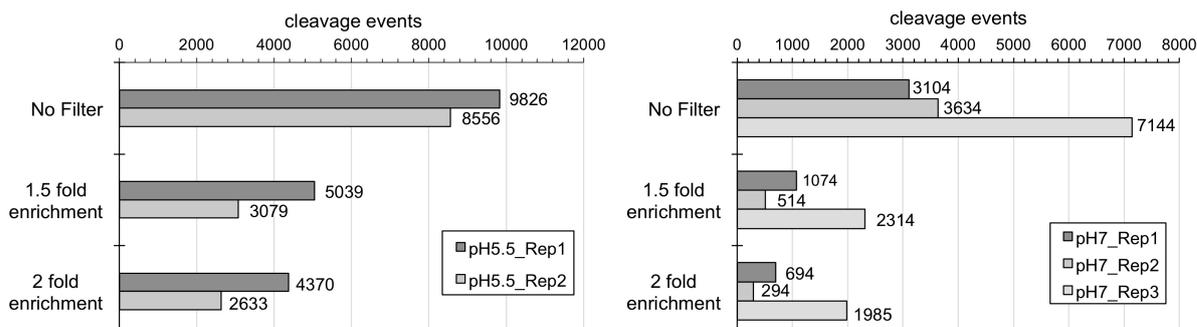
Figure 28: Filtering of cleavage events identified by TAILS. Results of filtering of raw data from experimental replicates at pH 5.5 (pH5.5_Rep) and pH 7 (pH7_Rep) are depicted. Filters were applied for 1.5-fold or 2-fold enrichment of cleavage events. Numbers of cleavage events after applying the various filters are depicted by bars and particular numbers are given next to bars.
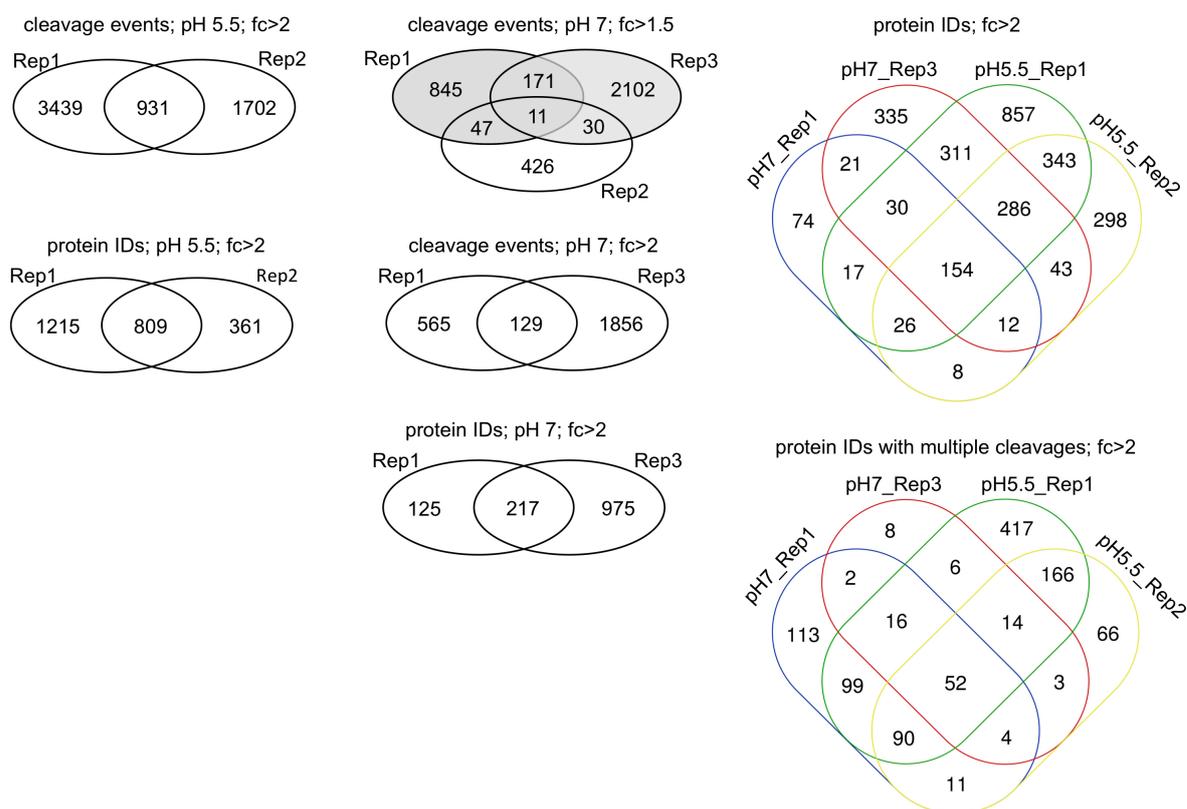
Figure 29: Venn diagrams depicting overlap of cleavage events and protein IDs among replicate experiments performed at pH 5.5 and pH 7. Overlap of cleavage events with 1.5-fold enrichment is shown on the top (middle) for three replicates at pH 7. Of those, replicate 1 and 3 (highlighted in grey) showed higher numbers of 1.5-fold enriched cleavage events and the greatest overlap. They were therefore considered for identification of replicated cleavage events at pH 7 with 2-fold enrichment. The other Venn diagrams considered cleavage events or protein IDs with 2-fold enrichment. The overlap among protein IDs detected in two pH 7 replicates and the two pH 5 replicates is depicted in the four-color Venn diagram at the top. The four-color Venn diagram at the bottom depicts overlap of proteins that were processed in more than one position.
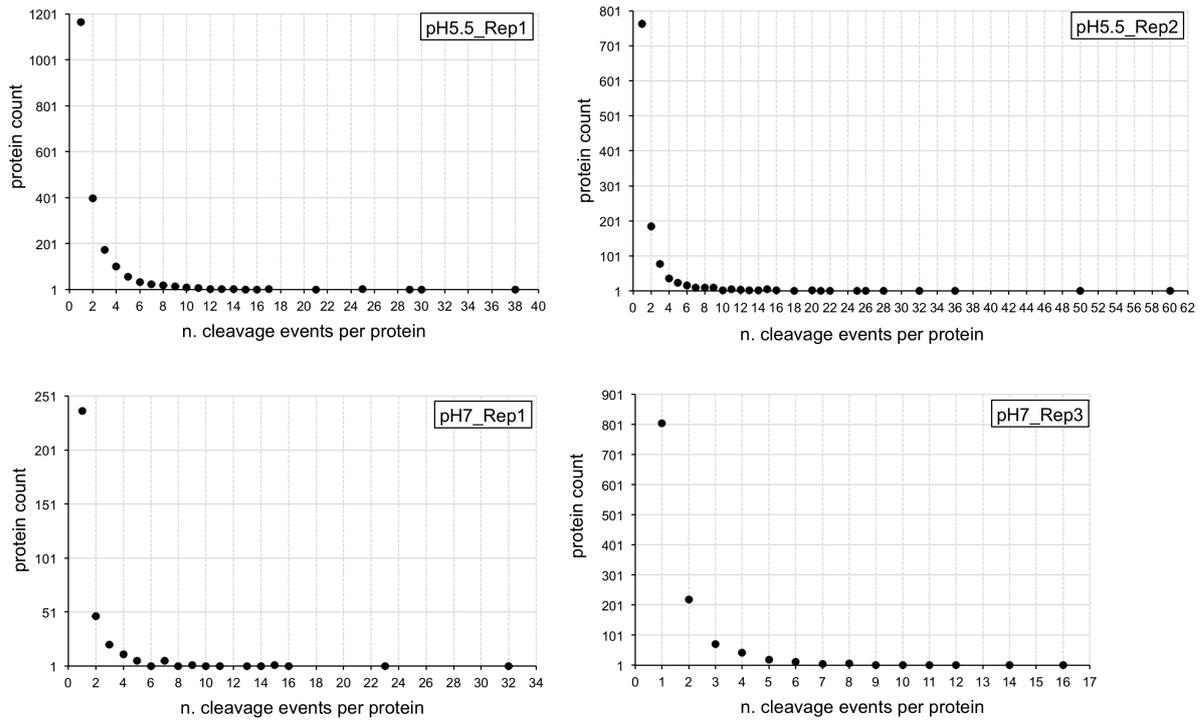
Figure 30: Numbers of cleavage events for proteins identified as substrates of HML-2 Pro. Results are depicted for the two TAILS experiments at pH 5.5 and for two TAILS experiments at pH 7. Only cleavage events with greater than 2-fold enrichment were considered. The majority of proteins was cleaved in a single position, fewer proteins were cleaved at more than one position, and a relatively small number of proteins were cleaved at many different positions within the particular protein.

Table 18: Examples of human proteins cleaved by HML-2 Pro in multiple positions [*]

pH5.5; Fc>2

Protein ID	gene symbol	protein name	Rep1	Rep2
P60709	*ACTB*	Actin beta	38	32
P08238	*HSP90AB1*	Heat shock protein 90 alpha family class B member 1	30	50
P22314	*UBA1*	Ubiquitin like modifier activating enzyme 1	29	21
P35579	*MYH9*	Myosin heavy chain 9	25	60
P68104	*EEF1A1*	Eukaryotic translation elongation factor 1 alpha 1	21	25
P07900	*HSP90AA1*	Heat shock protein 90 alpha family class A member 1	17	26
Q00839	*HNRNPU*	Heterogeneous nuclear ribonucleoprotein U	17	20
Q92598	*HSPH1*	Heat shock protein family H (Hsp110) member 1	17	15
P62937	*PPIA*	Peptidylprolyl isomerase A	16	16
P43487	*RANBP1*	RAN binding protein 1	15	13
P04406	*GAPDH*	Glyceraldehyde-3-phosphate dehydrogenase	14	11
P00558	*PGK1*	Phosphoglycerate kinase 1	13	18
P62917	*RPL8*	Ribosomal protein L8	13	14
P06744	*GPI*	Glucose-6-phosphate isomerase	12	10
P11142	*HSPA8*	Heat shock protein family A (Hsp70) member 8	11	36
Q13263	*TRIM28*	Tripartite motif containing 28	11	15
P63261	*ACTG1*	Actin gamma 1	11	14
Q04637	*EIF4G1*	Eukaryotic translation initiation factor 4 gamma 1	11	12
Q71U36	*TUBA1A*	Tubulin alpha 1a	11	11
P62753	*RPS6*	Ribosomal protein s6	10	12

pH7; Fc>2

Protein ID	gene symbol	protein name	Rep1	Rep3
P35579	*MYH9*	Myosin heavy chain 9	32	6
P11142	*HSPA8*	Heat shock protein family A (Hsp70) member 8	23	11
P55072	*VCP*	Transitional endoplasmic reticulum ATPase	15	12
P08238	*HSP90AB1*	Heat shock protein 90 alpha family class B member 1	15	10
P26038	*MSN*	Moesin	14	5
P14618	*PKM*	Pyruvate kinase M1/2	13	8
P06733	*ENO1*	Alpha-enolase	10	6
P60709	*ACTB*	Actin beta	9	14
P18206	*VCL*	Vinculin	9	8
O43707	*ACTN4*	Actinin alpha 4	8	6
P07437	*TUBB*	Tubulin beta class I	7	9
P07900	*HSP90AA1*	Heat shock protein 90 alpha family class A member 1	7	8
P0DMV9	*HSPA1B*	Heat shock protein family A (Hsp70) member 1B	7	6
P00558	*PGK1*	Phosphoglycerate kinase 1	7	5
P22314	*UBA1*	Ubiquitin like modifier activating enzyme 1	5	11
P78371	*CCT2*	T-complex protein 1 subunit beta	5	7
Q9Y490	*TLN1*	Talin-1	5	7
P29401	*TKT*	Transketolase	5	6
P63261	*ACTG1*	Actin gamma 1	4	6

[*] For the TAILS experiments at pH 5.5, proteins with 2-fold enrichment and 10 or more identified cleavage events in different positions in either replicate (Rep1, Rep2) are listed. For the TAILS experiments at pH 7, proteins with at least 4 observed cleavage events (Rep1, Rep3) and 2-fold enrichment for the two selected replicates are shown. Protein IDs, gene symbols and protein names are depicted.

4.1.5 Cellular substrates of HML-2 Pro: cellular localization, biological processes and association with diseases

In order to gain some insights into potential biological impact of an enzymatically active HML-2 Pro expressed in human cells, we determined cellular localization and biological function of human proteins that, on the basis of TAILS results, are substrates of HML-2 Pro. The 872 identified proteins (see the previous chapter) were subjected to Gene Ontology (GO) analysis. GO is a bioinformatic initiative that annotates and classifies genes and gene products using a set of predefined bins called terms [39]. The terms incorporate information related to cellular localization, molecular activity and biological processes in which a specific gene/protein is likely or factually involved. Each gene can be annotated through various terms organized hierarchically. Data provided by the GO system are associated with online tools that can apply different methods to profile sets of genes/proteins on the basis of various GO terms. For our profilings we used the PANTHER (Protein Analysis Through Evolutionary Relationships) classification database [140], a public website where our set of 872 protein IDs was loaded in order to retrieve for each protein in our list GO terms regarding cellular compartments and biological processes. Our analysis was primarily interested in the range of GO categories associated with our 872-protein list and numbers of proteins for each category of interest. It needs to be considered here that human proteins are dynamic and multifunctional elements inside a cell that are not necessarily associated with a single compartment or process. Thus, a protein can be related to multiple GO terms. The GO analysis also included an overrepresentation test that used a binomial statistics tool to compare the input set with a reference list (composed of all human genes included in the PANTHER database) and statistically determine over- or under- representation of PANTHER classification categories. However, that particular test was not relevant for our purpose. We ran two GO term analyses through PANTHER. Selected annotation datasets were "GO cellular component complete" and "GO biological processes complete".

The PANTHER analysis related to cellular components revealed that potential substrates of HML-2 Pro locate to diverse cellular compartments (Figure 31). The majority of proteins were associated with cytoplasm and nucleus, with ~580 and ~550 proteins, respectively. Many proteins were associated with plasma membrane (~150), membrane part (~130), extracellular space (~300), and cell projection (~125). Subcellular organelles enclosed by membrane were also associated with a number of proteins: ~70 proteins with secretory granules, ~20 with primary lysosome, and ~300 with exosome. Also, ~90 and ~50 proteins were associated with mitocondrion and ribosome, respectively.

The PANTHER analysis related to biological processes revealed that proteins cleaved by HML-2 Pro are involved in several processes. Relevant biological processes associated with HML-2 Pro cellular substrates included response to stress (~230 proteins), immune system process (~160 proteins), cell death (~120 proteins), cell differentiation (~100 proteins), cytoskeleton organization (~100 proteins),

regulation of cell cycle (90 proteins), DNA repair (~40 proteins), DNA replication (~20 proteins), mRNA processing (~60 proteins), translation (~100 proteins), protein folding (~60 proteins), mitochondrion organization (~20 proteins), and nervous system process (~10 proteins). The complete list of GO terms for biological processes considered by us, with respective number of HML-2 Pro substrates, is depicted in Figure 31. That list provides a picture of biological processes associated with HML-2 Pro substrates, thus indicating cellular processes that could be potentially influenced by HML-2 Pro activity.
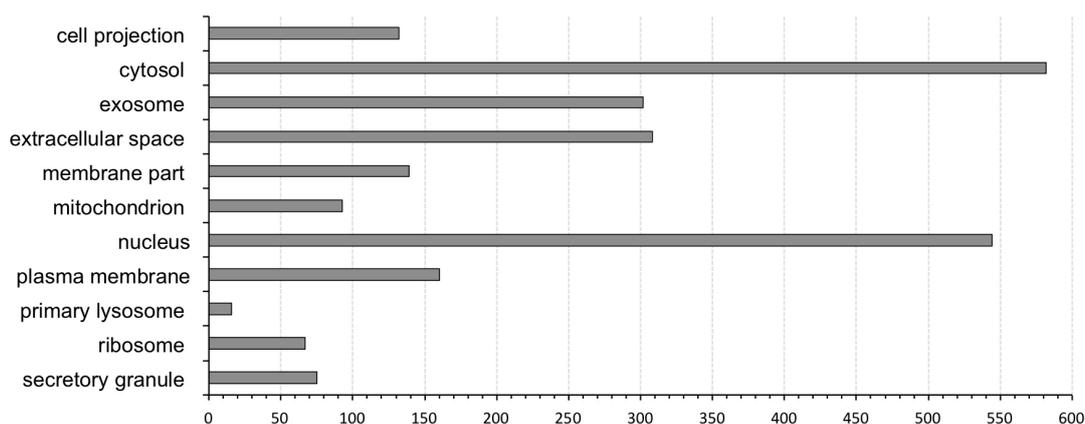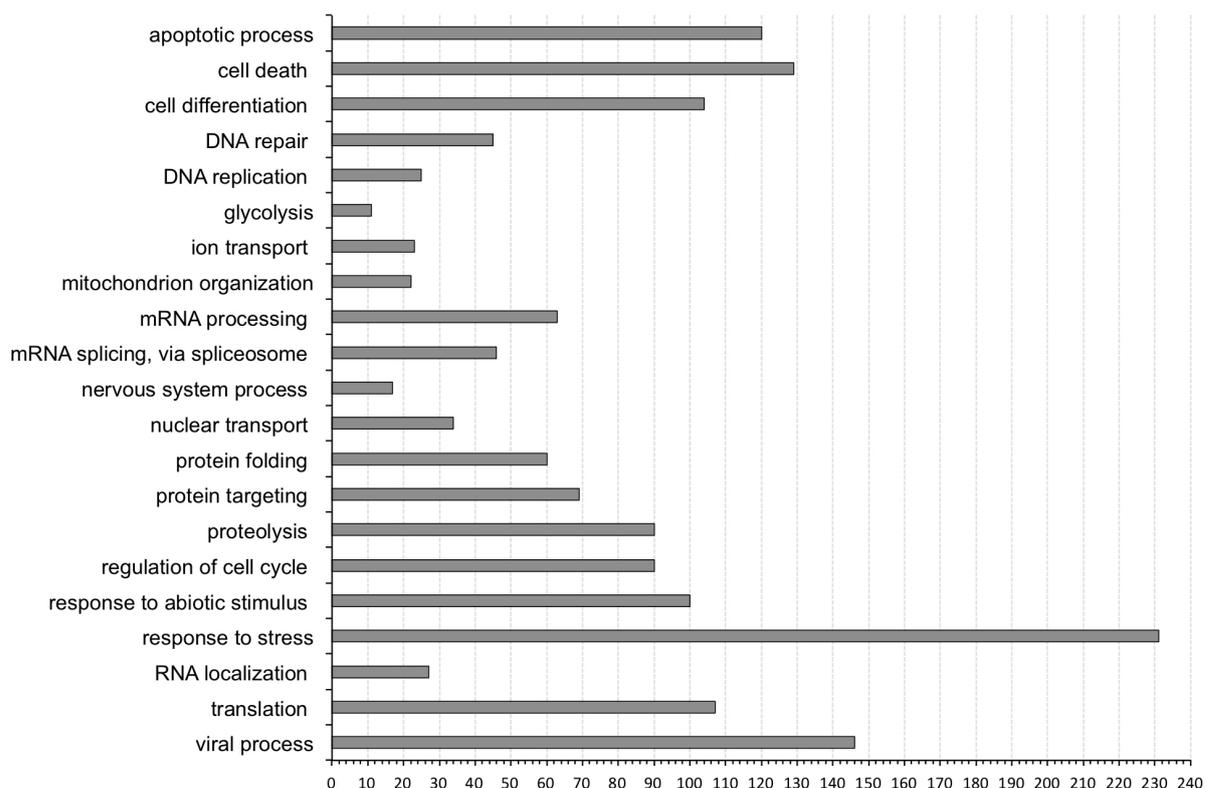
Figure 31: Summary of Gene Ontology term analysis of human proteins identified as substrates of HML-2 Pro by TAILS. Selected cellular compartments and biological processes are depicted. Numbers of proteins associated with the depicted Gene Ontology (GO) terms were obtained using the PANTHER (Protein ANalysis THrough Evolutionary Relationships) database as provided at http://www.pantherdb.org.

Furthermore, we aimed at evaluating the involvement of the identified HML-2 Pro substrates in human cancer, a condition in which HML-2 is implicated [171]. HML-2 upregulation has been previously associated with several tumor types [58,64]. Also, HML-2-encoded proteins have been detected in some of those tumor tissues [171].

For our analysis, the human genes corresponding to the selected 872 human proteins were intersected with genes included in the Catalogue Of Somatic Mutations In Cancer (COSMIC) database (https://cancer.sanger.ac.uk). COSMIC is a detailed and comprehensive resource collecting information on human genes/proteins and somatic mutations associated with cancer [201]. Our analysis identified 64 genes/proteins (in the following referred to as proteins) with an established relevance in oncology. Fifty-six of those proteins were associated with somatic mutations in cancer (see Table 19), 4 proteins (BUB1B, ERCC5, FH and SMARCE1) were associated with germline mutations that predispose to cancer, and 4 proteins (CDC73, MSH6, POLD1, MSH6) were associated with both. A total of 59 tumor types were associated with the 64 proteins. There were proteins associated with only one tumor type and proteins associated with various tumor types. Some tumors were more represented than others, with cases of more than 10 different proteins associated with a certain tumor, for instance, leukemia and lymphoma. Several proteins were associated with tumors of interest in the context of HML-2. Such proteins are reported in the following. B2M, ERCC5 and MAP2K2 were associated with melanoma, a type of cancer where HML-2 is found upregulated [29,183]. DAXX and IDH1 were associated with glioblastoma. FUBP1 was associated with oligodendroglioma. Glioblastoma and oligodendroglioma are brain tumors. A previous expression analysis of various brain tumor types, including glioblastoma, indicated that quite a number of HML-2 proviruses are transcriptionally active in those tissues [58]. MSH6 and PPP2R1A were associated with ovarian tumor, DDX5 and HNRNPA2B1 with prostate tumor, CTCF and ZMYM3 with breast cancer, three tumor types with reported altered HML-2 transcription [58,214]. Twelve proteins were associated with lymphoma, which could be of interest considering that HML-2 RNA was found at high levels in the plasma of patients with lymphoma [42].

Table 19: Involvement of potential HML-2 Pro cellular substrates in human cancers [*]

Gene symbol	Protein name	Tumour types (somatic)	Tumour types (germline)
ARAF	A-Raf proto-oncogene, serine/threonine kinase	cholangiocarcinoma, adenocarcinoma, Langerhans cell histiocytosis	
ATIC	5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase	ALCL	
B2M	beta-2-microglobulin	DLBCL, melanoma, colorectal adenocarcinoma	
BCL3	B-cell CLL/lymphoma 3	CLL	
BCL9L	B-cell CLL/lymphoma 9-like	colorectal cancer, endometrial carcinoma, gastric cancer	
BUB1B	BUB1 budding uninhibited by benzimidazoles 1 homolog beta (yeast)		rhabdomyosarcoma
CALR	calreticulin	MPN, MDS	
CDC73	cell division cycle 73	parathyroid adenoma	parathyroid adenoma, multiple ossifying jaw fibroma
CLIP1	CAP-GLY domain containing linker protein 1	Spitzoid tumour	
CNBP	CCHC-type zinc finger, nucleic acid binding protein	aneurysmal bone cyst	
CTCF	CCCTC-binding factor	endometrial, breast, head and neck cancer	
CTNND1	catenin delta 1	large intestine carcinoma	
DAXX	death-domain associated protein	pancreatic neuroendocrine tumour, paediatric glioblastoma	
DDX3X	DEAD-box helicase 3, X-linked	CLL, medulloblastoma	
DDX5	DEAD (Asp-Glu-Ala-Asp) box polypeptide 5	prostate	
EIF3E	eukaryotic translation initiation factor 3, subunit E	colorectal	
ERCC5	excision repair cross-complementing rodent repair deficiency, complementation group 5 (xeroderma pigmentosum, complementation group G (Cockayne syndrome))		skin basal cell, skin squamous cell, melanoma
EZR	ezrin	NSCLC	
FH	fumarate hydratase		leiomyomatosis, renal
FLNA	filamin A	phyllodes tumour of the breast	
FUBP1	far upstream element (FUSE) binding protein 1	oligodendroglioma	
GMPS	guanine monphosphate synthetase	AML	
HNRNPA2B1	heterogeneous nuclear ribonucleoprotein A2/B1	prostate	
HSP90AA1	heat shock protein 90kDa alpha (cytosolic), class A member 1	NHL	
HSP90AB1	heat shock protein 90kDa alpha (cytosolic), class B member 1	NHL	
IDH1	isocitrate dehydrogenase 1 (NADP+), soluble	glioblastoma	
KIF5B	kinesin family member 5B	NSCLC, Spitzoid tumour	
LASP1	LIM and SH3 protein 1	AML	
MAP2K2	mitogen-activated protein kinase kinase 2	NSCLC, melanoma	
MSH6	mutS homolog 6 (E. coli)	colorectal	colorectal, endometrial, ovarian
MSN	moesin	ALCL	
MYH9	myosin, heavy polypeptide 9, non-muscle	ALCL	
NACA	nascent-polypeptide-associated complex alpha polypeptide	NHL	
NONO	non-POU domain containing, octamer-binding	papillary renal	
NPM1	nucleophosmin (nucleolar phosphoprotein B23, numatrin)	NHL, APL, AML	
NT5C2	5'-nucleotidase, cytosolic II	relapse ALL	
NUMA1	nuclear mitotic apparatus protein 1	APL	
NUP214	nucleoporin 214kDa (CAN)	AML, T-ALL	
NUP98	nucleoporin 98kDa	AML	
PABPC1	poly(A) binding protein cytoplasmic 1	HNSCC, biliary tract carcinoma	
PCBP1	poly(rC) binding protein 1	CRC	
PCM1	pericentriolar material 1 (PTC4)	papillary thyroid, CML, MPN	
PHF6	PHD finger protein 6	ETP ALL	
PML	promyelocytic leukemia	APL, ALL	
POLD1	DNA polymerase delta 1, catalytic subunit	CRC	CRC
PPP2R1A	protein phosphatase 2, regulatory subunit A, alpha	clear cell ovarian carcinoma	
RANBP2	RAN binding protein 2	inflammatory myofibroblastic tumour	
RPL22	ribosomal protein L22 (EAP)	AML, CML	
RPL5	ribososomal protein L5	T-ALL	
SET	SET translocation	T-ALL	

Gene	Description	Somatic/Germline tumors	Second column
SF3B1	splicing factor 3b, subunit 1, 155kDa	MDS	
SFPQ	splicing factor proline/glutamine rich(polypyrimidine tract binding protein associated)	papillary renal	
SMARCE1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily e, member 1		meningioma
SND1	staphylococcal nuclease and tudor domain containing 1	pancreas acinar carcinoma	
STAT3	signal transducer and activator of transcription 3 (acute-phase response factor)	T-cell large granular lymphocytic leukaemia	paediatric large granular lymphocytic leukaemia
TFG	TRK-fused gene	papillary thyroid; ALCL; NSCLC; extraskeletal myxoid chondrosarcoma	
THRAP3	thyroid hormone receptor associated protein 3 (TRAP150)	aneurysmal bone cyst	
TPM3	tropomyosin 3	papillary thyroid, ALCL, NSCLC, Spitzoid tumour	
TPM4	tropomyosin 4	ALCL	
TPR	translocated promoter region	papillary thyroid, NSCLC	
TRIM33	tripartite motif-containing 33 (PTC7,TIF1G)	papillary thyroid	
UBR5	ubiquitin protein ligase E3 component n-recognin 5	mantle cell lymphoma, gastric, colorectal	
YWHAE	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, epsilon polypeptide (14-3-3 epsilon)	endometrial stromal sarcoma	
ZMYM3	zinc finger MYM-type containing 3	CRC, breast cancer	

* Human genes corresponding to proteins identified by TAILS as substrates of HML-2 Pro were intersected with cancer-relevant genes as compiled in COSMIC (Catalogue Of Somatic Mutations In Cancer; https://cancer.sanger.ac.uk/cosmic) [60]. The resulting genes list is reported, specifying associated tumor types. Tumor types in somatic cells and/or germ cells are indicated. ALCL: anaplastic large-cell lymphoma; ALL: acute lymphocytic leukaemia; AML: acute myeloid leukaemia; APL: acute promyelocytic leukaemia; CLL: chronic lymphocytic leukaemia; CML: chronic myeloid leukemia; CRC: colorectal cancer; DLBCL: diffuse large B-cell lymphoma; ETP-ALL: early T-cell precursor acute lymphoblastic leukemia; HNSCC: head and neck squamous cell carcinoma; MDS: myelodysplastic syndrome; MPN: myeloproliferative neoplasm; NHL: non-Hodgkin lymphoma; NSCLC: non small cell lung cancer; T-ALL: T-cell acute lymphoblastic leukaemia.

For an additional evaluation of an involvement of the identified HML-2 Pro substrates in diseases, our dataset of 872 proteins was also intersected with genetic disorders included in the Online Mendelian Inheritance in Man (OMIM) database [2], an online catalog of human genes and genetic phenotypes. OMIM focuses on gene-phenotype relationships, thus it provides information to understand how alteration of certain genes (and respective proteins) could lead to specific diseases. Our analysis identified 289 different disorder phenotypes of which approximately 250 were described as inherited. Detailed results from OMIM analyses are provided along with the electronic version of this dissertation. The identified phenotypes included various tumors, for example, neuroblastoma, colorectal cancer, prostate cancer, endometrial cancer, leukemia, and meningioma. Some of those tumor types were also identified by COSMIC analysis (see above). Among many other phenotypes, several neurological disorders were identified. For example, ALS, mental retardation, peripheral neuropathy, neurodegeneration with brain iron accumulation, Parkinson disease, and Huntington disease. Neurological disorders as well as different tumor types are of particular interest as previously pointed out in the context of HML-2 [35].

Taken together, our analyses indicated that among cellular substrates of HML-2 Pro there are many proteins for which their impairment has been associated with human diseases.

4.1.6 Further validation of potential substrates of HML-2 Pro

The TAILS analysis provided initial evidence of specific cleavage of many human proteins by HML-2 Pro. To provide further support to this, we considered verification of HML-2 Pro activity by additional, more specific experiments that could demonstrate processing of human proteins *in vitro* and *in vivo*. Eventually confirmed cleavages of cellular substrates could thus lend further support to a potential impact of HML-2 Pro in cellular processes in which cleaved proteins participate.

4.1.6.1 Verification experiments *in vitro*

Our *in vitro* experimental approach involved 1) selection of candidate proteins for verification experiments, 2) *in vitro* generation of selected candidate proteins tagged with an HA tag or a radioactive label, 3) incubation of candidate proteins with purified HML-2 Pro and analysis of cleavage of candidate proteins by means of either Western blot, or SDS-PAGE and phosphorimaging.

4.1.6.1.1 Selection of candidate proteins for verifications *in vitro*

The 2-fold enriched cleavage events identified in the two TAILS experiments at pH 5.5 were used as a starting point for selection of HML-2 Pro substrates to be tested. To reduce the initially long list of proteins to a reasonable number of candidates, we applied a series of criteria for filtering. Our criteria considered the amino acid specificities of HML-2 Pro cleavage sites (Figure 32 A), as recently reported [17]. The specificity profile of HML-2 Pro revealed P1 and P1', which refer to the two amino acids comprising the processed peptide bond, as major determinants of specificity of HML-2 Pro proteolysis. In particular, a preference for phenylalanine (F), tryptophan (W), tyrosine (Y) and glycine (G), thus primarily aromatic aa residues, was observed for P1. Whereas, phenylalanine (F), isoleucine (I), leucine (L), valine (V) and tryptophan (W), thus aromatic and aliphatic aa residues, were observed for P1'. Based on those published findings, we filtered the initial list of cleavage events by considering only proteins with F, G, Y, or W in P1, and F, I, L, V, or W in P1' (Figure 32 B). After applying filters for P1 and P1', 4370 (replicate 1) and 2633 (replicate 2) cleavage events were reduced to 355 and 260, respectively. Those cleavage events corresponded to 305 and 219 different human proteins, with 145 human proteins common to both datasets. Those 145 proteins were subjected to further filtering (Figure 32 C). Molecular mass of proteins had to be compatible with an *in vitro* coupled transcription/translation system. In general, proteins of larger molecular mass tend to require a longer time to be transcribed/translated or are produced relatively inefficiently by that system. That might have resulted in a (too) low yield of protein for our experiments. To ensure relatively high yield, we therefore considered only proteins with a length of less than 800 aa, or approximately 90 kDa, resulting in 107 different human proteins. Furthermore, we selected proteins that, based on associated GO terms, located to various cellular compartments, e.g., cytoplasm, cytosol, nucleus, endoplasmic reticulum, mitochondrion, and were furthermore known to be involved in relevant cellular processes, e.g.,

transcriptional regulation, growth control, protein folding, regulation of cell cycle. We also included one protein (C15orf57) for which cellular location and function were unknown. Eventually, we selected 14 different human proteins (Table 20) for subsequent experiments.
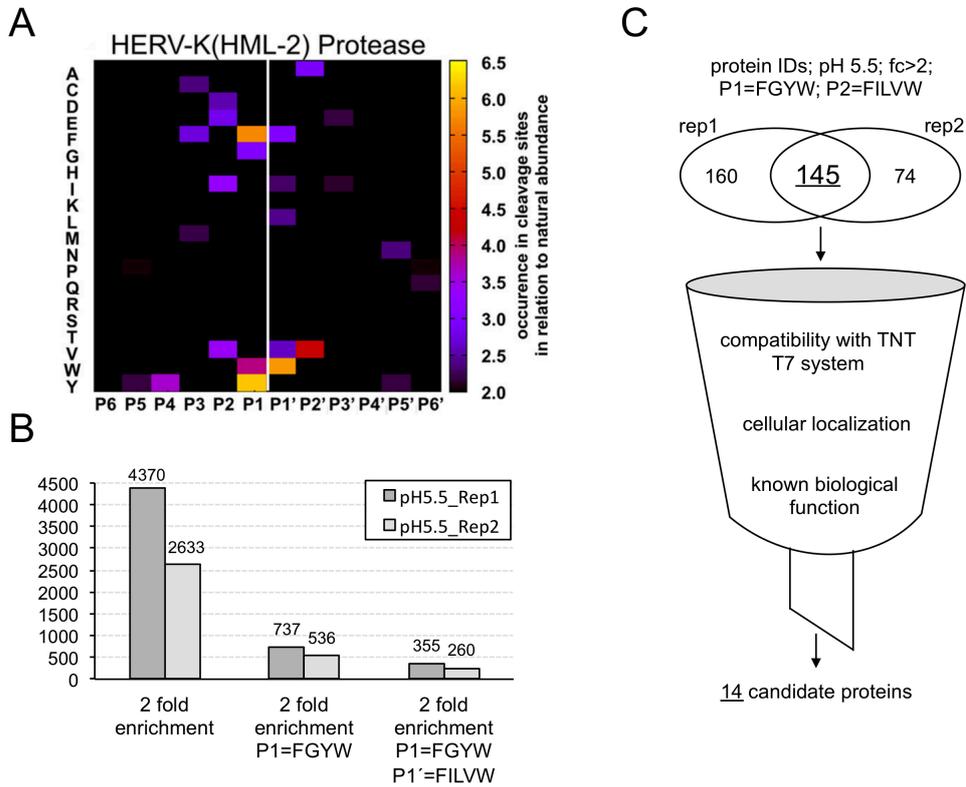
Figure 32: **Criteria for filtering of candidate proteins for *in vitro* verification experiments. A)** Specificity profile of HML-2 Pro as published recently [17]. The heatmap, retrieved from Biniossek et al. (2016) [17], sums up for observed cleavage sites the aa occurrences for each position, also taking into account frequencies of the various aa in human proteins. Aa specificities in positions P1 and P1´ were used as criteria for selection of candidate proteins for subsequent experiments *in vitro*. **B)** 2-fold enriched cleavage events identified in the two TAILS experiments at pH 5.5 were filtered by considering only cleavage events with F, G, Y, or W in P1, and F, I, L, V, or W in P1'. Bars depict numbers of cleavage events before and after filtering. Precise numbers of cleavage events are given above each bar. **C)** Further criteria used for selection of eventually 14 different human candidate proteins.

Table 20: Candidate proteins selected for verification of processing by HML-2 Pro *in vitro* *

Protein ID Gene/protein (protein full name)	GO cellular component	GO biological process
P27797 CALR (calreticulin)	acrosomal vesicle; cell surface; cytoplasm; cytosol; endocytic vesicle lumen; endoplasmic reticulum; endoplasmic reticulum lumen; endoplasmic reticulum-Golgi intermediate compartment membrane; external side of plasma membrane; extracellular exosome; extracellular region; extracellular space; focal adhesion; Golgi apparatus; integral component of lumenal side of endoplasmic reticulum membrane; intracellular; membrane; MHC class I peptide loading complex; nucleus; perinuclear region of cytoplasm; phagocytic vesicle membrane; polysome; proteinaceous extracellular matrix; sarcoplasmic reticulum lumen; smooth endoplasmic reticulum.	antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent; antigen processing and presentation of peptide antigen via MHC class I; ATF6-mediated unfolded protein response; cardiac muscle cell differentiation; cellular calcium ion homeostasis; cellular response to lithium ion; cellular senescence; chaperone-mediated protein folding; cortical actin cytoskeleton organization; glucocorticoid receptor signaling pathway; negative regulation of cell cycle arrest; negative regulation of intracellular steroid hormone receptor signaling pathway; negative regulation of neuron differentiation; negative regulation of retinoic acid receptor signaling pathway; negative regulation of transcription from RNA polymerase II promoter; negative regulation of transcription, DNA-templated; negative regulation of translation; negative regulation of trophoblast cell migration; peptide antigen assembly with MHC class I protein complex; positive regulation of cell cycle; positive regulation of cell proliferation; positive regulation of dendritic cell chemotaxis; positive regulation of DNA replication; positive regulation of endothelial cell migration; positive regulation of gene expression; positive regulation of NIK/NF-kappaB signaling; positive regulation of phagocytosis; positive regulation of substrate adhesion-dependent cell spreading; protein export from nucleus; protein folding; protein folding in endoplasmic reticulum; protein localization to nucleus; protein maturation by protein folding; protein stabilization; receptor-mediated endocytosis; regulation of apoptotic process; regulation of meiotic nuclear division; regulation of transcription, DNA-templated; response to drug; response to estradiol; response to testosterone; sequestering of calcium ion; spermatogenesis; vesicle fusion with endoplasmic reticulum-Golgi intermediate compartment (ERGIC) membrane.
O00571 DDX3X (DEAD-Box helicase 3, X-linked)	cytoplasm; cytoplasmic stress granule; cytosol; extracellular exosome; extracellular region; ficolin-1-rich granule lumen; mitochondrial outer membrane; nuclear speck; nucleus; secretory granule lumen.	cellular response to arsenic-containing substance; cellular response to osmotic stress; chromosome segregation; extrinsic apoptotic signaling pathway via death domain receptors; innate immune response; intracellular signal transduction; intrinsic apoptotic signaling pathway; mature ribosome assembly; negative regulation of apoptotic process; negative regulation of cell growth; negative regulation of cysteine-type endopeptidase activity involved in apoptotic process; negative regulation of intrinsic apoptotic signaling pathway; negative regulation of protein complex assembly; negative regulation of translation; neutrophil degranulation; positive regulation of apoptotic process; positive regulation of cell growth; positive regulation of chemokine (C-C motif) ligand 5 production; positive regulation of cysteine-type endopeptidase activity involved in apoptotic process; positive regulation of G1/S transition of mitotic cell cycle; positive regulation of gene expression; positive regulation of interferon-beta production; positive regulation of transcription from RNA polymerase II promoter; positive regulation of translation; positive regulation of translational initiation; positive regulation of viral genome replication; protein localization to cytoplasmic stress granule; response to virus; RNA secondary structure unwinding; stress granule assembly; transcription, DNA-templated; Wnt signaling pathway.
P06733 ENO1 (enolase 1)	cell cortex region; cell surface; cytoplasm; cytosol; extracellular exosome; extracellular space; M band; membrane; nucleus; phosphopyruvate hydratase complex; plasma membrane.	canonical glycolysis; gluconeogenesis; negative regulation of cell growth; negative regulation of hypoxia-induced intrinsic apoptotic signaling pathway; negative regulation of transcription from RNA polymerase II promoter; negative regulation of transcription, DNA-templated; positive regulation of ATP biosynthetic process; positive regulation of muscle contraction; positive regulation of plasminogen activation; response to virus; transcription, DNA-templated.
P08238 HSP90AB1 (heat shock protein 90 alpha family class B member 1)	apical plasma membrane; aryl hydrocarbon receptor complex; basolateral plasma membrane; brush border membrane; cell surface; cytoplasm; cytosol; extracellular exosome; extracellular region; ficolin-1-rich granule lumen; inclusion body; lysosomal membrane; melanosome; membrane; mitochondrion; nucleoplasm; nucleus; ooplasm; protein complex; secretory granule lumen; sperm head plasma membrane.	cellular response to drug; cellular response to interleukin-4; cellular response to organic cyclic compound; chaperone-mediated protein complex assembly; Fc-gamma receptor signaling pathway involved in phagocytosis; negative regulation of cell cycle arrest; negative regulation of complement-dependent cytotoxicity; negative regulation of neuron apoptotic process; negative regulation of proteasomal ubiquitin-dependent protein catabolic process; negative regulation of transforming growth factor beta activation; neutrophil degranulation; placenta development; positive regulation of cell differentiation; positive regulation of cell size; positive regulation of nitric oxide biosynthetic process; positive regulation of phosphoprotein phosphatase activity; positive regulation of protein binding; positive regulation of protein import into nucleus, translocation; positive regulation of protein localization to cell surface; positive regulation of protein serine/threonine kinase activity; positive regulation of telomerase activity; positive regulation of transforming growth factor beta receptor signaling pathway; protein folding; protein stabilization; regulation of cellular response to heat; regulation of interferon-gamma-mediated signaling pathway; regulation of protein ubiquitination; regulation of type I interferon-mediated signaling pathway; response to cocaine; response to salt stress; response to unfolded protein; supramolecular fiber organization; telomerase holoenzyme complex assembly; telomere maintenance via telomerase; virion attachment to host cell; xenobiotic metabolic process.
P30101 PDIA3 (protein disulfide isomerase family A member 3)	cell surface; endoplasmic reticulum; endoplasmic reticulum lumen; extracellular exosome; extracellular space; focal adhesion; melanosome; myelin sheath; nucleus; phagocytic vesicle; recycling endosome membrane.	antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent; antigen processing and presentation of peptide antigen via MHC class I; cell redox homeostasis; positive regulation of extrinsic apoptotic signaling pathway; protein folding; protein folding in endoplasmic reticulum; protein import into nucleus; protein retention in ER lumen; response to endoplasmic reticulum stress; signal transduction.

P43686 PSMC4 (proteasome 26S subunit, ATPase 4)	cytosol; cytosolic proteasome complex; inclusion body; membrane; nuclear proteasome complex; nucleoplasm; nucleus; proteasome accessory complex; proteasome complex; proteasome regulatory particle, base subcomplex; synapse.	anaphase-promoting complex-dependent catabolic process; antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent; blastocyst development; Fc-epsilon receptor signaling pathway; MAPK cascade; negative regulation of canonical Wnt signaling pathway; negative regulation of G2/M transition of mitotic cell cycle; negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle; NIK/NF-kappaB signaling; positive regulation of canonical Wnt signaling pathway; positive regulation of RNA polymerase II transcriptional preinitiation complex assembly; positive regulation of ubiquitin-protein ligase activity involved in regulation of mitotic cell cycle transition; proteasome-mediated ubiquitin-dependent protein catabolic process; protein deubiquitination; protein polyubiquitination; proteolysis; regulation of cellular amino acid metabolic process; regulation of mRNA stability; regulation of transcription from RNA polymerase II promoter in response to hypoxia; SCF-dependent proteasomal ubiquitin-dependent protein catabolic process; stimulatory C-type lectin receptor signaling pathway; T cell receptor signaling pathway; transmembrane transport; tumor necrosis factor-mediated signaling pathway; ubiquitin-dependent ERAD pathway; Wnt signaling pathway, planar cell polarity pathway.
P43487 RANBP1 (RAN binding protein 1)	centrosome; cytoplasm; cytosol; nuclear envelope; nucleus.	G1/S transition of mitotic cell cycle; positive regulation of mitotic centrosome separation; protein import into nucleus; RNA export from nucleus; signal transduction; spindle organization; ubiquitin-dependent protein catabolic process; viral process.
Q5TBB1 RNASEH2B (ribonuclease H2 subunit B)	nucleus; ribonuclease H2 complex.	in utero embryonic development; negative regulation of gene expression; positive regulation of fibroblast proliferation; regulation of DNA damage checkpoint; regulation of G2/M transition of mitotic cell cycle; ribonucleotide metabolic process; RNA catabolic process.
P26447 S100A4 (S100 calcium binding protein A4)	extracellular exosome; extracellular space; neuron projection; nucleus; perinuclear region of cytoplasm.	epithelial to mesenchymal transition; positive regulation of I-kappaB kinase/NF-kappaB signaling.
Q9BV29 C15orf57 (CCDC32; coiled-coil domain containing 32)	unknown	unknown
Q6FI81 CIAPIN1 (cytokine induced apoptosis inhibitor 1)	cytoplasm; mitochondrial intermembrane space; mitochondrion; nucleolus; nucleoplasm.	apoptotic process; hemopoiesis; iron-sulfur cluster assembly; negative regulation of apoptotic process.
P07900 HSP90AA1 (heat shock protein 90 alpha family class A member 1)	cytoplasm; cytosol; endocytic vesicle lumen; extracellular exosome; extracellular region; ficolin-1-rich granule lumen; lysosomal lumen; melanosome; membrane; myelin sheath; nucleoplasm; nucleus; plasma membrane; ruffle membrane; secretory granule lumen.	chaperone-mediated autophagy; chaperone-mediated protein complex assembly; ciliary basal body docking; ERBB2 signaling pathway; Fc-gamma receptor signaling pathway involved in phagocytosis; G2/M transition of mitotic cell cycle; mitochondrial transport; neutrophil degranulation; positive regulation of nitric oxide biosynthetic process; positive regulation of telomerase activity; protein import into mitochondrial outer membrane; protein refolding; protein stabilization; protein unfolding; receptor-mediated endocytosis; regulation of cellular response to heat; regulation of nitric-oxide synthase activity; regulation of protein complex assembly; regulation of protein ubiquitination; response to antibiotic; response to cold; response to heat; response to unfolded protein; signal transduction; telomerase holoenzyme complex assembly; telomere maintenance via telomerase; vascular endothelial growth factor receptor signaling pathway.
P36507 MAP2K2 (mitogen-activated protein kinase kinase 2)	cell-cell junction; cytoplasmic side of plasma membrane; cytosol; early endosome; endoplasmic reticulum; extracellular region; focal adhesion; Golgi apparatus; late endosome; microtubule; mitochondrion; nucleus; perinuclear region of cytoplasm; peroxisomal membrane; plasma membrane.	activation of MAPK activity; ERK1 and ERK2 cascade; MAPK cascade; negative regulation of gene expression; peptidyl-serine autophosphorylation; positive regulation of ERK1 and ERK2 cascade; positive regulation of production of miRNAs involved in gene silencing by miRNA; positive regulation of protein serine/threonine kinase activity; positive regulation of transcription, DNA-templated; proteolysis in other organism; regulation of early endosome to late endosome transport; regulation of Golgi inheritance; regulation of stress-activated MAPK cascade.
Q71U36 TUBA1A (tubulin alpha 1a)	cytoplasmic microtubule; cytoplasmic ribonucleoprotein granule; cytosol; extracellular exosome; membrane raft; microtubule; microtubule cytoskeleton; myelin sheath; nucleus; recycling endosome.	cell division; ciliary basal body docking; cytoskeleton-dependent intracellular transport; G2/M transition of mitotic cell cycle; microtubule-based process.

* A representative protein ID, approved gene/protein symbols and full names are given each. GO compartments and GO processes for respective genes/proteins were compiled from UniProt [40].

4.1.6.1.2 Generation of candidate proteins in an *in vitro* transcription/translation system

Candidate proteins were produced *in vitro* employing TNT® T7 Quick for PCR DNA, referred to as TNT T7 system in the following. The TNT T7 system consisted of a rabbit reticulocyte cell-free extract (hereafter designated as TNT T7 Master Mix), containing a T7 RNA polymerase and the essential molecular machinery for protein translation. For each selected candidate protein, a DNA template encompassing the protein coding region, as present in a respective coding sequence-harboring commercially available plasmid, was amplified by PCR. The PCR added to the DNA template, by means of specifically designed PCR primers, a T7 promoter, a spacer and a Kozak sequence upstream of the protein start codon. To the 3' end of the protein coding sequence were added an HA-tag and a stop codon (see Materials and Methods, Figure 12). PCR-generated DNA templates were added to the TNT T7 Master Mix in presence of either L-^{35}S-methionine or L-methionine to produce radiolabeled or "cold", respectively, proteins. Several parameters can influence the yield of protein produced by the utilized *in vitro* coupled transcription/translation system, for instance, DNA template characteristics, like the aa sequence encompassing the ATG translation initiation codon, amount of DNA template, and template purity. Time and temperature of incubation of reactions are also relevant. Thus, before setting up reactions for generation of candidate proteins, we confirmed that our experimental conditions (see also chapter 3.2.2.2) ensured an adequate yield of candidate protein. To do so, we produced an HA-tagged EGFP protein by using the TNT T7 system and we verified, by Western blot, that the protein was produced at a readily detectable amount (Figure 33 A). Indeed, EGFP was detected at reasonable amount with both an α-GFP and an α-HA antibody. The approximately 30 kDa band observed with both antibodies was of the expected size. No considerable amount of bands of side products was detected. As the HA-tag was present at the C-terminus, detection via the α-HA antibody furthermore confirmed that the protein was transcribed/translated in full-length. Subsequently, radiolabeled and "cold" candidate proteins were generated with established TNT T7 system conditions. Production of desired proteins was verified by SDS-PAGE followed by phosphorimaging or Western blot depending on ^{35}S or HA-only protein labels, respectively (Figure 33 B, C). The produced full-length candidate proteins matched the expected sizes. HML-2 Gag was also generated *in vitro* to be used as a positive control for cleavage by HML-2 Pro in the following experiments.
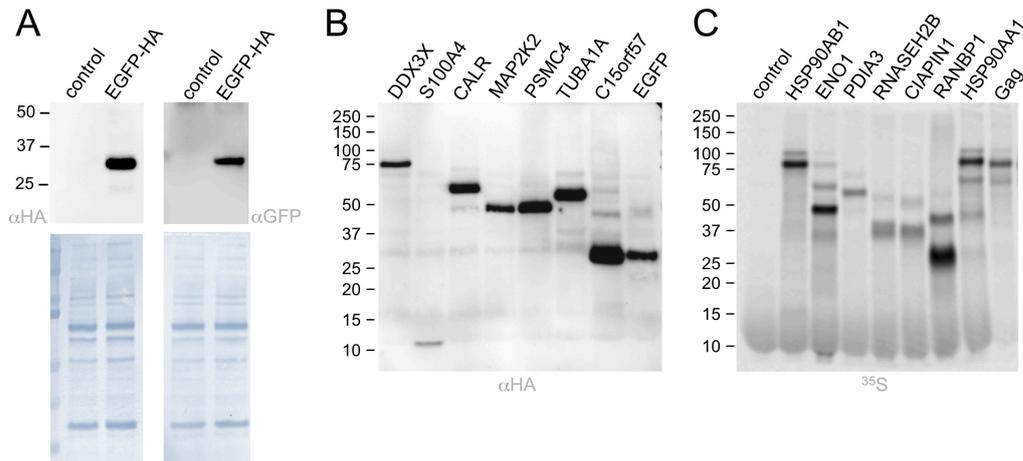
Figure 33: Generation of candidate proteins using an *in vitro* transcription/translation system. Proteins were produced via a TNT T7 system (see text) and detected by Western blot and phosphorimaging. **A)** HA-tagged EGFP produced in the TNT T7 system was detected by α-HA and α-GFP antibodies. Controls with TNT T7 mixtures without DNA templates are included. In the bottom part, blot membranes were stained with Coomassie after the detection procedure. **B)** Representative Western blot result with different candidate proteins produced *in vitro* and detected via HA-tag. **C)** Representative result of a dried SDS-PAGE gel with different ^{35}S-labeled candidate proteins produced *in vitro* and detected by phosphorimaging. HML-2 Gag was also included (rightmost lane). Note that in some cases, besides bands corresponding to full-length proteins, slower migrating, lower intensity bands were detected. We attributed those extra bands to incomplete denaturation of proteins before SDS-PAGE.

4.1.6.1.3 Incubation of candidate proteins with HML-2 Pro and verification of protein cleavage

In vitro transcribed/translated, radioactively or HA-labeled candidate proteins were incubated with purified HML-2 Pro to confirm processing by HML-2 Pro *in vitro*. For each candidate protein, 1 μl of the TNT T7 *in vitro* reaction was incubated with 400 nM purified HML-2 Pro. Two control reactions were included for each candidate protein tested: a reaction without HML-2 Pro and a reaction with HML-2 Pro enzymatic activity inhibited by presence of 400 μM pepstatin A. The buffer composition was the same as for the TAILS reactions, specifically 1 M NaCl and 0.1 M PIPES, pH 5.5. Reactions were incubated for 180 min at 37°C in a final volume of 16 μl, then the entire reaction volume was each loaded onto a PAA-gel (10% or 12% polyacrylamide) and subjected to SDS-PAGE. Protein detection was by phosphorimaging of dried gels for radiolabeled target proteins or Western blot for "cold" target proteins.

For our verification experiments, evidence of processing of candidate proteins by HML-2 Pro included (1) a more or less reduced amount of full-length candidate protein compared to amounts of full-length protein in control reactions without HML-2 Pro and with HML-2 Pro plus pepstatin A; (2) one or several additional protein bands of size(s) smaller than full-length protein in the reaction with HML-2 Pro compared to control reactions without HML-2 Pro or inhibited HML-2 Pro activity; (3) such additional protein bands also being present in the reaction with HML-2 Pro plus pepstatin A, yet at (much) lower amounts compared to the reaction with HML-2 Pro. Various combinations of those

criteria were observed in our verification experiments. For 8 candidate proteins, specifically C15orf57, CIAPIN1, HSP90AA1, HSP90AB1, MAP2K2, RANBP1, RNASEH2 and TUBA1A, we observed processing of candidate proteins along with generation of one or more cleavage products not visible in control reactions (Figure 34). For those candidate proteins we also observed a more or less pronounced reduction of full-length protein in presence of HML-2 Pro. In some cases (for instance HSP90AA1 and C15orf57), cleavage products were detectable also in the control reaction with HML-2 Pro plus pepstatin A, yet at much lower amounts compared to the reaction with HML-2 Pro. The same phenomenon was visible in reactions with HML-2 Gag, a specific substrate of HML-2 Pro. Such a phenomenon could result from higher specificity of HML-2 Pro for those substrates, thus it would be impossible to completely avoid such cleavages with the amount of pepstatin A included in controls. For PDIA3, evidence of cleavage was supported only by reduction of the full-length protein without appearance of detectable cleavage products. For DDX3X, the *in vitro* verification produced inconclusive results as no reduction of the amount of full-length protein was seen via the ^{35}S-label, although a lower molecular weight protein band of approximately 60 kDa increased in intensity in presence of HML-2 Pro. In contrast, when detecting the HA-tag, a slight reduction of full-length DDX3X protein was visible, yet without appearance of cleavage products. No evidence of processing by HML-2 Pro was obtained for CALR, ENO1, PSMC4 and S100A, for which no reduction of full-length protein or appearance of additional bands was observed.

Analysis of cleavage products generated by HML-2 Pro in combination with *in silico* analysis could provide indications about cleavage positions within the candidate protein of interest. The software Geneious provided functions necessary for this purpose. The precise amino acid sequence of a specific protein could be analyzed and the expected size in kDa for each desired portion of the protein could be predicted. Thus, using Geneious, we evaluated in which position of a candidate protein a cleavage might have generated fragments with a size also observed in *in vitro* verifications. For cleavage products detected via HA-tag, the C-terminus of the fragment was evident (visualized by detection of the HA-tag), thus the size of those products could provide unambiguous information concerning cleavage positions. On the contrary, for fragments detected via the ^{35}S-label the fragment termini were not known, thus the size of fragments provided rather ambiguous information for predicting cleavage positions. In order to reduce ambiguities, considering that TAILS experiments had provided information on actual cleavage site positions in candidate proteins, we evaluated similarities between data from the TAILS experiments at pH 5.5 and results from *in vitro* verifications (Figure 34 B).
As for C15orf57, the three cleavage sites identified by TAILS could effectively match with the products observed via the ^{35}S-label (Figure 34 B). A cleavage product of ~20 kDa detected via the ^{35}S-label was confirmed by HA-tag detection. It should be noted that, as indicated by TAILS results, for several candidate proteins cleavages occured in the proximity of the C-terminus, thus close to the HA-tag. Cleavages in those positions eventually produced 1) C-terminal HA-fragments of small size that,

for technical limitations, could not be detected; 2) N-terminal fragments that could be detected only via the ^{35}S-label. Two cases that likely represent such a situation are CIAPIN1 and RANBP1, whose N-terminal fragments after cleavage were visible via ^{35}S-label detection, while no cleavage product was visible by HA-tag detection. However, for both candidates, observed cleavage products matched with the cleavage sites identified by TAILS. As for MAP2K2, the cleavage product of ~40 kDa observed via ^{35}S-label and HA-tag, matched clearly with the unique cleavage site identified by TAILS. For TUBA1A the only cleavage product of ~50 kDa observed by HA-tag could correspond to one of the two cleavage sites identified by TAILS in the N-terminal part of the protein and replicated in pH 5.5 TAILS experiments. For RNASEH2B a cleavage product of ~28 kDa was observed by ^{35}S-label. Such cleavage product did not perfectly match with the cleavage product identified in both pH 5.5 TAILS experiments and predicted as a ~22 kDa peptide. The ~6 kDa discrepancy between expected and observed could be due to specific (structural or chemical) properties of that specific peptide, likely affecting migration during electrophoresis. In two cases, HSP90AA1 and HSP90AB1, processing products observed via the ^{35}S-label were difficult to assign due to too many cleavage sites reported by TAILS, although the only cleavage product observed via HA-tag matched with cleavage sites identified by TAILS.

Taken together, our *in vitro* verification experiments confirmed processing by HML-2 Pro for 9 out of 14 candidate proteins. Moreover, analysis of cleavage products generated by HML-2 Pro provided insight into cleavage site positions. For 6 of the candidate proteins, molecular weights of additional protein bands detected *in vitro* appeared to match sizes of processing products predicted by cleavage sites as identified in TAILS experiments.
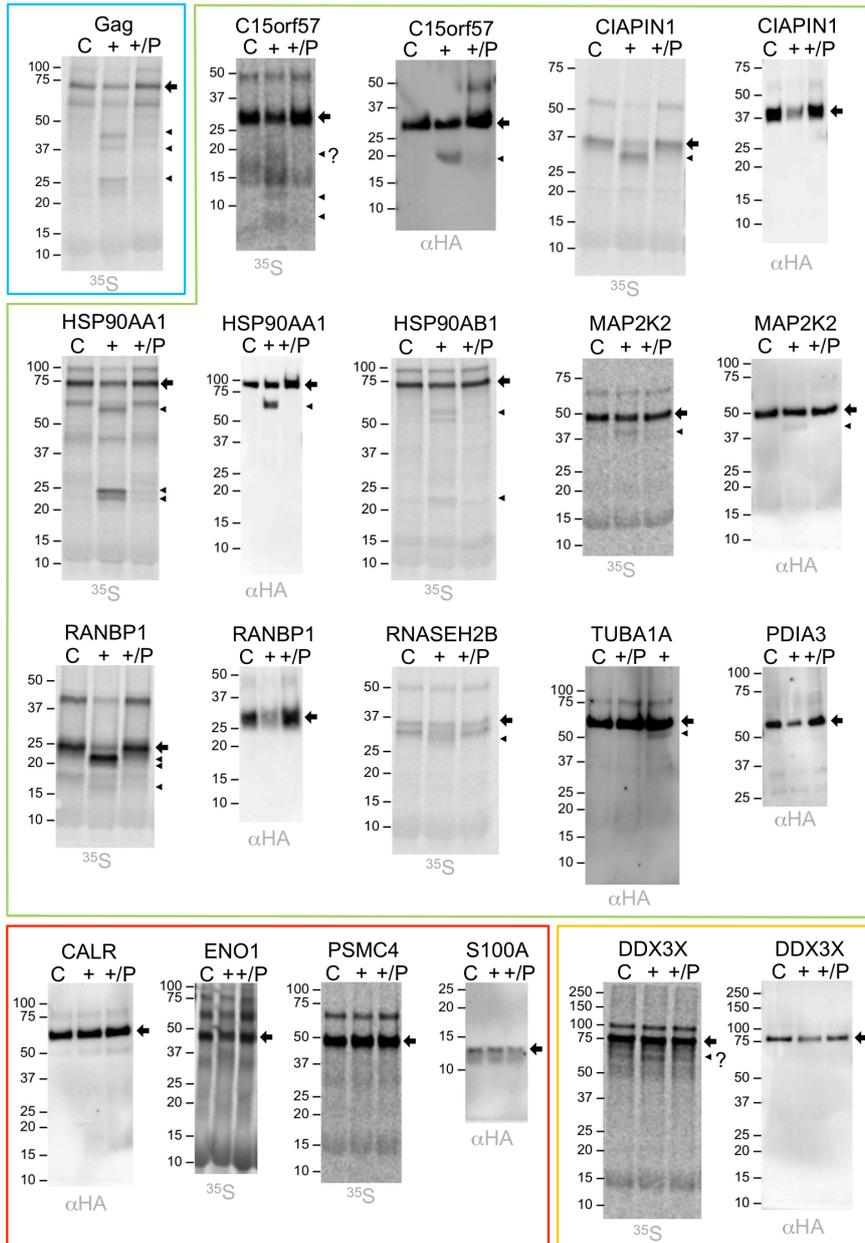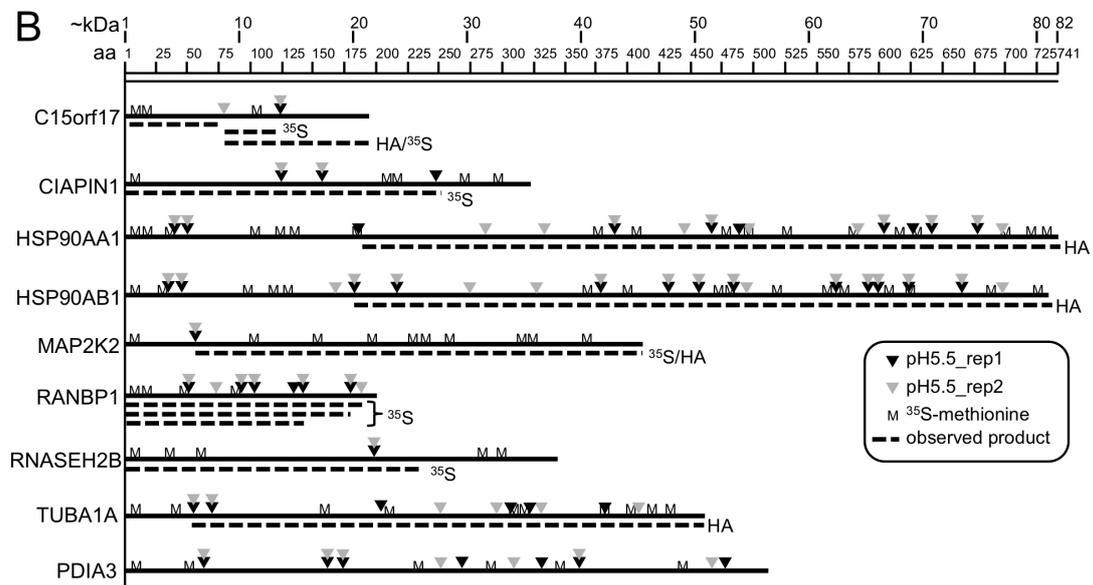
Figure 34: Verification of cleavage of candidate proteins by HML-2 Pro *in vitro*. A) Selected Western blot and phoshporimaging results from incubation of candidate proteins with HML-2 Pro. Each candidate protein was incubated for 180 min and with three different reactions: one without protease ("C"), one with protease ("+"), and one with protease plus pepstatin A ("+/P"). Reactions were subjected to SDS-PAGE in 10% or 12% polyacrylamide gels and further processed for subsequent phosphorimager analysis or HA-tag-specific Western blots depending on the protein label ("^{35}S" or "HA"). Full-length candidate proteins are indicated by an arrow, while arrowheads indicate additional protein bands ascribed to HML-2 Pro activity. Processing of Gag by HML-2 Pro was included as a control (blue frame). Results for candidate proteins with evidence of specific cleavage by HML-2 Pro are in a green frame. Results for candidate proteins with no evidence of specific cleavage by HML-2 Pro are in a red frame. Experiments with uncertain evidence of cleavage by HML-2 Pro (see text) are in an orange frame. For PDIA3 no lower-sized processing products could be observed, although specific cleavage by HML-2 Pro was supported by reduction of the amount of full-length protein. Note that C15orf57 migrated slower in gel electrophoresis than predicted by molecular mass. **B)** Depiction of candidate proteins that were confirmed to be processed by HML-2 Pro. Each protein is depicted as a line with number of amino acids and molecular mass indicated by scales on the top. Indicated for each protein are positions of methionines ("M") and cleavage sites (grey and black arrowheads), the latter as identified by TAILS experiments at pH 5.5. Dashed lines indicate processing products observed experimentally for candidate proteins either labeled by ^{35}S-methionine ("^{35}S") or HA-tag ("HA"). Note that the latter protein label will only detect C-terminal processing products. For the two HSP90 proteins, cleavage products observed via ^{35}S-label were difficult to assign because of too many cleavage sites identified by TAILS.

4.1.6.2 Verification experiments *in vivo*

The *in vitro* verification experiments had confirmed the ability of HML-2 Pro to cleave several of the human proteins tested. We next were interested in whether such a proteolytic activity of HML-2 Pro could also be observed *in vivo*. To do so, five different HA-tagged human candidate proteins that we had confirmed to be substrates of HML-2 Pro *in vitro* were co-expressed together with HML-2 Pro in HEK293T cells. Processing of candidate proteins was then monitored by Western blot through detection of the HA-tag.

The five selected HA-tagged candidate proteins were as follows: four candidate proteins (C15orf57-HA, HSP90AA1-HA, MAP2K2-HA, TUBA1A-HA) that had produced processing products detectable via the HA-tag in *in vitro* experiments; one candidate protein (CIAPIN1A-HA) that had produced a processing product that was detectable only via the ^{35}S-label in *in vitro* experiments (see above).

Of note, the HML-2 Pro precursor expressed in mammalian cells was slightly different from the HML-2 Pro precursor expressed in bacteria for Pro purification (compare Figure 7 and Figure 8). In particular, the Pro ORF subregion cloned into constructs for expression in mammalian cells included 1) 33 bp upstream of the Pro core domain, thus a shorter N-terminal region in comparison with Pro constructs for Pro purification; 2) the region downstream of the Pro core domain up to the end of the Pro ORF, with that region not included in Pro constructs for Pro purification. The known Pro self-cleavage site in the N-terminus was still present. Since the complete C-terminal portion of the Pro ORF was included, the Pro self-cleavage site in the C-terminus was also present, although its position was not precisely known, since it has not been mapped so far.

For verifications *in vivo*, HEK293T cells were seeded at a density of 2×10^5 cells per well in 12-well plates. Cells were transfected the following day. We employed two transfection set-ups. In a first set-up, cells were co-transfected with pcDNA3 plasmid producing an HA-tagged candidate protein and a phCMV plasmid producing either wild type (Pro-wt) or mutant (Pro-mut) HML-2 Pro. As a control, cells were co-transfected with plasmid producing the candidate protein and an empty phCMV vector. In a second set-up, cells were co-transfected with pcDNA3 plasmid producing an HA-tagged candidate protein and a pEGFP-Pro plasmid producing either wild type (EGFP-Pro) or mutant (EGFP-Pro-mut) HML-2 Pro, each fused in frame to an EGFP ORF at the N-terminus. As a control, cells were co-transfected with plasmid encoding a candidate protein and a pEGFP plasmid, producing sole EGFP. Transfection set-ups for verifications of processing of candidate proteins by HML-2 Pro *in vivo* were as follows:

Set-up 1

Pro-wt + candidate:	0.5 µg phCMV-Pro + 0.5 µg pcDNA3-candidate protein
Pro-mut + candidate:	0.5 µg phCMV-Pro mut + 0.5 µg pcDNA3-candidate protein
Control:	0.5 µg phCMV + 0.5 µg pcDNA3-candidate protein

Set-up 2

EGFP-Pro + candidate protein:	0.5 µg pEGFP-Pro + 0.5 µg pcDNA3-candidate protein
EGFP-Pro-mut + candidate protein:	0.5 µg pEGFP-Pro-mut + 0.5 µg pcDNA3-candidate protein
Control:	0.5 µg pEGFP + 0.5 µg pcDNA3-candidate protein

Cells were harvested 24 hpt and lysed in RIPA Buffer enriched with a protease inhibitor cocktail. Since the cocktail of protease inhibitors employed did not include aspartyl protease inhibitors, pepstatin A was supplemented to the lysis buffer at a final concentration of 1 µg/µl. Cell lysates were subjected to reducing SDS-PAGE. 15 µg total protein per gel lane were loaded onto PAA-gels (10% or 12% polyacrylamide). Proteins of interest were detected by Western blot. HA-tagged candidate protein was detected using an α-HA antibody. Expression of HML-2 Pro and EGFP-Pro was monitored using either a polyclonal α-HML-2 Pro antibody 9367 [185] or an α-GFP antibody.

Criteria that we applied for verification of candidates were as for the verification experiments *in vitro* (see above). For all blots, staining of proteins by Coomassie was used as a loading control, confirming no significant difference of total protein loaded among samples.

Results of verification experiments *in vivo* can be summarized as follows. When expressing Pro and Pro-mut, HML-2 Pro could be detected as ~18 kDa and ~19 kDa protein bands representing self-processed Pro and unprocessed Pro, respectively (Figure 35 Aa, bottom blot). When HML-2 Pro was expressed as EGFP-Pro or EGFP-Pro-mut fusion protein, proteins of ~30 kDa and ~47 kDa, representing processed and unprocessed EGFP(-Pro), respectively, could be detected with an α-GFP

antibody (Figure 35 Ab, middle blot). The 30 kDa processed EGFP included a short N-terminal flanking region of Pro (see Figure 8). Self-processed Pro and unprocessed EGFP-Pro(-mut) of ~18 kDa and ~50 kDa, respectively, could be detected when using an α-HML-2 Pro antibody (Figure 35 Ab, bottom blot; Ac).

The verification experiments provided evidence of processing of candidate proteins by HML-2 Pro *in vivo*. For HSP90AA1, a slight reduction of full-length protein and the appearance of an additional fragment of approximately 65 kDa was observed only when co-expressing wild-type HML-2 Pro. Interestingly, the size of the processing product generated *in vivo* corresponded with the one observed *in vitro* (Figure 35 A, compare proteins labeled with arrowheads in panel Aa and Ab with that in Ad). For C15orf57-HA, the amount of full-length protein was reduced to below detection limits in presence of active HML-2 Pro and a cleavage product of ~20 kDa was generated, that furthermore coincided in size with the one observed *in vitro*. For MAP2K2-HA, there was a discernible reduction of full-length protein in presence of active HML-2 Pro and appearance of a ~40 kDa processing product. A processing product of the same size was also observed in the *in vitro* verification experiment. For CIAPIN1-HA, there was evidence of processing by HML-2 Pro *in vivo* from a considerable reduction of full-length protein, as already observed *in vitro* through HA-tag detection. Also, for TUBA1A-HA, reduction of full-length protein was considerable but the cleavage product observed *in vitro* could not be deteced *in vivo,* possibly due to reduced stability of that product *in vivo* (compare Figure 35 A and B).

Taken together, we obtained experimental evidence of self-processing of active HML-2 Pro and processing of candidate proteins by HML-2 Pro *in vivo*. Evidence for the latter came from more or less pronounced reduction of the amount of full-length candidate proteins and, for three candidate proteins, also from processing products of lower molecular weight. Observed processing products were furthermore similar in size to processing products detected *in vitro*.
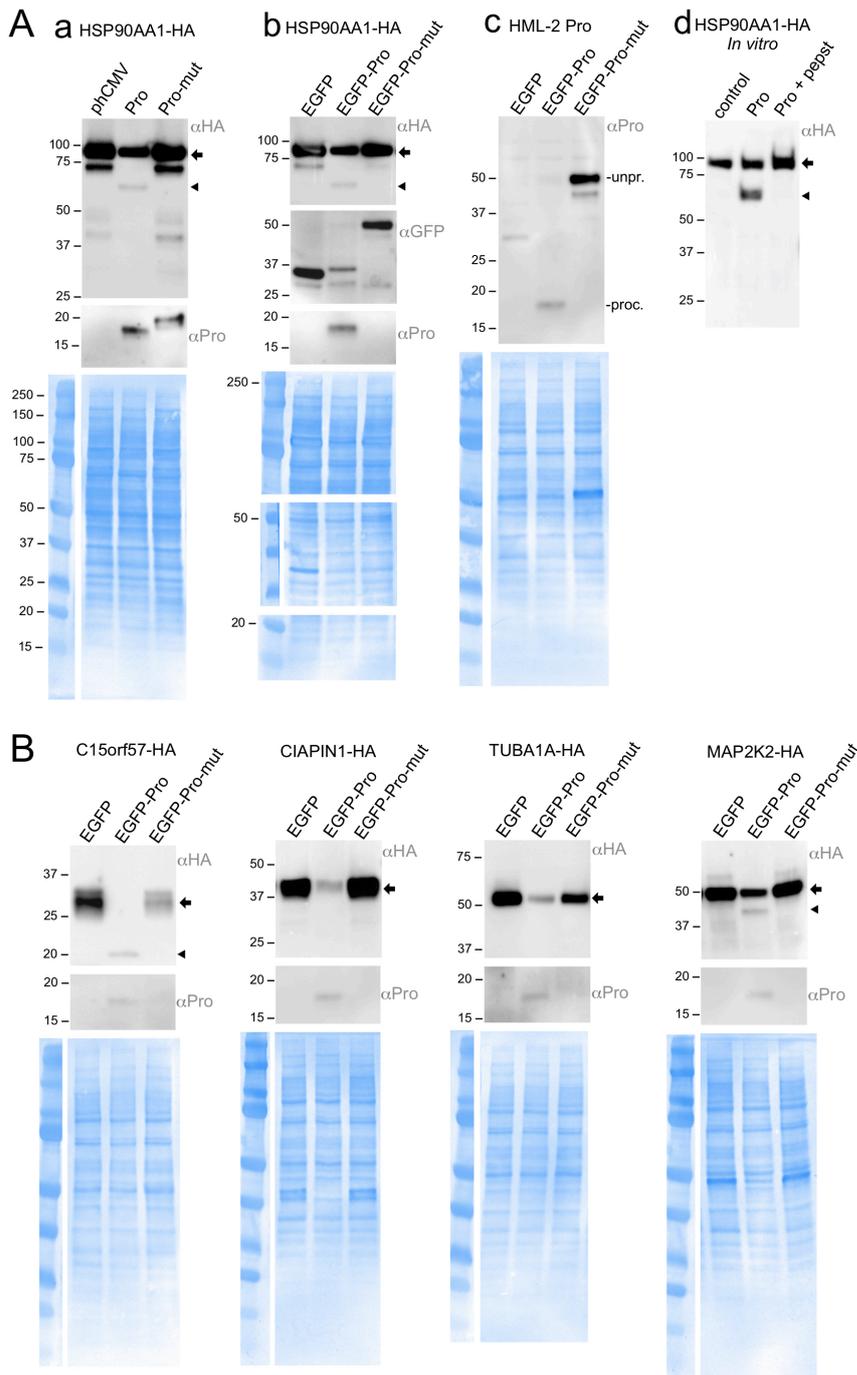
Figure 35: Experimental verification of cleavage of candidate proteins by HML-2 Pro *in vivo.* Five different candidate proteins were co-expressed together with HML-2 Pro in HEK293T cells and subsequently detected by Western blot in order to monitor their processing *in vivo*. For each blot, the left-most lane was a control co-transfected with plasmid encoding candidate protein and either a GFP-encoding plasmid or empty phCMV. Loaded onto middle and right-most lanes were candidate proteins co-expressed with wild-type Pro (Pro) or mutant Pro (Pro-mut), respectively. Pro was expressed as either sole Pro or EGFP-fused Pro. Full-length candidate proteins are indicated by arrows. Arrowheads indicate clearly discernible cleavage products. Blots were probed with α-HA, α-GFP, or α-Pro as indicated. **A)** Representative results from experiments co-expressing HSP90AA1 with either HML-2 Pro (**Aa**) or EGFP-Pro (**Ab** and **Ac**). Relevant blot regions are shown. Note the slight reduction of amounts of full-length HSP90AA1 and formation of a processing product corresponding to the product observed *in vitro*. For the sake of convenience, a blot from *in vitro* verification (already depicted in Figure 34 A) is shown (**Ad**). **B)** Selected Western blot results from co-expression of candidate proteins and HML-2 Pro. Note the reduction of amounts of full-length candidate protein, and sometimes processing products, in lanes with co-expressed HML-2 Pro. For experiments shown in both A and B, blotted membranes were stained with Coomassie after the ECL step to document same amounts of protein loaded onto gel lanes.

4.1.6.3 Additional cellular proteins verified as substrates of HML-2 Pro

During our project, we established a collaboration with Dr. John Goodier (McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA). An additional set of potential substrates of HML-2 Pro was subjected to verification experiments *in vivo* in the laboratory of Dr. Goodier. Additional tested proteins were of functional interest and readily available in that laboratory as cloned cDNAs. Of note, tested proteins were labelled with tags other than HA-tag. In particular, proteins tagged with FLAG, Myc, T7 or HA tag were included.

The additional tested proteins had been identified in our TAILS analysis with replicated evidence of processing by HML-2 Pro and at least 2-fold enrichment. Although, most of those proteins were not included in the list of 145 proteins filtered on the basis of favorable amino acids (for cleavage by HML-2 Pro) in P1 and P1´ (see chapter 4.1.6.1.1). This indicated that those additional verification experiments could also provide crucial information on cleavage of candidate proteins *in vivo* with no favorable amino acids in P1 and P1´ positions of cleavage sites as identified by TAILS.

As in our set-up for verification experiments, each protein was co-expressed with HML-2 Pro (or the respective mutant) in HEK293T cells. Different from our set-up, during expression of recombinant proteins, cells were treated approximately 18 hpt with proteasome inhibitor MG132 (Millipore-Sigma), then incubated for another 4–5 h before cell lysis. MG132 was included with the intent of reducing instability of potential cleavage products, thus facilitating validation of additional HML-2 Pro substrates by detecting cleavage products. Processing of proteins was monitored by SDS-PAGE and Western blot. Proteins were detected via the specific tag present in each protein.

Results produced in the laboratory of Dr. Goodier provided evidence of processing of 16 additional cellular proteins by HML-2 Pro *in vivo* (see Table 22 and [164]). Fourteen of those proteins had no favorable amino acids in P1 and P1´ positions of cleavage sites as identified by TAILS. Selected examples of relevant proteins are reported in the following. Among the 16 validated substrates were ribonuclease H2 (RNASEH2) subunit A and subunit B (RNASEH2A and RNASEH2B) that are two out of the three subunits that comprise the functional RNASEH2 protein. Notably, the processing of RNASEH2B by HML-2 Pro had been verified already by us *in vitro*. The list further included: endoplasmic reticulum chaperone BiP (HSPA5) and heat shock protein beta-1 (HSPB1), members of the family of the heat shock proteins [40], which also includes HSP90AB1 and HSP90AA1, the latter verified by us *in vitro* and *in vivo*; STIP1 homology and U-Box containing protein 1 (STUB1), that targets misfolded chaperone substrates towards proteasomal degradation; eukaryotic translation initiation factors 4 gamma 1 (EIF4GI) and 4B (EIF4B), two members of the eIF4 translation initiation factor family; transcription intermediary factor 1-beta (TRIM28), that mediates transcriptional control.

For all 16 validated candidates, a significant reduction of full-length protein was observed in presence of active HML-2 Pro. In some cases, cleavage products could also be visualized.

A supplementary experiment was performed in the laboratory of Dr. Goodier in order to document processing of constitutively expressed HSP90 protein by HML-2 Pro. Specifically, HSP90 was monitored via specific α-HSP90 antibody in cells expressing recombinant HML-2 Pro. Cleavage of endogenous HSP90 by HML-2 Pro was confirmed by generation of two cleavage products, one of which showed a size similar to that of a cleavage product detected in our experiments testing HSP90AA1 and HSP90AB1.

Overall, results of experiments performed in the laboratory of Dr. Goodier significantly expanded the number of cellular substrates of HML-2 Pro verified in the cellular context. The analysis also provided evidence of processing by HML-2 Pro of candidate proteins having no favorable amino acids in P1 and P1′ positions of cleavage sites as identified by TAILS. The additional findings corroborated our TAILS results and provided further support to our *in vivo* verifications. Small variations between our experimental set up and the one of Dr. Goodier, for instance, monitoring of protein cleavage using various protein tags, further corroborated reliability of our observations *in vivo*.

4.1.7 Initial investigation into the potential involvement of HML-2 Pro in cell death

When expressing active HML-2 Pro cells in cell culture, we noticed that approximately 24 hpt an unusually high number of detached rounded cells were floating in the medium. Controls without HML-2 Pro or with inactive HML-2 Pro did not produce such a phenotype. This suggested a cytotoxic effect of HML-2 Pro, as already observed for other retroviral proteases, e.g., HIV-1 Pro [18]. To further address this phenomenon, we performed a number of experiments to gain further insight into a potential involvement of HML-2 Pro in cell death.

4.1.7.1 Analysis of cell morphology during transient expression of HML-2

In one preliminary experiment investigating HML-2 Pro cytotoxicity, we used fluorescence microscopy for direct observation of Pro-expressing cells identified by staining with an α-HML-2 Pro antibody 9367 [185]. HeLa cells were used in this experiment because they were found more suitable than HEK293 because of much stronger adhesion to the cell culture chamber employed for this procedure. HeLa cells were seeded in an 8-well cell culture chamber suited for immunofluorescence staining. After 24 h, cells were transfected with phCMV-Pro and phCMV-Pro mut plasmids for expression of wild-type or mutant HML-2 Pro, respectively. Cells were fixed 24 hpt and immunostained with an α-HML-2 Pro polyclonal antibody 9367 [185]. Cells were mounted in presence of DAPI, which stains cell nuclei, followed by fluorescence microscopy (Figure 36).

Among cells transfected with expression plasmid encoding wild-type Pro, some cells displayed an altered rounded morphology with a condensed nucleus, implying cell death potentially due to apoptosis. Immunostaining confirmed that cells with altered morphology harbored HML-2 Pro (Figure

36), although not all Pro-expressing cells displayed the particular morphology. Cell morphology and viability appeared normal when expressing mutant HML-2 Pro, thus further suggesting that cell death was associated with HML-2 Pro activity.
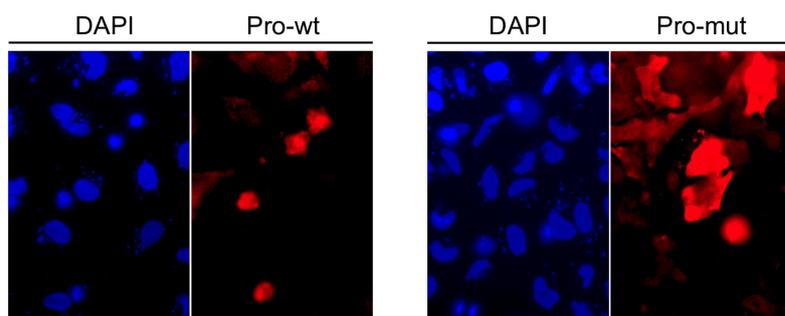
Figure 36: Altered cell morphology following expression of HML-2 Pro in HeLa cells. HeLa cells were transfected with plasmids expressing either enzymatically active (Pro-wt) or inactive (Pro-mut) HML-2 Pro. Cells were fixed 24 hpt and immunostained with an α-HML-2 Pro polyclonal antibody n. 9367 [185]. Cells were mounted in presence of DAPI. Cell nuclei (blue) and HML-2 Pro (red) were visualized by fluorescence microscopy.

4.1.7.2 Analysis of apoptosis markers during transient expression of HML-2 Pro

We further investigated whether cell death caused by HML-2 Pro involved apoptosis, a type of programmed cell death that can be triggered by different stimuli. In rather different ways, apoptosis initiating events lead up to activation of so-called caspases (cysteine aspartases), members of a family of proteases that carry out the process of cell death in a controlled manner through an irreversible cascade of proteolytic cleavages. In order to be activated, caspases need to be released through proteolytic cleavage of inactive precursors called procaspases. Some procaspases are designated as initiator procaspases because they are activated at the beginning of the proteolytic cascade. So-called executioner procaspases are activated in an advanced stage of the apoptotic process. Once activated, all caspases target several hundred proteins for proteolysis [202]. Poly [ADP-ribose] polymerase 1 (PARP), a DNA single strand repair enzyme, is a known substrate of caspases [23]. PARP cleavage, as well as the cleavage of procaspases to produce active caspases, provide evidence of apoptosis activation. Therefore, PARP and (pro)caspases are commonly used as apoptosis markers (for instance see [125]).

In our experiments, HeLa cells were transfected with phCMV-Pro plasmid to express wild-type HML-2 Pro. As negative controls, transfections with empty phCMV plasmid and plasmid expressing mutant, enzymatically inactive HML-2 Pro were performed. Moreover, as a positive control for apoptosis activation, HeLa cells were treated for 6 h with staurosporine, a strong inducer of apoptosis [12]. A sample with untreated cells was also included. Apoptosis biomarkers (pro)caspase 3 and PARP were monitored by Western blot at 6 and 24 h post transfection to evaluate apoptosis activation. Procaspase 3 is one of the executioner procaspases that during apoptosis is proteolytically cleaved releasing active caspase 3. Both procaspase 3 and caspase 3 were detected using a caspase 3 monoclonal antibody. PARP was monitored using a monoclonal antibody for specific detection of the 89 kDa PARP fragment

(referred as cleaved PARP) that is generated from the full-length PARP by active caspases [202]. Antibodies were provided as a cocktail that included also an α-muscle actin rabbit antibody as a loading control for normalization purposes.

As for Western blot results, induction of apoptosis could be inferred from increase of cleaved PARP, decrease of procaspase 3 and increase of caspase 3, the three events being clearly visible in samples treated with staurosporine (Figure 37) that strongly activated apoptosis of the entire cell population. No apoptosis was induced in untreated cells. Concerning cells expressing HML-2 Pro, levels of cleaved PARP, procaspase 3, and caspase 3 were not significantly different at 6 hpt when comparing cells expressing wild-type HML-2 Pro with negative controls, thus indicating that during the 6 h following transfection, apoptosis was not specifically induced by active HML-2 Pro, probably due to low level of HML-2 Pro at that time point. However, at 24 hpt, the level of cleaved PARP and caspase 3 in cells expressing wild-type HML-2 Pro was ~2-fold higher than that of negative controls, thus providing evidence of activation of apoptosis due to HML-2 Pro activity. As expected, in the presence of active HML-2 Pro, the increase in caspase 3 was also accompanied by a decrease in procaspase 3. Taken together, apoptosis marker proteins indicated that HML-2 Pro activity could activate apoptotic processes.
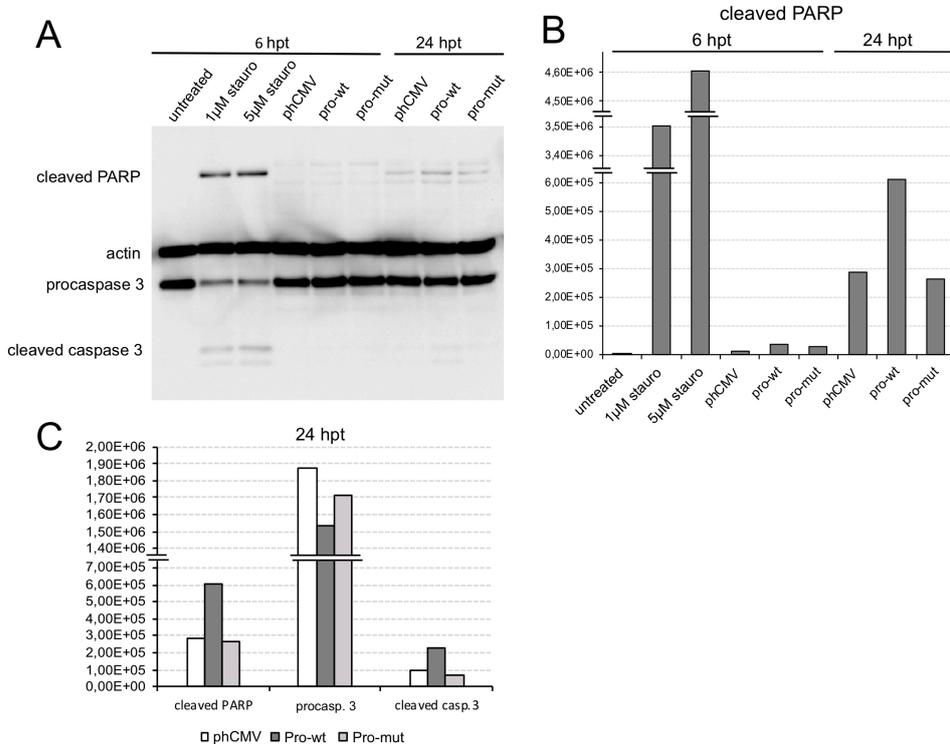
Figure 37: Activation of apoptosis marker proteins in HeLa cells expressing HML-2 Pro. HeLa cells were transfected with plasmid encoding either HML-2 Pro (Pro-wt) or an enzymatically inactive mutant HML-2 Pro (Pro-mut), or empty plasmid (phCMV). Cells were harvested and lysed 6 and 24 hpt. As a positive control for apoptosis, cells were treated for 6 h with 1 μM or 5 μM staurosporine. **A)** Cleaved PARP, procaspase 3, cleaved caspase 3, and actin, were detected by Western blot using specific monoclonal antibodies in an antibody cocktail. **B)** Bar chart depicting, for all samples, normalized intensities of proteins bands corresponding to cleaved PARP at 6 and 24 hpt. **C)** Bar chart comparing normalized intensities of protein bands corresponding to cleaved PARP, procaspase 3 and cleaved caspase 3 for samples at 24 hpt. Note that the amount of cleaved PARP and caspase 3 increases while the amount of procaspase 3 decreases when expressing wild-type HML-2 Pro. For relative quantification, band intensities of proteins of interest were normalized by actin signal intensities.

4.1.7.3 Monitoring of timing and intensity of HML-2 Pro cytotoxic effect using FACS

To further characterize a cytotoxic effect of transiently expressed HML-2 Pro, we evaluated in more detail timing and intensity of such an effect using Fluorescence-Activated Cell Sorting (FACS). For this analysis, we employed plasmids expressing EGFP-fused Pro (EGFP-Pro) and the respective mutant (EGFP-Pro-mut). The same plasmid constructs were employed for verifications of cleavage of candidate proteins *in vivo* (see chapter 4.1.6.2), where it was shown that enzymatically active HML-2 Pro, fused in-frame with EGFP, could self-process with similar efficiency as "sole" Pro. Self-processing of EGFP-Pro released self-processed HML-2 Pro and EGFP (see Figure 35 Ab), the latter including a short C-terminal protein portion of the Pro ORF upstream of the N-terminal self-processing site (see Figure 8). As a consequence, and important for interpretation of experiments, EGFP-positive cells also contained free, enzymatically active HML-2 Pro.

We firstly evaluated by fluorescence microscopy the impact of EGFP-Pro expression on cell viability. To do so, HEK293T cells were seeded in a 12-well plate and transfected with plasmid constructs for expression of EGFP-Pro, with or without 1 µM indinavir, a known strong inhibitor of HIV-1 Pro. As a control, cells were transfected with EGFP-Pro-mut. EGFP-positive cells were examined under the fluorescence microscope 30 hpt (Figure 38). Cell viability was altered when EGFP-Pro was expressed in cells. Seemingly dead cells of a rounded morphology, in part detached and floating in the medium were observed. The majority of those cells appeared EGFP-positive, although not all green, thus Pro-expressing cells appeared to have undergone cell death until 30 hpt. Cell death appared considerably reduced when EGFP-Pro was expressed in presence of 1 µM indinavir, that hampered Pro activity. Cell viability appeared unaffected when expressing EGFP-Pro-mut, that is proteolytically inactive. Overall, the experiment provided evidence of a cytotoxic effect induced by active HML-2 Pro. Even with a set-up involving EGFP-Pro, the effect of HML-2 Pro on cell viability and cell phenotype was in line with observations from experiments expressing "sole" Pro (see chapter 4.1.7.1).
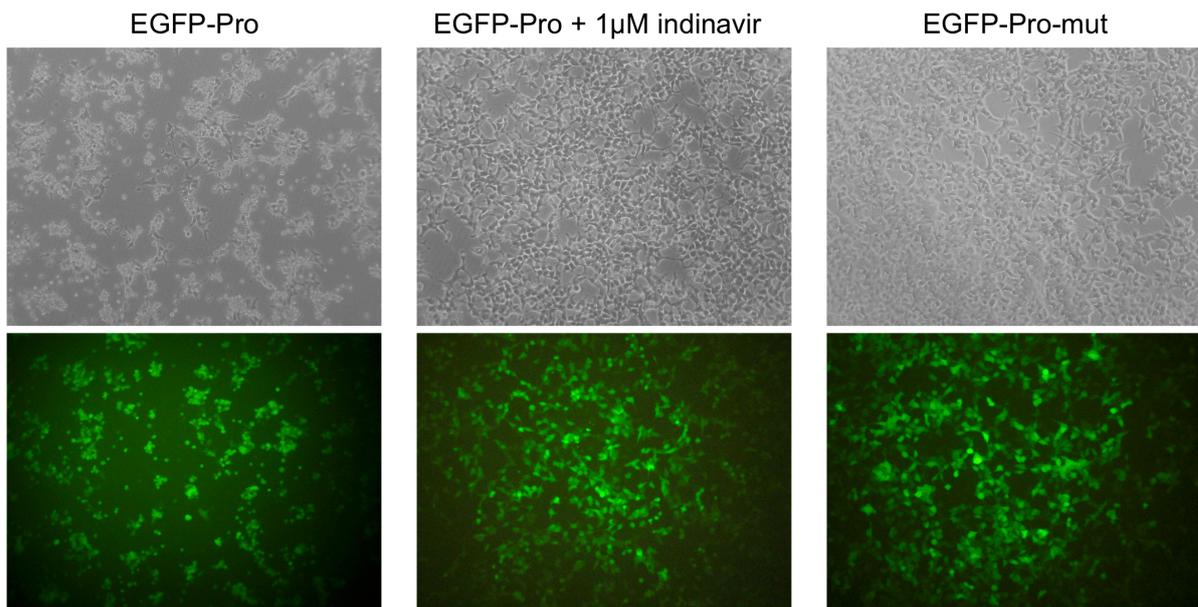
Figure 38: Microscopy images documenting cytotoxic effects following expression of EGFP-Pro in HEK293T cells. HEK293T cells were transfected with plasmid constructs for expression of EGFP-fused HML-2 Pro (EGFP-Pro) or respective inactive mutant (EGFP-Pro-mut). EGFP-Pro was also expressed in presence of 1 μM indinavir, a known inhibitor of HIV-1 Pro. EGFP-positive cells were examined under the fluorescence microscope 30 hpt. Note the less dense cell layer when expressing EGFP-Pro.

We next monitored at different time points, using FACS, the percentage of dead cells in a population of cells transiently expressing HML-2 Pro. We transfected HEK293T cells with plasmids expressing either EGFP, EGFP-Pro, or EGFP-Pro-mut. Cells were harvested 5, 10, 24, 30, and 48 hpt, fixed at each time point and subsequently analyzed by FACS. A control with transfection reagent only, but no plasmid DNA, was also included.

Employing forward and side scatter density plots (FSC-A/SSC-A), we gated the main population of cells, within size and granularity ranges for live HEK293T cells (Figure 39 A). Cells with FSC-A and/or SSC-A values above the range of values of the main population probably represented single events consisting of 2 or more independent cells (likely not properly detached). Such events were excluded from the analysis. The population of cells with FSC-A values below the range of values of the main population represented dead cells. That cell population was gated and quantified relative to the total number of cells counted. The percentage of EGFP-positive cells in the population of live and dead cells was also determined (Figure 39 B).

FACS results showed that at 24, 30, and 48 hpt, dead cells (P1 population) gradually accumulated in EGFP-Pro samples, representing approximately 6%, 12% and 15%, respectively, of the total cell count (Figure 40 A). At 5 h and 10 h, no significant difference in numbers of dead cells was observed when comparing samples. This indicated that in samples transfected with EGFP-Pro, but not in samples with EGFP-Pro mut or EGFP-only, cells began to die between 10 and 24 h following transfection. The majority of dead cells were EGFP-positive (Figure 40 B), further corroborating the association between HML-2 Pro and cell death.

During the course of the experiment many cells expressing active HML-2 Pro remained alive, as evidenced by the relative numbers of green live cells (Figure 40 C). The percentage of EGFP-positive (thus Pro-positive) living cells reached a plateau around 24 hpt. Those cells represented approximately 50% of the main cell population. Probably as a consequence of cell death affecting a portion of the cell population, the percentage of HML-2 Pro-positive living cells started to decline around 48 hpt, contrary to cells expressing EGFP-Pro-mut, for which number of EGFP-positive cells was still increasing until the last time point (Figure 40 C).
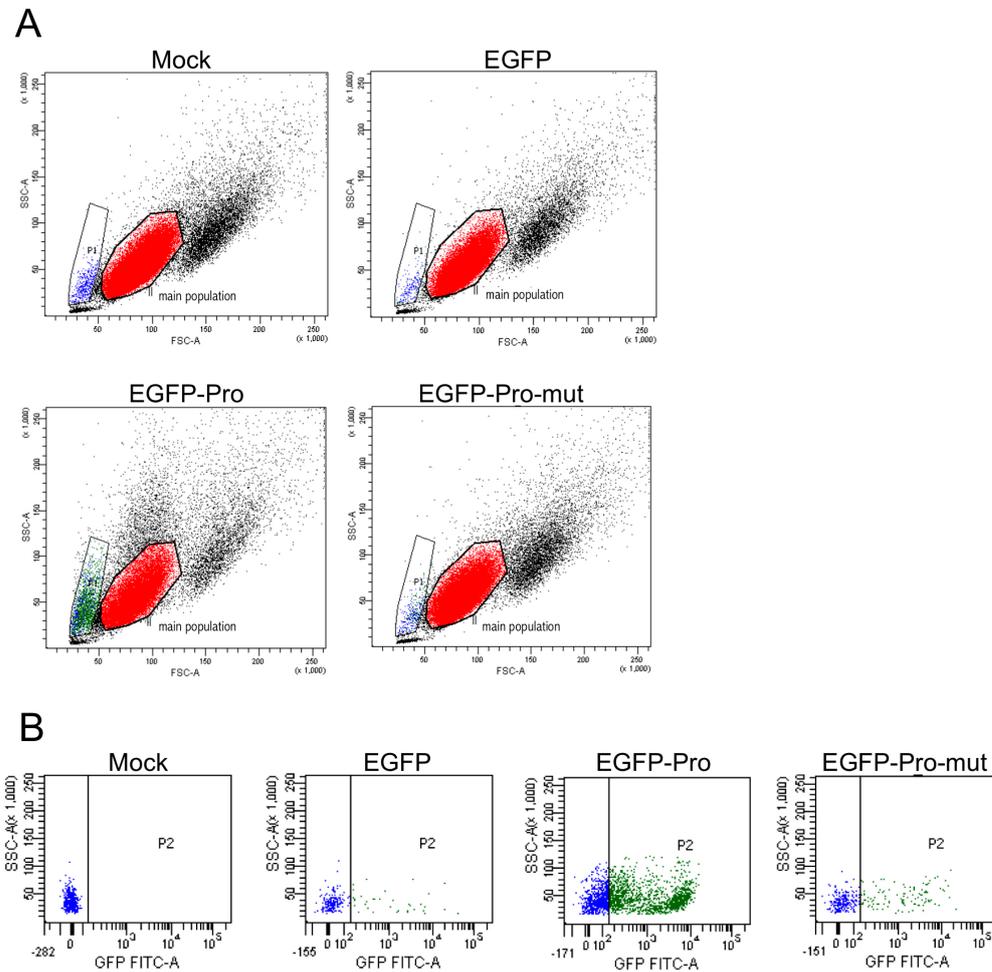
Figure 39: Representative FACS results of cells expressing EGFP-Pro. HEK293T cells were transfected with plasmids expressing either EGFP, EGFP-Pro, or EGFP-Pro-mut. A control sample treated with transfection reagent in absence of plasmid DNA (Mock) was included. Cells were fixed 30 hpt and subjected to FACS analysis. **A**) Forward scatter area (FSC-A) vs side scatter area (SSC-A) plots gating mock-transfected cells as a reference for live HEK293T cells. The population of smaller sized dead cells (P1) was gated to examine the effect of HML-2 Pro on cell viability. **B**) EGFP fluorescence vs side scatter plots gating EGFP-positive cells (P2) for the P1 population. The fluorescence of cells in a mock sample was used as reference for gating EGFP-negative cells, indicated by blue dots. EGFP-positive cells, indicated by green dots, present variable levels of EGFP. The number of EGFP-positive cells detected in the P1 population from EGFP-Pro sample is much higher than the number of EGFP-positive cells detected in the P1 population from EGFP and EGFP-Pro-mut samples, thus indicating an association between presence of active HML-2 Pro and cell death.
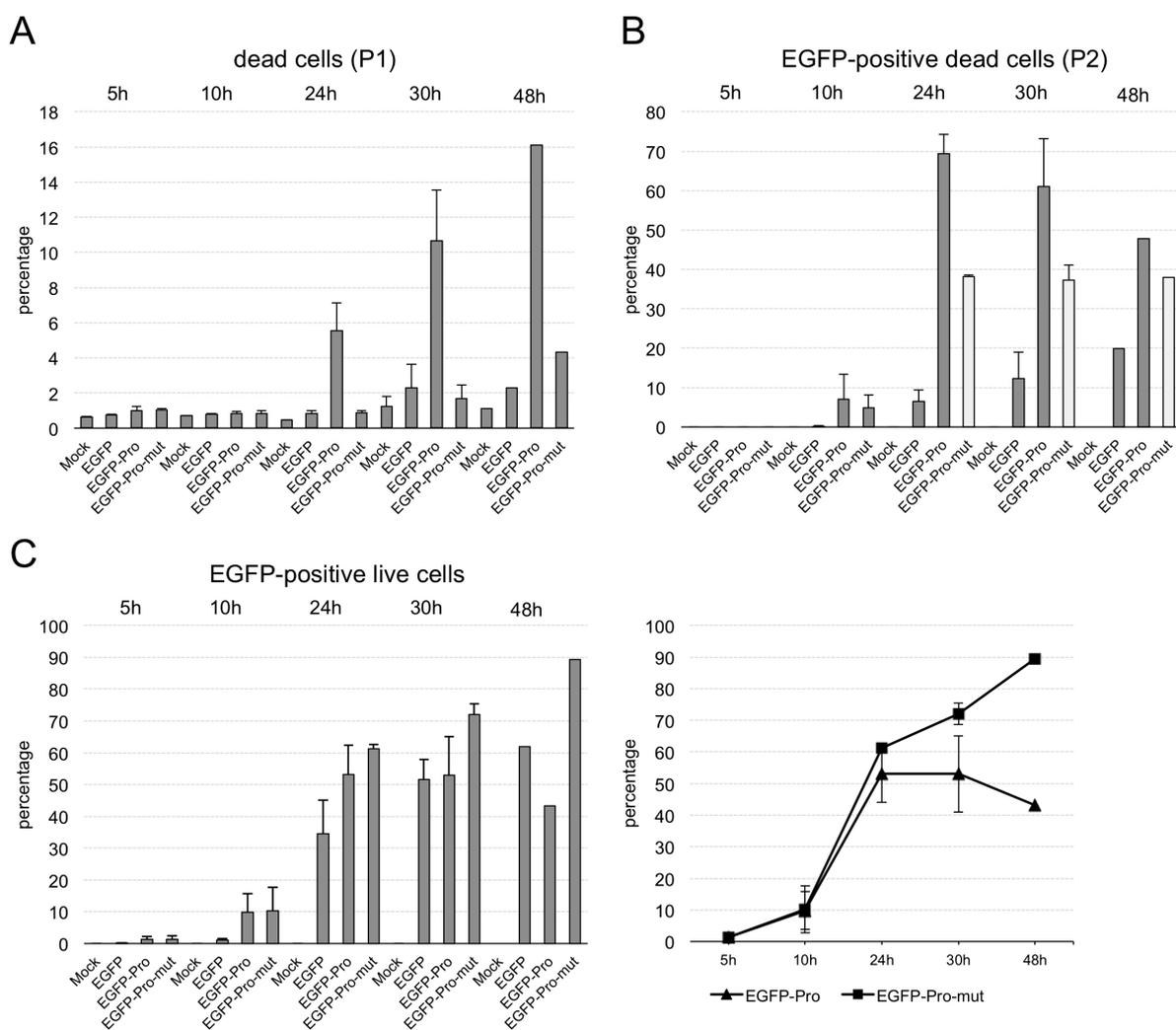
Figure 40: Monitoring cell viability during expression of HML-2 Pro. HEK293T cells were transiently transfected with plasmids expressing EGFP, EGFP-Pro-wt, or EGFP-Pro-mut. A mock transfection with transfection reagent only was also included. Cells were fixed after the indicated time periods and analyzed by FACS. Means and standard deviations from two experiments are given. Between 9000 and 150,000 cells were gated. **A)** Quantification of dead cells (P1 population, see also Figure 37 A) expressed as percentage of the total number of cells analyzed in each sample. **B)** Percentage of EGFP-positive, thus HML-2 Pro expressing cells, in the P1 population. Note that percentages can be misleading, e.g., white bars in EGFP-Pro-mut, when not taking into account relatively small cell numbers in respective P1 populations. **C)** Quantification of EGFP-positive live cells expressed as percentage of the total number of cells analyzed in each sample. On the right, an alternative depiction of cell numbers of EGFP-positive cells for cells expressing EGFP-Pro or EGFP-Pro-mut is shown for the sake of a better comparison of cell viabilities.

4.1.8 Processing of candidate proteins tested *in vivo* was not due to caspase activity

As demonstrated above, expression of HML-2 Pro can trigger cell death going along with activation of apoptotic processes. Therefore, we wondered whether the cleavage of candidate proteins by HML-2 Pro *in vivo* was due to activation of caspases following apoptosis induction, that would then process respective candidate proteins. To address this issue, processing of candidate proteins was monitored in cells driven into apoptosis and compared with processing in presence of HML-2 Pro. Since detection of specific cleavage products would be helpful for discerning HML-2 Pro activity from apoptotic

activities, we tested candidate proteins that had generated stable cleavage products during verification experiments *in vivo*, specifically HSPA90AA1, MAP2K2, and C15orf57 (see chapter 4.1.6.2).

The particular candidate proteins were expressed in HEK293T cells under six different conditions: 1) co-expression with EGFP; 2) co-expression with EGFP and addition of 2 μM staurosporine at 20 hpt; 3) co-expression with EGFP, addition of pan-caspase inhibitor Q-VD at 18 hpt and addition of 2 μM staurosporine at 20 hpt; 4) co-expression with EGFP-Pro; 5) co-expression with EGFP-Pro and addition of pan-caspase inhibitor Q-VD at 18 hpt; 6) co-expression with EGFP-Pro-mut. Staurosporine induced strong activation of apoptosis, thus activation of caspases. Q-VD was included at a concentration of 25 μM to block caspase activity. Cells were harvested 25 hpt and lysed. Cell lysates (15 μg total protein per lane) were subjected to reducing SDS-PAGE followed by Western blot and detection of HA-tagged candidate proteins.

When examining Western blot results, no significant reduction of full-length protein as well as no cleavage products could be observed for the three candidate proteins in presence of 2 μM staurosporine (with or without caspase inhibitor) (Figure 41). On the contrary, a reduction of full-length candidate proteins as well as appearance of cleavage products, could be observed in presence of active HML-2 Pro. Cleavage products in presence of HML-2 Pro were of sizes as expected from previous experiments *in vitro* and *in vivo* (see chapters 4.1.6.1 and 4.1.6.2). Also, the amount of cleavage products was not reduced in the presence of Q-VD.

Taken together, our observations confirmed that processing of candidate proteins was not due to caspase activity since no significant reduction of full-length proteins or detection of stable cleavage products were documented after activation of caspases. Also, processing products observed when co-expressing HML-2 Pro likely were not due to caspase but due to HML-2 Pro activity, as indicated by the specific size of cleavage products and by the fact that broadly inhibiting activity of caspases with Q-VD did not reduce amount of those processing products.
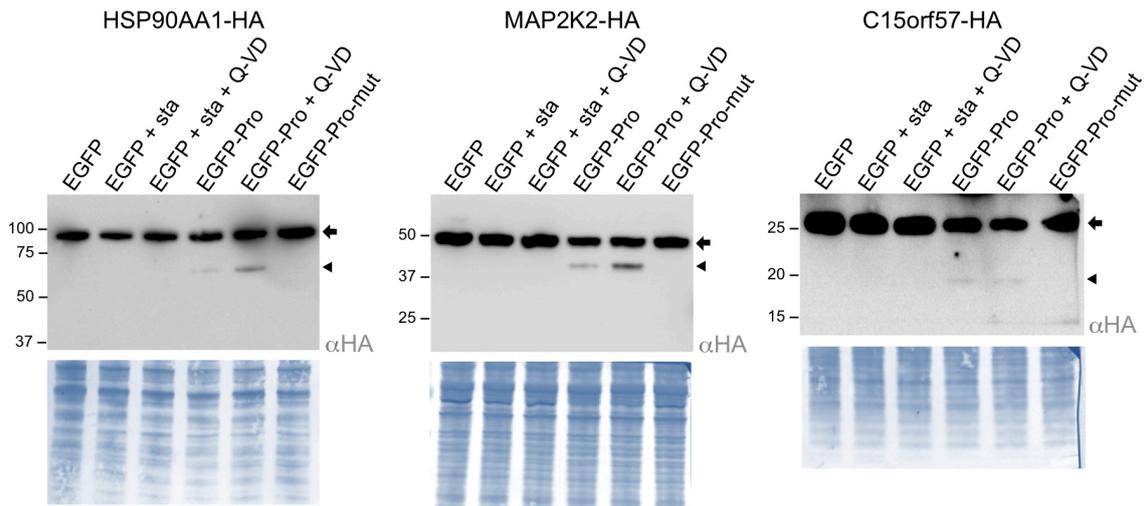
Figure 41: Rebuttal of processed protein products due to caspase activity. HEK293T cells were seeded in 12-well plates and transfected with plasmids encoding candidate protein HSP90AA1-HA, MAP2K2-HA, or C15orf57-HA. Cells were co-transfected with plasmids encoding EGFP, EGFP-fused Pro (EGFP-Pro) or EGFP-fused Pro inactive mutant (EGFP-Pro-mut). 18 hpt, selected wells were supplemented with pan-caspase inhibitor Q-VD (final concentration 25 µM) and/or staurosporine (final concentration 2 µM), the latter added 2 h after Q-VD addition. Cells were harvested and lysed 25 hpt, thus 5 h after addition of staurosporine. Samples were subjected to SDS-PAGE and Western blot detecting the HA-tag. Note that cleavage products of candidate proteins were observed only in samples with active HML-2 Pro (see text). Blot membranes subsequently stained with Coomassie are shown at the bottom.

4.1.9 Preliminary data providing evidence of endogenous HML-2 Pro activity in tumor-derived cell lines

In the experiments performed in cell culture, expression of HML-2 Pro in cells was driven from specific plasmid constructs. We wanted to gain initial evidence of endogenous proteolytic activity from endogenous HML-2 Pro. To do so, we investigated presence of endogenous HML-2 Pro activity in cell lines derived from germ cell tumors (GCTs) and melanoma, two tumor tissues for which expression of HERV-K(HML-2) was documented previously [28,171,179]. We employed the following cell lines: Tera-1 and NCCIT, two cell lines derived from human teratocarcinoma [59,203]; SK-MEL-28 and MeWo, two cell lines derived from human melanoma [28,144].

We first examined presence of endogenous HML-2 Pro in those tumor cell lines. HEK293T cells transiently expressing wild-type HML-2 Pro (HEK pro-wt) were used as a positive control for presence of self-processed HML-2 Pro. Untransfected HEK293T and HeLa cells were included as negative controls. The choice of those negative controls considered results from previous investigations (not shown) that did not detect HML-2 Pro or Pro precursor in those cells. For our investigation, approximately 20 µg of total protein lysates of each cell line were subjected to SDS-PAGE and Western blot followed by detection of HML-2 Pro using a polyclonal α-HML-2 Pro antibody 9367 [185].

Results of Western blots were as follows. An ~18 kDa protein band detected in HEK pro-wt represented self-processed HML-2 Pro, as already observed in previous experiments (for instance, see

Figure 35). Proteins of sizes of ~14 and ~12 kDa were also detected in HEK pro-wt. The origin of those protein bands could not be explained with certainty although their relatively strong intensity in HEK pro-wt and their absence in untransfected HEK cells suggested that they might also represent HML-2 Pro portions (Figure 42). The Pro bands visualized in HEK pro-wt served as reference for evaluating expression of endogenous self-processed HML-2 Pro in the other cell lines. None of the abovementioned protein bands was visible in untransfected HEK293T cells. Instead, protein bands of ~18 kDa were detected in Tera-1, MEWO and NCCIT cell lines, thus providing evidence of HML-2 Pro expression in those cell lines. In SK-MEL-28 cells, the ~18 kDa protein band was rather faint, although a protein band of ~14 kDa was more pronounced (Figure 42, arrowhead). Such an ~14 kDa band could represent one of the smaller proteins detected in HEK pro-wt. In this case, the ~14 kDa band could also support expression of HML-2 Pro in SK-MEL-28 cells. In HeLa cells, a faint protein band of ~18 kDa was detected. This might indicate low-level expression of HML-2 Pro also in this cell line.
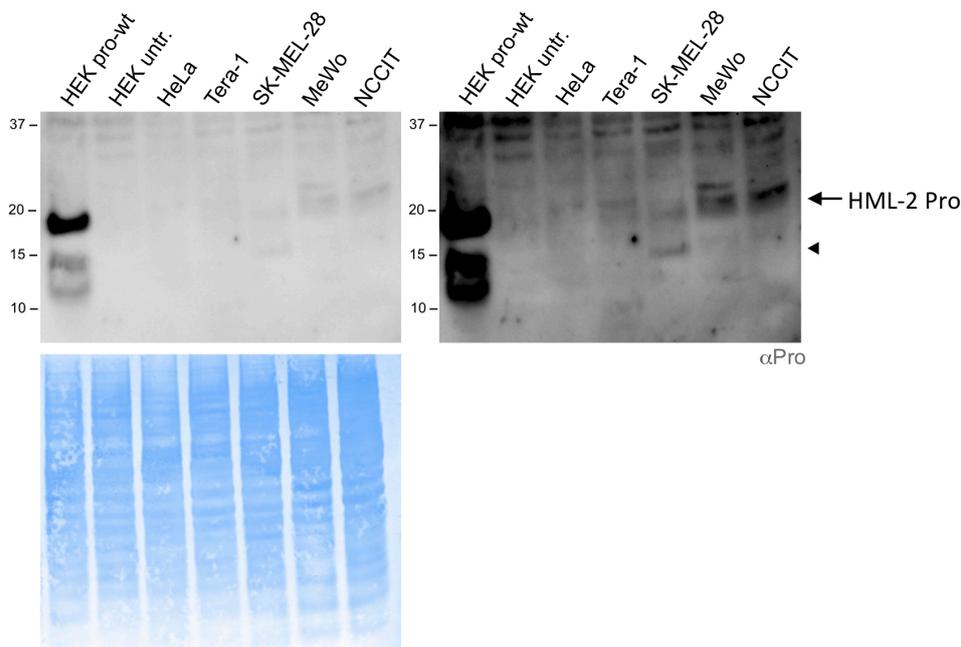
Figure 42: Detection of endogenous HML-2 Pro in cell lines known to overexpress HERV-K(HML-2). Total protein lysates from cell lines Tera-1, SK-MEL-28, MeWo and NCCIT were subjected to SDS-PAGE and Western blotting followed by detection of HML-2 Pro using a polyclonal α-HML-2 Pro antibody 9367 [185]. Total protein lysates from cell lines HEK293T and HeLa served as negative controls. Wild-type HML-2 Pro transiently expressed in HEK293T cells was included as a positive control (HEK pro-wt). The ~18 kDa protein detected in HEK pro-wt was used as main reference (arrow) for presence of self-processed HML-2 Pro in the other cell lines. HML-2 Pro bands of sizes of ~14 and ~12 kDa, thus smaller than expected, were also present in HEK pro-wt. The origin of those protein bands could not be explained with certainty. The ~14 kDa protein (arrowhead) was detected also in SK-MEL-28. On the right, the same blot was subjected to a longer exposure time. The blot membrane was stained with Coomassie after development to document equal amounts of protein loaded.

We next aimed at establishing a relatively straightforward assay for specific detection of proteolytic activity from endogenous HML-2 Pro. Our strategy involved transfection of cell lines Tera-1, SK-MEL-28, and MeWo with a plasmid expressing EGFP-Pro-mut, a fusion protein consisting of EGFP

and mutant Pro. EGFP-Pro-mut harbors sites for HML-2 Pro self-processing, but it is unable to self-process (see chapter 4.1.6.2). In contrast, EGFP-Pro-mut can be cleaved *in trans* by an active HML-2 Pro. Thus, EGFP-Pro-mut represented in this context a specific HML-2 Pro substrate, the cleavage of which was expected only in presence of proteolytic activity of endogenous HML-2 Pro. We also expressed EGFP-Pro-mut in presence of retroviral protease inhibitor indinavir in order to see if inhibition of endogenous HML-2 Pro activity effectively reduced EGFP-Pro-mut processing.

In a control experiment, wild-type HML-2 Pro (sole, not fused to EGFP) was co-expressed in HEK293T cells together with EGFP-Pro-mut to show processing of EGFP-Pro-mut *in trans*. Co-expression in HEK293T cells was also performed in presence of indinavir at concentrations of 1 µM, 10 µM, and 50 µM. In cell lines Tera-1, SK-MEL-28 and MeWo, EGFP-Pro-mut was expressed with or without 50 µM indinavir. Cells were harvested and lysed 24 hpt. Approximately 15 µg of total protein lysates were subjected to SDS-PAGE and Western blot followed by detection of EGFP. Transfection efficiency in NCCIT cells was too low for proper evaluation of results, thus results for NCCIT are not shown.

The control experiment in HEK293T cells showed that EGFP-Pro-mut was processed *in trans* when wild-type, active HML-2 Pro was co-expressed. Processing became evident by appearance of "free" EGFP of ~30 kDa (Figure 43 A). The amount of "free" EGFP appeared reduced in presence of increasing concentrations of indinavir, although indinavir did not completely block HML-2 Pro activity at the highest concentration tested (Figure 43 A). Importantly, transiently expressed EGFP-Pro-mut was likewise processed in Tera-1, SK-MEL-28 and MeWo, likewise releasing the ~30 kDa EGFP portion (Figure 43 B). Moreover, for SK-MEL-28 cells the amount of such processing was visibly reduced in presence of indinavir. For Tera-1 and MeWo cells, reduction of processing in presence of indinavir was not as pronounced as in SK-MEL-28 cells. In view of these results, it is reasonable to conclude that endogenous HML-2 Pro activity is present in cell lines Tera-1, SK-MEL-28, and MeWo.
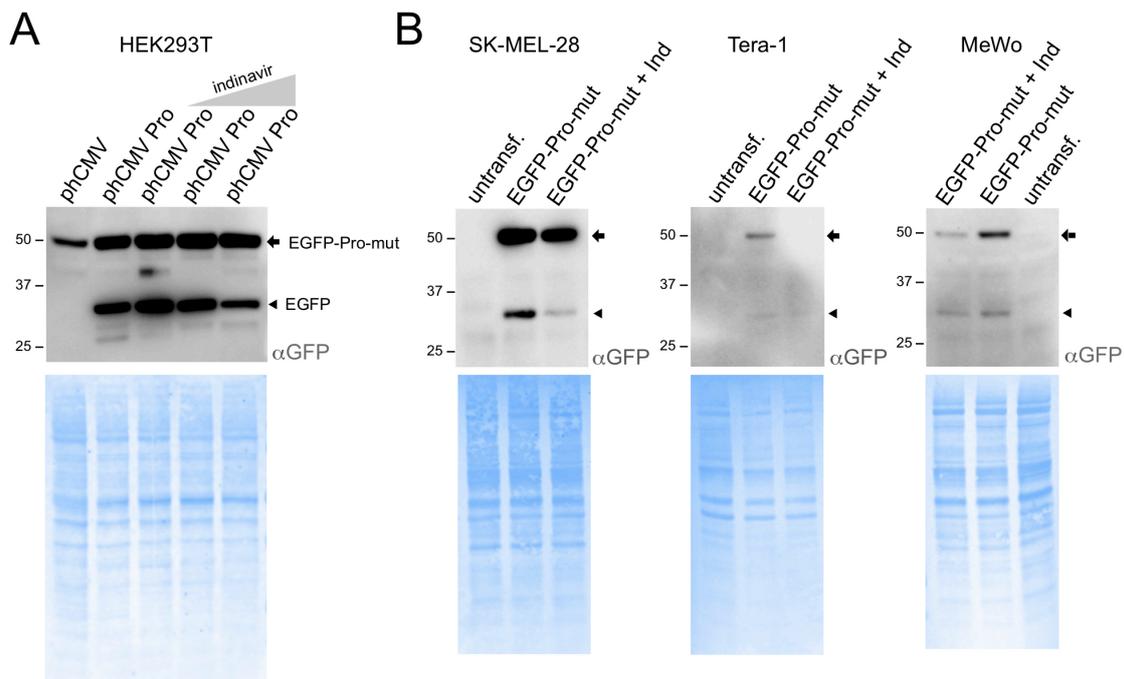
Figure 43: Evidence for presence of enzymatically active HML-2 Pro in cell lines known to overexpress HERV-K(HML-2). A) Control experiment showing the processing of EGFP-Pro-mut by HML-2 Pro and the inhibition of such activity by increasing concentrations of indinavir. HEK293T cells were co-transfected with pEGFP-Pro-mut, encoding an EGFP-fused mutant Pro (EGFP-Pro-mut) and either empty phCMV or phCMV-Pro, the latter encoding wild-type HML-2 Pro. EGFP-Pro-mut is unable to self-process, thus its cleavage, with release of the expected 30 kDa "free" EGFP, was used as a reporter of HML-2 Pro proteolytic activity. Co-transfection of EGFP-Pro-mut with phCMV-Pro was performed also in presence of three different concentrations (1 µM, 10 µM, 50 µM) of retroviral protease inhibitor indinavir. Processing of EGFP-Pro-mut was monitored by Western blot. **B)** Cell lines known to overexpress HERV-K(HML-2) were transiently transfected with pEGFP-Pro-mut, with or without indinavir at 50 µM. Untransfected cells were included as a control. For experiments shown in both A and B, cells were harvested 24 hpt and lysed. Total protein lysates were subjected to SDS-PAGE and Western blot followed by detection of EGFP by an α-GFP antibody. Processing of full-length EGFP-Pro-mut (arrow) generated the expected smaller-sized, "free" EGFP (arrowhead), indicating presence of proteolytic activity from endogenous HML-2 Pro. Blot membranes were stained with Coomassie after ECL to document equal amounts of protein loaded.

4.2 HML-2 integrase as potential inducer of DNA damage

The integrase (IN) of HERV-K(HML-2) is the second enzymatic protein that we considered for potential biological implications. Retroviral INs are multifunctional enzymes that catalyze the retroviral integration, a molecular process that disrupts the integrity of the host genome [120]. A non-sequence specific endonuclease activity (nonspecific alcoholysis) has been demonstrated for HIV-1 IN and other retroviral INs, capable of generating DNA strand-breaks independent from the process of retroviral integration [151,193]. For sites of DNA damage induced by IN activity, DNA double-strand breaks (DSBs) could be caused [176]. DSBs contribute to genome instability which is associated with, for instance, tumorigenesis [210]. If not properly repaired by the cellular machinery, DSBs can also induce activation of programmed cell death [25]. Thus, IN activity could potentially cause detrimental effects in the cell. In this context, HML-2 IN has particular relevance since HML-2 expression is occasionally upregulated in human cells. Several HML-2 loci that potentially encode an intact IN exist in the human genome [25] and some of these HML-2 loci are already known to be expressed in various

tumor tissues [58,183]. Previous experiments *in vitro* confirmed that HML-2 IN is functional [105], thus supporting the notion that, if expressed, this enzyme could affect genome integrity through its enzymatic activities.

We aimed at a preliminary evaluation of the impact of HML-2 IN on the genome of cultured human cells. The HML-2 IN employed in this investigation is derived from the HERV-K(HML-2.HOM) provirus [134], from which HML-2 Pro (see above) was derived as well. Experiments involving HML-2 IN employed a specific method for quantification of DSBs by fluorescence microscopy [198]. DSB formation was indirectly monitored through immunostaining of the 53 Binding Protein 1 (53BP1), one of the key proteins recruited as scaffold protein by the cellular DSB repair system and widely used as a marker of DNA damage [150]. 53BP1 molecules can be visualized by fluorescence microscopy as discrete foci formed in the cell nucleus in co-localization with DSBs where the cellular DNA repair machinery is assembled temporarily. As reported in the literature, 53BP1 foci formation and the associated kinetics aspects were described mainly in response to ionizing radiation (IR) [132,186]. IR is well-known to induce formation of DSBs in genomic DNA. Therefore, control assays establishing experimental conditions monitored 53BP1 foci formation in HeLa cells after exposure to 1 Gy of IR (Figure 44). Following irradiation, cells were incubated for 30 min at 37° C in order to allow for recovery, and then fixed with formaldehyde. Based on previous reports involving different cell lines (including HeLa) the average number of 53BP1 foci per cell reaches a peak after 30 min [132,186]. Cells were immunostained using a primary α-53BP1 polyclonal antibody produced in rabbit and a secondary antibody α-rabbit IgG coupled with Alexa Fluor 488. 53BP1 foci were analyzed under a fluorescence microscope after mounting cells in presence of DAPI that stained the DNA in cell nuclei. Numbers of 53BP1 foci within cells were evaluated by eye, counting foci directly under the microscope or in images captured with a camera.

Fluorescence microscopy revealed that irradiated cells presented 53BP1 protein different from that in unirradiated control cells. In irradiated cells, there were approximately 25 53BP1 foci per cell, whereas the majority of unirradiated cells showed between 3 foci and none at all, with some unirradiated cells sporadically presenting higher numbers of foci. In both irradiated and unirradiated samples, 53BP1 foci were of variable size and morphology, ranging from very small dots to larger rounded granules. Control assays thus suggested that experimental conditions established for detection of DSBs were adequate for the documentation of such DNA damage induced in nuclear DNA by a genotoxic agent other than IR.
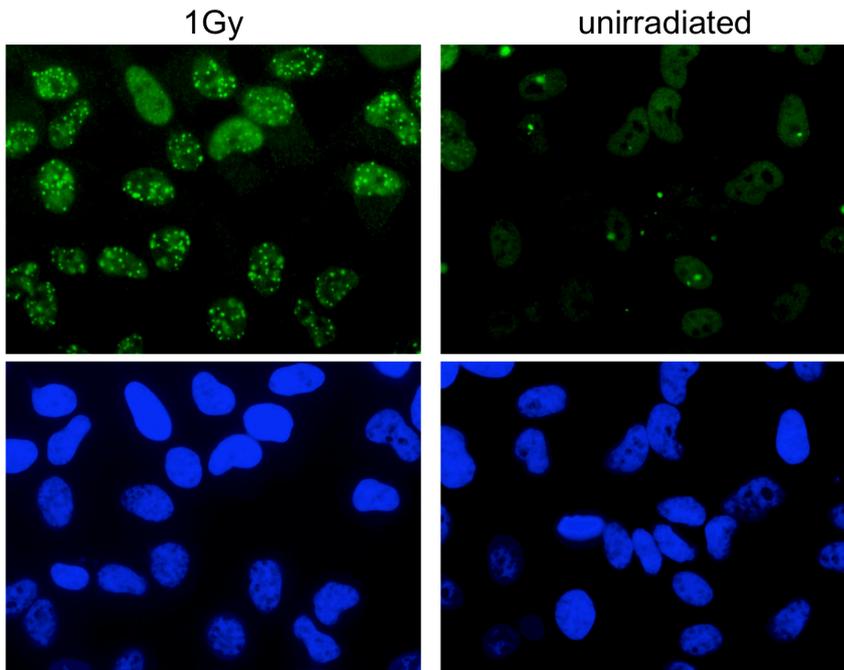
Figure 44: **Control assay for monitoring DSBs via detection of 53BP1 foci.** 53BP1 foci were visualized by immunostaining and fluorescence microscopy in HeLa cells exposed to 1 Gy of ionizing radiation. After irradiation, cells were incubated for 30 min at 37°C, fixed for 15 min with 4% [v/v] formaldehyde. Cells were permeabilized by a 5-min incubation with 0.2% [v/v] TritonX-100 on ice. Cells were blocked for 30 min with 1% [w/v] BSA and incubated overnight at 4°C with an α-53BP1 polyclonal antibody produced in rabbit. Following wash steps, cells were incubated for 1 h with a secondary α-rabbit IgG antibody conjugated with Alexa Fluor 488, washed again and mounted in presence of DAPI. Cell nuclei (in blue) and 53BP1 foci (in green) were visualized by fluorescence microscopy.

We then performed experiments utilizing HML-2 IN as an agent potentially causing DSBs. We aimed in those experiments at evaluating if presence of HML-2 IN in human cells could affect the number of 53BP1 foci, thus the number of DSBs. To do so, HeLa cells were transfected with plasmids encoding either pSG5 IN or pSG5 IN-mut to induce expression of wild-type HML-2 IN and an enzymatically inactive mutant (mutated in the catalytic domain), respectively, with both IN proteins labeled with an HA-tag at the N-terminus. Experiments also included untransfected cells as controls. Cells were fixed 24 hpt and stained with antibodies for subsequent immunofluorescence analysis. We established experimental conditions for a multiple immunostaining procedure that permitted examination of different proteins simultaneously. We detected the HA-tag in order to visualize cells specifically expressing IN, thus directly focusing our attention on cells in the cell culture that were transfected and harbored HML-2 IN. Detection of 53BP1, thus visualization of 53BP1 foci present in cell nuclei, allowed us to evaluate numbers of DSBs. 53BP1 signaling may be antagonized by cellular factors at different stages of the cell cycle [56]. Thus, we also took into account that 53BP1 foci might have presented variable levels of activation depending on the cell cycle phase. To minimize potential biases due to variable cell cycle phases, we considered cells that had been fixed in the same phase of the cell cycle. We therefore stained the protein Ki-67, a nuclear protein that is variably expressed during different phases of the cell cycle [195] (Figure 45 B). Ki-67 is typically used as proliferation marker since it is undetectable specifically in non-dividing cells [141]. More specifically, Ki-67 is degraded

continuously in G0 and G1 and accumulates during S to M phases [141]. It appears that in actively proliferating cells the Ki-67 level is at a minimum at the end of the G1 phase [141]. We therefore considered for the analysis of 53BP1 foci only cells with Ki-67 levels below the detection limit (Ki-67 negative cells), as those cells likely represented cells in G1 or G0 phase (leftmost in Figure 45 B). We thus reduced the number of cells considered in our analysis to potentially improve accuracy of results. As for the immunostaining procedure, primary antibodies raised in different species and directed against different target proteins of interest were added to fixed cells in parallel. The mixture of primary antibodies included: 1) a monoclonal α-HA antibody produced in rat; 2) an α-53BP1 antibody produced in rabbit and conjugated with fluorescent dye Alexa Fluor 488 (green), the latter allowing direct detection without the need for a secondary antibody; 3) a monoclonal α-Ki67 antibody produced in mouse. Secondary antibodies were: α-rat IgG and α-mouse IgG antibodies, both produced in goat and labelled with fluorescent dyes Alexa Fluor 594 (red) and Alexa Fluor 647 (infrared), respectively. Secondary antibodies were likewise incubated in parallel. Cell nuclei were stained with DAPI during mounting. Multi-color detection of the various probed proteins was performed using a fluorescence microscope equipped with a filter set for visualization of DAPI (blue), FITC (green), Texas-Red (red) and infrared. 53BP1 foci were analyzed and counted by eye in acquired images. Examination of 53BP1 foci was performed for a total of 210 Ki-67-negative cells. Those 210 cells analyzed included 70 cells expressing HA-tagged wild-type IN (IN-wt), 70 cells expressing HA-tagged IN mutant (IN-mut) and 70 untransfected cells.

When analyzing acquired images, we observed that IN localized in the nucleus and the cytoplasm of transfected cells, although in most of the cells IN appeared to be more enriched in the cytoplasm than in the nucleus. IN appeared to be distributed homogeneously both in the cell nucleus and cytoplasm of transfected cells. Although, IN also appeared to be enriched at discrete loci in the cytoplasm of some cells. Those higher-density spots may represent intracellular precipitates of IN as observed in experiments involving transiently expressed HIV-1 IN [155]. We did not notice differences in cellular localization and distribution when comparing IN-wt and IN-mut. Since IN localized in the cell nucleus, there was evidence to believe that IN could have interacted with the host DNA. However, analysis of 53BP1 foci did not reveal significant differences in numbers of 53BP1 foci when comparing IN-wt, IN-mut and untransfected cells (Figure 45 C, D). In each group of cells, there were approximately 3 foci per cell. Also, no differences in sizes of foci were noticed when comparing cells in the different groups. In all groups, 53BP1 foci were of variable dimensions, with a majority of small-sized foci and occasionally foci of larger size (Figure 45 A). Thus, a particular effect regarding the number, size, and morphology of 53BP1 foci in presence of HML-2 IN activity could not be documented in our initial experiments. Therefore, the employed experimental conditions did not produce evidence to conclude that HML-2 IN can damage DNA in a way that causes formation of DSBs.
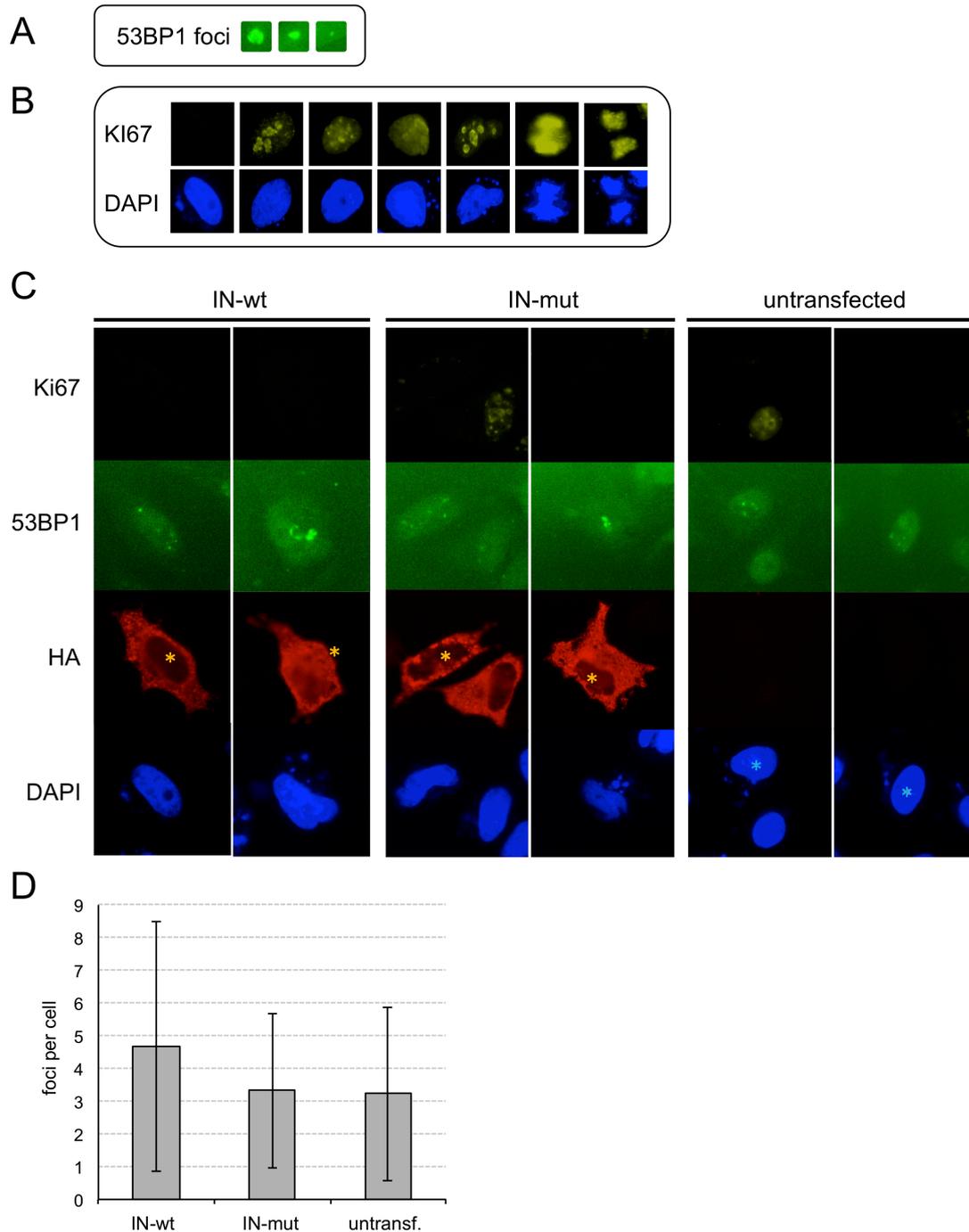
Figure 45: Evaluating the ability of HML-2 IN to induce formation of DSBs. A) Representative sizes and morphologies of 53BP1 foci detected by immunofluorescence. Foci with variable dimensions, ranging from small-sized foci to large-sized foci were counted. **B)** Representative examples of cell nuclei stained with Ki-67 (yellow). Ki-67 was distributed in the cell nucleus (blue) in variable patterns corresponding to different phases of the cell cycle. Ki-67-negative cells (leftmost image), possibly cells in G1 or G0 phase, were considered for analysis of 53BP1 foci. **C)** Representative cells (indicated with a star) fulfilling selection criteria for inclusion in the analysis of 53BP1 foci. Cells expressing HA-tagged IN or IN-mut (red) were detected with a monoclonal α-HA antibody produced in rat. Ki-67 (yellow) was detected with a monoclonal α-Ki-67 antibody produced in mouse. For the detection of 53BP1 (green), a polyclonal α-53BP1 antibody produced in rabbit and conjugated with Alexa fluor 488 was employed. Cell nuclei (blue) were stained with DAPI. **D)** Averages of 53BP1 foci number per cell based on analysis of 70 HA-IN-expressing cells, 70 HA-IN mutant expressing cells and 70 untransfected cells, with all of them fixed in the same cell cycle phase. Standard deviations are indicated.

5. Discussion

In the context of a retroviral infection, inside the cell, viral proteins interact with the cellular machinery to coordinate a process that results in viral spreading through production of viral particles. In the course of this process, the cellular integrity is not necessarily preserved. On the contrary, viral proteins perform multiple functions with several consequences in the cellular environment. Side effects, not necessarily relevant to the viral strategy, are also induced. Altogether, activities of viral proteins alter cellular functionalities resulting in detrimental, if not lethal, effects.

In view of this, HERV-K(HML-2) (hereafter HML-2), a group of HERVs that include proviruses retaining particularly intact genomes, represents a source of retroviral proteins that, whenever expressed, could exhibit enzymatic activities with a negative impact on cell biology. All human cells have the potential to produce such retroviral proteins. Currently, HML-2 transcription is known to be upregulated in various disease tissues and expression of HML-2 proteins was observed in tumor cells. In those pathological contexts, HML-2 proteins could represent initiators or promoters of cellular impairment and cell transformation, although the underlying mechanisms are still unexplored.

This work focused on HML-2 protease (Pro) and HML-2 integrase (IN), two HML-2 encoded proteins little investigated so far, whose enzymatic activities could potentially interfere with cellular processes. In the following, findings will be discussed in two sections: the first (5.1) related to HML-2 Pro and the second (5.2) to HML-2 IN.

5.1 HML-2 protease

The proteolytic activity of HML-2 Pro was demonstrated in previous studies, in which self-processing of HML-2 Pro and processing of HML-2 Gag proteins were directly observed through expression in *E. coli* [134] and through *in vitro* cleavage assays [185]. The presence of a functional HML-2 Pro has been inferred in germ cell tumors (GCTs), where properly processed HML-2 Gag proteins were detected [112,179]. Previous studies have observed the budding of HML-2 viral particles with mature morphology in GCT cells. Mature morphology is attributed to the presence of active HML-2 Pro [20]. Despite the evidence of a functional HML-2 Pro in some human cells, the proteolytic activity of this enzyme was only demonstrated against retroviral substrates. Whereas, we considered crucial to investigate the proteolytic activity of HML-2 Pro as a potentially harmful factor to which cellular proteins could be exposed. Some studies already demonstrated processing of cellular proteins by retroviral proteases, although those studies involved exogenous retroviruses (for instance, see [194,212]). Thus, in our study we aimed to know whether HML-2 Pro could be able to process cellular proteins and, by doing so, induce detrimental effects in the cell. Confirming such activity for an enzyme that is inducible in all human cells has potentially important implications. It was our specific intent to

identify potential cellular substrates of HML-2 Pro, assess their roles in the cell and eventually confirm their cleavage by HML-2 Pro under physiological conditions.

5.1.1 Expression of HML-2 Pro in bacteria and protein purification

Most of our investigation on HML-2 Pro was based on an *in vitro* setup, which allowed the study of HML-2 Pro activity at well-defined conditions. Prerequisite for our experimental approach was the generation of purified, functional HML-2 Pro, that could be included at proper concentrations to reactions. HML-2 Pro was successfully purified following a protocol for isolation of HML-2 Pro from inclusion bodies as previously published [112]. That protocol is based on a method described by Wondrak et al. (1991) [221] for HIV-1 Pro purification. Of note, Kuhelj et al. (2001) [112] reported purification of two forms of HML-2 Pro: 1) a truncated form, including the complete functional Pro core domain, but lacking 50 aa in the C-terminus that were demonstrated as not necessary for catalytic activity; 2) a full-length his-tagged Pro. Kuhelj et al. (2001) [112] observed identical specificities for the two Pro variants, although the truncated HML-2 Pro displayed a catalytic efficiency twice as high as that of full-length Pro. An important aspect of our study was to test HML-2 Pro activity against cellular substrates whose specificity for processing by HML-2 Pro was unknown. Since processing of candidates could have been hardly detectable if occurring at low level, we considered favorable for our experiments *in vitro* to employ the HML-2 Pro form that could display higher proteolytic activity. Thus, we purified a C-terminally truncated form of HML-2 Pro. Although truncated in the C-terminus, HML-2 Pro was expressed in bacteria as a precursor that included the complete HML-2 Pro core domain and an N-terminal region flanking the N-terminal autocatalytic cleavage site, the latter important for monitoring HML-2 Pro activity through its self-processing.

There were two noteworthy modifications with regard to the protocol of Kuhelj et al. (2001) [112]. First, for lysis of *E. coli* cells expressing HML-2 Pro, before sonication we homogenized bacteria by using a dounce homogenizer for improved lysis of bacteria in an inexpensive and practical way. Of note, this step was quick, did not require addition of reagents and could be performed in the cold. Thus, it can be assumed that that modification did not affect HML-2 Pro integrity. Second, chromatography for HML-2 isolation was performed by gravity flow instead of using a fast protein liquid chromatography (FPLC) system. Our method worked well, thus it represents an alternative for HML-2 Pro purification laboratories that do not have access to an FPLC system.

HML-2 Pro was purified at a sufficient level of purity. Only a contaminant protein of ~45 kDa appeared co-isolated during the elution steps. The HML-2 Pro eluates in which the contaminant protein was barely detectable (eluates 3 and 4), or not at all detectable (eluate 5) (see Figure 17), were pooled into an HML-2 Pro stock for further experiments. We assumed that the ~45 kDa contaminant protein, as

well as other contaminant proteins below detection limits, would have been further diluted during set up of further assays, thus becoming even less relevant.

The yield of purified HML-2 Pro could be considered reasonable. Inducing HML-2 Pro expression in 1 L of bacterial culture (absorbance at 600 nm was 0.6) and incubating the culture for 3 h after induction yielded a total of 540 µg of HML-2 Pro at the end of the purification procedure, that was aliquoted and stocked at -80°C. Approximately 90% of those 540 µg of purified HML-2 Pro were consumed in the course of the various experiments. Some modifications might have further improved purity of HML-2 Pro and consequently the amount of (pure) HML-2 Pro to be used for further experiments. For example, during purification the eluate containing the higher amount of HML-2 Pro could not be used because of higher amounts of the ~45 kDa contaminant protein (see Figure 17, eluate 2). Pro of higher purity may be obtained by additional washing steps to remove the contaminant 45 kDa protein before elution of HML-2 Pro. Considering that modification of ionic strength (1M NaCl in the binding buffer, no NaCl in the elution buffer) and pH (pH 6.5 in the binding buffer, pH 8.0 in the elution buffer) are key factors for elution of retroviral proteins specifically bound to pepstatin A [221], additional washing steps could employ buffers with intermediate concentrations of NaCl and pH, optimized towards elution of the 45 kDa contaminant protein without dramatically destabilizing the binding of HML-2 Pro to pepstatin A. Washing buffers with different concentrations of NaCl between 0 and 1 M and/or pH between 6.5 and 8 should be tested empirically.

Alternatively, considering that the molecular weight of the 45 kDa contaminant protein is significantly higher than the molecular weight of self-processed HML-2 Pro (~12 kDa), a filtering step after elution, for instance by size-exclusion chromatography, could also represent a good strategy for removal of the contaminant protein and thus purer HML-2 Pro.

We have shown that the HML-2 Pro from the provirus HERV-K(HML-2.HOM) is enzymatically active, corroborating a previous report about this specific protease [134]. HML-2 Pro exhibited enzymatic activity during expression in bacteria, as already observed in other studies in which bacterial expression of HML-2 Gag-Pro polyprotein led to specific cleavage into individual Gag domains [134,143]. In our experiments, two hours after protein induction, self-processing of HML-2 Pro precursor (~18 kDa) with release of the mature HML-2 Pro (~12 kDa) could be observed readily (see Figure 16). It can be assumed that, at the time when bacteria were pelleted for subsequent lysis, expression of HML-2 Pro was still ongoing, thus a certain amount of the 18 kDa Pro precursor was subjected to purification. Kuhelj et al. (2001) [112] reported that during purification of HML-2 Pro, the complete processing of the precursor took place during dialysis, when Pro renaturated. In our case, through monitoring of Coomassie-stained proteins following SDS-PAGE (see Figure 17) we could not discriminate during which step(s) of the purification procedure HML-2 Pro precursor eventually self-processed. In any case, only the 12 kDa HML-2 Pro was detected in eluates, thus confirming complete processing of the 18 kDa Pro precursor during the purification/renaturation procedure.

The procedure for HML-2 Pro purification and subsequent concentration, aliquoting and storage at -80°C preserved the enzymatic functionality. This was confirmed through a reaction in which purified, frozen and thawed HML-2 Pro processed *in vitro*-translated HML-2 Gag efficiently. Of note, this verification experiment also confirmed that our purified HML-2 Pro could process a substrate *in trans*. Such Pro activity could not be specifically discriminated during expression in bacteria, when monitoring just the cleavage of the 18 kDa HML-2 Pro precursor. Because of the efficient cleavage observed in those experiments, HML-2 Gag was employed as a positive control for processing by HML-2 Pro in several of our subsequent experiments.

5.1.2 No Purification of HML-2 Pro mutants

We could not purify enzymatically inactive HML-2 Pro mutants because of inefficient binding of mutant HML-2 Pro with pepstatin A. In particular, we tried to purify two Pro mutants that could have been employed as negative controls in *in vitro* assays for studying HML-2 Pro activity. One Pro mutant harbored an amino acid substitution D→N within the catalytically important DTG-motif. The D→N mutation is known to be critical for the active site geometry of retroviral proteases [180]. Thus, it was not entirely unexpected that such a mutant would not interact with pepstatin as well as wild-type HML-2 Pro. The second mutant was motivated by a study that reported successful purification, by pepstatin-A affinity chromatography, of an inactive HIV-1 protease with an R→K substitution in the GRNLL region [221], that is highly conserved and relevant for enzyme activity and structure [85]. We introduced an analogous substitution in the GRDLL region of HML-2 Pro (see chapter 4.1.1), but that mutation, in our case, did not permit efficient binding of the mutant HML-2 Pro to pepstatin A. The discrepant outcome between Wondrak et al. (1991) [221] and our study is currently not clear but may be due to variations in the purification procedure or different behavior of the two proteases involved. For instance, in Wondrak et al. (1991) [221] the buffer used for resuspension and binding of Pro with pepstatin-A-agarose has a relatively high concentration of ammonium sulfate, whereas ammonium sulfate was not present in the buffer employed in our purifications. That component could thus be considered relevant for optimized purification of HML-2 Pro mutants.

The substituted R residue within the GRNLL region is highly conserved among retroviral proteases and it is known to contribute to the protein structure and activity of HIV-1 Pro [85]. Since the R→K substitution introduces a relatively minor change in the GRNLL region, maintaining a positive charge on the side chain, one could speculate that, with regard to binding ability, the micro-environment of the cleavage site of HML-2 Pro could be less tolerant towards changes than that of HIV-1 Pro. Specific investigations would be required to produce further evidence. In any case, alternative HML-2 Pro mutants could be tested with regard to their suitability for purification. For example, substitution of the G residue in the GRDLL region could be tested. Mutations in other conserved regions of HML-2 Pro, for example, in the "flap" motif, could also be tested.

Use of a purified mutant of HML-2 Pro may represent an optimal control for *in vitro* assays. However, as we observed, purification of such mutants may require optimizations, with no success guaranteed. Our experimental strategy did not necessarily depend on such purified Pro mutants. Instead, as an alternative strategy for our assays, we considered it sufficient to inhibit HML-2 Pro activity by means of a specific inhibitor. For that purpose, we employed pepstatin A, a well-known inhibitor of aspartyl proteases that we confirmed to also efficiently inhibit purified HML-2 Pro. In particular, we found that in our experimental system, 200 μM pepstatin A could efficiently inhibit HML-2 Pro present at 460 nM, reducing its activity by 98% compared to controls without inhibitor. We also showed that in presence of 200 μM pepstatin A, complete inhibition (100%) of HML-2 Pro activity could be achieved when pre-incubating pepstatin A with HML-2 Pro for 10 min before addition of the protease substrate. We suppose that the pre-incubation step favored a more stable interaction of pepstatin A with HML-2 Pro before setting off interaction with substrate. This could be an important aspect to be considered when setting up negative controls for assays testing HML-2 Pro activity against substrates that are processed with high efficiency.

5.1.3 Characterization of HML-2 Pro and optimization of reaction condition

Enzymatic activity of HML-2 Pro was assessed and partially characterized in previous studies that monitored processing of synthetic, fluorescent substrates by means of fluorescence assays *in vitro* [112,205]. In our study, we used a similar approach to characterize enzymatic activity of purified HML-2 Pro, corroborating previous findings concerned with biochemical aspects of that enzyme. At the same time, we aimed at establishing reaction conditions suited for subsequent TAILS experiments, which required samples prepared in a buffer that, ideally, consisted of relatively few buffer components.

It is known that, besides HML-2 Gag, HML-2 Pro can also process HIV-1 Gag polyprotein [205]. In fact, in previous studies, the fluorogenic peptide substrates Lys-Ala-Arg-Val-Tyr-/-Phe-(NO$_2$)-Glu-Ala-Nle-NH$_2$ [205] and 2-aminobenzoyl-Ala-Thr-His-Gln-Val-Tyr-/-Phe-(NO$_2$)-Val-Arg-Lyr-Ala, derived from HIV-1 Gag and originally designed as substrates for HIV-1 Pro, have been shown to be efficiently cleaved by HML-2 Pro. When we established the strategy for our fluorescence assay, both peptides were not available commercially at reasonable prices. We therefore chose another HIV-1 Gag derived peptide, the hexapeptide FRET substrate 2-aminobenzoyl-Thr-Ile-Nle-/-Phe-(NO$_2$)-Gln-Arg-NH$_2$ (hereafter Anthranilyl-substrate), that was commercially available from Bachem AG. The Anthranilyl-substrate was likely also processed by HML-2 Pro, since it includes in its sequence aa residues favored for P1 and P1'. Specifically, Nle (Norleucine, an isomer of Leu) is present at the putative P1 site. Nle/Leu is hydrophobic and unbranched at the β-carbon as is the majority of aa residues found in P1 of cleavage sites of HML-2 Pro within HML-2 Gag [66]. Furthermore, Phe and

Ile are present in the putative P1' and P2 sites, respectively, which are aa residues favored for those positions, as per recently published specificity profile of HML-2 Pro cleavage sites [17].

Monitoring the processing of the Anthranilyl-substrate by HML-2 Pro at different buffer conditions corroborated that buffers with low pH and high ionic strength are particularly favorable for HML-2 Pro activity, as also shown in previous studies [112,205].

Different buffer systems have been used in previous studies on HML-2 Pro or other retroviral proteases (see [112,156,205]). Since our protocol for HML-2 Pro purification was based on the protocol published by Kuhelj et al. (2001) [112], we considered that study as main reference for buffer conditions for our *in vitro* fluorescence assays. We also tested a buffer used by Porter et al. (2002) [156] for assaying HIV-1 Pro activity against Anthranilyl-substrate.

One buffer retrieved from Kuhelj et al. (2001) [112] and composed of 50 mM MES, 1 M NaCl, 20% [v/v] glycerol, 1 mM EDTA, pH 5.0, turned out to be suitable for cleavage of Anthranilyl-substrate by HML-2 Pro. In the same buffer, but without glycerol, HML-2 Pro activity was ~30% higher. This indicated that lack of glycerol increased HML-2 Pro activity in our experimental system. The buffer reported by Porter et al. (2002) [156] had a simpler composition, specifically 100 mM MES-TRIS, 1.25 M NaCl, pH 6.0, and also appeared as suitable for cleavage of Anthranilyl-substrate by HML-2 Pro. This indicated that, apart from glycerol, other components commonly included in reaction buffers, like EDTA and DTT, were dispensable for assaying HML-2 Pro activity. We also demonstrated a negative effect of DMSO on Pro activity. DMSO is a solvent agent that is often used to resuspend inhibitors or other chemicals, reducing or avoiding such a component in assays of HML-2 Pro will thus increase proteolytic activity. Moreover, we evaluated in a specific experiment that use of PIPES instead of MES could favor HML-2 Pro activity (see also later). Altogether, our experiments provided relevant information for setting up *in vitro* assays for HML-2 Pro *in vitro*. A buffer composed of 100 mM PIPES and 1 M NaCl was eventually favored for assaying enzymatic activity of purified HML-2 Pro towards Anthranilyl-substrate and for subsequent TAILS experiments.

Another buffer reported by Kuhelj et al. (2001) [112], composed of 20 mM PIPES, 100 mM NaCl, 1 mM DTT, 10% [v/v] glycerol, pH 6.5, appeared as unsuited for cleavage of Anthranilyl-substrate by HML-2 Pro. HML-2 Pro was completely inactive in that buffer. Of note, the Kuhelj et al. (2001) study [112] used that buffer specifically for cleavage of *in vitro* translated HML-2 Gag. In line with the Kuhelj et al. (2001) study [112], the same buffer was suitable for cleavage of *in vitro* translated HML-2 Gag also in our study (see Figure 21). One can hypothesize that because of a combination of buffer features negatively affecting the enzyme activity (presence of glycerol, low ionic strength, high pH), HML-2 Pro could not efficiently process the Anthranilyl-substrate in that buffer. However, a complete absence of proteolytic activity was unexpected. Unsuitability of that buffer for our fluorescence assays may be due to the specific substrate involved. Additional investigations would be necessary to identify

the cause of that inactivity. Since we had other suited buffer compositions for HML-2 Pro activity at hand such investigations were not required though. We therefore did not further consider that unsuited buffer in our subsequent experiments.

5.1.4 Activity of HML-2 Pro at different pHs

We defined more precisely the pH required for optimal HML-2 Pro activity by monitoring the processing of the Anthranilyl-substrate in a set of reactions with acidic pH ranging from 4.5 to 6.5. A previous study [205] that assayed HML-2 Pro activity at variable pHs reported that the pH profile of HML-2 Pro activity was bell-shaped with a peak activity at approximately pH 4.5. In our experiments, the pH profile of HML-2 Pro activity was likewise bell-shaped, but with a peak at pH ~5.5. Our result is partially in disaccord with the previous report, although the slight difference in the observed HML-2 Pro pH optimum could be due to differences in buffer composition and/or fluorogenic substrate employed. The factual reason for the different pH optimum would have to be investigated in specific experiments; it appeared less relevant for our subsequent studies.

In a second set of experiments, HML-2 Pro activity was assayed along an extended pH range, specifically from pH 5.5 to 8.0, in order to further evaluate HML-2 Pro activity at higher pH. HML-2 Pro activity around neutral pH was of particular relevance since that pH range is found in major cellular compartments, namely nucleus and cytoplasm [31]. Knowledge of HML-2 Pro activity at different pHs becomes relevant concerning susceptibility of proteins from different cellular compartments for cleavage by HML-2 Pro. Remarkable stability of HML-2 Pro activity over a wider pH range was pointed out by Towler et al. (1998) [205]. For instance, it was shown that more than 60% of HML-2 Pro activity could be recovered after prolonged incubation at room temperature over a pH range from pH 3 to 9, both in presence and absence of 1 M NaCl [205]. However, in that particular experiment, residual HML-2 Pro activity was measured at a fixed pH 6.2. Concerning tolerance of HML-2 Pro activity towards pH variations, in that experiment, no HML-2 Pro activity was observed at pH > 6.5, both in presence or absence of 1 M NaCl. In our experiments (in presence of 1 M NaCl), we found that HML-2 Pro activity was not restricted to acidic pH. In fact, HML-2 Pro was still active at neutral pH, with approximately 10% of the activity observed at the optimal pH, and furthermore at basic pH (up to pH 8.0), with approximately 1% of the activity observed at the optimal pH.

Towler et al. (1998) [205] reported HML-2 Pro activity up to pH 8.5 when tested in a buffer with 1.25 M ammonium sulfate [205]. The authors of that study attributed higher HML-2 Pro activity to the nature of the salt, assuming that ammonium sulfate is more favorable than NaCl for stabilization of hydrophobically driven interactions between the substrate and Pro, thus increasing dimer stability and consequently the enzyme's efficiency even at basic pH. In our study, we found that the nature of the

buffering agent could also influence HML-2 Pro activity. In particular, we observed that at pHs ranging from 5.5 to 8.0 HML-2 Pro was more active in PIPES buffer than MES buffer. Comparing the two buffering agents the difference in Pro activity was gradually more pronounced at pH > 6.0. Pro activity was approximately 55%, 70%, 80%, 90% lower in MES buffer than in PIPES buffer when monitored at pH 6.5, 6.9, 7.5, 8.0, respectively. Thus, in our system PIPES buffer appeared as more suited than MES buffer, especially at pH > 6.0.

Considering that the cellular localization of HML-2 Pro is currently unknown, our own and previous findings suggest that HML-2 Pro would be enzymatically highly active when located in cellular compartments with acidic pHs, for example, organelles of the secretory and endocytic pathways (pH 6.7-4.7 [31]). However, there is also evidence that HML-2 Pro could potentially be active in cellular compartments with pH close to neutral, for example, nucleus, endoplasmic reticulum, and cytoplasm (the three compartments having pH 7.2 [31]), and in cellular compartments with alkaline pH, for example, mitochondrial matrix (pH 8.0 [31]).

5.1.5 Preparation of cellular protein mixture for incubation with HML-2 Pro

Apart from having a purified functional HML-2 Pro available and having knowledge of optimal *in vitro* conditions for its activity, another important aspect for identification of potential HML-2 Pro substrates concerned cellular proteins reacting with HML-2 Pro. This concerned in particular 1) a proper source for human proteins; 2) preparation of a protein mixture that preserved protein integrity and employing buffer components compatible with the subsequent TAILS procedure.

Human proteins subjected to HML-2 Pro processing derived from cell lysates of HeLa cells. HeLa is a well known human cell line that is widely employed in laboratories. That cell line was readily available in our laboratory. Considering that HeLa cells divide relatively quickly and can be maintained easily, it was considered a suited source for extraction of a higher amount of cellular proteins for use in our experiments. Expression of endogenous HML-2 Pro was not reported in the literature for HeLa cells. We verified in Western blot experiments (not shown) that endogenous HML-2 Pro was not expressed, at least not at a detectable level, in the HeLa cell line employed in our experiments. This was relevant in order to ensure that 1) in the *in vitro* reactions for TAILS experiments, HML-2 Pro activity derived from the exogenous HML-2 Pro added to the TAILS reaction; 2) exogenous Pro activity could be then inhibited efficiently by established concentrations of pepstatin A.

In a more recent experiment, using a more sensitive Western blot procedure, we gained preliminary evidence (requiring further corroboration) of very low level HML-2 Pro expression in HeLa cells (also discussed below). In any case, a low level of catalytically active endogenous HML-2 Pro was not considered an issue in our TAILS experiments. Even if present, such endogenous HML-2 Pro activity would have been inhibited by pepstatin A in control reactions as well.

Our procedure for preparation of a protein mixture from HeLa cells released proteins from their cellular compartments under native conditions, thus preserving their native state. That aspect was important in the context of subsequent incubations with HML-2 Pro, where we wanted to establish interactions with native proteins as they could also occur within the cell. Also, not properly folded proteins probably would have presented protein regions that in a native state would not have interacted with HML-2 Pro. Thus, incubating HML-2 Pro with denatured proteins would have produced flawed results not representative of cleavage events that could occur in a cellular context.

To preserve the native state of proteins, our procedure for lysis of HeLa cells was based on a physical disruption method that avoided factors inducing protein denaturation, e.g. denaturing (ionic) detergents or higher temperature. The cell lysis was achieved by freeze-thaw cycles in a detergent-free buffer, combined with short vortexing in presence of glass beads. Non-ionic detergents, that typically do not denature proteins [91], could have been employed to further promote disruption of cellular membranes without dramatically affecting protein integrity. Nevertheless, those detergents may have interfered with the subsequent TAILS procedure. We therefore omitted non-ionic detergents during cell lysis.

After cell lysis, all proteins in solution were, in principle, susceptible to protein degradation due to activity of cellular proteases released from cellular compartments [163]. Our procedure aimed at preserving protein integrity by reducing protein degradation due to cellular proteases. This aspect was relevant in order to better monitor the specific proteolytic activity exhibited by HML-2 Pro. The buffer employed for cell lysis therefore included a cocktail of protease inhibitors that inhibited a broad spectrum of cellular proteases. Of note, the cocktail of protease inhibitors employed did not contain inhibitors of aspartyl proteases, which was fundamental in order to not block HML-2 Pro activity in subsequent reactions.

5.1.6 Set up of reactions for TAILS

We verified in a small-scale experiment (see Figure 27 B) reaction conditions for TAILS documenting the processing of *in vitro*-translated HML-2 Gag by HML-2 Pro at pH 5.5 and pH 7. As expected, HML-2 Pro exhibited higher activity at pH 5.5, although proteolysis of HML-2 Gag could be observed also at pH 7.0. We also noticed that processing of HML-2 Gag by HML-2 Pro seemed to be slightly reduced in presence of HeLa-derived cellular proteins. We speculate that a lower level of processing was due to reduced interaction of HML-2 Pro with HML-2 Gag caused by competition of HML-2 Gag with other cellular substrates of HML-2 Pro.

Reaction conditions verified in small-scale experiments were employed for lager-scale reactions for TAILS. In TAILS reactions, 2 mg of HeLa-derived cellular proteins were incubated with 200 nM purified HML-2 Pro for 75 min at 37° C. One set of reactions was performed at pH 5.5, another set at pH 7. TAILS reactions at pH 5.5 could be expected to identify an as high as possible number of cellular proteins processed by HML-2 Pro because of high HML-2 Pro activity. Reactions at pH 7 could be expected to provide evidence of processing of cellular proteins by HML-2 Pro also at a pH typical for the cytoplasmic compartment, where HML-2 Pro locates most likely. Controls included pepstatin A at 200 µM, a concentration that, based on optimization experiments, was sufficient to block HML-2 Pro at 200 nM efficiently. At the end of incubations, we stopped HML-2 Pro activity by freezing reactions down on dry-ice and subsequently storing them at -80°C. This ensured that no additional reaction components for inactivation of HML-2 Pro that may have interfered with the subsequent TAILS procedure were added to reactions. Reactions were shipped on dry-ice to the laboratory of Dr. Schilling for the TAILS procedure.

5.1.7 TAILS analysis

Cellular substrates of proteases encoded by ERVs were not investigated so far. In accord, there were no known cellular proteins processed by HML-2 Pro. As for proteases encoded by exogenous retroviruses, most studies focused on non-viral substrates of HIV-1 Pro (for instance, see [82,87,187]). Such studies give an idea of the approaches used to identify cellular targets of HIV-1 Pro. In earlier studies, identification of substrates of HIV-1 Pro was based on observations of cellular effects during HIV-1 infections or transient expression of HIV-1 Pro in cell culture. For instance, altered cell shape during HIV-1 infection prompted researchers to investigate ability of HIV-1 Pro to process intermediate filament proteins [187]. Cell damage and lysis during HIV-1 infection prompted studies whether HIV-1 Pro could degrade microtubule proteins [213]. Induction of apoptosis during HIV-1 infection or cytotoxic effects of transiently expressed HIV-1 Pro prompted the investigation of cleavage of apoptosis-regulating proteins by HIV-1 Pro [146]. Altogether, results from various such studies identified several substrates of HIV-1 Pro, although the number of candidate proteins investigated in each study was very limited since selection of (relatively few) candidate proteins was just hypothesis-driven.

More recently, several advanced approaches were developed to screen protease substrates in a more systematic fashion [14]. Among such approaches, those based on proteomics technologies are known to facilitate high-throughput identification of protease substrates [14]. In the context of substrates of HIV-Pro, the great potential of a large-scale proteomics approach was demonstrated by a study by Impens et al. (2012) [82], that employed a proteomics technology termed Combined Fractional Diagonal Chromatography (COFRADIC) and identified 123 potential host cell substrates of HIV-1 Pro. That study reported, from a single analysis, a collection of HIV-1 Pro substrates much more

comprehensive than those of previous investigations. The reported catalogue of putative HIV-1 Pro host cell substrates provides information that allows to speculate on unknown roles of HIV-1 Pro in infected cells, and it can be used as an authentic reference for studies that aim at validating cellular targets of HIV-1 Pro and involvement of HIV-1 Pro in specific cellular processes.

Use of a large-scale proteomics approach was also recommended for our investigation of HML-2 Pro for which processing of cellular proteins was completely unknown so far. Similar to the Impens et al. (2012) [82] study, we also employed a large-scale proteomics approach, yet a slightly different, newer methodology developed for identification of substrates of proteases. Terminal Amine Isotopic Labeling of Substrates (TAILS) is a sophisticated proteomics technique that, like COFRADIC, is a so-called N-terminomics method, as it is based on labeling and identification of N-termini of proteins processed by a protease [106]. We had access to this methodology through a collaboration with Prof. Dr. Oliver Schilling (University of Freiburg), whose laboratory has vast expertise in TAILS and established protocols suited for our experimental set-up.

The TAILS technique discovers protease substrates by identification of protein processing fragments, specifically N-terminal protein fragments, that are quantified comparing a protease-treated sample with control reactions [14]. Substrates of various proteases have been investigated by TAILS [36,73,80], including substrates of viral proteases [86].

As described above, in our experimental design, samples for TAILS consisted of reactions where HML-2 Pro had been incubated with a protein mixture derived from HeLa cells. Control reactions differed from protease-treated samples by presence of pepstatin A that blocked HML-2 Pro activity [112]. N-terminal protein fragments generated during the reactions were isolated and identified by established protocols in the laboratory of Prof. Dr. Schilling. Different abundances of N-terminal protein fragments in protease-treated versus control reactions, as determined by TAILS, identified N-terminal protein fragments enriched in presence of HML-2 Pro, thus identifying actual human proteins cleaved by HML-2 Pro.

Our TAILS analysis demonstrated for the first time that HML-2 Pro activity can induce cleavage of hundreds of cellular proteins. As a consequence of HML-2 Pro activity, a considerable number of N-terminal protein fragments were identified as enriched in TAILS reactions. Since N-terminal protein fragments are products of unique cleavages, in the following we will refer to "cleavage events", i.e., processing of a specific protein at a specific aa position. When regarding cleavage events with at least 2-fold enrichment, we identified approximately 4300 and 2600 cleavage events in the two TAILS experiments performed at pH 5.5, whereas 690 and 1900 cleavage events were identified in two TAILS experiments performed at pH 7.0. The great number of cleavage events enriched at both pHs confirmed ability of HML-2 Pro to process cellular proteins under variable pH conditions. The lower number of cleavage events at pH 7.0 was probably due to lower activity of HML-2 Pro, which was expected

considering that through biochemical characterization of HML-2 Pro we documented approximately 10-fold lower Pro activity at pH 7.0 compared to pH 5.5. Also, small-scale reactions for TAILS showed a lower level of processing of HML-2 Gag at pH 7.0 compared to pH 5.5. Despite lower activity of HML-2 Pro at pH 7, the number of enriched cleavage events was still very high at this pH, indicating that even under non-optimal pH conditions HML-2 Pro could impact the integrity of many cellular proteins.

The multitude of cleavage events documented by TAILS amounted to hundreds of human proteins processed by HML-2 Pro. However, when regarding numbers, a discrepancy was observed between cleavage events and actual proteins cleaved. For instance, 931 cleavage events common to both TAILS pH 5.5 replicates corresponded to only 548 cleaved proteins. Such discrepancies were also noted in other studies that used TAILS or similar proteomics approaches [157][82]. Such discrepancies could be attributed to multiple cleavages in many of the cellular proteins cleaved in presence of protease. We counted approximately 300 proteins common to the two replicates at pH 5.5 and approximately 70 proteins common to two replicates at pH 7.0 with evidence of at least 2-fold enriched multiple cleavages (not necessarily in the same positions among replicates). In several instances, we observed a higher number of cleavages within the same protein, with some proteins being processed in more than 30 different positions (see Table 18).

Cleavage of a protein in only one position could be sufficient to knock down or at least alter its functionality. If one hypothesizes that interaction of HML-2 Pro with a specific cleavage site in a substrate occurs by chance, one could speculate that proteins with multiple cleavage sites have a higher chance of being processed when in contact with HML-2 Pro as several cleavage sites would be presented during the encounter. Consequently, one could speculate that proteins having many HML-2 Pro cleavage sites could be more susceptible to impairment due to HML-2 Pro proteolytic activity. Those proteins may deserve specific attention in the context of HML-2 Pro cellular targets. The relevance of multiple cleavages in particular proteins may be investigated in detail through further studies.

Combining results from TAILS at pH 5.5 and pH 7.0, we generated a list of 872 cellular proteins (see Table 21) for which TAILS provided evidence of processing by HML-2 Pro *in vitro*. Such a collection of potential targets of HML-2 Pro could be of interest for further investigations. For instance, studies validating cellular substrates of HML-2 Pro could employ that list as a reliable starting point for selection of candidate proteins. For studies that aim at elucidating biological roles of HML-2 Pro, that list could be a source of interesting hints when linking HML-2 Pro to biological pathways associated with identified substrates. Studies investigating cellular substrates of other retroviral proteases could also benefit from our list for comparative analyses.

The above-mentioned study by Impens et al. (2012) [82] that used COFRADIC to identify cellular substrates of HIV-1 Pro, reported 120 human proteins processed *in vitro*. In view of that study, our results have shown a higher number of proteins cleaved by HML-2 Pro. Differences could derive from the different design of experiments and different methodologies employed. Moreover, our TAILS-based experimental approach can be expected to be more sensitive and to thus identify more proteins than the approach employed by Impens et al. (2012) [82]. Interestingly, 60 proteins identified in our study were also identified by Impens et al. (2012) [82] likely because of overall similar specificity profiles of HIV-1 and HML-2 Pro [17]. Note that in Impens et al. (2012) [82] the total cellular protein for screening experiments derived from Jurkat T cells, while it derived from HeLa cells in our case. We assume that an even higher overlap between the identified substrates would have been observed if the same cell type had been used in both studies.

5.1.8 Even more human proteins than observed may be substrates of HML-2 Pro

The number of potential HML-2 Pro substrates identified in our experiments supports the assumption that HML-2 Pro activity could impact cell biology if expressed. It suffices to consider that altering the amount of a single cellular protein could potentially impact cell homeostasis. Some limitations of the experimental system utilized imply that even more human proteins than observed in our experiments could be substrates of HML-2 Pro. Three main factors limiting our screening are considered in the following.

First, lysis conditions likely yielded an incomplete cellular proteome. Thus, we assume that not the entire HeLa proteome has been subjected to HML-2 Pro activity. As discussed above, our procedure for extraction of proteins from HeLa cells was based on a physical cell lysis method that prioritized preservation of protein integrity rather than aiming to extract as many proteins as possible. In particular, use of detergents was avoided because of suspected incompatibility with TAILS. Detergents break lipid-lipid and protein-lipid interactions, promoting disruption of membranes and solubilization of membrane proteins [91]. We can assume that due to inadequate membrane disruption proteins associated with cellular membranes and proteins contained in subcellular compartments had a lower chance than, for instance, cytosolic proteins, to be resolved in the supernatant and not being precipitated with cell debris during centrifugation steps for protein mixture clarification. For a more complete protein extraction, protocols including mild non-denaturing detergents could be employed. In such experiments, detergents should be removed before TAILS by, for instance, dialysis, in order to avoid interferences with the TAILS procedure.

A second aspect is represented by the analysis of human proteins that are expressed at very low level in the cell. A comprehensive transcriptomic and proteomic study involving around 13,000 protein coding genes reported that approximately 10% of proteins could not be detected likely due to abundance below detection limits [216]. We suppose that, although present in the protein extract and

exposed to HML-2 Pro activity, proteins present at very low concentration (or their eventual cleavage products) went undetected in the TAILS analysis. HML-2 Pro activity against proteins expressed at very low level may be of particular interest. Alterations of their amount in the cell could induce stronger effects since the total number of respective proteins would be reduced considerably and only few remaining copies of respective proteins could compensate the degradation. Such aspects should be addressed in the future when further improved experimental strategies will enable analyses of proteins expressed at very low level in the cell.

Third, our analysis certainly missed human proteins that are not expressed in HeLa cells. Although all the different cell types of an organism contain the same DNA, they do not necessarily express the same proteins. In fact, each type of specialized cell type has its own pattern of gene expression and, consequently, it contains a specific set of proteins [216]. Therefore, TAILS experiments using protein extracts from human cell lines other than HeLa, or from different human tissues, can be expected to identify additional proteins as (candidate) substrates of HML-2 Pro. Novel detections could involve hundreds of proteins, as can be inferred from studies comparing protein expression profiles across different cell lines. For instance, a recent comparative analysis of expression profiles of cancer cell lines derived from nine human tissue types reported that ~5000 proteins could be identified in at least one cell line of every single tissue group, while another ~5000 proteins showed a distinct expression pattern between tissues [69].

After all, the three above described limitations were not considered an issue in our experiments because we did not intend to identify all the cellular substrates of HML-2 Pro, instead we aimed to provide proof of principle of HML-2 Pro processing of cellular proteins and produce a representative list of potential substrates of HML-2 Pro that could represent a starting point for future investigations.

5.1.9 Validation of substrates of HML-2 Pro identified by TAILS

As our own and other studies (for instance, see [82][86]) demonstrated, proteomics-based technologies for protease substrate discovery produce meaningful results. However, it needs to be considered that such technologies have limitations with regard to reproducibility and sensitivity. Because of this, proteomics-based analysis may be regarded as screening strategies requiring biochemical and cell-based assays for validation of putative substrates of proteases [182].

In our study, our compilation of data from TAILS experiments provided a list of proteins that we defined as "potential" substrates of HML-2 Pro since further experimental evidence will be required to validate their processing by HML-2 Pro. Other studies that employed TAILS or similar proteomics approaches for protease substrate identification typically reported specific experiments for validation of candidate proteins *in vitro* and, in some cases, in the cellular context (for instance, see [86,106,107]).

Since our study provided for the first time a collection of potential substrates of HML-2 Pro, it was particularly important to validate results from TAILS experiments. Therefore, we performed a number of specific experiments that aimed at verifying processing of candidate proteins both *in vitro* and *in vivo*.

Independent from the approach employed, verification experiments for hundreds of proteins could not be performed for practical reasons. Various criteria can be applied for narrowing down proteins for further studies. For instance, some criteria applied by other studies prioritized 1) proteins with family members that are known substrates of the protease of interest or of similar proteases [157]; 2) proteins with specific aa residues in P1 and P1' of identified cleavage sites [86,107]; 3) proteins involved in molecular pathways of particular interest in the specific context of the study [86]; 4) proteins considered biologically relevant on the basis of their cellular localization [106]. Sometimes, proteins were chosen randomly (for instance, see [86]).

We aimed at a reasonable number of candidate proteins for verification experiments. TAILS data included detailed information related to the proteins identified, for instance, position of the cleavage site, specific aa residues present at the cleavage site, molecular mass, GO terms for cellular localization, processes, functions. We applied several criteria to that additional information in order to filter a starting list of potential candidate proteins.

To initially facilitate selection of proteins efficiently cleaved by HML-2 Pro, we started the filtering based on our list of 809 potential substrates identified in TAILS experiments at pH 5.5, for which the higher activity of HML-2 Pro favored higher confidence identification of substrates. The selection of candidates further involved a filter for certain aa residues in positions P1 and P1' of observed cleavage sites. After filtering for favored aa residues in positions P1 and P1' to be required for more efficient cleavage by HML-2 Pro, still 145 human proteins had to be considered as potentially processed by HML-2 Pro. Further filtering criteria therefore included a specific molecular mass range due to technical limitations of the *in vitro* transcription/translation system used for verification purposes, as well as cellular localization and potentially disease-relevant biological functions of candidates. The entire filtering procedure resulted in 14 candidate human proteins (see Table 20) to be further tested in *in vitro* and *in vivo* experiments.

5.1.10 Verification experiments *in vitro*

Employing specific cleavage assays *in vitro*, we investigated whether the 14 selected candidate proteins identified by TAILS were *bona fide* substrates of HML-2 Pro. I*n vitro* cleavage assays are typically performed to validate targets of a protease of interest biochemically [86,106,157]. When considering other studies reported in the literature, approaches employed for such cleavage assays share a common

denominator, specifically they incubate a full-length candidate substrate with active or inactive protease of interest and document, taking into account control reactions, a reduced amount of full-length candidate substrate and appearance of cleavage products, with these two criteria indicating processing of the full-length protein.

For our verifications *in vitro*, selected candidate proteins were transcribed/translated *in vitro* introducing ^{35}S-labelled methionines or, alternatively, an HA-tag in their C-terminus. Candidate proteins generated *in vitro* were incubated with purified HML-2 Pro under buffer conditions established for optimal HML-2 Pro activity. Control reactions lacked HML-2 Pro, or HML-2 Pro was inhibited by pepstatin A. Processing of HA-tagged or ^{35}S-labelled candidate proteins was visualized through SDS-PAGE/Western blot or phosphor imaging, respectively.

Our experimental approach was convenient regarding several aspects as briefly discussed in the following. The convenience of our system prompted us to subject a subset of 14 candidate proteins to *in vitro* validations, a number considerably higher compared to numbers of candidate proteins examined in similar studies [106,157,196]. The experimental system (TNT T7 system) used for generation of candidate proteins *in vitro* (see Material and Methods, chapter 3.2.2.2), allowed for expression of sufficient amounts of proteins, in a practical way while preserving the proteins' native state. As an alternative approach employed in other studies (for instance, see [86,157]), candidate proteins were transiently expressed in cells from specific expression plasmids, and cell lysates were used for incubations with the protease of interest. Instead, the TNT T7 system was suited for transcription/translation in an Eppendorf tube using PCR products as DNA templates. Generation of proteins *in vitro* required only 90 min of incubation and yielding sufficient amounts of proteins for our experiments. Thus, that approach did not require cloning steps for generation of specific expression plasmid constructs that would have been necessary for protein expression in cells. Moreover, protein extractions after transient expression in cells were not required, thus also reducing exposure of proteins to cellular factors causing denaturation or degradation of proteins. Altogether, the experimental procedure for verifications *in vitro* was less time consuming and more straightforward than procedures involving expression of candidate proteins in cells.

Our approach allowed us to biochemically validate results from TAILS analysis using two detection approaches, HA-tag and a ^{35}S-label. Two detection approaches proved advantageous for better visualization of processing products. Because the HA-tag was included at the C-terminus of proteins, the HA-tag approach could not detect processing products consisting of internal regions of candidate proteins lacking the HA-tag. However, such processing products could be detected via ^{35}S-labeled methionine provided that the processing product contained a methionine and the signal was above detection limits.

Sets of candidate proteins marked with the same tag/label were furthermore advantageous for experiments because it did not require specific antibodies for detecting candidate proteins in Western

blots. Antibodies are relatively expensive and may require optimization experiments. When validating 14 different candidate proteins the use of different primary antibodies would have been relatively costly and more time consuming than the favored approach.

Our verification assays confirmed processing by HML-2 Pro for 9 out of 14 candidate proteins *in vitro*. For 8 candidate proteins we documented formation of cleavage products accompanied by a more or less pronounced reduction of full-length protein. In two cases (HSP90AA1 and C15orf57) cleavage products were also observed in controls with pepstatin A, although at much lower level. The same was observed in positive control reactions monitoring processing of HML-2 Gag, a specific substrate of HML-2 Pro. Presence of cleavage products in controls with pepstatin A may suggest high specificity of HML-2 Pro for HSP90AA1 and C15orf57, the cleavage of which likely could not be completely inhibited by the amount of pepstatin A present in the reaction. For one candidate protein (PDIA3), cleavage was evidenced only by reduction of full-length protein. In that case, absence of detectable cleavage products could not be explained with certainty, although it could be due to small (undetectable) sizes of cleavage products. Verification of one candidate protein (DDX3X) remained ambiguous.

The 9 verified substrates corresponded to 65% of candidate proteins selected. Projecting that percentage to the list of 145 proteins filtered for selection of candidates with favored amino acids in positions P1 and P1' of TAILS-identified cleavage sites, one would expect approximately 80 other proteins with a high chance of successful verification of cleavage by HML-2 Pro. Moreover, it is conceivable that a multitude of potential substrates, initially excluded by filtering for favored aa residues in P1 and P1', could also be verified. As also discussed further below, certain experiments *in vivo* verified cleavage of proteins with no favored amino acids in P1 and P1' as per TAILS results. One would expect verification of those proteins also *in vitro,* where reaction conditions are highly favorable. In view of that, when projecting *in vitro* verification results to the unfiltered list of 809 proteins used for selection of candidates, one can expect not just 80, but hundreds of proteins with a high chance of successful verification of cleavage by HML-2 Pro.

As speculated above, proteins with multiple cleavages could be particularly susceptible to HML-2 Pro processing. Out of the 9 validated candidates, 7 proteins showed multiple cleavages based on TAILS data, and 2 proteins showed a single cleavage. The majority of validated candidates showing multiple cleavages would be in line with our hypothesis at first sight. However, there was another aspect, described in the following, that was not in line with our hypothesis. Levels of reduction of full-length protein may provide indications of greater susceptibility of candidates with multiple cleavage sites to processing by HML-2 Pro. One may have expected greater reduction (in presence of HML-2 Pro) of full-length protein for those candidates with a higher number of cleavage sites. However, in our assays, reduction (in presence of HML-2 Pro) of full-length proteins showed variable levels among candidates,

with no obvious association between the level of reduction and the number of cleavage sites as reported by TAILS. This suggested that among the candidates tested susceptibility to cleavage by HML-2 Pro was independent of the number of cleavage sites. Certainly, additional specific investigations will be required in order to further elucidate this.

As for verified candidate proteins with multiple cleavage sites, most of the time there was at least one cleavage product observed in verification experiments that was also predicted by cleavage sites predicted by TAILS. Such an overlap of results from different experimental approaches corroborated both experimental approaches. However, not all the cleavage products predicted by TAILS were observed in verification experiments. This incongruence could be due to higher sensitivity of TAILS in comparison with verifications *in vitro*. It is conceivable that several processing products generated in verification assays *in vitro* were below the detection limit of Western blot and phosphor imaging. Also, some processing products could not be detected even if present at high level, specifically processing products without the C-terminal HA could not be detected via Western blot, and processing products without methionines could not be detected via the ^{35}S-label. Molecular weight of processing products must also be considered as a factor influencing detection. In particular, detection of cleavage products of low molecular weight was probably difficult in light of a relatively low concentration of polyacrylamide (10% or 12%) used in gels for SDS-PAGE. It is likely that that polyacrylamide matrix did not effectively separate smaller-sized protein fragments. Moreover, for experiments involving Western blot, it is likely that small-size proteins were not immobilized efficiently at the blot membrane during transfer. As can be seen in, for instance, Figure 35, Coomassie-stained blotted small-sized proteins appear much fainter in comparison with larger-sized proteins. Specific experiments aimed at identification of small-sized cleavage products could make use of gels with a higher concentration of polyacrylamide or gradient gels. Time of transfer during Western blot could also be optimized to avoid loss of small-sized proteins.

5.1.11 Verification experiments *in vivo*

So far, experiments demonstrating cleavage of human proteins by HML-2 Pro were based on an *in vitro* set-up. However, experiments *in vitro* are not necessarily representative of protein processing in the cellular, *in vivo* context. Within a cell, reaction conditions may vary. Proteins are not distributed homogeneously but may be enriched in cellular compartments. Conditions can differ considerably between cellular compartments. Proteins are subjected to protein turnover, that is, continuous synthesis and breakdown. Also, a multitude of cofactors play important roles by influencing molecular interactions. Thus, experimental verifications *in vivo* were fundamental to confirm that proteolytic activities observed for HML-2 Pro *in vitro* also applied to the complex environment of the human cell.

For our validations *in vivo*, we focused on five candidate proteins (CIAPIN1, HSP90AA1, MAP2K2, TUBA1A, C15orf57) for which processing by HML-2 Pro had been already verified by experiments *in vitro*. Those candidates included four proteins with well-known crucial functions in the human cell. CIAPIN1 is involved in the negative control of cell death [40]. HSP90AA1 is a molecular chaperone that promotes maturation, structural maintenance and proper regulation of specific target proteins involved, for instance, in cell cycle control and signal transduction [40]. MAP2K2 is a protein kinase known to play critical roles in important signal pathways involved in cell cycle regulation and proliferation [40]. TUBA1A is the major constituent of microtubules [40]. We also included C15orf57, a protein with currently unknown cellular functions [40].

The experimental approach used for validations *in vivo* consisted of transient co-expression of selected HA-tagged candidate proteins together with HML-2 Pro in HEK293T cells. As controls, candidate proteins were expressed without HML-2 Pro, or they were co-expressed with a mutant, enzymatically inactive HML-2 Pro. Cells were lysed 24 hpt and processing of candidate proteins was monitored via detection of the HA-tag by Western blot. Criteria for validating candidate proteins *in vivo* were the same as in verification experiments *in vitro*, i.e., document a reduced amount of full-length candidate substrate, ideally accompanied by appearance of cleavage products.

Our experiments provided evidence of processing by HML-2 Pro for the five tested candidate proteins, providing first documentation of cellular substrate of HML-2 Pro in the cellular context. Moreover, in the course of our project, a collaboration with Dr. John Goodier (McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA) allowed us to increase the number of cellular proteins confirmed as substrates of HML-2 Pro *in vivo*. Results from this collaboration have been published in Rigogliuso et al. (2019) [164]. In the laboratory of Dr. Goodier, a set of proteins of biological and clinical interest that were readily available as cloned cDNAs in that laboratory were co-expressed with HML-2 Pro in HEK293T cells. The HML-2 Pro co-expressed was identical to the one used in our own *in vivo* experiments as plasmid constructs for wild-type HML-2 Pro and respective mutant Pro were provided by us. Tested proteins were each labeled with tags such as FLAG, T7, Myc, or HA.

Experiments conducted in the laboratory of Dr. Goodier provided evidence of processing by HML-2 Pro for 16 additional human proteins (see Table 22), so that a total of 21 different cellular proteins could be demonstrated as substrates of HML-2 Pro *in vivo*. The proteins verified in our own experiments and those of Dr. Goodier further validated our TAILS analysis as they were included in the list of 872 potential substrates of HML-2 Pro identified by TAILS. Moreover, 14 proteins verified in Dr. Goodier's laboratory did not present favored aa residues (for processing by HML-2 Pro) in P1 and P1' as per TAILS results. Such evidence suggested that, among the 872 proteins identified as

potential substrates of HML-2 Pro, even proteins with unfavored aa residues in HML-2 Pro cleavage sites as identified by TAILS should be considered of interest for further validations.

The 21 different human proteins validated *in vivo* can be considered as primary candidates for further studies aiming at elucidating potential roles of HML-2 Pro within the cell. For example, heat shock protein HSP 90-alpha (HSP90AA1), heat shock protein family A member 5 (HSPA5) and heat shock protein beta-1 (HSPB1) could represent interesting candidates for further studies. As apparent from their names, the three proteins are members of the same protein family, heat shock proteins (HSPs), and it is conceivable that even more proteins from the same family could be susceptible to processing by HML-2 Pro. We already verified processing of HSP90AB1 *in vitro*, observing similar cleavage products generated by the processing of HSP90AA1 both *in vitro* and *in vivo*. HSP90AB1 displays high aa sequence similarity with HSP90AA1 and HML-2 Pro cleavage site sequences, as identified by TAILS, were nearly identical in both proteins. Thus, there is significant information to believe that HSP90AB1 could be processed by HML-2 Pro *in vivo* as well. In any case, there is evidence to assume that cellular processes involving HSPs could be vulnerable to HML-2 Pro activity. Many HSPs (including the four mentioned above) are chaperones, i.e., factors assisting in protein folding, maintenance of proteome integrity, and protein homeostasis (proteostasis) [104]. Chaperones are upregulated upon stress conditions but also expressed under normal conditions to maintain proteostasis [104]. In view of our results, the network of molecular chaperones could potentially be deregulated in the case of HML-2 Pro expression. It is known that disruption of proteostasis is implicated in the pathogenesis of numerous diseases, including neurodegenerative diseases like Parkinson's disease, Huntington's disease and Alzheimer's disease, but also other aging-related diseases like cancer [104]. Deregulation of HML-2 proviruses in several diseases including cancer and neurologic diseases has been already pointed out [35,64,171]. Further studies could investigate an association between HML-2 Pro activity and deregulated proteostasis contributing to degenerative diseases. Such investigations could involve, for example, transient expression of HML-2 Pro in cells and monitoring of protein misfolding with associated formation of intracellular deposits of aggregated proteins, that is typical for several neurodegenerative diseases [104]. Techniques for monitoring protein misfolding and aggregation in living cells are available [74].

In the context of HSPs as targets of HML-2 Pro, an important piece of information was provided by an experiment performed in the laboratory of Dr. Goodier that documented processing of endogenous HSP90 protein by transiently expressed HML-2 Pro. HSP90 processing was monitored specifically by an α-HSP90 antibody. Documenting experimentally an altered level of constitutively expressed endogenous HSP90 protein in presence of HML-2 Pro provided additional support to the notion that endogenous counterparts of candidate proteins transiently expressed by us in verification experiments are also processed by HML-2 Pro.

5.1.12 Instability of processing products generated *in vivo*

In our own and the experiments in the laboratory of Dr. Goodier, levels of processing of candidate proteins *in vivo* ranged from slight to complete reduction of full-length candidate protein, sometimes accompanied by presumed processing products. Different levels of reduction of full-length proteins and respective cleavage products could be due to variable specificity, thus cleavage efficiency, of HML-2 Pro against each candidate protein. Protein levels could have also been influenced by cellular factors, for instance, different turnover rates of candidate proteins within the cell.

When considering experiments conducted in our laboratory, no processing products were detected for two candidates cleaved by HML-2 Pro *in vivo*. In the case of CIAPIN1, this was somewhat expected since no cleavage products could be detected via an HA-tag in an *in vitro* assay, likely due to cleavage in proximity of the HA-tag. For candidates with processing products detected *in vivo*, those products were of low amounts based on Western blot intensities, especially when compared with stronger signals observed in assays *in vitro*. Absence or low level of processing products could be a consequence of an instability of such processing products in the cellular environment. Improved experimental set-ups could decrease a suspected instability of processing products, for instance, treatment of cells with proteasome inhibitors, thus inhibiting degradation of processing products.

Assays performed in the laboratory of Dr. Goodier included proteasome inhibitor MG132, a peptide aldehyde commonly used for inhibiting multiple peptidase activities of the proteasome [90]. Several of the verification experiments done by Dr. Goodier documented reduction of full-length proteins without generation of detectable cleavage products, which may argue for instability of processing products even in presence of MG132. Other cellular factors could influence detection of cleavage products, for instance, processing of proteins in proximity of the tag with generation of small tagged peptides undetectable by Western blot due to inefficient separation during electrophoresis and binding to the Western blot membrane. Other studies employing MG132 in HEK293T cells applied different MG132 concentrations and different time points of cell exposure to MG132 [90]. In experiments by Dr. Goodier, HEK293T cells were treated with MG132 for 4-5 h at a final concentration of 10 µM. Alternative conditions for treatment of cells with MG132 could be tested to verify whether more optimal parameters may promote stability of cleavage products. In such experiments, inclusion of controls monitoring the proteasome activity under different conditions may be of benefit. A GFP-based proteasome functionality reporter detectable via Western blot might be employed in this context [122].

5.1.13 Processing products *in vitro* vs *in vivo*

Processing products observed *in vivo* were similar in size to processing products observed in verification experiments *in vitro*. This overlap corroborated results, firmly indicating that proteolytic events verified *in vitro* were not just due to optimized *in vitro* conditions, instead such specific cleavages occurred in the same way under physiological conditions.

Such an overlap could not be verified for processing products identified in experiments in the laboratory of Dr. Goodier since respective proteins had not been subjected to specific verification experiments *in vitro* by us. However, we could verify that the size of processing products detected in experiments in the laboratory of Dr. Goodier matched well with cleavage products predicted by cleavage sites as identified in TAILS experiments. Those findings provide further evidence of the overlap between processing products detected in *in vitro* and *in vivo* experiments.

5.1.14 Novel protein fragments generated by HML-2 Pro activity could have biological implications in cellular processes

It was shown that HIV-1 Pro directly cleaves procaspase 8 with formation of a novel Casp8p41 fragment which, in turn, induces apoptosis [147]. In a similar way, processing products generated by HML-2 Pro activity could have biological implications in the cellular processes. Concerning procaspase 8, the protein was present in our complete, unfiltered TAILS result list. However, the position of the cleavage was not located in the protein region necessary for release of the Casp8p41, suggesting that processing of procaspase 8 by HML-2 Pro would not have the biological implications reported for HIV-1 Pro. In any case, other cellular proteins could potentially produce biologically active products due to cleavage by HML-2 Pro.

For example, caspase 1 was represented in the 872-protein list and is likewise an enzyme translated as a physiologically inactive zymogen that requires cleavages in order to become activated. Caspase 1 is activated by autoproteolysis which results in the generation of the two characteristic subunits of the catalytically active enzyme, as well as the removal of the N-terminal CARD (caspase activation and recruitment domain) [27]. As for the region within caspase 1 cleaved by HML-2 Pro, according to TAILS results caspase 1 was cleaved in the so-called "propeptide", a region that represents a modulator of functional activity of proteases [47] and also harboring the CARD domain in the case of caspase 1. Maintaining enzymes in an inactive state is among the best-known functions of propeptides [47]. Caspase 1 is also known as interleukin-1 converting enzyme (ICE) for its ability to activate interleukin-1 that plays a central role in the inflammatory response. One can hypothesize that, by cleaving the propeptide, HML-2 Pro could potentially activate caspase 1/ICE, thus indirectly also activating interleukin-1. Interleukin-1 is considered to play an important role in autoinflammatory diseases [94]. By indirect activation of interleukin-1, HML-2 Pro might trigger uncontrolled inflammatory response, that is a cellular process for which impairment was previously associated with HML-2 upregulation [62]. On the other side, HML-2 Pro processed caspase 1 at aa 34, while the CARD domain is usually released by cleavage at aa 103 [27]. Only approximately one third of the CARD domain is thus removed from caspase 1 by HML-2 Pro. The functional consequences of such a cleavage of caspase 1 by HML-2 Pro are currently unknown and thus need to be investigated in more detail. Specific experiments to verify potential mechanisms of modulation of caspase 1 and interleukin-1 by HML-2

Pro would thus be of great interest. Specific analyses of other zymogens potentially activated by HML-2 Pro substrates would be equally interesting but remain to be done as well.

5.1.15 Use of HEK293T cells in verification experiments

For verification experiments *in vivo*, we have used HEK293T cells because they were readily available to us and were advantageous regarding three important aspects.

First, HEK293T cells do not express detectable amounts of endogenous HML-2 Pro, as we initially confirmed in preliminary experiments (not shown) detecting endogenous HML-2 Pro in HEK293T cells by employing an α-HML-2 Pro polyclonal antibody 9367 [185]. Absence of HML-2 Pro in HEK293T cells was also evident in all the verification experiments *in vivo* (see Figure 35) where HML-2 Pro could not be detected (via α-HML-2 Pro antibody 9367 [185]) in HEK293T cells transfected with control plasmids, whereas it was clearly present in HEK293T transfected with HML-2 Pro-encoding expression plasmid. Moreover, we attempted in another experiment detection of endogenous HML-2 Pro in various tumoral cell lines (see Figure 42). There was no endogenous HML-2 Pro detectable in untransfected HEK293T cells, the latter also included in that specific experiment as a negative control.

Second, HEK293T could be readily transfected in order to transiently express proteins of interest, with relatively high transfection efficiencies and no detectable cytotoxicity.

Third, expression of proteins of interest in HEK293T proved very efficient when driven from a CMV promoter [158] that was present in plasmid constructs employed.

5.1.16 Experiments *in vivo* did not involve exactly the same subregion of HML-2 Pro ORF used for experiments *in vitro*.

HML-2 Pro expressed in *in vivo* experiments was not identical to HML-2 Pro expressed in *in vitro* experiments. In particular, Pro expressed in *in vitro* experiments consisted of a subregion of the HML-2 Pro ORF excluding a C-terminal region that was included in HML-2 Pro expressed *in vivo*. Since the C-terminal region was suggested to exert a regulatory/inhibitory effect *in cis* against HML-2 Pro activity [205] the *in vivo* experimental design may be closer to physiological conditions in this regard. We therefore tested HML-2 Pro activity *in vivo* in presence of the C-terminal region. In view of the potentially relevant experimental differences, a quantitative comparison of proteolytic activity of HML-2 Pro between experiments *in vitro* and *in vivo* was not attempted. However, we noticed that presence of the C-terminal region did not alter HML-2 Pro activity considerably, as the enzyme was still capable of efficient self-cleavage and cleavage *in trans*. Moreover, examples of candidate proteins cleaved in the same positions by both HML-2 Pro variants, evidenced by similar-sized cleavage products, indicate that identical cleavages could be catalyzed with or without the C-terminal region.

5.1.17 HML-2 Pro was expressed at levels sufficient for observing its activity

In order to document HML-2 Pro activity within cells, we had to avoid undetectable processing of candidate proteins because of low-level expression of HML-2 Pro. Two expedients were considered to express HML-2 Pro at a level sufficient for observing its activity.

First, our *in vivo* experimental set-up relied on plasmid constructs that enabled high expression of HML-2 Pro from a human cytomegalovirus (CMV) immediate-early promoter present. A CMV promoter is commonly used to drive gene expression in mammalian expression vectors, and although its promoter efficiency can be variable from cell to cell, CMV is known to induce strong expression in HEK293T cells [158].

Second, with regard to HML-2 proviral structure it is known that HML-2 Pro is translated as a Gag-Pro polyprotein from which HML-2 Pro self-processes [78,185]. Since HML-2 Gag-Pro precursor protein is translated via an occasional ribosomal frameshift between Gag and Pro ORFs, lesser amounts of Pro are thus likely produced in cells in comparison with Gag. To promote high levels of HML-2 Pro, HML-2 Pro was not expressed as Gag-Pro precursor in our experiments but as a precursor comprised of a subregion of the Pro ORF, from which Pro then self-processed. We assumed that, independent of the Pro precursor HML-2 Pro would, once released, exhibit identical proteolytic activity. This assumption was supported by equivalent proteolytic activity exhibited by HML-2 Pro and EGFP-Pro, the two different HML-2 Pro precursors transiently expressed in our experiments *in vivo* (Figure 35, compare Aa and Ab).

5.1.18 Monitoring of HML-2 Pro activity while expressed in cells

Our experimental design ensured monitoring of HML-2 Pro activity within cells and not during or after isolation of cell lysates. Following protein expression, cell lysis was performed in presence of a cocktail of protease inhibitors, that prevent degradation of proteins by cellular proteases during cell lysis. Notably, the lysis buffer employed included pepstatin A to ensure that candidate proteins were not processed by HML-2 Pro during and after cell lysis. The same precaution applied to experiments performed in the laboratory of Dr. Goodier.

5.1.19 HA-tag is not processed by HML-2 Pro

During our verification experiments, we ruled out the possibility that HML-2 Pro may have cleaved the HA-tag added to the C-teminus of candidate proteins, thus misleading interpretation of results. That hypothesis arose from considerable reduction of some full-length proteins in presence of HML-2 Pro but no apparent cleavage products. We ruled out processing of the HA-tag for the following reasons. First, if the HA-tag would have been processed by HML-2 Pro, all candidate proteins incubated with active HML-2 Pro should have shown a similar reduction of full-length protein due to processing of the HA-tag. This was not the case, specifically levels of full-length protein reduction were very variable

among candidates, with candidate proteins not affected by presence of active HML-2 Pro. Second, experiments performed by Dr. Goodier showed reduced levels of full-length protein for candidate proteins tagged with epitopes other than HA. This supports the conclusion that the observed cleavage of candidate proteins was not due to specificity of HML-2 Pro against the HA-tag of candidate proteins.

5.1.20 Processing of candidates or HA-tag is not due to activation of apoptosis

We were able to document cell death following expression of HML-2 Pro in cells (see later). As for verification experiments *in vivo,* we therefore had to consider activation of caspases. It was previously reported that the HA-tag loses immunoreactivity during apoptosis by cleavage by caspase-3 and-7 [181]. Also, candidate proteins might have been processed by caspases rather than HML-2 Pro.

We ruled out those concerns. First, verification experiments *in vivo* involved candidate proteins for which processing had been verified in experiments *in vitro.* Experiments *in vitro* did not involve caspases; thus, processing of candidate proteins could be attributed solely to HML-2 Pro, the specificity of which was further demonstrated by reduced processing in the presence of pepstatin A. Moreover, processing products observed *in vivo* were of sizes similar to those of processing products observed *in vitro*, providing additional strong evidence that they were generated specifically by HML-2 Pro activity and not by caspase activity.

Second, FACS data (see Figure 40) showed that 24 hpt, the time point when cell lysates for verifications *in vivo* were collected, only a minor proportion (~6%) of the cell population expressing HML-2 Pro was undergoing cell death. Thus, if processing of candidate proteins in verification *in vivo* was due to activation of apoptosis, one would expect only a slight reduction (~6%) of full-length protein signal for all proteins tested. This was not the case in our verifications *in vivo*, in fact, amounts of full-length candidate proteins were reduced considerably (greater than 6%) in presence of HML-2 Pro, with complete disappearance of the full-length protein in one case (C15orf57).

Third, measuring levels of cleaved PARP and caspase 3, both cellular markers of apoptotic proteolytic processes, amounts of those proteins were reduced considerably following treatment of cells with apoptosis-inducing staurosporin. However, activation of apoptosis by staurosporin, together with activation of caspases did not reduce amounts of full-length candidate proteins and did not induce formation of processing products derived from candidate proteins.

Fourth, verification experiments performed in the laboratory of Dr. Goodier also showed reduced levels of full-length protein for candidate proteins carrying epitope-tags other than HA.

5.1.21 Establishing conditions for more physiological expression levels of HML-2 Pro

As mentioned above, high level expression of HML-2 Pro was important for the purpose of our investigation. Further analyses may establish more physiological expression levels of HML-2 Pro in cell culture experiments. This could be achieved by, for instance, plasmid constructs carrying

promoters inducing lower expression levels of cloned coding sequences compared to CMV. Human ubiquitin C promoter (UBC), mouse phosphoglycerate kinase 1 promoter (PGK) and herpes simplex virus thymidine kinase (TK) promoter are examples of promoters that have been shown to be much weaker than CMV [63,158]. Employing conventional cloning strategies, CMV promoter could be replaced in our so far expression plasmids by one of those weaker promoters, or HML-2 Pro coding sequences could be cloned into eukaryotic expression vectors harboring respective promoters.

In order to establish more physiological expression systems, knowledge of physiological amounts of endogenous HML-2 Pro in cells in the disease-context would be very helpful. In preliminary experiments (discussed in more detail further below), we investigated presence of endogenous HML-2 Pro in cell lines derived from tumor tissues with known upregulated HML-2 transcription. Endogenous HML-2 Pro levels in those cell lines were above Western blot detection limits when using an α-HML-2 Pro polyclonal antibody 9367 [185]. Levels of endogenous HML-2 Pro was much lower in those cell lines than the level of HML-2 Pro in transiently transfected HEK293T cells. However, we also established experimentally that endogenous HML-2 Pro present in those cell lines efficiently processes a transiently expressed HML-2 Pro precursor (unable to self-process) *in trans*. This suggested that, proteolytic activity of HML-2 Pro at physiological levels was sufficient to detect its proteolytic activity in Western blot experiments (this aspect is discussed in more detail further below). Levels of endogenous HML-2 Pro present in those cell lines thus may be employed as a reference for establishing additional tools to express HML-2 Pro at physiological levels.

5.1.22 Advantages of an EGFP-fused HML-2 Pro

HML-2 Pro was expressed as a sole HML-2 Pro or as an EGFP-fused HML-2 Pro (EGFP-Pro). Both forms consisted of the HML-2 Pro precursor including the HML-2 Pro self-processing sites that allowed for monitoring of HML-2 Pro autocatalytic activity. EGFP could be readily detected through Western blot, providing a suited system for monitoring the self-processing of EGFP-Pro. The EGFP-Pro set-up was favorable when monitoring transfected cells under the microscope regarding transfection efficiency and cell viability of EGFP-positive cells. In both set-ups HML-2 Pro exhibited autocatalytic activity, releasing itself from the precursor protein as expected. Also, the HML-2 Pro activity against cellular substrates was not different between the two HML-2 Pro variants. Thus, the EGFP-tag did not seem to affect HML-2 Pro activities. In view of the experimental advantages, the EGFP-Pro may be preferable for further studies on HML-2 Pro *in vivo*.

5.1.23 Potential involvement of HML-2 Pro in cell death

This work comprises an initial investigation of a potential involvement of HML-2 Pro in cell death. During cell culture experiments, we noted that expression of HML-2 Pro in HeLa or HEK293T cells

caused altered cell morphology accompanied by detachment of cells from the surface of tissue culture flasks. Accumulation of cells floating in the culture medium became particularly evident around 24 hpt, when HML-2 Pro expression was ongoing for several hours. Floating cells were of a rounded morphology, suggesting that they were undergoing cell death, potentially apoptosis. Such a phenotype appeared considerably reduced when HML-2 Pro was expressed in presence of 1 μM indinavir, a potent inhibitor of HIV-1 Pro [209]. Moreover, cells were not affected in a similar way when an HML-2 Pro inactive mutant was expressed. This lead us to conclude that the observed cell death was a direct consequence of HML-2 Pro activity.

Analysis of apoptosis markers showed enhanced activation of apoptotic processes likely due to HML-2 Pro activity. In particular, 24 hpt, levels of cleaved PARP and caspase 3, two commonly used apoptosis markers, were significantly higher in cells expressing HML-2 Pro when compared with cells expressing inactive HML-2 Pro and untreated cells. This observation suggested that apoptosis may be the actual mechanism of cell death that is activated by HML-2 Pro. This would be in line with a previous report that HIV-1 Pro induces cell death via the apoptotic pathway [169]. For COS7 cells expressing HIV-1 Pro, morphological changes typical of necrosis have been observed [18], suggesting that mechanisms of cell death other than apoptosis could also be triggered by HIV-1 Pro. As for HML-2 Pro, one cannot rule out involvement of different mechanisms of cell death based on the currently very limited state of knowledge.

Using fluorescence microscopy, we focused on the Pro-positive cell population present in *in vivo* expression experiments. In one experiment, after HML-2 Pro expression in HeLa cells, Pro-positive cells were immunostained with α-HML-2 Pro polyclonal antibody 9367 [185]. In another experiment, we expressed EGFP-Pro, that self-processes releasing EGFP and Pro, in HEK293T cells and visualized EGFP-positive (thus Pro-positive) cells. In both experiments, we visualized by fluorescence microscopy the Pro-positive cell population. Many cells expressing HML-2 Pro were of a normal morphology, even 24 hpt. This suggested that only a small proportion of the Pro-expressing cell population was driven into cell death during the course of the expression experiment. We quantified by FACS analysis the relative amount of cell death following HML-2 EGFP-Pro expression in HEK293T cells. We found that 24, 30 and 48 hpt, the proportion of HML-2 Pro-expressing dying or dead cells represented approximately 6%, 12% and 15%, respectively, of the entire cell population. However, these numbers indicate that many Pro-positive cells were still alive up to 48 hpt. Those cells represented approximately 50% of gated live cells. Based on those findings, HML-2 Pro does not seem to exert a cytotoxic effect that induces cell death inevitably. The reasons for why not all cells respond in the same way to HML-2 Pro expression remain to be clarified. Experiments monitoring Pro-expressing cells later than 48 hpt should be addressed specifically in order to elucidate whether cells can survive even after overexpressing HML-2 Pro for longer time periods.

Among potential cellular substrates of HML-2 Pro, approximately 100 proteins were involved in cell death (discussed also below). Those proteins included negative and positive regulators of apoptosis. One can speculate that HML-2 Pro activity might impair levels of apoptosis regulators, thus promoting or suppressing induction of cell death. Involvement of HML-2 Pro in cell death may have pathological implications. Alteration of apoptotic processes is related to cancer, neurological disorders, cardiovascular disorders and autoimmune diseases [55]. Among diseases that have been associated with HML-2 [64], there are several cell-death related diseases including various types of cancer and amyotrophic lateral sclerosis. By influencing apoptotic processes, HML-2 Pro may have crucial roles in development of diseases associated with cell death.

More specific investigations will be required to elucidate in more detail the mechanism(s) by which HML-2 Pro causes cell death.

5.1.24 Preliminary experiments showing presence of endogenous HML-2 Pro in human cells

We demonstrated presence of enzymatically active endogenous HML-2 Pro in cell lysates from two cell lines derived from teratocarcinoma, namely Tera-1 and NCCIT, and two cell lines derived from human melanoma, namely SK-MEL-28 and MeWo. We focused on those cell lines because they are derived from tumor types with upregulated HML-2 expression [28,171,179]. Our findings corroborated in particular findings for Tera-1 and NCCIT cell lines, for which presence of endogenous HML-2 Pro had been reported before [112,179]. In Sauter et al. (1995) [179] expression of endogenous HML-2 Pro in Tera-1 cells was demonstrated indirectly by detection of processed HML-2 Gag-derived proteins. Similarly, Kuhelj et al. (2001) [112] observed processing of endogenous HML-2 Gag in NCCIT cells, moreover demonstrating that Gag processing could be blocked by synthetic Pro inhibitors, further supporting the HML-2 Pro involvement. In our experiments, HML-2 Pro was detected by Western blot using α-HML-2 Pro polyclonal antibody 9367 [185], that we also employed for detection of HML-2 Pro transiently expressed in bacteria and human cells. Our experiments provided direct evidence for self-processed, mature HML-2 Pro, corroborating previous reports on presence of enzymatically active HML-2 Pro in teratocarcinoma cells. Presence of enzymatically active HML-2 Pro in melanoma cell lines was not reported so far, thus such an enzyme activity was demonstrated in this study for the first time. Based on our results, HeLa cells may also express endogenous HML-2 Pro since a protein with a size of self-processed HML-2 Pro (~18 kDa), was detected at very low level also in this cell line. Such an observation was not in line with a previous Western blot experiment (not shown) that we performed to identify a cell line suitable (likely not expressing HML-2 Pro at high level) for preparation of total protein lysates for TAILS reactions (discussed above). This discrepancy probably was due to enhanced chemiluminescence (ECL) reagents employed in the two experiments. SignalFire™ ECL Reagent was employed in the first experiment

and SignalFire^{TM} Elite ECL Reagent (both from Cell Signaling Technology) was employed in subsequent experiments. The Elite reagent is particularly suited for detection of very low abundance proteins.

We did not attempt absolute quantification of endogenous HML-2 Pro detected in investigated tumor cell lines. However, we noticed that HML-2 Pro levels were much lower than levels of recombinant HML-2 Pro overexpressed in HEK293T cells. This was not surprising as HML-2 Pro expression from plasmid constructs was driven by a strong CMV promoter, very likely not comparable with physiological protein expression.

Our experimental findings indicate that endogenous HML-2 Pro is active and capable of proteolytic activity. In fact, transient expression of EGFP-Pro-mut (the precursor that harbors sites for HML-2 Pro self-processing but that is unable to self-process) in Tera-1, MeWo and SK-Mel cells lines resulted in processing of EGFP-Pro-mut, with release of Pro and EGFP subregions, as expected in presence of endogenous HML-2 Pro activity. The amount of such processing was reduced in the presence of HIV Pro inhibitor indinavir. It thus can be concluded that active HML-2 Pro present in those cells processed EGFP-Pro-mut through proteolytic activity *in trans*. Speculations about involvement of HML-2 Pro in tumor diseases benefit from our results showing presence of mature and functional HML-2 Pro in various tumor cell lines. For further studies, it would be interesting to investigate endogenous HML-2 Pro in a greater number of cell lines with known or suspected HML-2 expression. Various healthy and disease-derived tissues show transcription of HML-2 [113]. For a screening experiment examining presence of HML-2 Pro, involvement of cell lysates from normal and disease-derived human tissues could be also of interest. For an investigation of cell/tissue types potentially expressing HML-2 Pro, development of an α-HML-2 Pro monoclonal antibody could be of great benefit. The α-HML-2 Pro polyclonal antibody 9367 [185] employed in this study is currently the only one available to us for probing HML-2 Pro. Use of a monoclonal antibody would likely increase sensitivity and specificity of HML-2 Pro detection, reducing the risk of ambiguities that arise from polyclonal antibodies cross-reacting with unspecific proteins. An α-HML-2 Pro monoclonal antibody could also be valuable to investigate with higher specificity cellular localization of HML-2 Pro via immunofluorescence microscopy.

5.1.25 Proteins processed by HML-2 Pro locate to various cellular compartments and participate in diverse cellular processes

The list of human proteins that we identified by TAILS as potential HML-2 Pro substrates represents an important reference for further investigations regarding the impact of HML-2 Pro in cell biology. We employed that protein list to evaluate cellular compartments in which potential Pro substrates

localize and biological processes in which they participate. For such analysis we used Gene Ontology (GO), a system that is used to classify proteins in protein databases. GO classifies proteins via specific hierarchically-organized terms and thus annotates information associated with proteins. We subjected the 872 proteins in our list to GO term analysis via PANTHER (Protein Analysis Through Evolutionary Relationships) classification database [140]. Information on cellular localization and biological processes of potential HML-2 Pro substrates was retrieved.

Cellular proteins potentially susceptible to HML-2 Pro activity are localized in major cellular compartments including cytosol, nucleus and plasma membrane. A number of proteins also localized to exosome, secretory granules and primary lysosome. Of note, those organelles are characterized by acidic pH, thus one could suppose that if HML-2 Pro locates to such compartments, particular organelle-associated proteins could be cleaved relatively efficiently. Also, several proteins are associated with the mitochondrion, which could be relevant in view of the fact that HIV-1 Pro has been reported to localize to and cause biological effects in the mitochondrion [169]. Considering that our experimental approach for substrate identification involved a protein mixture derived from a total cell lysate, it was not surprising that proteins cleaved by HML-2 Pro derive from different cellular compartments. Since there is no published information regarding HML-2 Pro cellular localization yet, it is unknown whether HML-2 Pro could locate in the various above-mentioned compartments. In case of HML-2 Pro expression, proteins identified by us and associated with those cellular compartments could be likely subjected to HML-2 Pro proteolysis if HML-2 Pro locates in those compartments. Although, proteins located in cellular compartments where HML-2 Pro is not specifically located could be processed anyway during trafficking to those compartments. As a consequence, HML-2 Pro could potentially have a direct or indirect impact in those cellular compartments.

Besides the observation that potential HML-2 Pro substrates located to different cellular compartments, it was not unexpected that those proteins were also involved in a multitude of biological processes including response to stress, immune response, cell death, cell cycle regulation, DNA repair, DNA replication, translation, mRNA modification, protein folding. Thus, various processes that are crucial for cellular functions could be potentially affected by enzymatically active HML-2 Pro. Some biological processes that could be influenced by HML-2 Pro activity appear as particularly relevant. Approximately 200 different human proteins were associated with response to stress. Stress conditions trigger a global reprogramming of gene transcription [216], that could potentially also increase transcription of HML-2 proviruses, thus expression of HML-2 Pro. Results from our experiments allow to speculate that HML-2 Pro expression in stressed cells could affect the stress response by altering levels of stress response proteins. In the light of a list of 200 potential protein substrates it is currently difficult to evaluate if such a modulation would be negative or positive for the cell. Specific experiments investigating such aspects would be of great value in order to understand the impact of

HML-2 Pro on the stress response. As mentioned above, HML-2 Pro likely processes various heat shock proteins *in vivo*, that have crucial roles in the stress response.

Approximately 100 human proteins being potential substrates of HML-2 Pro *in vitro* were associated with organization of the cytoskeleton. This observation is in line with findings from previous studies that identified cellular substrates of other retroviral proteases. For instance, there is evidence that HIV-1 Pro cleaves cytoskeletal proteins actin, myosin, tropomyosin [189], tubulin [82], and microtubule associated proteins [213], all of which were also identified in our screening as potential substrates of HML-2 Pro. Of note, we verified processing of tubulin alpha-1A chain (TUBA1A) *in vivo*, providing further evidence that HML-2 Pro could potentially influence the cytoskeleton. Since the cytoskeleton is associated with a multitude of cellular functions including mechanical resistance, intracellular transport, cellular signaling, chromosome segregation, one can speculate that, through processing of proteins involved in the organization of the cytoskeleton, HML-2 Pro activity could influence many cellular processes.

Approximately 100 potential substrates of HML-2 Pro were associated with regulation of cell death. This finding is also in line with previous reports on HIV-1 Pro. For instance, there is documented cleavage of mitochondrial proteins by HIV-1 Pro followed by induction of molecular events leading to apoptosis [169]. Several proteins in our list were associated with apoptosis and/or mitochondrial organization, with the latter being involved in the activation of the intrinsic apoptosis pathway. Among cell death mechanisms, apoptosis could be effectively influenced by HML-2 Pro, as our specific investigation in that regard has demonstrated (see above). Among the apoptosis factors identified, cytokine induced apoptosis inhibitor 1 (CIAPIN1) was verified by us in experiments *in vivo*. CIAPIN1 is involved in negative control of cell death via an anti-apoptotic effect. One can speculate that processing of this anti-apoptotic factor by HML-2 Pro may reduce its cellular levels, thus potentially promoting activation of apoptosis. It is known that both reduced and increased apoptosis can have pathological relevance [55]. Various types of cancer, neurological disorders, infectious diseases, autoimmune diseases are among the diseases influenced by altered apoptotic mechanisms. Impairment of the apoptotic pathway due to HML-2 Pro activity may have a role in those diseases.

5.1.26 Overlap of potential substrates of HML-2 Pro with cancer-relevant genes/proteins

HML-2 has been implicated in tumor development in previous studies, and its upregulation has been associated with several tumor types [6,15,76,171]. HML-2 proteins Env, Rec and Np9 have been investigated regarding their tumorigenic properties [19,33,119]. The potential role of HML-2 Pro in cancer remains unexplored, although endogenous HML-2 Pro activity has been observed in various tumor tissues and derived cell lines [112,179]. We therefore performed more specific analyses to

evaluate the involvement of HML-2 Pro substrate proteins in tumor diseases. The genes corresponding to the 872 human proteins identified as potential substrates of HML-2 Pro were intersected with genes included in the Catalogue Of Somatic Mutations In Cancer (COSMIC) database, an online database that contains up-to-date information on known cancer genes and associated tumor types. Our analysis identified 64 human genes/proteins (in the following referred to as proteins for the sake of simplicity) with a known relevance in oncology. One can speculate that altered amount or altered functionality of those proteins, due to processing by HML-2 Pro, could represent a mechanism contributing to tumor development. Based on information retrieved from the COSMIC database, the 64 identified proteins were associated with a total of 59 tumor types. Many proteins were associated with more than just one tumor type (see Table 19), thus having stronger evidence of an involvement in tumor diseases. Those proteins could deserve particular attention in the context of HML-2 Pro as a potential promoter of tumor development. Several of the tumor types identified in our analysis were associated with HML-2 previously, i.e, melanoma [29,183], ovarian cancer [215], prostate tumor [113], breast cancer [214], brain cancer [58], and lymphoma [42]. Endometrial carcinoma, also included in our results, was not directly associated with HML-2 before, although HERV-encoded proteins derived from other HERV groups were found in such tumor type [171]. The HML-2 Pro substrates overlapping with the COSMIC collection could represent a key to elucidate molecular mechanisms behind the association of HML-2 upregulation and development of abovementioned tumor types. For instance, one can speculate that expression of HML-2 Pro in melanocytes could impair B2M, ERCC5 and MAP2K2 proteins, three potential substrates of HML-2 Pro that are associated with melanoma, thus potentially triggering molecular mechanisms leading to melanoma. More specific investigations will be necessary and of great interest in order to reveal such unknown aspects behind tumor diseases.

Our analysis revealed many tumor types that, to our knowledge, have not been associated with HML-2 so far, like colorectal cancer, papillary thyroid cancer, glioblastoma, epithelial lung cancer. In view of the notion that HML-2 Pro could potentially negatively affect proteins involved in those diseases, those tumor types could be investigated for presence of HML-2 transcripts and HML-2 Pro. Such investigations could eventually reveal unknown associations between HML-2 and specific tumor types.

5.1.27 Overlap of potential substrates of HML-2 Pro with OMIM database

We also intersected the list of 872 potential substrates of HML-2 Pro with genes and phenotypes compiled in the Online Mendelian Inheritance in Man (OMIM) database. A total of 289 diseases was identified in which HML-2 Pro could potentially have a role. Diseases of particular interest in the context of HML-2 included, apart from several types of tumors, various neurological diseases. Notably, five proteins (SOD1, VCP, HNRNPA1, TUBA4A, SQSTM1) have been associated with ALS, a disease that has been associated with HML-2 before [121]. Though, the role of HML-2 in the

pathophysiology of ALS is not clear at the moment, HML-2 Pro and ALS may be pathologically connected by an activity of HML-2 Pro against human proteins having a potential or factual role in ALS.

Overall, above described disease associations currently remain speculative, but provide rationale for further investigations concerning the role of HML-2 Pro in cancer and other diseases.

5.1.28 Conclusions for HML-2 Pro

The approach used in our investigation successfully combined TAILS with profiling of potential substrates of HML-2 Pro and experimental verification of candidate proteins *in vitro* and *in vivo*. As our results have shown, such an experimental approach represents a valuable strategy to get insight into potential substrates of under-investigated proteases. Our study revealed that hundreds of human proteins are potential substrates of HML-2 Pro. The vast list of HML-2 Pro substrates that we obtained is a precious source of information for further research in the context of retrovirus-encoded proteases. We provided experimental evidence to believe that, if expressed, HML-2 Pro could potentially affect levels of proteins involved in important cellular processes and disease-relevance. We obtained preliminary evidence for expression of functional HML-2 Pro in various tumoral cell lines, although the physiological consequences of an active HML-2 Pro within cells need to be elucidated. We initially characterized cell death following transient expression of HML-2 Pro and observed that apoptotic processes may be activated in the presence of HML-2 Pro activity. Our observations call for further studies that could elucidate the impact of HML-2 Pro expression on cell biology. Experiments investigating the potential involvement of HML-2 Pro in human diseases should be addressed specifically.

5.2 HML-2 integrase

5.2.1 Investigations of HML-2 IN as a potential genotoxic agent

The second part of our studies on biological relevance of enzymatically active HML-2 proteins focused on HML-2 IN. Kitamura et al. (1996) [105] demonstrated that IN of HERV-K10 (an HML-2 provirus) is active and capable of typical catalytic activities of retroviral INs (see Introduction). That finding raises the question of cellular consequences of an active retroviral IN potentially expressed in human cells. As an enzyme that specifically interacts and cleaves DNA molecules, HML-2 IN could contribute to genome instability. Bray et al. (2016) [25] remarked that several HML-2 loci are able to encode an intact IN. Transcripts derived from various HML-2 loci having an intact IN ORF were demonstrated in tumor tissues and derived cell lines [58,183]. Such evidence lets one hypothesize that an active HML-2 IN could be present in human cells and cause DNA damage, thus contribute to human disease.

Retroviral INs are known to induce DNA lesions during the process of provirus formation, during which a double-strand break (DSB) is created in the host cell DNA, leaving single-strand DNA (ssDNA) gaps that flank the site of viral genome integration [172]. DNA damage could be induced also via the so-called nonspecific c [100], an activity that has been shown for retroviral INs from different genera [151,193]. Nonspecific alcoholysis by a retroviral IN can potentially nick a DNA sequence in any position, independently of the presence of viral LTRs. That activity of transiently expressed HIV-1 IN may cause mutagenic effects leading to a lethal phenotype in yeast cells [151]. Besides typical retroviral activities, HML-2 IN may exert nonspecific alcoholysis as well, although this has not been investigated so far.

As a consequence of genotoxic stress potentially induced by HML-2 IN, DNA double-strand breaks (DSBs) might occur in the cellular genome [192]. In particular, at the level of DNA single strand damage, nuclease activity or spontaneous hydrolysis could induce generation of DSBs [193]. DSBs are particularly severe DNA lesions because DSB repair by the host machinery is intrinsically more complex than that of other types of DNA damage [103]. Induction of DSBs by HML-2 IN could have an impact on the cell biology considering that DSBs stimulate inflammatory immune response, trigger programmed cell death and cause deleterious mutations increasing the risk of tumorigenesis [103,210]. Until now, there are no published studies having investigated the ability of HML-2 IN, or any other ERV-derived IN, to induce DNA lesions in the host genome. In this study, we attempted an experiment to initially investigate genotoxicity of HML-2 IN *in vivo*. The strategy that we adopted counted DSBs induced in cells transiently expressing wild-type HML-2 IN (IN-wt) and compared numbers with those in cells expressing an inactive HML-2 IN (IN-mut) due to a mutated catalytic site. Untransfected cells were also examined as another control. All the cells included in the analysis were in the same phase of the cell cycle (see below). DSBs were examined by probing tumor suppressor p53-binding protein 1 (53BP1) and by immunofluorescence microscopy. 53BP1 is a key player in the DSB repair pathway

[150] that is assembled at sites of DNA damage as a scaffold for recruitment of other DSB-responsive proteins [150]. 53BP1 is a commonly used marker for DNA damage that, when probed with specific antibodies, can be visualized under the fluorescence microscope as discrete foci that co-localize with DSBs in the host genome [132]. An abnormal number of 53BP1 foci in presence of a genotoxic agent is indicative of induction of DSBs by that particular agent. DNA damage induced by different types of genotoxic stress, for instance, ionizing and UV radiation, camptothecin, etoposide and alkylating agent methylmethanesulfonate, have been shown to trigger recruitment of 53BP1 to discrete nuclear foci [159].

In this work, we examined HML-2 IN activity as a potential genotoxic agent. The number of 53BP1 foci counted in IN-wt were not significantly different from numbers of 53BP1 foci in control cells (IN-mut and untransfected). We observed a relatively large standard deviation for foci per cell (see Figure 45 D), with approximately 3 foci per cell for each group (70 cells per group). A previous study [132] investigating the dose-response of 53BP1 foci to IR reported that control unirradiated HeLa cells contained one or two 53BP1 foci when analyzed 24 h post-irradiation. In view of that report, presence of some 53BP1 foci in untreated HeLa cells could be expected in our experiments. With regard to 53BP1 foci number, we observed a seemingly slightly higher standard deviation for IN-wt than controls. It remains to be seen whether an effect of HML-2 IN activity on DSBs formation could be potentially revealed by analysis of a larger number of cells than in our initial experiments. Rappold et al. (2001) [159] documented a variation in size of 53BP1 foci depending on the genotoxic agent applied. We also examined sizes of 53BP1 foci in treated cells, as sizes of foci may have been affected by HML-2 IN activity. We did not notice significant differences in size of 53BP1 foci when comparing IN-wt with controls. Most of the 53BP1 foci showed a very small size in all the cells examined, although sporadically larger foci were observed. Taken together, our experiments did not provide evidence to conclude that HML-2 IN can damage DNA in a way causing formation of DSBs.

5.2.2 Caveats and pitfalls of our initial experiments assaying HML-2 IN

Results from our initial experiments assaying HML-2 IN should be interpreted with caution. Several factors might have influenced the outcome of our analysis. Some aspects limiting our analysis and possible improvements for further studies shall be discussed in the following.

In cells treated with IR, the dose of IR and the time until fixation of cells represent important factors affecting numbers of 53BP1 foci [132]. When considering a transiently expressed enzyme as a potential genotoxic agent, one could suppose that 1) the amount of HML-2 IN expressed and 2) the time of exposure to HML-2 IN, thus the interval between cell transfection (with a plasmid expressing IN) and cell fixation may represent relevant factors in the 53BP1 assay. We fixed cells 24 hpt, a time point when transiently expressed HML-2 IN was present at considerably high level in the cytoplasm

and nucleus of cells. We cannot exclude that DSBs induced by the HML-2 IN during the transient expression period may have been largely repaired by 24 h post transfection. Thus, DSBs may have been detectable more readily at earlier time points. On the other side, we cannot exclude that longer exposure to IN activity, for instance, fixing cells only at 48 hpt, could have revealed higher numbers of DSBs due to accumulation of DNA damage induced by HML-2 IN. Monitoring numbers of 53BP1 foci at different time periods of transient expression of HML- IN may be recommended. Use of an inducible expression system, for instance, a system based on a tetracycline-controlled operator [93], could be of great benefit to synchronize expression of HML-2 IN in cells.

Molecular processes behind 53BP1 signaling involve factors that may behave differently at different stages of the cell cycle [56]. As a consequence, 53BP1 focus formation might be influenced by the phase of the cell cycle. In view of this, we aimed in our 53BP1 foci analysis at reducing biases from cell cycle variations. To do so, we counted solely those cells that, based on Ki-67 staining, appeared to be in the same phase of the cell cycle when fixed. Ki-67 is a widely used marker of cell proliferation, with the expression level within the cell correlating with the cell-cycle phase. Ki-67 is known to be degraded continuously in G0 and G1 phase [195]. For our analysis we considered only cells with a Ki-67 level below the detection limit (Ki-67 negative cells), likely representing cells in G0 or G1 phase. Among cells with noticeable Ki-67 staining (Ki-67 positive cells), the level and distribution of Ki-67 was rather variable and not easily interpretable with regard to the actual cell cycle phase.

An analysis focused on cells in other phases of the cell cycle when fixed might be worthwhile. Considering that DNA single-strand breaks (SSBs) can collapse into DSBs if met by a replication fork during DNA replication [172] one might expect during the S-phase of the cell cycle more IN-induced SSBs eventually converted into DSBs. Thus, further investigations evaluating DSBs in presence of HML-2 IN activity could be meaningful when focusing attention on cells fixed during the S-phase. On the basis of our observations (see above), Ki-67 is not well suited for unambiguous discrimination of such cells. As an alternative, the proliferating cell nuclear antigen (PCNA) could be employed since cells not in S-phase (G0 included) stain only weakly with α-PCNA antibodies [32].

Our experiments employed asynchronous cell populations. This was not favorable with regard to selection of certain cell cycle phases. In fact, a large number of images of cells had to be taken and analyzed when evaluating 53BP1 foci. A relatively low number of HML-2 IN (or its mutant) positive cells were also Ki-67 negative. Identification of cells meeting our selection criteria (HML-2 IN positive and Ki-67 negative) was laborious and time consuming. Therefore, the total number of cells analyzed was limited to a total of approximately 200. For further experiments, use of a system to synchronize cells or arrest them in a certain phase of the cell cycle could be of great benefit to increase the number of positive cells examined, thus increase consistency of the analysis. For instance, synchronization of cells could be achieved by chemical blockade. Various inhibitors exist that block cells in specific

phases of the cell cycle [9]. Before transfection, cells could be arrested in a certain cell cycle phase with one or more (through parallel experiments) drugs [155]. Synchronized cells could be fixed at different time points after transfection and subjected to immunostaining for DNA damage analysis.

Our investigation required acquisition and analysis of hundreds of images. For those images, counting of 53BP1 foci was done by eye. Although certainly accurate, evaluation by eye is time consuming and may bias results. Software for analyzing foci have been reported (for instance, see [132]). Use of computerized *in silico* systems for foci analysis could save time and allow analysis of much larger numbers of cells. The latter is expected to produce more accurate mean values, thus (better) results that could potentially change our overall conclusions on DNA damage induced by HML-2 IN. In future studies, a combination of the established immunostaining approach and computerized foci counting systems could be of great benefit.

Repair kinetics of DSBs, thus of 53BP1 formation, may cause differences depending on the cell line utilized [98]. In our experiments, we utilized HeLa cells because they adhered firmly to the surface of tissue culture plates used for immunofluorescence assays. However, it is currently unknown whether other cell lines may be more suited for monitoring DNA damage induced by HML-2 IN activity. Use of alternative cell lines should thus be considered for further investigations of DNA damage induced by HML-2 IN.

Apart from 53BP1, other proteins involved in the repair of DSBs, that are part of foci observable by immunofluorescence, could be used as DSB markers [217], e.g., histone H2A variant (H2AX), which is specifically phosphorylated in response to DSB induction [217]. Phosphorylated H2AX, also known as gamma (γ)-H2AX, has been initially considered as DSB marker for our experiments since it showed to be highly sensitive and well characterized in previous studies [217], and it was already shown to locate at sites of retroviral integration during viral infection [150]. Moreover, this γ-H2AX has been used in a preliminary investigation conducted in our lab in the course of a bachelor thesis (Brill 2011), with some experiments having indicated a higher number of γ-H2AX foci in cells expressing transiently expressed HML-2 IN compared to cells expressing controls or mock-transfected cells. However, in that preliminary work, no quantification by counting of γ-H2AX foci was attempted. In our experiments, when establishing conditions for γ-H2AX assays, γ-H2AX foci could not be quantified reliably due to presence of too small (uncountable) foci and high background staining (not shown). Those issues might have been due to cross-reacting antibodies. Therefore, we focused on quantification of 53BP1 foci, because 53BP1 is known to co-localize with γ-H2AX at DSB sites [217] and also because 53BP1 foci could be visualized well in our experiments.

Since the phosphorylation of H2AX (at residue 139) is a highly specific event that takes place (and is amplified involving many H2AX molecules) in response to DNA damage [166] the levels of γ-H2AX

within the cell provide information about the amount of DSBs [166]. Although microscopy-based foci quantitation is among the most sensitive assays to measure DSB levels [89], other less laborious immunological methods could be attempted to study the genotoxicity of HML-2 IN via γ-H2AX. In particular, following transient expression of wild-type and mutant HML-2 IN in cell culture, a comparative analysis of γ-H2AX levels between two cell samples could be done by flow cytometry, which is a rapid and technically relatively simple method for an automated high throughput analysis [24]. Alternatively, γ-H2AX levels could be measured in cell lysates by Western blot or enzyme-linked immunosorbent assay (ELISA) [89]. Other proteins participating in DSB repair that are phosphorylated in response to DSBs, e.g., serine/threonine kinases ATM and Chk2, as well as 53BP1 [217], could be included in a comprehensive study monitoring those phosphorylated proteins during HML-2 IN expression employing above-mentioned immunological technologies.

Considering that DNA damage induced by retroviral IN consists of SSBs that, if not repaired, can transform into DSBs only secondarily [192], one might envisage evaluation of formation of SSBs as an alternative approach when investigating an ability of HML-2 IN to induce DNA damage. We focused in our investigations on the evaluation of DSBs since their detection by immunofluorescence assays using markers such as 53BP1 has been successfully applied and is well characterized [217]. Moreover, preliminary work by Brill (2011) mentioned above had provided indications in favor of HML-2 IN causing DSBs. In any case, evaluation of SSBs should be considered when further investigating genotoxicity of HML-2 IN. For such an investigation, X-ray repair cross complementing 1 protein (XRCC1) could represent a good candidate marker for monitoring SSBs. XRCC1 is required for DNA SSB repair in human cells [26] and has already been employed for detection of unrepaired SSBs via immunofluorescence assays [145].

HML-2 IN activity has been previously confirmed *in vitro*, using oligonucleotides that simulate the LTRs of the viral dsDNA genome [105]. However, experiments assaying HML-2 IN activity (or activity of IN from other ERVs) in a cellular context have not been attempted so far, to the best of our knowledge. Our *in vivo* experimental conditions did not investigate actual enzymatic activity of HML-2 IN transiently expressed in cell culture. Therefore, whether HML-2 IN is enzymatically active *in vivo* remains an unresolved question that requires further investigations employing specific experimental strategies that were not the focus of our initial studies.

Another aspect of our experimental set-up raised another question related to HML-2 IN activity. HML-2 IN was expressed in our experiments with an HA-tag in its N-terminus. One might wonder whether presence of an HA-tag may have hindered enzymatic activity. In support of the HA tag not impacting IN activity, it was reported before [153] that infectivity of HIV-1 harboring an IN fused to a C-terminal HA-tag was not affected, suggesting that a C-terminal HA-tag does not inhibit normal activities of

HIV-1 IN. We had employed an HA-tag because we had a well-functioning α-HA monoclonal antibody available in our laboratory and we also had well-established conditions for HA detection by immunofluorescence. Although activity of HA-tagged HML-2 IN was not assayed by us, it was previously demonstrated that a His-tag in the N-terminus of HML-2 IN does not hinder its activity [105]. One can thus assume that the 9 aa HA-tag, like a 6xHis tag, does not dramatically affect IN activity. Certainly, that assumption requires confirmation by specific experiments that, for example, could employ established assays *in vitro* [105], assaying HML-2 IN with at least an N-terminal HA-tag.

On the basis of HML-2 proviral organization, HML-2 IN is expected to be expressed as part of a Gag-Pro-Pol polyprotein and then released by HML-2 Pro activity. Retroviral proteases are often activated after virion budding [200]. One may wonder whether HML-2 IN can be released from its precursor protein within the cell, outside or inside retroviral particles. We have demonstrated for HML-2 Pro that HML-2 Pro can be activated independent from assembly of viral particles (see above). HML-2 Pro was thereby able to process its own viral and other cellular substrates in the cytoplasm. Therefore, there is evidence to believe that in case of HML-2 expression, HML-2 IN could be released from its precursor by active HML-2 Pro. Such a release could take place in the cytoplasm of cells, from which HML-2 IN could then translocate into the cell nucleus. A simple experiment to document the proper release of HML-2 IN from its precursor could involve transient expression of an HML-2 Pro-Pol protein with an engineered tag (for instance, HA) at the 3' end of the Pol. Detection of HA by Western blot would then show a band of a size expected for the HML-2 IN domain that was generated by HML-2 Pro activity.

In our analysis we observed HML-2 IN localizing in the cytoplasm and in the nucleus of cells, that was not reported before. HML-2 IN localizing in the cell nucleus indicated that this enzyme could have karyophilic properties already known for HIV-1 IN [155]. Such an observation allows for speculations about interaction of HML-2 IN with the host genomic DNA. Therefore, it can be considered a piece of evidence that supports (independent from our results obtained through 53BP1 foci analysis) the hypothesis that HML-2 IN could potentially influence host genome stability.

5.2.3 Conclusions for HML-2 IN

Our investigation of HML-2 IN addressed the potential ability of the enzyme to induce DNA damage in the cellular genome. Experiments performed can be considered pioneering as the biological impact of HML-2 IN in cells has not been investigated experimentally so far according to the published literature. Our analysis involved stringent criteria for evaluation of DSBs induced in presence of HML-2 IN. However, several factors not considered in this first investigation may have limited our analysis. Although we did not obtain evidence that allowed us to conclude that HML-2 IN activity could contribute to induction of DSBs in the host genome we have collected initial information that will be useful for further studies exploring biological roles of HML-2 IN. This first investigation provided an example of an experimental approach that can be further improved in future analysis to elucidate potential impact of HML-2 IN on the host genome.

6. References

1. Álvarez E, Menéndez-Arias L, Carrasco L (2003) The Eukaryotic Translation Initiation Factor 4GI Is Cleaved by Different Retroviral Proteases. J Virol 77:12392–12400

2. Amberger JS, Bocchini CA, Scott AF, Hamosh A (2018) OMIM.org: leveraging knowledge across phenotype–gene relationships. Nucleic Acids Res 47:1038–1043

3. Andersson A-C, Venables PJW, Tonjes RR, Scherer J, Eriksson L, Larsson E (2002) Developmental expression of HERV-R (ERV3) and HERV-K in human tissue. Virology 297:220–225

4. Andersson ML, Westley B, Medstrand P, Blomberg J, Lindeskog M, May F (1999) Diversity of human endogenous retrovirus class II-like sequences. J Gen Virol 80:255–260

5. Ariza M-E, Williams M V (2011) A human endogenous retrovirus K dUTPase triggers a TH1, TH17 cytokine response: does it have a role in psoriasis? J Invest Dermatol 131:2419–2427

6. Armbruester V, Sauter M, Krautkraemer E, Meese E, Kleiman A, Best B, Roemer K, Mueller-Lantzsch N (2002) A novel gene from the human endogenous retrovirus K expressed in transformed cells. Clin Cancer Res 8:1800–1807

7. Armbruester V, Sauter M, Roemer K, Best B, Hahn S, Nty A, Schmid A, Philipp S, Mueller A, Mueller-Lantzsch N (2004) Np9 protein of human endogenous retrovirus K interacts with ligand of numb protein X. J Virol 78:10310–10319

8. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25:25–29

9. Banfalvi G (2011) Overview of cell synchronization. Methods Mol Biol 761:1–23

10. Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A, Tristem M (2004) Long-term reinfection of the human genome by endogenous retroviruses. Proc Natl Acad Sci U S A 101:4894–4899

11. Belshaw R, Dawson ALA, Woolven-Allen J, Redding J, Burt A, Tristem M (2005) Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): implications for present-day activity. J Virol 79:12507–12514

12. Bertrand R, Solary E, O'Connor P, Kohn KW, Pommier Y (1994) Induction of a common pathway of apoptosis by staurosporine. Exp Cell Res 211:314–321

13. Beyer U, Moll-Rocek J, Moll UM, Dobbelstein M (2011) Endogenous retrovirus drives hitherto unknown proapoptotic p63 isoforms in the male germ line of humans and great apes. Proc Natl Acad Sci U S A 108:3624–3629

14. Bhagwat SR, Hajela K, Kumar A (2018) Proteolysis to Identify Protease Substrates: Cleave to Decipher. Proteomics 18:1–17

15. Bhardwaj N, Montesion M, Roy F, Coffin JM (2015) Differential expression of HERV-K (HML-2) proviruses in cells and virions of the teratocarcinoma cell line Tera-1. Viruses 7:939–968

16. Bieda K, Hoffmann A, Boller K (2001) Phenotypic heterogeneity of human endogenous retrovirus particles produced by teratocarcinoma cell lines. J Gen Virol 82:591–596

17. Biniossek ML, Niemer M, Maksimchuk K, Mayer B, Fuchs J, Huesgen PF, McCafferty DG, Turk B, Fritz G, Mayer J, Haecker G, Mach L, Schilling O (2016) Identification of Protease Specificity by Combining Proteome-Derived Peptide Libraries and Quantitative Proteomics. Mol Cell Proteomics 15:2515–2524

18. Blanco R, Carrasco L, Ventoso I (2003) Cell killing by HIV-1 protease. J Biol Chem 278:1086–1093

19. Boese A, Sauter M, Galli U, Best B, Herbst H, Mayer J, Kremmer E, Roemer K, Mueller-Lantzsch N (2000) Human endogenous retrovirus protein cORF supports cell transformation and associates with the promyelocytic leukemia zinc finger protein. Oncogene 19:4328–4336

20. Boller K, König H, Sauter M, Mueller-Lantzsch N, Löwer R, Löwer J, Kurth R (1993) Evidence that HERV-K is the endogenous retrovirus sequence that codes for the human teratocarcinoma-derived retrovirus HTDV. Virology 196:349–53

21. Bonnaud B, Bouton O, Oriol G, Cheynet V, Duret L, Mallet F (2004) Evidence of selection on the domesticated ERVWE1 env retroviral element involved in placentation. Mol Biol Evol 21:1895–1901

22. Bonner WA, Hulett HR, Sweet RG, Herzenberg LA (1972) Fluorescence Activated Cell Sorting. Rev Sci Instrum 43:404–409

23. Boulares AH, Yakovlev AG, Ivanova V, Stoica BA, Wang G, Iyer S, Smulson M (1999) Role of poly(ADP-ribose) polymerase (PARP) cleavage in apoptosis. Caspase 3-resistant PARP mutant increases rates of apoptosis in transfected cells. J Biol Chem 274:22932–22940

24. Bourton EC, Plowman PN, Zahir SA, Senguloglu GU, Serrai H, Bottley G, Parris CN (2012) Multispectral imaging flow cytometry reveals distinct frequencies of gamma-H2AX foci induction in DNA double strand break repair defective human cell lines. Cytometry A 81:130–137

25. Bray S, Turnbull M, Hebert S, Douville RN (2016) Insight into the ERVK Integrase – Propensity for DNA Damage. Front Microbiol 7:1941

26. Brem R, Hall J (2005) XRCC1 is required for DNA single-strand break repair in human cells. Nucleic Acids Res 33:2512–2520

27. Broz P, von Moltke J, Jones JW, Vance RE, Monack DM (2010) Differential requirement for Caspase-1 autoproteolysis in pathogen-induced cell death and cytokine processing. Cell Host Microbe 8:471–483

28. Büscher K, Trefzer U, Hofmann M, Sterry W, Kurth R, Denner J (2005) Expression of Human Endogenous Retrovirus K in Melanomas and Melanoma Cell Lines. Cancer Res 65:4172–4180

29. Büscher K, Hahn S, Hofmann M, Trefzer U, Özel M, Sterry W, Löwer J, Löwer R, Kurth R, Denner J (2006) Expression of the human endogenous retrovirus-K transmembrane envelope, Rec and Np9 proteins in melanomas and melanoma cell lines. Melanoma Res 16:223–234

30. Buzdin A, Ustyugova S, Khodosevich K, Mamedov I, Lebedev Y, Hunsmann G, Sverdlov E (2003) Human-specific subfamilies of HERV-K (HML-2) long terminal repeats: three master genes were active simultaneously during branching of hominoid lineages. Genomics 81:149–156

31. Casey JR, Grinstein S, Orlowski J (2010) Sensors and regulators of intracellular pH. Nat Rev Mol Cell Biol 11:50–61

32. Celis JE, Madsen P, Nielsen S, Celis A (1986) Nuclear patterns of cyclin (PCNA) antigen distribution subdivide S-phase in cultured cells — Some applications of PCNA antibodies. Leuk Res 10:237–249

33. Chen T, Meng Z, Gan Y, Wang X, Xu F, Gu Y, Xu X, Tang J, Zhou H, Zhang X, Gan X, Van Ness C, Xu G, Huang L, Zhang X, Fang Y, Wu J, Zheng S, Jin J, Huang W, Xu R (2013) The viral oncogene Np9 acts as a critical molecular switch for co-activating β-catenin, ERK, Akt and Notch1 and promoting the growth of human leukemia stem/progenitor cells. Leukemia 27:1469–1478

34. Chow SA, Vincent KA, Ellison V, Brown PO (1992) Reversal of integration and DNA splicing mediated by integrase of human immunodeficiency virus. Science 255:723–726

35. Christensen T (2016) Human endogenous retroviruses in neurologic disease. APMIS 124:116–126

36. Coffey MJ, Dagley LF, Seizova S, Kapp EA, Infusini G, Roos DS, Boddey JA, Webb AI, Tonkin CJ (2018) Aspartyl Protease 5 Matures Dense Granule Proteins That Reside at the Host-Parasite Interface in Toxoplasma gondii. MBio 9:e01796-18

37. Cohen CJ, Lock WM, Mager DL (2009) Endogenous retroviral LTRs as promoters for human genes: a critical assessment. Gene 448:105–114

38. Cohen M, Larsson E (1988) Human endogenous retroviruses. Bioessays 9:191–196

39. Consortium TGO (2007) The Gene Ontology project in 2008. Nucleic Acids Res 36:D440–D444

40. Consortium TU (2018) UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 47:D506–D515

41. Contreras-Galindo R, Kaplan MH, Markovitz DM, Lorenzo E, Yamamura Y (2006) Detection of HERV-K(HML-2) Viral RNA in Plasma of HIV Type 1-Infected Individuals. AIDS Res Hum Retroviruses 22:979–984

42. Contreras-Galindo R, Kaplan MH, Leissner P, Verjat T, Ferlenghi I, Bagnoli F, Giusti F, Dosik MH, Hayes DF, Gitlin SD, Markovitz DM (2008) Human endogenous retrovirus K (HML-2) elements in the plasma of people with lymphoma and breast cancer. J Virol 82:9329–9336

43. Daber R, Stayrook S, Rosenberg A, Lewis M (2007) Structural analysis of lac repressor bound to allosteric effectors. J Mol Biol 370:609–619

44. Daniel R, Greger JG, Katz RA, Taganov KD, Wu X, Kappes JC, Skalka AM (2004) Evidence that stable retroviral transduction and cell survival following DNA integration depend on components of the nonhomologous end joining repair pathway. J Virol 78:8573–8581

45. Dauber DS, Ziermann R, Parkin N, Maly DJ, Mahrus S, Harris JL, Ellman JA, Petropoulos C, Craik CS (2002) Altered Substrate Specificity of Drug-Resistant Human Immunodeficiency Virus Type 1 Protease. J Virol 76:1359–1368

46. Delelis O, Carayon K, Saïb A, Deprez E, Mouscadet J-F (2008) Integrase and integration: biochemical activities of HIV-1 integrase. Retrovirology 5:114

47. Demidyuk I V, Shubin A V, Gasanov E V, Kostrov S V (2010) Propeptides as modulators of functional activity of proteases. Biomol Concepts 1:305–322

48. Denne M, Sauter M, Armbruester V, Licht JD, Roemer K, Mueller-Lantzsch N (2007) Physical and functional interactions of human endogenous retrovirus proteins Np9 and rec with the promyelocytic leukemia zinc finger protein. J Virol 81:5607–5616

49. Devroe E, Silver PA, Engelman A (2005) HIV-1 incorporates and proteolytically processes human NDR1 and NDR2 serine-threonine kinases. Virology 331:181–189

50. Dewannieux M, Harper F, Richaud A, Letzelter C, Ribet D, Pierron G, Heidmann T (2006) Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. Genome Res 16:1548–1556

51. Douville R, Liu J, Rothstein J, Nath A (2011) Identification of active loci of a human endogenous retrovirus in neurons of patients with amyotrophic lateral sclerosis. Ann Neurol 69:141–151

52. Dunn BM, Goodenow MM, Gustchina A, Wlodawer A (2002) Retroviral proteases. Genome Biol 3:REVIEWS3006

53. Engelman AN, Cherepanov P (2017) Retroviral intasomes arising. Curr Opin Struct Biol 47:23–29

54. Escalera-Zamudio M, Greenwood AD (2016) On the classification and evolution of endogenous retrovirus: human endogenous retroviruses may not be "human" after all. APMIS 124:44–51

55. Favaloro B, Allocati N, Graziano V, Di Ilio C, De Laurenzi V (2012) Role of apoptosis in disease. Aging (Albany NY) 4:330–349

56. Feng L, Li N, Li Y, Wang J, Gao M, Wang W, Chen J (2015) Cell cycle-dependent inhibition of 53BP1 signaling by BRCA1. Cell Discov 1:15019

57. Ferreira CMH, Pinto ISS, Soares E V, Soares HMVM (2015) (Un)suitability of the use of pH buffers in biological, biochemical and environmental studies and their interaction with metal ions – a review. RSC Adv 5:30989–31003

58. Flockerzi A, Ruggieri A, Frank O, Sauter M, Maldener E, Kopper B, Wullich B, Seifarth W, Müller-Lantzsch N, Leib-Mösch C, Meese E, Mayer J (2008) Expression patterns of transcribed human endogenous retrovirus HERV-K(HML-2) loci in human tissues and the need for a HERV Transcriptome Project. BMC Genomics 9:354

59. Fogh J (ed), Trempe G (1975) Human Tumor Cells In Vitro. Plenum Press, New York

60. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, Cole CG, Ward S, Dawson E, Ponting L, Stefancsik R, Harsha B, Kok CY, Jia M, Jubb H, Sondka Z, Thompson S, De T, Campbell PJ (2016) COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res 45:D777–D783

61. Frank O, Giehl M, Zheng C, Hehlmann R, Leib-Mosch C, Seifarth W (2005) Human endogenous retrovirus expression profiles in samples from brains of patients with schizophrenia and bipolar disorders. J Virol 79:10890–10901

62. Freimanis G, Hooley P, Ejtehadi HD, Ali HA, Veitch A, Rylance PB, Alawi A, Axford J, Nevill A, Murray PG, Nelson PN (2010) A role for human endogenous retrovirus-K (HML-2) in rheumatoid arthritis: investigating mechanisms of pathogenesis. Clin Exp Immunol 160:340–347

63. Fukuchi K-I, Hearn MG, Deeb SS, Smith AC, Dang N, Miyazaki J-I, Bothwell M, Martin GM (1994) Activity assays of nine heterogeneous promoters in neural and other cultured cells. Vitr Cell Dev Biol - Anim 30:300–305

64. Garcia-Montojo M, Doucet-O'Hare T, Henderson L, Nath A (2018) Human endogenous retrovirus-K (HML-2): a comprehensive review. Crit Rev Microbiol 44:715–738

65. Garson JA, Usher L, Al-Chalabi A, Huggett J, Day EF, McCormick AL (2019) Quantitative analysis of human endogenous retrovirus-K transcripts in postmortem premotor cortex fails to confirm elevated expression of HERV-K RNA in amyotrophic lateral sclerosis. Acta Neuropathol Commun 7:45

66. George M, Schwecke T, Beimforde N, Hohn O, Chudak C, Zimmermann A, Kurth R, Naumann D, Bannert N (2011) Identification of the protease cleavage sites in a reconstituted Gag polyprotein of an HERV-K(HML-2) element. Retrovirology 8:30

67. Gérard A, Soler N, Ségéral E, Belshan M, Emiliani S (2013) Identification of low molecular weight nuclear complexes containing integrase during the early stages of HIV-1 infection. Retrovirology 10:13

68. Gevaert K, Goethals M, Martens L, Van Damme J, Staes A, Thomas GR, Vandekerckhove J (2003) Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. Nat Biotechnol 21:566–569

69. Gholami AM, Hahne H, Wu Z, Auer FJ, Meng C, Wilhelm M, Kuster B (2013) Global proteome analysis of the NCI-60 cell line panel. Cell Rep 4:609–620

70. Gifford R, Kabat P, Martin J, Lynch C, Tristem M (2005) Evolution and distribution of class II-related endogenous retroviruses. J Virol 79:6478–6486

71. Gifford RJ, Blomberg J, Coffin JM, Fan H, Heidmann T, Mayer J, Stoye J, Tristem M, Johnson WE (2018) Nomenclature for endogenous retrovirus (ERV) loci. Retrovirology 15:59

72. Goering W, Schmitt K, Dostert M, Schaal H, Deenen R, Mayer J, Schulz WA (2015) Human endogenous retrovirus HERV-K(HML-2) activity in prostate cancer is dominated by a few loci. Prostate 75:1958–1971

73. Gordon MH, Anowai A, Young D, Das N, Campden RI, Sekhon H, Myers Z, Mainoli B, Chopra S, Thuy-Boun PS, Kizhakkedathu J, Bindra G, Jijon HB, Heitman S, Yates R, Wolan DW, Edgington-Mitchell LE, MacNaughton WK, Dufour A (2019) N-Terminomics/TAILS Profiling of Proteases and Their Substrates in Ulcerative Colitis. ACS Chem Biol 14:2471–2483

74. Gregoire S, Irwin J, Kwon I (2012) Techniques for Monitoring Protein Misfolding and Aggregation in Vitro and in Living Cells. Korean J Chem Eng 29:693–702

75. Griffiths DJ (2001) Endogenous retroviruses in the human genome sequence. Genome Biol 2:REVIEWS1017.1-REVIEWS1017.5

76. Herbst H, Sauter M, Mueller-Lantzsch N (1996) Expression of human endogenous retrovirus K elements in germ cell and trophoblastic tumors. Am J Pathol 149:1727–1735

77. Heyne K, Kolsch K, Bruand M, Kremmer E, Grasser FA, Mayer J, Roemer K (2015) Np9, a cellular protein of retroviral ancestry restricted to human, chimpanzee and gorilla, binds and regulates ubiquitin ligase MDM2. Cell Cycle 14:2619–2633

78. Hohn O, Hanke K, Bannert N (2013) HERV-K(HML-2), the Best Preserved Family of HERVs: Endogenization, Expression, and Implications in Health and Disease. Front Oncol 3:246

79. Höner B, Shoeman RL, Traub P (1991) Human immunodeficiency virus type 1 protease microinjected into cultured human skin fibroblasts cleaves vimentin and affects cytoskeletal and nuclear architecture. J Cell Sci 100:799–807

80. Huesgen PF, Overall CM (2012) N- and C-terminal degradomics: new approaches to reveal biological roles for plant proteases from substrate identification. Physiol Plant 145:5–17

81. Hughes JF, Coffin JM (2004) Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. Proc Natl Acad Sci U S A 101:1668–1672

82. Impens F, Timmerman E, Staes A, Moens K, Ariën KK, Verhasselt B, Vandekerckhove J, Gevaert K (2012) A catalogue of putative HIV-1 protease host cell substrates. Biol Chem 393:915–931

83. Indik S, Gunzburg WH, Salmons B, Rouault F (2005) A novel, mouse mammary tumor virus encoded protein with Rev-like properties. Virology 337:1–6

84. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

85. Ishima R, Gong Q, Tie Y, Weber IT, Louis JM (2010) Highly conserved glycine 86 and arginine 87 residues contribute differently to the structure and activity of the mature HIV-1 protease. Proteins 78:1015–1025

86. Jagdeo JM, Dufour A, Klein T, Solis N, Kleifeld O, Kizhakkedathu J, Luo H, Overall CM, Jan E (2018) N-Terminomics TAILS Identifies Host Cell Substrates of Poliovirus and Coxsackievirus B3 3C Proteinases That Modulate Virus Infection. J Virol 92:e02211-17

87. Jäger S, Cimermancic P, Gulbahce N, Johnson JR, McGovern KE, Clarke SC, Shales M, Mercenne G, Pache L, Li K, Hernandez H, Jang GM, Roth SL, Akiva E, Marlett J, Stephens M, D'Orso I, Fernandes J, Fahey M, Mahon C, O'Donoghue AJ, Todorovic A, Morris JH, Maltby DA, Alber T, Cagney G, Bushman FD, Young JA, Chanda SK, Sundquist WI, Kortemme T, Hernandez RD, Craik CS, Burlingame A, Sali A, Frankel AD, Krogan NJ (2012) Global landscape of HIV-human protein complexes. Nature 481:365–370

88. Jern P, Sperber GO, Blomberg J (2005) Use of endogenous retroviral sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy. Retrovirology 2:50

89. Ji J, Zhang Y, Redon CE, Reinhold WC, Chen AP, Fogli LK, Holbeck SL, Parchment RE, Hollingshead M, Tomaszewski JE, Dudon Q, Pommier Y, Doroshow JH, Bonner WM (2017) Phosphorylated fraction of H2AX as a measurement for DNA damage in cancer cells and potential applications of a novel assay. PLoS One 12:e0171582–e0171582

90. Johnson M (2012) Proteasome Inhibitors. Labome MATER METHODS 2:133

91. Johnson M (2013) Detergents: Triton X-100, Tween-20, and More. Labome MATER METHODS 3:163

92. Jordan SP, Zugay J, Darke PL, Kuo LC (1992) Activity and dimerization of human immunodeficiency virus protease as a function of solvent composition and enzyme concentration. J Biol Chem 267:20028–20032

93. Kallunki T, Barisic M, Jäättelä M, Liu B (2019) How to Choose the Right Inducible Gene Expression System for Mammalian Studies? Cells 8:796

94. Kaneko N, Kurata M, Yamamoto T, Morikawa S, Masumoto J (2019) The role of interleukin-1 in general pathology. Inflamm Regen 39:12

95. Kang M, Kim SY, Lee S, Lee YK, Lee J, Shin HS, Kim YS (2005) Human β-globin second intron highly enhances expression of foreign genes from murine cytomegalovirus immediate-early promoter. J Microbiol Biotechnol 15:544–550

96. Kaplan AH, Swanstrom R (1991) Human immunodeficiency virus type 1 Gag proteins are processed in two cellular compartments. Proc Natl Acad Sci U S A 88:4528–4532

97. Karacostas V, Wolffe EJ, Nagashima K, Gonda MA, Moss B (1993) Overexpression of the HIV-1 Gag-Pol Polyprotein Results in Intracellular Activation of HIV-1 Protease and Inhibition of Assembly and Budding of Virus-like Particles. Virology 193:661–671

98. Kashiwagi H, Shiraishi K, Sakaguchi K, Nakahama T, Kodama S (2018) Repair kinetics of DNA double-strand breaks and incidence

of apoptosis in mouse neural stem/progenitor cells and their differentiated neurons exposed to ionizing radiation. J Radiat Res 59:261–271

99. Katzman M, Sudol M (1995) Mapping domains of retroviral integrase responsible for viral DNA specificity and target site selection by analysis of chimeras between human immunodeficiency virus type 1 and visna virus integrases. J Virol 69:5687–5696

100. Katzman M, Sudol M (1996) Nonspecific alcoholysis, a novel endonuclease activity of human immunodeficiency virus type 1 and other retroviral integrases. J Virol 70:2598–2604

101. Katzman M, Sudol M (1998) Mapping Viral DNA Specificity to the Central Region of Integrase by Using Functional Human Immunodeficiency Virus Type 1/Visna Virus Chimeric Proteins. J Virol 72:1744–1753

102. Katzourakis A, Gifford RJ (2010) Endogenous Viral Elements in Animal Genomes. PLOS Genet 6:e1001191

103. Khanna KK, Jackson SP (2001) DNA double-strand breaks: signaling, repair and the cancer connection. Nat Genet 27:247–254

104. Kim YE, Hipp MS, Bracher A, Hayer-Hartl M, Ulrich Hartl F (2013) Molecular Chaperone Functions in Protein Folding and Proteostasis. Annu Rev Biochem 82:323–355

105. Kitamura Y, Ayukawa T, Ishikawa T, Kanda T, Yoshiike K (1996) Human endogenous retrovirus K10 encodes a functional integrase. J Virol 70:3302–3306

106. Kleifeld O, Doucet A, auf dem Keller U, Prudova A, Schilling O, Kainthan RK, Starr AE, Foster LJ, Kizhakkedathu JN, Overall CM (2010) Isotopic labeling of terminal amines in complex samples identifies protein N-termini and protease cleavage products. Nat Biotechnol 28:281–288

107. Klein T, Fung S-Y, Renner F, Blank MA, Dufour A, Kang S, Bolger-Munro M, Scurll JM, Priatel JJ, Schweigler P, Melkko S, Gold MR, Viner RI, Régnier CH, Turvey SE, Overall CM (2015) The paracaspase MALT1 cleaves HOIL1 reducing linear ubiquitination by LUBAC to dampen lymphocyte NF-κB signalling. Nat Commun 6:8777

108. Koczorowska MM, Tholen S, Bucher F, Lutz L, Kizhakkedathu JN, De Wever O, Wellner UF, Biniossek ML, Stahl A, Lassmann S, Schilling O (2016) Fibroblast activation protein-α, a stromal cell surface protease, shapes key features of cancer associated fibroblasts through proteome and degradome alterations. Mol Oncol 10:40–58

109. Komurian-Pradel F, Paranhos-Baccala G, Bedin F, Ounanian-Paraz A, Sodoyer M, Ott C, Rajoharison A, Garcia E, Mallet F, Mandrand B, Perron H (1999) Molecular cloning and characterization of MSRV-related sequences associated with retrovirus-like particles. Virology 260:1–9

110. Kozak M (1986) Influences of mRNA secondary structure on initiation by eukaryotic ribosomes. Proc Natl Acad Sci U S A 83:2850–2854

111. Kraus B, Boller K, Reuter A, Schnierle BS (2011) Characterization of the human endogenous retrovirus K Gag protein: identification of protease cleavage sites. Retrovirology 8:21

112. Kuhelj R, Rizzo CJ, Chang CH, Jadhav PK, Towler EM, Korant BD (2001) Inhibition of human endogenous retrovirus-K10 protease in cell-free and cell-based assays. J Biol Chem 276:16674–16682

113. Kurth R, Bannert N (2010) Beneficial and detrimental effects of human endogenous retroviruses. Int J cancer 126:306–314

114. Lange PF, Overall CM (2013) Protein TAILS: when termini tell tales of proteolysis and function. Curr Opin Chem Biol 17:73–82

115. Lavialle C, Cornelis G, Dupressoir A, Esnault C, Heidmann O, Vernochet C, Heidmann T (2013) Paleovirology of "syncytins", retroviral env genes exapted for a role in placentation. Philos Trans R Soc Lond B Biol Sci 368:20120507

116. Lavie L, Kitova M, Maldener E, Meese E, Mayer J (2005) CpG methylation directly regulates transcriptional activity of the human endogenous retrovirus family HERV-K(HML-2). J Virol 79:876–883

117. Lawyer FC, Stoffel S, Saiki RK, Chang SY, Landre PA, Abramson RD, Gelfand DH (1993) High-level expression, purification, and enzymatic characterization of full-length Thermus aquaticus DNA polymerase and a truncated form deficient in 5' to 3' exonuclease activity. PCR Methods Appl 2:275–287

118. Lee YN, Bieniasz PD (2007) Reconstitution of an infectious human endogenous retrovirus. PLoS Pathog 3:e10–e10

119. Lemaître C, Tsang J, Bireau C, Heidmann T, Dewannieux M (2017) A human endogenous retrovirus-derived gene that can contribute to oncogenesis by activating the ERK pathway and inducing migration and invasion. PLOS Pathog 13:e1006451

120. Lesbats P, Engelman AN, Cherepanov P (2016) Retroviral DNA Integration. Chem Rev 116:12730–12757

121. Li W, Lee M-H, Henderson L, Tyagi R, Bachani M, Steiner J, Campanac E, Hoffman DA, von Geldern G, Johnson K, Maric D, Morris HD, Lentz M, Pak K, Mammen A, Ostrow L, Rothstein J, Nath A (2015) Human endogenous retrovirus-K contributes to motor neuron disease. Sci Transl Med 7:307ra153

122. Liu Y, Hettinger CL, Zhang D, Rezvani K, Wang X, Wang H (2014) The proteasome function reporter GFPu accumulates in young brains of the APPswe/PS1dE9 Alzheimer's disease mouse model. Cell Mol Neurobiol 34:315–322

123. Löwer R, Boller K, Hasenmaier B, Korbmacher C, Müller-Lantzsch N, Löwer J, Kurth R (1993) Identification of human endogenous retroviruses with complex mRNA expression and particle formation. Proc Natl Acad Sci U S A 90:4480–4484

124. Lowry OH, Rosenbrough NJ, Farr AL, Randall RJ (1951) Protein measurement with the Folin phenol reagent. J Biol Chem 193:265–275

125. Lübbehüsen C, Lüke J, Seeling C, Mellert K, Marienfeld R, von Baer A, Schultheiss M, Möller P, Barth TFE (2019) Characterization of Three Novel H3F3A-mutated Giant Cell Tumor Cell Lines and Targeting of Their Wee1 Pathway. Sci Rep 9:6458

126. Lyn D, Deaven LL, Istock NL, Smulson M (1993) The polymorphic ADP-ribosyltransferase (NAD+) pseudogene 1 in humans interrupts an endogenous pol-like element on 13q34. Genomics 18:206–211

127. Ma L, Yang F, Zheng J (2014) Application of fluorescence resonance energy transfer in protein studies. J Mol Struct 1077:87–100

128. Mager DL, Stoye JP (2015) Mammalian Endogenous Retroviruses. Microbiol Spectr 3:MDNA3-0009–2014

129. Magin C, Lower R, Lower J (1999) cORF and RcRE, the Rev/Rex and RRE/RxRE homologues of the human endogenous retrovirus family HTDV/HERV-K. J Virol 73:9496–9507

130. Magin C, Hesse J, Lower J, Lower R (2000) Corf, the Rev/Rex homologue of HTDV/HERV-K, encodes an arginine-rich nuclear localization signal that exerts a trans-dominant phenotype when mutated. Virology 274:11–16

131. Mallet F, Bouton O, Prudhomme S, Cheynet V, Oriol G, Bonnaud B, Lucotte G, Duret L, Mandrand B (2004) The endogenous retroviral locus ERVWE1 is a bona fide gene involved in hominoid placental physiology. Proc Natl Acad Sci U S A 101:1731–1736

132. Marková E, Schultz N, Belyaev IY (2007) Kinetics and dose-response of residual 53BP1/γ-H2AX foci: Co-localization, relationship with DSB repair and clonogenic survival. Int J Radiat Biol 83:319–329

133. Martin MA, Bryan T, Rasheed S, Khan AS (1981) Identification and cloning of endogenous retroviral sequences present in human DNA. Proc Natl Acad Sci U S A 78:4892–4896

134. Mayer J, Sauter M, Racz A, Scherer D, Mueller-Lantzsch N, Meese E (1999) An almost-intact human endogenous retrovirus K on human chromosome 7. Nat Genet 21:257–258

135. Mayer J, Meese EU (2003) Presence of dUTPase in the various human endogenous retrovirus K (HERV-K) families. J Mol Evol 57:642–649

136. Mayer J, Ehlhardt S, Seifert M, Sauter M, Muller-Lantzsch N, Mehraein Y, Zang K-D, Meese E (2004) Human endogenous retrovirus HERV-K(HML-2) proviruses with Rec protein coding capacity and transcriptional activity. Virology 322:190–198

137. Mayer J, Meese E (2005) Human endogenous retroviruses in the primate lineage and their influence on host genomes. Cytogenet Genome Res 110:448–456

138. Mayer J, Harz C, Sanchez L, Pereira GC, Maldener E, Heras SR, Ostrow LW, Ravits J, Batra R, Meese E, Garcia-Perez JL, Goodier JL (2018) Transcriptional profiling of HERV-K(HML-2) in amyotrophic lateral sclerosis and potential implications for expression of HML-2 proteins. Mol Neurodegener 13:39

139. Medstrand P, Blomberg J (1993) Characterization of novel reverse transcriptase encoding human endogenous retroviral sequences similar to type A and type B retroviruses: differential transcription in normal human tissues. J Virol 67:6778–6787

140. Mi H, Muruganujan A, Thomas PD (2013) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res 41:D377-86

141. Miller I, Min M, Yang C, Tian C, Gookin S, Carter D, Spencer SL (2018) Ki67 is a Graded Rather than a Binary Marker of Proliferation versus Quiescence. Cell Rep 24:1105-1112.e5

142. Montesion M, Bhardwaj N, Williams ZH, Kuperwasser C, Coffin JM (2018) Mechanisms of HERV-K (HML-2) Transcription during Human Mammary Epithelial Cell Transformation. J Virol 92:e01258-17

143. Mueller-Lantzsch N, Sauter M, Weiskircher A, Kramer K, Best B, Buck M, Grasser F (1993) Human endogenous retroviral element K10 (HERV-K10) encodes a full-length gag homologous 73-kDa protein and a functional protease. AIDS Res Hum Retroviruses 9:343–350

144. Muster T, Waltenberger A, Grassauer A, Hirschl S, Caucig P, Romirer I, Födinger D, Seppele H, Schanab O, Magin-Lachmann C, Löwer R, Jansen B, Pehamberger H, Wolff K (2003) An Endogenous Retrovirus Derived from Human Melanoma Cells. Cancer Res 63:8735–8741

145. Nassour J, Martien S, Martin N, Deruy E, Tomellini E, Malaquin N, Bouali F, Sabatier L, Wernert N, Pinte S, Gilson E, Pourtier A, Pluquet O, Abbadie C (2016) Defective DNA single-strand break repair is responsible for senescence and neoplastic escape of epithelial cells. Nat Commun 7:10399

146. Nie Z, Phenix BN, Lum JJ, Alam A, Lynch DH, Beckett B, Krammer PH, Sekaly RP, Badley AD (2002) HIV-1 protease processes procaspase 8 to cause mitochondrial release of cytochrome c, caspase cleavage and nuclear fragmentation. Cell Death Differ 9:1172–1184

147. Nie Z, Bren GD, Rizza SA, Badley AD (2008) HIV Protease Cleavage of Procaspase 8 is Necessary for Death of HIV-Infected Cells. Open Virol J 2:1–7

148. Ono M, Yasunaga T, Miyata T, Ushikubo H (1986) Nucleotide sequence of human endogenous retrovirus genome related to the mouse mammary tumor virus genome. J Virol 60:589–598

149. Padow M, Lai L, Fisher RJ, Zhou YC, Wu X, Kappes JC, Towler EM (2000) Analysis of human immunodeficiency virus type 1 containing HERV-K protease. AIDS Res Hum Retroviruses 16:1973–1980

150. Panier S, Boulton SJ (2014) Double-strand break repair: 53BP1 comes into focus. Nat Rev Mol Cell Biol 15:7–18

151. Parissi V, Caumont A, de Soultrait VR, Desjobert C, Calmels C, Fournier M, Gourgue G, Bonneu M, Tarrago-Litvak L, Litvak S (2003) The lethal phenotype observed after HIV-1 integrase expression in yeast cells is related to DNA repair and recombination events. Gene 322:157–168

152. Perot P, Mugnier N, Montgiraud C, Gimenez J, Jaillard M, Bonnaud B, Mallet F (2012) Microarray-based sketches of the HERV transcriptome landscape. PLoS One 7:e40194

153. Petit C, Schwartz O, Mammano F (1999) Oligomerization within virions and subcellular localization of human immunodeficiency virus type 1 integrase. J Virol 73:5079–5088

154. Pettit SC, Simsic J, Loeb DD, Everitt L, Hutchison CA 3rd, Swanstrom R (1991) Analysis of retroviral protease cleavage sites reveals two types of cleavage sites and the structural requirements of the P1 amino acid. J Biol Chem 266:14539–14547

155. Pluymers W, Cherepanov P, Schols D, De Clercq E, Debyser Z (1999) Nuclear Localization of Human Immunodeficiency Virus Type 1 Integrase Expressed as a Fusion Protein with Green Fluorescent Protein. Virology 258:327–332

156. Porter DJT, Hanlon MH, Furfine ES (2002) HIV-1 Protease: Characterization of a Catalytically Competent Enzyme−Substrate Intermediate. Biochemistry 41:1302–1307

157. Prudova A, Gocheva V, Auf dem Keller U, Eckhard U, Olson OC, Akkari L, Butler GS, Fortelny N, Lange PF, Mark JC, Joyce JA, Overall CM (2016) TAILS N-Terminomics and Proteomics Show Protein Degradation Dominates over Proteolytic Processing by Cathepsins in Pancreatic Tumors. Cell Rep 16:1762–1773

158. Qin JY, Zhang L, Clift KL, Hulur I, Xiang AP, Ren B-Z, Lahn BT (2010) Systematic comparison of constitutive promoters and the doxycycline-inducible promoter. PLoS One 5:e10611

159. Rappold I, Iwabuchi K, Date T, Chen J (2001) Tumor suppressor p53 binding protein 1 (53BP1) is involved in DNA damage-signaling pathways. J Cell Biol 153:613–620

160. Reus K, Mayer J, Sauter M, Zischler H, Müller-Lantzsch N, Meese E (2001) HERV-K(OLD): ancestor sequences of the human endogenous retrovirus family HERV-K(HML-2). J Virol 75:8917–8926

161. Reus K, Mayer J, Sauter M, Scherer D, Müller-Lantzsch N, Meese E (2001) Genomic Organization of the Human Endogenous Retrovirus HERV-K(HML-2.HOM) (ERVK6) on Chromosome 7. Genomics 72:314–320

162. Richards AD, Roberts R, Dunn BM, Graves MC, Kay J (1989) Effective blocking of HIV-1 proteinase activity by characteristic

inhibitors of aspartic proteinases. FEBS Lett 247:113–117

163. Richtie C (2013) Protease Inhibitors. Labome MATER METHODS 3:169

164. Rigogliuso G, Biniossek ML, Goodier JL, Mayer B, Pereira GC, Schilling O, Meese E, Mayer J (2019) A human endogenous retrovirus encoded protease potentially cleaves numerous cellular proteins. Mob DNA 10:36

165. Rivière Y, Blank V, Kourilsky P, Israël A (1991) Processing of the precursor of NF-κB by the HIV-1 protease during acute infection. Nature 350:625–626

166. Rogakou EP, Pilch DR, Orr AH, Ivanova VS, Bonner WM (1998) DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. J Biol Chem 273:5858–5868

167. Romanish MT, Lock WM, van de Lagemaat LN, Dunn CA, Mager DL (2007) Repeated recruitment of LTR retrotransposons as promoters by the anti-apoptotic locus NAIP during mammalian evolution. PLoS Genet 3:e10

168. Ruggieri A, Maldener E, Sauter M, Mueller-Lantzsch N, Meese E, Fackler OT, Mayer J (2009) Human endogenous retrovirus HERV-K(HML-2) encodes a stable signal peptide with biological properties distinct from Rec. Retrovirology 6:17

169. Rumlová M, Křížová I, Keprová A, Hadravová R, Doležal M, Strohalmová K, Pichová I, Hájek M, Ruml T (2014) HIV-1 protease-induced apoptosis. Retrovirology 11:37

170. Ruprecht K, Ferreira H, Flockerzi A, Wahl S, Sauter M, Mayer J, Mueller-Lantzsch N (2008) Human endogenous retrovirus family HERV-K(HML-2) RNA transcripts are selectively packaged into retroviral particles produced by the human germ cell tumor line Tera-1 and originate mainly from a provirus on chromosome 22q11.21. J Virol 82:10008–10016

171. Ruprecht K, Mayer J, Sauter M, Roemer K, Mueller-Lantzsch N (2008) Endogenous retroviruses and cancer. Cell Mol Life Sci 65:3366–3382

172. Ryan EL, Hollingworth R, Grand RJ (2016) Activation of the DNA Damage Response by RNA Viruses. Biomolecules 6:2

173. Ryan FP (2011) Human endogenous retroviruses in multiple sclerosis: potential for novel neuro-pharmacological research. Curr Neuropharmacol 9:360–369

174. Ryu W-S (2017) Retroviruses. Academic Press, Seoul, Korea

175. Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science (80-) 230:1350–1354

176. Sakurai Y, Komatsu K, Agematsu K, Matsuoka M (2009) DNA double strand break repair enzymes function at multiple steps in retroviral infection. Retrovirology 6:114

177. Samuelson LC, Phillips RS, Swanberg LJ (1996) Amylase gene structures in primates: retroposon insertions and promoter evolution. Mol Biol Evol 13:767–779

178. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74:5463–5467

179. Sauter M, Schommer S, Kremmer E, Remberger K, Dölken G, Lemm I, Buck M, Best B, Neumann-Haefelin D, Mueller-Lantzsch N (1995) Human endogenous retrovirus K10: expression of Gag protein and detection of antibodies in patients with seminomas. J Virol 69:414–421

180. Sayer JM, Liu F, Ishima R, Weber IT, Louis JM (2008) Effect of the active site D25N mutation on the structure, stability, and ligand binding of the mature HIV-1 protease. J Biol Chem 283:13459–13470

181. Schembri L, Dalibart R, Tomasello F, Legembre P, Ichas F, De Giorgi F (2007) The HA tag is cleaved and loses immunoreactivity during apoptosis. Nat Methods 4:107–108

182. Schilling O, Overall CM (2007) Proteomic discovery of protease substrates. Curr Opin Chem Biol 11:36–45

183. Schmitt K, Reichrath J, Roesch A, Meese E, Mayer J (2013) Transcriptional profiling of human endogenous retrovirus group HERV-K(HML-2) loci in melanoma. Genome Biol Evol 5:307–328

184. Schmitt K, Heyne K, Roemer K, Meese E, Mayer J (2015) HERV-K(HML-2) rec and np9 transcripts not restricted to disease but present in many normal human tissues. Mob DNA 6:4

185. Schommer S, Sauter M, Kräusslich HG, Best B, Mueller-Lantzsch N (1996) Characterization of the human endogenous retrovirus K proteinase. J Gen Virol 77 (Pt 2:375–379

186. Schultz LB, Chehab NH, Malikzay A, Halazonetis TD (2000) p53 binding protein 1 (53BP1) is an early participant in the cellular response to DNA double-strand breaks. J Cell Biol 151:1381–1390

187. Shoeman RL, Höner B, Stoller TJ, Kesselmeier C, Miedel MC, Traub P, Graves MC (1990) Human immunodeficiency virus type 1 protease cleaves the intermediate filament proteins vimentin, desmin, and glial fibrillary acidic protein. Proc Natl Acad Sci U S A 87:6336–6340

188. Shoeman RL, Kesselmeier C, Mothes E, Höner B, Traub P (1991) Non-viral cellular substrates for human immunodeficiency virus type 1 protease. FEBS Lett 278:199–203

189. Shoeman RL, Sachse C, Höner B, Mothes E, Kaufmann M, Traub P (1993) Cleavage of human and mouse cytoskeletal and sarcomeric proteins by human immunodeficiency virus type 1 protease. Actin, desmin, myosin, and tropomyosin. Am J Pathol 142:221–230

190. Shoeman RL, Hartig R, Hauses C, Traub P (2002) Organization of focal adhesion plaques is disrupted by action of the HIV-1 protease. Cell Biol Int 26:529–539

191. Singh A, Upadhyay V, Upadhyay AK, Singh SM, Panda AK (2015) Protein recovery from inclusion bodies of Escherichia coli using mild solubilization process. Microb Cell Fact 14:41

192. Skalka AM, Katz RA (2005) Retroviral DNA integration and the DNA damage response. Cell Death Differ 12 Suppl 1:971–978

193. Skinner LM, Sudol M, Harper AL, Katzman M (2001) Nucleophile selection for the endonuclease activities of human, ovine, and avian retroviral integrases. J Biol Chem 276:114–124

194. Snášel J, Shoeman R, Hořejší M, Hrušková-Heidingsfeldová O, Sedláček J, Ruml T, Pichová I (2000) Cleavage of Vimentin by Different Retroviral Proteases. Arch Biochem Biophys 377:241–245

195. Sobecki M, Mrouj K, Colinge J, Gerbe F, Jay P, Krasinska L, Dulic V, Fisher D (2017) Cell-Cycle Regulation Accounts for Variability in Ki-67 Expression Levels. Cancer Res 77:2722–2734

196. Starr AE, Bellac CL, Dufour A, Goebeler V, Overall CM (2012) Biochemical characterization and N-terminomics analysis of

leukolysin, the membrane-type 6 matrix metalloprotease (MMP25): chemokine and vimentin cleavages enhance cell migration and macrophage phagocytic activities. J Biol Chem 287:13382–13395

197. Stengel S, Fiebig U, Kurth R, Denner J (2010) Regulation of human endogenous retrovirus-K expression in melanomas by CpG methylation. Genes Chromosomes Cancer 49:401–411

198. Su Y, Meador JA, Geard CR, Balajee AS (2010) Analysis of ionizing radiation-induced DNA damage and repair in three-dimensional human skin model system. Exp Dermatol 19:e16-22

199. Subramanian RP, Wildschutte JH, Russo C, Coffin JM (2011) Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. Retrovirology 8:90

200. Swanstrom R, Wills J (1997) Synthesis, Assembly, and Processing of Viral Proteins. In: Coffin JM, Hughes SH, Varmus HE (eds). Retroviruses. Cold Spring Harbor Laboratory Press, Cold Spring Harbor New York

201. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, Fish P, Harsha B, Hathaway C, Jupe SC, Kok CY, Noble K, Ponting L, Ramshaw CC, Rye CE, Speedy HE, Stefancsik R, Thompson SL, Wang S, Ward S, Campbell PJ, Forbes SA (2018) COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res 47:D941–D947

202. Taylor RC, Cullen SP, Martin SJ (2008) Apoptosis: controlled demolition at the cellular level. Nat Rev Mol Cell Biol 9:231–241

203. Teshima S, Shimosato Y, Hirohashi S, Tome Y, Hayashi I, Kanazawa H, Kakizoe T (1988) Four new human germ cell tumor cell lines. Lab Invest 59:328–336

204. Tomasselli AG, Hui JO, Adams L, Chosay J, Lowery D, Greenberg B, Yem A, Deibel MR, Zurcher-Neely H, Heinrikson RL (1991) Actin, troponin C, Alzheimer amyloid precursor protein and pro-interleukin 1 beta as substrates of the protease from human immunodeficiency virus. J Biol Chem 266:14548–14553

205. Towler EM, Gulnik S V, Bhat TN, Xie D, Gustchina E, Sumpter TR, Robertson N, Jones C, Sauter M, Mueller-Lantzsch N, Debouck C, Erickson JW (1998) Functional Characterization of the Protease of Human Endogenous Retrovirus, K10: Can It Complement HIV-1 Protease? Biochemistry 37:17137–17144

206. Tristem M (2000) Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. J Virol 74:3715–3730

207. Turnbull MG, Douville RN (2018) Related Endogenous Retrovirus-K Elements Harbor Distinct Protease Active Site Motifs. Front Microbiol 9:1577

208. Turner G, Barbulescu M, Su M, Jensen-Seaman MI, Kidd KK, Lenz J (2001) Insertional polymorphisms of full-length endogenous retroviruses in humans. Curr Biol 11:1531–1535

209. Tyagi R, Li W, Parades D, Bianchet MA, Nath A (2017) Inhibition of human endogenous retrovirus-K by antiretroviral drugs. Retrovirology 14:21

210. Vamvakas S, Vock EH, Lutz WK (1997) On the Role of DNA Double-Strand Breaks in Toxicity and Carcinogenesis. Crit Rev Toxicol 27:155–174

211. Vogt PK (1997) Historical Introduction to the General Properties of Retroviruses. In: Coffin JM, Hughes SH, Varmus HE (eds). Retroviruses. Cold Spring Harbor Laboratory Press, Cold Spring Harbor New York

212. Wagner RN, Reed JC, Chanda SK (2015) HIV-1 protease cleaves the serine-threonine kinases RIPK1 and RIPK2. Retrovirology 12:74

213. Wallin M, Deinum J, Goobar L, Danielson UH (1990) Proteolytic cleavage of microtubule-associated proteins by retroviral proteinases. J Gen Virol 71:1985–1991

214. Wang-Johanning F, Frost AR, Jian B, Epp L, Lu DW, Johanning GL (2003) Quantitation of HERV-K env gene expression and splicing in human breast cancer. Oncogene 22:1528–1535

215. Wang-Johanning F, Liu J, Rycaj K, Huang M, Tsai K, Rosen DG, Chen D-T, Lu DW, Barnhart KF, Johanning GL (2007) Expression of multiple human endogenous retrovirus surface envelope proteins in ovarian cancer. Int J cancer 120:81–90

216. Wang D, Eraslan B, Wieland T, Hallström B, Hopf T, Zolg DP, Zecha J, Asplund A, Li L, Meng C, Frejno M, Schmidt T, Schnatbaum K, Wilhelm M, Ponten F, Uhlen M, Gagneur J, Hahne H, Kuster B (2019) A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. Mol Syst Biol 15:e8503

217. Wang H, Adhikari S, Butler BE, Pandita TK, Mitra S, Hegde ML (2014) A Perspective on Chromosomal Double Strand Break Markers in Mammalian Cells. Jacobs J Radiat Oncol 1:3

218. Werner T, Brack-Werner R, Leib-Mosch C, Backhaus H, Erfle V, Hehlmann R (1990) S71 is a phylogenetically distinct human endogenous retroviral element with structural and sequence homology to simian sarcoma virus (SSV). Virology 174:225–238

219. Wildschutte JH, Williams ZH, Montesion M, Subramanian RP, Kidd JM, Coffin JM (2016) Discovery of unfixed endogenous retrovirus insertions in diverse human populations. Proc Natl Acad Sci U S A 113:E2326-34

220. Wlodawer A, Gustchina A (2000) Structural and biochemical studies of retroviral proteases. Biochim Biophys Acta 1477:16–34

221. Wondrak EM, Louis JM, Mora PT, Oroszlan S (1991) Purification of HIV-1 wild-type protease and characterization of proteolytically inactive HIV-1 protease mutants by pepstatin A affinity chromatography. FEBS Lett 280:347–350

222. Yoshikawa R, Takeuchi JS, Yamada E, Nakano Y, Misawa N, Kimura Y, Ren F, Miyazawa T, Koyanagi Y, Sato K (2017) Feline Immunodeficiency Virus Evolutionarily Acquires Two Proteins, Vif and Protease, Capable of Antagonizing Feline APOBEC3. J Virol 91:e00250-17

223. Zábranský A, Hadravová R, Štokrová J, Sakalian M, Pichová I (2009) Premature processing of mouse mammary tumor virus Gag polyprotein impairs intracellular capsid assembly. Virology 384:33–37

224. Zhou F, Li M, Wei Y, Lin K, Lu Y, Shen J, Johanning GL, Wang-Johanning F (2016) Activation of HERV-K Env protein is essential for tumorigenesis and metastasis of breast cancer cells. Oncotarget 7:84093–84117

7. Appendix

Table 21: Gene/protein names of potential cellular substrates of HML-2 Pro identified by TAILS analysis

AARS1	BOLA2	CPS1	EIF3G	GPN3	IPO5	NCL	PFN2	PSMD2	RPS15A	SRM	UBA1
ACAT1	BORCS5	CRCP	EIF3I	GPS1	IPO7	NCOA7	PGAM1	PSMD3	RPS16	SRP19	UBA2
ACAT2	BTF3	CSDE1	EIF3J	GPT2	IQGAP1	NDRG3	PGAM4	PSMD4	RPS18	SRP9	UBA3
ACBD6	BUB1B	CSNK1E	EIF3L	GRIPAP1	IVNS1ABP	NDUFA5	PGD	PSMD7	RPS19	SSB	UBA5
ACIN1	C11orf98	CSRP1	EIF3M	GRSF1	JPT1	NDUFS6	PGK1	PSME2	RPS2	SSBP3	UBAC1
ACLY	C12orf29	CSTB	EIF4A1	GSK3A	JPT2	NEDD8	PGM1	PSMG1	RPS20	SSNA1	UBAP2L
ACO1	C12orf57	CSTF2	EIF4B	GSK3B	KCTD16	NELFE	PHACTR4	PTBP1	RPS21	ST13	UBE2L3
ACTA1	C15orf57	CTCF	EIF4G1	GSPT1	KDM3B	NHP2	PHF5A	PTGES3	RPS23	STAT3	UBE2O
ACTA2	C1orf198	CTDP1	EIF4G2	GSPT2	KHSRP	NME1	PHF6	PTMS	RPS24	STIM1	UBE2Q1
ACTB	C1orf52	CTNND1	EIF5	GSR	KIF5B	NME2P1	PHGDH	PUF60	RPS25	STIP1	UBE2S
ACTBL2	C2orf49	CTPS1	EIF5A	GSTM3	KLC1	NMT1	PHPT1	PUS7	RPS27	STMN1	UBE3A
ACTC1	C8orf33	CTR9	ELAC2	GTF2E1	KLC2	NOLC1	PIN1	PWP1	RPS27A	STUB1	UBQLN1
ACTG1	C9orf40	CTTN	ELF1	GTF2I	KLF13	NONO	PITPNM1	PYM1	RPS28	SUB1	UBQLN4
ACTG2	CACYBP	CWF19L1	ELOA	GTF3C5	KLHDC4	NOSIP	PKM	QARS1	RPS3A	SUGT1	UBR4
ACTL6B	CALD1	CYRIB	ELOB	H2AZ1	KPNA2	NPM1	PLCB3	RAD23A	RPS4X	SURF6	UBR5
ACTN1	CALM1	DAB2	ELOC	H2BC9	KPNA6	NSFL1C	PLEKHG2	RAD23B	RPS5	SWAP70	UBR7
ACTN4	CALR	DARS1	ENAH	H3C15	KPNB1	NSUN2	PLIN3	RAI14	RPS6	TACC2	UBXN1
ACTR1A	CALU	DAXX	ENO1	HARS1	KRT18	NT5C2	PLS3	RAN	RPS6KA3	TACC3	UBXN4
ACTR3	CAMK1D	DBN1	ENO3	HDGF	KRT8	NT5DC1	PML	RANBP1	RPS7	TAGLN2	UCHL3
ADAMTS7	CAPNS1	DBNL	ENSA	HDGFL2	KTI12	NUB1	PMM2	RANBP2	RPS8	TARS1	UCHL5
ADH5	CAPZB	DCAF1	EPRS1	HECTD1	LAMTOR5	NUBP1	POLD1	RANBP3	RPS9	TBCB	UGDH
ADRM1	CARHSP1	DCAF6	EPS15L1	HERC1	LARP1	NUCKS1	POLE3	RANGAP1	RPSA	TCERG1	UHRF1
AHCYL1	CASP1	DCTN2	ERCC5	HEXIM1	LARP4	NUDC	POTEE	RARS1	RRBP1	TCOF1	UPF1
AHSA1	CAST	DDB1	ERCC6L	HGH1	LARS1	NUFIP2	POTEI	RBBP5	RRM1	TCP1	USF1
AIMP2	CAVIN1	DDI2	ERH	HHATL	LASP1	NUMA1	POTEKP	RBBP7	RRM2	TDP2	USP14
AJUBA	CBR1	DDTL	ERP44	HINT1	LENEP	NUP153	PPA1	RCN1	RRP15	TECPR2	USP15
AK1	CBX3	DDX1	ESD	HINT2	LIMA1	NUP214	PPIA	RCN3	RSL1D1	TFG	USP5
AKAP12	CCAR2	DDX17	EXOSC1	HMGB1	LIMCH1	NUP98	PPID	REPS1	RTCB	TGM2	UTP3
AKAP13	CCDC124	DDX21	EZR	HMGB3	LRRFIP1	OBSL1	PPIE	REXO4	RTN4	THG1L	VARS1
AKAP2	CCDC43	DDX3X	FABP5	HMGN5	LSM2	OGA	PPIF	RFC4	RUVBL2	THOP1	VASP
AKR7A2	CCDC50	DDX46	FAM114A1	HNRNPA0	LTV1	OGFR	PPIL1	RGCC	S100A10	THRAP3	VCL
AKT1S1	CCT2	DDX5	FAM120A	HNRNPA1	LZIC	OLA1	PPIP5K2	RIOK1	S100A11	TKFC	VCP
ALDH9A1	CCT3	DFFA	FAM136A	HNRNPA1L2	MACROD1	ORC2	PPM1B	RIPOR1	S100A4	TKT	VTA1
ALDOA	CCT4	DHX15	FAM172A	HNRNPA2B1	MAGED1	OSBPL10	PPM1G	RNASEH2A	S100A6	TLN1	WAPL
ALDOC	CCT5	DHX9	FASN	HNRNPAB	MAP2K2	OTUB1	PPP1R14B	RNASEH2B	SAE1	TMOD3	WASHC2A
ANAPC16	CCT6A	DLGAP5	FDPS	HNRNPD	MAP4	OTULIN	PPP1R2B	RNF146	SAMD4B	TMSB10	WBP4
ANKZF1	CCT7	DMD	FH	HNRNPH1	MAP7	OXSR1	PPP2CB	RNF25	SARS1	TNKS1BP1	WDHD1
ANP32A	CCT8	DNAJA1	FHL1	HNRNPK	MAP7D1	P4HB	PPP2R1A	RNH1	SDF2L1	TNPO1	WDR1
ANP32E	CD2BP2	DNAJA2	FKBP1A	HNRNPR	MAPRE1	PA2G4	PPP2R1B	RPAP1	SEC16A	TNRC6B	WIPI2
ANXA1	CDC37	DNAJC21	FKBP2	HNRNPU	MAT2A	PABPC1	PPP3R1	RPAP3	SERPINH1	TOE1	WNK1
ANXA2	CDC42	DNAJC7	FKBP4	HNRNPUL1	MBD3	PABPC4	PPP4R2	RPL10	SET	TP53BP1	XPNPEP1
ANXA3	CDC73	DNAJC8	FLNA	HP1BP3	MCM2	PABPN1	PPP6R3	RPL10A	SF1	TPD52	XRCC1
ANXA5	CDK11B	DNM1L	FLNB	HSD17B10	MCM4	PAICS	PPT1	RPL10L	SF3B1	TPD52L2	XRCC4
AP1B1	CDV3	DNMT1	FLYWCH2	HSD48	MCM6	PAIP2	PRAG1	RPL11	SF3B2	TPI1	XRCC5
APPL1	CEP170	DOHH	FOXK1	HSP90AA1	MDC1	PALM2	PRDX1	RPL13	SF3B5	TPM2	XRCC6
APRT	CEP170B	DPP7	FPGT	HSP90AB1	MDH2	PANK2	PRDX2	RPL17	SFN	TPM3	XRN2
ARAF	CFL1	DPYSL2	FSCN1	HSP90B1	MFAP1	PAPOLA	PRDX5	RPL19	SFPQ	TPM4	YAP1
ARAP1	CHAC2	DSTN	FUBP1	HSPA1B	MICAL3	PAPSS1	PRDX6	RPL22	SGTA	TPR	YARS1
ARCN1	CHAMP1	DUT	G3BP1	HSPA2	MIF	PARD3	PRKCSH	RPL23	SHTN1	TRAFD1	YBX1
ARHGAP40	CHML	DYNC1I2	G3BP2	HSPA4	MKI67	PARK7	PRKDC	RPL24	SKP1	TRAP1	YBX3
ARHGDIA	CHMP4B	DYNC1LI1	GANAB	HSPA4L	MOB4	PASK	PRMT5	RPL26L1	SLC4A1AP	TRIM28	YJU2
ARHGEF1	CHMP5	DYNC1LI2	GAPDH	HSPA5	MPG	PAWR	PROSER2	RPL27	SLK	TRIM33	YOD1
ARIH1	CHORDC1	EARS2	GAPVD1	HSPA6	MSH6	PC	PRPF4	RPL27A	SMAP	TRMT112	YWHAB
ARPC4	CIAPIN1	EBNA1BP2	GARS1	HSPA8	MSN	PCBP1	PRPS1	RPL29	SMARCA2	TRMT2A	YWHAE
ARPIN	CKAP5	ECPAS	GART	HSPA9	MTA2	PCBP2	PRPS2	RPL3	SMARCE1	TRMT5	YWHAG
ASS1	CKB	EDC3	GBF1	HSPB1	MTHFD1	PCLAF	PRPSAP1	RPL30	SMC2	TRMT6	YWHAQ
ATIC	CKS1B	EDF1	GCLM	HSPB8	MTRR	PCM1	PRRC2A	RPL32	SMC4	TRUB1	YWHAZ
ATP5F1B	CKS2	EEF1A1	GDI1	HSPD1	MYBBP1A	PCNP	PRRC2B	RPL34	SNAP29	TSNAX	YY1
ATP6V1G1	CLIC1	EEF1A2	GDI2	HSPE1	MYH10	PDAP1	PRRC2C	RPL35	SNCG	TTC1	ZC3H14
ATP6V1H	CLIC4	EEF1B2	GEMIN5	HSPH1	MYH9	PDCD10	PSAT1	RPL36	SND1	TTC9C	ZC3H8
ATXN2L	CLIP1	EEF1D	GFER	HTT	MYL12A	PDCD4	PSMA1	RPL36AL	SNRPB	TTLL12	ZFYVE16
ATXN7L3B	CLNS1A	EEF1G	GFM1	HUWE1	MYL6	PDCD6IP	PSMA3	RPL39P5	SNRPC	TUBA1A	ZFYVE19
B2M	CMPK1	EEF2	GID8	HYOU1	MYL9	PDCL	PSMA5	RPL4	SNUPN	TUBA1C	ZGPAT
BAG2	CNBP	EFHC2	GIGYF2	ICAM1	MYPN	PDE12	PSMA7	RPL5	SNX1	TUBA4A	ZMYM3
BAG3	CNDP2	EFHD2	GKAP1	IDE	NAA10	PDE3A	PSMB1	RPL7	SOD1	TUBB	ZNF428
BAG6	CNN2	EIF2D	GLCCI1	IDH1	NACA	PDIA3	PSMB4	RPL8	SORD	TUBB4B	ZNF622
BAIAP2L1	CNPY2	EIF2S2	GLRX3	IGBP1	NADK	PDLIM1	PSMC2	RPLP0P6	SP1	TUFM	ZPR1
BANF1	COBL	EIF3A	GMPS	IGF2BP1	NANS	PEX19	PSMC3	RPP25	SP2	TWF2	ZWINT
BCL3	COPB2	EIF3B	GOLGA2P5	ILRUN	NAP1L1	PFAS	PSMC4	RPRD2	SPAG9	TXNDC5	ZYX
BCL9L	COPG1	EIF3CL	GOLGA3	IMPDH2	NAP1L4	PFDN2	PSMC5	RPS10	SPINDOC	TXNDC9	
BIN1	COPZ1	EIF3D	GOLGA4	INF2	NAPRT	PFDN6	PSMC6	RPS12	SPTBN1	TXNRD1	
BLVRB	CORO1B	EIF3E	GPI	INPPL1	NASP	PFKP	PSMD11	RPS13	SQSTM1	TYMS	
BOD1L1	CPNE1	EIF3F	GPKOW	INTS3	NAXE	PFN1	PSMD12	RPS14	SREK1		

* Gene/protein names were retrieved from UniProt [40] providing, as identifiers, protein IDs of potential cellular substrates identified by TAILS analysis.

Table 22: Additional cellular proteins verified *in vivo* as substrates of HML-2 Pro

Protein ID Gene (protein)	GO cellular component	GO biological process
O95817 BAG3 (BAG family molecular chaperone regulator 3)	aggresome; cell; chaperone complex; cytoplasm; cytosol; neuron projection; nucleus; plasma membrane; stress fiber; Z disc	aggresome assembly; autophagosome assembly; brain development; cellular response to heat; cellular response to mechanical stimulus; cellular response to unfolded protein; chaperone-mediated autophagy; chaperone-mediated protein transport; extrinsic apoptotic signaling pathway in absence of ligand; extrinsic apoptotic signaling pathway via death domain receptors; muscle cell cellular homeostasis; negative regulation of apoptotic process; negative regulation of protein targeting to mitochondrion; negative regulation of striated muscle cell apoptotic process; negative regulation of transcription from RNA polymerase II promoter in response to stress; positive regulation of aggrephagy; positive regulation of protein export from nucleus; positive regulation of protein import into nucleus; protein folding; protein stabilization; protein transport along microtubule; regulation of cellular response to heat; spinal cord development
Q9NR30 DDX21 (nucleolar RNA helicase 2)	cytosol; membrane; mitochondrion; nucleolus; nucleoplasm	defense response to virus; innate immune response; osteoblast differentiation; positive regulation of gene expression, epigenetic; positive regulation of I-kappaB kinase/NF-kappaB signaling; positive regulation of myeloid dendritic cell cytokine production; response to exogenous dsRNA; rRNA processing; transcription by RNA polymerase II
P23588 EIF4B (eukaryotic translation initiation factor 4B)	cytosol; eukaryotic translation initiation factor 4F complex; polysome	eukaryotic translation initiation factor 4F complex assembly; regulation of translational initiation; translational initiation
Q04637 EIF4GI (eukaryotic translation initiation factor 4 gamma 1)	cytoplasm; cytosol; eukaryotic translation initiation factor 4F complex; membrane; nucleus; polysome	behavioral fear response; cap-dependent translational initiation; cellular macromolecule biosynthetic process; cellular response to nutrient levels; developmental process; energy homeostasis; negative regulation of autophagy; negative regulation of neuron death; negative regulation of peptidyl-threonine phosphorylation; nuclear-transcribed mRNA catabolic process, nonsense-mediated decay; positive regulation of cell death; positive regulation of cell growth; positive regulation of cellular protein metabolic process; positive regulation of eukaryotic translation initiation factor 4F complex assembly; positive regulation of G1/S transition of mitotic cell cycle; positive regulation of miRNA mediated inhibition of translation; positive regulation of mRNA cap binding; positive regulation of neuron differentiation; positive regulation of peptidyl-serine phosphorylation; positive regulation of translation in response to endoplasmic reticulum stress; regulation of cellular response to stress; regulation of gene silencing by miRNA; regulation of mRNA stability; regulation of polysome binding; regulation of presynapse assembly; regulation of translational initiation; translation; translational initiation; viral process
Q13283 G3BP1 (ras GTPase-activating protein-binding protein 1)	cytoplasm; cytoplasmic stress granule; cytosol; focal adhesion; nucleus; perikaryon; ribonucleoprotein complex	defense response to virus; innate immune response; negative regulation of canonical Wnt signaling pathway; positive regulation of stress granule assembly; Ras protein signal transduction; stress granule assembly; type I interferon production; viral process
P11021 HSPA5 (endoplasmic reticulum chaperone BiP)	cell surface; cytoplasm; cytosol; endoplasmic reticulum; endoplasmic reticulum chaperone complex; endoplasmic reticulum-Golgi intermediate compartment; endoplasmic reticulum lumen; endoplasmic reticulum membrane; extracellular exosome; focal adhesion; integral component of endoplasmic reticulum membrane; intracellular membrane-bounded organelle; melanosome; membrane; midbody; mitochondrion; nucleus; plasma membrane; protein-containing complex; smooth endoplasmic reticulum	ATF6-mediated unfolded protein response; cellular response to antibiotic; cellular response to calcium ion; cellular response to cAMP; cellular response to drug; cellular response to gamma radiation; cellular response to glucose starvation; cellular response to interleukin-4; cellular response to manganese ion; cellular response to nerve growth factor stimulus; cellular response to unfolded protein; cerebellar Purkinje cell layer development; cerebellum structural organization; chaperone cofactor-dependent protein refolding; endoplasmic reticulum unfolded protein response; ER overload response; IRE1-mediated unfolded protein response; luteolysis; maintenance of protein localization in endoplasmic reticulum; negative regulation of apoptotic process; negative regulation of IRE1-mediated unfolded protein response; negative regulation of protein-containing complex assembly; negative regulation of transforming growth factor beta receptor signaling pathway; neuron apoptotic process; neuron differentiation; PERK-mediated unfolded protein response; positive regulation of cell migration; positive regulation of neuron projection development; positive regulation of protein ubiquitination; positive regulation of transcription from RNA polymerase II promoter in response to endoplasmic reticulum stress; posttranslational protein targeting to membrane, translocation; protein folding in endoplasmic reticulum; protein refolding; regulation of ATF6-mediated unfolded protein response; regulation of IRE1-mediated unfolded protein response; regulation of PERK-mediated unfolded protein response; regulation of protein folding in endoplasmic reticulum; response to cocaine; response to methamphetamine hydrochloride; response to unfolded protein; stress response to metal ion; substantia nigra development; toxin transport; ubiquitin-dependent ERAD pathway
P04792 HSPB1 (heat shock protein beta-1)	axon cytoplasm; cytoplasm; cytoskeleton; cytosol; extracellular exosome; extracellular space; focal adhesion; nucleus; plasma membrane; proteasome complex; spindle; Z disc	anterograde axonal protein transport; cellular response to vascular endothelial growth factor stimulus; chaperone-mediated protein folding; intracellular signal transduction; negative regulation of apoptotic process; negative regulation of oxidative stress-induced intrinsic apoptotic signaling pathway; negative regulation of protein kinase activity; platelet aggregation; positive regulation of angiogenesis; positive regulation of blood vessel endothelial cell migration; positive regulation of endothelial cell chemotaxis; positive regulation of endothelial cell chemotaxis by VEGF-activated vascular endothelial growth factor receptor signaling pathway; positive regulation of interleukin-1 beta production; positive regulation of tumor necrosis factor biosynthetic process; regulation of autophagy; regulation of I-kappaB kinase/NF-kappaB signaling; regulation of mRNA stability; regulation of protein phosphorylation; regulation of translational initiation; response to unfolded protein; response to virus; retina homeostasis; viral process
P06748 NPM1 (nucleophosmin)	centrosome; cytoplasm; cytosol; focal adhesion; membrane; nucleolus; nucleoplasm; nucleus; protein-containing complex; protein-DNA complex; ribonucleoprotein	cell aging; cellular response to UV; CENP-A containing nucleosome assembly; centrosome cycle; chromatin remodeling; DNA repair; intracellular protein transport; negative regulation of apoptotic process; negative regulation of cell population proliferation; negative regulation of centrosome duplication; negative regulation of protein kinase activity by regulation of protein phosphorylation; nucleocytoplasmic transport; nucleosome assembly; positive regulation of cell cycle G2/M phase transition; positive regulation of cell population proliferation; positive regulation of NF-kappaB transcription factor activity; positive regulation of transcription, DNA-templated; positive regulation of transcription by RNA polymerase II; positive regulation of translation; protein localization; regulation of

Protein ID / Gene / Name	GO compartments	GO processes
	complex; spindle pole centrosome	centriole replication; regulation of centrosome duplication; regulation of eIF2 alpha phosphorylation by dsRNA; regulation of endodeoxyribonuclease activity; regulation of endoribonuclease activity; regulation of mRNA stability involved in cellular response to UV; regulation of transcription by RNA polymerase II; ribosomal large subunit biogenesis; ribosomal large subunit export from nucleus; ribosomal small subunit biogenesis; ribosomal small subunit export from nucleus; ribosome assembly; rRNA export from nucleus; signal transduction; viral process
P11940 PABPC1 (polyadenylate-binding protein 1)	catalytic step 2 spliceosome; cell leading edge; cytoplasm; cytoplasmic ribonucleoprotein granule; cytoplasmic stress granule; cytosol; extracellular exosome; focal adhesion; membrane; nucleus; ribonucleoprotein complex	gene silencing by RNA; mRNA polyadenylation; mRNA splicing, via spliceosome; mRNA stabilization; negative regulation of nuclear-transcribed mRNA catabolic process, nonsense-mediated decay; nuclear-transcribed mRNA catabolic process, nonsense-mediated decay; positive regulation of nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay; positive regulation of nuclear-transcribed mRNA poly(A) tail shortening; positive regulation of viral genome replication; regulation of mRNA stability; translational initiation
P47897 QARS (glutamine-tRNA ligase)	aminoacyl-tRNA synthetase multienzyme complex; cytoplasm; cytosol; mitochondrial matrix; protein-containing complex	brain development; glutaminyl-tRNA aminoacylation; negative regulation of apoptotic signaling pathway; negative regulation of protein kinase activity; negative regulation of stress-activated MAPK cascade; negative regulation of transcription, DNA-templated; tRNA aminoacylation for protein translation
O75792 RNASEH2A (ribonuclease H2 subunit A)	cytosol; nucleoplasm; ribonuclease H2 complex	DNA replication; DNA replication, removal of RNA primer; mismatch repair; RNA catabolic process
Q5TBB1 RNASEH2B (ribonuclease H2 subunit B)	nucleoplasm; nucleus; ribonuclease H2 complex	in utero embryonic development; negative regulation of gene expression; positive regulation of fibroblast proliferation; regulation of DNA damage checkpoint; regulation of G2/M transition of mitotic cell cycle; ribonucleotide metabolic process; RNA catabolic process
Q9Y230 RUVBL2 (RuvB-like 2)	centrosome; cytoplasm; cytosol; extracellular exosome; Ino80 complex; intracellular; membrane; MLL1 complex; NuA4 histone acetyltransferase complex; nuclear euchromatin; nuclear matrix; nucleoplasm; nucleus; R2TP complex; ribonucleoprotein complex; Swr1 complex	box C/D snoRNP assembly; cellular response to estradiol stimulus; cellular response to UV; chromatin remodeling; DNA recombination; DNA repair; establishment of protein localization to chromatin; histone acetylation; histone H2A acetylation; histone H4 acetylation; negative regulation of canonical Wnt signaling pathway; negative regulation of estrogen receptor binding; positive regulation of histone acetylation; positive regulation of telomerase RNA localization to Cajal body; positive regulation of transcription by RNA polymerase II; protein folding; regulation of growth; regulation of transcription by RNA polymerase II; transcriptional activation by promoter-enhancer looping
Q9UNE7 STUB1 (E3 ubiquitin-protein ligase CHIP)	chaperone complex; cytoplasm; cytosol; endoplasmic reticulum; nuclear inclusion body; nucleoplasm; nucleus; ubiquitin conjugating enzyme complex; ubiquitin ligase complex; Z disc	cellular response to heat; cellular response to hypoxia; cellular response to misfolded protein; chaperone-mediated autophagy; DNA repair; endoplasmic reticulum unfolded protein response; ERBB2 signaling pathway; negative regulation of protein binding; negative regulation of transforming growth factor beta receptor signaling pathway; positive regulation of chaperone-mediated protein complex assembly; positive regulation of proteasomal ubiquitin-dependent protein catabolic process; positive regulation of protein ubiquitination; positive regulation of ubiquitin-protein transferase activity; proteasome-mediated ubiquitin-dependent protein catabolic process; protein autoubiquitination; protein K63-linked ubiquitination; protein maturation; protein polyubiquitination; protein quality control for misfolded or incompletely synthesized proteins; protein ubiquitination; regulation of glucocorticoid metabolic process; regulation of protein stability; response to ischemia; ubiquitin-dependent ERAD pathway; ubiquitin-dependent protein catabolic process; ubiquitin-dependent SMAD protein catabolic process
Q13263 TRIM28 (transcription intermediary factor 1-beta)	chromatin; nuclear euchromatin; nuclear heterochromatin; nucleoplasm; nucleus; protein-containing complex; RNA polymerase II transcription factor complex	chromatin organization; convergent extension involved in axis elongation; DNA methylation involved in embryo development; DNA repair; embryo implantation; embryonic placenta morphogenesis; epithelial to mesenchymal transition; innate immune response; negative regulation of DNA demethylation; negative regulation of single stranded viral RNA replication via double stranded DNA intermediate; negative regulation of transcription, DNA-templated; negative regulation of transcription by RNA polymerase II; negative regulation of viral release from host cell; positive regulation of DNA binding; positive regulation of DNA repair; positive regulation of methylation-dependent chromatin silencing; positive regulation of protein import into nucleus; positive regulation of transcription, DNA-templated; protein autophosphorylation; protein sumoylation; Ras protein signal transduction; regulation of genetic imprinting; transcription initiation from RNA polymerase II promoter; viral process
O95071 UBR5 (E3 ubiquitin-protein ligase UBR5)	cytosol; membrane; nucleoplasm; nucleus; perinuclear region of cytoplasm; protein-containing complex	cellular response to DNA damage stimulus; DNA repair; negative regulation of double-strand break repair; negative regulation of histone H2A K63-linked ubiquitination; positive regulation of canonical Wnt signaling pathway; positive regulation of gene expression; positive regulation of protein import into nucleus; progesterone receptor signaling pathway; protein K48-linked ubiquitination; protein polyubiquitination; regulation of double-strand break repair; viral process

* Experiments validating additional cellular substrates of HML-2 Pro were conducted by Dr. John Goodier's laboratory (McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA). A representative protein ID, approved gene/protein symbols and full names are given each. GO compartments and GO processes for respective genes/proteins were compiled from UniProt.

8. Abbreviations

°C	Degree Celsius
~	Approximately
53BP1	53 binding protein 1
aa	Amino acid(s)
ALS	Amyotrophic lateral sclerosis
bp	Base pair
BSA	Bovine serum albumine
CMV	Cytomegalovirus
DMEM	Dulbecco's modified eagle medium
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DPBS	Dulbecco's phosphate-buffered saline
DSB	DNA double-strand break
dsDNA	Double-stranded DNA
dUTPase	Deoxyuridine triphosphatase
E. coli	*Escherichia coli*
e.g.	*Exempli gratia* (for example)
ECL	Enhanced chemiluminescence
EGFP	Enhanced green fluorescent protein
Env	Envelope
ERV	Endogenous retrovirus
etc	Et cetera
FACS	Fluorescence-activated cell sorting
g	Relative centrifugal force
Gag	Group-specific antigen
GCT	Germ cell tumor
GFP	Green fluorescent protein
GO	Gene ontology
h	Hours
H_2O_{dd}	Double distilled water
HERV	Human endogenous retrovirus
HIV-1	Human immunodeficiency virus type 1
HML	Human MMTV-like
hpt	Hours post-transfection
HSP	Heat shock protein
HTLV	Human T-lymphotropic virus
i.e.	*id est* (that is)
IgG	Immunoglobulin G
IN	Integrase
IPTG	Isopropyl β-D-1-thiogalaktopyranoside
IR	Ionizing radiation
kb	Kilobase
kDa	Kilodaltons
L	Liter
LTR	Long terminal repeat
m	Milli (10^{-3})
M	Molar (mol/l)
min	Minutes
MMTV	Mouse mammary tumor virus
mRNA	Messenger-RNA

n	Nano (10^{-9})
NF-κB	Nuclear factor kappa-light-chain-enhancer of activated B cells
nt	Nucleotide
ORF	Open reading frame
PAA-gel	Polyacrylamide gel
PAGE	Polyacrylamide gel electrophoresis
PARP	Poly [ADP-ribose] polymerase 1
PBS	Primer binding site
PCR	Polymerase chain reaction
Pro	Protease
RNA	Ribonucleic acid
RSV	Rous sarcoma virus
RT	Reverse transcriptase
SA	Splice acceptor
SD	Splice donor
SDS	Sodium dodecyl sulfate
sec	Second
SP	Signal peptide
SSB	DNA single-strand break
ssDNA	Single-stranded DNA
SU	Surface
TAE	Tris-acetate-EDTA
Taq	*Thermus aquaticus*
tRNA	Transfer-RNA
UV	Ultraviolet
V	Volt
V/V	Volume/volume percentage
W/V	Weight/volume percentage
μ	Micro (10^{-6})

Amino acids

1-letter code	3-letter code	name	1-letter code	3-letter code	name
A	Ala	Alanine	K	Lys	Lysine
R	Arg	Arginine	M	Met	Methionine
N	Asn	Asparagine		Nle	Norleucine
D	Asp	Aspartic acid	F	Phe	Phenylalanine
C	Cys	Cysteine	P	Pro	Proline
Q	Gln	Glutamine	S	Ser	Serine
E	Glu	Glutamic acid	T	Thr	Threonine
G	Gly	Glycine	W	Trp	Tryptophan
H	His	Histidine	Y	Tyr	Tyrosine
I	Ile	Isoleucine	V	Val	Valine
L	Leu	Leucine			

Nucleobases

A	Adenine
C	Cytosine
G	Guanine
T	Thymine

9. List of figures

10. List of tables

11. Publications

Rigogliuso G, Biniossek ML, Goodier JL, Mayer B, Pereira GC, Schilling O, Meese E, Mayer J (2019) A human endogenous retrovirus encoded protease potentially cleaves numerous cellular proteins. Mob DNA 10:1–22

Poster presentation

28th Annual Meeting of the Society for Virology, 14-17 March 2018, Würzburg, Germany

Rigogliuso G, Biniossek ML, Meese E, Schilling O, Mayer J. A Human Endogenous Retrovirus-encoded Protease Potentially processes numerous cellular proteins.

12. Acknowledgements

I would like to thank all my friends, in particular my former flatmates Nicole, Jonas and David for providing me a roof and a lot of affection from the first moment I set foot in Germany. I'd also like to thank my current flatmates (and almost siblings) Amir and Jenny for making me feel every day at home, filling my days with joy and craziness.

A special thank you to my dear Ann-Kathrin, a constant source of love and inspiration in my life. I also say thank you to her beautiful family, which I now feel part of.

Last but not least, I thank my parents for being always present, believing in me without ever hindering my dreams. Finally, many thanks to my brother and sister for lifting me up every time I fell, encouraging me to face life with a smile.