Auto-tethering as a selection mechanism for recognition of multimeric substrates by the AAA+ unfoldase ClpX

by

Aliaa H. Abdelhakim
B.Sc. Biochemistry
McGill University, 2001

Submitted to the Department of Biology in
Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY IN BIOCHEMISTRY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

SEPTEMBER 2008

Signature of Author _____
Department of Biology
September 2, 2008

Certified by _____
Tania A. Baker
E. C. Whitehead Professor of Biology
Thesis Supervisor

Accepted by _____
Tania A. Baker
E. C. Whitehead Professor of Biology
Chair, Committee for Graduate Students

CONTENTS

ABSTRACT

Auto-tethering as a substrate selection mechanism for recognition of multimeric substrates by the AAA+ unfoldase ClpX

By

Aliaa H. Abdelhakim

Submitted to the Department of Biology August 2008
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy in Biochemistry

The Clp/Hsp100 enzymes, which belong to the AAA+ family of ATPases, use their unfoldase activity to degrade and remodel multimeric substrates in the bacterial cell. The mechanical energy exerted by Clp/Hsp100 enzymes drives forward essential transitions in important biological processes. However, with a potentially destructive and energetically expensive enzymatic activity, mechanisms must be employed to ensure that Clp/Hsp100 enzymes act on the desired substrate at the right time and in the right location.

The remodeling of stable complexes by Clp/Hsp100 enzymes must be directed toward the correctly assembled form of substrates in the cell, and therefore strategies must exist that guide Clp/Hsp100 enzymes to correctly distinguish between multimeric and monomeric forms of a substrate. In this work I explore how substrate multimerization modulates recognition by the enzyme, using the AAA+ unfoldase ClpX and its multimeric substrate the Mu transpososome. Phage Mu transposase tetramerizes in the cell to form the Mu transpososome, which mediates replicative transposition of the phage. After transposition is completed, the Mu transpososome forms an extremely stable tetramer that needs to be destabilized by ClpX to allow it to facilitate phage Mu genome amplification. How ClpX is guided to the correctly assembled stable transpososome is the subject of my work. I find that multimerization of the phage Mu transposase to form the tetrameric Mu transpososome exposes residues that make contact with the ClpX only in the context of the tetrameric complex. These unique contacts recruit ClpX to the stable transpososome with high affinity. The dual role of subunits in the transpososome in providing high affinity ClpX binding sites as well as ClpX substrate degradation signals is referred to in this work as auto-tethering. Additionally, I show that the N terminal domain of ClpX, which plays a role in substrate selection, is important in facilitating discrimination between different multimeric forms of MuA by ClpX.

ClpX destabilizes the tetrameric transpososome by unfolding only one of the subunits within the complex. However, it is not known which subunit within the transpososome is unfolded, nor is it clear whether it is the same or different subunits that facilitate high affinity binding of ClpX to the complex. I am currently performing experiments to determine the geometry of unfolding and auto-tethering using an altered specificity mutant of MuA, which binds to Mu DNA binding sites in the transpososome containing compensatory mutations. This work can shed light on the division of labor required to mediate auto-tethering in the transpososome as well as in other multimeric substrates of ClpX.

Thesis Supervisor: Professor Tania A. Baker
Title: E.C. Whitehead Professor of Biology, MIT

ACKNOWLEDGMENTS

I would like to thank the following people, without whom this work and my development as a scientist would not have been possible.

I would like to thank my advisor Tania Baker, for taking me into her lab when I needed a new home, for providing an excellent training environment where I could develop as a scientist, for her deep dedication to her students and for her encouragement in the face of adversity. Joining your lab has been my best decision in graduate school.

I would like to thank my co-advisor Bob Sauer, for his truly insightful and continual input and help with my projects over the years and for making SJ meetings a comfortable environment to share and discuss data, both good and bad.

I would like to thank my thesis committee for their help over the years: Uttam RajBhandary, Thomas Schwartz, Stephen Bell, David Jeruzalmi and Frank Solomon.
I would especially like to thank Frank Solomon for his help and encouragement, and for always listening. Getting through graduate school may well have been impossible without your help.

I would like to thank the Baker lab for being such a wonderful group of people. Your support and humour over the years have been invaluable. I hope that in the future I will be lucky enough to work with people as friendly and caring as you are.

Finally I would like to thank my family: My parents, my siblings and my husband. Words cannot describe how essential your love and support have been to me.

CHAPTER 1

INTRODUCTION: SUBSTRATE SELECTION MECHANISMS OF THE CLP/HSP100 UNFOLDASES

INTRODUCTION

The inside of a cell is an incredibly densely packed and complex environment, harboring many thousands of proteins that need to remain soluble and functional despite extreme molecular crowding effects. Protein interactions must be highly dynamic to respond to rapidly changing extracellular environments. It is therefore imperative to maintain the specificity of these protein interactions in order to support life. Keeping intracellular order has high entropic cost, and therefore much of the cell's energy is invested in ensuring that biochemical reactions are well-timed and well-located. Central to the highly regulated environment of the cell is the implementation of enzyme-substrate specificity. Cellular enzymes have evolved to use binding energy to specifically recognize specific structural elements within their substrates, and in this way, biological reactions can go forward in a productive fashion. Elucidating the rules of substrate specificity is therefore important to understanding how the cell orchestrates the myriad of biochemical pathways necessary for life.

As the cellular environment is dynamic and changing, so must be the structures of the proteins that populate it. Many reactions in the cell require significant conformational changes from the native, fully-folded structure of protein substrates to achieve the desired biological outcomes. This includes processes such as degradation of proteins, remodeling or disassembly of higher order complexes and translocation of protein substrates across intracellular membranes. Specialized remodeling enzymes exist in the cell to guide substrates through these significant structural changes in a way that is specific and which minimizes unwanted interactions with surrounding macromolecules.

One of the most studied families of such remodeling enzymes is the ATPases associated with diverse cellular activities (AAA+) family of proteins. These all operate by promoting conformational changes or remodeling of their substrates using energy derived from ATP hydrolysis (Sauer et al., 2004). The AAA+ proteins are defined by the presence of a domain known as the AAA domain, a 200-250 residue domain that binds and hydrolyzes ATP and promotes multimerization of the enzyme. Different AAA+ enzymes adopt a variety of multimeric architectures and catalyze many essential cellular functions in all kingdoms of life (Erzberger and Berger, 2006). The mechanical energy exerted by these enzymes drives forward many diverse biological reactions. These reactions include unfolding of substrates for regulated degradation, protein quality control, sporulation and competence, regulation of bacterial cell cycles, regulation of stress responses, thermotolerance, DNA replication, membrane fusion and disaggregation (Baker and Sauer, 2006; Davey et al., 2002; Hanson and Whiteheart, 2005; Jenal and Hengge-Aronis, 2003; Lazazzera and Grossman, 1997; Neuwald et al., 1999). In all these cases, AAA+ enzyme action on the substrate leads to significant and often irreversible structural changes in the substrate, resulting in unfolding, degradation, remodeling, disaggregation or disassembly. Therefore, activity of the AAA+ enzymes can be potentially destructive if left unchecked. Regulation of substrate selection at multiple levels is thus of great importance for the activity of AAA+ enzymes *in vivo*.

How has enzyme-substrate specificity evolved to guide AAA+ enzymes to their targets in the cell? This thesis work explores this question by examining the enzymatic action of the AAA+ unfoldase ClpX. ClpX belongs to bacterial Clp/Hsp100 family of enzymes, a subfamily of the AAA+ proteins for which substrate selection rules are most well understood. Most of these

enzymes facilitate substrate degradation using their unfoldase activity. Because their enzymatic action is inherently destructive, Clp/Hsp100 enzymes employ an arsenal of different mechanisms to keep their activity in check. These mechanisms include regulation at the level of direct binding of the substrate to enzyme, timing of exposure of degradation signals, and the use of adaptor and anti-adaptor proteins to enhance or inhibit recognition of substrates (Figure 1). For all reactions catalyzed by the Clp/Hsp100 enzymes, a combination of these substrate selection tools is used to recognize the desired substrate in a specific and energy-efficient manner.

In addition to supporting proteolytic activities, many Clp/Hsp100 enzymes must remodel extremely stable multimeric complexes to bring about a change in the function of these complexes and hence to drive biological processes forward. Remodeling enzymes must specifically recognize their substrates in the correctly assembled multimeric form and avoid nonproductive binding to substrate monomers, which contain the same binding peptides as the multimers. How this selection is achieved is of fundamental importance to AAA+ enzymes in both prokaryotic and eukaryotic cells, as many of them must act on correctly assembled protein-protein or protein-nucleic acid complexes to promote essential transitions in biological pathways. The work in this thesis explores substrate selection mechanisms that modulate Clp/Hsp100 enzyme binding to multimeric complexes, and specifically how substrate multimerization can create internal adaptor binding sites that function to enhance the affinity of ClpX for an assembled substrate complex. We refer to this substrate selection mechanism as "auto-tethering". Investigating the mechanistic basis behind auto-tethering will shed light on substrate selection strategies for a large number of multimeric ClpX substrates, such as the

tubulin-like FtsZ and the stationary phase DNA protecting enzyme Dps, as well as provide insight into how other remodeling AAA+ enzymes recognize and bind their substrates.

All mechanisms of substrate selection by Clp/Hsp100 enzymes, including auto-tethering, complement each other to provide a robust yet versatile regulatory repertoire that caters to the nature and function of AAA+ enzyme action in the cell. The goal of this introduction will be to provide a comprehensive view of how Clp/Hsp100 substrate selection mechanisms modulate enzyme function (Figure 1), and thus to put into perspective how auto-tethering fits as an important substrate selection tool for Clp/Hsp100 enzymes.

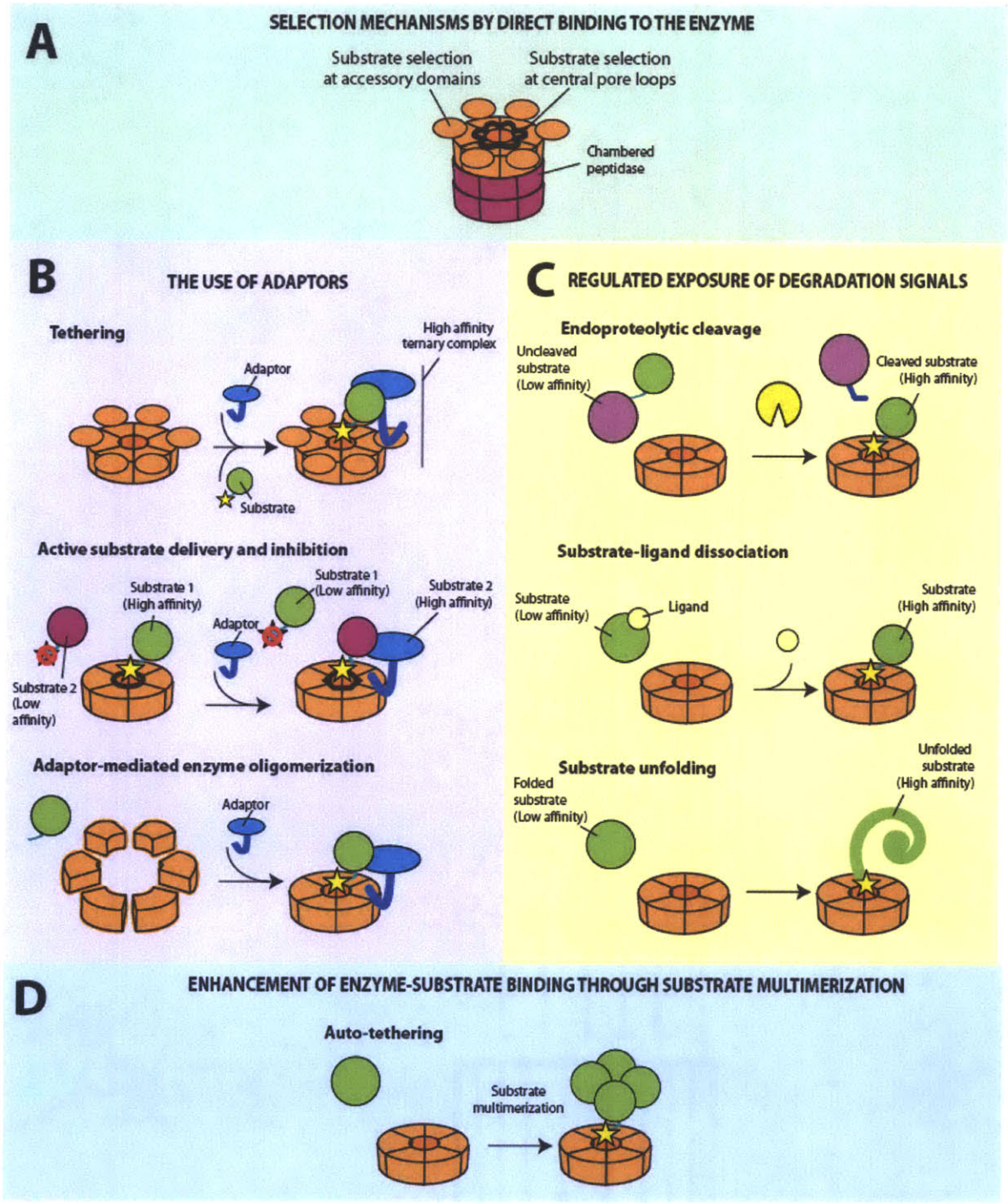


Figure 1. Clp/Hsp100 enzymes use a wide variety of substrate selection strategies. The Clp/Hsp100 unfoldase is depicted as an orange ringed hexamer. A star indicates that the degradation tag is efficiently recognized by the enzyme. In this introduction, mechanisms to achieve substrate specificity will be discussed at several levels: A. Clp/Hsp100 enzymes can discriminate between different classes of substrates at the level of the enzyme processing pore

and accessory domains, which act together to confer upon the enzyme a unique repertoire of degradation tag preferences. B. Adaptors mediate substrate binding to Clp/Hsp100 enzymes by a variety of mechanisms. Tethering occurs when the adaptor binds substrate and enzyme simultaneously to form a ternary complex, delivering the substrate with high affinity to the enzyme processing pore. Adaptors can also modulate substrate specificity by increasing the affinity of the enzyme for one class of substrate and decreasing the affinity for a different class. These "active" delivery adaptors may also promote allosteric conformational changes in the enzyme and/or substrate, as denoted by the change in the shape of the enzyme processing pore. Adaptors can also regulate degradation by promoting oligomerization of the Clp/Hsp100 enzyme, which is necessary for recognition of the substrate. C. Regulated exposure of substrate degradation signals control the time and location of substrate recognition by Clp/Hsp100 enzymes. This can occur through an endoproteolytic cleavage event which exposes N and C termini in the substrate, through dissociation of a ligand that exposes degradation signals in the substrate, or through unfolding of the substrate. D. This thesis explores auto-tethering as a further mechanism that is used in the toolbox of substrate selection strategies for Clp/Hsp100 enzymes. In this model, multimerization of the substrate promotes high affinity binding by active mechanisms that expose signals not present in the monomer form of the substrate.

Structural features of the bacterial Clp/Hsp100 enzymes

The structural features and domains of Clp/Hsp100 enzymes are geared toward achieving multiple goals, including ATP hydrolysis, enzyme multimerization and substrate selection. In this section, I introduce the structural aspects of the Clp/Hsp100 family that are relevant to substrate selection.

General architecture of the Clp/Hsp100 enzymes

The bacterial Clp/Hsp100 family of proteins all form hexameric rings that are stabilized by ATP and participate in energy-dependent proteolysis, disaggregation and remodeling of substrates in the cell (Figure 2) (Ito and Akiyama, 2005; Lee and Suzuki, 2008; Schirmer et al., 1996; Zolkiewski, 2006). All energy-dependent proteases in the bacterial cell, including the enzymes ClpX, ClpA, HslU, Lon and FtsH, have a similar architecture. Primarily, this structure contains a multimeric barrel-shaped peptidase, otherwise known as a chambered peptidase. To avoid indiscriminate degradation of proteins in the cell, access is restricted to the peptidase via a 10Å portal on either end of the chamber. Substrates therefore need to be unfolded to enter to chambered peptidase compartment for degradation (Figure 3). Substrate selection and binding is achieved by the Clp/Hsp100 enzymes, and ATP hydrolysis by these enzymes drives substrate denaturation and translocation into the chambered peptidase (Burton et al., 2003; Kenniston et al., 2003; Kim et al., 2000; Weber-Ban et al., 1999). The unfolding and peptidase functions can be present on the same or separate polypeptide chains. For example, subunits of the enzymes Lon and FtsH contain both the AAA+ and serine peptidase domains within the same polypeptide. The enzymes ClpX, ClpA and HslU bind to multimeric chambered

peptidase encoded by a separate gene. In the case of both ClpX and ClpA, the associated serine peptidase is the protein ClpP, and the resulting complexes are referred to as ClpXP and ClpAP, respectively. The Clp/Hsp100 family of proteins can also be subdivided further into classes depending on the number of AAA+ domains that the enzyme possesses. For example, Class I Clp enzymes such as ClpA and ClpB contain two AAA+ domains, whereas Class II enzymes such as ClpX contain only one. Despite such differences however, these enzymes function using similar mechanisms, as described below.

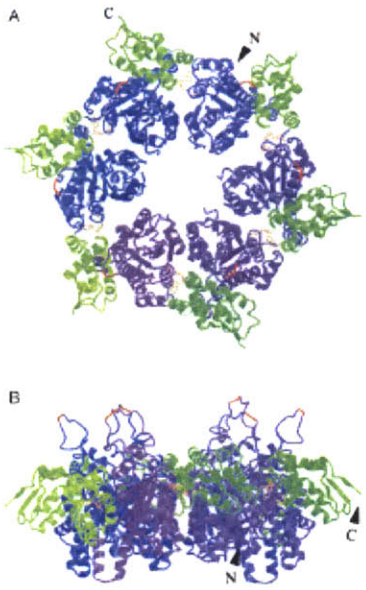


Figure 2. Clp/Hsp100 enzymes adopt a ring hexamer structure. Shown above is an example of the ring hexamer formation demonstrated by a modeled hexameric structure of ClpX. Figure from (Kim and Kim, 2003). A. Top view of ClpX hexamer; B. Side view of the ClpX hexamer. N – N terminal residue of one subunit in the hexamer; C – C terminal residue of the same subunit.

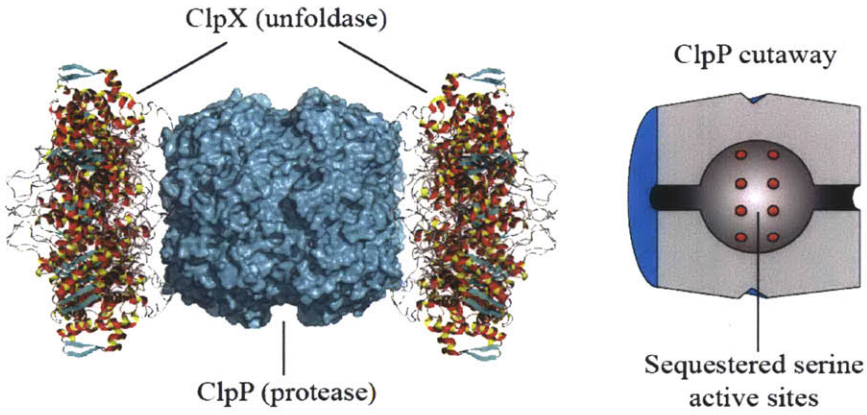


Figure 3. Two ClpX ring hexamers (side view) bound to the chambered serine peptidase ClpP. ClpP, a barrel-like complex consisting of two heptameric rings stacked back-to-back, can associate with one or two ClpX hexamers. A cartoon cutaway representation of ClpP shows the multiple active sites that are sequestered within the complex. Substrates must be unfolded by ClpX for access to the serine protease active sites. Structure from (Kim and Kim, 2003) and (Wang et al., 1997). Figure from Greg Hersch.

Hexamer pore and substrate binding loops

The central pore of the ringed hexamers of the Clp/Hsp100 family of enzymes is lined with amino acid loops which play an important role in substrate binding and processing (See for example Figure 5). Certain motifs located in the pore such as the "GYVG" motif (also known as pore-1 loop) are conserved among all Clp/Hsp100 enzymes and play an important role in substrate binding and selection, as discussed below. In addition to the conserved sequences in the central pore, some enzymes also possess pore loops that are specific to the subfamily of Clp/Hsp100 to which they belong. ClpX, for example, contains the "RKH" loop, found only in bacterial ClpX homologs, which plays a role in discriminating between different classes of ClpX substrates (Figure 5) (Farrell et al., 2007).

Accessory domains

In addition to AAA+ domains, unfoldases of the Clp/Hsp100 family possess "accessory" domains that aid in oligomerization of the enzyme, ATPase activity (Lo et al., 2001; Wojtyra et al., 2003), complex formation with the chambered protease (Hinnerwisch et al., 2005b), feeding substrate into the enzyme (Thibault et al., 2006a) and substrate selection (Barnett et al., 2005; Siddiqui, 2004; Wojtyra et al., 2003). Many of these accessory domains are located at the N-terminus of the Clp/Hsp100 enzyme, and when purified as isolated domains form monomers or dimers. ClpA and ClpX, for example, contain N terminal domains that *in vitro* are monomeric and dimeric, respectively (Donaldson et al., 2003; Guo et al., 2002a; Guo et al., 2002b; Xia et al., 2004; Zeth et al., 2002). Some Class I Clp/Hsp100 enzymes contain accessory domains located in between the two AAA+ domains. For example, ClpB, in addition to possessing an N-terminal

domain, carries a middle or "M" domain located in between the first and second AAA+ domains, which aids in substrate disaggregation (Haslberger et al., 2007). In a similar fashion, ClpC from *B. subtilis* contains a middle "linker" domain which plays a role in binding to adaptor proteins that aid in substrate selection (Kirstein et al., 2006).

MECHANISMS OF SUBSTRATE SELECTION BY THE CLP/HSP100 ENZYMES

Substrate recognition by direct binding to the enzyme

The bacterial Clp/Hsp100 family of enzymes recognizes substrates via exposed peptide sequences often located at the N or C terminus of their substrates; these sequences are referred to as recognition or degradation tags (Figure 4). Recognition of degradation tags is best understood for ClpX. A proteomic study revealed that ClpX recognizes at least five different classes of recognition tags – two classes located at the C-terminus of substrates (C motifs 1 and 2) and three classes located at the N-terminus (N motifs 1,2 and 3) (Flynn et al., 2003). Of these classes, the binding determinants are most well characterized for the ssrA tag (AANDENYALAA-$^{COO-}$), which is representative of the C terminal motif-1 class of degradation tags. ssrA is a peptide that is added co-translationally to the C terminus of polypeptides on stalled ribosomes (Gottesman et al., 1998; Keiler et al., 1996). ClpX recognizes the ssrA tag primarily via the C terminal di-alanine in the peptide and the free C terminal carboxyl group (Flynn et al., 2001).

In addition to directly recognizing binding determinants in tags, some Clp/Hsp100 enzymes also use other strategies to select substrates. For example, ClpA, in addition to recognizing specific degradation tags, can recognize unfolded proteins that do not contain degradation tags as well as N end rule substrates, whose stability in the cell is determined by

the identity of the first N-terminal residue (Hoskins et al., 2000a; Hoskins et al., 2000b; Tobias et al., 1991; Varshavsky, 1992). Lon can recognize specific sequence motifs located at the termini of a substrate (Shah and Wolf, 2006), but it also possesses the ability to recognize degradation signals interior to the protein (E. Gur, personal communication).

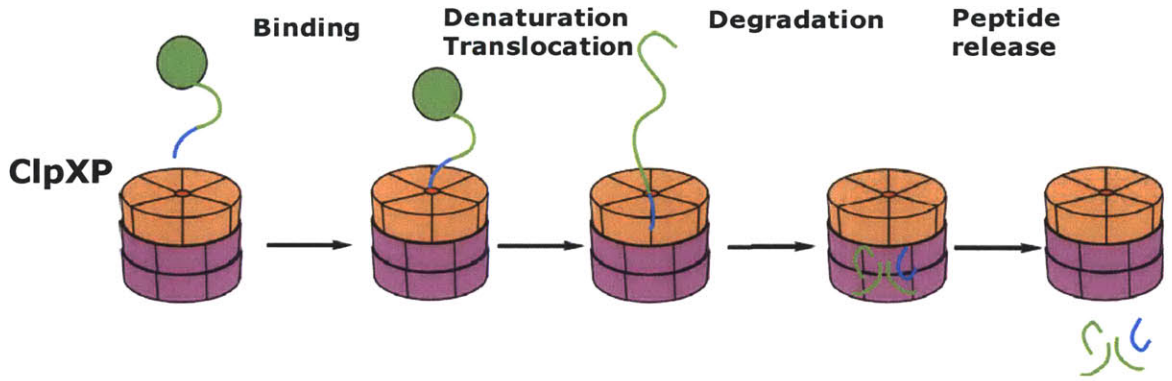


Figure 4. Schematic of a cycle of degradation of a substrate, catalyzed by the proteolytic machine ClpXP. $ClpX_6$ is shown in orange, and $ClpP_{14}$ is shown in purple. ClpX binds to the substrate via a degradation tag, located at the N or C terminus of the substrate (shown in blue). Once stably bound, ClpX unfolds the substrate using energy derived from ATP, and translocates it into ClpP, where the serine active sites act to degrade the substrates into small peptides. Peptides are then released from the ClpXP machine, thus completing the degradation reaction. Similar reactions are catalyzed by other Clp/Hsp100 enzymes, including ClpAP and HslUV.

Despite additional substrate selection mechanisms that may precede binding to the substrate tag, all Clp/Hsp100 enzymes must eventually make direct contact with the substrate for degradation, remodeling or disaggregation. How do Clp/Hsp100 enzymes discriminate between different substrates despite high levels of sequence conservation in the enzyme processing pores and high similarity between different substrate tags? Selection mechanisms exist to regulate these fundamental interactions at the level of the enzyme sequence and structure. In this part of the introduction, I will discuss how loops in the central pore of the ringed hexameric Clp/Hsp100 enzymes act to distinguish between different classes of substrates, as well as the role of accessory domains in cooperating with the enzyme processing pore to ensure high specificity of action by the Clp/Hsp100 enzymes.

Direct binding of substrates to Clp/Hsp100 central pore

Loops that line the central pore of the hexameric, ringed Clp/Hsp100 enzymes through which unfolded substrates are translocated play an important role in recognizing and binding substrates. Despite the conservation of multiple pore loop motifs, such as the GYVG motif, Clp/Hsp100 enzymes have the ability to discriminate between different substrates at the level of the enzyme pore. Studies investigating the roles that pore loop residues play in the substrate selection process of individual enzymes are beginning to shed light on how such highly similar enzymes achieve distinct substrate specificities.

ClpX: Studies have shown that residues in the GYVG motif located in the ClpX hexamer pore binds the five ClpX substrate tag classes differently. Mutation of the conserved Val154 to

phenylalanine severely inhibits recognition of ssrA-like recognition tags, but affects other classes of recognition tag to a much smaller extent (Siddiqui et al., 2004).

The ssrA tag has been used extensively as a model substrate with ClpX to dissect substrate selection at the level of the ClpX pore. The GYVG motif is not the sole binding determinant for ssrA binding in ClpX, and several recent studies have shown that two more pore loops are involved in recognizing ssrA: a pore loop present only in bacterial ClpX homologs, referred to as the "RKH loop", and the "pore-2" loop located toward the face of ClpX which faces ClpP (Figure 5) (Farrell et al., 2007; Martin et al., 2007, 2008). Transplanting the RKH and pore-2 loops into human ClpX, which does not normally bind to ssrA, allows this enzyme to efficiently recognize ssrA targets, confirming the importance of these two loops in the recognition of this class of substrate (Martin et al., 2008). The RKH loop recognizes the free carboxyl group at the C-terminus of the ssrA tag (Farrell et al., 2007). Interestingly, mutating the arginine in the RKH loop not only decreases affinity to ssrA-tagged substrates, but also dramatically increases the affinity of ClpX for classes of recognition tag that contain an overall positive charge. These observations suggest a model whereby the positively charged arginine in the loop acts to electrostatically attract the negative charge of the carboxyl group of ssrA at the expense of higher binding affinity for positively charged recognition tags.

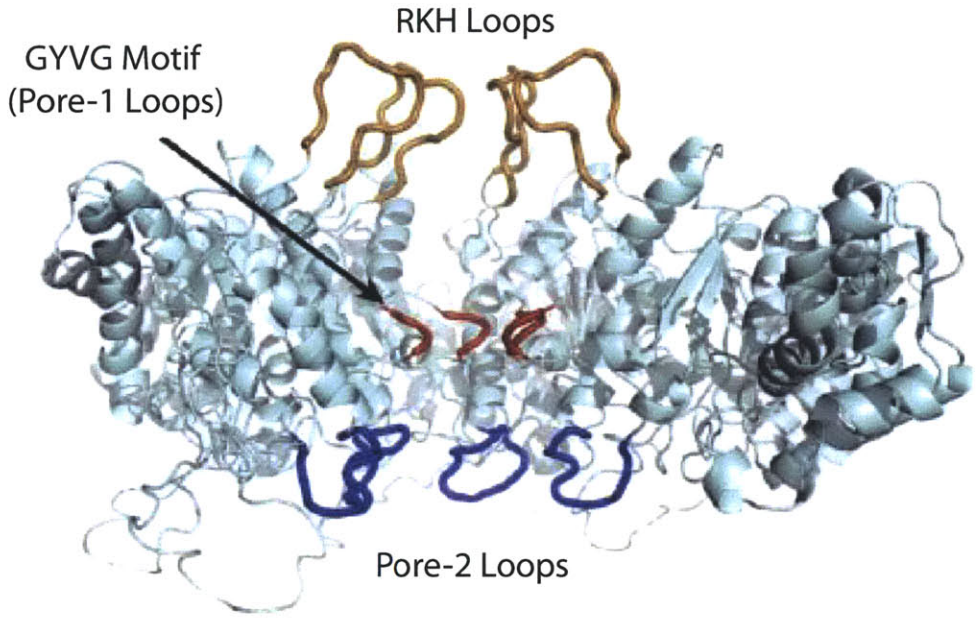


Figure 5. Location of ClpX-specific substrate binding loops within the ClpX hexamer. Two subunits within the hexamer have been removed for clarity. Figure from Martin et al. 2008.

ClpA: Although fewer substrates have been identified for ClpA relative to ClpX, recent analyses have shown that the pore residues of ClpA bind differentially to different classes of ClpA substrates (Hinnerwisch et al., 2005a). Like ClpX, ClpA recognizes ssrA substrates, although the molecular contacts that mediate the ClpA-ssrA interaction are different from those for ClpX-ssrA (Flynn et al., 2001; Gottesman et al., 1998). A recent study has shown that the ssrA tag binds directly to three distinct regions in the ClpA pore (Hinnerwisch et al., 2005a). This study also shows that other non-ssrA substrate classes require the same pore residues as ssrA for eventual substrate processing, although they may initiate binding with ClpA using different interactions.

ClpB: ClpB, the prokaryotic homolog of Hsp104 in eukaryotes, is a Clp/Hsp100 enzyme that is required for thermotolerance (Parsell et al., 1991; Sanchez and Lindquist, 1990; Weibezahn et al., 2005). ClpB functions *in vivo* and *in vitro* in conjunction with the DnaKJE system, and is thought to act by disaggregating and solubilizing important proteins in the cell upon heat shock. Binding studies using arrays of cellulose-immobilized peptides derived from multiple proteins show that ClpB prefers to bind to peptides containing aromatic and basic residues (Schlieker et al., 2004). This preference overlaps with that found for DnaK, and may be necessary for the cooperation of ClpB with the DnaKJE system. Moreover, the conserved residue Tyr251 in the GYVG motif plays an important role in the initial step of substrate binding.

ClpB does not associate with any ClpP-like peptidases and is not thought to function in the cell by promoting degradation like other Clp/Hsp100 enzymes. However, experiments using ClpB engineered to bind to ClpP (referred to as "BAP") show that ClpB functions, like ClpX and

other Clp/Hsp100 unfoldases, by unfolding substrates through its central pore (Weibezahn et al., 2004). Although specific substrates for ClpB were not identified, this study shows that ClpB binds specifically to only a subset of protein substrates during heat shock *in vivo*, as indiscriminate degradation of substrates by BAP was not observed.

Direct binding of substrates to accessory domains of Clp/Hsp100 enzymes

In addition to substrate specificity at the level of the enzyme pore, Clp/Hsp100 enzymes also bind substrates via accessory domains located at the N terminus or in the interior of the protein. For example, ClpX possesses an accessory domain consisting of 61 residues, located at the N terminus of the enzyme. This domain adopts a treble clef zinc finger structure, and forms a dimer when purified as the isolated domain (Donaldson et al., 2003; Wojtyra et al., 2003). In the context of the ClpX hexamer, it is thought to form a trimer of dimers. Although consensus sequences for the N domain are not well understood, it is known that the N domain possesses a general preference for basic and hydrophobic residues in a wide variety of sequence combinations, consistent with its preference for binding ClpX recognition tags that contain an overall positive charge (Thibault et al., 2006b).

The ClpX N domain facilitates the binding of certain classes of ClpX substrates to the enzyme. Deleting the N domain, for example, decreases efficiency of binding to ClpX substrates such as the transposase MuA and the phage λ O replication enzyme (Wojtyra et al., 2003). Previous studies have also shown that the N domain plays differential roles in recognizing the five ClpX recognition tag classes. Whereas the ssrA tag does not require the N domain for full binding to ClpX, the remaining four classes of recognition tags require the N domain to

maximum affinity binding (Siddiqui et al., 2004). Interestingly, the non-ssrA tag classes seem to bind with higher affinity to ClpX in the context of a dimer than a monomer in an N domain-dependent manner, suggesting that the N domain may play a general role in recognizing substrates when they are in a multimeric form (A. Abdelhakim, unpublished results) (Siddiqui, 2004). This ability of the N domain to aid in the discrimination between different multimer forms of substrates may be an important feature for mediating auto-tethering, as is discussed in Chapter 2.

ClpA also uses its N domain, which is a monomer in the context of the ClpA hexamer, for recognition of at least a subset of its substrates. For example, the ClpA-specific RepA recognition tag, derived from the phage P1 RepA replication protein, must first bind to the N domain of ClpA for recognition (Hinnerwisch et al., 2005a). Other studies have shown than deleting the N domain of ClpA results in defects in recognition of the ssrA tag, although this has been attributed to weaker association with ClpP, rather than a direct binding defect between ssrA and the ClpA N domain (Hinnerwisch et al., 2005b). The ClpA N domain may also play a role in stabilizing enzyme-substrate interactions while unfolding of the substrate is being initiated (Hinnerwisch et al., 2005a).

Much less is known about how accessory domains in other Clp/Hsp100 enzymes bind to substrates. It is, however, becoming apparent that many of these domains play some role in substrate recognition and selection. For example, although there is some controversy regarding its specific roles, some studies have suggested that the ClpB N domain plays a role in binding large aggregates in preference to smaller ones (Barnett et al., 2005).

Substrate recognition by adaptors

In addition to binding directly to substrates, substrate recognition by Clp/Hsp100 enzymes can be controlled with the help of an adaptor protein. This type of interaction provides an additional level of regulation of substrate selection for Clp/Hsp100 enzymes, thus ensuring that high priority substrate unfolding is executed with high affinity and specificity. Adaptors modulate the enzyme-substrate interaction by increasing or decreasing the affinity of the substrate for the enzyme (Ades, 2004; Baker and Sauer, 2006; Dougan et al., 2002a). In most cases, the adaptor protein must bind simultaneously to the substrate and to an accessory domain of the Clp/Hsp100 enzyme to deliver the substrate to the enzyme processing pore. Different adaptors function using a diverse number of mechanisms, as described below.

Adaptors that function by tethering the substrate

Adaptors can increase the affinity of a substrate for the enzyme by simultaneously binding the substrate and the enzyme. Formation of this ternary complex increases the "effective concentration" of the substrate, and in this way the enzyme and the substrate interact with an affinity higher than in the absence of the adaptor. This mechanism of adaptor function is commonly referred to as "tethering" and is by far the most well understood mode of adaptor function (Sauer et al., 2004; Ades, 2004; Baker and Sauer, 2006; Dougan et al., 2002). Tethering as a mechanism of substrate delivery has been described in most detail for the enzyme ClpX, which binds adaptors via its N domain.

The ClpX-specific adaptor SspB is the most well understood adaptor that uses tethering to enhance the affinity of substrates for ClpX (Figure 6) (Levchenko et al., 2000). SspB is a

homodimer and consists of a stably folded N terminal substrate binding and dimerization domain, followed by a flexible unstructured C terminal region containing ClpX-interacting residues located at the extreme C terminus, known as the ClpX binding (XB) motif (Levchenko et al., 2003; Song and Eck, 2003). SspB enhances the affinity of substrates for ClpX by simultaneously binding a peptide sequence in the substrate via its substrate binding domain and the N domain of ClpX via its XB motif (Figure 6) (Bolon et al., 2004b; Dougan et al., 2003; McGinness et al., 2007; Wah et al., 2003). To date, only two natural substrates have been confirmed to require SspB for enhanced binding to ClpX: these are ssrA-tagged substrates and NRseA, the N-terminal cleavage product of the anti-sigma factor RseA (Flynn et al., 2004; Levchenko et al., 2000).

Interactions that mediate the ternary complex formation and substrate delivery by SspB of ssrA substrates have been well characterized (Bolon et al., 2004a; Bolon et al., 2004b; Flynn et al., 2001; Hersch et al., 2004; Martin et al., 2008; McGinness et al., 2007; Park et al., 2007; Wah et al., 2003). Interestingly, the peptide sequences recognized by SspB in ssrA and NRseA are not similar, and even bind the SspB substrate binding domain in opposite orientations (Levchenko et al., 2005). Despite these different substrate binding modes, both ssrA and NRseA are delivered to ClpX via tethering mediated by SspB.

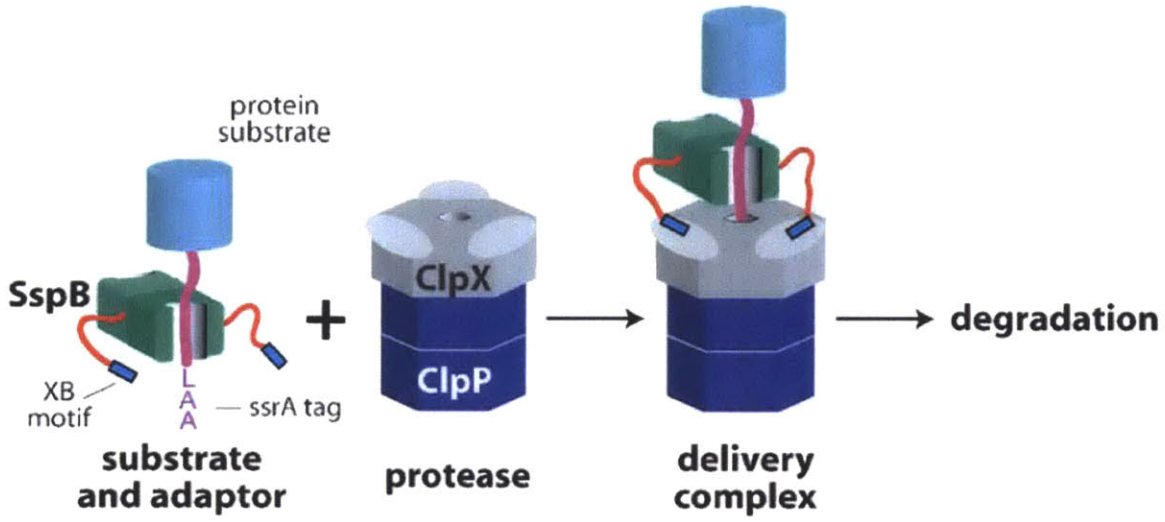


Figure 6. Schematic of a tethering reaction mediated by the adaptor SspB and the proteolytic machine ClpXP. SspB binds to the ssrA substrate via its substrate binding domain (shown in green) and to the N domain of ClpXP via the XB motif (shown in blue; ClpX N domain shown as ovals). The resulting tethered complex delivers the substrate with very high affinity (K_D 15nM; K_D for ssrA binding to ClpXP without SspB is 1μM). Figure from (McGinness et al., 2007).

Although not as well characterized as SspB, two other ClpX-specific adaptors, UmuD and RssB, are also thought to deliver substrates to ClpX by tethering. RssB is a response regulator that delivers the starvation and stationary phase-specific sigma factor, σ^S, to ClpXP for degradation during bacterial exponential phase and during nutritional excess (Becker et al., 1999; Klauck et al., 2001; Muffler et al., 1996; Pratt and Silhavy, 1996; Zhou et al., 2001). Although less is known about the peptide sequences in RssB that bind ClpX, it is known that RssB needs to bind both ClpX and σ^S simultaneously to deliver the substrate to ClpXP with high affinity (Studemann et al., 2003). Additionally, there are multiple layers of regulation that modulate the delivery of σ^S by RssB to ClpXP. For example, at low substrate and adaptor concentration, RssB needs to be phosphorylated to bind efficiently to σ^S and deliver it for degradation (Zhou et al., 2001). This requirement for phosphorylation of RssB can be overcome at sufficiently high substrate and adaptor concentrations (Ebrahim, 2007).

UmuD is a component of DNA polymerase V, the error-prone polymerase in bacteria (Gonzalez and Woodgate, 2002). UmuD is induced upon DNA damage and is inefficiently processed at its N terminus to form UmuD'. UmuD' is the form of UmuD that is active in translesional DNA synthesis, and must be cleared from the cell when the DNA damage response is resolved. To this end, UmuD binds to UmuD' to form the heterodimer UmuD/D', and delivers UmuD' to ClpXP for degradation (Frank et al., 1996; Gonzalez et al., 2000). This reaction occurs because UmuD binds to the ClpX N domain via an XB-like motif located within the residues that have been cleaved in UmuD' (Gonzalez et al., 2000; Neher et al., 2003b). In this way, UmuD acts as an adaptor and tethers the UmuD' substrate to delivers it with high affinity to ClpXP. Within the UmuD/UmuD' heterodimer, UmuD does not get degraded by ClpXP, however some

degradation of UmuD can occur in the context of a UmuD/UmuD homodimer, supporting a model where UmuD can become its own adaptor (Neher et al., 2003b).

Adaptors that function by "active" or allosteric mechanisms

In addition to increasing the affinity of substrates for their enzyme, adaptors can also alter substrate specificity by increasing affinity of one class of substrates for the enzyme, while simultaneously decreasing the affinity of a different class. ClpS, a ClpA-specific adaptor, functions in this way. ClpS binds as a monomer to the N domain of ClpA and inhibits recognition of ssrA tagged substrates as well as ClpA autodegradation. However, by binding to ClpA, ClpS increases the affinity of N-end rule and aggregated substrates for ClpA, thus altering ClpA substrate preference (Dougan et al., 2002b; Erbse et al., 2006; Wang et al., 2007). ClpS can switch the substrate preferences of ClpA by a simple change in the binding stoichiometry of ClpS to ClpA, with the maximum stoichiometry being one ClpS molecule to one ClpA subunit (Hou et al., 2008). There may be mechanisms that make use of these changes in stoichiometry *in vivo*, as it has been demonstrated in *E. coli* that intracellular ClpS to ClpA ratios change as cells transition from exponential into stationary phase (Farrell et al., 2005).

ClpS consists of a flexible N terminal linker followed by a stably folded domain (Guo et al., 2002a; Zeth et al., 2002). Although ClpS increases the affinity of N end rule substrates by binding both substrate and enzyme simultaneously, it does not deliver the substrate to ClpA by simple tethering. Deletion of the flexible ClpS N terminal linker does not affect binding of substrate or enzyme to ClpS, however this truncated version of ClpS fails to deliver N end rule substrates for degradation (Hou et al., 2008). Additionally, the N terminally truncated ClpS also

affects the ATPase rates of ClpA differently from full-length ClpS. This observation eliminates the simple model whereby ClpS "passively" tethers N end rule substrate to ClpA, and suggests instead a mechanism whereby ClpS actively or allosterically delivers N end rule substrates for degradation by ClpA with high affinity.

Mechanisms also exist whereby degradation is activated by an adaptor-like molecule through allosteric conformational changes promoted in a substrate. For example, the Mu repressor (Rep), a multimeric protein involved in the development of lysogeny in phage Mu and containing an intrinsic C terminal ClpXP tag, is on its own a poor ClpXP substrate (Krause and Higgins, 1986; Welty et al., 1997). Rep can be targeted for degradation by ClpXP by a dominant negative form of the Rep protein, known as Vir, which contains a frameshift mutation producing an altered C terminal peptide sequence (Geuskens et al., 1992; Welty et al., 1997). Interestingly, the targeting activity does not require the N domain of ClpX, and therefore Vir does not target Rep for degradation by a simple tethering mechanism (Marshall-Batty and Nakai, 2008b). Instead, Vir is thought to deliver Rep for fast degradation by ClpXP via a "trans-targeting" mechanism, inducing a conformational change which results in high local flexibility at the Rep C terminus, which in turn makes Rep a better substrate for ClpXP (Marshall-Batty and Nakai, 2003, 2008a). How Vir induces high flexibility in the C terminus of Rep, and how this translates into making Rep a better ClpXP substrate is yet to be determined.

Adaptors that function by promoting oligomerization of the Clp/Hsp100 enzyme

In addition to increasing or modulating affinities of substrates for their enzymes, some adaptors function as essential components of proteolytic machines in a much more binary "on-

off" mechanism. In this model, the enzyme is inactive in the absence of the adaptor and becomes active for unfolding and degradation only upon binding the required adaptor. The Clp/Hsp100 enzyme ClpC from the soil bacterium *Bacillus subtilis* functions in this manner. Like ClpX and ClpA, ClpC can bind to the ClpP to form a proteolytic machine known as ClpCP. ClpCP is involved in protein quality control and the regulation of developmental processes in *B. subtilis* including sporulation, competence and stress responses (Krause and Higgins, 1986; Kruger et al., 2000; Kruger et al., 2001; Turgay et al., 1998). ClpC can only form ringed hexamers in the presence of its adaptor proteins and therefore degradation of substrates can only occur in the presence of their respective adaptor (Kirstein et al., 2007; Kirstein et al., 2006). Like ClpXP adaptors, ClpCP adaptors must bind the enzyme accessory domains and the substrate simultaneously to mediate substrate delivery, and therefore may also play an additional tethering role (Kirstein et al., 2007; Kirstein et al., 2006; Persuh et al., 1999).

Several ClpCP substrates have been identified that absolutely require the function of an adaptor for degradation. For example, the transcription factor ComK, which is the "master regulator" in the development of competence in *B. subtilis*, is degraded by ClpCP when not needed and only in the presence of the adaptor MecA (Kirstein et al., 2006; Schlothauer et al., 2003). Another ClpCP substrate, CtsR, a transcriptional repressor that is encoded within and regulates the *clpC* operon, requires the phosphorylated form of the tyrosine kinase McsB to act as an adaptor and promote its degradation (Kirstein et al., 2007; Kirstein et al., 2005). Interestingly, both the MecA and McsB adaptors are degraded by ClpCP in the absence of substrate. This suggests an auto-regulatory mechanism whereby the protease is activated by the adaptor only in the presence of its cognate substrate, thereby conserving cellular energy

and resources when not needed (Kirstein et al., 2007; Kirstein et al., 2006; Schlothauer et al., 2003).

Adaptors that function by an unknown mechanism

Not surprisingly, there are adaptor-like proteins for Clp/Hsp100 enzymes that utilize molecular mechanisms that do not fall within any of the classes described above and for which a molecular mechanism has yet to be fully elucidated. Such adaptors include RcdA in *Caulobacter crescentus*, which is required for degradation of the master regulator of the cell cycle CtrA (McGrath et al., 2006). In *C. crescentus*, ClpXP is localized to the poles of stalk cells, where degradation of CtrA occurs, allowing for the G1 to S transition (Jensen et al., 2002; Ryan and Shapiro, 2003). RcdA is required to localize CtrA to the poles and to activate its degradation by ClpXP. However, CtrA alone is a substrate for ClpXP *in vitro*, and addition of RcdA to the reaction has no effect on the rate of degradation by ClpXP, ruling out its function as a conventional adaptor (Chien et al., 2007). Therefore, the molecular mechanisms that underlie the necessity of RcdA *in vivo* remain to be established.

Inhibition of adaptor function: Anti-adaptors

A new class of degradation regulators, known as anti-adaptors, has recently been shown to inhibit degradation of substrates by directly binding to adaptors and inhibiting delivery of substrate to the enzyme. This is seen for example with degradation of the starvation-specific sigma factor σ^S. σ^S on its own is a poor ClpXP substrate, and as mentioned above requires the function of the phosphorylatable adaptor RssB for delivery. RssB is responsible for continually

delivering σ^S for degradation during exponential phase. Upon entry into stationary phase however, RssB levels do not decrease. This led to the discovery of factors responsible for the stabilization of σ^S upon entry into stationary phase. IraP (Inhibitor of RssB activity under phosphate starvation) stabilizes σ^S by binding directly to RssB and inhibiting delivery of the substrate by the adaptor to ClpXP (Bougdour et al., 2006). Different inhibitors of RssB function in this way and are specific to different types of cell starvation or stress: IraP inhibits RssB upon phosphate starvation, whereas IraM and IraD inhibit activity of RssB under conditions of magnesium starvation and DNA damage, respectively (Bougdour et al., 2008). The exact molecular mechanism of inhibition by the Ira proteins has not yet been elucidated.

Another example of proteins that function as anti-adaptors is illustrated with the degradation of the competence regulator ComK by ClpCP in *B. subtilis*. ComK is continually degraded by ClpCP and depends on the adaptor MecA for delivery, as described above. However, upon entry into stationary phase, the small protein ComS, which is synthesized during quorum response, binds to MecA, releasing ComK and allowing it to engage in transcription of genes required for bacterial competence (Ogura et al., 1999; Turgay et al., 1998; Turgay et al., 1997). ComS displaces ComK from MecA by direct competition for the adaptor, binding to MecA via a peptide sequence similar to the MecA-interacting peptide sequence of ComK (Prepiak and Dubnau, 2007). Interestingly, ComS itself is delivered by MecA and degraded by ClpCP, potentially as a homeostatic mechanism within the cell (Turgay et al., 1998).

Regulated exposure of degradation signals

Multiple substrate selection mechanisms function by regulating the timing and location of exposure of degradation tags in the desired substrate. For example, endoproteolytic cleavage of substrates can expose new N and C termini that are recognized by Clp proteases as a degradation signal. This mode of regulation is seen with the LexA repressor, where upon DNA damage signals, RecA-stimulated autocleavage of LexA produces an N-terminal domain fragment containing an ssrA-like degradation tag at the C terminus (Neher et al., 2003a). In this way, the repressor is degraded and transcription of SOS genes can ensue. Another example is the endoproteolytic cleavage of the anti-sigma factor RseA, which under non-stress conditions sequesters the σ^E sigma factor required for the transcription of genes involved in the extracytoplasmic stress response (Flynn et al., 2004). Like for LexA, endoproteolytic cleavage of RseA triggered by the stress response results in an N terminal domain with a new C-terminal degradation tag (termed NRseA), allowing ClpXP to degrade this domain and release σ^E. In addition to the endoproteolytic cleavage event, NRseA is delivered to ClpXP by the adaptor molecule SspB, as described below (Levchenko et al., 2005).

Other mechanisms of exposing degradation signals include the dissociation of a protein, nucleic acid or small ligand, allowing the substrate to become recognized by the protease. Examples of this include degradation of the phage Mu transposase, MuA, by ClpXP. In this example, MuA contains a C-terminal degradation tag that can be "masked" by the phage Mu activator and target immunity protein MuB. MuB and ClpX bind to overlapping regions in the C terminal domain of MuA, thus ClpXP can only efficiently degrade MuA in the absence of MuB

(Levchenko et al., 1997). Another example is the degradation of zinc-binding transcriptional regulator ZntR by ClpXP and Lon (Pruteanu et al., 2007). ZntR controls transcription of genes involved in zinc export in bacteria (Outten et al., 1999). Recognition of ZntR by ClpXP and Lon is inhibited by binding to zinc and DNA, and therefore high intracellular zinc concentrations prevent the degradation of ZntR, allowing the expression of zinc export genes in the cell. However, during low intracellular zinc concentrations, dissociation of ZntR from zinc or DNA allows this transcriptional regulator to become a good degradation substrate. In this way, the cell ensures that ZntR is degraded only when it is no longer needed.

REMODELING OF STABLE COMPLEXES BY AAA+ UNFOLDASES

In addition to catalyzing degradation and disaggregation reactions, AAA+ unfoldases can also use energy from ATP to remodel stable protein complexes to drive biological reactions forward (Burton and Baker, 2005). The primary functional outcome of the remodeling reaction by Clp/Hsp100 enzymes is to bring about a change in the structure of the stable multimeric substrate that results in a change in its function, and not degradation *per se* (Figure 7). Examples of remodeling by AAA+ enzymes include, for example, the action of the AAA+ enzyme NSF on stable SNARE complexes to drive forward membrane fusion reactions. SNAREs are molecules that are required for membrane fusion in eukaryotic cells. Upon completion of the fusion reaction, SNAREs form very stable complexes that are resistant to high temperatures and harsh detergents. To complete the membrane fusion reaction and recycle free SNARE molecules, hexameric NSF uses its adaptor α-SNAP to disassemble the stable SNARE complexes, thus freeing the subunits for another round of membrane fusion (Jahn et al., 2003; Whiteheart

et al., 2001). Similar reactions include the remodeling of the inactive dimeric RepA replication protein from bacteriophage P1 by ClpA to form replication-active monomers, and the severing of microtubules by the AAA+ enzyme katanin (McNally and Vale, 1993; Pak and Wickner, 1997).

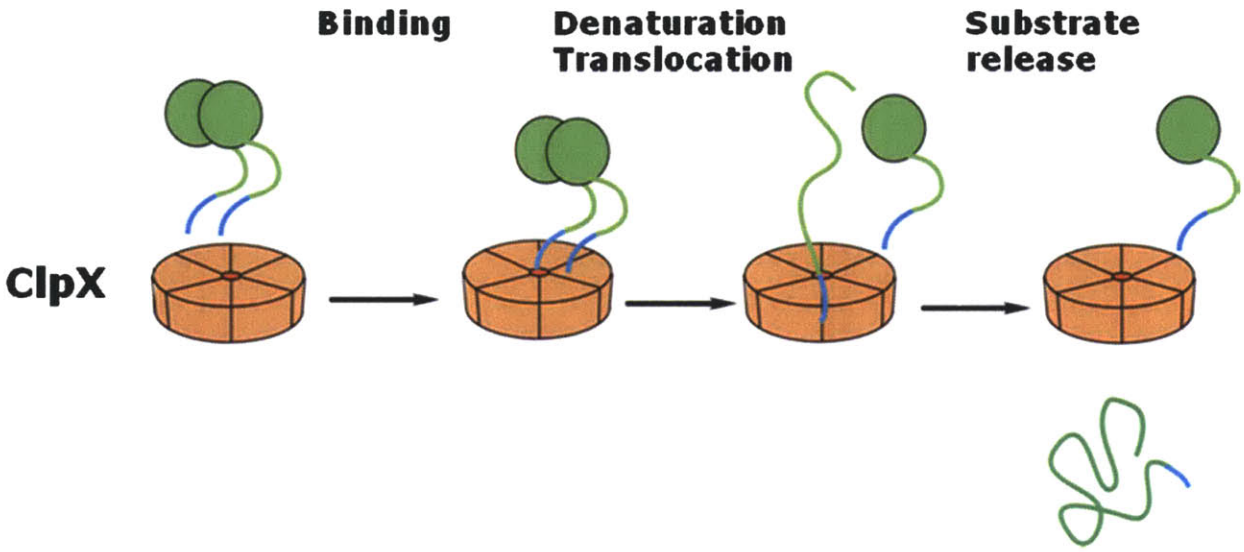


Figure 7. Schematic of a disassembly reaction by ClpX. In this figure, $ClpX_6$ (in orange) is shown disassembling a dimeric substrate by binding to the degradation tag on one subunit and unfolding it. In this way, the dimer is remodeled into its constituent monomers. The unfolded monomer can in some cases refold into its native structure (Burton and Baker, 2003).

Remodeling of the Mu transpososome by ClpX: Paradigm for disassembly

The most well understood substrate for remodeling by a Clp/Hsp100 unfoldase is the tetrameric form of MuA transposase known as the Mu transpososome, which comes from the phage Mu. Phage Mu is a bacteriophage that amplifies its genome in the host via replicative transposition (Figure 8). MuA catalyzes the transposition reaction by binding to Mu DNA sites at the terminal repeats of the Mu genome (the "left" and "right" sides of the genome) and forming a sequential series of tetrameric complexes known as transpososomes (Surette et al., 1987). The transpososomes catalyze the DNA cleavage and joining reactions necessary for transposition, resulting in an exceedingly stable transpososome referred to as the strand transfer complex. Transposition catalyzed by MuA results in the formation of forked DNA intermediates, to which phage Mu subsequently recruits the host DNA replication machinery for amplification of its genome (Nakai et al., 2001; Nakai and Kruklitis, 1995).

Although the strand transfer complex is the last transpososome complex in the DNA transesterfication reaction catalyzing transposition, it is also the most stable complex and, if not destabilized, becomes a hindrance to the replication machinery of the host. ClpX is essential for the replication of phage Mu in the host bacterial cell. It is recruited by the phage to remodel the strand transfer complex into a less stable form; the remodeled fragile complex, in turn, recruits the host replication machinery to complete amplification of the Mu genome (Jones et al., 1998; Kruklitis et al., 1996; Levchenko et al., 1995; Nakai et al., 2001; Nakai and Kruklitis, 1995). It is the remodeling function of ClpX rather than degradation that is necessary to drive forward replicative transposition, as deleting ClpX inhibits replication of phage Mu *in vivo* almost completely, but deleting ClpP has little or no effect (Mhammedi-Alaoui et al., 1994).

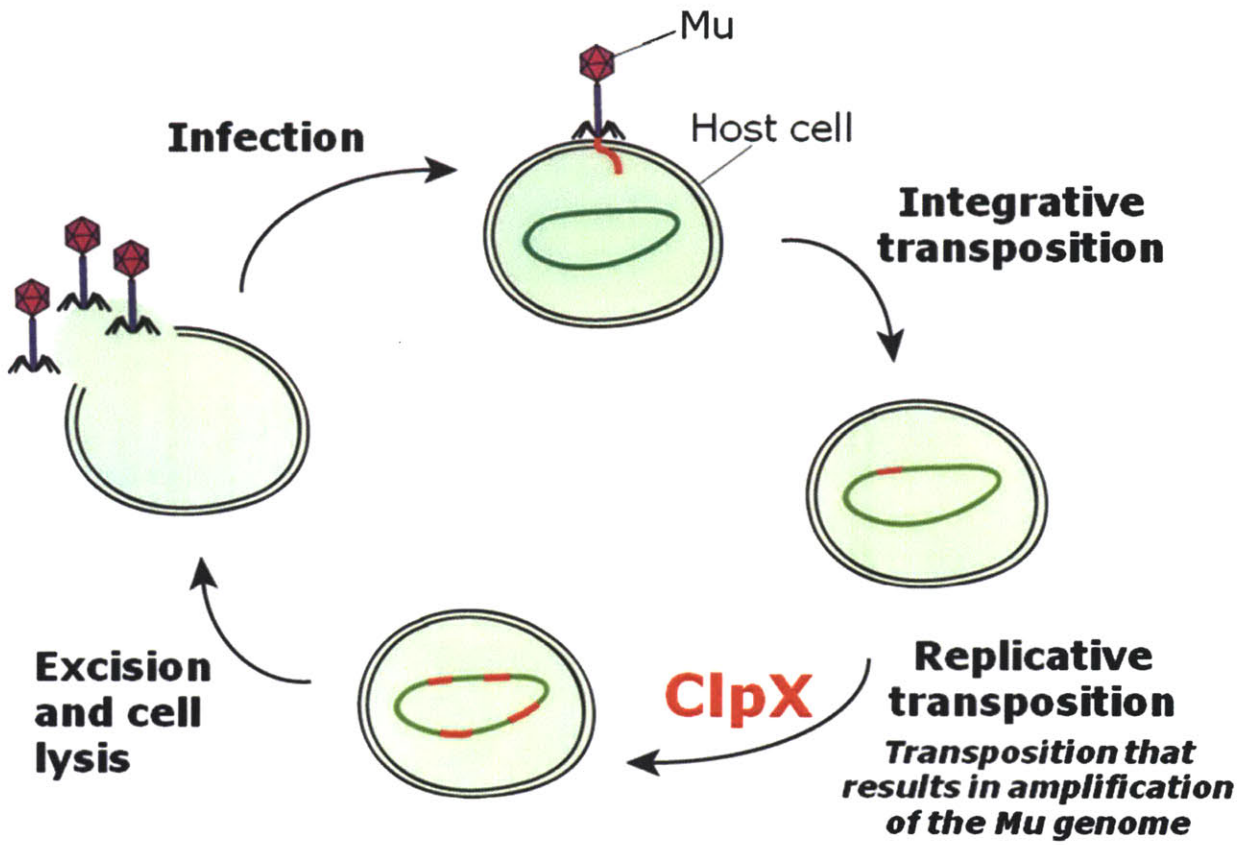


Figure 8. Simplified representation of the life cycle of phage Mu. Phage Mu infects its host and integrates into the genome via a process known as integrative transposition. After integration, replicative transposition ensues, which results in more than 100 copies of the Mu genome *in vivo*. ClpX is essential for the replication of phage Mu in the host.

MuA is a 75 kDa protein consisting of three major folded domains and belongs to the DDE family of recombinases, which includes HIV integrase and RAG recombinase (Brandt and Roth, 2004; Rice and Baker, 2001). The transposition reaction leading to the formation of the Mu transpososome can be catalyzed *in vitro* using supercoiled DNA containing MuA binding sites. In the absence of Mu DNA binding sites *in vitro*, MuA remains a monomer (Kuo et al., 1991). Transpososomes *in vitro* can be remodeled by ClpX or ClpXP to form complexes known as strand transfer complex II or fragile complexes (Burton et al., 2001; Jones et al., 1998;

Levchenko et al., 1995). The fragile complexes are relatively unstable and dissociate upon native agarose gel electrophoresis, and therefore the remodeling reaction by ClpX *in vitro* is often referred to as a disassembly reaction (Burton and Baker, 2003; Levchenko et al., 1995). ClpXP can also recognize and degrade the monomer form of MuA *in vitro* (Levchenko et al., 1997). Both reactions, the disassembly of the Mu transpososome by ClpX or ClpXP and the degradation of monomer MuA by ClpXP, require the binding of ClpX to the MuA ClpX-specific recognition tag located at the extreme C terminus of the transposase (Figure 9) (Levchenko et al., 1997).

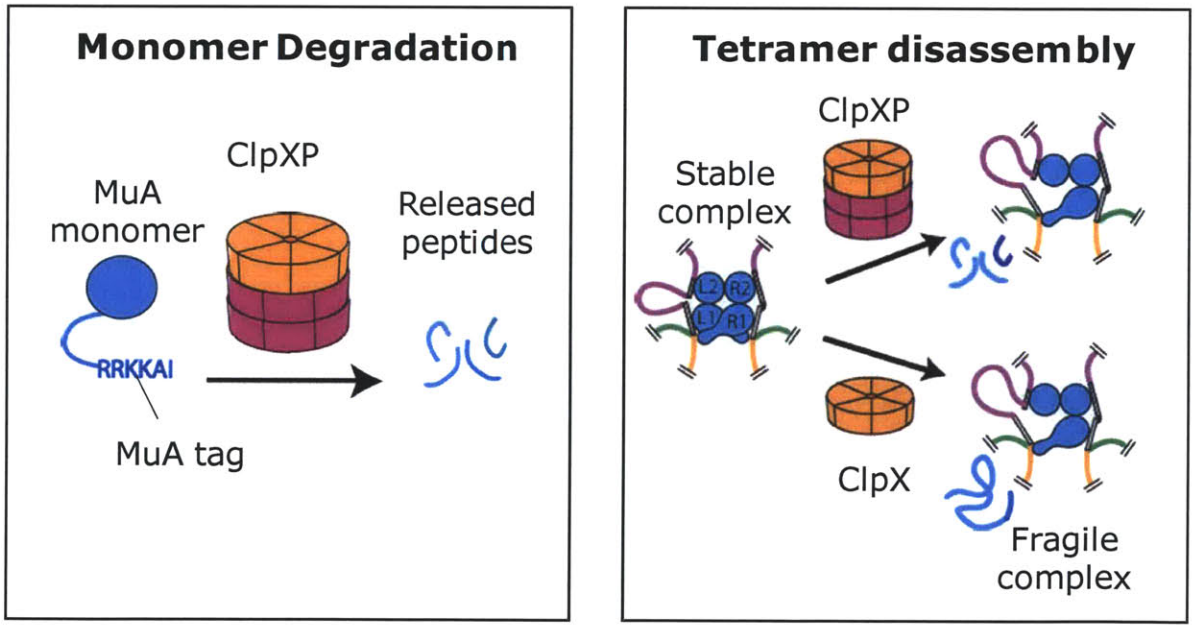


Figure 9. MuA monomer degradation by ClpXP and Mu transpososome disassembly by ClpX or ClpXP. In both reactions, ClpX must bind to the MuA tag to catalyze unfolding and/or degradation. The disassembly reaction proceeds by the unfolding of one subunit from the left side of the transpososome. This results in the formation of the fragile complex, which *in vivo* recruits the replication machinery, and *in vitro* falls apart upon gel electrophoresis to create a family of differently supercoiled plasmid DNA molecules.

The Mu transpososome is an asymmetric complex (Figure 10). Subunits bound to the catalytic Mu DNA binding sites R1 and L1 catalyze DNA cleavage and joining *in trans*, adopting an interwoven structure with numerous intersubunit contacts (Aldaz et al., 1996; Savilahti and Mizuuchi, 1996; Yuan et al., 2005). MuA subunits bound to R2 and L2 make relatively few intersubunit contacts and do not adopt the interwoven structure seen with R1 and L1 subunits. In addition, the sequences of the Mu DNA binding sites are not identical, and the spacing between the right end binding sites is different from the spacing between binding sites of the left end. Previous work has shown that ClpX destabilizes the transpososome not by unfolding all subunits in the tetramer but rather by unfolding one subunit from the left side of the transpososome (Figure 9) (Burton and Baker, 2003; Burton et al., 2001). The unfolding of a subset of subunits within the transpososome is biologically relevant, as it is important that ClpX not unfold the remaining subunits in the complex that are required to recruit the host replication machinery to the Mu DNA fork intermediate (Jones and Nakai, 1997). Additionally, previous studies have suggested that replication of the Mu genome initiates at the left end of the genome, and therefore it make sense structurally to unfold the transpososome on the left side of the complex (Jones and Nakai, 1997; Wijffelman and Lotterman, 1977).

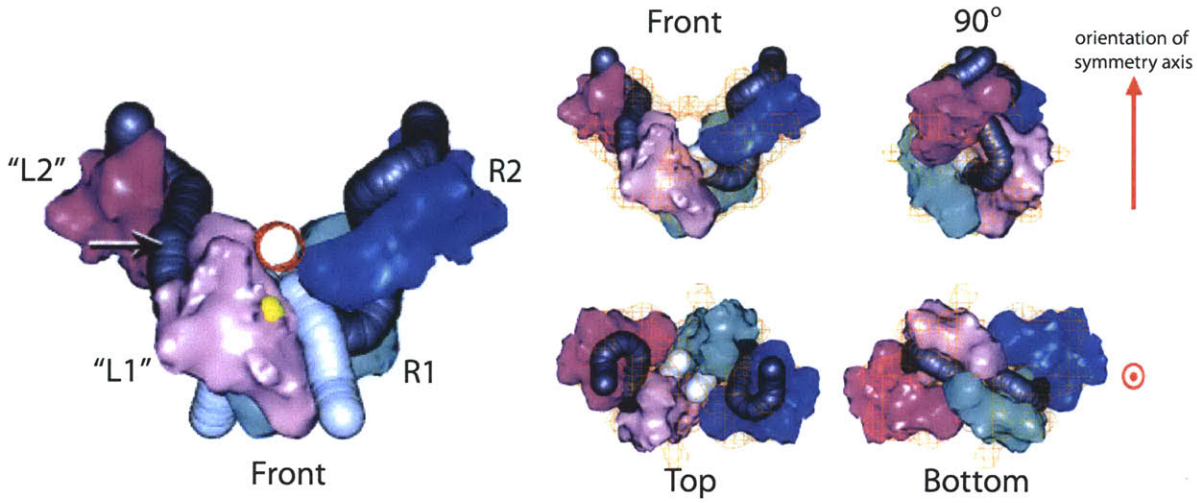


Figure 10. The transpososome is an asymmetric structure. Shown is a 20Å cryoelecton microscope structure of a transpososome assembled on two right ends ("L2" is an R2 site, "L1" is an R1 site), however this structure can be used to approximate what a transpososome on one left and one right end could look like (Yuan et al., 2005). The arrow on the left-most figure represents the location of the DNA spacer that would be present on a left end DNA binding site. Also shown is views of the transpososome at 90°, from the top and from the bottom of the structure. Figure adapted from Yuan et al., 2005.

Auto-tethering as an important selection mechanism for multimeric substrates

To remodel or disassemble complexes, AAA+ unfoldases must specifically recognize the assembled form of the substrate in preference to the free subunits in the cell. How do AAA+ enzymes achieve this specificity, and what specific structural cues guide them to bind to the multimeric substrate? This thesis work explores how AAA+ unfoldases achieve such specificity using the remodeling action of ClpX on the Mu transpososome as a model reaction. This is an ideal system to parse the mechanism of such reactions, as ClpX recognizes both the monomeric and assembled forms of MuA, and the MuA degradation tag is required for recognition for both reactions. This work sheds light on the mechanistic details of substrate auto-tethering, whereby subunits within the complex act as internal adaptors to tether the enzyme to the

multimeric form of the substrate. This mode of binding therefore allows the enzyme to bind to the multimer with higher affinity compared to free subunits. Auto-tethering is now being recognized in other systems, including recognition by ClpX of substrates such as the tubulin-like protein FtsZ and the bacterial DNA stationary phase protector Dps. Mechanistic investigation of these and other multimeric substrates will allow us to further understand this level of regulation as another important tool in the regulatory repertoire of AAA+ enzymes.

REFERENCES

Ades, S.E. (2004). Proteolysis: Adaptor, adaptor, catch me a catch. Curr Biol *14*, R924-926.

Aldaz, H., Schuster, E., and Baker, T.A. (1996). The interwoven architecture of the Mu transposase couples DNA synapsis to catalysis. Cell *85*, 257-269.

Baker, T.A., and Sauer, R.T. (2006). ATP-dependent proteases of bacteria: recognition logic and operating principles. Trends Biochem Sci *31*, 647-653.

Barnett, M.E., Nagy, M., Kedzierska, S., and Zolkiewski, M. (2005). The amino-terminal domain of ClpB supports binding to strongly aggregated proteins. J Biol Chem *280*, 34940-34945.

Becker, G., Klauck, E., and Hengge-Aronis, R. (1999). Regulation of RpoS proteolysis in Escherichia coli: the response regulator RssB is a recognition factor that interacts with the turnover element in RpoS. Proc Natl Acad Sci U S A *96*, 6439-6444.

Bolon, D.N., Grant, R.A., Baker, T.A., and Sauer, R.T. (2004a). Nucleotide-dependent substrate handoff from the SspB adaptor to the AAA+ ClpXP protease. Mol Cell *16*, 343-350.

Bolon, D.N., Wah, D.A., Hersch, G.L., Baker, T.A., and Sauer, R.T. (2004b). Bivalent tethering of SspB to ClpXP is required for efficient substrate delivery: a protein-design study. Mol Cell *13*, 443-449.

Bougdour, A., Cunning, C., Baptiste, P.J., Elliott, T., and Gottesman, S. (2008). Multiple pathways for regulation of sigmaS (RpoS) stability in Escherichia coli via the action of multiple anti-adaptors. Mol Microbiol *68*, 298-313.

Bougdour, A., Wickner, S., and Gottesman, S. (2006). Modulating RssB activity: IraP, a novel regulator of sigma(S) stability in Escherichia coli. Genes Dev *20*, 884-897.

Brandt, V.L., and Roth, D.B. (2004). V(D)J recombination: how to tame a transposase. Immunol Rev *200*, 249-260.

Burton, B.M., and Baker, T.A. (2003). Mu transpososome architecture ensures that unfolding by ClpX or proteolysis by ClpXP remodels but does not destroy the complex. Chem Biol *10*, 463-472.

Burton, B.M., and Baker, T.A. (2005). Remodeling protein complexes: insights from the AAA+ unfoldase ClpX and Mu transposase. Protein Sci *14*, 1945-1954.

Burton, B.M., Williams, T.L., and Baker, T.A. (2001). ClpX-mediated remodeling of mu transpososomes: selective unfolding of subunits destabilizes the entire complex. Mol Cell *8*, 449-454.

Burton, R.E., Baker, T.A., and Sauer, R.T. (2003). Energy-dependent degradation: Linkage between ClpX-catalyzed nucleotide hydrolysis and protein-substrate processing. Protein Sci *12*, 893-902.

Chien, P., Perchuk, B.S., Laub, M.T., Sauer, R.T., and Baker, T.A. (2007). Direct and adaptor-mediated substrate recognition by an essential AAA+ protease. Proc Natl Acad Sci U S A *104*, 6590-6595.

Davey, M.J., Jeruzalmi, D., Kuriyan, J., and O'Donnell, M. (2002). Motors and switches: AAA+ machines within the replisome. Nat Rev Mol Cell Biol *3*, 826-835.

Donaldson, L.W., Wojtyra, U., and Houry, W.A. (2003). Solution structure of the dimeric zinc binding domain of the chaperone ClpX. J Biol Chem *278*, 48991-48996.

Dougan, D.A., Mogk, A., Zeth, K., Turgay, K., and Bukau, B. (2002a). AAA+ proteins and substrate recognition, it all depends on their partner in crime. FEBS Lett *529*, 6-10.

Dougan, D.A., Reid, B.G., Horwich, A.L., and Bukau, B. (2002b). ClpS, a substrate modulator of the ClpAP machine. Mol Cell *9*, 673-683.

Dougan, D.A., Weber-Ban, E., and Bukau, B. (2003). Targeted delivery of an ssrA-tagged substrate by the adaptor protein SspB to its cognate AAA+ protein ClpX. Mol Cell *12*, 373-380.

Ebrahim, S.E. (2007). Binding with intent to destroy : RssB adaptor function in Clp-XP-mediated proteolysis of sigmaS. Ph.D. Thesis, Deptarment of Biology, Massachusetts Institute of Technology.

Erbse, A., Schmidt, R., Bornemann, T., Schneider-Mergener, J., Mogk, A., Zahn, R., Dougan, D.A., and Bukau, B. (2006). ClpS is an essential component of the N-end rule pathway in Escherichia coli. Nature *439*, 753-756.

Erzberger, J.P., and Berger, J.M. (2006). Evolutionary relationships and structural mechanisms of AAA+ proteins. Annu Rev Biophys Biomol Struct *35*, 93-114.

Farrell, C.M., Baker, T.A., and Sauer, R.T. (2007). Altered specificity of a AAA+ protease. Mol Cell *25*, 161-166.

Farrell, C.M., Grossman, A.D., and Sauer, R.T. (2005). Cytoplasmic degradation of ssrA-tagged proteins. Mol Microbiol *57*, 1750-1761.

Flynn, J.M., Levchenko, I., Sauer, R.T., and Baker, T.A. (2004). Modulating substrate choice: the SspB adaptor delivers a regulator of the extracytoplasmic-stress response to the AAA+ protease ClpXP for degradation. Genes Dev *18*, 2292-2301.

Flynn, J.M., Levchenko, I., Seidel, M., Wickner, S.H., Sauer, R.T., and Baker, T.A. (2001). Overlapping recognition determinants within the ssrA degradation tag allow modulation of proteolysis. Proc Natl Acad Sci U S A *98*, 10584-10589.

Flynn, J.M., Neher, S.B., Kim, Y.I., Sauer, R.T., and Baker, T.A. (2003). Proteomic discovery of cellular substrates of the ClpXP protease reveals five classes of ClpX-recognition signals. Mol Cell *11*, 671-683.

Frank, E.G., Ennis, D.G., Gonzalez, M., Levine, A.S., and Woodgate, R. (1996). Regulation of SOS mutagenesis by proteolysis. Proc Natl Acad Sci U S A *93*, 10291-10296.

Geuskens, V., Mhammedi-Alaoui, A., Desmet, L., and Toussaint, A. (1992). Virulence in bacteriophage Mu: a case of trans-dominant proteolysis by the Escherichia coli Clp serine protease. EMBO J *11*, 5121-5127.

Gonzalez, M., Rasulova, F., Maurizi, M.R., and Woodgate, R. (2000). Subunit-specific degradation of the UmuD/D' heterodimer by the ClpXP protease: the role of trans recognition in UmuD' stability. EMBO J *19*, 5251-5258.

Gonzalez, M., and Woodgate, R. (2002). The "tale" of UmuD and its role in SOS mutagenesis. Bioessays *24*, 141-148.

Gottesman, S., Roche, E., Zhou, Y., and Sauer, R.T. (1998). The ClpXP and ClpAP proteases degrade proteins with carboxy-terminal peptide tails added by the SsrA-tagging system. Genes Dev *12*, 1338-1347.

Guo, F., Esser, L., Singh, S.K., Maurizi, M.R., and Xia, D. (2002a). Crystal structure of the heterodimeric complex of the adaptor, ClpS, with the N-domain of the AAA+ chaperone, ClpA. J Biol Chem *277*, 46753-46762.

Guo, F., Maurizi, M.R., Esser, L., and Xia, D. (2002b). Crystal structure of ClpA, an Hsp100 chaperone and regulator of ClpAP protease. J Biol Chem *277*, 46743-46752.

Hanson, P.I., and Whiteheart, S.W. (2005). AAA+ proteins: have engine, will work. Nat Rev Mol Cell Biol *6*, 519-529.

Haslberger, T., Weibezahn, J., Zahn, R., Lee, S., Tsai, F.T., Bukau, B., and Mogk, A. (2007). M domains couple the ClpB threading motor with the DnaK chaperone activity. Mol Cell *25*, 247-260.

Hersch, G.L., Baker, T.A., and Sauer, R.T. (2004). SspB delivery of substrates for ClpXP proteolysis probed by the design of improved degradation tags. Proc Natl Acad Sci U S A *101*, 12136-12141.

Hinnerwisch, J., Fenton, W.A., Furtak, K.J., Farr, G.W., and Horwich, A.L. (2005a). Loops in the central channel of ClpA chaperone mediate protein binding, unfolding, and translocation. Cell *121*, 1029-1041.

Hinnerwisch, J., Reid, B.G., Fenton, W.A., and Horwich, A.L. (2005b). Roles of the N-domains of the ClpA unfoldase in binding substrate proteins and in stable complex formation with the ClpP protease. J Biol Chem *280*, 40838-40844.

Hoskins, J.R., Kim, S.Y., and Wickner, S. (2000a). Substrate recognition by the ClpA chaperone component of ClpAP protease. J Biol Chem *275*, 35361-35367.

Hoskins, J.R., Singh, S.K., Maurizi, M.R., and Wickner, S. (2000b). Protein binding and unfolding by the chaperone ClpA and degradation by the protease ClpAP. Proc Natl Acad Sci U S A *97*, 8892-8897.

Hou, J.Y., Sauer, R.T., and Baker, T.A. (2008). Distinct structural elements of the adaptor ClpS are required for regulating degradation by ClpAP. Nat Struct Mol Biol *15*, 288-294.

Ito, K., and Akiyama, Y. (2005). Cellular functions, mechanism of action, and regulation of FtsH protease. Annu Rev Microbiol *59*, 211-231.

Jahn, R., Lang, T., and Sudhof, T.C. (2003). Membrane fusion. Cell *112*, 519-533.

Jenal, U., and Hengge-Aronis, R. (2003). Regulation by proteolysis in bacterial cells. Curr Opin Microbiol *6*, 163-172.

Jensen, R.B., Wang, S.C., and Shapiro, L. (2002). Dynamic localization of proteins and DNA during a bacterial cell cycle. Nat Rev Mol Cell Biol *3*, 167-176.

Jones, J.M., and Nakai, H. (1997). The phiX174-type primosome promotes replisome assembly at the site of recombination in bacteriophage Mu transposition. EMBO J *16*, 6886-6895.

Jones, J.M., Welty, D.J., and Nakai, H. (1998). Versatile action of Escherichia coli ClpXP as protease or molecular chaperone for bacteriophage Mu transposition. J Biol Chem *273*, 459-465.

Keiler, K.C., Waller, P.R., and Sauer, R.T. (1996). Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA. Science *271*, 990-993.

Kenniston, J.A., Baker, T.A., Fernandez, J.M., and Sauer, R.T. (2003). Linkage between ATP consumption and mechanical unfolding during the protein processing reactions of an AAA+ degradation machine. Cell *114*, 511-520.

Kim, D.Y., and Kim, K.K. (2003). Crystal structure of ClpX molecular chaperone from Helicobacter pylori. J Biol Chem *278*, 50664-50670.

Kim, Y.I., Burton, R.E., Burton, B.M., Sauer, R.T., and Baker, T.A. (2000). Dynamics of substrate denaturation and translocation by the ClpXP degradation machine. Mol Cell *5*, 639-648.

Kirstein, J., Dougan, D.A., Gerth, U., Hecker, M., and Turgay, K. (2007). The tyrosine kinase McsB is a regulated adaptor protein for ClpCP. EMBO J *26*, 2061-2070.

Kirstein, J., Schlothauer, T., Dougan, D.A., Lilie, H., Tischendorf, G., Mogk, A., Bukau, B., and Turgay, K. (2006). Adaptor protein controlled oligomerization activates the AAA+ protein ClpC. EMBO J *25*, 1481-1491.

Kirstein, J., Zuhlke, D., Gerth, U., Turgay, K., and Hecker, M. (2005). A tyrosine kinase and its activator control the activity of the CtsR heat shock repressor in B. subtilis. EMBO J *24*, 3435-3445.

Klauck, E., Lingnau, M., and Hengge-Aronis, R. (2001). Role of the response regulator RssB in sigma recognition and initiation of sigma proteolysis in Escherichia coli. Mol Microbiol *40*, 1381-1390.

Krause, H.M., and Higgins, N.P. (1986). Positive and negative regulation of the Mu operator by Mu repressor and Escherichia coli integration host factor. J Biol Chem *261*, 3744-3752.

Kruger, E., Witt, E., Ohlmeier, S., Hanschke, R., and Hecker, M. (2000). The clp proteases of Bacillus subtilis are directly involved in degradation of misfolded proteins. J Bacteriol *182*, 3259-3265.

Kruger, E., Zuhlke, D., Witt, E., Ludwig, H., and Hecker, M. (2001). Clp-mediated proteolysis in Gram-positive bacteria is autoregulated by the stability of a repressor. EMBO J *20*, 852-863.

Kruklitis, R., Welty, D.J., and Nakai, H. (1996). ClpX protein of Escherichia coli activates bacteriophage Mu transposase in the strand transfer complex for initiation of Mu DNA synthesis. Embo J *15*, 935-944.

Kuo, C.F., Zou, A.H., Jayaram, M., Getzoff, E., and Harshey, R. (1991). DNA-protein complexes during attachment-site synapsis in Mu DNA transposition. Embo J *10*, 1585-1591.

Lazazzera, B.A., and Grossman, A.D. (1997). A regulatory switch involving a Clp ATPase. Bioessays *19*, 455-458.

Lee, I., and Suzuki, C.K. (2008). Functional mechanics of the ATP-dependent Lon protease-lessons from endogenous protein and synthetic peptide substrates. Biochim Biophys Acta *1784*, 727-735.

Levchenko, I., Grant, R.A., Flynn, J.M., Sauer, R.T., and Baker, T.A. (2005). Versatile modes of peptide recognition by the AAA+ adaptor protein SspB. Nat Struct Mol Biol *12*, 520-525.

Levchenko, I., Grant, R.A., Wah, D.A., Sauer, R.T., and Baker, T.A. (2003). Structure of a delivery protein for an AAA+ protease in complex with a peptide degradation tag. Mol Cell *12*, 365-372.

Levchenko, I., Luo, L., and Baker, T.A. (1995). Disassembly of the Mu transposase tetramer by the ClpX chaperone. Genes Dev *9*, 2399-2408.

Levchenko, I., Seidel, M., Sauer, R.T., and Baker, T.A. (2000). A specificity-enhancing factor for the ClpXP degradation machine. Science *289*, 2354-2356.

Levchenko, I., Yamauchi, M., and Baker, T.A. (1997). ClpX and MuB interact with overlapping regions of Mu transposase: implications for control of the transposition pathway. Genes Dev *11*, 1561-1572.

Lo, J.H., Baker, T.A., and Sauer, R.T. (2001). Characterization of the N-terminal repeat domain of Escherichia coli ClpA-A class I Clp/HSP100 ATPase. Protein Sci *10*, 551-559.

Marshall-Batty, K.R., and Nakai, H. (2003). Trans-targeting of the phage Mu repressor is promoted by conformational changes that expose its ClpX recognition determinant. J Biol Chem *278*, 1612-1617.

Marshall-Batty, K.R., and Nakai, H. (2008a). Activation of a dormant ClpX recognition motif of bacteriophage Mu repressor by inducing high local flexibility. J Biol Chem *283*, 9060-9070.

Marshall-Batty, K.R., and Nakai, H. (2008b). Trans-targeting of protease substrates by conformationally activating a regulable ClpX-recognition motif. Mol Microbiol *67*, 920-933.

Martin, A., Baker, T.A., and Sauer, R.T. (2007). Distinct static and dynamic interactions control ATPase-peptidase communication in a AAA+ protease. Mol Cell *27*, 41-52.

Martin, A., Baker, T.A., and Sauer, R.T. (2008). Diverse pore loops of the AAA+ ClpX machine mediate unassisted and adaptor-dependent recognition of ssrA-tagged substrates. Mol Cell *29*, 441-450.

McGinness, K.E., Bolon, D.N., Kaganovich, M., Baker, T.A., and Sauer, R.T. (2007). Altered tethering of the SspB adaptor to the ClpXP protease causes changes in substrate delivery. J Biol Chem *282*, 11465-11473.

McGrath, P.T., Iniesta, A.A., Ryan, K.R., Shapiro, L., and McAdams, H.H. (2006). A dynamically localized protease complex and a polar specificity factor control a cell cycle master regulator. Cell *124*, 535-547.

McNally, F.J., and Vale, R.D. (1993). Identification of katanin, an ATPase that severs and disassembles stable microtubules. Cell *75*, 419-429.

Mhammedi-Alaoui, A., Pato, M., Gama, M.J., and Toussaint, A. (1994). A new component of bacteriophage Mu replicative transposition machinery: the Escherichia coli ClpX protein. Mol Microbiol *11*, 1109-1116.

Muffler, A., Fischer, D., Altuvia, S., Storz, G., and Hengge-Aronis, R. (1996). The response regulator RssB controls stability of the sigma(S) subunit of RNA polymerase in Escherichia coli. EMBO J *15*, 1333-1339.

Nakai, H., Doseeva, V., and Jones, J.M. (2001). Handoff from recombinase to replisome: insights from transposition. Proc Natl Acad Sci U S A *98*, 8247-8254.

Nakai, H., and Kruklitis, R. (1995). Disassembly of the bacteriophage Mu transposase for the initiation of Mu DNA replication. J Biol Chem *270*, 19591-19598.

Neher, S.B., Flynn, J.M., Sauer, R.T., and Baker, T.A. (2003a). Latent ClpX-recognition signals ensure LexA destruction after DNA damage. Genes Dev *17*, 1084-1089.

Neher, S.B., Sauer, R.T., and Baker, T.A. (2003b). Distinct peptide signals in the UmuD and UmuD' subunits of UmuD/D' mediate tethering and substrate processing by the ClpXP protease. Proc Natl Acad Sci U S A *100*, 13219-13224.

Neuwald, A.F., Aravind, L., Spouge, J.L., and Koonin, E.V. (1999). AAA+: A class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes. Genome Res *9*, 27-43.

Ogura, M., Liu, L., Lacelle, M., Nakano, M.M., and Zuber, P. (1999). Mutational analysis of ComS: evidence for the interaction of ComS and MecA in the regulation of competence development in Bacillus subtilis. Mol Microbiol *32*, 799-812.

Outten, C.E., Outten, F.W., and O'Halloran, T.V. (1999). DNA distortion mechanism for transcriptional activation by ZntR, a Zn(II)-responsive MerR homologue in Escherichia coli. J Biol Chem *274*, 37517-37524.

Pak, M., and Wickner, S. (1997). Mechanism of protein remodeling by ClpA chaperone. Proc Natl Acad Sci U S A *94*, 4901-4906.

Park, E.Y., Lee, B.G., Hong, S.B., Kim, H.W., Jeon, H., and Song, H.K. (2007). Structural basis of SspB-tail recognition by the zinc binding domain of ClpX. J Mol Biol *367*, 514-526.

Parsell, D.A., Sanchez, Y., Stitzel, J.D., and Lindquist, S. (1991). Hsp104 is a highly conserved protein with two essential nucleotide-binding sites. Nature *353*, 270-273.

Persuh, M., Turgay, K., Mandic-Mulec, I., and Dubnau, D. (1999). The N- and C-terminal domains of MecA recognize different partners in the competence molecular switch. Mol Microbiol *33*, 886-894.

Pratt, L.A., and Silhavy, T.J. (1996). The response regulator SprE controls the stability of RpoS. Proc Natl Acad Sci U S A *93*, 2488-2492.

Prepiak, P., and Dubnau, D. (2007). A peptide signal for adapter protein-mediated degradation by the AAA+ protease ClpCP. Mol Cell *26*, 639-647.

Pruteanu, M., Neher, S.B., and Baker, T.A. (2007). Ligand-controlled proteolysis of the Escherichia coli transcriptional regulator ZntR. J Bacteriol *189*, 3017-3025.

Rice, P.A., and Baker, T.A. (2001). Comparative architecture of transposase and integrase complexes. Nat Struct Biol *8*, 302-307.

Ryan, K.R., and Shapiro, L. (2003). Temporal and spatial regulation in prokaryotic cell cycle progression and development. Annu Rev Biochem *72*, 367-394.

Sanchez, Y., and Lindquist, S.L. (1990). HSP104 required for induced thermotolerance. Science *248*, 1112-1115.

Sauer, R.T., Bolon, D.N., Burton, B.M., Burton, R.E., Flynn, J.M., Grant, R.A., Hersch, G.L., Joshi, S.A., Kenniston, J.A., Levchenko, I., *et al.* (2004). Sculpting the proteome with AAA(+) proteases and disassembly machines. Cell *119*, 9-18.

Savilahti, H., and Mizuuchi, K. (1996). Mu transpositional recombination: donor DNA cleavage and strand transfer in trans by the Mu transposase. Cell *85*, 271-280.

Schirmer, E.C., Glover, J.R., Singer, M.A., and Lindquist, S. (1996). HSP100/Clp proteins: a common mechanism explains diverse functions. Trends Biochem Sci *21*, 289-296.

Schlieker, C., Weibezahn, J., Patzelt, H., Tessarz, P., Strub, C., Zeth, K., Erbse, A., Schneider-Mergener, J., Chin, J.W., Schultz, P.G., *et al.* (2004). Substrate recognition by the AAA+ chaperone ClpB. Nat Struct Mol Biol *11*, 607-615.

Schlothauer, T., Mogk, A., Dougan, D.A., Bukau, B., and Turgay, K. (2003). MecA, an adaptor protein necessary for ClpC chaperone activity. Proc Natl Acad Sci U S A *100*, 2306-2311.

Shah, I.M., and Wolf, R.E., Jr. (2006). Sequence requirements for Lon-dependent degradation of the Escherichia coli transcription activator SoxS: identification of the SoxS residues critical to

proteolysis and specific inhibition of in vitro degradation by a peptide comprised of the N-terminal 21 amino acid residues. J Mol Biol *357*, 718-731.

Siddiqui, S.M. (2004). Dissecting the steps of substrate processing by the energy-dependent protease ClpXP. Ph.D. Thesis, Department of Biology, Massachusetts Institute of Technology.

Siddiqui, S.M., Sauer, R.T., and Baker, T.A. (2004). Role of the processing pore of the ClpX AAA+ ATPase in the recognition and engagement of specific protein substrates. Genes Dev *18*, 369-374.

Song, H.K., and Eck, M.J. (2003). Structural basis of degradation signal recognition by SspB, a specificity-enhancing factor for the ClpXP proteolytic machine. Mol Cell *12*, 75-86.

Studemann, A., Noirclerc-Savoye, M., Klauck, E., Becker, G., Schneider, D., and Hengge, R. (2003). Sequential recognition of two distinct sites in sigma(S) by the proteolytic targeting factor RssB and ClpX. EMBO J *22*, 4111-4120.

Surette, M.G., Buch, S.J., and Chaconas, G. (1987). Transpososomes: stable protein-DNA complexes involved in the in vitro transposition of bacteriophage Mu DNA. Cell *49*, 253-262.

Thibault, G., Tsitrin, Y., Davidson, T., Gribun, A., and Houry, W.A. (2006a). Large nucleotide-dependent movement of the N-terminal domain of the ClpX chaperone. EMBO J *25*, 3367-3376.

Thibault, G., Yudin, J., Wong, P., Tsitrin, V., Sprangers, R., Zhao, R., and Houry, W.A. (2006b). Specificity in substrate and cofactor recognition by the N-terminal domain of the chaperone ClpX. Proc Natl Acad Sci U S A *103*, 17724-17729.

Tobias, J.W., Shrader, T.E., Rocap, G., and Varshavsky, A. (1991). The N-end rule in bacteria. Science *254*, 1374-1377.

Turgay, K., Hahn, J., Burghoorn, J., and Dubnau, D. (1998). Competence in Bacillus subtilis is controlled by regulated proteolysis of a transcription factor. EMBO J *17*, 6730-6738.

Turgay, K., Hamoen, L.W., Venema, G., and Dubnau, D. (1997). Biochemical characterization of a molecular switch involving the heat shock protein ClpC, which controls the activity of ComK, the competence transcription factor of Bacillus subtilis. Genes Dev *11*, 119-128.

Varshavsky, A. (1992). The N-end rule. Cell *69*, 725-735.

Wah, D.A., Levchenko, I., Rieckhof, G.E., Bolon, D.N., Baker, T.A., and Sauer, R.T. (2003). Flexible linkers leash the substrate binding domain of SspB to a peptide module that stabilizes delivery complexes with the AAA+ ClpXP protease. Mol Cell *12*, 355-363.

Wang, J., Hartling, J.A., and Flanagan, J.M. (1997). The structure of ClpP at 2.3 A resolution suggests a model for ATP-dependent proteolysis. Cell *91*, 447-456.

Wang, K.H., Sauer, R.T., and Baker, T.A. (2007). ClpS modulates but is not essential for bacterial N-end rule degradation. Genes Dev *21*, 403-408.

Weber-Ban, E.U., Reid, B.G., Miranker, A.D., and Horwich, A.L. (1999). Global unfolding of a substrate protein by the Hsp100 chaperone ClpA. Nature *401*, 90-93.

Weibezahn, J., Schlieker, C., Tessarz, P., Mogk, A., and Bukau, B. (2005). Novel insights into the mechanism of chaperone-assisted protein disaggregation. Biol Chem *386*, 739-744.

Weibezahn, J., Tessarz, P., Schlieker, C., Zahn, R., Maglica, Z., Lee, S., Zentgraf, H., Weber-Ban, E.U., Dougan, D.A., Tsai, F.T., *et al.* (2004). Thermotolerance requires refolding of aggregated proteins by substrate translocation through the central pore of ClpB. Cell *119*, 653-665.

Welty, D.J., Jones, J.M., and Nakai, H. (1997). Communication of ClpXP protease hypersensitivity to bacteriophage Mu repressor isoforms. J Mol Biol *272*, 31-41.

Whiteheart, S.W., Schraw, T., and Matveeva, E.A. (2001). N-ethylmaleimide sensitive factor (NSF) structure and function. Int Rev Cytol *207*, 71-112.

Wijffelman, C., and Lotterman, B. (1977). Kinetics of Mu DNA synthesis. Mol Gen Genet *151*, 169-174.

Wojtyra, U.A., Thibault, G., Tuite, A., and Houry, W.A. (2003). The N-terminal zinc binding domain of ClpX is a dimerization domain that modulates the chaperone function. J Biol Chem *278*, 48981-48990.

Xia, D., Esser, L., Singh, S.K., Guo, F., and Maurizi, M.R. (2004). Crystallographic investigation of peptide binding sites in the N-domain of the ClpA chaperone. J Struct Biol *146*, 166-179.

Yuan, J.F., Beniac, D.R., Chaconas, G., and Ottensmeyer, F.P. (2005). 3D reconstruction of the Mu transposase and the Type 1 transpososome: a structural framework for Mu DNA transposition. Genes Dev *19*, 840-852.

Zeth, K., Ravelli, R.B., Paal, K., Cusack, S., Bukau, B., and Dougan, D.A. (2002). Structural analysis of the adaptor protein ClpS in complex with the N-terminal domain of ClpA. Nat Struct Biol *9*, 906-911.

Zhou, Y., Gottesman, S., Hoskins, J.R., Maurizi, M.R., and Wickner, S. (2001). The RssB response regulator directly targets sigma(S) for degradation by ClpXP. Genes Dev *15*, 627-637.

Zolkiewski, M. (2006). A camel passes through the eye of a needle: protein unfolding activity of Clp ATPases. Mol Microbiol *61*, 1094-1100.

CHAPTER 2

UNIQUE CONTACTS DIRECT HIGH-PRIORITY RECOGNITION OF THE TETRAMERIC TRANSPOSASE-DNA COMPLEX BY THE AAA+ UNFOLDASE CLPX

Aliaa H. Abdelhakim, Elizabeth S. C. Oakes, Robert T. Sauer and Tania A. Baker

A modified version of this manuscript is published in Molecular Cell, April 11 2008.

This work was done in collaboration with Dr. Elizabeth S.C. Oakes. I contributed the data in Figures 1, 2, 3, 4, 5, 6 and 8, whereas Elizabeth contributed the data for Figure 7.

SUMMARY

Clp/Hsp100 ATPases remodel and disassemble multiprotein complexes, yet little is known about how they preferentially recognize these complexes rather than their constituent subunits. We explore how substrate multimerization modulates recognition by the ClpX unfoldase using a natural substrate, MuA transposase. MuA is initially monomeric but forms a stable tetramer when bound to transposon DNA. Destabilizing this tetramer by ClpX promotes an essential transition in the phage Mu recombination pathway. We show that ClpX interacts more tightly with tetrameric than with monomeric MuA. Residues exposed only in the MuA tetramer are important for enhanced recognition—which requires the N-domain of ClpX—as well as for a high maximal disassembly rate. We conclude that an extended set of potential enzyme contacts are exposed upon assembly of the tetramer and function as internal guides to recruit ClpX, thereby ensuring that the tetrameric complex is a high-priority substrate.

INTRODUCTION

AAA+ ATPase proteins are present in all kingdoms of life and orchestrate many important cellular processes (Neuwald et al., 1999; Hanson and Whiteheart, 2005). Their activities often involve the remodeling of higher-order complexes to promote activities as diverse as protein disaggregation, membrane fusion, microtubule severing, and DNA replication. Because these enzymes direct so many key biochemical reactions, it is crucial that their substrate specificity be exquisitely regulated both spatially and temporally. Control is especially important for protein-unfolding ATPases as their activity, if left unchecked, could easily inappropriately destroy proteins.

E. coli ClpX is a member of the protein-unfolding Clp/Hsp100 family, a subgroup of AAA+ enzymes. ClpX acts alone as an unfolding chaperone and, in conjunction with the peptidase ClpP, forms an essential part of the ClpXP proteolytic machine (Burton and Baker, 2005). ClpX unfolds proteins to target them for degradation or to alter their structures, a process known as remodeling. To recognize its substrates, ClpX often binds to unstructured peptide sequences known as recognition or degradation tags located near the N- or C- terminus of the target protein. Tags can be intrinsic to the substrate or added co-translationally as is the case for the ssrA tag, which is added during ribosome stalling and targets the nascent chain for degradation (Gottesman et al., 1998; Flynn et al., 2003).

A variety of regulatory strategies operate at the level of ClpX substrate selection. One strategy is the recognition of latent signals that are exposed upon endoproteolytic cleavage of the

substrate (Neher et al., 2003a; Flynn et al., 2004). In other cases, substrates associate with adaptor proteins that, in turn, mediate formation of a ternary complex with ClpX (Gonzalez et al., 2000; Levchenko et al., 2000; Zhou et al., 2001; Neher et al., 2003b). This mechanism of enhanced substrate recognition is referred to as tethering (Baker and Sauer, 2006). In this study, we explore substrate multimerization as a potential mechanism of substrate-recognition control. Because ClpX catalyzes the remodeling/disassembly of multi-protein complexes, we reasoned that mechanisms might exist to allow preferential recognition of assembled complexes relative to the free, unassociated subunits.

To explore the role of multimerization in substrate selection, we used the MuA transposase, a natural disassembly substrate for ClpX. MuA consists of three domains (Figure 1a) and, together with HIV integrase and RAG recombinase, belongs to the DDE family of recombinases (van Gent et al., 1996; Curcio and Derbyshire, 2003). Phage Mu duplicates its genome by replicative transposition. During this multi-step reaction, MuA binds DNA sites located in the terminal repeats of the Mu genome, forms a tetramer, and catalyzes specific DNA cleavage and joining required for transposition (Craigie et al., 1984; Kuo et al., 1991; Lavoie et al., 1991). Because the transesterification reactions that join the DNA molecules during recombination are isoenergetic, MuA and other members of the DDE family must employ mechanisms to drive recombination toward formation of the DNA products. MuA solves this problem by forming a sequential set of increasingly stable nucleoprotein complexes known as transpososomes, allowing stability of the complexes to drive the recombination pathway forward. However, the final product-bound complex is an exceedingly stable transpososome called the strand transfer

complex I (STCI) (Surette et al., 1987), which must be disassembled to allow initiation of phage DNA replication (Mhammedi-Alaoui et al., 1994; Nakai and Kruklitis, 1995). The host ClpX unfoldase remodels and destabilizes the STCI to form the fragile STCII complex, which, in turn, helps recruit DNA-replication machinery and is finally removed from the DNA (Levchenko et al., 1995; Nakai and Kruklitis, 1995; Jones et al., 1998; Nakai et al., 2001). Genetic studies establish that transpososome remodeling by ClpX is essential for phage Mu growth (Mhammedi-Alaoui et al., 1994).

MuA is a monomer in solution but efficiently assembles into transpososomes *in vitro*. The STCI is an asymmetric complex generated by incubating MuA with supercoiled plasmid DNA containing "right" and "left" Mu DNA binding sites (R1, R2, L1 and L2) (Figure 1b) (Mizuuchi, 1983). We will refer to this complex as the MuA tetramer or the transpososome. Both the MuA monomer and the transpososome are ClpX substrates (Levchenko et al., 1995; Levchenko et al., 1997). ClpXP degrades MuA monomers, whereas either ClpX or ClpXP can remodel stable STCI to form fragile STCII (Figure 1b,c,d). Unlike STCI, STCII is unstable to gel electrophoresis, and the transpososome remodeling reaction is often referred to as disassembly (Figure 1b) (Kruklitis et al., 1996). MuA contains a C-terminal recognition signal known as the MuA tag (RRKKAI) that is necessary for ClpX recognition of both monomeric MuA and transpososomes (Levchenko et al., 1997). According to a current model for disassembly, ClpX destabilizes STCI by unfolding and extracting one subunit from the left side the transpososome to form STCII (Figure 1b) (Burton and Baker, 2003). Given that, by this model, ClpX unfolds one MuA molecule both during degradation and disassembly, the same interactions could mediate MuA recognition in

both reactions. Alternatively, contacts made specifically with the MuA tetramer might guide recognition of the complex.

Here, we show that ClpX contacts MuA monomers and Mu transpososomes differently. We find that ClpX makes more extensive protein-protein contacts with the transpososome, resulting in higher affinity for the tetramer, and show that the N-domain of ClpX is critical only for recognition of the transpososome. Furthermore, we provide evidence that the regions of MuA that interact with the ClpX N-domain are more exposed and more accessible in the tetramer. Thus, this work reveals multiple strategies that can be used by disassembly enzymes to ensure specific recognition of the assembled state of a substrate.

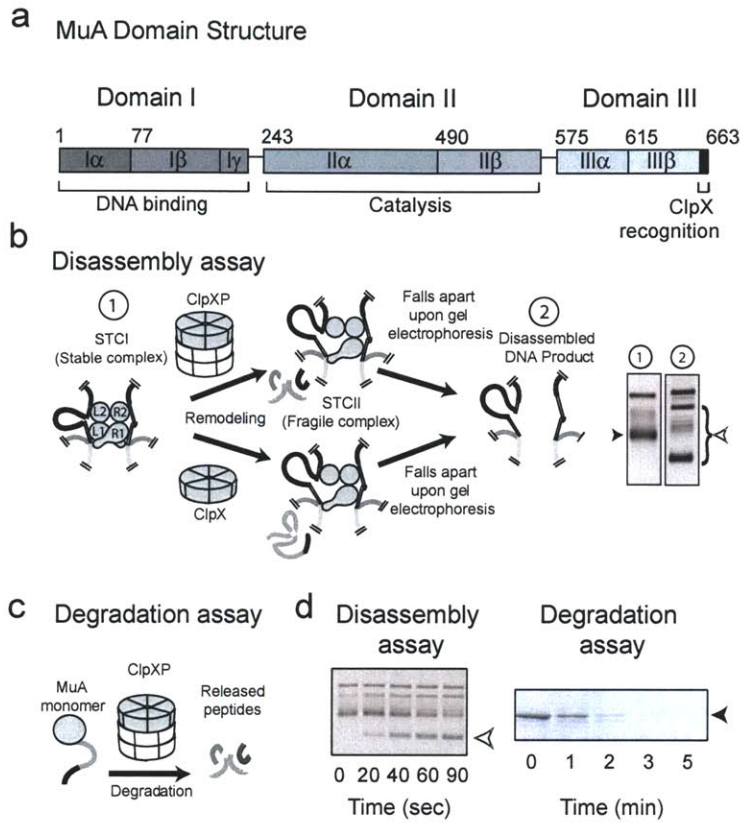


Figure 1: Schematic of Degradation and Disassembly Reactions and Assays

A. Schematic representation of full-length MuA domain structure. B. Transpososome remodeling by ClpX or ClpXP. The stable complex (STCI) is a tetramer of the transposase MuA assembled onto plasmid DNA pMK586 ("mini-Mu") *in vitro*. Assembly was done in the presence of the host protein, HU, but without the MuA-activating protein, MuB. MuB was omitted from the reactions because the binding sites on MuA for MuB and ClpX partially overlap. This overlap allows MuB to control the timing/access of ClpX to the transpososome for disassembly (Levchenko et al., 1997). As this study is focused on MuA-ClpX affinities, we omitted MuB from all reactions. When visualized on a native agarose gel, the stable complex appears as a band that migrates slower than supercoiled plasmid DNA alone (1). ClpX remodels the stable complex by unfolding a subunit on the left side of the transpososome to produce the fragile complex (STCII). If ClpX is bound to ClpP, then the subunit that is unfolded is also degraded. The STCII falls apart upon gel electrophoresis and produces a characteristic series of differently supercoiled DNA disassembly products (2). C. Schematic of ClpXP-mediated degradation of monomeric MuA. D. Rates of disassembly of MuA tetramer were assayed by measuring the rate of appearance of the lower most DNA disassembly product on a native agarose gel (open arrow). For each timepoint, this band was quantified as a percent of the total counts in the lane and normalized to the +SDS control, which was used as the "100% disassembly" control. Rates of MuA monomer degradation were measured by measuring the rate of disappearance of MuA by SDS-PAGE (closed arrow).

RESULTS

Amino-acid substitutions reveal two classes of ClpX-MuA contacts.

ClpX poorly recognizes a MuA mutant lacking the C-terminal tag, as assayed both by monomer degradation and tetramer disassembly (Figure 2; Levchenko et al., 1997). It is unknown, however, whether ClpX makes the same contacts with this C-terminal tag in both reactions. To address this question, we used site-directed mutagenesis to singly mutate each of the C-terminal six residues of MuA (RRKKAI). In each case, we introduced a glutamic acid (E) or an aspartic acid (D), because acidic residues disrupt ClpX contacts with other recognition tags (Flynn et al., 2001).

The MuA variants were purified, shown to be active in transposition *in vitro*, and assayed for monomer degradation by ClpXP and for disassembly by either ClpX or ClpXP (Figure 2; see Figure 1 for description of assays). Most tag substitutions (R659D, K660E, K661E, A662D and I663D) inhibited both degradation and disassembly. The wild-type residues at these positions presumably mediate "core" contacts that are important for ClpX recognition of both forms of MuA. By contrast, substitution of R658 with glutamate (R658E) did not affect degradation but slowed disassembly by more than 50% (Figure 2; Figure 3, a and c). As a control, we repeated this experiment at various substrate concentrations and even when the mutant transpososomes where present at 10 nM (~100-fold below the K_M for their recognition, see below), they were still disassembled more slowly than the wild-type complexes. These results suggest that R658 plays no role in ClpX recognition of monomeric MuA but contributes to ClpX recognition of the transpososome. We refer to interactions of this type as "extended" contacts.

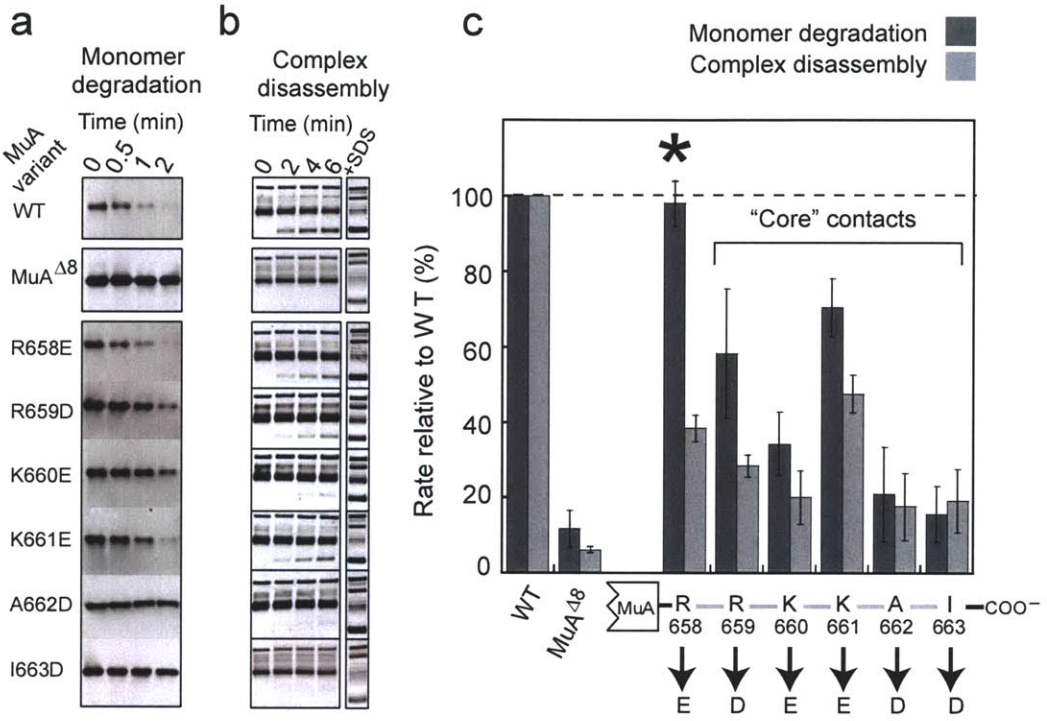


Figure 2. Point mutations that affect both monomer degradation and complex disassembly. A. Degradation rates by ClpXP of wild-type MuA, MuA$^{\Delta 8}$ (MuA lacking the last eight C-terminal residues) and "core" residue mutations in monomeric form. The initial MuA concentration was kept low (0.1 μM) to ensure that differences in affinity for ClpX are reflected in the rate of the reaction (see Figure 4 for substrate K_M values). Degradation reactions were visualized by Western blotting. B. Disassembly of transpososomes assembled with wild-type MuA, MuA$^{\Delta 8}$ and MuA containing "core" residue substitutions by ClpX. The initial transpososome concentration was 0.1 nM. The "+SDS" controls shows the pattern of plasmid migration upon complete disassembly. C. Quantification of differences in degradation and disassembly rates relative to wild type MuA. Degradation of all MuA variants was compared to a degradation reaction for wild-type MuA on the same gel and Western blot to control for variations. Experiments were repeated in triplicate, and error bars represent the standard deviation of the average. Residue mutations that affect both degradation and disassembly are labeled "core" residues. R658E, which affects only disassembly, is labeled with an asterisk.

Previous results show that ClpX can bind to many peptides that correspond to sequences within the MuA protein (Thibault et al., 2006). To search for additional "extended" contact residues, we focused on arginines within the C-terminal MuA domain IIIβ, because arginine is frequently

found in ClpX recognition motifs (Flynn et al., 2003). Several mutations with the anticipated properties were identified (Figure 3, a and b). For example, the R616A and R622A transpososome variants were disassembled slowly by ClpX, but the monomeric mutants were degraded at the same rate as wild-type MuA. Other arginine mutations (R635A and R643A) had no effect on either degradation or disassembly (Figure 3a and b). The apparent K_M for disassembly of transpososomes, a measure of the functional affinity of ClpX for the complex (see below), was increased ~2 to 5 fold for the extended-contact R616A, R622A and R658E MuA variants (see Figure 4a). These results reinforce the conclusion that ClpX makes protein-protein contacts with specific residues in the transpososome tetramer in addition to contacts that it also makes with the MuA monomer.

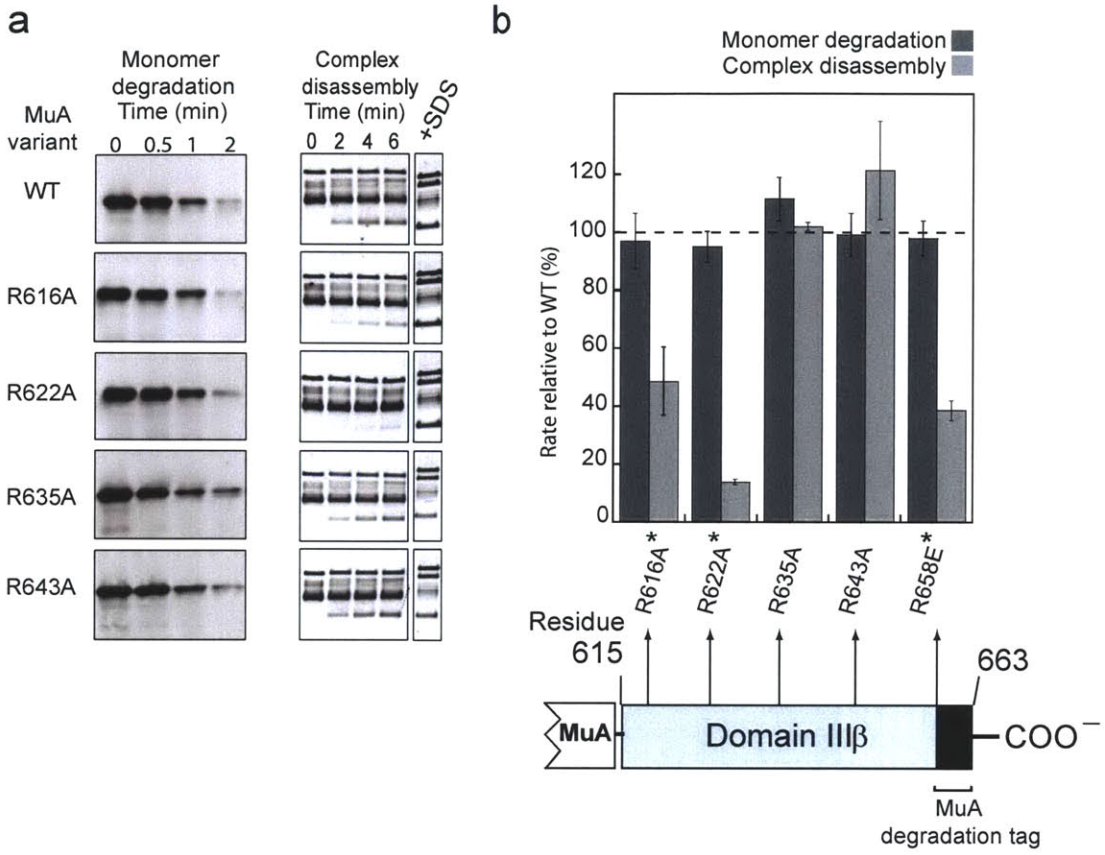


Figure 3. Point mutations that affect only complex disassembly. A. Degradation and disassembly rates of internal point mutants R616A, R622A, R635A and R643A relative to wild-type MuA. B. Quantification of differences in degradation and disassembly rates for each mutant relative to wild-type MuA, including variants in Figure 3a and MuA R658E (Figure 2). Experiments were repeated in triplicate, and error bars represent the standard deviation of the average. Residue mutations that affect disassembly but not degradation are labeled with an asterisk.

ClpX interacts with transpososomes more strongly than MuA monomers.

The finding that ClpX makes more extensive interactions with MuA in the tetramer than the monomer suggested that ClpX might bind the MuA tetramer more strongly. To test this model, we measured the functional interaction of ClpX with MuA during degradation and disassembly by determining the concentration required to obtain half-maximal velocity for the two reactions (K_M^{app}). Because it is difficult to obtain transpososomes at high concentration, we started with a fixed substrate concentration, varied the concentration of ClpXP, measured the rate of appearance of free DNA released by disassembly (see Figure 1d), and analyzed the data as previously described to obtain apparent K_M^{app} values (Herschlag and Cech, 1990; Pyle and Green, 1994). K_M^{app} for the wild-type ClpXP-transposase interaction was 1.0 ± 0.3 μM (Figure 4a). We then analyzed ClpXP-mediated MuA monomer degradation in the standard fashion, by measuring degradation rates as a function of MuA concentration. K_M^{app} for MuA monomer degradation by ClpXP was 10.5 ± 2.7 μM (Figure 4b). To establish that varying enzyme concentration is a valid method for determining K_M^{app}, we also measured MuA monomer degradation rates by increasing ClpXP concentration. The resulting rates were superimposable on the curve obtained by varying substrate concentration (Figure 4b). Thus, ClpX appears to interact with the MuA monomer substantially more weakly than with the tetramer.

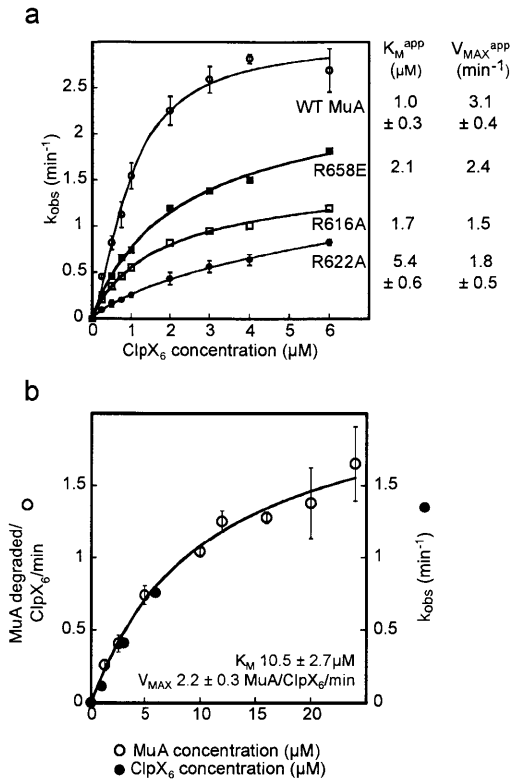


Figure 4. Half-maximal velocity determination for monomeric and multimeric forms of MuA. A. ClpXP-mediated disassembly rate curves for wild-type complexes and complexes with "extended" mutations as a function of enzyme concentration. The initial concentration of transpososomes was 0.1 nM. The curve for wild type MuA transpososomes was best fit by including a modest cooperativity term or "apparent" Hill coefficient compared to the transpososomes assembled with the mutant variants of MuA and to monomeric MuA. This observation is consistent with models with more than one binding site on the wild-type transpososome for ClpX. Apparent K_M curves for transpososomes were fit to a modified Hill equation (reaction rate = $(V_{max}^{app})/(1+(K_M^{app}/[ClpXP])^n)$, where n is the Hill coefficient. The Hill value for which the wild type curve fit best was 1.5 ; for mutants R658E, R616A and R622A, the apparent Hill value for which curves fit best was 1. The R^2-value for each fit was >0.99. Curves for wild type and R622A variant were repeated in triplicate. B. K_M curve for monomeric MuA. Open circles represent K_M curve as a function of MuA concentration. $ClpX_6$ concentration was kept constant for each MuA concentration at 0.4 μM for this curve. Each data point was repeated in triplicate. Binding curve was fit to the equation rate = $(V_{max}^{app})/(1+(K_M^{app}/[MuA]))$. The R^2-value for the fit was >0.99. Superimposed is the apparent K_M curve obtained as a function of ClpXP concentration (closed circles). The initial MuA concentration for this experiment was 1 μM. Apparent rates were fit to the equation rate = $(V_{max}^{app})/(1+(K_M^{app}/[ClpXP]))$. The value at each ClpXP concentration on the curve was divided by the resulting V_{max}^{app} for this fit. Each data point was then multiplied by the V_{max}^{app} calculated for the K_M curve where substrate concentration was varied.

These titration experiments suggest that the functional interaction of ClpX with transpososomes is about 10-times tighter than the interaction with the MuA monomer. As shown by the reduced "ClpX-affinities" of the R616A, R622A and R658E MuA transpososome variants (Figure 4a), extended contacts between ClpX and MuA in the transpososome play roles in stabilizing of the enzyme-substrate complex.

Transpososome recognition requires the N-domain of ClpX.

The differential interaction of ClpX with residues R616, R622 and R658 in the MuA monomer and tetramer demonstrates that the increased affinity of the tetramer for ClpX is not simply a multivalent avidity effect. To test for other differences in enzyme-substrate interactions, we examined the role of the ClpX N-domain, which binds MuA *in vitro* and is essential for the replication of phage Mu *in vivo* (Wojtyra et al., 2003). This ClpX domain functions in selection of many substrates, often by binding adaptor proteins (Neher et al., 2003b; Wojtyra et al., 2003; Bolon et al., 2004; Siddiqui, 2004). For these studies, we used a $ClpX^{\Delta N}$ variant that supports ClpP degradation of ssrA-tagged substrates with the same efficiency as wild-type ClpX (Neher et al., 2003b).

We compared the activities of $ClpX^{\Delta N}$ and ClpX in degradation and disassembly assays. In the presence of ClpP, $ClpX^{\Delta N}$ supported degradation of the MuA monomer at a rate roughly comparable to wild-type ClpX and with a K_M (21 μM) twice the wild-type value (Figure 5a; data not shown). We conclude that the N-domain of ClpX contributes to, but is not essential for, efficient recognition of MuA monomers. By contrast, $ClpX^{\Delta N}$ was severely defective in

disassembly (Figure 5b). ClpX alone disassembled over half of the transpososome complexes within 1 min, whereas $ClpX^{\Delta N}$ disassembled less than 20% of the transpososomes after 1 hour, an ~150-fold difference in disassembly rates. To ensure that $ClpX^{\Delta N}$ was not defective at disassembling stable complexes, we used transpososomes made with the chimeric protein MuA-ssrA, which contains the ssrA tag fused to the C-terminus of MuA^{1-615} (Burton and Baker, 2003). SsrA-tag recognition is unaffected by the N-domain (Wojtyra et al., 2003). Indeed, $ClpX^{\Delta N}$ disassembled MuA-ssrA transpososomes ~60-fold faster than wild-type transpososomes (Figure 5c), showing that this enzyme has no inherent defect in remodeling complexes. Rather, the N-domain of ClpX appears to make crucial contacts with wild-type transpososomes that are required for robust disassembly.

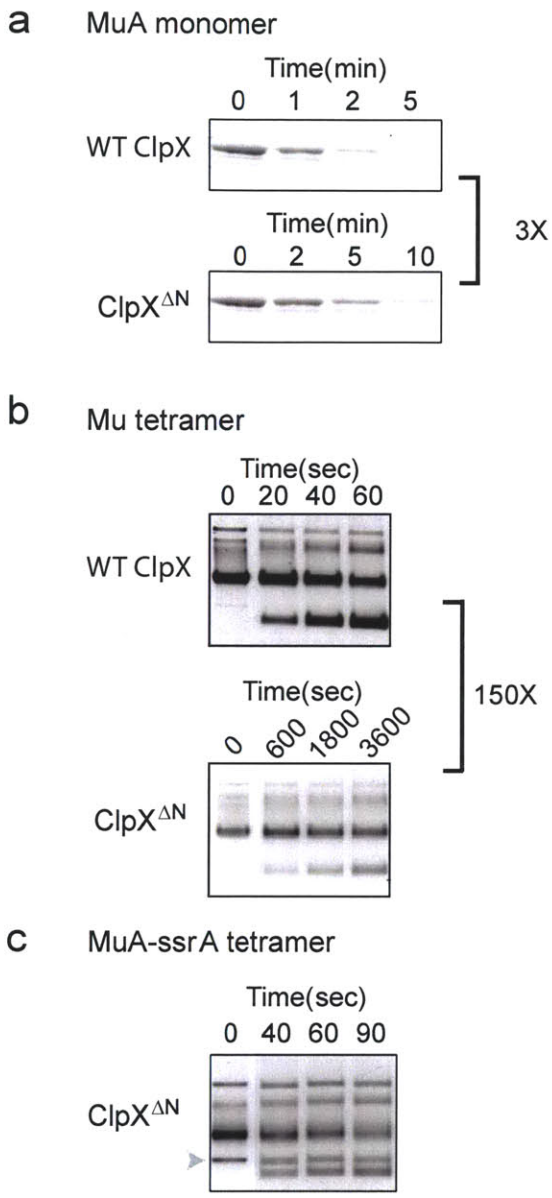


Figure 5. Monomer degradation and complex disassembly have differential requirements for the ClpX N-domain. A. ClpP-mediated degradation of monomeric MuA supported by wild type ClpX and ClpX$^{\Delta N}$. Starting concentrations were MuA (1 μM), ClpX (0.3 μM), and ClpP (0.8 μM). Proteins were visualized using Sypro Orange protein stain. B. ClpP-mediated disassembly of wild type transpososomes supported by wild type ClpX and ClpX$^{\Delta N}$. Starting concentrations were transpososomes (1 nM), ClpX or ClpX$^{\Delta N}$ (1 μM). C. Disassembly of MuA$^{1\text{-}615}$-ssrA complexes by ClpX$^{\Delta N}$. Gray arrow points to supercoiled plasmid DNA not associated with MuA.

MuA transposase contains cryptic recognition determinants buried in the monomer.

Our results reveal that the ClpX N-domain is more important for recognition of the MuA tetramer than the monomer. One possibility is that cryptic N-domain dependent signals may only become exposed upon assembly of the MuA tetramer. These signals could serve to increase the affinity of ClpX for the transpososome via N-domain interactions. To test this model, we purified truncated variants of MuA based on the known domain boundaries (Figure 6a). These proteins were then assayed for N-domain interactions by comparing the rates of ClpP degradation supported by ClpX and $ClpX^{\Delta N}$ (Figure 6b). ClpXP degraded the truncated MuA^{1-492} and MuA^{1-574} variants more slowly than full-length MuA, as expected because these substrates lack the C-terminal MuA tag (Figure 6b). Importantly, however, ClpXP degradation of these variants depended strongly on the ClpX N-domain. At substrate concentrations of 1 μM, deletion of the N-domain reduced the rate of degradation of MuA about 3-fold but reduced degradation of MuA^{1-574} and MuA^{1-492} almost 60-fold (Figure 6b). Similarly, N-domain deletion reduced degradation of isolated domain III of MuA ($MuA^{575-663}$) about 10-fold, showing that recognition of this polypeptide also benefits from the ClpX N-domain interactions (Figure 6b). These data suggest that separation of MuA domains I and II from domain III exposes signals that are recognized by the ClpX N-domain.

MuA^{1-574} and the transpososome make the same N-domain contacts.

Although numerous studies have probed contacts between MuA domain III and ClpX, our data establish that ClpX can also contact MuA variants lacking domain III. The strong N-domain dependence of ClpX recognition of the monomeric variants (MuA^{1-574} and MuA^{1-492}) mirrors the

requirements for the N-domain exhibited by the tetramer (see above). To probe whether a truncated monomeric variant and the tetramer make similar contacts with the ClpX N-domain we used a competition approach. If the MuA^{1-574} monomer makes N-domain contacts similar to those needed for recognition of the MuA tetramer, then it would be expected to compete efficiently for ClpX in disassembly assays. By contrast, if only the contacts with the C-terminus of intact MuA are important for degradation, then the MuA^{1-574} monomer should not compete efficiently for degradation. Strikingly, this pattern of competition was observed (Figure 6c, d). These data suggest that the ClpX N-domain binds MuA determinants, located between residues 1 and 574, which are exposed in the tetrameric transpososome but not in the MuA monomer. Assembly-dependent interactions of this type could account for the large dependence of tetramer recognition on the N-domain of ClpX. Moreover, the full-length MuA monomer inhibited disassembly only ~30% as effectively as MuA^{1-574} (data not shown), further supporting a model in which important ClpX recognition contacts only become exposed upon MuA tetramer assembly or deletion of domain III.

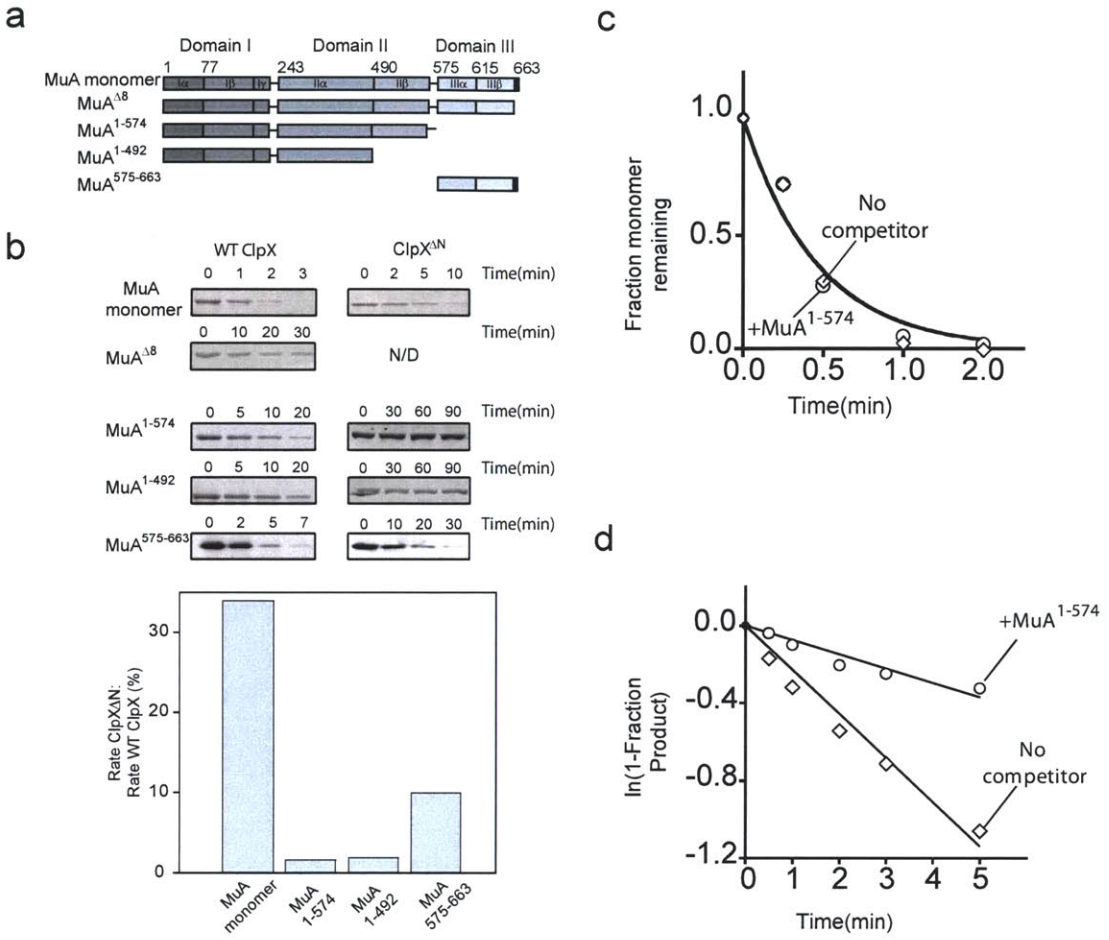


Figure 6. MuA truncations contain cryptic recognition determinants that compete only with the Mu tetramer for ClpX binding sites. A. Schematic of truncations used to determine ClpXP interaction with full-length MuA. The black box at the end of full length MuA and truncation 575-663 represents residues 656-663 of the MuA C-terminal tag. B. ClpP-mediated degradation of MuA monomer and truncation variants supported by ClpX and ClpX$^{\Delta N}$. The bar graph expresses dependence of MuA monomer and truncation products MuA$^{575-663}$, MuA$^{1-574}$, and MuA$^{1-492}$ on the ClpX N-domain as percent rate of reaction with ClpX$^{\Delta N}$P compared to wild-type ClpXP. Substrate concentrations (1 μM); ClpXP concentration (0.3 μM). MuA$^{575-663}$ degradation is visualized by Western blotting; all other gels are visualized by Sypro Orange. C. Addition of MuA$^{1-574}$ as a competitive inhibitor to degradation of monomeric MuA (no inhibitor, open diamonds; plus inhibitor, open circles). Starting concentrations of proteins were full-length MuA (0.1 μM), ClpX (0.1 μM), ClpP (0.2 μM), and MuA$^{1-574}$ (1 μM). The degradation reaction was separated by SDS-PAGE and probed using Western blotting. D. Addition of MuA$^{1-574}$ as a competitive inhibitor to disassembly of Mu complexes. Starting concentrations were transpososome (0.1 nM), ClpX (0.1 μM), and MuA$^{1-574}$ (1 μM).

Increased lysine exposure upon MuA tetramer formation.

Do conformational changes result in enhanced exposure of some regions of MuA upon tetramer formation? To address this question, we performed lysine-acetylation footprinting experiments on MuA monomers or tetramers and analyzed the extent of chemical modification by tandem mass spectrometry. Lysine residues exposed to solvent should be more readily acetylated than those buried in the structure of the protein. These experiments revealed a set of lysines that were acetylated to a greater extent in MuA tetramers than in monomers. For example, lysines at positions 95, 109, 461, 579, 582, 586, and 631 were fully acetylated in the tetramer but only partially acetylated in the monomer, whereas lysine 530 was partially acetylated in the tetramer but was not acetylated in the monomer (Figure 7a). These results strongly suggest that conformational changes exist between MuA monomers and tetramers that increase the accessibility of eight lysines, six of which were located in domains II and III; several of these exposed residues are near residues we identified as making "extended" contacts with ClpX (Figures 3, 6, 7a). As a result, our chemical-modification experiments support a model in which "extended" contacts are exposed only in the tetramer, allowing ClpX to bind with high affinity specifically to the transpososome.

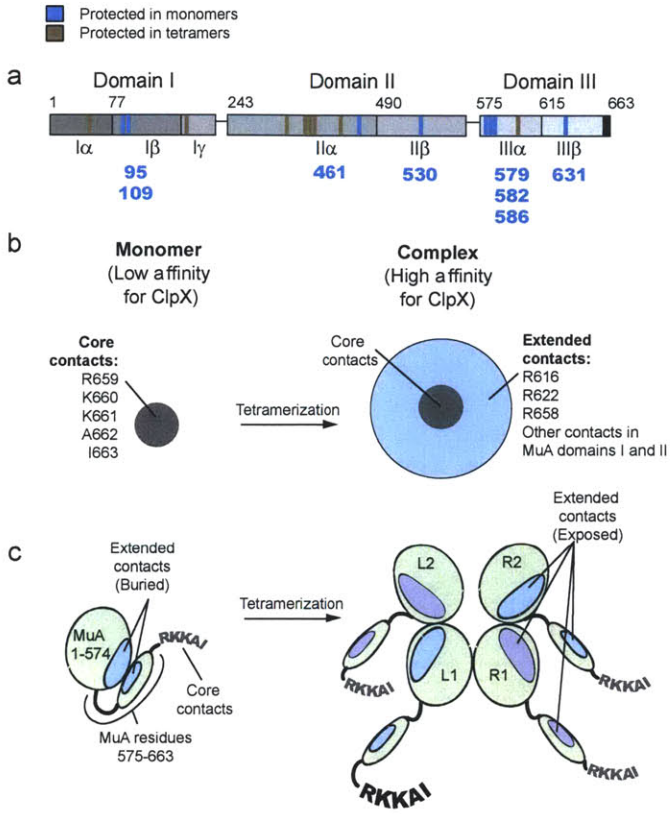


Figure 7. Model of differential presentation of monomeric MuA and Mu complex to ClpX/ClpXP. A. Acetylation experiments reveal lysines in MuA that are more exposed in the tetramer than the monomer (marked as blue lines; corresponding residue numbers are labeled in blue). Lysine residues that are more exposed in monomers than tetramers are marked in brown (lysines 44, 177, 299, 357, 364, 373, 406, 590). Protection of lysines in the tetramer can be attributed to intersubunit contacts and interactions with DNA. B. Venn diagram representation of MuA interactions with ClpX in the monomer and tetramer form. In this representation, the total area of the circles is proportional to the number of contacts, and hence to the affinity, of MuA to ClpX. C. Schematic representation of differential interactions of MuA with ClpXP. In the MuA monomer, surfaces that interact with ClpX and the N-domain in the tetramer are buried. Formation of the transpososome is accompanied by changes in quaternary structure and conformation, leading to the exposure of ClpX-interacting regions that serve to increase the affinity of the complex for ClpX. ClpX is guided by these interactions to unfold a subunit on the left side of the transpososome. For the purposes of this illustration the L1 degradation tag is highlighted for unfolding, although current data does not allow us to distinguish whether it is the L1 or L2 subunit that is unfolded by ClpX. Light blue areas represent tetramer-specific ClpX interacting regions facing out of the page; dark blue areas represent the same type of regions facing into the page.

DISCUSSION

ClpX recognizes the transpososome by an autotethering mechanism.

Our results establish that the mode of recognition by an AAA+ unfoldase can be altered upon multimerization of a substrate. ClpX makes important "core" contacts with specific residues of a C-terminal recognition tag both in MuA monomers and in the tetrameric MuA transpososome. However, ClpX makes a larger set of "extended" protein-protein contacts with residues that are exposed specifically within the transpososome, resulting in a higher affinity for the tetramer (Figure 7b). It is unlikely that we have identified all of the residues within MuA that make extended transpososome-ClpX contacts, however just a handful of such contacts, made by residues in MuA domain III and in other domains, could easily account for ClpX's 10-fold higher affinity for the MuA tetramer relative to the monomer.

In principle, the higher ClpX affinity for the MuA tetramer could result from a simple multivalent avidity effect. However, an avidity model does not account for our finding that the ClpX N-domain plays a substantially more important role in tetramer disassembly than in monomer degradation nor for our observation that some MuA mutations affect tetramer disassembly but not monomer degradation. Additionally, if simple multivalent recognition of the MuA tag were responsible for the enhanced recognition of the tetramer, then tag mutations would be expected to have greater effects on remodeling than on monomer degradation. This result was not observed for the A662D and I663D mutations, which slowed disassembly and monomer degradation to comparable extents. By contrast, the existence of assembly-dependent "extended" contacts, mediated at least in part by the ClpX N-domain, can

account for these observations and for enhanced tetramer affinity. Hence, ClpX recognition of the tetrameric MuA transpososome is more complex than recognition of the MuA monomer.

What mechanism could result in the distinct modes of recognition of the two Mu substrates by ClpX? An intriguing model, consistent with our deletion, competition, and lysine-modification experiments, is that MuA tetramerization leads to conformational changes that expose otherwise cryptic recognition determinants within one or more MuA subunits. Conceptually similar changes have been proposed to explain preferential ClpXP degradation of Mu-repressor multimers containing a mixture of wild-type and mutant subunits (Marshall-Batty and Nakai, 2003). In the case of MuA, these conformational changes could involve assembly-dependent rearrangements in domain-domain contacts, as depicted in Figure 7c. Indeed, we observed that truncating larger portions of domain III in MuA resulted in faster degradation, suggesting that domain III in the monomer masks ClpX recognition determinants in the remaining domains of MuA (Figure 8). Consistent with our observations, cryo-electron microscopy studies reveal that MuA monomers and tetramers adopt different conformations (Yuan et al., 2005). Thus, we propose that upon formation of the tetramer, MuA undergoes conformational changes affecting domain III that expose residues in the remainder of MuA, which subsequently contribute to ClpX binding. This model explains why removal of a few C-terminal residues of MuA severely inhibits degradation, whereas removal of much larger portions of domain III alleviates some of this defect in recognition of the MuA monomer. This mechanism also explains why the truncated MuA^{1-574} variant inhibits tetramer disassembly but not monomer degradation.

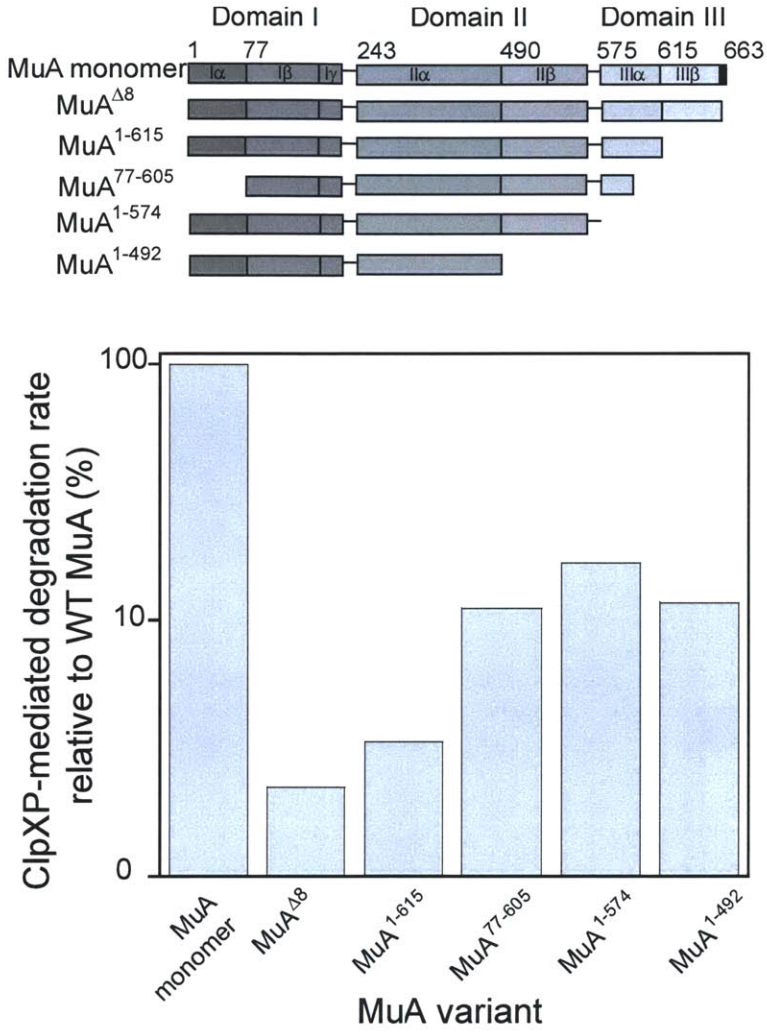


Figure 8

Rates of degradation of MuA monomer and truncations by ClpXP, expressed as percentage of MuA monomer degradation rate. Controls were performed to ensure that the absence of MuA residues 1-77 in the truncation MuA^{77-605} does not affect the rate of degradation by ClpXP.

ClpX recognition of the MuA transpososome, compared to the monomer, is reminiscent of adaptor-enhanced substrate delivery in having a lower K_M, a higher V_{max}, and a larger dependence on the ClpX N-domain. These features suggest that contacts on subunits within

the complex act as internal protein "adaptors" that tether the subunit to be unfolded to ClpX via its N-domain. The inherent asymmetry and geometry of the transpososome could provide ClpX with a mechanism to unfold only one subunit from the complex. In this model, "adaptor" contacts with the transpososome guide ClpX to the "substrate" subunit with high affinity and specificity. Once the "substrate" subunit is unfolded and the STCII is formed, the geometry optimal for high-affinity binding by ClpX is lost. This "autotethering" mechanism would therefore be a self-limiting reaction that prevents ClpX from completely destroying STCII, which is required to complete phage DNA replication (Kruklitis et al., 1996). It is currently unclear whether the L1 or L2 subunit of the transpososome is unfolded by ClpX and whether the same or different subunits within the transpososome act as adaptor subunits. Experiments are currently underway to determine the architecture of these ClpX-MuA transpososome interactions.

The N-domain of ClpX is known to bind peptide sequences, for example LRVVK at the C-terminus of the SspB adaptor (Dougan et al., 2003; Wah et al., 2003; Park et al., 2007). Our data are inconsistent with binding of a single short peptide within MuA to the N-domain of ClpX, as the "extended" contacts span large regions of the primary sequence of MuA, including multiple residues in MuA domains II and III. It is possible that these extended contact residues are close in space and create a surface on the Mu tetramer that is recognized by the N-domain of ClpX. Alternatively, several different peptide sequences in a MuA tetramer might interact with different N-domains in the ClpX hexamer. By either model, eliminating an "extended" contact

by mutation would weaken but not abolish binding of the tetramer to ClpX, because other "extended" and core contacts would maintain some binding of the enzyme to the substrate.

Comparisons with ssrA-substrate recognition

The best-characterized ClpX recognition signal is the 11-residue ssrA tag sequence, which is co-translationally added to nascent chains during ribosome stalling and targets the truncated polypeptide for degradation. The ssrA tag has a comparatively strong affinity for ClpX (~1 μM) and the residues within this sequence that interact with ClpX have been characterized. The last two residues (AA) and the C-terminal carboxylate of the ssrA tag make the most important contacts with ClpX (Kim et al., 2000; Flynn et al., 2001). Our studies show that ClpX recognition of the MuA tag is different from ssrA-tag recognition. The MuA monomer has weaker apparent affinity for ClpX (~ 10 μM), and the C-terminal carboxylate of the MuA tag does not appear to contribute substantially to recognition (A.A., unpublished results). In the context of the transpososome, however, diffuse signals within MuA subunits function synergistically, resulting in an affinity that rivals that of the ssrA tag. In fact, recent results using model substrates reveal that weak tags benefit the most from adaptor-mediated tethering (McGinness et al., 2006), and thus we suggest that the intrinsically weak MuA tag is important in allowing stimulation by other sequence-contacts upon multimer formation.

Why are ssrA-tagged proteins and MuA recognized by different mechanisms? We propose that the answer lies in how these two substrates evolved to fulfill their functions. The function of the ssrA tag is to target a highly diverse collection of failed translation products for rapid

degradation. Because there is no guarantee that these proteins contain additional degradation determinants, the ssrA tag must contain all of the determinants for strong binding to ClpX within a short sequence. On the other hand, phage Mu has evolved to efficiently use bacterial enzymes without eliciting a "host response" that could potentially destroy essential viral proteins such as MuA. The low ClpX affinity of MuA subunits should help prevent ClpXP from degrading the transposase before it has assembled into complexes. In contrast, the high affinity of the transpososome serves to focus the enzymatic activity of ClpX on a critical phage Mu substrate, as failure to disassemble the transpososome prevents phage DNA replication. This mechanism also provides a means to achieve complex remodeling rather than its complete destruction. Differential substrate recognition through substrate multimerization provides a versatile regulatory mechanism for the MuA monomer to switch from weak to strong ClpX binding upon formation of the Mu tetramer, without the need for covalent modification or the binding of heterologous adaptors.

Complex-specific recognition signals may be a widespread recognition mechanism.

There are many protein complexes in addition to MuA that are candidates to be recognized by ClpX in a manner more similar to the transpososome than to ssrA-tagged substrates. A proteomic study has identified many ClpX substrates that contain endogenous degradation tags (Flynn et al., 2003). Almost 60% of these substrates perform their biological function as subunits within a complex, either as a homomultimer (~25% of total substrates) or as a heteromultimer (~35%). Homomultimeric substrates include proteins such as the DNA-protective Dps and the tubulin-like protein FtsZ, whereas heteromultimeric substrates include

the F_1F_0-ATP synthase. Many of these substrates are known to have different structural and functional characteristics in their monomeric versus multimeric forms. Therefore, it is possible that these proteins also interact with ClpXP differentially as multimers.

There are many other examples where it becomes important for AAA+ enzymes to preferentially recognize a substrate in its multimeric, biologically active form. For example, the monomeric phage lambda replication protein λO, which tetramerizes to form the "O-some," has diffuse signals throughout the protein that target it for degradation by ClpXP (Gonciarz-Swiatek et al., 1999). Studies have shown that ClpXP recognizes the O-some differently than the monomer (Zylicz et al., 1998). Moreover, the distinct recognition modes for the O-some and λO monomers may be mediated by differential exposure of ClpX recognition signals in the complex. In eukaryotes, the AAA enzyme katanin uses ATP to sever and disassemble microtubules into tubulin αβ dimers. Tubulin dimers in solution cannot compete with microtubules for binding sites on katanin, suggesting that the microtubules have a higher affinity for the enzyme than the dimers (McNally and Vale, 1993). Therefore, similar to MuA, microtubules may present a different, high-affinity binding surface compared to tubulin dimers that guides katanin to specifically disassemble the multimeric form of the substrate.

In addition, the role of accessory domains such as the N-domain in mediating specific recognition of large protein complexes may be a common mechanism among AAA+ unfoldases that interact with higher order oligomers. Recent studies show that in some cases, the N-domain of the bacterial Hsp104 homolog ClpB is more strongly required for recognition and

disaggregation of large protein aggregates than smaller ones (Barnett et al., 2005). Studies on membrane fusion reveal that the N-domain of N-ethyl-maleimide sensitive factor (NSF) is necessary to mediate the binding of NSF to the large multimeric SNAP-SNARE complexes, an association that is required to mediate specific disassembly of SNAREs and to complete membrane fusion (Nagiec et al., 1995). As for ClpX, ClpB and NSF may specifically recognize the multimeric substrates by binding to complex-specific recognition signals in an N-domain-dependent manner. Taken together, our model of differential recognition of monomer and multimer substrates by AAA+ ATPases provides a mechanism to explain how these enzymes focus their activity on biologically relevant complexes. Higher resolution structural data and further analysis of substrate recognition signals will indicate that internal recognition motifs in other substrates play similar roles controlling substrate choice by AAA+ unfoldases and proteases.

EXPERIMENTAL PROCEDURES

DNA for transposition and cloning

$ClpX^{\Delta N}$ was constructed using the pET3a plasmid containing the *clpX* gene as the template and primers that amplified a region between residue 47 of ClpX and an internal RsrII site in the *clpX* gene. The resulting PCR fragment was subcloned into pET3a plasmid containing the *clpX* gene digested with NdeI and RsrII.

For variant MuA proteins, point mutations were introduced using the Quikchange kit (Strategene). For MuA truncation mutants 1-492, 1-574 and 1-655, the Quikchange kit was

used to introduce a stop codon at the corresponding C-terminal residue for each variant. The constructs for MuA truncations 575-663 and 575-659 were a gift from the Chaconas laboratory (Wu and Chaconas, 1995).

Protein purification

Wild type, mutant variants and truncations of MuA (Baker et al., 1991; Wu and Chaconas, 1995) and ClpP (Kim et al., 2000) and ClpX (Neher et al., 2003b) were purified as previously described. $ClpX^{\Delta N}$ was purified as described for wild-type ClpX.

Transpososome assembly

Transpososomes were assembled *in vitro* in the following solution: 25 mM Hepes (pH 7.6), 1 mM $MgCl_2$, 140 mM NaCl, 1 mM DTT, 15% glycerol, 20 μg/ml BSA and 12% DMSO. Transposition reactions contained 30 μg/ml pMK586, 130 nM HU protein, and 1 μM MuA. The reaction was carried out at 30 ^{o}C for 90 min.

Degradation and disassembly assays

ClpX alone or ClpX and ClpP were preincubated with ATP regeneration mix (ATP, creatine phosphate and creatine kinase) for 90 s at 30 ^{o}C prior to addition of substrate. Degradation was stopped by addition of 2.5X SDS loading buffer and freezing in liquid nitrogen. After SDS-PAGE, products were visualized with Sypro Orange (Invitrogen/ Molecular Probes) or transferred for Western Blotting. Disassembly reactions were stopped by addition of 100 mM EDTA, and DNA products were separated and visualized as described (Burton and Baker, 2003).

Determination of steady-state kinetic parameters

Transpososomes were made as a stock at 10 nM and diluted for addition to disassembly reaction in buffer containing 25 mM Hepes, 0.1 mM EDTA, 1 mM DTT, 10% glycerol, and 500 mM KCl. All reactions were carried out at 30 °C. For each disassembly reaction, $ClpP_{14}$ was added to $ClpX_6$ at a 2:1 ratio. All reactions contained 25 mM Hepes buffer (pH 7.6), 1 mM $MgCl_2$, 2.5 mM DTT, 110 mM KCl, 15% glycerol, 5 mM ATP, and ATP regeneration mix, including 0.5 mg/ml creatine kinase (Roche) and 20 mM disodium creatine phosphate (Roche). For each timepoint, the reaction was stopped by addition of EDTA to 50 mM and urea to 1 M, and DNA products were separated as described above. Controls were performed to ensure that the intensities of bands quantified were proportional to the amount of DNA disassembly product in the gel and that any remaining free MuA was present at too low a concentration to affect the rate of disassembly. K_M curves to determine half-maximal velocity for MuA monomer as a function of ClpXP concentration were carried out as described for transpososomes. K_M curves determined as a function of MuA substrate concentration were carried out as above, except the concentration of ClpXP was kept constant at 0.4 μM (0.4 μM $ClpX_6$, 0.8 μM $ClpP_{14}$). Reaction conditions were kept the same as described above for all concentrations of MuA tested.

MuA lysine modification and mass spectrometry

Stable transpososome complexes were purified on a Biogel-A column after high-salt challenge as described (Burton et al., 2001). Transpososomes or MuA monomers were reacted for various periods of time (from 1 minute to 1 hour) at room temperature with 10 to 30 mM sulfo NHS acetate in 100 mM $NaHCO_3$ (pH 8.5). The reaction was quenched with 50 mM Tris-HCl (pH

8). Native agarose-gel electrophoresis revealed no acetylation-dependent transpososome dissociation, even for the 1-hour reactions. Guanidine hydrochloride (GuHCl) was then added to 6 M final concentration, the sample was heated to 95 °C for 20 min, and then dialyzed overnight into 1 M GuHCl, 50 mM ammonium bicarbonate (pH 7.8), 1 mM $CaCl_2$, and then into the same buffer without GuHCl for tryptic digests. Samples were then digested with trypsin overnight using the Promega solution digest protocol and analyzed by LCMS. Alternatively, after denaturation, samples were dialyzed into 1 M GuHCl, 20 mM sodium acetate (pH 3.5) and then into the same buffer without GuHCl for peptic digests. Equal volumes of a 50% immobilized pepsin bead slurry (Pierce) equilibrated in the 20 mM sodium acetate buffer were added to the samples. After 1-5 minutes, the solutions were spun through a 2 micron filter to remove the pepsin beads. LC was a 1 or 2-hour reverse phase gradient from 5 to 55% acetonitrile with constant 0.1% formic acid. Tandem MS spectra were analyzed using SEQUEST and the MuA amino-acid sequence for peptides without modification and for peptides with an additional 42 daltons on lysine residues, expected for acetylation.

REFERENCES

Baker, T.A., Mizuuchi, M., and Mizuuchi, K. (1991). MuB protein allosterically activates strand transfer by the transposase of phage Mu. Cell *65*, 1003-1013.

Baker, T.A., and Sauer, R.T. (2006). ATP-dependent proteases of bacteria: recognition logic and operating principles. Trends Biochem Sci *31*, 647-653.

Barnett, M.E., Nagy, M., Kedzierska, S., and Zolkiewski, M. (2005). The amino-terminal domain of ClpB supports binding to strongly aggregated proteins. J Biol Chem *280*, 34940-34945.

Bolon, D.N., Wah, D.A., Hersch, G.L., Baker, T.A., and Sauer, R.T. (2004). Bivalent tethering of SspB to ClpXP is required for efficient substrate delivery: a protein-design study. Mol Cell *13*, 443-449.

Burton, B.M., and Baker, T.A. (2003). Mu transpososome architecture ensures that unfolding by ClpX or proteolysis by ClpXP remodels but does not destroy the complex. Chem Biol *10*, 463-472.

Burton, B.M., and Baker, T.A. (2005). Remodeling protein complexes: insights from the AAA+ unfoldase ClpX and Mu transposase. Protein Sci *14*, 1945-1954.

Burton, B.M., Williams, T.L., and Baker, T.A. (2001). ClpX-mediated remodeling of mu transpososomes: selective unfolding of subunits destabilizes the entire complex. Mol Cell *8*, 449-454.

Craigie, R., Mizuuchi, M., and Mizuuchi, K. (1984). Site-specific recognition of the bacteriophage Mu ends by the Mu A protein. Cell *39*, 387-394.

Curcio, M.J., and Derbyshire, K.M. (2003). The outs and ins of transposition: from mu to kangaroo. Nat Rev Mol Cell Biol *4*, 865-877.

Dougan, D.A., Weber-Ban, E., and Bukau, B. (2003). Targeted delivery of an ssrA-tagged substrate by the adaptor protein SspB to its cognate AAA+ protein ClpX. Mol Cell *12*, 373-380.

Flynn, J.M., Levchenko, I., Sauer, R.T., and Baker, T.A. (2004). Modulating substrate choice: the SspB adaptor delivers a regulator of the extracytoplasmic-stress response to the AAA+ protease ClpXP for degradation. Genes Dev *18*, 2292-2301.

Flynn, J.M., Levchenko, I., Seidel, M., Wickner, S.H., Sauer, R.T., and Baker, T.A. (2001). Overlapping recognition determinants within the ssrA degradation tag allow modulation of proteolysis. Proc Natl Acad Sci U S A *98*, 10584-10589.

Flynn, J.M., Neher, S.B., Kim, Y.I., Sauer, R.T., and Baker, T.A. (2003). Proteomic discovery of cellular substrates of the ClpXP protease reveals five classes of ClpX-recognition signals. Mol Cell *11*, 671-683.

Gonciarz-Swiatek, M., Wawrzynow, A., Um, S.J., Learn, B.A., McMacken, R., Kelley, W.L., Georgopoulos, C., Sliekers, O., and Zylicz, M. (1999). Recognition, targeting, and hydrolysis of the lambda O replication protein by the ClpP/ClpX protease. J Biol Chem *274*, 13999-14005.

Gonzalez, M., Rasulova, F., Maurizi, M.R., and Woodgate, R. (2000). Subunit-specific degradation of the UmuD/D' heterodimer by the ClpXP protease: the role of trans recognition in UmuD' stability. Embo J *19*, 5251-5258.

Gottesman, S., Roche, E., Zhou, Y., and Sauer, R.T. (1998). The ClpXP and ClpAP proteases degrade proteins with carboxy-terminal peptide tails added by the SsrA-tagging system. Genes Dev *12*, 1338-1347.

Hanson, P.I., and Whiteheart, S.W. (2005). AAA+ proteins: have engine, will work. Nat Rev Mol Cell Biol *6*, 519-529.

Herschlag, D., and Cech, T.R. (1990). Catalysis of RNA cleavage by the Tetrahymena thermophila ribozyme. 1. Kinetic description of the reaction of an RNA substrate complementary to the active site. Biochemistry *29*, 10159-10171.

Jones, J.M., Welty, D.J., and Nakai, H. (1998). Versatile action of Escherichia coli ClpXP as protease or molecular chaperone for bacteriophage Mu transposition. J Biol Chem *273*, 459-465.

Kim, Y.I., Burton, R.E., Burton, B.M., Sauer, R.T., and Baker, T.A. (2000). Dynamics of substrate denaturation and translocation by the ClpXP degradation machine. Mol Cell *5*, 639-648.

Kruklitis, R., Welty, D.J., and Nakai, H. (1996). ClpX protein of Escherichia coli activates bacteriophage Mu transposase in the strand transfer complex for initiation of Mu DNA synthesis. Embo J *15*, 935-944.

Kuo, C.F., Zou, A.H., Jayaram, M., Getzoff, E., and Harshey, R. (1991). DNA-protein complexes during attachment-site synapsis in Mu DNA transposition. Embo J *10*, 1585-1591.

Lavoie, B.D., Chan, B.S., Allison, R.G., and Chaconas, G. (1991). Structural aspects of a higher order nucleoprotein complex: induction of an altered DNA structure at the Mu-host junction of the Mu type 1 transpososome. Embo J *10*, 3051-3059.

Levchenko, I., Luo, L., and Baker, T.A. (1995). Disassembly of the Mu transposase tetramer by the ClpX chaperone. Genes Dev *9*, 2399-2408.

Levchenko, I., Seidel, M., Sauer, R.T., and Baker, T.A. (2000). A specificity-enhancing factor for the ClpXP degradation machine. Science *289*, 2354-2356.

Levchenko, I., Yamauchi, M., and Baker, T.A. (1997). ClpX and MuB interact with overlapping regions of Mu transposase: implications for control of the transposition pathway. Genes Dev *11*, 1561-1572.

Marshall-Batty, K.R., and Nakai, H. (2003). Trans-targeting of the phage Mu repressor is promoted by conformational changes that expose its ClpX recognition determinant. J Biol Chem *278*, 1612-1617.

McGinness, K.E., Baker, T.A., and Sauer, R.T. (2006). Engineering controllable protein degradation. Mol Cell *22*, 701-707.

McNally, F.J., and Vale, R.D. (1993). Identification of katanin, an ATPase that severs and disassembles stable microtubules. Cell *75*, 419-429.

Mhammedi-Alaoui, A., Pato, M., Gama, M.J., and Toussaint, A. (1994). A new component of bacteriophage Mu replicative transposition machinery: the Escherichia coli ClpX protein. Mol Microbiol *11*, 1109-1116.

Mizuuchi, K. (1983). In vitro transposition of bacteriophage Mu: a biochemical approach to a novel replication reaction. Cell *35*, 785-794.

Nagiec, E.E., Bernstein, A., and Whiteheart, S.W. (1995). Each domain of the N-ethylmaleimide-sensitive fusion protein contributes to its transport activity. J Biol Chem *270*, 29182-29188.

Nakai, H., Doseeva, V., and Jones, J.M. (2001). Handoff from recombinase to replisome: insights from transposition. Proc Natl Acad Sci U S A *98*, 8247-8254.

Nakai, H., and Kruklitis, R. (1995). Disassembly of the bacteriophage Mu transposase for the initiation of Mu DNA replication. J Biol Chem *270*, 19591-19598.

Neher, S.B., Flynn, J.M., Sauer, R.T., and Baker, T.A. (2003a). Latent ClpX-recognition signals ensure LexA destruction after DNA damage. Genes Dev *17*, 1084-1089.

Neher, S.B., Sauer, R.T., and Baker, T.A. (2003b). Distinct peptide signals in the UmuD and UmuD' subunits of UmuD/D' mediate tethering and substrate processing by the ClpXP protease. Proc Natl Acad Sci U S A *100*, 13219-13224.

Neuwald, A.F., Aravind, L., Spouge, J.L., and Koonin, E.V. (1999). AAA+: A class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes. Genome Res *9*, 27-43.

Park, E.Y., Lee, B.G., Hong, S.B., Kim, H.W., Jeon, H., and Song, H.K. (2007). Structural basis of SspB-tail recognition by the zinc binding domain of ClpX. J Mol Biol *367*, 514-526.

Pyle, A.M., and Green, J.B. (1994). Building a kinetic framework for group II intron ribozyme activity: quantitation of interdomain binding and reaction rate. Biochemistry *33*, 2716-2725.

Siddiqui, S.M. (2004). Dissecting the steps of substrate processing by the energy-dependent protease ClpXP. PhD Thesis, Department of Biology, Massachusetts Institute of Technology.

Surette, M.G., Buch, S.J., and Chaconas, G. (1987). Transpososomes: stable protein-DNA complexes involved in the in vitro transposition of bacteriophage Mu DNA. Cell *49*, 253-262.

Thibault, G., Yudin, J., Wong, P., Tsitrin, V., Sprangers, R., Zhao, R., and Houry, W.A. (2006). Specificity in substrate and cofactor recognition by the N-terminal domain of the chaperone ClpX. Proc Natl Acad Sci U S A *103*, 17724-17729.

van Gent, D.C., Mizuuchi, K., and Gellert, M. (1996). Similarities between initiation of V(D)J recombination and retroviral integration. Science *271*, 1592-1594.

Wah, D.A., Levchenko, I., Rieckhof, G.E., Bolon, D.N., Baker, T.A., and Sauer, R.T. (2003). Flexible linkers leash the substrate binding domain of SspB to a peptide module that stabilizes delivery complexes with the AAA+ ClpXP protease. Mol Cell *12*, 355-363.

Wojtyra, U.A., Thibault, G., Tuite, A., and Houry, W.A. (2003). The N-terminal zinc binding domain of ClpX is a dimerization domain that modulates the chaperone function. J Biol Chem *278*, 48981-48990.

Wu, Z., and Chaconas, G. (1995). A novel DNA binding and nuclease activity in domain III of Mu transposase: evidence for a catalytic region involved in donor cleavage. Embo J *14*, 3835-3843.

Yuan, J.F., Beniac, D.R., Chaconas, G., and Ottensmeyer, F.P. (2005). 3D reconstruction of the Mu transposase and the Type 1 transpososome: a structural framework for Mu DNA transposition. Genes Dev *19*, 840-852.

Zhou, Y., Gottesman, S., Hoskins, J.R., Maurizi, M.R., and Wickner, S. (2001). The RssB response regulator directly targets sigma(S) for degradation by ClpXP. Genes Dev *15*, 627-637.

Zylicz, M., Liberek, K., Wawrzynow, A., and Georgopoulos, C. (1998). Formation of the preprimosome protects lambda O from RNA transcription-dependent proteolysis by ClpP/ClpX. Proc Natl Acad Sci U S A *95*, 15259-15263.

ACKNOWLEDGMENTS

We wish to dedicate this paper to the memory of Arthur Kornberg. We would like to thank B.M. Burton, B.O. Cezairliyan, A.S. Meyer and F. Solomon for critical reading of the manuscript and members of the Baker and Sauer labs for reagents and helpful discussions. This work was supported by NIH grants GM-49224 and AI-16892. A.H.A. was supported by an International Fellowship from the American Association of University Women. T.A.B. is an employee of the Howard Hughes Medical Institute.

CHAPTER 3

DISCUSSION: AUTO-TETHERING AS A SUBSTRATE SELECTION MECHANISM FOR RECOGNITION OF MULTIMERIC SUBSTRATES BY THE AAA+ UNFOLDASE CLPX

The work in this thesis has aimed to explore and expand upon the myriad of mechanisms that are utilized by Clp/Hsp100 enzymes to regulate when and where they act upon their substrates in the cell. Specifically, my work with the Clp/Hsp100 unfoldase ClpX and its substrate the MuA transposase has revealed mechanistic insight into auto-tethering, one of the less well understood substrate selection strategies utilized by this family of enzymes. This work has implications not only for other multimeric substrates of ClpX but also for other AAA+ remodeling enzymes that need to specifically bind to correctly assembled multimeric complexes.

Auto-tethering can be thought of as a specialized type of tethering, mediated not by an adaptor protein that is different from the substrate, but rather by the substrate itself when it is in multimeric form. Because auto-tethering requires that the substrate play a dual role in recognition by the enzyme – it must tether itself and also be unfolded – mechanisms must exist to adjust the affinity of the Clp/Hsp100 enzyme for different multimeric forms of the substrate in accordance with the requirements for remodeling or degradation in the cell. Although UmuD' is a processed form of UmuD, the delivery of UmuD' to ClpXP by UmuD can be classified as a type of auto-tethering mechanism. In this case, ClpXP preferentially recognizes the dimer UmuD/D', as the division of labor between subunits is determined by the differential presence of the tethering XB-like sequence in the UmuD and of enzyme-pore binding signals on the UmuD' subunit (Gonzalez et al., 2000; Neher et al., 2003b). This arrangement allows UmuD/D' to satisfy the dual role of adaptor and substrate accordingly.

The work in this thesis has shown that conformational changes that accompany multimerization of the enzyme also play a role in the division of labor among subunits of

multimeric substrates that mediate auto-tethering. In the case of MuA transposase, multimerization to form the Mu transpososome exposes residues within the subunits of the complex that allow ClpX to make favorable contacts that are unique to the tetramer. In this way, ClpX can preferentially bind to the form of MuA transposase that requires remodeling. This binding, like tethering, requires the presence of the ClpX N domain. The discovery of this requirement for the N domain sheds light on the variety of ways that the N domain participates in substrate selection. We now know that the N domain not only assists ClpX in discriminating between different classes of tags and binding to adaptors, but also allows ClpX to discern between different multimeric forms of a substrate.

Diverse modes of substrate-enzyme contacts mediate recognition

The ssrA tag has been used extensively to characterize enzyme-tag contacts that are required for substrate engagement and unfolding (Flynn et al., 2001; Gottesman et al., 1998; Hersch et al., 2004; Kenniston et al., 2003; Kim et al., 2000; Martin et al., 2007, 2008; Siddiqui et al., 2004). However, experiments show that a growing number of substrates bind to ClpX via diffuse, weak signals that act cooperatively to create stable substrate-enzyme complexes in a mechanism different from that described for ssrA (Abdelhakim et al., 2008; Hoskins and Wickner, 2006; Mettert and Kiley, 2005; Ryan et al., 2002; Studemann et al., 2003). Binding via multiple weak signals provides the potential for gradations of binding and also presents opportunities for timing degradation appropriately. Degradation will be signaled only upon the formation of a suitable binding surface concomitant with the assembly of the correct substrate complex. The mechanism of this type of binding is described in mechanistic detail for the ClpX-

transpososome interaction in this thesis. Like MuA, many of the substrates that likely use a combination of weak signals to bind to the enzyme form biologically relevant stable multimers, suggesting that diffuse degradation signals may be a common mechanism mediating substrate-enzyme interactions for this multimeric substrate class.

Whether a substrate has evolved to utilize diffuse, weak signals or strong-binding, localized ssrA-like signals may depend on the required timing and biological function of that substrate in the cell. ssrA is required to mediate strong binding within a short number of residues, as it may be appended to potentially toxic ribosome truncation products that do not contain intrinsic degradation signals. Other substrates containing ssrA-like degradation tags may also require similarly swift or continual degradation in the cell. This type of degradation signal is seen with several ClpXP substrates involved in the SOS response, where rapid changes in the levels of enzymes regulating the response are critical to prevent DNA damage and potential loss of cell viability. For example, upon auto-endoproteolytic cleavage of the LexA repressor induced by RecA, the DNA binding LexA N-terminal fragment must be cleared from the cell quickly and efficiently so as to allow expression of SOS genes as quickly as possible. The post-cleavage N-terminal fragment of LexA contains an ssrA-like degradation tag and is cleared by ClpXP rapidly *in vivo*, therefore preventing any potential interference with expression of SOS genes (Neher et al., 2003a). Failure to clear the LexA N-terminal fragment by ClpXP results in a modest decrease in cell viability at high doses of UV irradiation (Neher et al., 2003a). Similarly, RecN, another SOS enzyme involved in repair of double stranded DNA breaks, is continually degraded by ClpXP via an ssrA-like tag at its C-terminus (Nagashima et al., 2006; Neher et al., 2006). Increased levels of RecN during DNA damage are due to dramatic increases in levels of

expression of the *recN* gene rather than decreases in the rate of degradation (Nagashima et al., 2006; Neher et al., 2006). Upon resolution of damage, RecN must be cleared from the cell rapidly by ClpXP, as high levels of RecN result in loss of cell viability (Nagashima et al., 2006). Therefore, for both the LexA N-terminal autoproteolytic fragment and RecN, rapid and continual clearing of the substrate mediated by a strong-binding a localized tag is necessary for optimal responses to DNA damage. This mode of degradation by ClpXP is rapid and efficient, as it does not require intermediary steps such as substrate multimerization or conformational changes.

On the other hand, it is more functionally beneficial for some multimeric substrates to bind to ClpXP using weak, diffuse signals that act cooperatively to bind with high affinity to the enzyme, as has been described for MuA. The stationary-phase enzyme Dps, for example, is a ClpXP substrate that may use strategies similar to MuA to bind to ClpXP. During exponential phase, Dps is continually degraded by ClpXP. Upon entry into stationary phase however, Dps is stabilized and becomes one of the most highly abundant proteins in the cell, present in approximately 180,000 copies (Ali Azam et al., 1999). Under prolonged starvation conditions, Dps binds to DNA in a large, extremely stable multimeric "biocrystal", which functions to protect the DNA from oxidative damage during starvation and stress (Wolf et al., 1999). Upon exit from stationary phase, Dps must be rapidly degraded by ClpXP. Several lines of evidence have suggested that Dps requires the use of an adaptor for its degradation during exit from stationary phase (J. M. Flynn, unpublished results) (Stephani et al., 2003). In fact, like MuA, Dps may act as its own adaptor. Evidence for this comes from in vitro observations, which show that, like the transpososome, Dps has an almost absolute requirement for the ClpX N domain

for recognition and degradation by the enzyme (A. S. Meyer, personal communication). This requirement suggests that Dps may mediate its own degradation via auto-tethering. Considering the high levels of Dps requiring degradation, this strategy is economical in times of starvation and stress, as it saves the cell the energy costs of transcribing, translating and potentially degrading or inactivating similarly high levels of a heterologous adaptor protein. Additionally, this mechanism provides a simple yet effective method of regulating the timing of Dps regulation at entry of exponential phase. The Dps stationary phase biocrystal, which has been shown to have different physical properties from the non-DNA bound Dps dodecamer (Frenkiel-Krispin et al., 2004; Wolf et al., 1999), would minimally require a conformational change promoting auto-tethering and degradation upon receiving signals ushering exit from stationary phase. In this manner, degradation of Dps in the cell can be initiated in a timely and rapid manner.

Other multimeric substrates that bind to ClpXP using diffuse weak signals include the error-prone polymerase pol V subunit UmuD. Under conditions of DNA damage, UmuD is processed in the cell to form UmuD', an N-terminally self-cleaved version of UmuD that is formed upon stimulation by RecA. UmuD and UmuD' dimerize to form three types of dimers: two homodimers ($UmuD_2$ and $UmuD'_2$) and a heterodimer (UmuD/D') which, among the three dimers, is formed preferentially. ClpXP degrades the UmuD' subunit from the UmuD/D' heterodimer most efficiently, although it can also recognize $UmuD_2$ more weakly (Frank et al., 1996; Gonzalez et al., 2000; Neher et al., 2003b) An explanation for this hierarchy in preference of degradation by ClpXP can be found in the functions of these dimers in the cell. In vivo, $UmuD'_2$, UmuD/UmuD' or $UmuD_2$ can bind to UmuC, another essential subunit of the error-

prone polymerase. However, only UmuD'$_2$C is mutagenically active and able to mediate error-prone DNA replication. UmuC is the limiting factor in vivo for formation of the UmuD'$_2$C complex. Furthermore, UmuD is only inefficiently cleaved to form UmuD', making the non-mutagenically active UmuD$_2$ and UmuD/D' the dominant dimer forms in the cell.

Although the error-prone pol V increases overall survival during DNA damage, the polymerase introduces mutations into the genome during replication and therefore the cell must use this system sparingly. Additionally, any complexes that are not required in the cell after response to DNA damage must be removed, as the inaccuracies in DNA replication caused by pol V can potentially prove to be lethal. The cell must therefore balance its opposing needs to avoid the sequestration of UmuC in inactive complexes upon DNA damage and to keep the number of mutagenically active and potentially harmful error-prone polymerases to a minimum. Multiple mechanisms, including degradation, are used to keep the cell in balance. UmuD$_2$ homodimers are kept to minimum levels and are prevented from sequestering UmuC primarily via Lon-mediated degradation, and to a lesser extent by ClpXP (Frank et al., 1996; Gonzalez et al., 1998; Neher et al., 2003b). ClpXP mainly targets the UmuD/D' complexes for degradation (Gonzalez et al., 2000). In this case, it is advantageous for ClpXP to recognize the UmuD/D' complex using a combination of weak signals. Within the UmuD/D' dimers, the UmuD N-terminal peptide tethers the complex to ClpXP via the ClpX N domain. Upon tethering by UmuD, UmuD' binds to the processing pore via multiple weak signals and becomes degraded by ClpXP, leaving behind a UmuD monomer (Gonzalez et al., 2000; Neher et al., 2003b). ClpXP can therefore specifically target the UmuD/D' complex as it is the only heterodimer which provides a surface on the substrate containing signals for tethering and processing. This mode

of regulation would not be possible if both UmuD and UmuD' contained a strong localized signal, as this would potentially make all possible combinations of the UmuD or UmuD' dimer a good substrate for ClpXP. Therefore, in this way, the cell can keep error-prone polymerase complex levels to a minimum and also enrich for mutagenically active $UmuD'_2$ homodimers upon DNA damage by targeting $UmuD_2$ and UmuD/D' for degradation. Additionally, this provides a way for the cell to rid itself specifically of the mutagenically active UmuD' after DNA damage has been resolved.

A number of other substrates regulating complex biological pathways also bind to ClpXP using complex or bipartite degradation signals. For example, CtrA from *C. crescentus*, the master transcriptional regulator required for the transition from the G1 to S phase in this bacterium, is regulated at many levels including phosphorylation, multimerization, localization in the cell and degradation (Jensen et al., 2002; Ryan and Shapiro, 2003). CtrA contains a bipartite ClpXP degradation signal. This mode of binding to the protease may provide an additional opportunity for fine-tuning binding to the ClpXP protease and hence levels of CtrA in the cell in accordance with the cell's needs (Ryan et al., 2002). Another important transcriptional regulator, σ^S, which controls the transcription of stationary-phase specific genes, is also controlled at multiple levels in the cell. Although σ^S alone is a poor ClpXP substrate, it is thought to have at least a bipartite ClpXP degradation signal, and requires the use of the adaptor RssB for delivery to the protease (Studemann et al., 2003; Zhou et al., 2001). These multiple checks in the degradation pathway of σ^S may be necessary to ensure degradation of σ^S only when it is not needed in the cell. FNR is another transcriptional regulator that contains at least two degradation signals that are required for recognition by ClpXP, located at both the N

and the C termini of the substrate protein (Mettert and Kiley, 2005). FNR was identified as a substrate in a proteomic study utilizing a ClpP substrate trap (Flynn et al., 2003). This study has also identified many other substrates that potentially contain more than one degradation signal and whose degradation may be controlled at the level of multimerization.

Autotethering of MuA functions in conjunction with other regulatory mechanisms

Autotethering may exist as an additional layer of regulation that functions in conjunction with other regulatory mechanisms to ensure the correct timing and location of recognition of the transpososome by ClpXP. It is known for example that the MuA activator protein MuB binds to a region in the C-terminal domain of MuA that overlaps with the ClpX tag, and in fact can inhibit the recognition by ClpX of the monomer and complex in vitro (Levchenko et al., 1997). This overlap between the binding sites of ClpX and MuB may function to protect MuA from disassembly or degradation by ClpX until it has assembled into correctly formed transpososome complexes required to transition into Mu DNA replication. A potential model that coordinates MuB-mediated inhibition of MuA-ClpX recognition and MuA-ClpX autotethering may operate whereby the affinity of MuA for MuB decreases at each step of the transposition pathway, coupled with the increase in MuA-ClpX affinity provided by autotethering. In this way, both positive (autotethering) and negative (MuB inhibition) regulatory mechanisms ensure a robust regulatory pathway that ensures correct timing of remodeling in the cell.

The AAA+ protease FtsH also recognizes MuA in the cell and continuously degrades it throughout the phage lytic cycle (Gama et al., 1990). However, unlike a *clpX* null host, which almost completely inhibits replication of phage Mu in vivo, the absence of FtsH in vivo does not

have any effect on phage replication, suggesting that FtsH targets a form of MuA that is not required for the transition into the replicative cycle of the phage. FtsH may therefore target the monomeric form of MuA to keep it at minimum levels in the cell; in this way, the phage avoids occupying ClpX with excess MuA monomers that may inhibit disassembly of transpososomes, which are present at relatively low amounts in the cell (~100 transpososomes/ cell in one lytic cycle). Alternatively, FtsH may function to degrade incorrectly assembled transpososomes that cannot be remodeled into a functional replisome by ClpX, therefore allowing ClpX to focus on those correctly assembled complexes that can transition efficiently into replication.

It is becoming evident that ClpXP uses a variety of distinct binding modes to recognize substrates in a fashion compatible with the required regulation in the cell. This thesis has provided mechanistic insight into ClpXP-MuA interactions that are mediated via multiple weak degradation signals in a combinatorial manner, a mode of binding important for the correct timing and geometry of remodeling in vivo. MuA will continue to be a paradigmatic "auto-tethering" ClpXP substrate, and can offer deeper insights into how enzyme-substrate contacts modulate degradation via multimerization, conformational change and multiple cooperative contacts. There are still many questions that remain to be answered regarding how this type of complex recognition is mediated by ClpXP, and how it differs from recognition of ssrA-like degradation tags. A subset of these questions, and how MuA can be used to gain insight into function and mechanism, are listed and discussed below.

What role does the MuA tag play in mediating multimeric recognition?

This thesis has shown that a degradation tag residue in MuA, R658, and to a lesser extent K660, plays a role in mediating ClpX-MuA interactions only in the context of the transpososome. Furthermore, in vitro experiments using fusion proteins show that the MuA tag binds more tightly to ClpX in the context of a dimer than as a monomer (A. Abdelhakim, unpublished results; Kim et al., 2000). This enhancement in recognition is not seen with ssrA tag substrates. What physical characteristics of the MuA tag allow it to modulate recognition according to the multimeric state of the substrate? An answer to this question may be gleaned by examining properties such as the number and location of MuA binding sites on the ClpX enzyme and more careful determination of affinity constants for MuA tags in the context of monomers, dimers and higher order multimers not related to the transpososome. A possible explanation for the modulation of MuA tag binding by multimerization may be multiple MuA binding sites on the enzyme including the substrate processing pore and the N domain. This multivalent binding can also explain why ssrA, which does not require the N domain and binds to ClpX with a stoichiometry of 1 peptide:1 ClpX hexamer (Kim et al., 2000; Siddiqui, 2004; Wojtyra et al., 2003), cannot provide binding enhancement when present in multimeric form. Multivalent binding by the MuA tag may be a property that is shared by other ClpX-recognition tags that contain an overall positive charge.

How does the MuA tag contact ClpX?

Although this question is intimately related to the previous one, it is important to gain a full understanding of the binding mode of MuA to ClpX. Previous results have shown that MuA and

ssrA have some similarities in binding to ClpX, and also some key differences. MuA and ssrA are similar in that both require the terminal two hydrophobic residues for binding to ClpX (AA-$^{COO-}$ for ssrA; AI-$^{COO-}$ for MuA) (Abdelhakim et al., 2008; Flynn et al., 2001). It will therefore be interesting to determine why, despite this similarity, the ssrA and MuA monomer tags have such a different binding affinities to ClpX (1uM for ssrA tag fused to a model reporter protein; >100uM for the last 8 residues of MuA fused to a model reporter protein (λcI-MuA), A. Abdelhakim, unpublished results and Kim et al., 2000; the difference in affinity between the MuA tag in the context of the MuA monomer (~10uM) and λcI-MuA may be due to additional contacts that ClpX makes outside of the MuA tag in full-length MuA). The difference in the binding affinities of the MuA and ssrA tags may be explained by observed differences in the binding modes of these two recognition sequences. For example, only ssrA requires the free carboxyl group at the C-terminus for binding to the enzyme (S. Bissonette, personal communication; Flynn et al., 2001). Additionally, several studies have shown that the MuA and ssrA tags have differential requirements for ClpX substrate binding loops, including the GYVG motif and the RKH loop (Farrell et al., 2007; Siddiqui et al., 2004). Other experiments have shown that the residue antepenultimate to the two C-terminal hydrophobic residues can also affect binding affinity of the tag to ClpX (P. Chien, personal communication). Parsing out and quantifying the contributions of these differences to the relative affinities of ssrA and MuA to ClpX can be insightful in determining how ClpX can distinguish between different classes of tag despite superficial similarities in their sequence.

Where on the ClpX N-domain does the MuA tag interact?

It is very likely that the common denominator essential for mediating multimerization-dependent recognition, cooperativity of multiple weak signals on substrates and auto-tethering is the ClpX N-domain. The ClpX N-domain is a very intriguing domain that possesses the paradoxical abilities of specifically binding very defined sequences, such as the XB motif in SspB, as well as promiscuously binding a wide variety of sequences that may have little in common besides an overall positive charge or hydrophobic nature (Dougan et al., 2003; Park et al., 2007; Thibault et al., 2006; Wah et al., 2003). The MuA tag can be extremely helpful in determining how the N-domain can mediate these two seemingly opposing binding modes. Although fluorescence anisotropy has so far been unsuccessful in determining the molecular contacts of MuA and the N-domain due to the weak nature of these interactions, other methods can be used to determine the structural basis of multiple cooperative contacts mediated by the N-domain. For example, the N-domain has been crystallized in complex with the XB peptide, leading to a molecular map of the interactions necessary for this type of specific binding (Park et al., 2007). In a similar fashion, crystallography may be used to determine the molecular basis for MuA tag binding to the N-domain. The weak nature of MuA tag binding to the N-domain may be overcome by engineering a "better" MuA tag that binds more tightly to the N-domain but still retains binding properties unique to this tag. Alternatively, the peptide can be linked to the N-domain using crosslinking chemicals or recombinant techniques to increase the effective concentration of the peptide for the domain.

In addition, it has been shown that the N-domain plays some role in mediating binding of all other non-ssrA tags to ClpX, all of which contain an overall positive charge (Siddiqui, 2004). Inspection of the N-domain shows that there is a prominent negatively charged surface

composed of residues E37, E38, D41, D45, E49 and E50, present on each N-domain monomer in ClpX and making a total of 6 negatively charged surfaces in the ClpX hexamer (Figure 1). Mutation of E38, D41 and D45 has no effect on binding of SspB to ClpX (Thibault et al., 2006); however these residues may play an important role in the preference of the N-domain for positively charged residues and tags, including MuA. Such electrostatic interactions can be useful in mediation of multimeric and combinatorial recognition, due to the non-stringent and potentially cooperative nature of these types of interactions. Site directed mutagenesis of these residues, as well as perhaps other residues in the N-domain, and functional assays of these mutants with MuA as well as other tags and adaptors may be useful in creating a structural map to detail the repertoire of interactions that the N-domain can potentially mediate.

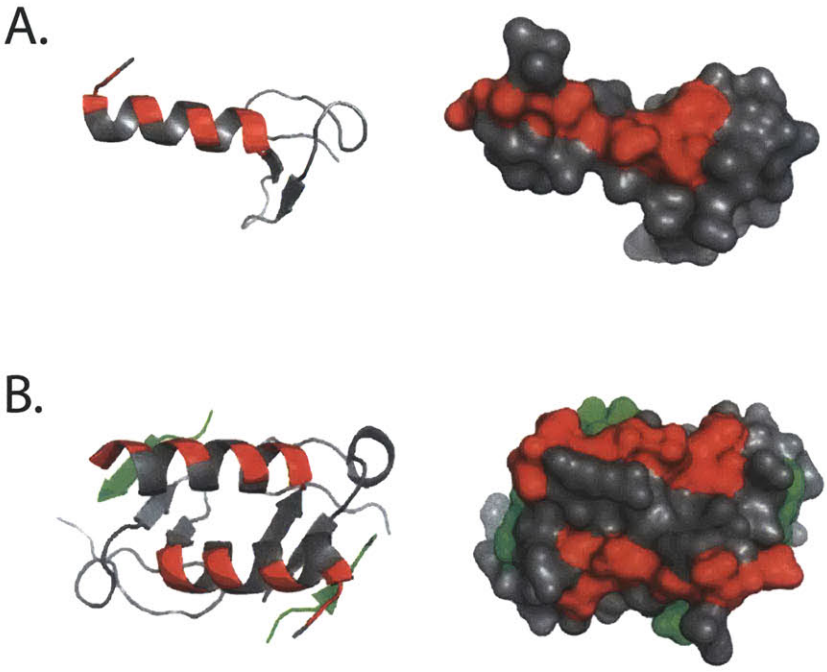


Figure 1. A. Ribbon and corresponding space-filling structures of the ClpX N-domain monomer, with E37, E38, D41, D45, E49 and E50 colored in red, to highlight the negatively charged surface of the N-domain. This surface may bind positively charged residues in degradation tags and substrates, and may explain the preference of the N-domain for positively charged residues. B. The residues highlighted in panel A are highlighted in the biologically relevant dimeric form of the ClpX N-domain. This structure also contains the XB peptide in green, which binds on the face of the dimer opposite to the negatively charged face. Structures from (Park et al., 2007)

REFERENCES

Abdelhakim, A.H., Oakes, E.C., Sauer, R.T., and Baker, T.A. (2008). Unique contacts direct high-priority recognition of the tetrameric Mu transposase-DNA complex by the AAA+ unfoldase ClpX. Mol Cell *30*, 39-50.

Ali Azam, T., Iwata, A., Nishimura, A., Ueda, S., and Ishihama, A. (1999). Growth phase-dependent variation in protein composition of the Escherichia coli nucleoid. J Bacteriol *181*, 6361-6370.

Dougan, D.A., Weber-Ban, E., and Bukau, B. (2003). Targeted delivery of an ssrA-tagged substrate by the adaptor protein SspB to its cognate AAA+ protein ClpX. Mol Cell *12*, 373-380.

Farrell, C.M., Baker, T.A., and Sauer, R.T. (2007). Altered specificity of a AAA+ protease. Mol Cell *25*, 161-166.

Flynn, J.M., Levchenko, I., Seidel, M., Wickner, S.H., Sauer, R.T., and Baker, T.A. (2001). Overlapping recognition determinants within the ssrA degradation tag allow modulation of proteolysis. Proc Natl Acad Sci U S A *98*, 10584-10589.

Flynn, J.M., Neher, S.B., Kim, Y.I., Sauer, R.T., and Baker, T.A. (2003). Proteomic discovery of cellular substrates of the ClpXP protease reveals five classes of ClpX-recognition signals. Mol Cell *11*, 671-683.

Frank, E.G., Ennis, D.G., Gonzalez, M., Levine, A.S., and Woodgate, R. (1996). Regulation of SOS mutagenesis by proteolysis. Proc Natl Acad Sci U S A *93*, 10291-10296.

Frenkiel-Krispin, D., Ben-Avraham, I., Englander, J., Shimoni, E., Wolf, S.G., and Minsky, A. (2004). Nucleoid restructuring in stationary-state bacteria. Mol Microbiol *51*, 395-405.

Gama, M.J., Toussaint, A., and Pato, M.L. (1990). Instability of bacteriophage Mu transposase and the role of host Hfl protein. Mol Microbiol *4*, 1891-1897.

Gonzalez, M., Frank, E.G., Levine, A.S., and Woodgate, R. (1998). Lon-mediated proteolysis of the Escherichia coli UmuD mutagenesis protein: in vitro degradation and identification of residues required for proteolysis. Genes Dev *12*, 3889-3899.

Gonzalez, M., Rasulova, F., Maurizi, M.R., and Woodgate, R. (2000). Subunit-specific degradation of the UmuD/D' heterodimer by the ClpXP protease: the role of trans recognition in UmuD' stability. EMBO J *19*, 5251-5258.

Gottesman, S., Roche, E., Zhou, Y., and Sauer, R.T. (1998). The ClpXP and ClpAP proteases degrade proteins with carboxy-terminal peptide tails added by the SsrA-tagging system. Genes Dev *12*, 1338-1347.

Hersch, G.L., Baker, T.A., and Sauer, R.T. (2004). SspB delivery of substrates for ClpXP proteolysis probed by the design of improved degradation tags. Proc Natl Acad Sci U S A *101*, 12136-12141.

Hoskins, J.R., and Wickner, S. (2006). Two peptide sequences can function cooperatively to facilitate binding and unfolding by ClpA and degradation by ClpAP. Proc Natl Acad Sci U S A *103*, 909-914.

Jensen, R.B., Wang, S.C., and Shapiro, L. (2002). Dynamic localization of proteins and DNA during a bacterial cell cycle. Nat Rev Mol Cell Biol *3*, 167-176.

Kenniston, J.A., Baker, T.A., Fernandez, J.M., and Sauer, R.T. (2003). Linkage between ATP consumption and mechanical unfolding during the protein processing reactions of an AAA+ degradation machine. Cell *114*, 511-520.

Kim, Y.I., Burton, R.E., Burton, B.M., Sauer, R.T., and Baker, T.A. (2000). Dynamics of substrate denaturation and translocation by the ClpXP degradation machine. Mol Cell *5*, 639-648.

Levchenko, I., Yamauchi, M., and Baker, T.A. (1997). ClpX and MuB interact with overlapping regions of Mu transposase: implications for control of the transposition pathway. Genes Dev *11*, 1561-1572.

Martin, A., Baker, T.A., and Sauer, R.T. (2007). Distinct static and dynamic interactions control ATPase-peptidase communication in a AAA+ protease. Mol Cell *27*, 41-52.

Martin, A., Baker, T.A., and Sauer, R.T. (2008). Diverse pore loops of the AAA+ ClpX machine mediate unassisted and adaptor-dependent recognition of ssrA-tagged substrates. Mol Cell *29*, 441-450.

Mettert, E.L., and Kiley, P.J. (2005). ClpXP-dependent proteolysis of FNR upon loss of its O2-sensing [4Fe-4S] cluster. J Mol Biol *354*, 220-232.

Nagashima, K., Kubota, Y., Shibata, T., Sakaguchi, C., Shinagawa, H., and Hishida, T. (2006). Degradation of Escherichia coli RecN aggregates by ClpXP protease and its implications for DNA damage tolerance. J Biol Chem *281*, 30941-30946.

Neher, S.B., Flynn, J.M., Sauer, R.T., and Baker, T.A. (2003a). Latent ClpX-recognition signals ensure LexA destruction after DNA damage. Genes Dev *17*, 1084-1089.

Neher, S.B., Sauer, R.T., and Baker, T.A. (2003b). Distinct peptide signals in the UmuD and UmuD' subunits of UmuD/D' mediate tethering and substrate processing by the ClpXP protease. Proc Natl Acad Sci U S A *100*, 13219-13224.

Neher, S.B., Villen, J., Oakes, E.C., Bakalarski, C.E., Sauer, R.T., Gygi, S.P., and Baker, T.A. (2006). Proteomic profiling of ClpXP substrates after DNA damage reveals extensive instability within SOS regulon. Mol Cell *22*, 193-204.

Park, E.Y., Lee, B.G., Hong, S.B., Kim, H.W., Jeon, H., and Song, H.K. (2007). Structural basis of SspB-tail recognition by the zinc binding domain of ClpX. J Mol Biol *367*, 514-526.

Ryan, K.R., Judd, E.M., and Shapiro, L. (2002). The CtrA response regulator essential for Caulobacter crescentus cell-cycle progression requires a bipartite degradation signal for temporally controlled proteolysis. J Mol Biol *324*, 443-455.

Ryan, K.R., and Shapiro, L. (2003). Temporal and spatial regulation in prokaryotic cell cycle progression and development. Annu Rev Biochem *72*, 367-394.

Siddiqui, S.M. (2004). Dissecting the steps of substrate processing by the energy-dependent protease ClpXP. Ph.D. Thesis, Department of Biology, Massachusetts Institute of Technology.

Siddiqui, S.M., Sauer, R.T., and Baker, T.A. (2004). Role of the processing pore of the ClpX AAA+ ATPase in the recognition and engagement of specific protein substrates. Genes Dev *18*, 369-374.

Stephani, K., Weichart, D., and Hengge, R. (2003). Dynamic control of Dps protein levels by ClpXP and ClpAP proteases in Escherichia coli. Mol Microbiol *49*, 1605-1614.

Studemann, A., Noirclerc-Savoye, M., Klauck, E., Becker, G., Schneider, D., and Hengge, R. (2003). Sequential recognition of two distinct sites in sigma(S) by the proteolytic targeting factor RssB and ClpX. EMBO J *22*, 4111-4120.

Thibault, G., Yudin, J., Wong, P., Tsitrin, V., Sprangers, R., Zhao, R., and Houry, W.A. (2006). Specificity in substrate and cofactor recognition by the N-terminal domain of the chaperone ClpX. Proc Natl Acad Sci U S A *103*, 17724-17729.

Wah, D.A., Levchenko, I., Rieckhof, G.E., Bolon, D.N., Baker, T.A., and Sauer, R.T. (2003). Flexible linkers leash the substrate binding domain of SspB to a peptide module that stabilizes delivery complexes with the AAA+ ClpXP protease. Mol Cell *12*, 355-363.

Wojtyra, U.A., Thibault, G., Tuite, A., and Houry, W.A. (2003). The N-terminal zinc binding domain of ClpX is a dimerization domain that modulates the chaperone function. J Biol Chem *278*, 48981-48990.

Wolf, S.G., Frenkiel, D., Arad, T., Finkel, S.E., Kolter, R., and Minsky, A. (1999). DNA protection by stress-induced biocrystallization. Nature *400*, 83-85.

Zhou, Y., Gottesman, S., Hoskins, J.R., Maurizi, M.R., and Wickner, S. (2001). The RssB response regulator directly targets sigma(S) for degradation by ClpXP. Genes Dev *15*, 627-637.

APPENDIX I

DIVISION OF LABOR AMONG SUBUNITS IN THE TRANSPOSOSOME FOR REMODELING BY CLPX

INTRODUCTION

The Mu transpososome is an extremely stable and asymmetric complex that is the product of transposition in the phage Mu (see Chapter 2). The transpososome is inhibitory to the host DNA replication machinery *in vivo* and must be remodeled by the AAA+ unfoldase ClpX to form a complex that promotes phage Mu genome replication (Jones and Nakai, 1997; Jones et al., 1998; Kruklitis et al., 1996; Levchenko et al., 1995). Although the transpososome is a tetramer, ClpX remodels and destabilizes the transpososome by unfolding only one subunit from the transpososome (Burton and Baker, 2003). Furthermore, footprinting experiments show that this subunit seems to be extracted from the "left" side of the complex. It is however unclear whether this subunit is extracted from the L1 or the L2 DNA site as footprinting patterns on both DNA sites are altered upon remodeling (Burton and Baker, 2003). In addition, it is possible that unfolding of a subunit on the right side of the complex could lead to observed footprinting changes on the left, due to the interwoven structure of the transpososome. Finally, it is known that one or more subunits in the transpososome act as internal "adaptors" that mediate binding to ClpX with high affinity, but it is not known which subunit fulfills this function, or how many (Abdelhakim et al., 2008).

We sought to determine the division of labor within the transpososome complex (which subunit on the left side is selected for unfolding? Which subunit mediates the auto-tethering "adaptor" function?) using an experimental set up previously described to target subunits to specific DNA sites within the tetrameric transpososome complex (Namgoong et al., 1998). In this system, an altered-specificity MuA mutant (MuA R146V) binds to Mu DNA sites engineered to contain a compensatory mutation facilitating MuA R146V binding and in the context of a

plasmid formation of transpososomes containing different MuA subunits (Mariconda et al., 2000; Namgoong and Harshey, 1998; Namgoong et al., 1998). In this way, specific MuA subunits can be targeted to any one of the four DNA sites in the complex (Figure 1A). This experimental design is ideal to parse apart the different roles which MuA subunits fulfill in the remodeling reaction by ClpX.

This project is currently at a stage where optimization of the system is still underway and some tentative conclusions about the division of labor of the transpososome subunits can be made (see below). The goal for the next two to three months will be to obtain more complete picture of the geometry of disassembly based on this experimental setup.

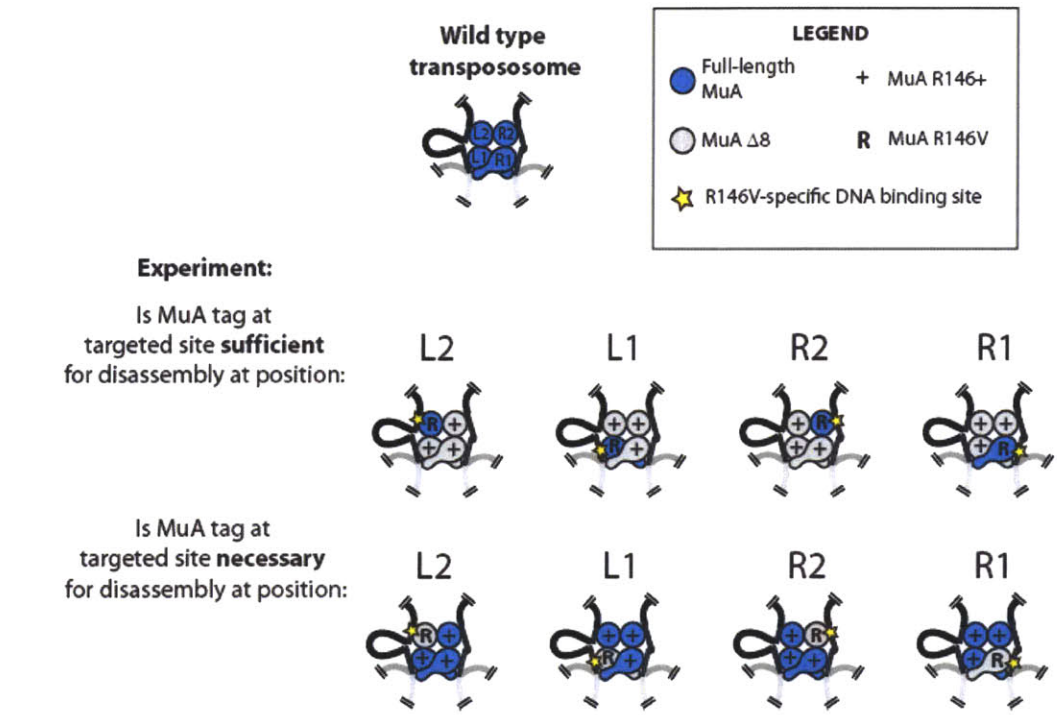
EXPERIMENTAL STRATEGY

Several MuA mutants are used in these experiments to determine which subunits in the transpososome mediate unfolding, and which mediate auto-tethering. To determine which subunits are required for unfolding, two complementary experiments are performed: first, a full length MuA is targeted to each site in the transpososome, while the remaining subunits in the complex are filled with MuA Δ8, a mutant MuA that does not promote unfolding by ClpX (Figure 1A). The rate of disassembly for each of the complexes described above is compared to the rate of disassembly of complexes containing only wildtype full-length MuA, complexes containing only MuA Δ8 and complexes with full-length R146V targeted to the site of investigation, with full-length wildtype MuA subunits filling the remaining Mu DNA sites.

The second experiment is the reverse of the first, whereby MuA Δ8 is targeted to each site in the complex, while full length MuA fills in the remaining sites in the complex (Figure 1A).

Again, complex disassembly is compared to the disassembly rates of appropriate control complexes. The former experiment determines whether a full-length MuA at the site under study is sufficient to mediate unfolding, whereas the latter determines whether the corresponding site is necessary to complete the reaction.

A.



B.

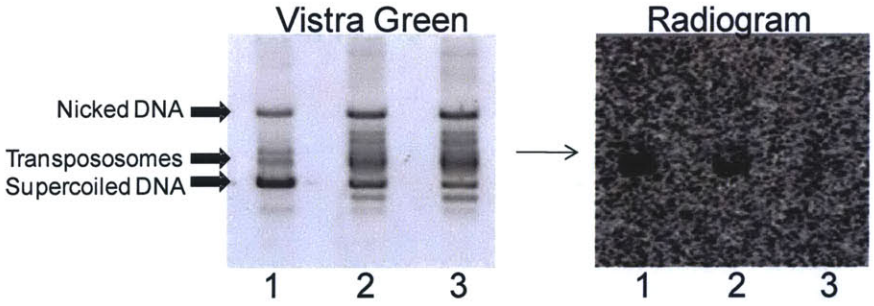


Figure 1. A. Experimental strategy to determine geometry of unfolding in the transpososome. Experiments are performed using ^{35}S-labeled R146V so that only complexes that have incorporated this variant are quantified. B. Transpososomes containing ^{35}S-labeled R146V MuA and unlabeled MuA subunits can be visualized using the DNA binding reagent Vistra Green, and also by radiogram. Lanes 1 contains transpososomes assembled with 50nM ^{35}S-labeled WT MuA only; lane 2 contains transpososomes assembled with 50nM unlabeled WT MuA and 70nM ^{35}S-labeled MuA R146V; lane 3 contains transpososomes assembled with 50nM unlabeled WT MuA, 70nM ^{35}S-labeled MuA R146V and 430nM unlabeled MuA R146V.

To determine which subunits in the transpososome mediate the auto-tethering "adaptor" function, experiments conceptually similar to the ones described above will be performed using full-length and "tethering" mutants instead. The tethering mutants contain substitutions in the "extended" contacts the transpososome uses to mediate high affinity binding to ClpX, specifically in the context of the tetramer (see Chapter 2). As the most severe extended contact mutant is defective in disassembly by 10-fold, this may not be a large enough difference to obtain clear-cut results using this system. The tethering mutant in these experiments will contain double rather than single substitutions that will function to produce larger difference in rates of disassembly. In this manner, we can specifically isolate the unfolding and auto-tethering functions of different subunits within the transpososome.

Disassembly of the mixed transpososome containing mutants with altered specificity from the same reaction mix are monitored using two assays: one monitoring rates of protein unfolding, and the other monitoring rates of formation of DNA disassembly products. To assay for protein unfolding, MuA variants are body-labeled using ^{35}S-methionine and cysteine, and the disappearance of a transpososome band on a radiogram would be indicative of disassembly by ClpX. By labeling all MuA R146V variants and filling the remaining DNA binding sites in the transpososome using unlabeled wildtype MuA subunits, we can specifically monitor only complexes which have incorporated MuA R146V (Figure 1B). DNA disassembly product formation is monitored using the sensitive DNA binding dye Vistra Green. Comparing the rate of DNA disassembly product formation to the rate of protein unfolding by radiogram for each time point allows us to determine whether unfolding of a subunit at a specific site results in

disassembly. If this is in fact the case, we should observe a concomitant rate of DNA disassembly product formation similar to the rate of ^{35}S protein unfolding.

RESULTS

MuA R146V was purified and the monomer form was tested for degradation by ClpXP. MuA R146V showed no defect in the rate of degradation compared to wildtype MuA, indicating that there is no severe defect in the recognition of the monomer form of MuA R146V by the protease (Figure 2). We then targeted variants of MuA R146V to different Mu DNA binding sites in the transpososomes and tested for disassembly, as described below.

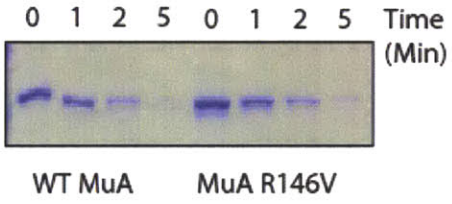


Figure 2. Degradation of MuA R146V monomer by ClpXP. MuA R146V is degraded at the same rate at WT MuA. MuA concentrations: 1μM. $ClpX_6$ concentration: 0.3μM. $ClpP_{14}$ concentration: 0.8μM.

MuA tag at site L2 is recognized by ClpX but does not result in disassembly

Disassembly reactions containing full-length ^{35}S-labeled MuA R146V targeted to the L2 site and unlabeled MuA Δ8 subunits filling the remaining Mu DNA sites were performed. When monitored for protein unfolding by radiogram, we observed that the rate of disassembly of these complexes were intermediate between the rate of disassembly of all full-length wildtype and the rate of disassembly of all MuA Δ8, suggesting that ClpX was able to unfold some subunits at the L2 site (Figure 3). However, observing the rate of disassembly by rate of formation of DNA disassembly products showed that the unfolding of this subunit at the L2 resulted in no appreciable disassembly (Figure 3). This observation suggests a model whereby the subunit at the L2 site is accessible for unfolding, but removal of this subunit does not result in productive disassembly. In other experiments, we observed that the ^{35}S-labeled MuA R146V targeted to the L2 site became dislodged when run on an agarose gel containing 1 M urea, suggesting that this subunit is not tightly bound to the transpososome complex (data not shown). These observations are consistent with previous studies showing that this subunit does not make intimate contacts with the L2 DNA site and is not required for the stability or formation of the transpososome (Kuo et al., 1991; Lavoie et al., 1991; Mizuuchi et al., 1991)

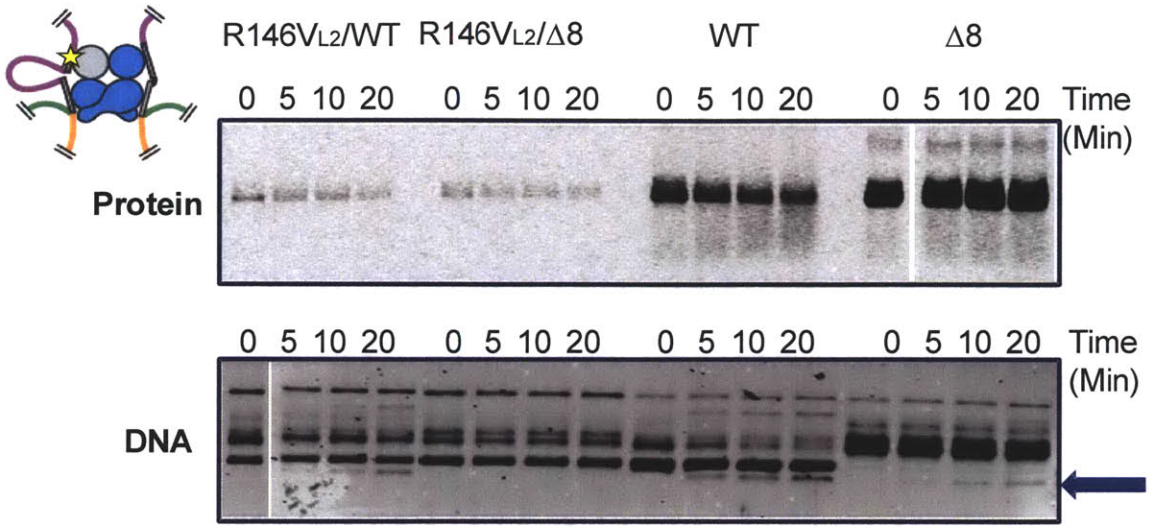


Figure 3. Disassembly of transpososomes with full-length 70 nM ^{35}S-labeled R146V targeted to the L2 mutant site and 50 nM unlabeled WT subunits at the remaining MuA DNA sites (R146VL2/WT), 70 nM full-length ^{35}S-labeled R146V targeted to the L2 mutant site and 50 nM unlabeled MuA Δ8 subunits at the remaining MuA DNA sites (R146VL2/Δ8), 50 nM full-length ^{35}S-labeled WT MuA on wildtype mini-Mu plasmid (WT), and 50 nM full-length ^{35}S-labeled MuA Δ8 on wildtype mini-Mu plasmid (Δ8). Panel labeled "Protein" visualizes the reaction by radiogram, and panel labeled "DNA" visualizes the reaction by Vistra Green. Blue arrow in DNA panel points to DNA disassembly products, whose rate of appearance is quantified and compared to the rate of disappearance of labeled protein on the radiogram.

MuA tag at site R2 does not support disassembly by ClpX

Disassembly reactions containing full-length ^{35}S-labeled MuA R146V targeted to the R2 site and unlabeled MuA Δ8 subunits filling the remaining Mu DNA sites were performed. Assays monitored by radiogram or by Vistra Green showed that targeting a full-length subunit to R2 does not allow subunit unfolding or for disassembly of complexes, suggesting that ClpX does not recognize subunits at the R2 site or is not able to mediate unfolding at that site (Figure 4).

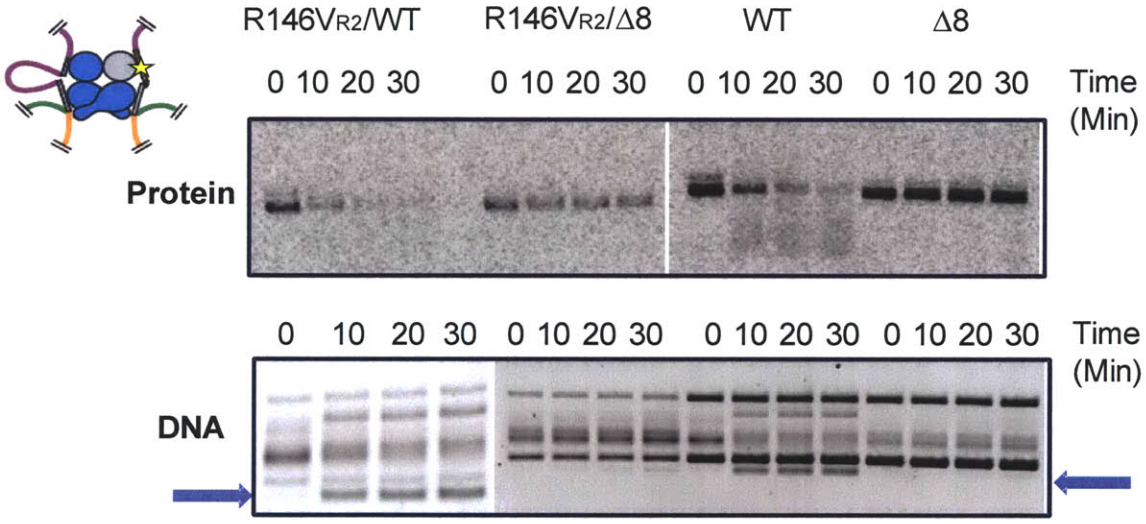


Figure 4. Disassembly of transpososomes with full-length 70 nM ^{35}S-labeled R146V targeted to the R2 mutant site and 50 nM unlabeled WT subunits at the remaining MuA DNA sites (R146VR2/WT), 70 nM full-length ^{35}S-labeled R146V targeted to the L2 mutant site and 50 nM unlabeled MuA Δ8 subunits at the remaining MuA DNA sites (R146VR2/Δ8), 50 nM full-length ^{35}S-labeled WT MuA on wildtype mini-Mu plasmid (WT), and 50 nM full-length ^{35}S-labeled MuA Δ8 on wildtype mini-Mu plasmid (Δ8). Panel labeled "Protein" visualizes the reaction by radiogram, and panel labeled "DNA" visualizes the reaction by Vistra Green. Blue arrows in DNA panel points to DNA disassembly products, whose rate of appearance is quantified and compared to the rate of disappearance of labeled protein on the radiogram.

MuA R146V-specific site at L1 does not support assembly of transpososomes

Targeting MuA R146V o the L1 site resulted in severe defects in assembly of transpososomes, suggesting that mutations required for the altered specificity complexes are not tolerated at this site (Figure 5). To overcome this technical problem, we engineered plasmids containing altered specificity Mu DNA sites at L2, R1 and R2, leaving the wild-type sequence at L1 (Figure 6A). In this way, we can target full-length, MuA Δ8 or tethering mutants to the L1 site while filling in the remaining sites with the MuA R146V variant of choice. These plasmids supported assembly of some transpososomes *in vitro* (Figure 6B). Experiments are under way to optimize assembly of transpososomes on this plasmid.

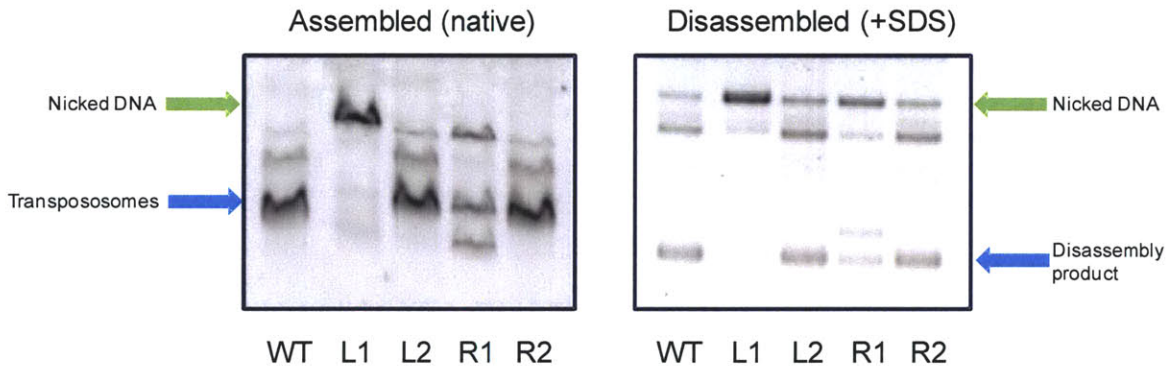


Figure 5. Plasmids containing R146V-specific targeting mutations assembled with different efficiencies; plasmid with mutation in the L1 site fails to assemble transpososomes and produces mostly nicked DNA. Right gel image shows native assembled transpososomes; left gel images shows corresponding transpososomes that are disassembled by addition of SDS.

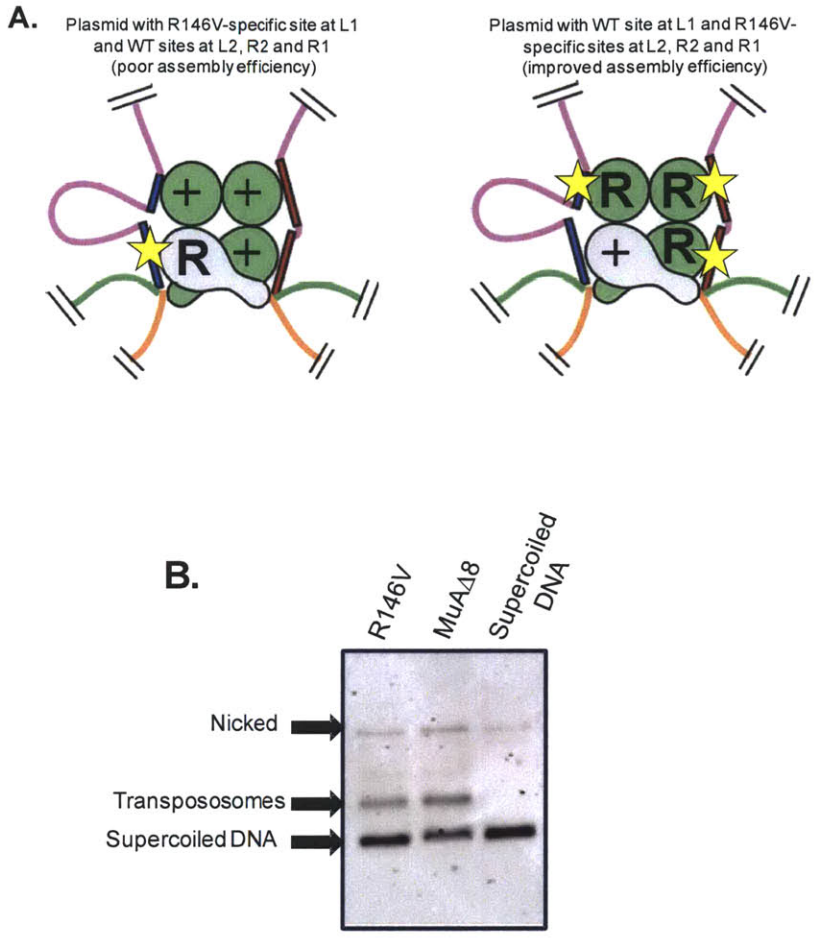


Figure 6. Strategy for targeting subunits to the L1 site. A. As mutating L1 site in mini-Mu to make it specific to MuA R146V does not support assembly of transpososomes, an alternative strategy is to mutate sites R1, R2 and L2 to R146V-specific sites, thus targeting wildtype MuA binding subunits to L1 and R146V MuA to the remaining Mu DNA sites. "R" denotes MuA R146V; "+" denotes R146+. B. This strategy results in a plasmid that can form transpososomes at much higher efficiency. Concentration of MuA used to assemble transpososomes in this panel is 200 nM for both WT and R146V variants.

MuA tag at R1 site is probably not recognized by ClpX

Targeting of MuA R146V to this site has so far only been successful by monitoring using the rate of formation of DNA disassembly products. Monitoring the disassembly reaction in this way showed that targeting the full length MuA R146V to the R1 site, while filling in the remaining Mu sites with MuA Δ8, does not allow for disassembly (data not shown). Experiments are underway to optimize the monitoring of protein unfolding using ^{35}S-labeled MuA R146V targeted to the R1 site.

DISCUSSION

Many ClpX substrates form multimers, and how ClpX interacts with subunits for unfolding and auto-tethering is still poorly understood. The transpososome, a ClpX substrate that has been studied for many years, is the ideal multimeric substrate to parse apart the molecular interactions required to mediate remodeling by ClpX. Although previous experiments have shown that ClpX likely selects a subunit from the left side of the transpososome for unfolding, these experiments could not resolve which of the two subunits on the left side resulted in disassembly, and whether any other subunits in the transpososome played an important role in binding to ClpX (Burton and Baker, 2003). The altered specificity experiments described above show promising potential to answer at least some of these remaining questions about how ClpX is able to select specific subunits for unfolding in the context of a multimer.

Although this set of experiments is far from complete, we can make some conclusions regarding the disassembly reaction mediated by ClpX. It is clear from these experiments that unfolding of the subunit at L2 by ClpX, although possible, does not result in a productive

disassembly reaction. This makes sense in light of our observation that L2 dissociates from the transpososome complex under conditions of mild urea challenge, but this dissociation does not result in the disassembly of the remaining complex. Additionally, previous studies have shown that L2 does not make intimate contacts with the L2 Mu DNA binding site (Kuo et al., 1991; Lavoie et al., 1991; Mizuuchi et al., 1991). This does not exclude the possibility that upon formation of the transpososome, the L2 site is held to the transpososome via protein-protein contacts rather than protein-DNA contacts. As it has been shown that the subunit that is unfolded likely originates from the left side of the transpososome, this leaves the subunit at L1 a promising candidate for unfolding for purposes of disassembly by ClpX. Experiments are currently underway to optimize the results at the L1 site for a more concrete result.

In addition to optimizing the experiments investigating which subunits are sufficient and which are necessary for unfolding, we are also engineering a suitable auto-tethering mutant to target to specific sites within the transpososome. We believe that, with some optimization, this system will be very useful in allowing us to determine the different roles that subunits in the transpososome play in the remodeling reaction.

EXPERIMENTAL PROCEDURES

DNA for transposition and cloning

For MuA R146V proteins, point mutations were introduced using the Quikchange kit (Strategene). For MuA Δ8, the Quikchange kit was used to introduce a stop codon after residues 655. For mini-Mu constructs containing R146V-specific Mu binding sites, the Quikchange kit was used to introduce the necessary substitutions. For DNA sequence of each altered Mu DNA binding site in phage Mu, refer to (Namgoong and Harshey).

Protein purification

Unlabeled wild type and mutant variants of MuA (Baker et al., 1991; Wu and Chaconas, 1995), labeled variants of MuA (Levchenko et al., 1997), and ClpX (Neher et al., 2003) were purified as previously described.

Transpososome assembly

Transpososomes were assembled *in vitro* in the following solution: 25 mM Hepes (pH 7.6), 1 mM $MgCl_2$, 140 mM NaCl, 1 mM DTT, 15% glycerol, 20 µg/ml BSA and 12% DMSO. Transposition reactions contained 30 µg/ml mini-Mu or mini-Mu altered specificity variant (plasmids are pMK586) and 130 nM HU protein. For reactions containing mixtures of R146V MuA and non-R146V MuA, R146V MuA was preincubated with mini-Mu DNA for 5 minutes at 30°C, to overcome the lower binding affinity of R146V MuA for its binding sites (Namgoong and Harshey, 1998; Namgoong et al., 1998). Transposition reactions were carried out at 30 °C for 90 min. Transpososomes were purified prior to disassembly using a phosphocellulose resin mini-spin column.

Degradation and disassembly assays

ClpX was preincubated with ATP regeneration mix (ATP, creatine phosphate and creatine kinase) for 90 s at 30 °C prior to addition of substrate. Disassembly reactions were stopped by addition of 100 mM EDTA, and DNA products were separated and visualized as described (Burton and Baker, 2003).

REFERENCES

Abdelhakim, A.H., Oakes, E.C., Sauer, R.T., and Baker, T.A. (2008). Unique contacts direct high-priority recognition of the tetrameric Mu transposase-DNA complex by the AAA+ unfoldase ClpX. Mol Cell *30*, 39-50.

Baker, T.A., Mizuuchi, M., and Mizuuchi, K. (1991). MuB protein allosterically activates strand transfer by the transposase of phage Mu. Cell *65*, 1003-1013.

Burton, B.M., and Baker, T.A. (2003). Mu transpososome architecture ensures that unfolding by ClpX or proteolysis by ClpXP remodels but does not destroy the complex. Chem Biol *10*, 463-472.

Jones, J.M., and Nakai, H. (1997). The phiX174-type primosome promotes replisome assembly at the site of recombination in bacteriophage Mu transposition. EMBO J *16*, 6886-6895.

Jones, J.M., Welty, D.J., and Nakai, H. (1998). Versatile action of Escherichia coli ClpXP as protease or molecular chaperone for bacteriophage Mu transposition. J Biol Chem *273*, 459-465.

Kruklitis, R., Welty, D.J., and Nakai, H. (1996). ClpX protein of Escherichia coli activates bacteriophage Mu transposase in the strand transfer complex for initiation of Mu DNA synthesis. Embo J *15*, 935-944.

Kuo, C.F., Zou, A.H., Jayaram, M., Getzoff, E., and Harshey, R. (1991). DNA-protein complexes during attachment-site synapsis in Mu DNA transposition. Embo J *10*, 1585-1591.

Lavoie, B.D., Chan, B.S., Allison, R.G., and Chaconas, G. (1991). Structural aspects of a higher order nucleoprotein complex: induction of an altered DNA structure at the Mu-host junction of the Mu type 1 transpososome. Embo J *10*, 3051-3059.

Levchenko, I., Luo, L., and Baker, T.A. (1995). Disassembly of the Mu transposase tetramer by the ClpX chaperone. Genes Dev *9*, 2399-2408.

Levchenko, I., Yamauchi, M., and Baker, T.A. (1997). ClpX and MuB interact with overlapping regions of Mu transposase: implications for control of the transposition pathway. Genes Dev *11*, 1561-1572.

Mariconda, S., Namgoong, S.Y., Yoon, K.H., Jiang, H., and Harshey, R.M. (2000). Domain III function of Mu transposase analysed by directed placement of subunits within the transpososome. J Biosci *25*, 347-360.

Mizuuchi, M., Baker, T.A., and Mizuuchi, K. (1991). DNase protection analysis of the stable synaptic complexes involved in Mu transposition. Proc Natl Acad Sci U S A *88*, 9031-9035.

Namgoong, S.Y., and Harshey, R.M. (1998). The same two monomers within a MuA tetramer provide the DDE domains for the strand cleavage and strand transfer steps of transposition. EMBO J *17*, 3775-3785.

Namgoong, S.Y., Sankaralingam, S., and Harshey, R.M. (1998). Altering the DNA-binding specificity of Mu transposase in vitro. Nucleic Acids Res *26*, 3521-3527.

Neher, S.B., Sauer, R.T., and Baker, T.A. (2003). Distinct peptide signals in the UmuD and UmuD' subunits of UmuD/D' mediate tethering and substrate processing by the ClpXP protease. Proc Natl Acad Sci U S A *100*, 13219-13224.

Wu, Z., and Chaconas, G. (1995). A novel DNA binding and nuclease activity in domain III of Mu transposase: evidence for a catalytic region involved in donor cleavage. Embo J *14*, 3835-3843.

APPENDIX II

THE MICRORNAS OF CAENORHABDITIS ELEGANS

Lee P. Lim*, Nelson C. Lau*, Earl G. Weinstein*, Aliaa Abdelhakim*, Soraya Yekta, Matthew W. Rhoades, Chris B. Burge, and David P. Bartel

Published in Genes and Development, April 15 2003.
** With equal contribution.*

This work was done during my time in Professor David Bartel's lab (May 2002-May 2003).

The microRNAs of *Caenorhabditis elegans*

Lee P. Lim,[1,2,3,4] Nelson C. Lau,[1,2,3] Earl G. Weinstein,[1,2,3] Aliaa Abdelhakim,[1,2,3] Soraya Yekta,[1,2] Matthew W. Rhoades,[1,2], Christopher B. Burge,[1,5] and David P. Bartel[1,2,6]

[1]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA, and [2]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA

MicroRNAs (miRNAs) are an abundant class of tiny RNAs thought to regulate the expression of protein-coding genes in plants and animals. In the present study, we describe a computational procedure to identify miRNA genes conserved in more than one genome. Applying this program, known as MiRscan, together with molecular identification and validation methods, we have identified most of the miRNA genes in the nematode *Caenorhabditis elegans*. The total number of validated miRNA genes stands at 88, with no more than 35 genes remaining to be detected or validated. These 88 miRNA genes represent 48 gene families; 46 of these families (comprising 86 of the 88 genes) are conserved in *Caenorhabditis briggsae*, and 22 families are conserved in humans. More than a third of the worm miRNAs, including newly identified members of the *lin-4* and *let-7* gene families, are differentially expressed during larval development, suggesting a role for these miRNAs in mediating larval developmental transitions. Most are present at very high steady-state levels—more than 1000 molecules per cell, with some exceeding 50,000 molecules per cell. Our census of the worm miRNAs and their expression patterns helps define this class of noncoding RNAs, lays the groundwork for functional studies, and provides the tools for more comprehensive analyses of miRNA genes in other species.

[*Keywords*: miRNA; noncoding RNA; computational gene identification; Dicer]

Supplemental material is available at http://www.genesdev.org.

Received January 13, 2003; accepted in revised form February 25, 2003.

Noncoding RNAs (ncRNAs) of ~22 nucleotides (nt) in length are increasingly recognized as playing important roles in regulating gene expression in animals, plants, and fungi. The first such tiny regulatory RNA to be identified was the *lin-4* RNA, which controls the timing of *Caenorhabditis elegans* larval development (Lee et al. 1993; Wightman et al. 1993). This 21-nt RNA pairs to sites within the 3′ untranslated region (UTR) of target mRNAs, specifying the translational repression of these mRNAs and triggering the transition to the next developmental stage (Lee et al. 1993; Wightman et al. 1993; Ha et al. 1996; Moss et al. 1997; Olsen and Ambros 1999). A second tiny riboregulator, *let-7* RNA, is expressed later in development and appears to act in a similar manner to trigger the transition to late-larval and adult stages (Reinhart et al. 2000; Slack et al. 2000). The *lin-4* and *let-7* RNAs are sometimes called small temporal RNAs (stRNAs) because of their important roles in regulating the timing of larval development (Pasquinelli et al. 2000). The *lin-4* and *let-7* stRNAs are now recognized as the founding members of a large class of ~22-nt ncRNAs termed microRNAs (miRNAs), which resemble stRNAs but do not necessarily control developmental timing (Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros 2001).

Understanding the biogenesis and function of miRNAs has been greatly facilitated by analogy and contrast to another class of tiny ncRNAs known as small interfering RNAs (siRNAs), first identified because of their roles in mediating RNA interference (RNAi) in animals and posttranscriptional gene silencing in plants (Hamilton and Baulcombe 1999; Hammond et al. 2000; Parrish et al. 2000; Zamore et al. 2000; Elbashir et al. 2001a; Klahre et al. 2002). During RNAi, long double-stranded RNA (either a bimolecular duplex or an extended hairpin) is processed by Dicer, an RNAse III enzyme, into many siRNAs that serve as guide RNAs to specify the destruction of the corresponding mRNA (Hammond et al. 2000; Zamore et al. 2000; Bernstein et al. 2001; Elbashir et al. 2001a). Although these siRNAs are initially short double-stranded species with 5′ phosphates and 2-nt 3′ overhangs characteristic of RNAse III cleavage products, they eventually become incorporated as single-stranded RNAs into a ribonucleoprotein com-

[3]**These authors contributed equally to this work.**
[4]**Present address: Rosetta Inpharmatics, 12040 115th Avenue NE, Kirkland, Washington 98034.**
Corresponding authors.
[5]**E-MAIL cburge@mit.edu; FAX (617) 452-2936.**
[6]**E-MAIL dbartel@wi.mit.edu; FAX (617) 258-6768.**
Article published online ahead of print. Article and publication date are at http://www.genesdev.org/cgi/doi/10.1101/gad.1074403.

plex, known as the RNA-induced silencing complex (RISC; Hammond et al. 2000; Elbashir et al. 2001a,b; Nykäken et al. 2001; Martinez et al. 2002; Schwarz et al. 2002). The RISC identifies target messages based on perfect (or nearly perfect) antisense complementarity between the siRNA and the mRNA, and then the endonuclease of the RISC cleaves the mRNA at a site near the middle of the siRNA complementarity (Elbashir et al. 2001a,b). Similar pathways have been proposed for gene silencing in plants and fungi, with siRNAs targeting mRNA for cleavage during posttranscriptional gene silencing and heterochromatic siRNAs targeting chromatin for histone methylation, triggering heterochromatin formation and consequent transcriptional gene silencing (Hamilton and Baulcombe 1999; Vance and Vaucheret 2001; Hall et al. 2002; Hamilton et al. 2002; Pickford et al. 2002; Reinhart and Bartel 2002; Volpe et al. 2002; Zilberman et al. 2003).

miRNAs have many chemical and functional similarities to the siRNAs. Like siRNAs they are processed by Dicer, and so they are the same length and possess the same 5′-phosphate and 3′-hydroxyl termini as siRNAs (Grishok et al. 2001; Hutvágner et al. 2001; Ketting et al. 2001; Lau et al. 2001; Park et al. 2002; Reinhart et al. 2002). They are also incorporated within a ribonucleoprotein complex, known as the miRNP, which is similar if not identical to the RISC (Caudy et al. 2002; Hutvágner and Zamore 2002; Ishizuka et al. 2002; Martinez et al. 2002; Mourelatos et al. 2002). In fact, many plant miRNAs match their predicted mRNA targets with near-perfect antisense complementarity, as if they were functioning as siRNAs within a RISC complex (Rhoades et al. 2002), and the plant miR171 and miR165/166 have been shown to specify cleavage of their mRNA targets (Llave et al. 2002b; Tang et al. 2003). The *C. elegans* and *Drosophila* miRNAs do not have as pronounced a tendency to pair with their targets with near-perfect complementarity (Rhoades et al. 2002). Nonetheless, some might still direct cleavage of their targets, as suggested by the observation that miRNAs and siRNAs with 3–4 mismatches with their targets can still direct cleavage in plant and animal lysates (Tang et al. 2003). Furthermore, the *let-7* miRNA is present within a complex that can cleave an artificial RNA target when such a target is perfectly complementary to the miRNA (Hutvágner and Zamore 2002). The known biological targets of *lin-4* and *let-7* RNAs have several mismatches within the central region of the miRNA complementary sites, perhaps explaining why in these particular cases, the miRNAs specify translational repression rather than mRNA cleavage during *C. elegans* larval development (Lee et al. 1993; Wightman et al. 1993; Ha et al. 1996; Moss et al. 1997; Olsen and Ambros 1999; Reinhart et al. 2000; Slack et al. 2000; Hutvágner and Zamore 2002).

Regulatory targets for most animal miRNAs have not yet been identified. Prediction of plant miRNA targets has led to the proposal that many plant miRNAs function to clear from differentiating cells mRNAs encoding key transcription factors, thereby facilitating plant development and organogenesis (Rhoades et al. 2002). Confident computational prediction of animal miRNA targets has relied on experimental evidence to first narrow the number of candidate mRNAs (Lai 2002). Nonetheless, as seen for the plant miRNAs, the sequences of the animal miRNAs are generally highly conserved in evolution. For example, 91 of the 107 miRNAs cloned from mammals are detected in the pufferfish (*Fugu rubripes*) genome, implying that they have important functions preserved during vertebrate evolution (Lim et al. 2003).

The first step in a systematic approach to identifying the biological roles of miRNAs is to find the miRNA genes themselves. Because gene-prediction programs had not been developed to identify miRNAs in genomic sequence, miRNA gene identification has been primarily achieved by cloning the small RNAs from size-fractionated RNA samples, sometimes specifically enriching in miRNAs by first immunoprecipitating the miRNP complex or by using a cloning protocol specific for the 5′ phosphate and 3′ hydroxyl found on Dicer products (Lagos-Quintana et al. 2001, 2002, 2003; Lau et al. 2001; Lee and Ambros 2001; Llave et al. 2002a; Mourelatos et al. 2002; Park et al. 2002; Reinhart et al. 2002). Once small RNAs have been cloned, the challenge is to differentiate the authentic miRNAs from other RNAs present in the cell, particularly from endogenous siRNAs. Because both miRNAs and siRNAs are Dicer products and both can act to specify mRNA cleavage, miRNAs cannot be differentiated based on their chemical composition or their functional properties. However, miRNAs can be distinguished from siRNAs based on their biogenesis and evolutionary conservation: (1) They are 20- to 24-nt RNAs that derive from endogenous transcripts that can form local RNA hairpin structures; (2) these hairpins are processed such that a single miRNA molecule ultimately accumulates from one arm of each hairpin precursor molecule; (3) the sequences of the mature miRNAs and their hairpin precursors are usually evolutionarily conserved; and (4) the miRNA genomic loci are distinct from and usually distant from those of other types of recognized genes, although a few are found within predicted introns but not necessarily in the same orientation as the introns. Endogenous siRNAs differ in that (1) they derive from extended dsRNA, (2) each dsRNA precursor gives rise to numerous different siRNAs, (3) they generally display less sequence conservation, and (4) they often perfectly correspond to the sequences of known or predicted mRNAs, transposons, or regions of heterochromatic DNA (Aravin et al. 2001; Djikeng et al. 2001; Elbashir et al. 2001a; Lau et al. 2001; Llave et al. 2002a; Mochizuki et al. 2002; Reinhart and Bartel 2002; Reinhart et al. 2002). Regarding this fourth criterion, miRNAs can also perfectly correspond to sequences of their mRNA targets, but when they do, they still derive from loci distinct from those of their mRNA targets (Llave et al. 2002a,b; Reinhart et al. 2002). Because miRNAs are primarily distinguished based on their biogenesis and evolutionary conservation, the current norms for identification and validation of miRNA genes include experimental evidence for endogenous expression of the miRNA, coupled with evidence of a hairpin precursor, preferably

one that is evolutionarily conserved (Ambros et al. 2003).

Some miRNAs might be difficult to isolate by cloning, due to their low abundance or to biases in cloning procedures. Thus, computational identification of miRNAs from genomic sequences would provide a valuable complement to cloning. Recent advances have been made in the computational identification of ncRNA genes through comparative genomics, and complex algorithms have been developed to identify ncRNAs in general (Argaman et al. 2001; Rivas et al. 2001; Wassarman et al. 2001), as well as specific ncRNA families such as tRNAs and snoRNAs (Lowe and Eddy 1997, 1999).

In the present study, we describe a computational procedure to identify miRNA genes. By using this procedure, together with extensive sequencing of clones (3423 miRNA clones were sequenced), we have detected 30 additional miRNA genes, including previously unrecognized *lin-4* and *let-7* homologs. Extrapolation of the computational analysis indicates that miRNA gene identification in *C. elegans* is now approaching saturation, and that no more than 120 miRNA genes are present in this species. We also identify those genes with intriguing expression patterns during larval development and conditions of nutrient stress, and we show that most miRNAs are expressed at very high levels, with some present in as many copies per cell as the highly abundant U6 snRNA. This extensive census of worm miRNAs and their expression patterns establishes the general properties of this gene class and provides resources and tools for studies of miRNA function in nematodes and other organisms.

Results

Computational prediction of C. elegans *miRNA genes*

We developed a computational tool to specifically identify miRNAs that are conserved in two genomes and have the features characteristic of known miRNAs. To identify miRNAs in nematodes, the *C. elegans* genome was first scanned for hairpin structures with sequences that were conserved in *Caenorhabditis briggsae*. About 36,000 hairpins were found that satisfied minimum requirements for hairpin structure and sequence conservation. This procedure cast a sufficiently wide net to capture 50 of the 53 miRNAs previously reported to be conserved in the two species (Lau et al. 2001; Lee and Ambros 2001). These 50 published miRNA genes served as a training set for the development of a program called MiRscan, which was then used to assign scores to each of the 36,000 hairpins, evaluating them based on their similarity to the training set with respect to the following features: base pairing of the miRNA portion of the fold-back, base pairing of the rest of the fold-back, stringent sequence conservation in the 5′ half of the miRNA, slightly less stringent sequence conservation in the 3′ half of the miRNA, sequence biases in the first five bases of the miRNA (especially a U at the first position), a tendency toward having symmetric rather than asymmetric internal loops and bulges in the miRNA region, and the presence of two to nine consensus base pairs between the miRNA and the terminal loop region, with a preference for 4–6 bp (Fig. 1A).

The distribution of MiRscan scores for the ~36,000 hairpins illustrated the ability of MiRscan to discern the 50 miRNA genes of the training set, which fell mostly in the high-scoring tail of the distribution (Fig. 2). Of the features evaluated by MiRscan, base-pairing potential and sequence conservation played primary roles in distinguishing known miRNAs (Fig. 1B). Some of the other conserved hairpins also scored highly; 35 had scores exceeding 13.9, the median score of the 58 known miRNAs (Fig. 2B). These 35 hairpins were carried forward as the top miRNA candidates predicted by MiRscan.

Molecular identification of miRNA genes

Our initial cloning and sequencing of small RNAs from mixed-stage *C. elegans* had identified 300 clones that represented 54 unique miRNA sequences (Lau et al. 2001). For the present study, this approach for identifying miRNAs was scaled-up ~10-fold. In an effort to identify miRNAs not normally expressed in mixed-stage logarithmically growing hermaphrodite worms, RNA was also cloned from populations of *him-8* worms, starved L1, and dauer worms. The *him-8* population was ~40% males, whereas the normal (N2) population was nearly all hermaphrodites (Broverman and Meneely 1994). Starved L1 and dauer worms are arrested in development at larval stages L1 and L3, respectively, with dauer worms having undergone morphological changes that enhance survival after desiccation or other harsh conditions.

As before, some clones matched *Escherichia coli*, the food source of the worms, others corresponded to fragments of annotated *C. elegans* RNAs. Nevertheless, 3423 clones were classified as miRNA clones (Table 1). Most of these represented the 58 miRNA genes previously identified in *C. elegans* (Lau et al. 2001; Lee and Ambros 2001). For example, *lin-4* was represented by 125 clones, *let-7* by 17 clones, and *mir-52* by 404 clones (Table 1). The remaining miRNA clones represented 23 newly identified miRNA loci.

In total, 80 loci were represented by cloned miRNAs (Table 1). Of these, 77 had the classical features of *C. elegans* miRNA genes, in that they had the potential to encode stereotypic hairpin precursor molecules with the 20- to 25-nt cloned RNAs properly positioned within an arm of the hairpin so as to be excised during Dicer processing, and their expression was manifested as a detectable Northern signal in the 20- to 25-nt range. Three other loci, *mir-41*, *mir-249*, and *mir-229*, were also included. The *mir-41* and *mir-249* RNAs were not detected on Northern blots but were still classified as miRNAs because these RNAs and their predicted hairpin precursors appear to be conserved in *C. briggsae*.

The *mir-229* locus was also classified as a miRNA gene, even though it appears to derive from an unusual fold-back precursor. Its precursor appears to be larger

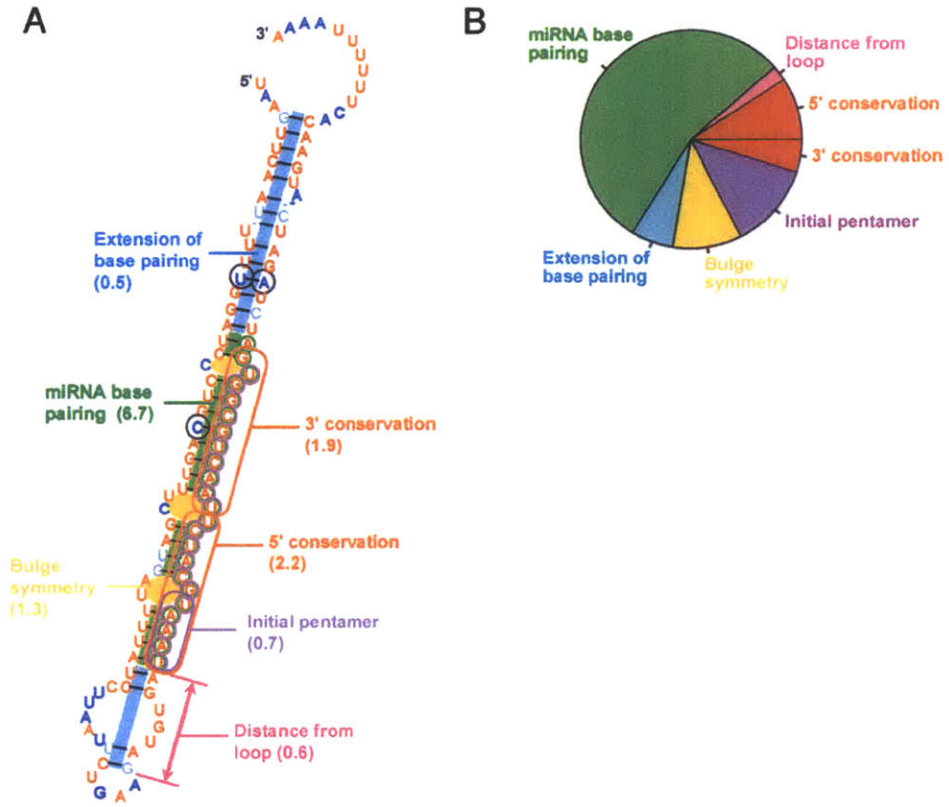


Figure 1. Criteria used by MiRscan to identify miRNA genes among aligned segments of two genomes. (*A*) The seven components of the MiRscan score for *mir-232* of *C. elegans*/*C. briggsae*. These components are annotated in the context of the MiRscan prediction for *mir-232*, with the residues of the predicted miRNA circled in purple and the residues of the validated miRNA (Table 2), circled in green. In parenthesis are the scores for each component, which were added together to give the total score of 13.9. MiRscan predictions are visualized within the consensus *C. elegans*/*C. briggsae* secondary structure, as generated by using ClustalW (Thompson et al. 1994) and Alidot (Hofacker and Stadler 1999). Shown is the *C. elegans* sequence with residues colored to indicate conserved sequence and pairing potential. Residues conserved in *C. briggsae* are red, residues that vary while maintaining their predicted paired or unpaired state are blue (with variant residues that maintain pairing also circled in black), and residues that maintain neither sequence nor pairing are in gray. (*B*) Estimated relative importance of each MiRscan criterion. Estimates were based on the relative entropy between the training set of 50 previously identified nematode miRNAs and the background set of ~36,000 potential stem loops. Because pairing and conservation were used to identify the potential stem loops, the total contributions of these types of criteria for distinguishing miRNA genes from non-protein-coding genomic sequence were underestimated. Likewise, the total contribution of the distance from the loop was underestimated because only those candidates 2–9 bp from the loop were evaluated.

than normal, possibly because of an extra 35-nt stem loop protruding from the 3′ arm of the precursor stem loop (Supplementary Fig. 1). Nonetheless, miR-229 was detectable as a ~25- to 26-nt species on Northern blots, and accumulation of its presumed precursor increased in the *dcr-1* mutant, suggesting that Dicer processes this precursor despite the unusual predicted secondary structure (Supplementary Fig. 1). Furthermore, *mir-229* is only 400 bp upstream of a previously recognized miRNA gene cluster, including *mir-64*, *mir-65*, and *mir-66*. miR-229 also has significant sequence identity with the miRNAs of this cluster. We provisionally classified *mir-229* as a miRNA and a member of this *C. elegans* cluster. Greater confidence would be warranted if its unusual precursor structure were conserved in another species. A weakly homologous cluster of two potential miRNAs was found in *C. briggsae*, but neither of the predicted *C. briggsae* homologs appeared to have an unusual precursor resembling that of miR-229.

Validation of computationally predicted miRNAs

Of the 23 newly cloned miRNAs, 20 received MiRscan scores, and these scores are indicated in yellow in Figure 2B. The other three were not scored because orthologous sequences in *C. briggsae* were not identified. A Mann-Whitney test showed that the distribution of scores for these recently cloned miRNAs was not significantly different from that of the previously cloned miRNAs. Because the recently cloned miRNAs were not known during the development of MiRscan, their high scores gave added assurance that MiRscan was not over-fitting its training set. Ten of the 23 newly cloned miRNAs were

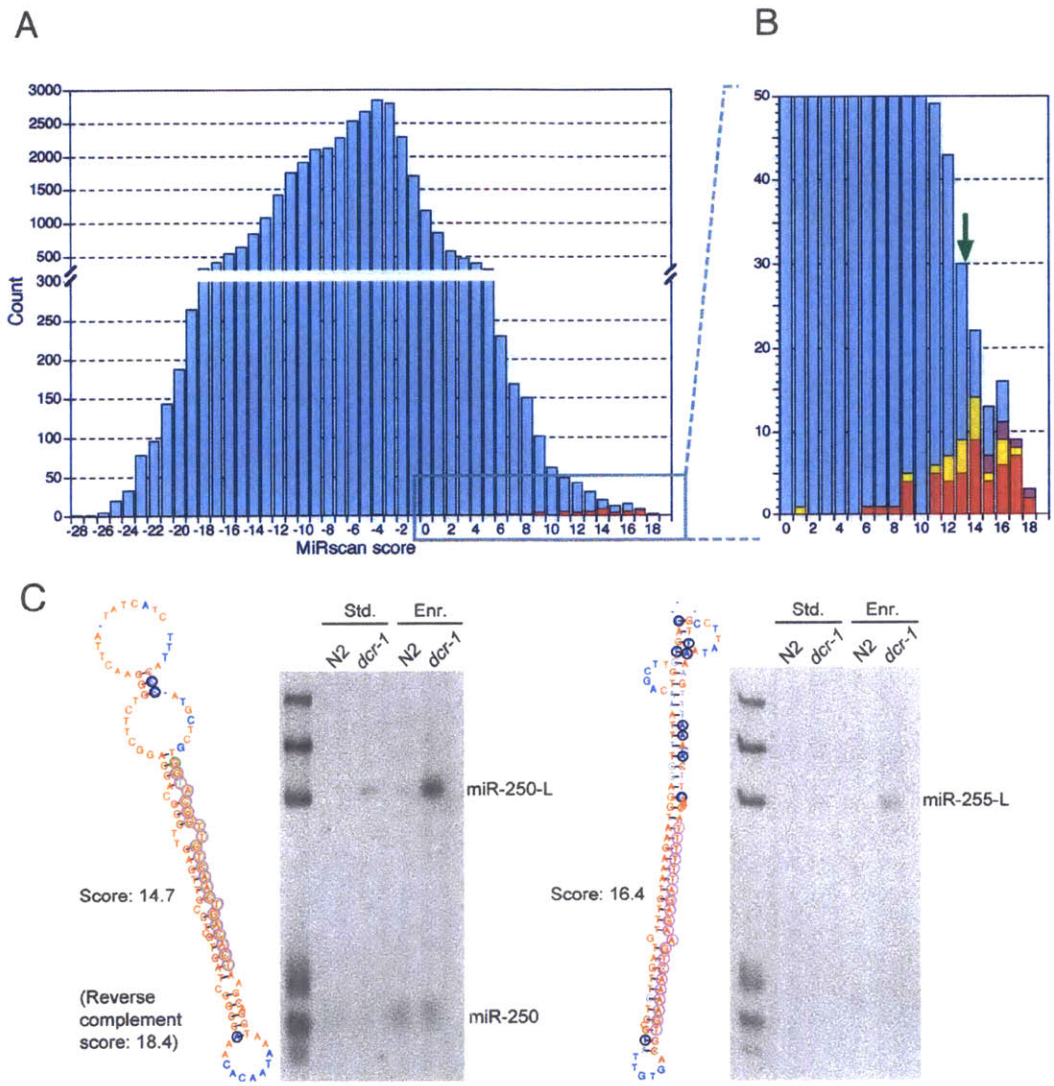


Figure 2. Computational identification of miRNA genes. (*A*) The distribution of MiRscan scores for 35,697 *C. elegans* sequences that potentially form stem loops and have loose conservation in *C. briggsae*. Note that the *Y*-axis is discontinuous so that the scores of the 50 previously reported miRNA genes that served as the training set for MiRscan can be more readily seen (red). Scores for these 50 genes were jackknifed to prevent inflation of their values because of their presence in the training set. (*B*) An expanded view of the high-scoring tail of the distribution. This view captures 49 of the 50 genes of the training set (red). The median score of the 58 previously reported miRNA loci that satisfy the current criteria for designation as miRNA genes (Ambros et al. 2003) is 13.9 (green arrow). Note that this median score was the midpoint between the scores of the 29th and 30th highest-scoring loci of the 50-member training set; namely, it was designated the median score after including the 8 previously reported miRNA genes that were not in the training set because they were lost during the identification of conserved hairpins, usually because they lacked sufficient *C. briggsae* homology. Scores of genes validated by cloning are indicated (yellow), as are scores of six genes that have not yet been cloned but were verified by Northern analysis (purple). (*C*) Examples of miRNA genes identified by MiRscan with the Northern blots that served to validate them. Stem-loops were annotated as in Figure 1A, except the DNA rather than RNA sequence is depicted. The Northern blots show analysis of RNA from either wild-type (N2) or *dcr-1* worms, isolated using either our standard protocol (Std.) or an additional polyethylene glycol precipitation step to enrich for small RNAs (Enr.). Homozygous worms of the *dcr-1* population have reduced Dicer activity, increasing the level of miRNA precursors (e.g., miR-250-L and miR-255-L), which facilitated the validation of miRNA loci, especially those for which the mature miRNA was not detected (e.g., miR-255). RNA markers (*left* lane) are 18, 21, 24, 60, 78, and 119 nt. The miR-250 stem loop shown received a MiRscan score of 14.7. The mir-250 reverse complement received an even greater score of 18.4, but was not detected by Northern analysis. Thus, the predicted *mir-250* gene was assigned the score of the higher-scoring, although incorrect, alternative stem loop (Table 1; Fig. 2B).

among the set of 35 high-scoring miRNA gene candidates and served to validate these 10 candidates.

The remaining 25 candidate miRNAs that had not been cloned were tested by Northern blots. RNA from

Table 1. *Cloning frequency and MiRscan scores of* Caenorhabditis elegans *miRNAs*

miRNA	MiRscan score	Number of sequenced clones: mixed stage	dauer	starved L1	*him*-8	total
let-7 RNA	13.8	15	0	0	2	17
lin-4 RNA	15.8	48	46	4	27	125
miR-1	14.7	43	17	7	9	76
miR-2	6.2	138	46	20	9	213
miR-34	14.1	13	25	5	9	52
miR-35	14.4	23	0	1	2	26
miR-36	14.6	21	0	1	5	27
miR-37	9.6	8	0	1	2	11
miR-38	8.9	10	0	1	0	11
miR-39	9.5	11	0	0	1	12
miR-40	15.4	12	0	4	2	18
miR-41	12.0	2	0	0	0	2
miR-42	9.5	10	4	3	1	18
miR-43	17.5	8	1	9	0	18
miR-44/45	16.6/17.4	22	3	3	4	32
miR-46	11.3	14	11	9	3	37
miR-47	16.5	19	7	4	5	35
miR-48	12.0	52	1	0	8	61
miR-49	13.1	1	0	1	1	3
miR-50	14.6	10	16	5	1	32
miR-51	12.0	16	5	2	2	25
miR-52	11.6	287	70	18	29	404
miR-53	12.4	20	6	1	4	31
miR-54	9.4	49	40	9	13	111
miR-55	13.8	47	32	16	15	110
miR-56	NS	40	16	9	6	71
miR-57	12.1	31	11	8	3	53
miR-58	17.5	181	51	27	31	290
miR-59	18.5	1	0	0	0	1
miR-60	14.1	20	6	3	7	36
miR-61	13.7	8	5	1	3	17
miR-62	15.1	4	4	6	0	14
miR-63	NS	7	1	0	1	9
miR-64	NS	11	4	8	3	26
miR-65	7.4	22	7	3	2	34
miR-66	NS	68	25	6	7	106
miR-67	16.8	3	0	0	0	3
miR-70	11.6	11	8	3	6	28
miR-71	17.9	53	72	23	22	170
miR-72	NS	49	22	10	9	90
miR-73	11.3	13	7	1	1	22
miR-74	17.9	35	12	6	7	60
miR-75	12.6	14	3	2	2	21
miR-76	14.9	1	2	6	3	12
miR-77	14.2	17	3	0	2	22
miR-78	NS	5	1	1	0	7
miR-79	14.2	14	3	3	3	23
miR-80	17.1	121	27	20	17	185
miR-81	18.8	32	24	6	12	74
miR-82	16.3	36	12	6	11	65
miR-83	15.2	12	12	2	8	34
miR-84	−3.3	12	2	1	4	19
miR-85	17.5	10	0	0	12	22
miR-86	16.3	46	57	30	17	150
miR-87	16.7	1	0	0	0	1
miR-88	−7.9					0
miR-90	14.0	5	37	14	9	65
miR-124	15.7	7	16	7	5	35
miR-228	17.5	1	13	8	3	25
miR-229	NS	2	1	0	0	3
miR-230	16.8	0	0	0	1	1
miR-231	14.1	1	2	0	0	3
miR-232	13.8	4	7	2	1	14
miR-233	16.4	1	8	4	0	13
miR-234	14.3	0	0	1	0	1
miR-235	1.9	5	21	1	8	35
miR-236	16.8	3	6	2	1	12
miR-237	11.9	3	0	0	0	3
miR-238	14.0	0	4	1	0	5
miR-239a	12.7	4	0	0	1	5
miR-239b	13.6					0
miR-240	12.5	0	0	0	1	1
miR-241	14.9	7	0	0	3	10
miR-242	9.9	0	0	1	1	2
miR-243	NS	1	0	1	0	2
miR-244	13.4	0	2	5	0	7
miR-245	13.8	0	1	0	0	1
miR-246	12.8	0	0	0	1	1
miR-247	NS	0	2	0	0	2
miR-248	14.6	0	2	0	0	2
miR-249	13.7	0	2	1	0	3
miR-250	18.4					0
miR-251	15.5					0
miR-252	17.7					0
miR-253	16.9					0
miR-254	15.7					0
miR-255	16.4					0
Total clones		1821	851	363	388	3423

A total of 3423 clones from logarithmically growing mixed-stage worms and worms from the indicated stages or mutant (dauer, starved L1, and *him-8*) represented 79 different miRNAs (and 80 different miRNA genes, because the miR-44/45 miRNA appears to be encoded at two loci). Genes not represented in the set of ~36,000 stem loops did not receive scores (NS). Note that the previously reported miR-68 clone is not included. This RNA was not detected on Northern blots, and neither it nor its predicted precursor appears to be conserved in another species. Accordingly, it is now classified as an endogenous siRNA. Two other *C. elegans* loci previously thought to encode miRNAs (*mir-69* and *mir-89*) also do not satisfy the current criteria for classification as miRNA genes (Ambros et al. 2003) and were not considered during the course of this study. One previously reported gene, *mir-88*, was not represented in our set of sequenced clones but is detected on Northern blots as a ~22-nt RNA (V. Ambros, pers. comm.) and thus satisfies the current criteria for classification as an miRNA gene.

dcr-1 worms was included on the blots to enhance detection of precursor hairpins. Dicer-dependent processing of ~70-nt precursors was detected for six candidates (as shown for miR-250 and miR-255; Fig. 2C), and ~22-nt miRNAs were detected for miR-250, miR-251, and miR-252 . Despite prolonged exposure times and enrichment for small RNA by size fractionation, the Northern signals were generally weak, perhaps explaining why

these miRNAs were missed in the current set of 3423 sequenced miRNA clones.

To investigate whether these miRNAs eventually would have been identified after further cloning and sequencing of our cDNA library of small RNA sequences, a PCR assay was used to detect the presence of these miRNAs in the library. By using a primer specific to the 3′ segment of the predicted miRNA, together with a second primer corresponding to the adapter sequence attached to the 5′ terminus of all the small RNAs, the 5′ segment of the miRNA was amplified, cloned, and sequenced. This procedure validated five of the six predicted miRNAs for which at least a precursor could be detected on Northerns, including two of the candidates (miR-253 and miR-254) for which a mature ~22-nt RNA was not detected on Northern blots. In addition, it identified the 5′ terminus of these five miRNAs, which is difficult to achieve with confidence when using only bioinformatics and hybridization.

Combining the cloning and expression data, 16 of the 35 computationally identified candidates were validated (10 from cloning, five from Northerns plus the PCR assay, and one from Northerns only, which validated the precursor but did not identify the mature miRNA). Of the remaining 19 candidates, four could be readily classified as false positives. They appear to be nonannotated larger ncRNA genes, in that probes designed to hybridize to these candidates hybridized instead to high-molecular-weight species that remained constant in the samples from *dcr-1* worms. The remaining 15 new candidates with high MiRscan scores but without any Northern signal might also be false positives, or they might be authentic miRNAs that are expressed at low levels or in only very specific cell types or circumstances. Considering the extreme case in which all the nonvalidated candidates are false positives, the minimum specificity of MiRscan for the *C. elegans*/*C. briggsae* analysis can be calculated as (29 + 16)/(29 + 35), or 0.70, at a sensitivity level that detects half of the 58 previously known miRNAs. A summary of the miRNA genes newly identified by validating computational candidates (16 genes) or by cloning alone (13 genes) is shown in Table 2, and predicted stem-loop precursors are shown in Supplemental Material. Table 2 also includes one additional gene, *mir-239b*, which was identified based on its homology with *mir-239a* and its MiRscan score of 13.6.

Evolutionary conservation of miRNAs

The 88 *C. elegans* miRNA genes identified to this point were grouped into 48 families, each comprising one to eight genes (data not shown). Within families, sequence identity either spanned the length of the miRNAs or was predominantly at their 5′ terminus. All but two of these families extended to the miRNAs of *C. briggsae*. The two families without recognizable *C. briggsae* orthologs each comprised a single miRNA (miR-78 and miR-243). Thus, nearly all (>97%) of the *C. elegans* miRNAs identified had apparent homologs in *C. briggsae*, and all but six of these *C. elegans* miRNAs (miR-72, miR-63, miR-64, miR-66, miR-229, and miR-247) had retained at least 75% sequence identity to a *C. briggsae* ortholog. Of the 48 *C. elegans* miRNA families, 22 also had representatives among the known human miRNA genes (Fig. 3). In that these 22 families included 33 *C. elegans* genes, it appears that at least a third (33/88) of the *C. elegans* miRNA genes have homologs in humans and other vertebrates.

Developmental expression of miRNAs

The expression of 62 miRNAs during larval development was examined and compiled together with previously reported expression profiles (Lau et al. 2001) to yield a comprehensive data set for the 88 *C. elegans* miRNAs (Fig. 4). RNA from wild-type embryos, the four larval stages (L1 through L4), and young adults was probed, as was RNA from *glp-4* (*bn2*) young adults, which are severely depleted in germ cells (Beanan and Strome 1992). Nearly two thirds of the miRNAs appeared to have constitutive expression during larval development (Fig. 4A). These miRNAs might still have differential expression during embryogenesis, or they might have tissue-specific expression, as has been observed for miRNAs of larger organisms in which tissues and organs can be more readily dissected and examined (Lee and Ambros 2001; Lagos-Quintana et al. 2002; Llave et al. 2002a; Park et al. 2002; Reinhart et al. 2002).

Over one third of the miRNAs had expression patterns that changed during larval development (Fig. 4B,C), and there were examples of miRNA expression initiating at each of the four larval stages (Fig. 4B). Expression profiles for miR-48 and miR-241 (which are within 2 kb of each other in the *C. elegans* genome) were similar to those previously reported for *let-7* RNA and miR-84 (Fig. 4B; Reinhart et al. 2000; Lau et al. 2001). In fact, these four miRNAs appear to be paralogs, with all four miRNAs sharing the same first eight residues (Fig. 3). Another newly identified miRNA, miR-237, is a paralog of the other canonical stRNA, *lin-4* RNA (Fig. 3), although miR-237 exhibited an expression pattern distinct from *lin-4* RNA (Fig. 4E). The existence of these paralogs, as well as other families of miRNAs with expression initiating at the different stages of larval development, supports the idea that *lin-4* and *let-7* miRNAs are not the only stRNAs with important roles in the *C. elegans* heterochronic pathway.

Expression usually remained constant once it initiated, as has been seen for *lin-4* and *let-7* miRNA expression (Fig. 4A,B). Exceptions to this trend included the miRNAs of the *mir-35–mir-41* cluster, which were expressed transiently during embryogenesis (Lau et al. 2001); miR-247, which was expressed transiently in larval stage 3 (and dauer); and miR-248, which was most highly expressed in dauer (Fig. 4C,D). miR-234 was expressed in all stages, but expression was highest in both L1 worms (which had been starved shortly before harvest to synchronize the worm developmental staging) and dauer worms, suggesting that this miRNA might be induced as a consequence of nutrient stress.

Table 2. *Newly identified* Caenorhabditis elegans *miRNA genes*

miRNA gene	ID method	miRNA sequence	miRNA length (nt)	*C. briggsae* homology	Fold-back arm	Chr.	Distance to nearest gene		
mir-124	MS, C, N	UAAGGCACGCGGUGAAUGCCA	21	+++	3′	IV	within intron of	C29E6.2	(s)
mir-228	MS, C, N	AAUGGCACUGCAUGAAUUCACGG	21–24	+++	5′	IV	0.2 kb downstream of	T12E12.5	(as)
mir-229	C, N	AAUGACACUGGUUAUCUUUUCCAUCG	25–27	–	5′	III	0.4 kb upstream of	*mir-64*	(s)
mir-230	MS, C, N	GUAUUAGUUGUGCGACCAGGAGA	23	++	3′	X	0.4 kb downstream of	F13D11.3	(as)
mir-231	MS, C, N	UAAGCUCGUGAUCAACAGGCAGAA	23–24	++	3′	III	10.4 kb upstream of	*lin-39*	(s)
mir-232	C, N	UAAAUGCAUCUUAACUGCGGUGA	23–24	+++	3′	IV	1.1 kb downstream of	F13H10.5	(as)
mir-233	MS, C, N	UUGAGCAAUGCGCAUGUGCGGGA	19–23	+++	3′	X	within intron of	W03G11.4	(s)
mir-234	MS, C, N	UUAUUGCUCGAGAAUACCCUU	21	+++	3′	II	1.5 kb downstream of	Y54G11B.1	(as)
mir-235	C, N	UAUUGCACUCUCCCCGGCCUGA	22	+	3′	I	0.6 kb upstream of	T09B4.7	(s)
mir-236	MS, C, N	UAAUACUGUCAGGUAAUGACGCU	21–25	+++	3′	II	0.3 kb downstream of	C52E12.1	(as)
mir-237	C, N	UCCCUGAGAAUUCUCGAACAGCUU	23–24	+	5′	X	3.4 kb upstream of	F22F1.2	(as)
mir-238	MS, C, N	UUUGUACUCCGAUGCCAUUCAGA	21–23	++	3′	III	2.0 kb upstream of	*mir-80*	(s)
mir-239a	C, N	UUUGUACUACACAUAGGUACUGG	22–23	++	5′	X	6.0 kb upstream of	C34E11.1	(s)
mir-239b	H	UUUGUACUACACAAAAGUACUGG	n.d.	++	5′	X	7.0 kb upstream of	C34E11.1	(s)
mir-240	C, N	UACUGGCCCCCAAAUCUUCGCU	22	++	3′	X	1.7 kb upstream of	C39D10.3	(s)
mir-241	MS, C, N	UGAGGUAGGUGCGAGAAAUGA	21	++	5′	V	1.8 kb upstream of	*mir-48*	(s)
mir-242	C, N	UUGCGUAGGCCUUUGCUUCGA	21	++	5′	IV	0.9 kb downstream of	*nhr-78*	(as)
mir-243	C, N	CGGUACGAUCGCGGCGGGAUAUC	22–23	–	3′	IV	1.0 kb upstream of	R08C7.1	(s)
mir-244	C, N	UCUUUGGUUGUACAAAGUGGUAUG	23–25	+++	5′	I	1.6 kb downstream of	T04D1.2	(as)
mir-245	C, N	AUUGGUCCCCUCCAAGUAGCUC	22	+++	3′	I	1.9 downstream of	F55D12.1	(s)
mir-246	C, N	UUACAUGUUUCGGGUAGGAGCU	22	++	3′	IV	0.4 kb downstream of	ZK593.8	(s)
mir-247	C, N	UGACUAGAGCCUAUUCUCUUCUU	22–23	–	3′	X	1.9 kb upstream of	C39E6.2	(as)
mir-248	MS, C, N	UACACGUGCACGGAUAACGCUCA	23	++	3′	X	within intron of	AH9.3	(s)
mir-249	C	UCACAGGACUUUUGAGCGUUGC	22–23	++	3′	X	2.7 kb upstream of	Y41G9A.6	(s)
mir-250	MS, N, PCR	UCACAGUCAACUGUUGGCAUGG	~22	++	3′	V	0.1 kb downstream of	*mir-61*	(s)
mir-251	MS, N, PCR	UUAAGUAGUGGUGCCGCUCUUAUU	~24	+++	5′	X	0.2 kb downstream of	F59F3.4	(as)
mir-252	MS, N, PCR	UAAGUAGUAGUGCCGCAGGUAAC	~23	+++	5′	II	1.8 kb downstream of	VW02B12L.4	(as)
mir-253	MS, D, PCR	CACACCUCACUAACACUGACC	n.d.	++	5′	V	within intron of	F44E7.5	(s)
mir-254	MS, D, PCR	UGCAAAUCUUUCGCGACUGUAGG	n.d.	++	3′	X	within intron of	ZK455.2	(s)
mir-255	MS, D	—	n.d.				1.5 kb upstream of	F08F3.9	(as)

For predicted stem-loop precursors, see Supplementary Fig. 2. Genes were identified and validated as indicated in the ID method column: MS, candidate gene had high MiRscan score (Table 1); C, miRNA was cloned and sequenced (Table 1); N, expression of the mature miRNA was detectable on Northern blots; D, the miRNA stem-loop precursor was detected on Northern blots and enriched in RNA from *dcr-1* animals, but the mature miRNA was not detected; PCR, targeted PCR amplification and sequencing detected the miRNA in a library of *C. elegans* small RNAs; H, the locus was closely homologous to that of a validated miRNA. For the miRNAs cloned and sequenced, some miRNAs were represented by clones of different lengths, due to heterogeneity at the miRNA 3′ terminus. The observed range in length is indicated, and the sequence of the most abundant length is shown. For the RNAs that have not been cloned, the 5′ terminus was determined by the PCR assay, but the 3′ terminus was not determined. For *mir-250*, *mir-251*, and *mir-252*, the length of the miRNA sequence shown was inferred from the Northern blots; for other miRNAs not cloned, the length was not determined (n.d.). For *mir-254*, the PCR assay detected ~22-nt RNAs from both sides of the fold-back, representing both the miRNA and the miRNA*. Their relative positions within the precursor suggest that the RNA from the 5′ arm is 22 nt and the RNA from the 3′ arm is 23 nt. The RNA from the 3′ arm was chosen as the miRNA because of its similarity to the human miR-19 gene family. The miR-255 gene is known only as the precursor, a conserved stem loop with Dicer-dependent processing (Fig. 2b). Comparison to *C. briggsae* shotgun traces from the *C. briggsae* Sequencing Consortium (obtained from www.ncbi.nlm.nih.gov) revealed miRNA orthologs with 100% sequence identity (+++) and potential orthologs with >90% (++) and >75% (+) sequence identity. To indicate the genomic loci of the genes, the chromosome (Chr.), distance to nearest annotated gene, and the orientation relative to that gene, sense (s) or antisense (as), are specified.

Molecular abundance of miRNAs

The very high cloning frequency of certain miRNAs (e.g., miR-52, represented by >400 clones) raised the question as to the molecular abundance of these and other miRNA species. In addition, there was the question of whether the actual molecular abundance of miRNAs in nematodes was proportionally reflected in the numbers of clones sequenced. To address these questions, quantitative Northerns were used to examine the molecular abundance of 12 representative miRNAs, picked so as to span the range of frequently and rarely cloned sequences and differing 3′ and 5′ terminal residues (Fig. 5).

To determine the molecular abundance of these 12 miRNAs in the adult worm soma, the hybridization signals for RNA from a known number of *glp-4* young adult worms were compared with standard curves from chemically synthesized miRNAs (Fig. 5; Hutvágner and Zamore 2002). Accounting for RNA extraction yields and dividing the number of miRNA molecules per worm by the total number of cells in the worms, yielded averages of up to 50,000 molecules per cell, with the most abundant miRNAs as plentiful as the U6 snRNA of the spliceosome (Fig. 5C). These are much higher numbers than those for the typical worm mRNAs, estimated to average ~100 molecules per cell for the 5000 most highly expressed genes in the cell. [This estimate was calculated based on our yield of 20 pg total RNA per worm cell, assuming that the 5000 most highly expressed genes have mRNAs averaging 2 kb in length and represent 3% of the total RNA in an adult worm; it was consistent with estimates based on hybridization kinetics of

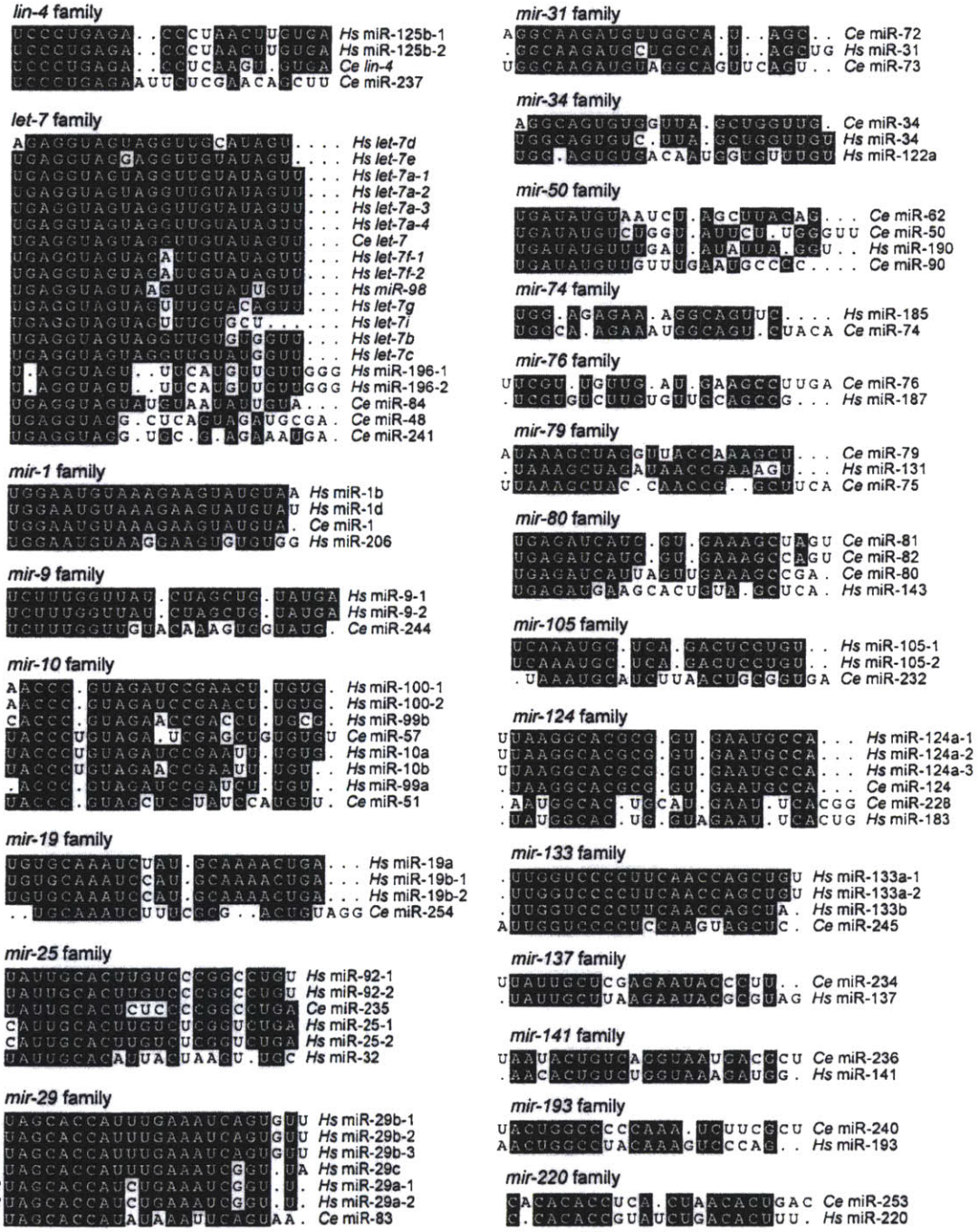


Figure 3. Alignments of *C. elegans* and human miRNA sequences that can be grouped together in families. Human miRNAs (*Hs*) are those identified in human cells (Lagos-Quintana et al. 2001; Mourelatos et al. 2002) or are orthologs of miRNAs identified in other vertebrates (Lagos-Quintana et al. 2002, 2003; Lim et al. 2003).

mRNAs from mouse tissues (Hastie and Bishop 1976).] Perhaps high concentrations of miRNAs are needed to saturate the relevant complementary sites within the target mRNAs, which might be recognized with low affinity because of the noncanonical pairs or bulges that appear to be characteristic of the animal miRNA–target interactions.

Because these numbers represent molecular abundance averaged over all the cells of the worm, including cells that might not be expressing the miRNA, there are

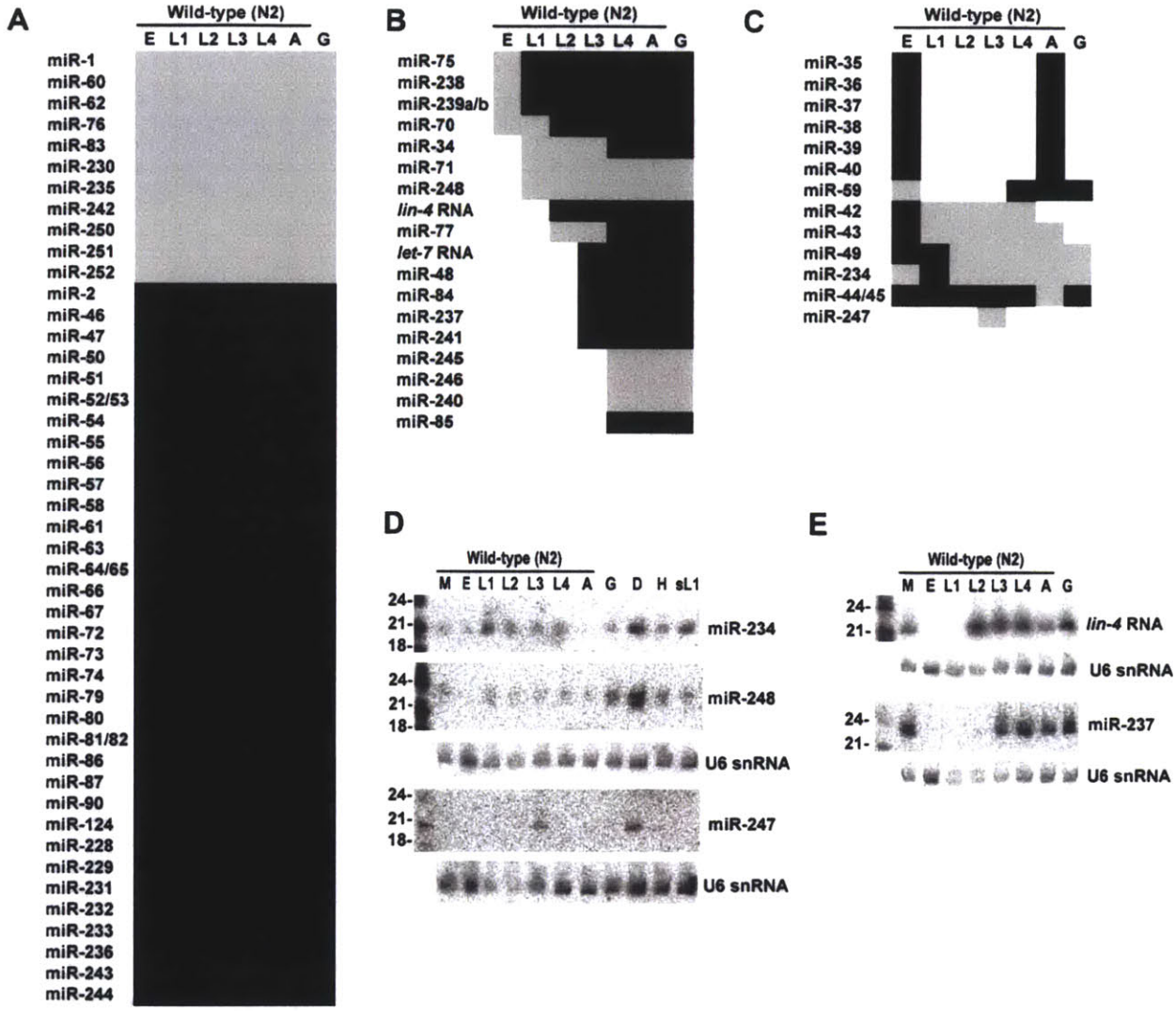


Figure 4. Expression of *C. elegans* miRNAs during larval development. Total RNA was analyzed from mixed-stage N2 worms (M), embryos (E), larval stages (L1, L2, L3, L4), adults (A), *glp-4(bn2)* adults (G), N2 dauers (D), mixed-stage *him-8(e1489)* worms (H), and N2 starvation-arrested L1 larvae (sL1). Intense signals are represented as black rectangles and faint signals are represented as gray rectangles. Of the 87 *C. elegans* miRNAs identified, 6 could not be detected on developmental Northerns (miR-41, miR-78, miR-249, miR-253, miR-254, and miR-255). (*A*) miRNAs constitutively expressed throughout nematode development. (*B*) stRNAs, *lin-4* and *let-7*, and similarly expressed miRNAs, which commence expression during larval development and remain expressed through adulthood. (*C*) miRNAs with discontinuous developmental expression patterns. (*D*) Northern analysis of miRNAs with enhanced expression in the dauer stage. To control for loading, the blot used for both miR-234 and miR-248 and the blot used for miR-247 were reprobed for the U6 snRNA (U6). Quantitation with a PhosphorImager showed that the lane-to-lane variation in U6 signal was as great as threefold. Normalizing to the U6 signal, the miR-248 signal was fourfold greater in dauer than in most other stages, except for *glp-4* adults, in which it was twofold greater, whereas the miR-234 signal was highest in dauer and L1, with a signal in these stages about twofold greater than the average of the other stages. (*E*) Northern analysis of the *lin-4* RNA and its paralog, miR-237.

likely to be some cells that express even more molecules of the miRNA. To examine the abundance in a single cell type, HeLa RNA was probed for representative human miRNAs, yielding a similar range of molecular abundance (Fig. 5C). The high number of miRNA molecules in human cells increases the mystery as to why miRNAs had gone undetected for so long, which raises the question of whether other classes of highly expressed ncRNAs might yet remain to be discovered. A recent large-scale analysis of full-length cDNAs from mouse indicates the possible existence of hundreds or thousands of expressed ncRNAs in vertebrates (Okazaki et al. 2002).

To address the extent to which the actual molecular abundance of miRNAs in nematodes is proportionally reflected in the numbers of clones sequenced, the abun-

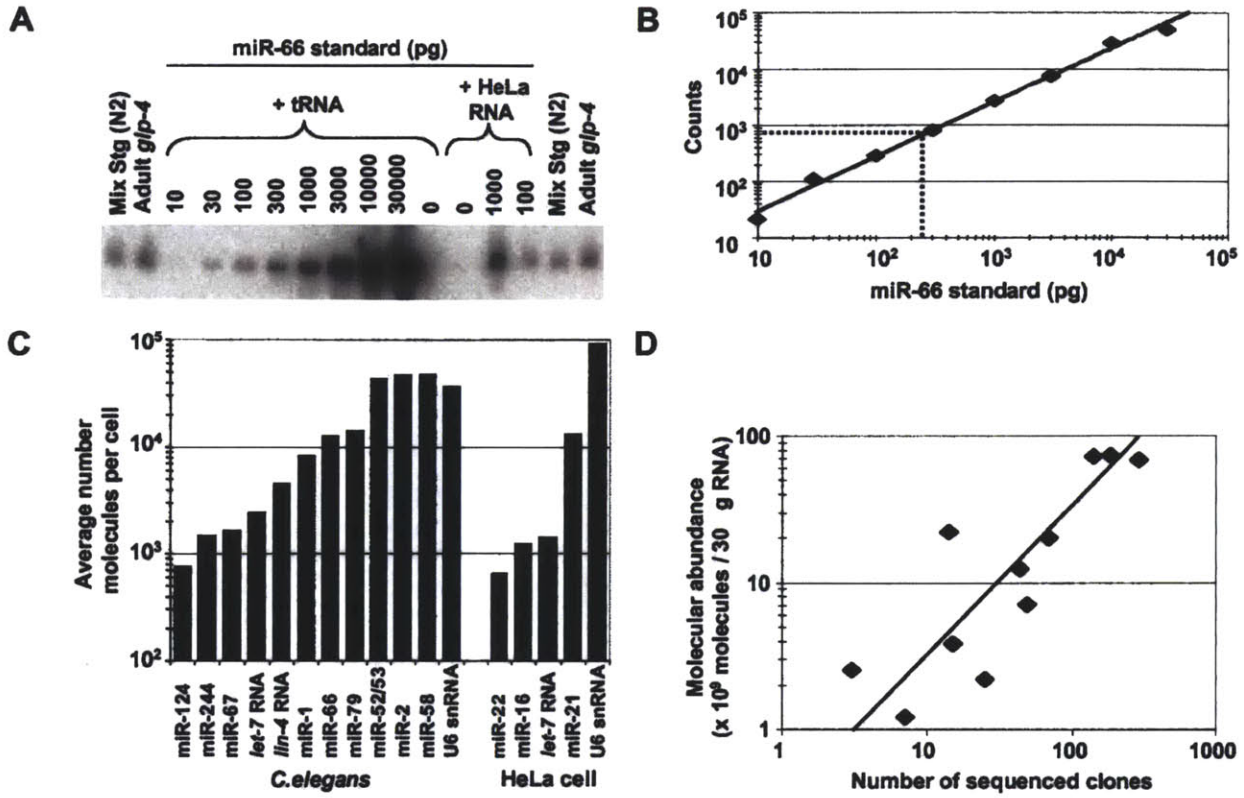


Figure 5. Quantitative analysis of miRNA expression. (*A*) Northern blot used to quantify the abundance of miR-66. RNA prepared from the wild-type (N2) mixed-stage worms used in cloning and from *glp-4(bn2)* young adult worms were run in duplicate with a concentration course of synthetic miRNA standard. The signal from the standard did not change when total RNA from HeLa cells replaced *E. coli* tRNA as the RNA carrier, showing that the presence of other miRNAs did not influence membrane immobilization of the miRNA or hybridization of the probe. (*B*) Standard curve from quantitation of miR-66 concentration course. The best fit to the data is a line represented by the equation $y = 3.3x^{0.96}$ ($R^2 = 0.99$). Interpolation of the average signal in the *glp-4* lanes indicates that the *glp-4* samples contain 240 pg of miR-66 (broken lines). (*C*) Molecular abundance of miRNAs and U6 snRNA. Amounts of the indicated RNA species in the *glp-4* samples were determined as shown in *A* and *B*. The average number of molecules per cell was then calculated considering the number of animals used to prepare the sample, and the yield of a radiolabeled miRNA spiked into the preparation at an early stage of RNA preparation. Analogous experiments were performed to determine the amounts of the indicated human miRNAs in HeLa RNA samples. (*D*) Correlation between miRNA molecular abundance and cloning frequency. The number of molecules in the mixed-stage RNA samples was determined as described for the *glp-4* samples and then plotted as a function of the number of times the miRNAs was cloned from this mixed-stage population (Table 1). The line is best fit to the data and is represented by the equation $y = 0.32x$ ($R^2 = 0.78$).

dance of the miRNA within the mixed-stage RNA preparation was compared with the number of clones generated from that preparation (Fig. 5D). The strong positive correlation observed between the molecular abundance and the number of times the miRNAs were cloned indicated that systematic biases in the cloning procedure were not major. At most, these miRNAs were over- or underrepresented fivefold in the sequenced set relative to their actual abundance as measured by quantitative Northerns. We cannot rule out the possibility that certain miRNAs not yet cloned might be refractory to our cloning procedure, for example, because of a propensity to form secondary structures that preclude adaptor ligation reactions. Nonetheless, on the whole, the cloning frequencies can be used to approximate the molecular abundance of the miRNAs, and we have no reason to suspect that the set of miRNAs identified by cloning differs in any substantive way, other than an overall higher steady-state expression level, from the complete set of *C. elegans* miRNAs.

Other endogenous ~22-nt RNAs of C. elegans

Of the 4078 *C. elegans* clones, a large majority represented authentic miRNAs (3423 clones, Table 1). The next most abundant class represented degradation fragments of larger ncRNAs, such as tRNA and rRNA (447 clones) and introns (18 clones). The remaining clones represented potential Dicer products that were not classified as miRNAs. Some corresponded to sense (18 clones) or antisense (23 clones) fragments of known or predicted mRNAs and might represent endogenous

siRNAs. Others (143 clones) corresponded to regions of the genome not thought to be transcribed; these might represent another type of endogenous siRNAs, known as heterochromatic siRNAs (Reinhart and Bartel 2002). The possible roles of the potential siRNAs and heterochromatic siRNAs in regulating gene expression are still under investigation. The remaining clones were difficult to classify because they matched more than one locus, and their loci were of different types (six clones).

A fourth class of potential Dicer products (38 clones, representing 14 loci) corresponded to miRNA precursors but derived from the opposite arm of the hairpin than the more abundantly expressed miRNA, as has been reported previously for miR-56 in *C. elegans*, miR156d and miR169 in plants, and several vertebrate miRNAs (Lau et al. 2001; Lagos-Quintana et al. 2002, 2003; Mourelatos et al. 2002; Reinhart et al. 2002). Our current data add another 13 examples of this phenomenon (Fig. 6). In all of our cases, the ~22-nt RNA from one arm of the fold-back was cloned much more frequently than that from the other and was far more readily detected on Northern blots. We designated the less frequently cloned RNA as the miRNA-star (miRNA*) fragment (Lau et al. 2001).

Discussion

We have developed a computational procedure for identifying miRNA genes conserved in two genomes. By using this procedure, together with extensive sequencing of clones from libraries of small RNAs, we have now identified 87 miRNA genes in *C. elegans* (Tables 1, 2). Together with *mir-88* (Lee and Ambros 2001), which we have not yet cloned or found computationally, the number of validated *C. elegans* genes stands at 88. More than a third of these genes have human homologs (Fig. 3), and a similar fraction, including previously unrecognized *lin-4* and *let-7* paralogs, is differentially expressed during larval development (Fig. 4). Most miRNAs accumulated to very high steady-state levels, with some at least as plentiful as the U6 snRNA (Fig. 5). Below, we discuss some implications of these results with regard to some of the defining features of miRNA genes in animals, the processing of miRNA precursors, and the number of miRNA genes remaining to be identified.

MiRscan accuracy and the defining features of miRNAs

As calculated in the Results section, the specificity of MiRscan was ≥0.70 at a sensitivity that detects half the previously known *C. elegans* miRNAs, when starting from an assembled *C. elegans* genome and *C. briggsae* shotgun reads. This accuracy was sufficient to identify new genes and obtain an upper bound on the total number of miRNA genes in the worm genome (described later). However, it was not sufficient to reliably identify all the conserved miRNA genes in *C. elegans*. The accuracy of MiRscan appears to be at least as high as that of general methods to identify ncRNA genes in bacteria (Argaman et al. 2001; Rivas et al. 2001; Wassarman et al. 2001), but is lower than that of algorithms designed to identify protein-coding genes or specialized programs that predict tRNAs and snoRNAs (Lowe and Eddy 1997, 1999; Burge and Karlin 1998). The relative difficulty in identifying miRNAs can be explained by the low information content inherent in their small size and lack of strong primary sequence motifs. The performance of

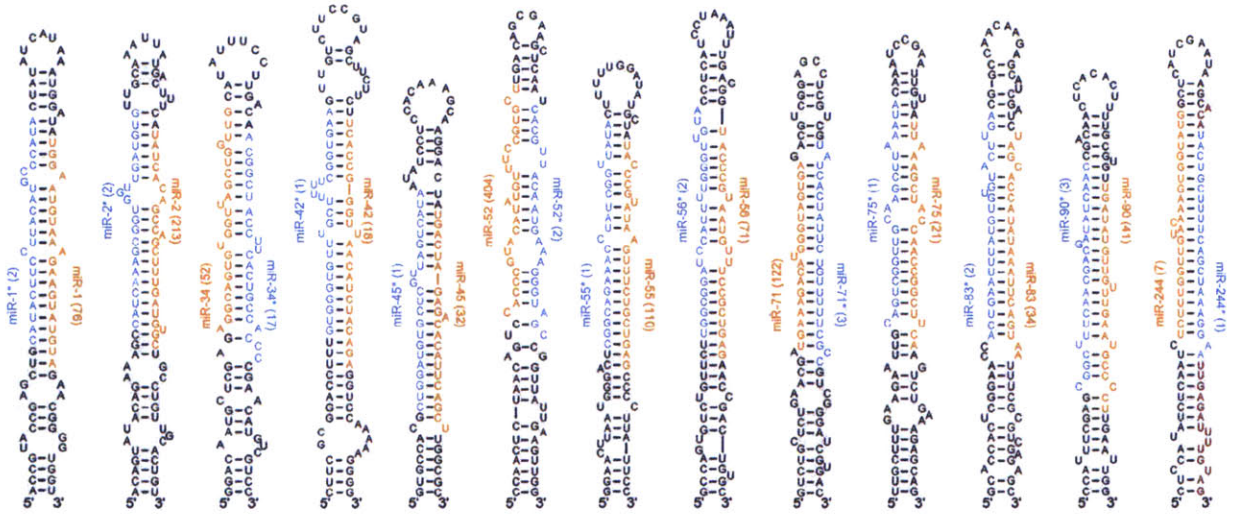


Figure 6. miRNA (red) and miRNA* (blue) sequences within the context of their predicted fold-back precursors. The number of sequenced clones is shown in parentheses. For each miRNA and miRNA*, colored residues are those for the most frequently cloned species. There was 3′ heterogeneity among the sequenced clones for some miRNA*s and most miRNAs. Heterogeneity at the 5′ terminus was not seen among the sequenced clones for the miRNA*s and was rare among those for the miRNAs; when it occurred, it was not observed for more than one of the many clones representing each miRNA.

MiRscan will improve with a more complete and assembled *C. briggsae* genome. We anticipate that using only those sequences conserved in a syntenic alignment of the two genomes would capture fewer of the background sequences, enabling the authentic miRNAs to be more readily distinguished from the false positives.

Improvement would also come from bringing in a third nematode genome, particularly a genome more divergent than those of *C. elegans* and *C. briggsae*. The advantage of such an additional genome is illustrated by our application of MiRscan to the identification of vertebrate miRNAs using three genomes. The version of MiRscan described here, which had been trained on the set of 50 miRNAs conserved in worms, was applied to the assembled human genome, shotgun reads of the mouse genome, and the assembled pufferfish (*Fugu*) genome (Lim et al. 2003). This analysis had a specificity of ≥0.71 at a sensitivity that detected three fourths of the previously known vertebrate miRNAs. The accuracy of the vertebrate analysis was therefore substantially improved over that of the *C. elegans*/*C. briggsae* analysis, even though the vertebrate genomes are 4–30 times larger than those of *C. elegans* and *C. briggsae*, and are expected to have a correspondingly higher number of background sequences. This improved performance can be attributed to using three genomes, as well as to the evolutionary distance between the mammalian and fish genomes, which are distant enough to reduce the number of fortuitously high scoring sequences, yet close enough to retain most of the known miRNAs.

Other improvements in the computational identification of miRNAs will come with the definition of additional sequence and structural features that specify which sequences are transcribed, processed into miRNAs, and loaded into the miRNP. With the exception of sequence conservation, the features that MiRscan currently uses to identify miRNAs (Fig. 1A) are among those that the cell also uses to specify the biogenesis of miRNAs and miRNPs. The utility of these parameters for MiRscan (Fig. 1B) is a function of both the degree to which these features are correctly modeled (or have already been used to restrict the number of miRNA candidates; see Fig. 1B legend) and their relative importance in vivo. Clearly, much of what defines a miRNA in vivo remains to be determined. Sequence elements currently unavailable for MiRscan include transcriptional promoter and termination signals. Additional sequence and structural features important for processing of the primary transcript and the hairpin precursors also remain to be identified (Lee et al. 2002).

miRNA biogenesis

The presence of miRNA* species, observed now for 14 of the *C. elegans* miRNAs (Fig. 6; Lau et al. 2001), provides evidence for the idea that Dicer processing of miRNA precursors resembles that of siRNA precursors (Hutvágner and Zamore 2002; Reinhart et al. 2002). We suspect that with more extensive sequencing of clones, miRNA* sequences will be found for a majority of the miRNA precursors, a notion supported by the identification of additional miRNA* sequences using our PCR assay (data not shown). As observed for both *MIR156d* and *MIR169* in plants (Reinhart et al. 2002), the miRNA:miRNA* segments are typically presented within the predicted precursor, paired to each other with 2-nt 3′ overhangs (Fig. 6)—a structure analogous to that of a classical siRNA duplex. This is precisely the structure that would be expected if both the miRNA and the miRNA* were excised from the same precursor molecule, and the miRNA* fragments were transient side-products of productive Dicer processing. An alternative model for miRNA biogenesis and miRNA* formation, which we do not favor but cannot rule out, is that the Dicer complex normally excises a ~22-nt RNA from only one side of a miRNA precursor but it sometimes binds the precursors in the wrong orientation and excises the wrong side. In an extreme version of the favored model, the production of the miRNA* would be required for miRNA processing and miRNP assembly; in a less extreme version, miRNA* production would be an optional off-pathway phenomenon. The idea that ~22-nt RNAs might be generally excised from both sides of the same precursor stem loop brings up the question of why the miRNAs and miRNA*s are present at such differing levels. With the exception of miR-34* (sequenced 17 times), none of the miRNA*s is represented by more than three sequenced clones. Perhaps the miRNAs are stabilized relative to their miRNA* fragments because they preferentially enter the miRNP/RISC complex. Alternatively, both the miRNA and the miRNA* might enter the complex, but the miRNA might be stabilized by interactions with its targets.

Five of the newly identified miRNAs are within annotated introns, all five in the same orientation as the predicted mRNAs. When considered together with the previously identified miRNAs found within annotated introns (Lau et al. 2001), 10 of 12 known *C. elegans* miRNAs predicted to be in introns are in the same orientation as the predicted mRNAs. This bias in orientation, also reported recently for mammalian miRNAs (Lagos-Quintana et al. 2003), suggests that some of these miRNAs are not transcribed from their own promoters but instead derive from the excised pre-mRNA introns (as are many snoRNAs), and it is easy to imagine regulatory scenarios in which the coordinate expression of a miRNA with an mRNA would be desirable.

The number of miRNA genes in C. elegans *and other animals*

In addition to providing a set of candidate miRNAs, MiRscan scoring provides a means to estimate the total number of miRNA genes in *C. elegans*. A total of 64 loci have scores greater than the median score of the 58 initially reported *C. elegans* miRNAs (Fig. 2B). Note that this set of 58 miRNAs includes not only the 50 conserved miRNAs of the training set but also the eight previously reported miRNAs that were not in our set of

36,000 potential stem loops, usually because they lacked easily recognizable *C. briggsae* orthologs. Thus, the estimate calculated below takes into account the poorly conserved miRNAs without MiRscan scores. Four of the 64 high-scoring loci are known to be false positives. Thus, the upper bound on the number of miRNA genes in *C. elegans* would be $2 \times (64 - 4)$, or 120. This upper bound of ~120 genes remained stable when extrapolating from points other than the median, ranging from the top 25th–55th percentiles. For this estimate, we made the assumption that the set of all *C. elegans* miRNAs has a distribution of MiRscan scores similar to the distribution of initially reported miRNAs. Such an assumption might be called into question, particularly when considering that the initially reported miRNAs served as a training set for the development of MiRscan (even though the scores of the training-set loci have been jackknifed to prevent overfitting). However, this assumption is supported by two observations. First, the set of newly cloned miRNAs did indeed have a distribution of scores indistinguishable from that of the training set of previously reported miRNAs (Fig. 2B). Second, there is no correlation between the number of times that a miRNA has been cloned and its MiRscan score (Fig. 7). The absence of a correlation between cloning frequency and MiRscan score lessens our concern that miRNAs that are difficult to clone, including those still not present in our set of 3423 sequenced clones, might represent a population of miRNAs that are refractory to computational analysis as well.

This estimate of 120 genes is an upper bound and would decrease if additional high-scoring candidates were shown to be false positives. The extreme scenario, in which all are false positives, places the lower bound of miRNA genes near the number of validated genes, adding perhaps another five genes to account for the low-scoring counterparts of the five computational candidates validated only by Northerns and PCR, yielding a lower bound on the number of *C. elegans* miRNAs of ~93.

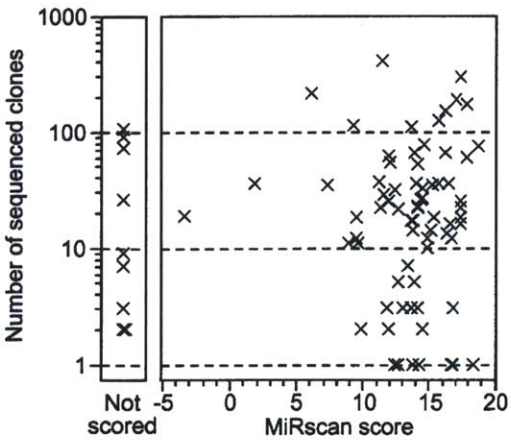


Figure 7. Plot illustrating the absence of a correlation between the MiRscan score of a cloned miRNA and the number of times that miRNA was cloned and sequenced. Nine of 80 cloned loci of Table 2 were not scored (*left*) because potential homologs of these genes were not identified among the available *C. briggsae* sequencing reads.

Our count of 105 ± 15 miRNA genes in *C. elegans* might underestimate the true count if there are miRNAs with unusual fold-back precursors that were cloned but dismissed as endogenous siRNAs or degradation fragments. To investigate this possibility, we examined the expression of each small RNA that was cloned more than once but did not appear to derive from a canonical miRNA precursor as predicted by RNAfold. Because most (72 of 88) of the authentic miRNAs identified to date were represented by multiple clones (Table 1), this analysis should uncover most of the miRNAs coming from nonconventional precursors. This broader analysis detected only a single additional miRNA, miR-229. All of the other sequences that we cloned more than once were minor degradation fragments or processing byproducts of larger ncRNAs (e.g., the 5′ leader sequence of a tRNA). Thus, the number of miRNAs that derive from nonconventional precursors is not sufficient to significantly influence the miRNA gene count.

The estimated number of miRNA genes represents between 0.5% and 1% of the genes identified in the *C. elegans* genome, a fraction similar to that seen for other very large gene families with presumed regulatory roles, such as those encoding nuclear hormone receptors (270 predicted genes), C2H2 Zinc-finger proteins (157 predicted genes), and homeodomain proteins (93 predicted genes; Chervitz et al. 1998; *C. elegans* Sequencing Consortium 1998). Extending our analysis to vertebrate genomes revealed that 230 ± 30 of the human genes are miRNAs, also nearly 1% of the genes in the genome (Lim et al. 2003). The miRNA genes are also among the most abundant of the ncRNA gene families in humans, comparable in number to the genes encoding rRNAs (~650–900 genes), tRNAs (~500 genes), snRNAs (~100 genes), and snoRNAs (~100–200 genes; Lander et al. 2001). For rRNAs, tRNAs, and snRNAs, the hundreds of gene copies in the human genome represent only relatively few distinct genes, probably <100 distinct genes for all three classes combined. For the miRNAs and snoRNAs, there are many more distinct genes, and each is present in only one or a few copies.

Unlike the other large ncRNA gene families and many of the transcription-factor gene families, there is no indication that miRNAs are present in single-celled organisms such as yeast. A pilot attempt to clone miRNAs from *Schizosaccharomyces pombe* did not detect any miRNAs (Reinhart and Bartel 2002), and there is no evidence that the proteins (such as Dicer) needed for miRNA accumulation in plants and animals are present in *Saccharomyces cerevisiae*. Given the known roles of miRNAs in *C. elegans* development (Lee et al. 1993; Wightman et al. 1993; Reinhart et al. 2000) and the very probable roles of miRNAs in plant development (Rhoades et al. 2002), it is tempting to speculate that the substantial expansion of miRNA genes in animals (and the apparent loss of miRNA genes in yeast) is related to

their importance in specifying cell differentiation and developmental patterning, and that the extra layer of gene regulation afforded by miRNAs was crucial for the emergence of multicellular body plans. The identification of most of the worm miRNAs and the quantitation of the number of genes remaining to be found are important steps toward understanding the evolution of this intriguing class of genes and placing them within the gene regulatory circuitry of these and other animals.

Materials and methods

Computational identification of stem loops

Potential miRNA stem loops were located by sliding a 110-nt window along both strands of the *C. elegans* genome (WormBase release 45, http://www.wormbase.org) and folding the window with the secondary structure-prediction program RNAfold (Hofacker et al. 1994) to identify predicted stem-loop structures with a minimum of 25 bp and a folding free energy of at least 25 kcal/mole ($\Delta G^\circ_{\text{folding}} \leq -25$ kcal/mole). Sequences that matched repetitive elements were discarded, as were those with skewed base compositions not observed in known miRNA stem loops and those that overlapped with annotated coding regions. Stem loops that had fewer base pairs than overlapping stem loops were also culled. *C. briggsae* sequences with at least loose sequence similarity to the remaining *C. elegans* sequences were identified among *C. briggsae* shotgun sequencing reads (November 2001 download from http://www.ncbi.nlm.nih.gov/Traces) using WU-BLAST with default parameters and a nonstringent cutoff of E < 1.8 (W. Gish, http://blast.wustl.edu). These *C. briggsae* sequences were folded with RNAfold to ensure that they met the minimal requirements for a hairpin structure as described above. This procedure yielded ~40,000 pairs of potential miRNA hairpins. For each pair of potential miRNA hairpins, a consensus *C. elegans*/*C. briggsae* structure was generated using the alidot and pfrali utilities from the Vienna RNA package (Hofacker et al. 1998; Hofacker and Stadler 1999; http://www.tbi.univie.ac.at/~ivo/RNA). To create RNA consensus structures, alidot and pfrali combine a Clustal alignment (Thompson et al. 1994) of a pair of sequences with either the minimum free energy structures of these sequences (alidot) derived using the Zuker algorithm (Zuker 1994) or the base pairing probability matrices of these sequences (pfrali) derived using the McCaskill algorithm (McCaskill 1990).

MiRscan

Of the ~40,000 pairs of hairpins, 35,697 had the minimal conservation and base pairing needed to receive a MiRscan score. Among this set were 50 of the 53 previously published miRNAs that were reported to be conserved between *C. elegans* and *C. briggsae* (Lau et al. 2001; Lee and Ambros 2001). [miR-53 is included as a previously reported conserved miRNA because it is nearly identical to miR-52, which has a highly conserved *C. briggsae* ortholog (Lau et al. 2001; Lee and Ambros 2001). The three conserved genes missing from the ~36,000 pairs of hairpins were *mir-56*, *mir-75*, and *mir-88*. The reverse complements of *mir-75* and *mir-88* were later observed among the ~36,000 hairpins and given scores (Table 1).] The MiRscan program was developed to discriminate these 50 known miRNA hairpins from background sequences in the set of ~36,000 hairpins. For a given 21-nt miRNA candidate, MiRscan makes use of the seven features derived from the consensus hairpin structure illustrated in Figure 1A: x_1, "miRNA base pairing," the sum of the base-pairing probabilities for pairs involving the 21-nt candidate miRNA; x_2, "extension of base pairing," the sum of the base-pairing probabilities of the pairs predicted to lie outside the 21-nt candidate miRNA but within the same helix; x_3, "5′ conservation," the number of bases conserved between *C. elegans* and *C. briggsae* within the first 10 bases of the miRNA candidate; x_4, "3′ conservation," the number of conserved bases within the last 11 bases of the miRNA candidate; x_5, "bulge symmetry," the number of bulged or mismatched bases in the candidate miRNA minus the number of bulged or mismatched bases in the corresponding segment on the other arm of the stem loop; x_6, "distance from loop," the number of base pairs between the loop of the stem loop and the closest end of the candidate; and x_7, "initial pentamer," the specific bases at the first five positions at the candidate 5′ terminus.

For a given feature i with a value x_i, MiRscan assigns a log-odds score

$$s_i(x_i) = \log_2\left(\frac{f_i(x_i)}{g_i(x_i)}\right),$$

where $f_i(x_i)$ is an estimate of the frequency of feature value x_i in miRNAs derived from the training set of 50 known miRNAs, and $g_i(x_i)$ is an estimate of the frequency of feature value x_i among the background set of ~36,000 hairpin pairs. The overall score assigned to a candidate miRNA is simply the sum of the log-odds scores for the seven features:

$$S = \sum_{i=1..7} s_i(x_i).$$

To score a given hairpin, MiRscan slides a 21-nt window representing the candidate miRNA along each arm of the hairpin, assigns a score to each window, and then assigns the hairpin the score of its highest-scoring window. In order to be evaluated, a window was required to be two to nine consensus base pairs away from the terminal loop.

For features x_1, x_3, x_4, x_5, and x_6, f_i and g_i were obtained by smoothing the empirical frequency distributions from the training and background sets, respectively, using the R statistical package (http://lib.stat.cmu.edu/R/CRAN) with a triangular kernel. Because x_1 and x_2 are not independent of each other, the relative contribution of x_2 was decreased by computing f_2 and g_2 separately subject to the conditions $x_1 \geq 9$ and $x_1 < 9$, in order to account for this dependence. For x_7, a weight matrix model (WMM) was generated for the five positions at the miRNA 5′ terminus. The background WMM, g_7, was set equal to the base composition of the background sequence set. The miRNA WMM, f_7, was derived from the position-specific base frequencies of the 50 training set sequences, using standard unit pseudo-counts and normalizing for the contributions of related miRNAs.

Because both strands of the *C. elegans* genome were analyzed, both a hairpin sequence and its reverse complement were sometimes included in the set of ~36,000 stem loops. For representation in Figure 2, in such cases both sequences were considered as a single locus that received the score of the higher scoring hairpin. Also, to prevent overscoring of the 50 known miRNA loci within the training set, each known miRNA locus was assigned a jackknife score calculated by using a training set consisting of the other 49 miRNAs. MiRscan is available for use (http://genes.mit.edu/mirscan).

RNA cloning and bioinformatic analyses

Small RNAs were cloned as described previously (Lau et al. 2001), using the protocol available on the Web (http://web.

wi.mit.edu/bartel/pub). Sequencing was performed by Agencourt Bioscience. Sequences of known *C. elegans* tRNA and rRNA were removed, and the remaining clones were clustered based on the location of their match to the *C. elegans* genome (*C. elegans* Sequencing Consortium 1998), downloaded from WormBase (http://www.wormbase.org). Genomic loci not previously reported to encode miRNAs were examined by using the RNA-folding program RNAfold (Hofacker et al. 1994). Two sequences were folded for each locus: one included 15 nt upstream and 60 nt downstream of the most frequently cloned sequence from that locus; the other included 60 nt upstream and 15 nt downstream. Sequences for which the most stable predicted folding resembled the stem-loop precursors of previously validated miRNAs were carried forward as candidate miRNA loci. Sequences without classical stem-loop precursors were also analyzed further (see Discussion), but only one, miR-229, was classified as a miRNA. The clones classified as representing potential fragments of mRNAs (18 clones) and potential antisense fragments of mRNAs (23 clones) corresponded to predicted ORFs (as annotated in GenBank) or probable UTR segments (100 bp upstream or 200 bp downstream of the predicted ORF).

Northerns

Expression of candidate miRNA loci was examined by using Northern blots and radiolabeled DNA probes (Lau et al. 2001). To maintain hybridization specificity without varying hybridization or washing conditions, the length of probes for different sequences was adjusted so that the predicted melting temperatures of the miRNA-probe duplexes did not exceed 60°C (Sugimoto et al. 1995). Probes not corresponding to the entire miRNA sequence were designed to hybridize to the 3′ region of the miRNA, which is most divergent among related miRNA sequences.

PCR validation

A PCR assay was performed to detect the sequences of predicted miRNAs within a cDNA library constructed from 18- to 26-nt RNAs expressed in mixed-stage worms. This library, the same as that used for cloning (Lau et al. 2001), consisted of PCR-amplified DNA that comprised the 18- to 26-nt sequences flanked by 3′- and 5′-adaptor sequences. For each miRNA candidate, a primer specific to the predicted 3′ terminus of the candidate and a primer corresponding to the 5′-adaptor sequence common to all members of the library (ATCGTAG GCACCTGAAA) were used at concentrations of 1.0 μM and 0.1 μM, respectively (100 μL PCR reaction containing 5 μL of a 400-fold dilution of the PCR reaction previously used to amplify all members of the cDNA library). The specific primer was added after the initial denaturation incubation had reached 80°C. After 20 PCR cycles, the reaction was diluted 20-fold into a fresh PCR reaction for another 20 cycles. PCR products were cloned and sequenced to both identify the 5′ terminus of the miRNA and ensure that the amplified product was not a primer-dimer or other amplification artifact. Specific primers for the reactions that successfully detected candidate miRNAs were ACCATGCCAACAGTTG (miR-250), TAAGAGCGGCACCA CTAC (miR-251), TACCTGCGGCACTACTAC (miR-252), GTCAGTGTTAGTGAGG (miR-253), TACAGTCGGAAAGA TTTG (miR-254), and GTGGAAATCTATGCTTC (miR-254*).

Quantitative Northerns

miRNA standards (purchased from Dharmacon) were diluted to appropriate concentrations in the presence of 1.0 μg/μL carrier RNA in the form of either *E. coli* tRNA or HeLa cell total RNA. Northern analysis was performed (Lau et al. 2001), loading 30 μg of RNA per lane, in the format shown for miR-66 (Fig. 5A). Signals were quantitated using phosphor imaging, standard curves (linear through at least three orders of magnitude, including the region of interpolation) were constructed, and absolute amounts of miRNAs per sample were determined, as illustrated for miR-66 (Fig. 5B). The average number of miRNA molecules per *glp-4* adult nematode was calculated using 19 ng as the average amount of total RNA extracted per worm. This number was determined as the average of three independent extraction trials, from known numbers of synchronized, 2-day-old adult *glp-4(bn2)* hermaphrodites, the same frozen worm population used for the quantitative Northern blots. All extractions were performed as described previously (Lau et al. 2001), except during two of the trials a radiolabeled miRNA was spiked into the preparation during worm lysis. At least 90% of this RNA was recovered, indicating near quantitative yield. Having calculated the number of each miRNA per worm, the average number of miRNAs per cell was calculated using 989 as number of cells per worm. The 989 cells per worm is based on the 959 somatic nuclei of the adult hermaphrodites plus the 30 germ nuclei of 2-day-old adult *glp-4(bn2)* animals (Sulston et al. 1983; Beanan and Strome 1992). Total RNA from known numbers of HeLa cells was determined in an analogous fashion.

Acknowledgments

We thank the *C. briggsae* Sequencing Consortium for the availability of sequencing reads, WormBase (http://www.wormbase.org) for annotation of the *C. elegans* genome, Compaq for computer resources, V. Ambros for communicating unpublished data, C. Mello for the *dcr-1* strain, S. Griffiths-Jones and the miRNA Gene Registry for assistance with gene names, P. Zamore for helpful comments on this manuscript, and R.F. Yeh, H. Houbaviy, and G. Ruvkun for advice and helpful discussions. Supported by grants from the NIH and the David H. Koch Cancer Research Fund (D.P.B) and a grant from the NIH (C.B.B.).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

Ambros, V., Bartel, B., Bartel, D.P., Burge, C.B., Carrington, J.C., Chen, X., Dreyfuss, G., Eddy, S., Griffiths-Jones, S., Matzke, M., et al. 2003. A uniform system for microRNA annotation. *RNA* **9:** 277–279.

Aravin, A.A., Naumova, N.M., Tulin, A.A., Rozovsky, Y.M., and Gvozdev, V.A. 2001. Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in *D. melanogaster* germline. *Curr. Biol.* **11:** 1017–1027.

Argaman, L.,Hershberg, R., Vogel, J., Bejerano, G., Wagner, E.G., Margalit, H., and Altuvia, S. 2001. Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol.* **11:** 941–950.

Beanan, M.J. and Strome, S. 1992. Characterization of a germline proliferation mutation in *C. elegans*. *Development* **116:** 755–766.

Bernstein, E., Caudy, A.A., Hammond, S.M., and Hannon, G.J. 2001. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409:** 295–296.

Broverman, S.A. and Meneely, P.M. 1994. Meiotic mutants that cause a polar decrease in recombination on the X chromosome in *Caenorhabditis elegans*. *Genetics* **136:** 119–127.

Burge, C.B. and Karlin, S. 1998. Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8:** 346–354.

Caudy, A.A., Myers, M., Hannon, G.J., and Hammond, S.M. 2002. Fragile X-related protein and VIG associate with the RNA interference machinery. *Genes & Dev.* **16:** 2491–2496.

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282:** 2012–2018.

Chervitz, S.A., Aravind, L., Sherlock, G., Ball, C.A., Koonin, E.V., Dwight, S.S., Harris, M.A., Dolinski, K., Mohr, S., Smith, T., et al. 1998. Comparison of the complete protein sets of worm and yeast: Orthology and divergence. *Science* **282:** 2022–2028.

Djikeng, A., Shi, H., Tschudi, C., and Ullu, E. 2001. RNA interference in *Trypanosoma brucei*: Cloning of small interfering RNAs provides evidence for retroposon-derived 24–26-nucleotide RNAs. *RNA* **7:** 1522–1530.

Elbashir, S.M., Lendeckel, W., and Tuschl, T. 2001a. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes & Dev.* **15:** 188–200.

Elbashir, S.M., Martinez, J., Patkaniowska, A., Lendeckel, W., and Tuschl, T. 2001b. Functional anatomy of siRNAs for mediating efficient RNAi in *Drosophila melanogaster* embryo lysate. *EMBO J.* **20:** 6877–6888.

Grishok, A., Pasquinelli, A.E, Conte, D., Li, N., Parrish, S., Ha, I., Baillie, D.L., Fire, A., Ruvkun, G., and Mello, C.C. 2001. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. *Cell* **106:** 23–34.

Ha, I., Wightman, B., and Ruvkun, G. 1996. A bulged lin-4/lin-14 RNA duplex is sufficient for *Caenorhabditis elegans* lin-14 temporal gradient formation. *Genes & Dev.* **10:** 3041–3050.

Hall, I.M., Shankaranarayana, G.D., Noma, K., Ayoub, N., Cohen, A., and Grewal, S.I. 2002. Establishment and maintenance of a heterochromatin domain. *Science* **297:** 2232–2237.

Hamilton, A.J. and Baulcombe, D.C. 1999. A novel species of small antisense RNA in posttranscriptional gene silencing. *Science* **286:** 950–952.

Hamilton, A., Voinnet, O., Chappell, L., and Baulcombe, D. 2002. Two classes of short interfering RNA in RNA silencing. *EMBO J.* **21:** 4671–4679.

Hammond, S.C., Bernstein, E., Beach, D., and Hannon, G.J. 2000. An RNA-directed nuclease mediates posttranscriptional gene silencing in *Drosophila* cells. *Nature* **404:** 293–296.

Hastie, N.D. and Bishop, J.O. 1976. The expression of three abundance classes of messenger RNA in mouse tissues. *Cell* **9:** 761–774.

Hofacker, I.L. and Stadler, P.F. 1999. Automatic detection of conserved base pairing patterns in RNA virus genomes. *Comput. Chem.* **15:** 401–414.

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., and Schuster, P. 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte f. Chemie* **125:** 167–188.

Hofacker, I.L., Fekete, M., Flamm, C., Huynen, M.A., Rauscher, S., Stolorz, P.E., and Stadler, P.F. 1998. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucleic Acids Res.* **26:** 3825–3836.

Hutvágner, G. and Zamore, P.D. 2002. A microRNA in a multiple-turnover RNAi enzyme complex. *Science* **297:** 2056–2060.

Hutvágner, G., McLachlan, J., Pasquinelli, A.E., Balint, E., Tuschl, T., and Zamore, P.D. 2001. A cellular function for the RNA-interference enzyme Dicer in the maturation of the *let-7* small temporal RNA. *Science* **293:** 834–838.

Ishizuka, A., Siomi, M.C., and Siomi, H. 2002. A *Drosophila* fragile X protein interacts with components of RNAi and ribosomal proteins. *Genes & Dev.* **16:** 2497–2508.

Ketting, R.F., Fischer, S.E.J., Bernstein, E., Sijen, T., Hannon, G.J., and Plasterk, R.H.A. 2001. Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C. elegans*. *Genes & Dev.* **15:** 2654–2659.

Klahre, U., Crete, P., Leuenberger, S.A, Iglesias, V.A., and Meins, F. 2002. High molecular weight RNAs and small interfering RNAs induce systemic posttranscriptional gene silencing in plants. *Proc. Natl. Acad. Sci.* **99:** 11981–11986.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. 2001. Identification of novel genes coding for small expressed RNAs. *Science* **294:** 853–858.

Lagos-Quintana, M., Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W., and Tuschl, T. 2002. Identification of tissue-specific microRNAs from mouse. *Curr. Biol.* **12:** 735–739.

Lagos-Quintana, M., Rauhut, R., Meyer, J., Borkhardt, A., and Tuschl, T. 2003. New microRNAs from mouse and human. *RNA* **9:** 175–179.

Lai, E.C. 2002. MicroRNAs are complementary to 3′UTR motifs that mediate negative post-transcriptional regulation. *Nat. Genet.* **30:** 363–364.

Lander E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitz Hugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294:** 858–862.

Lee, R.C. and Ambros, V. 2001. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* **294:** 862–864.

Lee, R.C., Feinbaum, R.L., and Ambros, V. 1993. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **75:** 843–854.

Lee, Y., Jeon, K., Lee, J.T., Kim, S., and Kim, V.N. 2002. MicroRNA maturation: Stepwise processing and subcellular localization. *EMBO J.* **21:** 4663–4670.

Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B., and Bartel, D.P. 2003. Vertebrate microRNA genes. *Science* **299:** 1540.

Llave, C., Kasschau, K.D., Rector, M.A., and Carrington, J.C. 2002a. Endogenous and silencing-associated small RNAs in plants. *Plant Cell* **14:** 1605–1619.

Llave, C., Xie, Z., Kasschau, K.D., and Carrington, J.C. 2002b. Cleavage of Scarecrow-like mRNA targets directed by a class of *Arabidopsis* miRNA. *Science* **297:** 2053–2056.

Lowe, T.M. and Eddy, S.R. 1997. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25:** 955–964.

———. 1999. A computational screen for methylation guide snoRNAs in yeast. *Science* **283:** 1168–1171.

Martinez, J., Patkaniowska, A., Urlaub, H., Luhrmann, R., and Tuschl, T. 2002. Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. *Cell* **110:** 563–574.

McCaskill, J.S. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29:** 1105–1119.

Mochizuki, K., Fine, N.A., Fujisawa, T., and Gorovsky, M.A. 2002. Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in Tetrahymena. *Cell* **110:** 689–699.

Moss, E.G., Lee, R.C., and Ambros, V. 1997. The cold shock domain protein LIN-28 controls developmental timing in *C. elegans* and is regulated by the lin-4 RNA. *Cell* **88:** 637–646.

Mourelatos, Z., Dostie, J., Paushkin, S., Sharma, A., Charroux, B., Abel, L., Rappsilber, J., Mann, M., and Dreyfuss, G. 2002. miRNPs: A novel class of ribonucleoproteins containing numerous microRNAs. *Genes & Dev.* **16:** 720–728.

Nykäken, A., Haley, B., and Zamore, P.D. 2001. ATP requirements and small interfering RNA structure in the RNA interference pathway. *Cell* **107:** 309–321.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420:** 563–573.

Olsen, P.H. and Ambros, V. 1999. The *lin-4* regulatory RNA controls developmental timing in *Caenorhabditis elegans* by blocking LIN-14 protein synthesis after the initiation of translation. *Dev. Biol.* **216:** 671–680.

Park, W., Li, J., Song, R., Messing, J., and Chen, X. 2002. CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in *Arabidopsis thaliana*. *Curr. Biol.* **12:** 1484–1495.

Parrish, S., Fleenor, J., Xu, S., Mello, C., and Fire, A. 2000. Functional anatomy of a dsRNA trigger: Differential requirement for the two trigger strands in RNA interference. *Mol. Cell* **6:** 1077–1087.

Pasquinelli, A.E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M., Maller, B., Srinivasan, A., Fishman, M., Hayward, D., Ball, E., et al. 2000. Conservation across animal phylogeny of the sequence and temporal regulation of the 21 nucleotide *let-7* heterochronic regulatory RNA. *Nature* **408:** 86–89.

Pickford, A.S., Catalanotto, C., Cogoni, C., andMacino, G. 2002. Quelling in Neurospora crassa. *Adv. Genet.* **46:** 277–303.

Reinhart, B.J. and Bartel, D.P. 2002. Small RNAs correspond to centromere heterochromatic repeats. *Science* **297:** 1831.

Reinhart, B.J., Slack, F.J., Basson, M., Bettinger, J.C., Pasquinelli, A.E., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G. 2000. The 21 nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* **403:** 901–906.

Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B., and Bartel, D.P. 2002. MicroRNAs in plants. *Genes & Dev.* **16:** 1616–1626.

Rhoades, M.W., Reinhart, B.J., Lim, L.P., Burge, C.B., Bartel, B., and Bartel, D.P. 2002. Prediction of plant microRNA targets. *Cell* **110:** 513–520.

Rivas, E., Klein, R.J., Jones, T.A., and Eddy, S.R. 2001. Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr. Biol.* **11:** 1369–1373.

Schwarz, D.S., Hutvágner, G., Haley, B., and Zamore, P.D. 2002. Evidence that siRNAs function as guides, not primers, in the *Drosophila* and human RNAi pathways. *Mol. Cell* **10:** 537–548.

Slack, F.J., Basson, M., Liu, Z., Ambros, V., Horvitz, H.R., and Ruvkun, G. 2000. The *lin-41* RBCC gene acts in the *C. elegans* heterochronic pathway between the *let-7* regulatory RNA and the LIN-29 transcription factor. *Mol. Cell* **5:** 659–669.

Sugimoto, N., Nakano, S., Katoh, M., Matsumura, A., Nakamuta, H., Ohmichi, T., Yoneyama, M., and Sasaki, M. 1995. Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. *Biochemistry* **34:** 11211–11216.

Sulston, J.E., Schierenberg, E., White, J.G., and Thomson, J.N. 1983. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* **100:** 64–119.

Tang, G., Reinhart, B.J., Bartel, D.P., and Zamore, P.D. 2003. A biochemical framework for RNA silencing in plants. *Genes & Dev.* **17:** 49–63.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.

Vance, V. and Vaucheret, H. 2001. RNA silencing in plants: Defense and counterdefense. *Science* **292:** 2277–2280.

Volpe, T., Kidner, C., Hall, I., Teng, G., Grewal, S., and Martienssen, R. 2002. Heterochromatic silencing and histone H3 lysine 9 methylation are regulated by RNA interference. *Science* **297:** 1833–1837.

Wassarman, K.M., Repoila, F., Rosenow, C., Storz, G., and Gottesman, S. 2001. Identification of novel small RNAs using comparative genomics and microarrays. *Genes & Dev.* **15:** 1637–1651.

Wightman, B., Ha, I., and Ruvkun, G. 1993. Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell* **75:** 855–862.

Zamore, P.D., Tuschl, T., Sharp, P.A., and Bartel, D.P. 2000. RNAi: Double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101:** 25–33.

Zilberman, D., Cao, X., and Jacobsen, S.E. 2003. ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. *Science* **299:** 716–719.

Zuker, M. 1994. Prediction of RNA secondary structure by energy minimization. *Methods Mol. Biol.* **25:** 267–294.