A GENETIC SCREEN TO ISOLATE "LARIAT" PEPTIDE INHIBITORS OF PROTEIN FUNCTION

A Thesis Submitted to the College of

Graduate Studies and Research

In Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy

In the Department of Biochemistry

University of Saskatchewan

Saskatoon

By

Kris Michael Barreto

PERMISSION TO USE

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Biochemistry

University of Saskatchewan

Saskatoon, Saskatchewan, S7N 0W0

ABSTRACT

Functional genomic analyses provide information that allows hypotheses to be formulated on protein function. These hypotheses, however, need to be validated using reverse genetic approaches, which are difficult to perform on a large scale and in diploid organisms. To address this problem, we developed a genetic screen to rapidly isolate "lariat" peptides that function as *trans* dominant inhibitors of protein function.

We engineered intein proteins to genetically produce lariats. A lariat consists of a lactone peptide covalently attached to a linear peptide. Cyclizing peptides with a lactone bond imposes a constraint even within the reducing environment found inside of cells. The covalently attached linear peptide provides a site for fusing protein moieties. We fused a transcriptional activation domain to a combinatorial lactone peptide, which allowed combinatorial lariat libraries to be screened for protein interactions using the yeast two-hybrid assay.

We confirmed that the intein processed in yeast using Western blot analysis. A chemoselective ring opening of the lactone bond with heavy water, followed by mass spectrometry analysis showed that ~ 44% of purified lariat contained an intact lactone bond. To improve the stability of the lactone bond, we introduced mutations into the engineered intein and analyzed their processing and stability by mass spectrometery. Several mutations were identified that increased the amount of intact lariat.

Combinatorial libraries of lactone peptides were generated and screened using the yeast-two-hybrid interaction trap. Lactone cyclic peptides that bound to a number of different targets including LexA, Jak2, and Riz1 were isolated. A lactone cyclic peptide isolated against the bacterial repressor protein LexA was characterized. LexA regulates bacterial SOS response and LexA mutants that cannot undergo autoproteolyis make bacteria more sensitive to, and inhibit resistance against cytotoxic reagents. The anti-LexA lariat interacted with LexA with a dissociation constant of 37 μM by surface plasmon resonance. The lactone constraint was determined to be required for the interaction of the anti-LexA L2 lariat with LexA in the yeast-two-hybrid assay. Alanine scanning showed that only two amino acids (G8 and E9) in the anti-LexA L2 sequence (1-SRSWDLPGEY-10) were not required for the interaction with LexA. The interaction

of the anti-LexA lariat with LexA *in vivo* was confirmed by chromatin precipitation of the lactone peptide-LexA-DNA complex. The anti-microbial properties of the anti-LexA lariat were also characterized. The anti-LexA lariat potentiated the activity of a DNA damaging agent mitomycin C and inhibited the cleavage of LexA, preventing the SOS response pathway from being activated.

In summary, lariats possess desired traits for characterizing the function and therapeutic potential of proteins. The ability to genetically and chemically synthesize lariats allows the lariat transcription activation domain to be replaced by other peptide and chemical moieties such as affinity tags, fluorescent molecules, localization sequences, *et cetera*, which give them advantages over "head to tail" cyclized peptides, which have no free end to attach moieties.

ACKNOWLEDGEMENTS

I would like to start by thanking my supervisor Dr. Ron Geyer. I appreciate the seriousness in which he has taken on his role in guiding and mentoring me through my entire Ph.D. I was his first graduate student and I appreciate the freedom he gave me in the lab to pursue many projects, and learn as many techniques as possible. I enjoyed how he treated me as an equal and listened to my many ideas, and I enjoyed hearing his ideas and his passion for science even when nothing was working. I can only hope I am able to inspire others to be as excited about science as he did for me.

I would like to thank my committee members, Dr. Roesler, Dr. Anderson, Dr. Harkness, Dr. Hong, Dr. Khandelwal for their guidance and support of my project.

I would like to thank all the past and present members of the Geyer lab, particularly Annette Kerviche who taught me how to sequence DNA, and Karen Mochoruk who put up with me and spent hours troubleshooting the mass spectrometer with me, and whose experience with HPLC was invaluable. I would also like to thank Mark Boyd of the cancer center who taught me how to use the flow cytometer.

Finally I would like to thank Wendy Bernhard, for helping me get through it all, making me smile everyday, and getting me in for the morning shift. I am so glad I met you and will always love you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Definition
ADE2	Phosphoribosylamino-imidazole-carboxylase
ADH1	Alcohol dehydrogenase promoter
amp	Ampicillin
bp	Base pairs
CD4	Cluster of differentiation 4
CDK2	Cyclin dependent kinase 2
CDR	Complementarity determining regions
CFU	Colony forming unit
ChIP	Chromatin immunoprecipitation
CTLA-4	Cytotoxic T-Lymphocyte Antigen 4
DARP	Domain ankyrin repeat protein
ddH_2O	Filtered and autoclaved water
DRP	Domain repeat protein
dsRNA	Double stranded RNA
DTT	1,4-dithiothreitol
E. coli	*Escherichia coli*
EDTA	Ethylenediaminetetraacetic acid
EGFR	Epidermal growth factor receptor
ELISA	Enzyme-linked immunosorbent assay
EMSA	Electrophoretic mobility Shift Assay
ESI	Electrospray ionization
Extein	External protein
Fab	Fragment responsible for antigen binding
FACS	Fluorescent activated cell sorting
FRET	Fluorescence resonance energy transfer
Fv	Antibody fragment
GFP	Green fluorescent protein
GST	Glutathione S-transferase
HA	Haemagglutinin tag
HIV-1	Human immunodeficiency virus
I_C	Intein C-terminal domain
Ig	Immunoglobulin
IHC	Immunohistochemistry
IL-2	Interleukin-2
I_N	Intein N-terminal domain
Intein	Internal protein
IP	Immunoprecipitation
IPTG	Isopropyl β-D-1-thiogalactopyranoside
Kan	Kanamycin
kDa	Kilo Dalton
LacZ	Beta-galactosidase

Abbreviation	Definition
LC	Liquid chromatography
LEU2	Beta-isopropylmalate dehydrogenase
m/z	Mass over charge ratio
MENSA	2-mercaptoethanesulfonic acid
miRNA	Micro-RNA
MS	Mass spectrometry
MW	Molecular weight
NEMO	NF-kappa-beta essential modulator
NLS	Nuclear localization signal
NMR	Nuclear magnetic resonance
Npu	*Nostoc punctiforme*
Nt	Nucleotide
NTA	Nitrilotriacetic acid
qPCR	Quantitative PCR
RNAi	RNA interference
S. cerevisiae	*Saccharomyces cerevisiae*
scFv	Single chain variable fragment
siRNA	small interfering RNA
SPR	Surface plasmon resonance
Ssp	*Synechocystis species*
STM	SteA triple mutant
TIC	Total Ion Count
TOF	Time-of-flight
TrxA	Thioredoxin A
VEGF	Vascular endothelial growth factor
Y2H	Yeast-two-hybrid

1 Literature review

1.1 Introduction

Many research and diagnostic applications exploit the ability of antibodies to bind their target antigen with high affinity and specificity. Antibodies have become standard laboratory reagents for a variety of applications including: (i) Western blot analysis, which is used to study protein expression, stability, and modifications; (ii) immunoprecipitation (IP), which is used to characterize protein complexes and protein activities; (iii) chromatin immunoprecipitation (ChIP) and electrophoretic mobility shift assays (EMSA), which are used to analyze protein-nucleic acid interactions; (iv) fluorescent-activated cell sorting (FACS), which is used to analyze, study, and sort cell populations; (v) immunohistochemistry (IHC), which is used to visualize antigen localization and distribution within cells and tissues; and (vi) enzyme-linked immunosorbent assay (ELISA), which is used to detect antigen in a sample. All of these technologies have been successfully used in basic research and diagnostic applications.

The success of antibodies in research and diagnostics has led to the development of antibody-based drugs. The U.S. Food and Drug Administration (FDA) has approved twenty-two monoclonal antibodies for therapy (Reichert, 2008; Aires da Silva *et al.*, 2008), including Bevacizumab (Avastin®, Genetech), Rituximab (Rituxan®, Genetech), and Trastuzumab (Herceptin®, Genentech). These drugs have reached 'blockbuster' status with sales of approximately $2.7, $2.6, and $1.4 billion, respectively in the U.S.A in 2008 (http://www.gene.com/gene/ir/financials/historical).

Despite their success, antibodies have a number of significant drawbacks. In general antibodies are difficult to isolate and synthesize. Antibodies are large (150 kDa) proteins that undergo post-translational modifications, specifically glycosylation. Therefore, antibodies cannot be chemically synthesized and must be produced using biological systems, which increases their cost. In addition, antibodies evolved to function outside of cells and are stabilized by disulfide bonds that do not form correctly in the reducing environment within cells (Wörn and Plückthun, 2001). Further, their large size limits oral availability and reduces tissue and cell penetration, which reduces efficacy (Jain, 1990; Yokota *et al.*, 1992). Finally,

antibody-based drugs are immunogenic and have short half-lives (Khazaeli *et al.*, 1994). As a result of these limitations, antibody-based drugs have only been used to inhibit extracellular targets and are generally delivered intravenously. This currently limits their usage to severe diseases where frequent hospital visits are acceptable.

The proven success of antibody-based therapeutics in the clinic and in basic research is driving research to develop methods to overcome these limitations. For example, strategies have been developed to reduce immunogenicity and increase half-lives of antibodies (Boulianne *et al.*, 1984; Morrison *et al.*, 1984; Jones *et al.*, 1986; Presta, 2006). Strategies have also been developed to lower the cost of production and to improve the stability of antibodies by engineering antibody fragments that contain only the antigen-binding fragment (Fv) (Bird *et al.*, 1988; Huston *et al.*, 1988; Visintin *et al.*, 1999; Visintin *et al.*, 2002). Relative to antibodies, Fvs have several advantages. First, some Fvs have the potential to be used as intracellular therapeutics, since many antibody fragments have been engineered to fold in the absence of disulfide bonds (Wörn and Plückthun, 1998; Tanaka and Rabbitts, 2008). Second, their small size increases their bioavailability, making them easier to produce and more likely to penetrate tissues. However, the majority of Fvs are not stable and a universal solution to stabilize all Fvs has still not been discovered. Substantial hurdles need to be overcome before Fvs will be readily available for a variety of intracellular and extracellular applications. The proven success of full-length antibodies in extracellular applications and the unrealized potential of Fvs in intracellular applications has generated considerable interest in alternative peptide affinity reagents that function like antibodies but do not share their limitations.

Currently, the majority of drugs are small molecules. This is in part because decades of research have created standard procedures to optimize small molecules for bioavailability and efficacy. The difficulties associated with oral delivery of peptide-based therapeutics have made them unpopular. However, antibodies have several significant advantages over small molecules that make them desirable for therapeutic applications. First, antibodies have the ability to bind a wide variety of targets. Second, current antibody isolation technologies make it relatively easy to generate antibodies against many targets. These properties combined with the fact that diseases are now being characterized at the molecular level and specific drug targets are identified makes antibodies extremely desirable therapeutic reagents.

Cyclic peptides are a class of peptide affinity reagents that are of particular interest as therapeutics. Cyclic peptides are small, stable, and can be synthesized using solid-phase peptide chemistry, which reduces their cost. Unlike antibodies, cyclic peptides can function inside of cells and some are capable of translocating across cell membranes (Fletcher and Hughes, 2009; D'Ursi *et al.*, 2006). Cyclic peptides can also be imported into cells using permeation tags (Heitz *et al.*, 2009; Derossi *et al.*, 1996) and liposomes to increase their bioavailability. In addition, even if a cyclic peptide cannot be directly used as a therapeutic reagent, it can be used as a small molecule lead. The small size of cyclic peptides allows their structures to be easily solved using nuclear magnetic resonance (NMR), X-ray crystallography, and *in silico* modeling techniques (Yongye *et al.*, 2009; Maupetit *et al.*, 2009). Crystal or NMR structures of peptide-protein complexes or *in silico* predicted complexes can be used to identify parts of the cyclic peptide that interact with the target. Sites on the cyclic peptide required for the interaction with its target are called "pharmacores". Pharmacores from a cyclic peptide can be grafted onto a small molecule using peptidomimetics (Vagner *et al.*, 2008), creating a small molecule with the same pharmaceutical activity as the cyclic peptide. The resulting small molecules can then be optimized for bioavailability using existing strategies available for small molecules.

Like antibodies, cyclic peptides are natural cell products. Cyclic peptides are produced by plants (Saether *et al.*, 1995), bacteria (Trabi and Craik, 2002), and fungi (Cheng and Walton, 2000) where they function as antimicrobials, toxins, and insecticides (Jennings *et al.*, 2001). Cyclic peptides are also produced in mammals. The cyclic peptide theta-defensin, which is present in the rhesus monkey, defends against HIV (Tang *et al.*, 1999). In humans, the theta-defensin cyclic peptide gene is inactivated by a point mutation, which introduces a premature stop codon and prevents translation (Venkataraman *et al.*, 2009). Repairing this stop codon results in the production of the cyclic peptide, indicating that a general mechanism to produce cyclic peptides in humans is still active (Venkataraman *et al.*, 2009). Since cyclic peptides are ubiquitous, natural defense agents they are good candidates for therapeutic reagents.

One advantage that antibodies have that cyclic peptides lack is the ability to easily select antibodies against a given target using reverse genetic screens. Many cyclic peptides are non-ribosomally produced and generally contain non-standard amino acids. Mechanisms used to

synthesize non-ribosomally produced cyclic peptides are gradually being deciphered, however methods to genetically encode libraries of non-ribosomally produced cyclic peptides have still not been reported. Without a genetically tractable way to make libraries of cyclic peptides, it is difficult to use genetic screens to isolate cyclic peptides that interact with a specific target. This is the major bottleneck in isolating cyclic peptides that inhibit a specific process or target.

Libraries of biologically derived or chemically synthesized cyclic peptides have been screened for a given phenotype using activity-based assays (reviewed by Baldwin, 1996; Edwards and Morrell, 2002; Huwe, 2006). The number of cyclic peptides that can be screened using these types of screens is limited by the number of assays that can be performed, which is typically around 10^3 to 10^5 (Tan, 2005; Webb, 2005). Furthermore, the target or mode of action of a cyclic peptide isolated from a phenotype-based screen cannot be directly determined. Biologically derived extracts containing cyclic peptides pose an additional problem since they are composed of complicated mixtures of peptides, proteins, and other molecules. Identifying the bioactive compound in the mixture can be a time consuming and difficult process (Escoubas *et al.*, 2008). As a result, there is a need for technologies that allow the selection of cyclic peptides using reverse genetic selection techniques.

1.2 Deciphering molecular interaction/functional networks

Peptide affinity reagents were originally designed as dominant agents to decipher molecular interaction and functional networks. The molecular interactions and functional modifications between biomolecules form complex networks that allow cells to respond appropriately to their environment, which is required for proper cell function (Pawson and Nash, 2000). Elucidating these interaction/functional networks is fundamental in understanding cellular processes and the biological basis of diseases.

Technological advances in DNA sequencing, protein and RNA expression profiling, and molecular interaction detection have generated large data sets that enable putative networks to be constructed. New DNA sequencing technologies have provided whole genome sequences for numerous organisms. Unknown gene sequences can be compared to known genes and functional assignments can be made based on sequence homology. Gene interactions can also be determined when genomic sequence data is combined with expression profiling (Pellegrini *et al.*, 1999). Expression profiling using microarrays that detect mRNA, proteins, or small molecules allows interactions to be predicted. Molecules whose expression profiles change in

response to the same stimuli allows one to infer using "guilt by association" logic that these molecules are involved in the same pathway and are connected (Pellegrini *et al.*, 1999). Large-scale protein interaction studies using technologies such as the yeast-two-hybrid (Y2H) assay have been used to elucidate whole organism protein-protein interaction networks (Li *et al.*, 2004; Rual *et al.*, 2005; Yu *et al.*, 2008; Makhnevych *et al.*, 2009; Schwartz *et al.*, 2009; Venkatesan *et al.*, 2009). Together, the information generated with these technologies can be used to construct interaction/functional networks. To verify these interaction/functional networks they need to be validated using directed experimentation.

Traditionally, strategies to validate protein function relied on reverse genetic approaches that delete the gene of interest at the DNA level, referred to as gene "knock-outs". These approaches are time consuming and difficult to perform in polyploid organisms, especially in a high-throughput manner. Alternatively, protein expression can be blocked by siRNA, which degrades RNA or by miRNA, which inhibits protein translation (reviewed in Carthew and Sontheimer, 2009). Methods that block protein expression at the mRNA level function in *trans*, making it easier to block protein expression in polyploid organisms where there are multiple gene targets. Thus, siRNA/miRNA are extremely powerful techniques for the reverse analysis of protein function in polyploid organisms.

However, there are limitations to blocking protein expression at the mRNA level. First, it is difficult to deliver siRNA or miRNA into cells. Second, these technologies rarely remove the target protein completely. Third, proteins with long half-lives are not good targets for siRNA or miRNA since even if the mRNA is degraded or protein translation is stalled, there is little or no turn-over of the protein that is already present. Finally, siRNA and miRNA off-target effects must be considered. Double stranded siRNA or miRNA designed to target the sense stand of a gene of interest may mis-target a natural anti-sense transcript (Rinn *et al.*, 2007; Faghihi and Wahlestedt, 2009). Partial matches between siRNA and 3`-untranslated regions of mRNA can lead to miRNA-mediated repression of nontargeted proteins, or can lead to the down regulation of alternatively spliced protein products generated from the same transcript that share much of the same sequence (Scacheri *et al.*, 2004; Jackson *et al.*, 2003; Birmingham *et al.*, 2006). Expression of siRNA or miRNA to target a gene of interest can also result in artifacts due to the overloading of the siRNA or miRNA machinery. Exogenous siRNA and miRNA compete with natural endogenous siRNAs and miRNAs for binding siRNA

and miRNA machinery, resulting in off-target effects (Baek *et al.*, 2008; Selbach *et al.*, 2008; Castanotto *et al.*, 2007). siRNA and miRNA can also produce misleading results since they can activate pathways in innate immune responses (Agrawal and Kandimalla, 2004; Hornung *et al.*, 2005; Judge *et al.*, 2005; Heinicke *et al.*, 2009).

In summary, traditional reverse approaches that block gene function at the DNA level are difficult to perform and the limits of these technologies that either degrade mRNA (siRNA) or prevent protein translation (miRNA) are not well defined. In either case, these methods block the expression of the target protein and thus they may not always correlate with the ability of the target to be inhibited by small molecules.

The difference between small molecule inhibitors and inhibition strategies that block protein expression can be illustrated using a hypothetical interaction network (Figure 1.1). In this network, nodes, represent biochemical molecules such as proteins. Vertices, represent actions such as modifications, enzymatic activities, or physical interactions. Reverse genetic strategies that block protein expression result in the removal of a node from an interaction network (Figure 1.1b). If a node has more than one vertex, then multiple pathways are affected (Figure 1.1b).

Vertices can be disrupted by inhibiting protein activities or interactions (Figure 1.1c). Point mutations are particularly useful for validating the therapeutic potential of a target protein since they have the ability to mimic the effect of a small molecule by knocking out a single function rather than removing the whole protein. In this way, point mutations often disrupt only a single vertex in an interaction network (Figure 1.1c). Homozygous point mutations however are difficult to generate in polyploidy organisms and usually arise as a result of natural disease states (Suzuki *et al.*, 1971). Furthermore, point mutations can result in the misfolding or degradation of proteins.

Peptide inhibitors can be used to block a single vertex in an interaction/functional network in a manner similar to a small molecule inhibitor or point mutation (Figure 1.1c). Peptide inhibitors function in *trans* with a dominant mode of action and are capable of interacting with a defined surface of the target. Peptide affinity reagents have the following advantages over miRNA/siRNA technologies and gene deletions. First, dosage effects of peptide inhibitors can be easily tested. Second, peptide affinity reagents can modify target

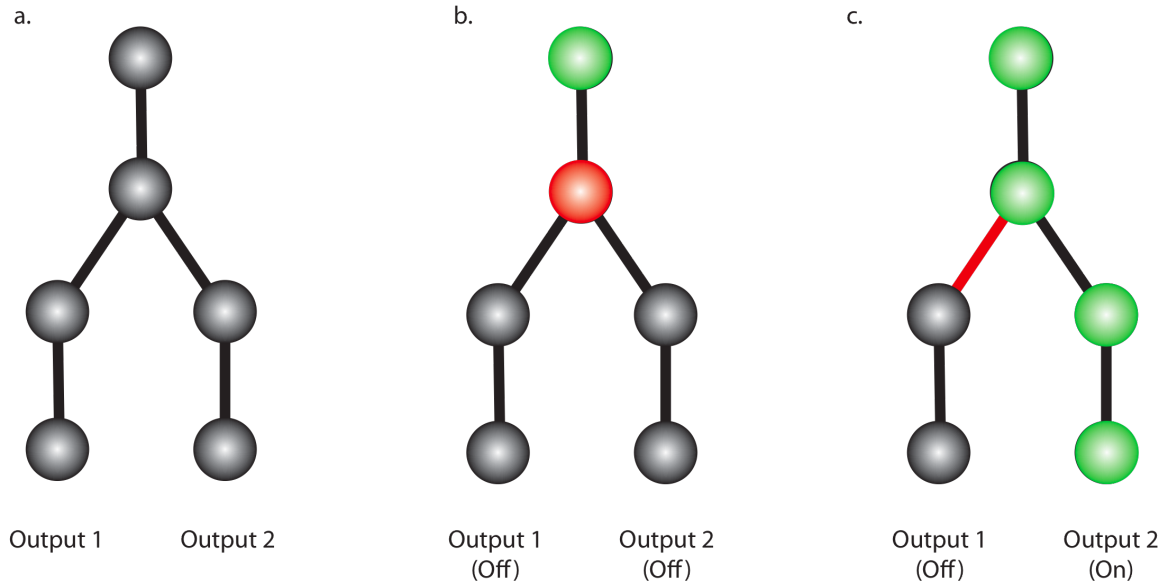


Figure 1.1 | Molecular interaction/functional networks
(a) Hypothetical interaction/functional network. Nodes (circles), represent biological molecules such as protein, RNA, DNA, *etc*. Vertices (lines), represent actions such as a physical interaction, modification, *etc*. Grey nodes represent unactivated biomolecules. **(b)** Effect of protein deletion on network output. A protein deletion (red node) either by gene 'knock-out' techniques or by a siRNA or miRNA prevents translation of the protein and eliminates a node and thus all protein functions (vertices), which prevents downstream nodes from being activated (grey nodes). Green nodes represent activated proteins. **(c)** Effect of a peptide, small molecule, or point mutation on network output. A peptide affinity reagent, small molecule or point mutations is used to block a specific function or interaction of a protein (red vertex), resulting in only one output being affected.

localization (Geda *et al.*, 2008) and stability (McLean *et al.*, 2009). Third, peptide inhibitors have the potential to target only a specific conformation, allele, or modified state of a protein (Kurtz *et al.*, 2003; Colas *et al.*, 1996). Fourth, peptide inhibitors can be designed against metabolites and other non-protein targets of interest, which are often considered "undruggable" by small molecules (Skerra, 2008).

The difference between removing a protein from a system and hence all its corresponding interactions/functions can be very different from blocking a single specific interaction/function. Abed *et al.* demonstrated this principle by showing that removing a protein using gene knock-out techniques has a greater impact on mRNA expression profiles than a dominant-negative mutation or a peptide affinity reagent (Abed *et al.*, 2007). Microarray studies comparing the effects of siRNA and small molecule inhibitors of BCR-ABL

showed that mRNA profiles between the siRNA and small molecule are substantially different (Zhelev *et al.*, 2004).

1.3 Design of peptide affinity reagent libraries

Peptide affinity reagents can either be rationally designed or selected from combinatorial pools of peptides. The majority of peptide affinity reagents have been generated against protein targets. This thesis focuses on protein targets but the concepts can be applied to other molecular targets. Rationally designed peptide affinity reagents are designed using structural information of known protein-protein interactions as a guide to design peptide affinity reagents. This information is used to design a peptide that mimics the interaction or alternatively that will bind to the target based on the structural information. Peptide affinity reagents can also be selected from large combinatorial pools called "libraries". Libraries contain many different peptides, where each peptide has specific positions in the peptide where random amino acids are substituted. In theory, if a random pool of peptides is complex enough, there should be one member that interacts with any given target.

Proteins contain many different surfaces and each surface can have a different function. For example, a protein surface can have enzymatic activity, or it can be involved in a protein-protein interaction, in a protein-molecule interaction, or in maintaining the protein's structure. Many of these surfaces are dynamic and change in response to modifications or interactions with other molecules. Combinatorial peptide libraries have been used to isolate peptides that bind to specific states of a protein (Pamonsinlapatham *et al.*, 2008; Davis *et al.*, 2009). This is an extremely powerful property of peptides and in theory an optimal combinatorial peptide library will contain members that bind to each surface and state of a protein. In this way, specific protein functions or conformations can be blocked while leaving all other functions unaffected.

To facilitate the isolation of peptides against a specific protein domain or surface, proteins can be split up and expressed as individual domains. This ensures a peptide will be generated against a specific domain and not another region of the protein. In some cases this is necessary since the domain of interest is inaccessible in the full-length protein due to the conformation of the protein. Domain specific peptides can be used to study the function/interaction of a specific domain while leaving other protein functions unaffected or to design specific inhibitors.

In practice a single type of peptide library is not likely to be capable of specifically disrupting every possible protein interaction, since protein-protein interaction surfaces vary from a few strong interactions in a small highly specific binding pocket to many weak interactions spread over a large surface areas. Multiple different peptide libraries may be required to develop inhibitors against a wide range of protein surfaces.

Antibodies are an example of a naturally occurring, genetically encoded peptide affinity reagent library. Lymphocytes in the human immune system have evolved a sophisticated method for generating molecular diversity in antibodies using recombination processes and somatic mutation. Multiple genes are used to encode different antibody variable and constant regions for both heavy and light chains. Each cell undergoes somatic recombination to produce a functional heavy and light chain. Somatic recombination is a process of imprecise joining of genes, which produces antibodies in naïve B-cells. Antibodies are then tested for their interaction with the target antigen inside of the human body. Those cells that express an antibody capable of binding the target antigen weakly have their affinity enhanced by somatic hypermutation, which involves the introduction of point mutations preferentially into the complementarity determining regions (CDRs). This strategy allows the human immune system to generate over 10^{11} different antibody combinations using relatively little DNA.

Synthetically designed genetically encoded peptide libraries typically have a 1:1 relationship of DNA to library members, where each DNA sequence encodes a single library member. There are four DNA nucleotides [A, C, T, G] and each amino acid is specified by a three-nucleotide codon, resulting in a total of 64 (4^3) possible codons. These codons represent three stop codons and 20 amino acids. Random peptide libraries are made by synthesizing oligonucleotides that contain random combinations of nucleotides at given positions. Random coupling of degenerate nucleotides can be used to make randomized codons that express combinatorial libraries of peptides. A completely random oligonucleotide synthesized with single nucleotides uses NNN at each codon position, where N = A, G, C, or T. A peptide library, which is ten amino acids long and constructed using NNN codons, will have a nucleic acid diversity of over 1×10^{18} and a peptide diversity of 1×10^{13}. These libraries are larger than the maximum screening size of most cell-based systems ($\sim 10^9$). Therefore, it is not feasible to generate and screen libraries containing all possible amino acids with lengths greater than seven or eight amino acids. Reducing the number of amino acids can drastically increase the

coverage of a library. Decreasing the amino acid diversity at each position in the peptide can also increase library coverage. This strategy has been used to generate and screen antibody fragment libraries with only two amino acids (tyrosine and serine) in the CDRs (Koide and Sidhu, 2009). Alternatively, libraries can be constructed by joining trinucleotide phosphoramidite repeats together during automated oligonucleotide synthesis (Sondek and Shortle, 1992; Kayushin *et al.*, 1996; Gaytán *et al.*, 1998; Yagodkin *et al.*, 2007; Mauriala *et al.*, 2004). Each trinucleotide phosphoramidite represents a single codon. By mixing trinucleotide phosphoramidites, the frequency of each codon and thus amino acid in the library can be directly controlled. Furthermore, stop codons can be eliminated from oligonucleotide libraries constructed with trinucleotide phosphoramidites. The main disadvantage is that these libraries are expensive to synthesize.

1.4 Selection of peptide affinity reagents

Peptide affinity reagents can be selected in both intra- and extra-cellular environments from combinatorial libraries. Assays that detect protein-protein, protein-RNA, and protein-DNA interactions within cells include: the split-ubiquitin system (Stagljar *et al.*, 1998), the protein-fragment complementation assay (Remy and Michnick, 1999), the repressor reconstitution assay (Hirst *et al.*, 2001), the SOS recruitment system (Broder *et al.*, 1998), and two-hybrid systems (reviewed in Vidal and Legrain, 1999). These assays occur inside the cell and detect protein-protein or protein-DNA interactions using reporter genes/proteins. In these assays, each cell acts as a test-tube to detect the interaction of one member of a library with a target. This type of assay is useful since there is no competition between peptide affinity reagents, resulting in the isolation of peptide affinity reagents that target many different surfaces in a single screen.

Assays that isolate peptide affinity reagents against a target outside of the cell couple DNA to the peptide affinity reagent directly or indirectly. Examples of this type of assay include: phage display (Smith, 1985), bacterial display (Francisco *et al.*, 1993), yeast display (Boder and Wittrup, 1997), ribosome display (Jermutus *et al.*, 1998), and mRNA display (Jermutus *et al.*, 1998). These assays typically include an amplification step, which enriches peptide affinity reagents with the highest binding affinity. The advantages of these systems are that high affinity binders are isolated and rounds of negative selection can be included. This allows the selection of peptide affinity reagents capable of discriminating between two

molecules. These selections, unlike intracellular screens, are not limited to biomolecules. Furthermore, peptide affinity reagents can be isolated that discriminate between complex mixtures such as two different cell types, without prior knowledge of what makes the two cell types different. Any of these assays or similar assays that couple DNA or RNA to its expressed protein can be used to isolate peptide affinity reagents.

1.4.1 Forward screens

In the context of this thesis, a forward screen will be used to describe peptide affinity reagents that are isolated based on their ability to confer the phenotype under investigation. The main advantage of a forward screen is that a variety of molecules can be used including small molecules, peptides, and nucleic acids. There is no requirement for the molecule used in a forward screen to be genetically encoded. For example, chemically synthesized cyclic peptides have been previously used in forward screens (Baldwin, 1996; Edwards and Morrell, 2002; Huwe, 2006; Tan, 2005; Webb, 2005).

Forward screens have several drawbacks. First, there is minimal target specificity; any disruption in the pathway, which gives the desired phenotype, is selected. This lack of target specificity can be an advantage when working with an unknown pathway or when trying to identify new molecules in a pathway. However, the lack of specificity is generally a disadvantage since therapeutic targets are usually identified by studying the disease and then molecules that inhibit these specific targets are sought out. Second, it is difficult to identify the target that is inhibited to cause the desired phenotype. Third, forward screens may result in inhibitors against primarily one target due to its role or accessibility in the pathway. Fourth, targets that do not have an easily observed phenotype are not likely to be identified. Fifth, forward screens are more prone to false positives since pre-existing and spontaneous mutations that overcome the phenotype can also be selected. To overcome these disadvantages and to isolate inhibitors against a protein of interest, reverse screens are often used first followed by a forward screen to select for molecules with desired activity.

1.4.2 Reverse screens

In this thesis, reverse screens are defined as assays that use the physical interaction of a peptide affinity reagent with its target as the basis for their selection. Reverse screens have several advantages over forward screens. Reverse screens do not need to be modified to isolate

peptide affinity reagents against most protein targets. In contrast, forward screens target a process or pathway that produces a phenotype, and thus a different screen must be designed to isolate peptide affinity reagents for targets in different pathways. Reverse screens also allow protein domains, modifications, alleles, or states to be specifically targeted.

Disadvantages of reverse screens are that the target must be known, expressed, and fold properly in the system used for screening. For *in vivo* systems, the target must also be genetically encoded. Screens that are based on the physical interaction of the peptide affinity reagent with the target rather than a phenotype-based screen may not isolate peptide affinity reagents that have an inhibitory function. However, peptide affinity reagents selected based soley on a physical interaction with the target can subsequently be tested for the desired activity. Compared to a forward screen this strategy reduces the number of peptides that need to be tested for activity and ensures the activity will be through the target of interest. Ideally, reverse screens are used to isolate potentially therapeutic molecules against a target of interest or as general affinity reagents for basic research.

1.4.2.1 Two-hybrid assays

Two-hybrid assays are based on the principle of splitting a functional protein into two separate or hybrid proteins. Association of the two fusion proteins results in reconstitution of a functional protein and a phenotypic output. Typically, cell growth, cell death, and colorimetric or fluorescence assays are linked to the association of the hybrid proteins. There are many alternative two-hybrid systems that take place inside mammalian cells (Ozawa *et al.*, 2000), bacterial cells (Joung *et al.*, 2000) and yeast cells (Fields and Song, 1989; Gyuris *et al.*, 1993). The Y2H system is the most commonly used two-hybrid reverse screening assay. In this assay, a DNA binding domain is fused to a target protein called the 'bait' and a transcriptional activation domain is fused to a second protein called the 'prey'. When the bait and prey interact, they reconstitute a functional transcription factor and induce expression of reporter genes (Figure 1.2).

The Y2H assay developed by Fields splits the GAL4 transcription factor into a DNA binding domain (bait) and transcription activation domain (prey) (Fields and Song, 1989). The Y2H assay developed by Brent constructs an artificial transcription factor using the *E. coli* LexA protein as the DNA binding domain (bait) and a bacterial protein, referred to as the B42 acid 'blob', as the transcription activation domain (prey) (Gyuris *et al.*, 1993). A variation of

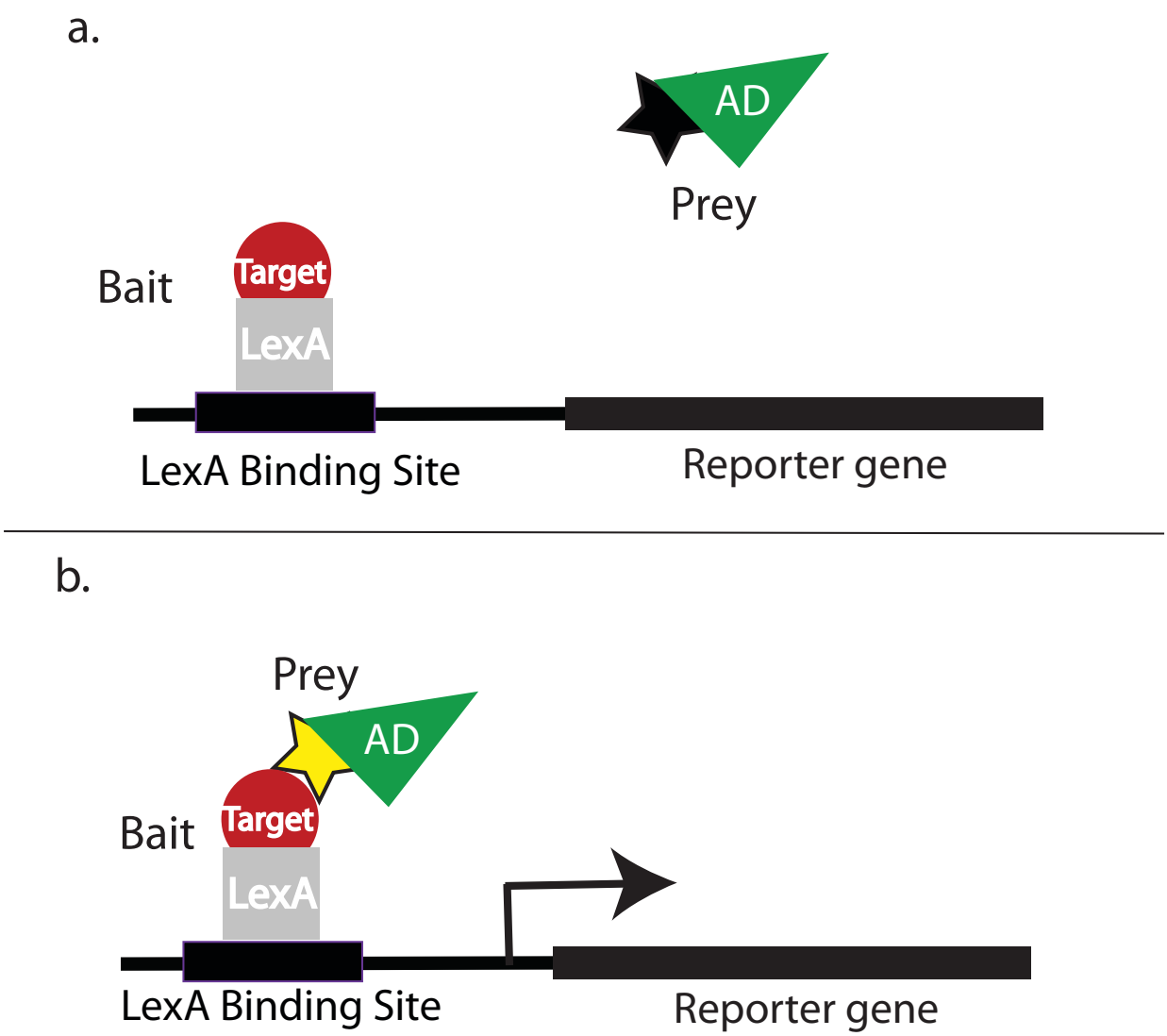


Figure 1.2 | Principle of the Y2H assay
(a) The LexA/target protein fusion (bait) binds to the LexA binding site, which is located at the 5`-end of the reporter gene. The reporter gene is not activated since the peptide affinity reagent (star) fused to the activation domain (AD) (prey) does not interact with the bait. **(b)** Interaction between the bait and prey results in reporter gene activation.

the LexA-based Y2H assay, referred to as the interaction mating assay, has been developed to facilitate the analysis of pairwise interactions and to screen protein/peptide libraries (Finley and Brent, 1994; Golemis *et al.*, 2009). In the mating interaction assay, libraries of yeast cells containing different preys are mated to yeast cells containing the target bait, which results in diploid cells that contains bait, prey, and reporter genes. Using yeast mating to combine the

bait and prey plasmids into the same yeast cell eliminates laborious co-transformation steps. An additional advantage of the interaction mating assay is that a single library of preys can be constructed and then mated to multiple different baits, which eliminates the need to transform the prey library into each different bait strain.

In this thesis, a variation of the LexA-based Y2H assay was used. In the modified Y2H assay the bait plasmid is maintained in one strain containing the Y2H reporters and the prey plasmid library is maintained in another strain under the control of an inducible promoter. The bait and prey strains are mated together on non-selective media to form diploid yeast cells containing reporter genes and bait and prey plasmids (Figure 1.3). Diploid yeast cells containing bait and prey proteins that interact induce expression of reporter genes, which allow the interaction to be visualized by growth and color on selection medium. Three different reporter genes are used in our system: *LEU2*, *ADE2*, and *LACZ*. *LEU2* reporter gene activation is detected by growth on medium lacking leucine. *ADE2* reporter gene activation can be detected by growth and color on medium lacking adenine. The *ADE2* gene encodes a protein responsible for the conversion of 5'-phosphoribosyl-5-aminoimidazole (AIR) to 5'-phosphoribosyl-5-aminoimidazole-4-carboxylateadenine (CAIR) in the adenine biosynthesis pathway. Without the *ADE2* gene, yeast buildup AIR, causing them to turn red. When the *ADE2* reporter gene is activated, yeast cells grow on medium lacking adenine and do not accumulate AIR. The *LACZ* gene, which encodes the enzyme beta-galactosidase, is used as an additional colorimetric reporter. *LACZ* cleaves 5-bromo-4-chloro-3-indolyl-b-D-galactopyranoside (X-gal) into galactose and 5-bromo-4-chloro-3-hydroxyindole. The 5-bromo-4-chloro-3-hydroxyindole is oxidized into 4,4'-dibromo-4,4'-dichloro-indigo, which is insoluble and forms a blue precipitate that is observable in the yeast agar medium.

The Y2H has several advantages and disadvantages for selecting peptide affinity agents from genetically encoded combinatorial peptide libraries. The main advantages of the Y2H system are: (i) there is no competition between peptide affinity reagents in a library since each interaction is tested separately inside a yeast cell; (ii) target proteins are expressed in an intracellular eukaryotic environment; (iii) the bait protein does not need to be purified. The main disadvantages of the Y2H assay are: (i) a smaller number of peptides are screened relative to *in vitro* selection strategies such as phage or mRNA display; (ii) significant false positive and false negative signals are observed in the Y2H assay.

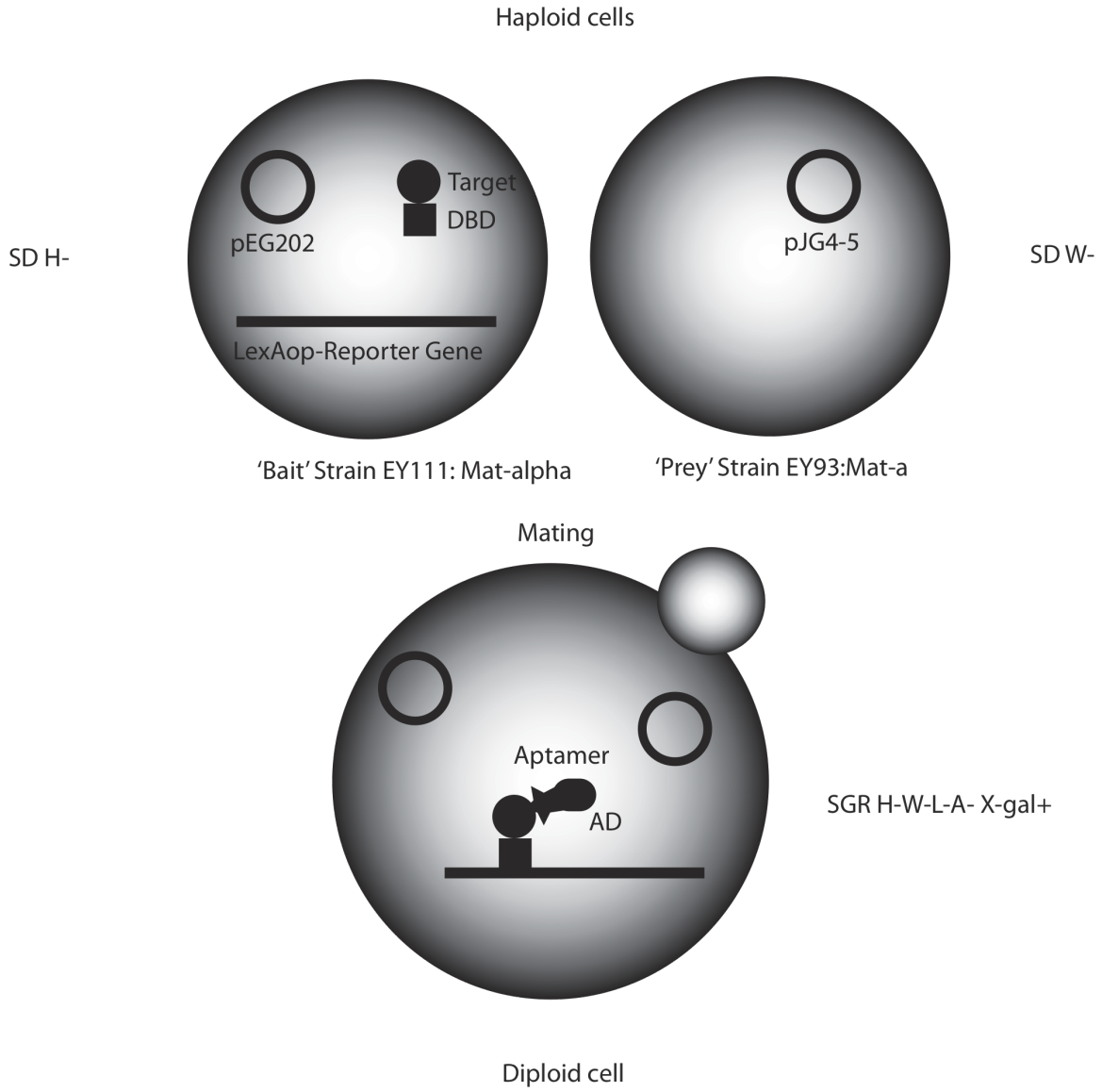


Figure 1.3 | Y2H interaction mating assay
The bait strain (EY111) containing the bait plasmid (pEG202, His^+ marker) expresses the bait protein fused to a DNA binding domain (DBD). The prey strain (EY93) containing the prey plasmid (pJG4-5, Trp^+ marker) expresses the prey protein fused to an activation domain. On glucose medium (SD), the prey strain does not express the prey protein. Mating the bait and prey strains results in a diploid yeast cell that expresses the bait and prey proteins when cultured on medium containing galactose (SGR). Prey plasmids encoding a protein that interacts with the bait protein are selected by their ability to induce expression of reporter genes, which allow the yeast cell to grow on media lacking leucine, adenine and induce a blue color in the presence of X-gal (SGR H-W-L-A- X-gal).

The antibiotic resistance reporter gene Aureobasidin A (Takesako *et al.*, 1993; Cerantola *et al.*, 2009) and pooling strategies (Bruno *et al.*, 1995; Jin *et al.*, 2006; Du *et al.*, 2006) have been used to reduce false positives and negatives in Y2H screens. The major limitation that has not been overcome in using the Y2H assay for screening combinatorial peptide libraries is the relative low affinity of the peptides that are isolated. This is most likely due to the relatively low number of peptides that can be screened.

1.4.2.2 Three-hybrid assays

Three-hybrid systems use a third molecule to facilitate or disrupt interactions between the two hybrid proteins (Figure 1.4). A variety of molecules can be screened in this assay including RNA (SenGupta *et al.*, 1996; Bernstein *et al.*, 2002), enzymatic substrates (Baker *et al.*, 2002), small molecules (Kley, 2004; Licitra and Liu, 1996; Becker *et al.*, 2004), and protein molecules (Vidal *et al.*, 1996; Huang and Schreiber, 1997; Kato-Stankiewicz *et al.*, 2002; Yin *et al.*, 2003). Advantages of three-hybrid systems are that non-protein molecules or protein molecules that are not compatible with two-hybrid systems can be used. For example, cyclic peptides are not compatible with two-hybrid systems, since there is no free terminus to attach a transcription activation domain. Three-hybrid systems have been used to isolate cyclic peptides that disrupt a given protein-protein interaction (Horswill and Benkovic, 2005). The disadvantage of three-hybrid systems is that only molecules that allow two-hybrid proteins to interact or that disrupt a known protein-protein interaction can be isolated.

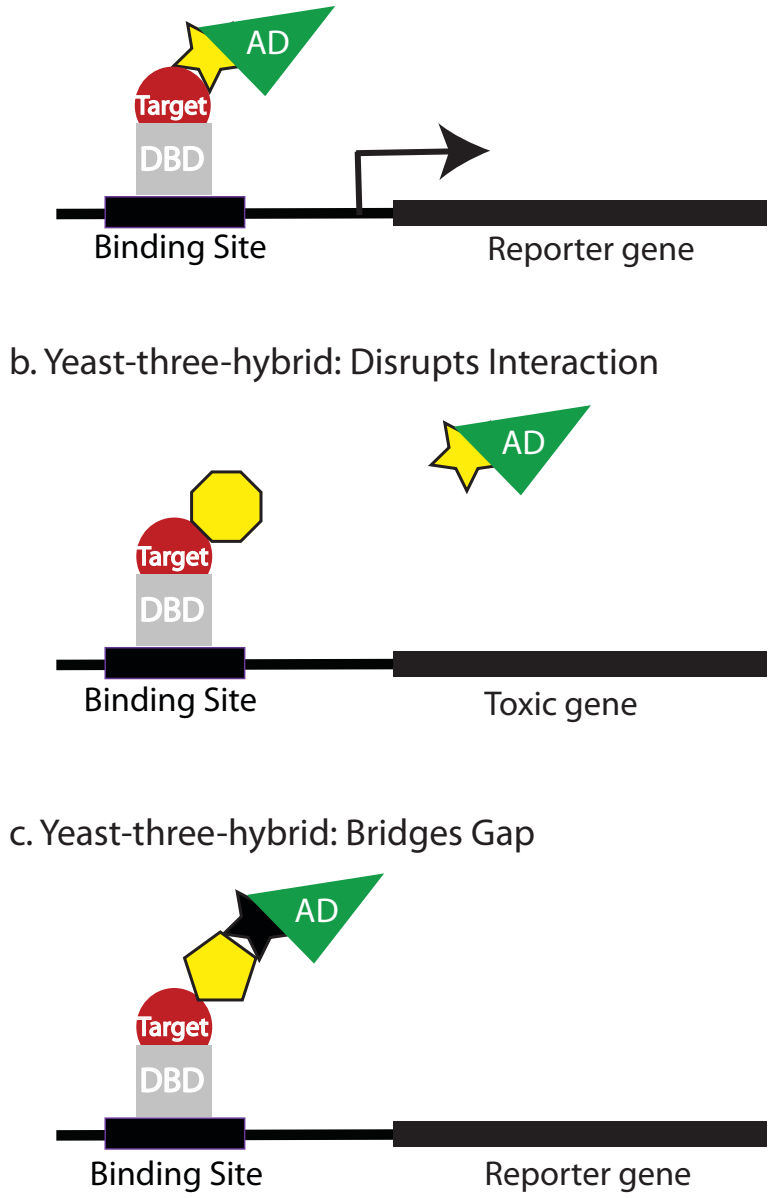


Figure 1.4 | Yeast three-hybrid assay
(a) In the Y2H system the bait and prey proteins interact, which allows the reporter gene to be expressed by localizing the activation domain to the promoter of the reporter gene. **(b)** The yeast three-hybrid disruption assay is used to detect molecules that prevent a protein-protein interaction using a toxic reporter gene. Disruption of the target interaction blocks gene expression by preventing the expression of a toxic reporter gene, which allows the cell to grow. **(c)** The yeast three-hybrid association assay is used to detect a molecule that bridges the interaction between two proteins that do not interact. One protein is fused to a DNA binding domain and the other protein is fused to a transcriptional activation domain. A third molecule is required to bridge the gap between the proteins and localize the transcriptional activation domain fusion protein to the reporter gene.

1.4.2.3 Surface display methods for screening combinatorial libraries

Combinatorial peptide libraries can be displayed on the surface of phage, yeast, and bacteria. In phage display, surface proteins orientate and present the peptide library on the outside of the phage (reviewed in Sidhu and Koide, 2007). The target is immobilized on a column and combinatorial libraries of peptides are passed over the column. Phage displaying peptides that interact with the immobilized target remain attached to the column, while phage that display proteins that do not bind the target are eluted in subsequent wash steps (Figure 1.5). After several rounds of selection, the phage that bound to the column are collected and the DNA is sequenced to determine the sequence of the peptide(s) that interact with the target. Yeast display systems exploit the Aga2 surface protein to display peptide libraries (Boder and Wittrup, 1997). Bacterial display systems fuse combinatorial peptide libraries to the first nine amino acids of the major outer membrane lipoprotein (Lpp) or the outer membrane protein (OmpA, amino acids 66 – 159) (Francisco *et al.*, 1993). Alternatively, peptide libraries have been displayed from a loop in TrxA fused to the bacterial flagellin gene (fliC) (Lu *et al.*, 1995).

The advantages of surface display systems are two-fold. First, the genetic material, which encodes the protein of interest, is contained within the organism. This facilitates the sequence identification of peptides that interact with the target of interest. Second, the organism itself offers a mechanism to amplify the protein of interest. By combining multiple rounds of amplification and selection, the tightest binders from a library are isolated. The ability to include rounds of negative selection, where peptides that bind a similar target are removed, makes this an extremely powerful technique for isolating peptides that can discriminate between similar targets. For example, phage that recognize a protein with a single amino acid mutation can be isolated by performing rounds of negative selection where the wild-type protein is used to remove peptides that bind the wild-type protein, leaving only the peptides that bind the mutant protein (Gao *et al.*, 2009; Chang *et al.*, 2006). In the negative selection step, phage that interact with wild-type protein remain bound to the column and phage that do not interact with the wild-type protein are found in the elutant, which is then used in the positive selection round to isolate phage that bind the mutant protein. Multiple rounds of selection performed in this manner result in the isolation of phage that only interact with the mutant protein. The biggest advantage of display systems is that the highest affinity peptide can be isolated from potentially billions of different peptides using multiple rounds of selection

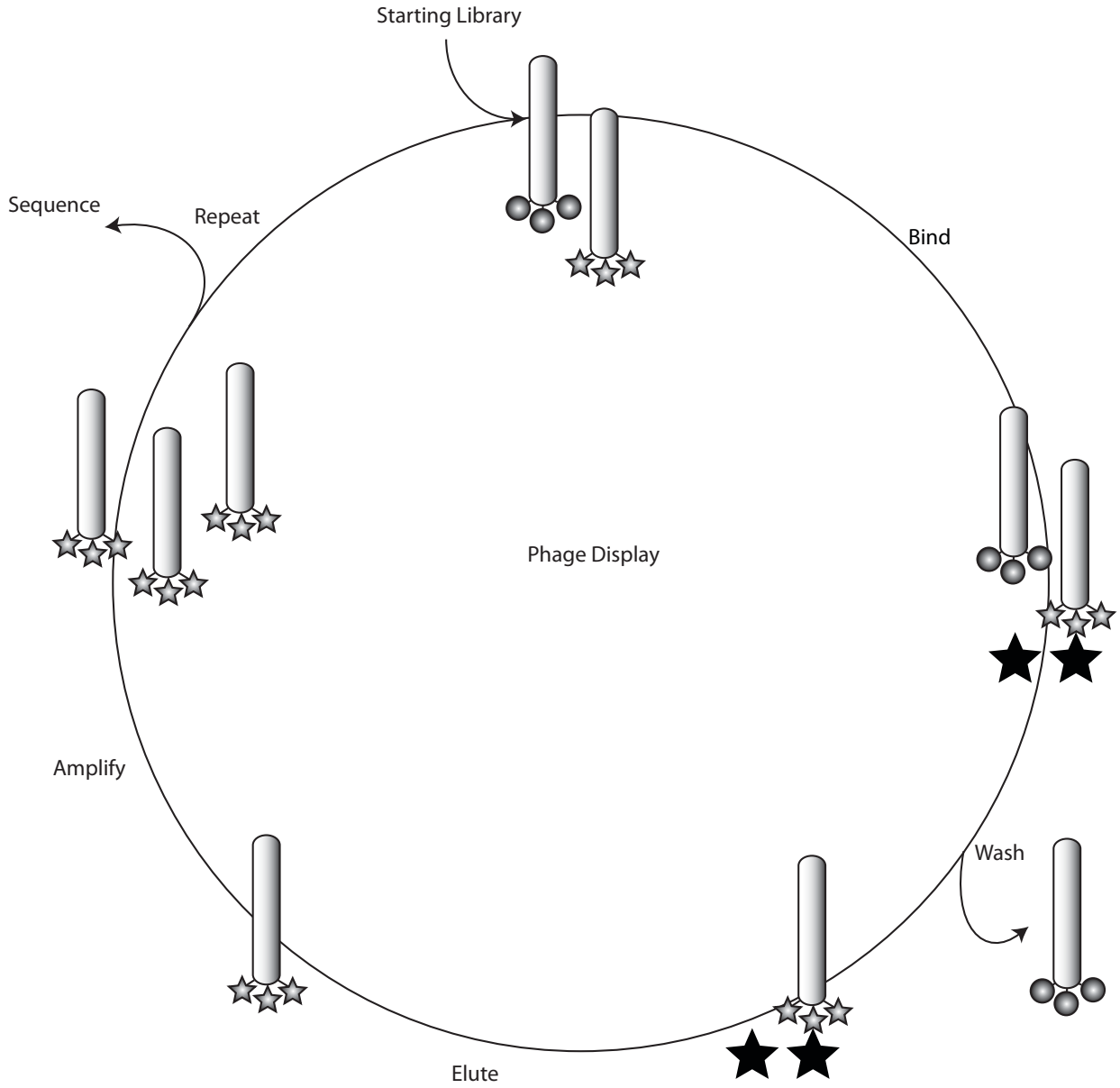


Figure 1.5 | Phage display
A library of peptides (Grey Circles/Stars) is displayed from the surface of a phage. The library is passed over the target (Black Stars) and only phage displaying peptides that bind to the target are retained in subsequent wash steps. Phage that display peptides, which bind the target are eluted and amplified. Subsequent rounds of selection are performed to isolate the strongest binding peptides.

and amplification. Negative selection rounds with the appropriate controls virtually eliminate the false positives associated with other screening technologies. The disadvantage of using multiple rounds of selection is that peptides that bind to low affinity sites may be missed. High affinity peptide reagents that bind to a non-therapeutic site are less valuable in drug design than a peptide affinity reagent that binds weakly to a therapeutic site. Other disadvantages are that the target protein must be purified or displayed on the surface of a cell and that the selection takes place in the extracellular environment. This is a concern since peptides selected in extracellular environments may not function within cells.

1.4.2.4 Ribosome and mRNA display

Ribosome and mRNA display are similar to the other display systems such as phage display with the main difference being that ribosome and mRNA display are performed in a cell-free system. Ribosome display links the mRNA encoding the protein directly to the protein (Hanes and Plückthun, 1997). In ribosome display, the ribosome is stalled, which causes the newly synthesized protein to remain non-covalently linked to the ribosome and mRNA. In mRNA display the protein is covalently cross-linked to the mRNA after protein synthesis (Roberts and Szostak, 1997). This covalent linkage allows the large ribosome to be removed from the mRNA:protein complex, simplifying subsequent interaction selections. mRNA/ribosome display share the advantages of other display systems by linking the protein to the mRNA that encodes them. Multiple rounds of selection and amplification are used to isolate high affinity interactions. Amplification is performed using PCR to enrich mRNA sequences that encode proteins that interact with the target after wash steps are performed.

Ribosome and mRNA display have additional advantages over the other cell-based display systems since they are performed *in vitro*, which allows non-natural amino acids to be incorporated using engineered tRNAs (Noren *et al.*, 1989; Forster *et al.*, 2003; Frankel *et al.*, 2003; Kawakami *et al.*, 2008). Since they are cell-free systems they have the largest library sizes, which can approach 10^{15} members. However, they share the same disadvantages as other display systems in that they typically isolate the affinity reagent with the strongest interactions and not necessarily ones that bind to the most important biological or therapeutic sites. Since the assay is performed *in vitro*, the target must be synthesized and fold correctly in the

extracellular environment while attached to a solid support, which can affect binding and stability.

1.4.2.5 Screening summary

The major obstacle in screening peptide libraries is correlating the phenotype/interaction with the active sequence. Genetically encoded libraries are typically used to address this issue, since large numbers of sequences can be screened and the sequence of the peptide that has activity is easily determined from the DNA encoding it. New technologies such as micro-fluidics combined with mass-spectrometry have the potential to replace genetically encoded peptide libraries (Shiau *et al.*, 2008). The combination of these technologies allows the direct identification of compounds that interact with the target using mass-spectrometry, circumventing the need to link the DNA encoding the peptide of interest to the interaction/phenotype. However, these techniques require expensive equipment and are still in their infancy.

Advantages of intracellular assays performed in yeast and mammalian cells are that no protein needs to be purified since the assay takes place in an intracellular environment. Also each cell acts as a test tube allowing individual interactions to be tested without competition. This allows peptides that interact with all binding sites, even weaker but potentially more therapeutically valuable ones, to be isolated. Intracellular assays also allow phenotypes to be assayed. For example, in yeast the kinase activity of human tyrosine kinases results in a growth suppression phenotype (Trible *et al.*, 2006), which can be assayed directly after initial reverse screens have been performed.

The advantages of extracellular assays are that they allow successive rounds of positive and negative selection, resulting in the selection of specific, high affinity interactions. *In vitro* systems share the advantages of the extracellular selections with the additional advantages that larger libraries can be screened and non-natural amino acids can be used.

Ideally, the optimal solution would be to first use a display system to screen large numbers of peptides and to isolate peptides that have high affinity for their target. This step would be followed by an intracellular screening assay to isolate peptides that function inside the cell.

1.5 Peptide affinity reagents and peptide scaffold proteins

Peptide affinity agents can be generated by screening linear peptides, cyclic peptides, or peptides constrained on the surface loop of a scaffold protein. Linear peptide affinity reagents are generally not preferred since they are readily degraded by proteases and are often not stable. Linear peptides are thought to be less structured and undergo a conformational change upon binding their target, which can reduce their binding affinities (Ladner, 1995). Peptides can be constrained by cyclizing them or displaying them in a surface loop of a protein. Proteins used to constrain peptides are referred to as scaffold proteins (Hosse *et al.*, 2006). The interaction of peptides constrained by a scaffold protein with their target has been shown to be ~ 1000-fold stronger than their linear counterparts (Cohen *et al.*, 1998). Many different proteins have been used to scaffold peptide affinity reagents including antibodies, peptide aptamers (Colas *et al.*, 1996), peptamers (Norman *et al.*, 1999), affibodies (Nord *et al.*, 1997), thioredoxin insert proteins (TIPS) (Böttger *et al.*, 1997), and perturbagens (Kamb and Teng, 2000). Skerra estimated that approximately 50 different scaffolds have been used to date (Skerra, 2007). Many of these scaffolds are used in academic settings, however several scaffolds have been commercialized and some are in clinical trials (Table 1.1).

1.5.1 Antibodies

Antibodies are naturally occurring affinity reagents produced by the immune system. Antibodies are made up of two heavy and two light chains (Figure 1.6). Each IgG heavy chain has four domains; three constant domains (C_{H1}, C_{H2}, C_{H3}) and one variable domain (V_H). IgG light chains have two domains; one constant domain (C_L) and one variable domain (V_L). Antibody molecule domains have a characteristic immunoglobulin fold, which consists of two beta-sheets that are packed tightly against each other and joined by a disulfide bond. Immunoglobulin folds are stabilized by hydrogen bonds between the beta strands of each sheet, by hydrophobic interactions between residues of opposite sheets, and by disulfide bonds between the sheets. The constant domains contain seven beta-strands arranged into two sheets of three and four strands. The variable domains are slightly larger and have nine beta-strands arranged in two sheets of four and five strands. The variable domains contain three CDRs that

Table 1.1 | Commercialized peptide affinity reagents
Molecule: the commercial or common name given to a peptide affinity reagent. **Scaffold:** the protein used to constrain and display the peptide. **Formally**: the company previously owning the patents to a given molecule.

Company	Molecule	Scaffold	Formally
Ablynx	Nanobodies	Llama antibody	
Adnexus	Adnectins	Fibronectin	
Affibody	Affibodies	Protein A (Z-domain)	
Amgen	Peptibodies	Antibodies	Avidia
Arana	Evibodies	Antibodies	Evogenix
Borean	Unnamed	Trimerized tetranectin	
Bristol-Myers-Squibb		Antibody	
ESBATech	scFV Fragment	Antibody	
Genmab	Unibodies	Antibody	
GlaxoSmithKline	Domain Antibodies	Antibody	Domantis
Imaxio	Peptide Aptamer	TrxA	Aptanomics
Immunocore		T-Cell Receptor	Avidex
Isogenica		CIS display	
Kalthera pty ltd	Cyclotide	Cyclotide	
Micromet	BiTEs	scFvs	
Molecular Partners	DARPins	Ankyrin	
NascaCell	Microbodies	Knottins	Selecore
Pfizer	Transbodies	Transferrin	Biorexis
Pieris Proteolab	Anticalins	Lipocalins	
Scil Proteins	Affilins	lens protein	
Trubion Pharmaceuticals	SMIPs	Antibody	

are separated by four framework regions. CDRs are responsible for the interaction of antibodies with their target antigen.

The vertebrate immune system is very effective at generating high affinity antibodies against a diverse range of epitopes. However, some epitopes cannot be targeted by antibodies raised in animals since they are recognized as self-antigens, are too small or do not elicit an immune response. These limitations can be overcome by using *in vitro* selection strategies (Section 1.4.2.3). In either case, antibodies isolated by these methods do not always bind their target *in vivo*. Since antibodies are selected outside of cells in both phage display and in the immune system, they may not fold correctly in an intracellular environment or may be prone to aggregation. Antigens may also fold differently than the native protein outside of cells, either due to accessory folding proteins or post-translation modifications depending on the expression system used to produce them. Unfortunately, antibodies cannot be selected within cells since they are not stable in the reducing intracellular environment. To address this problem, many groups have engineered antibody fragments to be more amendable to intracellular conditions, allowing technologies like the Y2H assay to be used for screening antibody-fragment libraries (Lobato and Rabbitts, 2004; Tanaka and Rabbitts, 2008) (Figure 1.6).

1.5.2 Fibronectin (Adnectin)

Fibronectin is a large extracellular protein involved in matrix and cell-cell interactions. Fibronectin does not contain disulfide bonds and thus it can be expressed within cells. NMR (Main *et al.*, 1992) and X-ray crystallography structures are available for fibronectin (Leahy *et al.*, 1992; Dickinson *et al.*, 1994), which facilitates scaffold design. The human tenth fibronectin type III domain is 94 amino acids long and forms an immunoglobulin fold consisting of seven beta-strands. This scaffold is referred to as either a monobody or an adnectin and has been commercialized by Adnexus (Table 1.1). Fibronectin interacts with other proteins via its two surface loops (Koide *et al.*, 1998). To create adnectin libraries, five amino acids in each of the two surface loops of fibronectin have been randomized. Adnectins have been isolated using phage display against ubiquitin (Koide *et al.*, 1998) and the Src SH3 (Src homology three) domain (Karatan *et al.*, 2004). mRNA display has been used to isolate Adnectins against tumor necrosis factor alpha (TNF-alpha) (Xu *et al.*, 2002), vascular endothelial growth factor receptor–2 (VEGFR-2) (CT-322, Adnexus), the nucleocapsid protein from the severe acute respiratory syndrome coronavirus (SARS-CoV) (Liao *et al.*, 2009), and

the phosphorylated I-kappa-B-alpha peptide (Olson *et al.*, 2008). Yeast surface display has been used to isolate adnectins against lysozyme (Hackel *et al.*, 2008) and the Y2H assay has been used to isolate adnectins against the active form of the estrogen receptor (Koide *et al.*, 2002; Huang *et al.*, 2006).

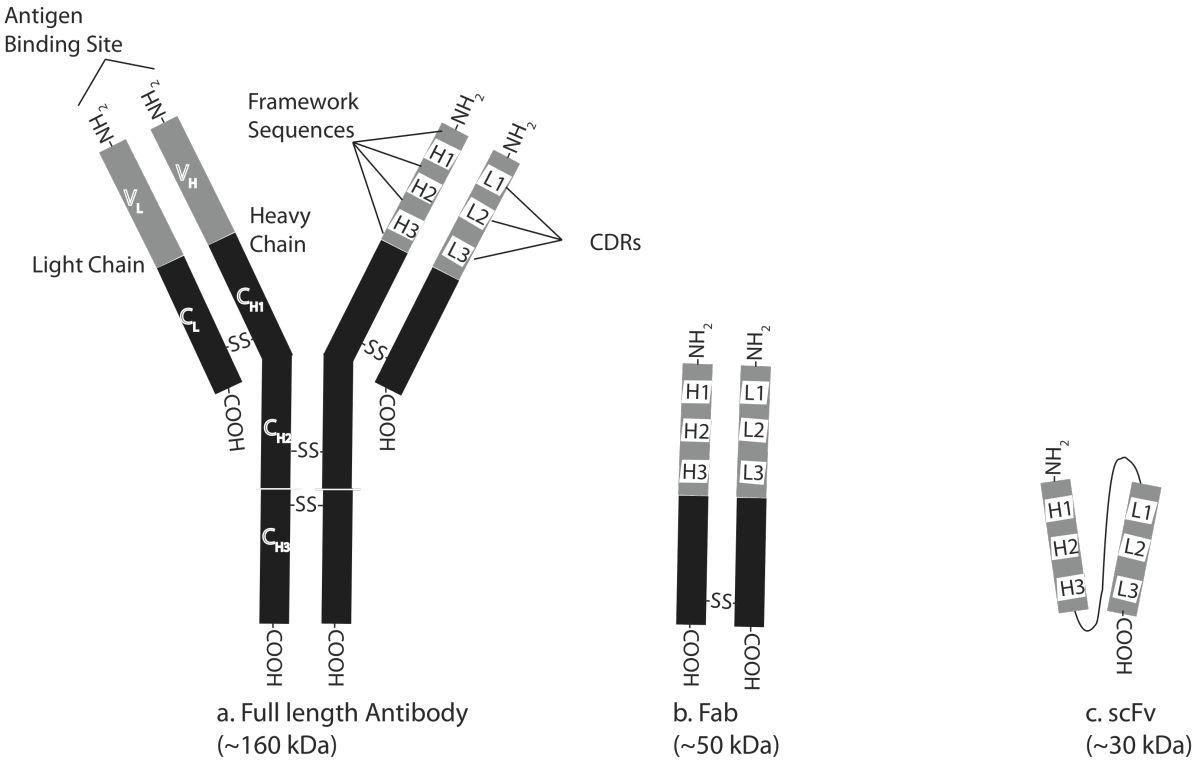


Figure 1.6 | Antibodies and antibody fragments
(a) Schematic of a full-length antibody consisting of two light chains and two heavy chains. The heavy chains are made up of three constant regions (C_{H1-3}) and one variable region (V_H). The light chains are made up of one constant region (C_L) and one variable region (V_L). The antigen-binding site is made up of variable regions from light and heavy chains. Variable regions are further subdivided into framework regions (FR), which function as scaffolds for displaying hypervariable regions (H1-3 or L1-3) called complementarity determining regions (CDRs). **(b)** Antigen binding fragment (Fab) is a truncated antibody comprised of the variable region and one constant region from each of the heavy and light chains. **(c)** Single chain variable fragment (scFv) contains variable regions from the heavy and light chains joined by a linker peptide.

1.5.3 Lipocalins (Anticalin)

Lipocalins are naturally occurring carrier proteins for a wide variety of different molecules including vitamins, lipids, and steroids (Grzyb *et al.*, 2006). Lipocalins form six or eight stranded beta-barrel structures and are approximately 20 kDa in size (Grzyb *et al.*, 2006). Lipocalins are excreted proteins and like antibodies they function outside of cells. Engineered lipocalins with modified or randomized binding sites are called "anticalins". Anticalins have many benefits over antibodies such as, their reduced size and superior expression in *E. coli* (Skerra, 2008). Anticalins have been commercialized by Pieris Proteolab (Table 1.1). Anticalins have been selected against small molecules (Skerra, 2008), Cytotoxic T-Lymphocyte Antigen 4 (CTLA-4)(PRS-010, http://www.pieris-ag.com) (Schlehuber and Skerra, 2005), Vascular endothelial growth factor (VEGF) (PRS-050, http://www.pieris-ag.com), and several undisclosed targets.

1.5.4 Ankyrin repeats (DARpins)

Ankyrins are a subset of a class of proteins called domain repeat proteins (DRPs) (Andrade *et al.*, 2001; Bork, 1993) and have been commercialized by Molecular Partners (Table 1.1). Ankyrins are naturally occurring proteins involved solely in protein-protein interactions. Ankyrin proteins consist of repeated domains, which are responsible for their binding activity. Each domain is made up of thirty-three amino acids, which form a beta-turn and two alpha-helices (Sedgwick and Smerdon, 1999). Repeating ankyrin domains called DARPs have been modified to isolate artificial ankyrin proteins that bind novel targets. These modified ankyrin proteins consist of a fixed N- and C-terminal capping module and two or three repeat domains. Residues in the repeat domain element have been randomized to create combinatorial libraries of ankyrin proteins that bind novel targets (Forrer *et al.*, 2003). Typically, six amino acids in each repeated domain are completely randomized (Binz *et al.*, 2004). Since ankyrins naturally exist inside of cells they have the potential to target intracellular targets. Ankyrin repeat proteins that bind to maltose binding protein (MBP) (Binz *et al.*, 2004), Jun N-terminal Kinase (JNK2), p38 kinase, the CC2-LZ domain of NF-Kappa-B Essential Modulator (NEMO) (Wyler *et al.*, 2007), and human Cluster of Differentiation 4 (CD4) (Schweizer *et al.*, 2008) have been selected using ribosome-display (Amstutz *et al.*, 2006). Ankyrin repeat proteins have also been isolated that bind proteinase inhibitors in *E. coli* using a two-step selection procedure. In this selction, combinatorial libraries of ankyrin repeat

proteins were first screened in an *in vitro* assay. Those that interacted with their target were then tested in a modified two-hybrid intracellular assay (Kawe *et al.*, 2006).

1.5.5 Thioredoxin

Thioredoxin (TrxA) is a small globular protein involved in intracellular redox reactions. TrxA has been engineered to display a constrained peptide loop from its surface that mimics a single CDR displayed from the framework of an antibody (Colas *et al.*, 1996). In contrast to antibodies (Figure 1.6), which have six binding loops, TxA peptide affinity reagents have only a single loop. The TrxA active site loop is tolerant to peptide insertions (LaVallie *et al.*, 1993). Crystal structures of TrxA show that the active site loop is part of an extended alpha helix and mutations at the active site loop have been shown to disrupt the helix and create a solvent exposed loop (Collet *et al.*, 2005). The active site loop of TrxA is used to display combinatorial libraries of peptides that are typically between five and thirty amino acids in length. The TrxA scaffold has been commercialized for intracellular applications by Aptanomics (Table 1.1)

Initially, the TrxA scaffold was used to display combinatorial libraries of peptides on the flagellum of *E. coli* (FLITRX) (Lu *et al.*, 1995). TrxA peptide aptamers were first used in intracellular screens using the Y2H assay to isolate peptide aptamers that interacted with cyclin-dependent kinase 2 (CDK2) (Colas *et al.*, 1996). Since then, TrxA peptide aptamers have been isolated against at least 30 targets intracellularily (see Table 1.2). Due to the success in isolating TrxA peptide aptamers against a variety of targets subsequent studies have examined its binding and clinical properties. In some cases, TrxA peptides could not be moved from the TrxA scaffold to another scaffold protein (Klevenz *et al.*, 2002), which indicates that the TrxA protein is involved in the interactions with at least some of its targets (Woodman *et al.*, 2005). This poses a problem since TrxA was not designed to be a therapeutic drug, but rather to function as a scaffold to stabilize and constrain random peptide loops. The ends of the peptide loop in TrxA contain cysteine at both ends, which can potentially be used to constrain the peptides via the formation of a disulfide bond. The bacterial TrxA scaffold is also immunogenic. Borghouts *et al.* developed a modified human TrxA to reduce immunogenicity, increase expression, and reduce non-specific interactions (Borghouts *et al.*, 2008). TrxA represents an important scaffold, which demonstrates that synthetic binding scaffold proteins are viable for peptide affinity reagents (Colas *et al.*, 1996).

Table 1.2 | Screens using the TrxA scaffold

Target	Class	Reference
hsCDK2	Kinase	(Colas *et al.*, 1996)
dmCDK2	Kinase	(Kolonin and Finley, 1998)
Phenotype		(Geyer *et al.*, 1999)
E2F	Transcription factor	(Fabbrizio *et al.*, 1999)
ThyA	Thymidylate synthetase	(Blum *et al.*, 2000)
HPV16 E6	Viral protein	(Butz *et al.*, 2000)
HBV core protein	Viral protein	(Butz *et al.*, 2001)
HPV16 E7	Viral protein	(Nauenburg *et al.*, 2001)
Rho-GEF	GEF	(Schmidt *et al.*, 2002)
EGFR	Receptor	(Buerger *et al.*, 2003)
Ras	GTPase	(Kurtz *et al.*, 2003)
HPV16 E6	Viral protein	(Butz *et al.*, 2000)
STAT3	Transcription factor	(Nagel-Wolfrum *et al.*, 2004)
BCL-6	Transcription factor	(Chattopadhyay *et al.*, 2006)
ErbB2	Receptor	(Kunz *et al.*, 2006)
TGMV AL1	Viral protein	(Lopez-Ochoa *et al.*, 2006)
CK2	Kinase	(Martel *et al.*, 2006)
DHBV core	Viral protein	(Tomai *et al.*, 2006)
Phenotype	Antiproliferation	(de Chassey *et al.*, 2007)
PrP sc	Prion	(Gilch *et al.*, 2007)
Nr-13	Anti-apoptosis	(Nouvion *et al.*, 2007)
Fur	Transcription regulator	(Abed *et al.*, 2007)
HIV-1 Integrase	Viral protein	(Armon-Omer *et al.*, 2008)
Ras-GAP	GTPase	(Pamonsinlapatham *et al.*, 2008)
HPV E6	Viral protein	(Dymalla *et al.*, 2009)
A20	Viral protein	(Saccucci *et al.*, 2009)
LMO2	Transcription regulator	(Appert *et al.*, 2009)
pUL84	Viral protein	(Kaiser *et al.*, 2009)
LIVIN	Inhibitor of Apoptosis	(Crnković-Mertens *et al.*, 2010)
ID1/3	Inhibitor differentiation	(Mern *et al.*, 2010)

1.5.6 Stefin A (STM)

Considering the problems identified with the TrxA scaffold protein, Woodman *et al.* set out to find a new scaffold protein with the following properties: (i) a crystal structure to guide placement of randomized loop; (ii) high stability to allow a variety of peptides to be displayed; (iii) folds correctly regardless of the inserted random peptide; (iv) a biologically neutral structure that does not interact or have activity that would complicate the selection or function assignment of peptide affinity reagents; and (v) fold the same in eukaryotic as well as prokaryotic cells (Woodman *et al.*, 2005). Eight different scaffolds were tested and human stefin A (SteA), a small (90 a.a.), lysosomal peptidase from the cystatin family, was ultimately selected. SteA is not post-translationally modified, does not contain disulfide bonds, and folds correctly in a variety of different organisms (Woodman *et al.*, 2005). The peptidase activity of SteA was blocked by introducing mutations G4W and V48D and inserting an *Rsr*II restriction site between amino acids 72-74 site to allow the insertion of random peptide libraries (Woodman *et al.*, 2005). The mutant steA scaffold is referred to as STM. The STM scaffold has been used to select peptide affinity agents against the yeast Sho1p SH3 domain (Woodman *et al.*, 2005). Although there are few reported selections using this scaffold, it has been used in a number of novel applications. The STM scaffold has been used in surface plasmon resonance assays to detect protein-protein interactions (Johnson *et al.*, 2008; Davis *et al.*, 2007) and in a protein microarray format that generates an electric signal when STM binds to its target protein (Evans *et al.*, 2008).

1.5.7 Green fluorescent protein

Green fluorescent protein (GFP) is an interesting scaffold since the expression level and the effect of peptide insertion on the scaffold can be monitored directly using the fluorescent properties of GFP. Abedi *et al.* identified Gln157-Lys158 as the ideal insertion site to display a peptide loop from GFP (Abedi *et al.*, 1998). A random fifteen amino acid library was constructed and used to select peptides capable of transcriptional transactivation (Abedi *et al.*, 2001). GFP has also been optimized as a scaffold in mammalian screens to isolate peptide affinity reagents using retroviral expression vectors (Peelle *et al.*, 2001b; Peelle *et al.*, 2001a). GFP peptide affinity reagents have been isolated based on their ability to block cell proliferation (Hitoshi *et al.*, 2003) and disrupt the pheromone-response pathway in *S. cerevisiae* (Caponigro *et al.*, 1998).

1.5.8 *Staphylococcal* nuclease (Peptamer)

Staphylococcal nuclease is a Ca^{2+}-dependent, extracellular enzyme that catalyzes the hydrolysis of RNA and DNA (Cotton *et al.*, 1979). Staphylococcal nuclease is 16.8 kDa and consists of three beta-sheets that fold to form a beta-barrel (Jacobs and Fox, 1994). Staphylococcal nuclease has been used as a scaffold to display random peptide libraries from an exposed surface loop of inactivated staphylococcal nuclease (Norman *et al.*, 1999). This engineered scaffold is called a Peptamer. Peptamer libraries containing 16 random amino acids have been used in forward screens to isolate peptamers that inhibit the spindle checkpoint and pheromone-response pathway in *S. cerevisiae* (Norman *et al.*, 1999).

1.5.9 Staphylococcal protein A (Affibody)

The alpha-helical receptor domain Z from bacterial *Staphylococcal* protein A was used to construct Affibodies (Nord *et al.*, 1997). Random libraries of Affibodies have been constructed by displaying combinatorial peptide libraries from a solvent exposed loop on the surface of the Z-domain (Nord *et al.*, 1997). Affibodies have been isolated against *Taq* DNA polymerase, human insulin, apolipoprotein A-1 (Nord *et al.*, 1997), epidermal growth factor receptor (EGFR) (Friedman *et al.*, 2008), interleukin-2 (IL-2) (Grönwall *et al.*, 2008), Alzheimer amyloid beta peptides (Grönwall *et al.*, 2007), HIV-1 gp120 (Wikman *et al.*, 2006), and human epidermal growth factor receptor 2 (HER2) (Wikman *et al.*, 2004) using phage display. Affibodies have been commercialized by Affibody (Table 1.1).

1.5.10 Cyclic peptide affinity reagents

Cyclic peptides are self-constrained peptides and thus do not require a scaffold to be constrained. Cyclic peptides have diminished proteolytic susceptibility relative to linear peptides. They also display enhanced binding to their target due to their restricted conformational space (Hortin and Murthy, 2002; Li and Roller, 2002), which decreases entropy loss upon binding (Williams *et al.*, 2002). Both naturally occurring and synthetically designed cyclic peptides have been successfully employed as drugs to treat human diseases (Horswill and Benkovic, 2005).

1.5.10.1 Disulfide bond constrained cyclic peptides

Linear peptides containing two cysteines can be constrained by forming a disulfide bond and thus producing a side-chain to side-chain 'cyclic' peptide (Giebel *et al.*, 1995; Meyer *et al.*, 2006). Disulfide bond constraints are reversible and depending on their environment they form either linear or cyclic peptides. Disulfide constrained peptides are limited since the sequence between the cysteines has to allow the two cysteines to get within disulfide bonding distance (~ 2.0 Å). Disulfide bonds also form predominantly in extracellular environments and thus are not amenable to intracellular applications. Disulfide bond constrained peptides have been exploited extensively by nature to produce venoms (Escoubas *et al.*, 2008).

Defensins are naturally occurring disulfide bond constrained peptides found in mammals. There are three main classes of defensins: alpha, beta, and theta. Knottins and Cyclotides are naturally occurring peptides found in plants that contain up to three disulfide bonds. Knottins are linear peptides that are constrained by up to six cysteines, which form a knot of three disulfide bonds, giving them increased stability and a fixed rigid structure. Selecore has commercialized Knottins (Table 1.1). Cyclotides, like knottins, have six cysteines, but they are cyclized through the formation of amide bond between their N- and C-termini (Craik *et al.*, 1999), which folds them into a highly structured protein (Craik *et al.*, 1999). Cyclotides have six loops and each loop can be modified independently of the other loops, making them interesting affinity agents as they can have multiple different binding motifs on the same molecule (Craik *et al.*, 1999). Kalthera Pty. Ltd. has commercialized cyclotides (Table 1.1). Naturally occurring cyclotides have many activities including potent insecticides (Jennings *et al.*, 2001).

1.5.10.2 N- to C-terminus and side-chain cyclized peptides

Many naturally occurring cyclic peptides are non-ribosomally synthesized. Non-ribosomally synthesized cyclic peptides are constructed by specialized proteins, where each protein performs a step in the synthesis of the peptide (Kleinkauf and Von Döhren, 1996). Many naturally occurring non-ribosomally synthesized cyclic peptides are lactones (Watanabe and Oikawa, 2007) (Figure 1.7). Although non-ribosomally synthesized peptides are extremely

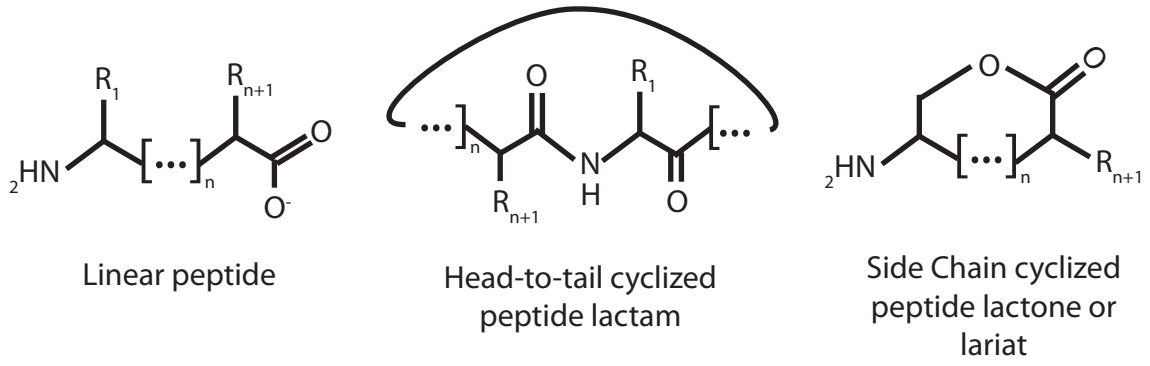


Figure 1.7 | Linear, lactam and lariat peptides.
Linear, lactam, and lactone peptides. R_1 represents the first amino acid side-chain and R_{n+1} represents the last amino acid side-chain in the peptide. In the lariat (lactone peptide) R_1 is serine.

interesting, it is difficult to genetically encode random libraries of non-ribosomally produced cyclic peptides due to their complex synthesis mechanisms. A better understanding of how these peptides are synthesized may allow genetically encoded libraries of non-ribosomally synthesized cyclic peptides to be produced.

A variety of methods have been used to mimic the constraints observed in non-ribosomally synthesized peptides. Chemical cross-linkers can be used to cyclize ribosomally produced proteins. For example, chemically cross-linking the N-terminus of a protein to a fixed lysine side chain using disuccinimidyl glutaratecyclization has been used to cyclize, and subsequently screen, peptide libraries using mRNA display (Millward *et al.*, 2005; Millward *et al.*, 2007). Other ways to generate cross-linked cyclic peptides include using non-naturally charged tRNA (Nakajima *et al.*, 2009), sugar molecules (Altamura *et al.*, 2009), or light cleavable cross-linkers.

Naturally occurring proteins that modify amino acid side chains or peptide bonds can be exploited to cyclize peptides inside cells. In theory, any protein modifying enzyme that introduces a side-chain:side-chain, a side-chain:N-terminus, a side-chain to C-terminus, or an N-terminus:C-terminus bond can be engineered to create cyclic peptides. For example, the mechanism for cyclotide cyclization is not known, but asparaginyl endopeptidase is linked to both the cleavage of the linear precursor and the ligation of the cyclic peptide (Saska *et al.*, 2007). Recently, the sortase protein has been used to cyclize proteins through an

intramolecular transpeptidation with the N-terminal glycine (Antos *et al.*, 2009). There are likely many other enzymes that have not been identified yet that can also be engineered to cyclize proteins. For example, a cyclic peptide with a glutamic acid to N-terminus linkage has been isolated, but the enzyme responsible has not been identified (Rosengren *et al.*, 2003). These enzymes will likely be exploited to produce cyclic peptides as the demand for genetically encoded libraries of cyclic peptides increases.

Naturally occurring split inteins have been permutated to express genetically encoded cyclic peptides (Scott *et al.*, 1999; Tavassoli and Benkovic, 2007; Scott *et al.*, 2001; Abel-Santos *et al.*, 2003; Horswill *et al.*, 2004; Naumann *et al.*, 2005; Horswill and Benkovic, 2005; Horswill and Benkovic, 2006). Combinatorial libraries of cyclic peptides have been created and genetically screened in forward and reverse assays to isolate inhibitors of biological processes (Kinsella *et al.*, 2002), enzyme activities (Kritzer *et al.*, 2009; Cheng *et al.*, 2007; Naumann *et al.*, 2008) and protein interactions (Tavassoli and Benkovic, 2005; Tavassoli *et al.*, 2008).

1.5.11 Scaffold summary

Antibodies are the most sophisticated scaffolds but they are expensive to produce, difficult to deliver, and are optimized for extracellular expression. Until the hurdles of using antibodies in intracellular environments are overcome, there will be motivation to develop alternate affinity reagents. Further, antibodies interact with antigen primarily through tyrosine (Koide and Sidhu, 2009). This could be a result of their finger like structure, which may sterically limit the type of surfaces an antibody can bind and prevent them from binding small clefts and protein active sites (Stijlemans *et al.*, 2004; De Genst *et al.*, 2006; Lauwereys *et al.*, 1998). Therefore, a mixture of different scaffolds or proteins may be required to obtain affinity agents against the entire spectrum of protein surfaces.

Many scaffolds have been commercialized and are being tested in clinical trials (Table 1.1). Most scaffolds are based on proteins whose natural function is to bind a variety of different molecules and are composed of a structured region and a variable region. Evolution has had years to optimize these scaffolds to bind their selected targets. Using these proteins capitalizes on the years of natural selection and modifying these proteins for new applications is often a very successful strategy.

Engineered monoclonal antibodies have been around for over thirty years and have only recently been used successfully as drugs (Laffly and Sodoyer, 2005). Engineered peptide aptamers based on the TrxA scaffold were initially developed fourteen years ago (Lu *et al.*, 1995; Colas *et al.*, 1996). Although few if any of these scaffolds are FDA approved drugs, many of them are now in clinical trials. Antibodies and antibody fragments have many significant drawbacks for use in therapeutic applications. These include overcoming their immunogenicity, poor cell and tissue penetrance, and difficulties in isolating them for therapeutic purposes. However, due to the huge potential of these molecules as therapeutics, many of these hurdles have been overcome and they are now successfully used to treat a variety of disease. As FDA approval of antibody-based drugs continues the number of antibody fragments and other peptide reagents in clinical trials should continue to increase (Aires da Silva *et al.*, 2008).

Cyclic peptides are desired therapeutic reagents since they are small, stable, and have high affinity for their targets (Finking and Marahiel, 2004; Craik, 2006). Cyclic peptides are typically isolated from natural sources or from chemically synthesized libraries. Natural sources of cyclic peptides include microorganisms, plants, and marine organisms (Craik *et al.*, 2002; Berer *et al.*, 2004). Cyclosporine is an example of a successful therapeutic cyclic peptide derived from a natural source (Borel, 2002). Cyclosporine is a potent, orally bioavailable cyclic peptide that suppresses the immune system. Although cyclic peptides have already been shown to be useful drugs, identifying therapeutic cyclic peptides is a difficult process, creating a bottleneck in their discovery and subsequent use.

Natural sources of cyclic peptides are composed of complex mixtures of peptides and small molecules. Identifying the active cyclic peptide from these complex mixtures is a difficult process (Escoubas *et al.*, 2008), which impedes the discovery of therapeutic cyclic peptides. To avoid this problem, chemically synthesized libraries of cyclic peptides are used (Baldwin, 1996; Edwards and Morrell, 2002; Huwe, 2006). Chemically synthesized libraries of cyclic peptides can be assayed individually or in small pools to facilitate the identification of active cyclic peptides. However, these screens are limited by the number of assays that can be performed (Tan, 2005; Webb, 2005).

A reverse genetic selection method to isolate cyclic peptides that interact with any given target *in vivo* would allow the rapid identification of potentially therapeutic cyclic peptides,

alleviating this bottleneck. Reverse genetic selection methods unlike conventional screening of biological derived or chemically synthesized cyclic peptides can be performed in a high-throughput manner and are amenable to a wide variety of targets. Reverse genetic screens offer four main advantages for isolating cyclic peptides against a given target. First, a reverse genetic screen unlike a forward screen selects directly for the ability of a cyclic peptide to bind the target of interest allowing specific proteins rather than phenotypes to be targeted. Second, the amino acid sequence of the active cyclic peptide is directly determined from the DNA encoding it simplifying the identification of the cyclic peptide with the desired activity. Third, synthetic oligonucleotides and current cloning techniques make it easy to produce large combinatorial peptide libraries allowing large numbers of cyclic peptides to be screened. Finally, *in vivo* selections allow peptides to be selected in an intracellular environment, which is useful for peptides that target intracellular proteins.

1.6 Intein-based synthesis of cyclic peptides

1.6.1 Intein background

Inteins are naturally occurring self-splicing proteins (Gogarten *et al.*, 2002). Inteins are thought to be parasitic genetic elements, which exist solely to propagate themselves (Gogarten and Hilario, 2006). Many inteins contain a homing endonuclease domain, which functions to propagate inteins by cleaving target genes lacking the intein and allowing the intein to be recombined into this site (Gogarten and Hilario, 2006). Inteins are the protein analogue of self-splicing introns, whereby introns interrupt mature mRNA and remove themselves post-transcriptionally, inteins interrupt mature proteins and remove themselves post-translationally. Exteins are the mature protein equivalent of an exon (Gogarten *et al.*, 2002). Inteins remove themselves from the precursor protein, which results in joining of exteins and production of the mature protein (Figure 1.8a). Naturally occurring inteins have also been found that act in *trans*. These inteins are referred to as split-inteins and they express the N-Extein:N-Intein as a separate protein from the C-Intein:C-Extein. The N-Intein (I_N) and the C-Intein (I_C) associate and ligate exteins together (Figure 1.8b). *Trans*-splicing inteins have been permutated to produce amide-cyclized (lactam) proteins and peptides (Scott *et al.*, 1999) (Figure 1.8c).

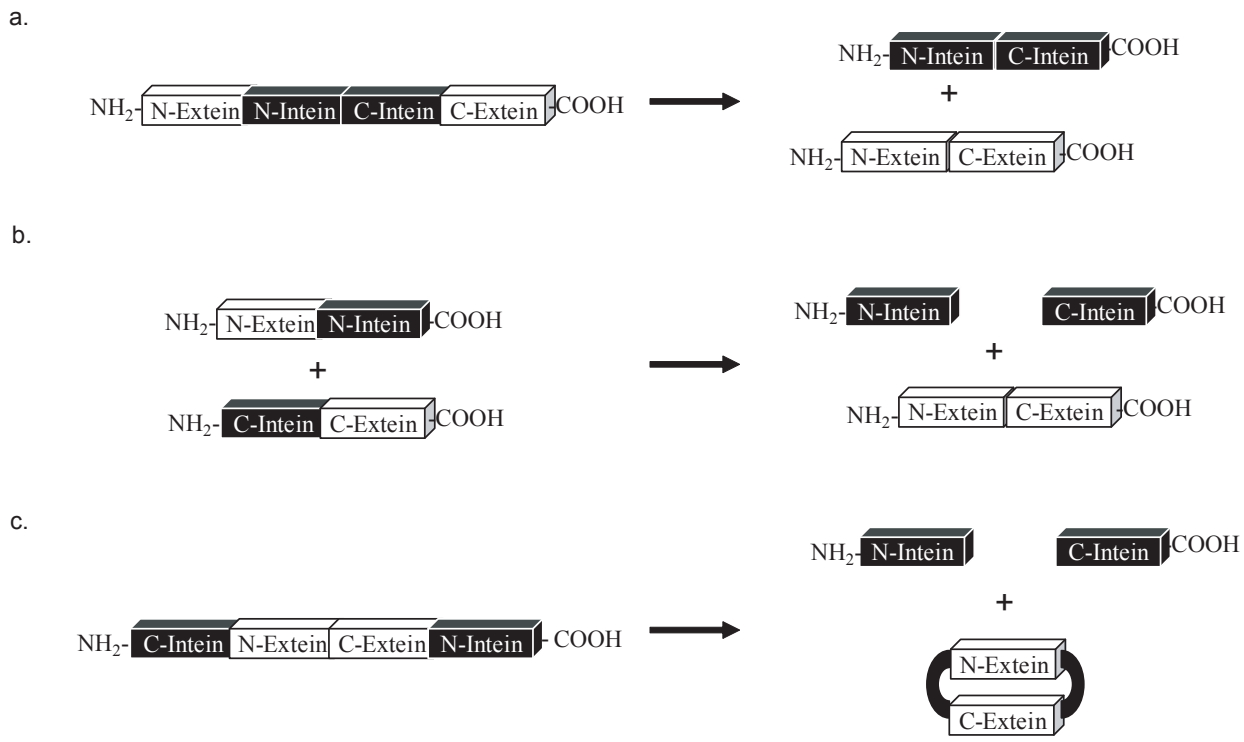


Figure 1.8 | Overview of intein-mediated protein splicing
(**a**) Self-splicing intein reaction. Intein domains (black) catalyze a self-splicing reaction that results in the extein domain fusion (white). (**b**) Split-intein reaction. Intein and extein domains are expressed separately. Interaction of the intein domains results in the ligation of exteins. (**c**) Intein-cyclization reaction. The orders of the intein domains relative to the extein domains are permutated. Intein domains fold together and catalyze cyclization of the extein domain.

1.6.2 Intein Nomenclature

1.6.2.1 Intein genes

Inteins are named by the organism the intein is found in, followed by the gene the intein disrupts. The organism's name is generally italicized and given a three-letter abbreviation. The first letter in the abbreviation comes from the first letter of the genus and the second and third letters are the first two letters of the species. For example, an intein found in *Synechocystis species* is abbreviated *Ssp* (letters used to form the abbreviation are underlined). The second part of the intein gene name is the abbreviation of the gene that is interrupted by the intein. For example, if in *Ssp* the DNA polymerase III alpha subunit (DnaE) gene is interrupted, then the full intein name is *Ssp* DnaE. If there is more than one intein found in the same gene, a suffix after the gene name is added to denote the order in the DNA sequence (5` – 3`) in which the intein occurs. For example, if two inteins disrupt the DnaE gene in *Synechocystis species* they would be named *Ssp* DnaE1-1 and *Ssp* DnaE1-2. For inteins that contain an endonuclease domain, where the endonuclease activity is confirmed, the convention is to add the prefix 'PI'. An additional suffix '-n' or '-c' is included for split inteins, '-n' refers to the N-terminal domain of the intein, and '-c' refers to the C-terminal domain of the intein. The *Ssp* DnaE intein is a split intein and thus the separate domains are referred to as *Ssp* DnaE-n and *Ssp* DnaE-c, which are abbreviated I_N and I_C, respectively.

1.6.2.2 Intein amino acid numbering

A numbering scheme has been developed to assist in comparing heterologous or foreign inteins to each other. This convention numbers amino acids in inteins sequentially from the N-terminus to the C-terminus beginning with the first residue of the intein and ending with the last residue of the intein. This becomes confusing when comparing amino acids far from splice sites or when comparing natural inteins to split-inteins and permutated inteins that are used to cyclize proteins. Therefore, the following numbering schemes will be used: The I_N intein domain will be numbered [I_{N+1}, I_{N+2}, I_{N+3}, …] from N-terminus to C-terminus. The extein will be numbered from the C-terminus to the N-terminus [I_{N-1}, I_{N-2}, I_{N-3}, …] or from the N-terminus to the C-terminus [I_{C+1}, I_{C+2}, I_{C+3}, …]. The I_C intein domain is thus numbered [I_{C-1}, I_{C-2}, I_{C-3}, …] going from the C-terminus to the N-terminus of the intein (Perler, 2002) (Figure 1.9). We developed an additional numbering system to compare conserved amino acids sites that are not

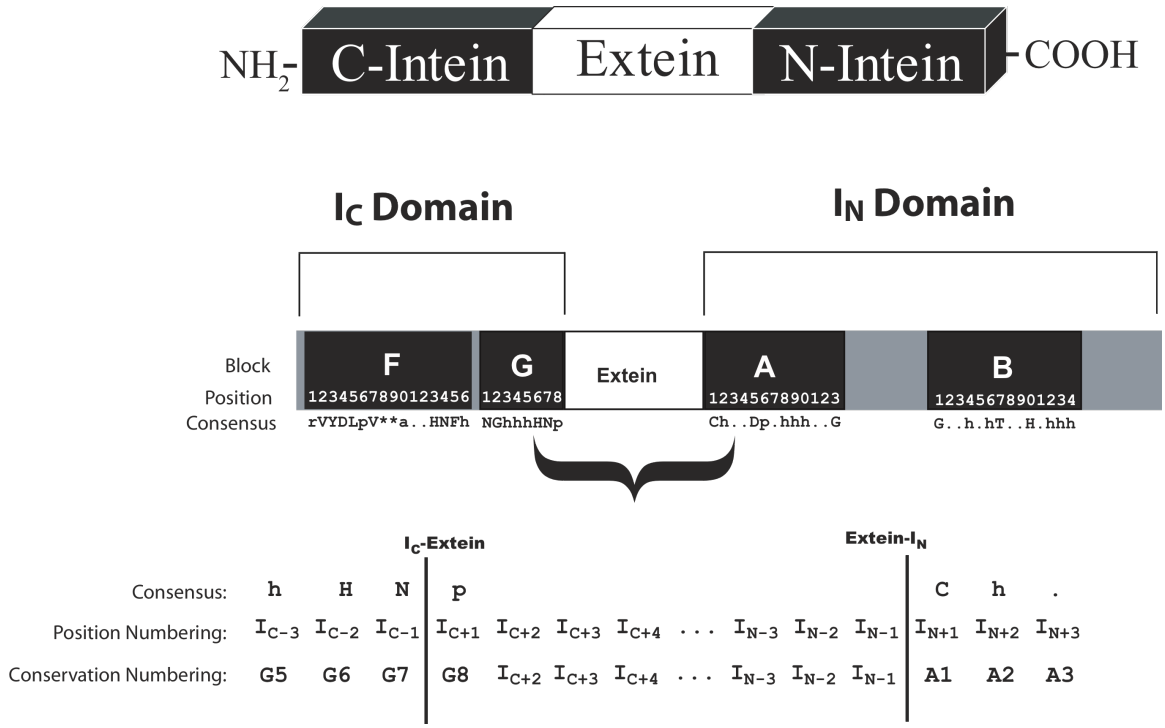


Figure 1.9 | Intein amino acid numbering scheme
The I_C domain contains conserved blocks F and G. The I_N domain contains conserved blocks A and B. Conserved amino acids are numbered according to their block number and position. An enlargement of the splice site at the I_C-Extein-I_N boundaries is shown. The I_C intein is numbered from C-terminus to N-terminus [I_{C-1}, I_{C-2}, I_{C-3} …] or according to block and position G1-8 and F1-16. The I_N intein is numbered from the N-terminus to the C-terminus [I_{N+1}, I_{N+2}, I_{N+3} …] or according to block and position A1-13 and B1-14. The extein is numbered from N-terminus to C-terminus [I_{C+1}, I_{C+2}, I_{C+3}, … , I_{N-3}, I_{N-2}, I_{N-1}]. The consensus sequence is derived from the analysis of more than 300 known inteins (Perler, 2002). Abbreviations are as follows: Uppercase: conserved amino acid from single letter amino acid code. Lowercase: r = Aromatic (F,Y,W); a = Acidic (D,E); h = Hydrophobic (G,V,I,L,A,M); p = Polar (S,T,C); . = Non-conserved; * = Gap used to align sequences.

close to the splice site between different inteins. To facilitate referring to these conserved amino acids, we used a numbering scheme based on conserved intein motifs. Several conserved motifs have been identified by comparing intein amino acid sequences. There are two nomenclatures for these motifs: (i) Blocks A, B, C, D, E, H, F, G (Pietrokovski, 1994; Telenti *et al.*, 1997) and (ii) Blocks N1, N3, EN1, EN2, EN3, EN4, C2 and C1 (Pietrokovski, 1998). We used the A, B, C, D, E, H, F, G nomenclature and each amino acid was assigned to a conserved block with a number from N-terminus to C-terminus. For example, I_{C+1} is the eighth amino acid from the N-terminus in block G and can also be labeled G8. Similarly, I_{N+1},

which is the first amino acid from the N-terminus of block A, can also be labeled A1. The I_N intein domain contains blocks A and B and the I_C intein domain contains blocks F and G. The region to be cyclized or the extein is numbered from N-terminus to C-terminus [I_{C+1}, I_{C+2}, I_{C+3},…, I_{N-3}, I_{N-2}, I_{N-1}] (See Figure 1.9 for overview of numbering scheme). For amino acids within five amino acids of the splice junctions, we named them using both conventions. Amino acids further than five amino acids from the splice site were referred to only using their conserved block letter and amino acid number. Although the I_N and I_C nomenclature is somewhat confusing when discussing permutated inteins where the I_C domain is found at the N-terminus and the I_N domain is found at the C-terminus, this nomenclature is preserved so amino acids within the permutated inteins can still be directly compared with their nonpermutated counterparts.

1.6.3 Intein-mediated protein splicing

There are two known mechanisms for intein-mediated protein splicing. The first mechanism is the most common and is referred to as the "standard" mechanism. The second less common mechanism is referred to as the "alternative" mechanism. These mechanisms are based on three amino acids that have been identified as crucial to intein splicing (I_{C+1} (G8), I_{C-1} (G7), and I_{N+1} (A1)).

1.6.3.1 Standard mechanism of intein-mediated protein splicing

In the standard intein mechanism for protein splicing there are four steps (Figure 1.10). Step 1 involves an N-X acyl shift where X is sulfur or oxygen, depending on whether cysteine or serine occurs at position I_{N+1} (A1). The acyl shift introduces a (thio)-ester into the amide backbone of the protein between I_{N+1} (A1) and I_{N-1}. Ester and thio-ester bonds are more labile than amide bonds and thus provide good leaving groups for the transesterification reaction in Step 2. Formation of the ester bond also positions the (thio)-ester bond for attack by the I_{C+1} (G8) cysteine, serine, or threonine nucleophile in Step 2 (Southworth *et al.*, 2000; Poland *et al.*, 2000). Step 2 involves a transesterification reaction with cysteine, serine, or threonine at position I_{C+1} (G8), acting as a nucleophile that reacts with the thioester or ester bond formed in the Step 1. This results in the cleavage of the I_N domain from the intein between amino acids at positions I_{N+1}(A1) and I_{N-1} and the formation of a branched intermediate via a thioester or ester bond between I_{C+1} (G8) and I_{N-1}. Step 3 involves an asparagine cyclization reaction that

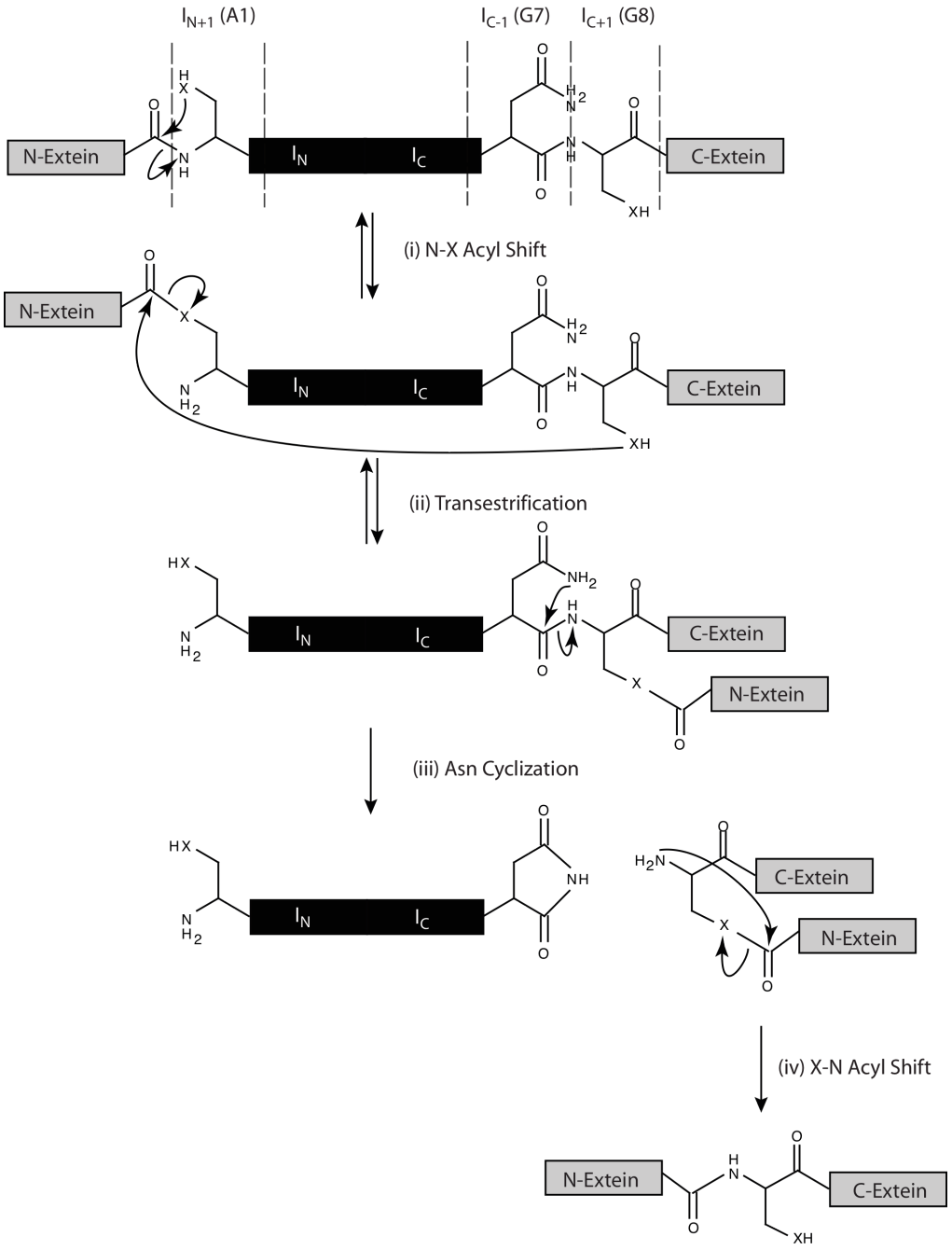


Figure 1.10 | Standard mechanism of intein-mediated protein splicing
(i) Cysteine or serine at the I_{N+1} (A1) position undergoes an N-X acyl shift. **(ii)** Serine, threonine, or cysteine at the I_{C+1} (G8) position undergoes a nucleophilic attack on the ester or thioester formed in step 1, which released the I_N domain is then released. **(iii)** Asparagine at the I_{C-1} (G7) position undergoes a cyclization reaction to release the I_C domain and produce the linear extein joined by an (thio)ester bond. **(iv)** The (thio)ester bond spontaneously rearranges to form a stable amide bond and the mature extein. "X" represents either sulfur or oxygen depending if side chain is cysteine, serine, or threonine.

cleaves the amide bond, connecting amino acids at positions I_{C+1} (G8) and I_{C-1} (G7), and releases the extein, which contains a thioester or ester bond between I_{C+1} (G8) and I_{N-1}. In Step 4, the ester bond is converted to an amide bond between I_{C+1} (G8) and I_{N-1} in the extein.

1.6.3.2 Alternative mechanism of intein-mediated protein splicing

In addition to the standard mechanism of intein splicing, there is at least one alternative mechanism that has been described (Southworth *et al.*, 2000). The alternative mechanism of intein-mediated protein splicing consists of only three steps. In the alternative mechanism there is no N-X acyl shift (Step 1 in standard mechanism). Inteins that use the alternative mechanism generally have serine or cysteine at position I_{N+1} (A1) replaced by alanine. Alanine occurs at I_{N+1} (A1) in 25/344 inteins in InBase (Perler, 2002) and these inteins likely undergo this alternative splicing mechanism. Inteins that have alanine at position I_{N+1} (A1) undergo a direct nucleophilic attack on the peptide backbone between I_{N-1} and I_{N+1} (A1) using the amino acid at position I_{C+1} (G8) (Figure 1.11) (Southworth *et al.*, 2000), which introduces an ester or thioester bond between the I_{C+1} amino acid and the I_{N-1} amino acid. The N-X acyl shift that occurs in Step 1 of the standard mechanism is not needed in inteins that use this alternative mechanism. The amide bond in these inteins is already aligned for direct attack by the nucleophile at I_{C+1} (G8) and therefore they do not need the extension in the backbone caused by the N-X acyl shift (Southworth *et al.*, 2000; Poland *et al.*, 2000). The remaining steps are the same as described for the standard mechanism.

1.6.4 Conserved amino acids in intein-mediated splicing

Based on 344 intein sequences in the InBase database (Perler, 2002), the majority (315/344) of sequences contain either cysteine (281 inteins) or serine (34 inteins) at position I_{N+1} (A1), which is involved in Step 1 of intein processing. These inteins most likely undergo the standard splicing mechanism. Twenty-five inteins (25/344) contain alanine at position I_{N+1} (A1) and therefore they most likely undergo the alternative mechanism of splicing. Two other amino acids that are highly conserved and are thought to be essential for proper intein splicing are located at positions I_{C+1} (G8) and I_{C-1} (G7), which are involved in Steps 2 and 3, respectively. The I_{C+1} (G8) amino acid is most frequently cysteine (139 inteins), serine (120 inteins), or threonine (81 inteins). The I_{C-1} (G7) amino acid is almost exclusively asparagine

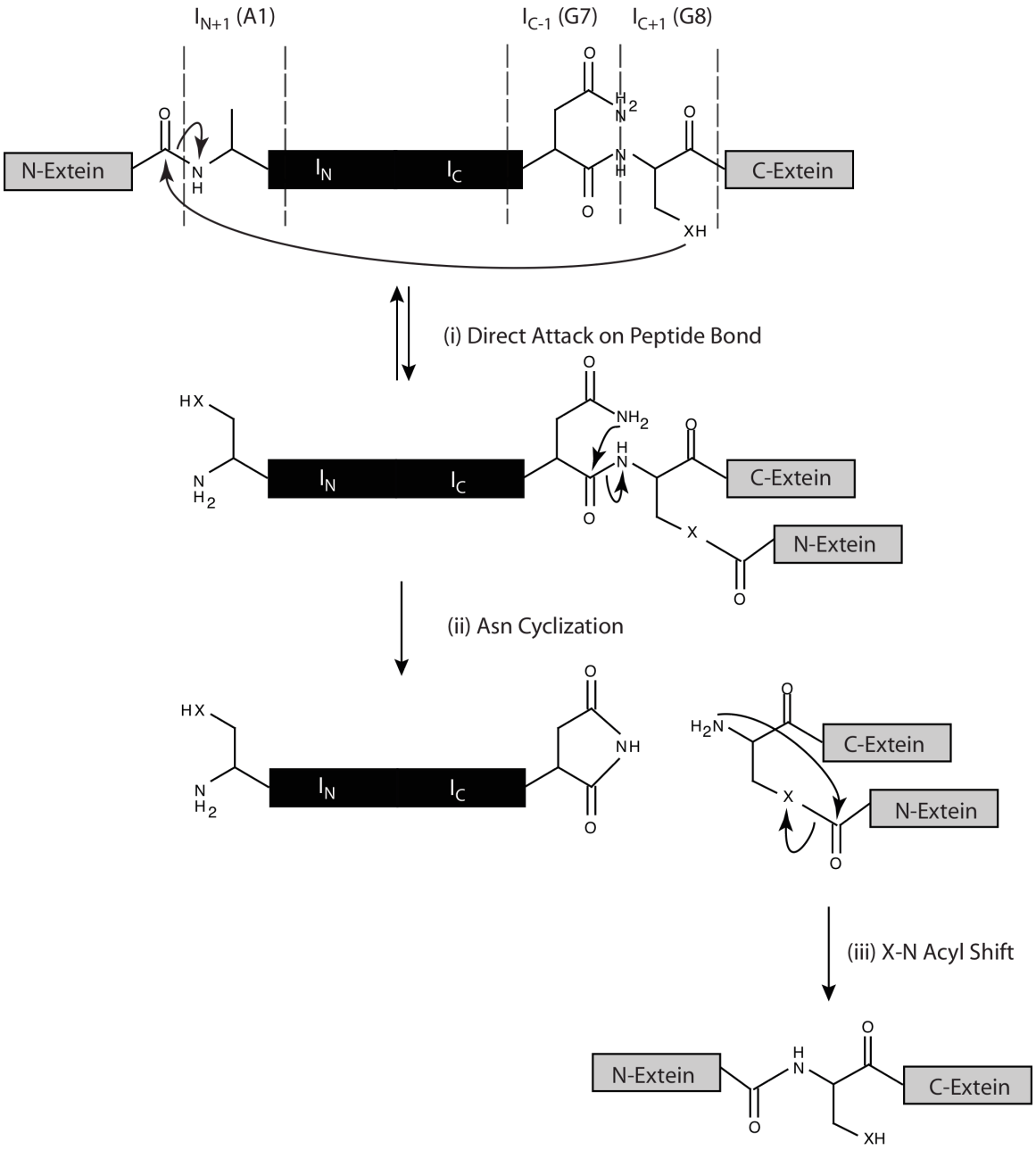


Figure 1.11 | Alternative mechanism of intein-mediated protein splicing
(i) In the alternative mechanism cysteine or serine at the I_{C+1} (G8) position directly attacks the carbonyl group of the I_{N+2} (A2) amino acid. The I_N domain is released and an (thio)ester bond is introduced between the extein domains. **(ii)** Asparagine at the I_{C-1} (G7) position undergoes a cyclization reaction to release the I_C domain and produce the linear extein joined by an (thio)ester bond. **(iii)** The (thio)ester bond spontaneously rearranges to form a stable amide bond and the mature extein. "X" represents either sulfur or oxygen depending if side chain is cysteine, serine, or threonine.

(327 inteins) although sometimes glutamine (15 inteins), which may also undergo cyclization (Pietrokovski, 1998) is present.

Other amino acids have also been observed at the three positions (I_{C+1} (G8), I_{C-1} (G7), and I_{N+1} (A1)) that are directly involved in splicing. These inteins may use a mechanism for intein splicing that is different from the standard or alternative mechanisms. For example, glutamine occurs at position I_{C-1} (G7) in place of asparagine and undergoes a similar cyclization reaction (Pietrokovski, 1998). Aspartic acid also occurs at I_{C-1} (G7) in place of asparagine in one intein but the intein still undergoes splicing (Amitai *et al.*, 2004). It was originally thought that aspartic acid like asparagine would undergo cyclization. However, this intein is capable of splicing even if aspartic acid I_{C-1} (G7) is mutated to alanine, which cannot undergo cyclization, indicating that there are still undetermined non-standard mechanisms for intein splicing (Amitai *et al.*, 2004).

1.6.5 Intein-mediated protein ligation

Inteins have been used to ligate peptides to other proteins (Xia *et al.*, 2008; Wood *et al.*, 2004; Evans *et al.*, 1999; Ludwig *et al.*, 2009). Protein ligation can be performed between two peptides; one containing a reactive cysteine at the N-terminus and one containing a thioester at the C-terminus. Proteins used in ligation reactions can be either synthesized or expressed and purified from cells. Cellular expression is typically cheaper and allows larger proteins to be used. The main problem with using proteins produced in cells is introducing the reactive groups. Proteases have been successfully used to introduce cysteines at the N-terminus of proteins (Erlanson *et al.*, 1996). Alternatively, inteins have been engineered to produce tagless peptides/proteins that have either a cysteine at their N-terminus or a reactive thioester at their C-terminus (Figure 1.12). Extracellular protein ligation using inteins has also been exploited for segmental isotopic labeling (Züger and Iwai, 2005; Muona *et al.*, 2008; Busche *et al.*, 2009; Aranko *et al.*, 2009). Segmental isotopic labeling is used to study large proteins and domains with NMR. The technique uses partially labeled proteins and NMR to visualize small, labeled segments present in the context of a large unlabeled protein.

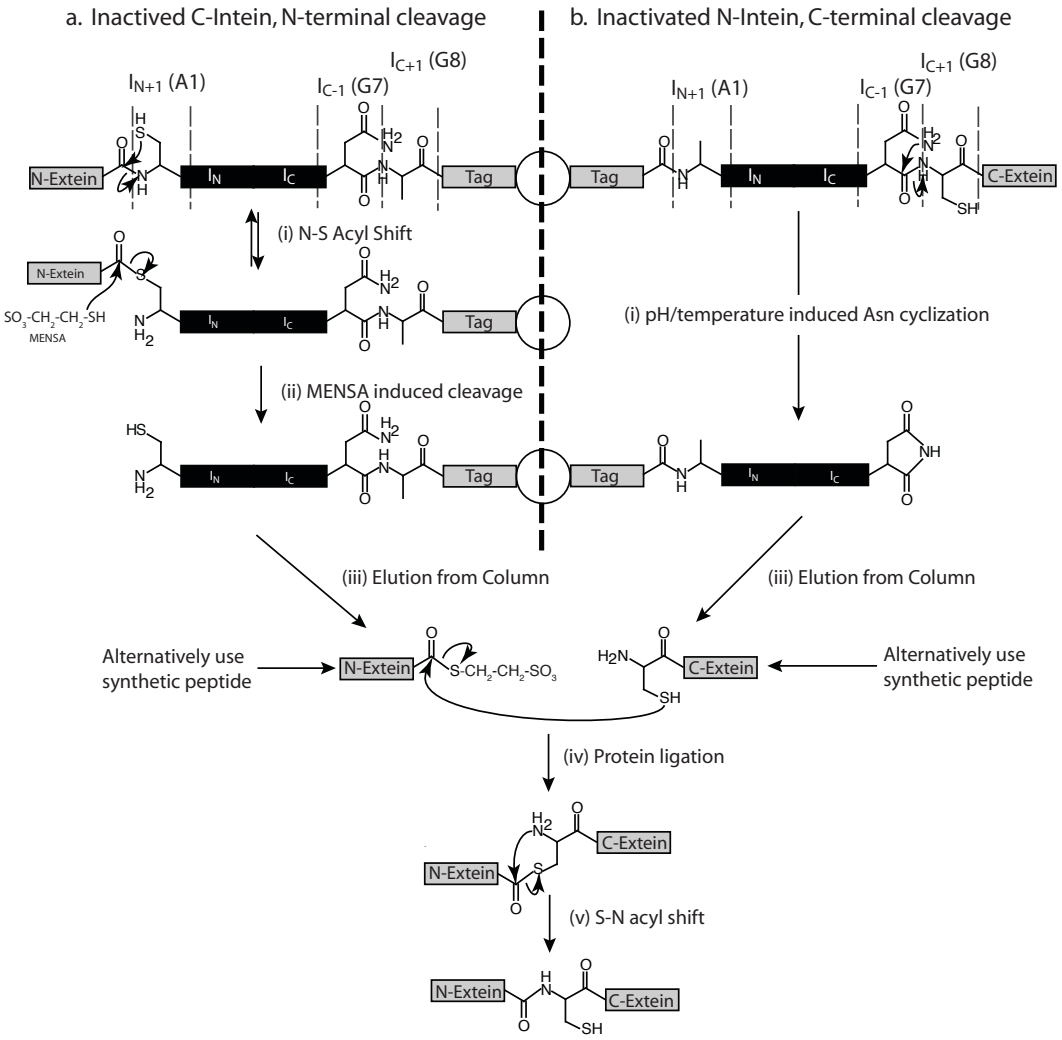


Figure 1.12 | *In vitro* intein-mediated protein ligation
(a) N-terminal fusion generating a peptide with a reactive C-terminal thioester. The protein target is fused to the N-terminus of an intein that cannot undergo asparagine cyclization since it is blocked by mutating asparagine (I_{C-1} (G7)) to alanine. (i) N-S acyl shift. Step 1 in intein processing introduces a thioester between the extein and the intein. (ii) MENSA-induced cleavage. The thioester formed in Step (i) is attacked by MENSA, which introduces a C-terminal thioester on the protein of interest. (iii) Elution from column. The intein remains attached to the column through an affinity tag. The modified extein is released from the column. (iv) Protein ligation. The extein containing the C-terminal thioester interacts with a N-terminal cysteine to produce a ligated protein. (v) S-N acyl shift. The final step is an S-N acyl shift resulting in the two proteins being joined by an amide bond. **(b)** C-terminal fusion, generating a peptide with a N-terminal cysteine. The protein target is fused to the C-terminus of an intein that cannot undergo an N-S acyl shift since it is blocked by mutating cysteine (I_{N+1}(A1)) to alanine. (i) Asparagine cyclization is induced by a shift in pH or temperature, resulting in the release of the extein with a N-terminal cysteine. The remaining steps are as described above. **(a)** or **(b)** can be used in conjunction with a synthetic peptide or both products can be generated via an intein reaction.

1.6.5.1 Extracellular intein-mediated protein ligation

Inteins can be used to ligate proteins outside of cells. Full-length inteins are mutated to allow either N-terminal or C-terminal cleavage (Figure 1.12). Engineered inteins that do not undergo N-terminal cleavage but allow C-terminal cleavage can be used to incorporate a reactive cysteine at the N-terminus of a protein. The cysteine at the N-terminus can be used for intein-mediated protein ligation (IPL) (Figure 1.12) (Wu *et al.*, 1998; Chong *et al.*, 1998; Chong *et al.*, 1998). The N to S acyl shift reaction in the intein is blocked by mutating the cysteine at I_{N+1} to alanine. The intein still undergoes inducible C-terminal cleavage through asparagine cyclization (Mathys *et al.*, 1999), which results in the release of the target protein with a cysteine at the N-terminus.

Alternatively, a reactive thioester at the C-terminus of the target protein can be generated by creating an N-terminal fusion to modified inteins (Telenti *et al.*, 1997). Modified inteins are fused to the C-terminus of the target protein. These inteins have their C-terminus asparagine (I_{C+1}) mutated to alanine, which blocks asparagine cyclization. The thioester linkage at the extein-I_N junction forms, but the intein does not process further. The thioester is cleaved using 1,4-dithiothreitol (DTT), beta-mercaptoethanol, cysteine, or 2-mercaptoethanesulfonic acid. Using 2-mercaptoethanesulfonic acid (MENSA) results in a reactive thioester at the C-terminus of the protein (Chong *et al.*, 1997; Chong *et al.*, 1998; Evans *et al.*, 1998; Southworth *et al.*, 1999).

1.6.5.2 Intracellular intein-mediated protein ligation

Inteins can be used for intracellular protein ligation (Giriat and Muir, 2003). Engineered split inteins, where the exteins are replaced with foreign genes, allow expression of two separate proteins under the control of different promoters. Each protein is fused to one half of the split intein. Co-expression of the two proteins results in ligation of the two target proteins *in vivo* using natural intein processing (Figure 1.8). Alternatively, a synthetic peptide can be created as a fusion with one split intein *in vitro,* which can then be delivered into the cell where protein ligation occurs *in vivo* (Giriat and Muir, 2003).

1.6.6 Intein-mediated protein cyclization

1.6.6.1 *In vitro* protein cyclization

Cyclic peptides or proteins can be generated *in vitro* using the methods described for *in vitro* protein ligation (Section 1.6.5.1) (Figure 1.13). In this system the target gene is flanked at the N-terminus by the full length *Ssp* DnaB intein, which has been modified so that the N-S acyl shift does not occur (Step 1 in Figure 1.13). The target gene is flanked at the C-terminus by a second intein (*Mxe* or *Mth*) that is modified to prevent asparagine cyclization (Step 3 in Figure 1.13). The C-terminal intein also has an affinity tag that allows it to be immobilized on a column. The protein is purified using an affinity column and then the *Ssp* DnaB intein is activated to undergo C-terminal cleavage, either through a change in pH or temperature, which exposes a cysteine at the N-terminus of the protein. The protein is still attached to the column through the C-terminal fusion with the *Mxe* or *Mth* intein affinity tag. The second intein undergoes the N-S acyl shift to produce the thioester between the extein-I_N junction, which is then cleaved by MENSA and releases the target protein with a C-terminal reactive thioester. An intramolecular reaction between the N-terminal reactive cysteine and the C-terminal reactive thioester results in the cyclization of the target protein (Figure 1.13).

1.6.6.2 Intracellular protein cyclization

Methods have been developed that use engineered inteins to produce lactam peptides *in vivo* (Horswill *et al.*, 2004; Tavassoli and Benkovic, 2005). Intracellular protein cyclization uses naturally occurring split inteins. The I_N and I_C domains are permutated to flank the peptide region to be cyclized. The affinity of these two domains brings the splice regions close together, facilitating and improving the efficiency of protein splicing and allowing it to occur *in vivo*. The final gene construct consists of an I_C domain:target-protein:I_N domain fusion. The intein undergoes the standard splicing mechanism (Figure 1.14). In Step 1, the side chain cysteine at I_{N+1} (A1) attacks the I_{N-1} carbonyl group, which results in a thioester bond at the extein-intein junction. In Step 2, the serine at I_{C+1} attacks the thioester bond, resulting in a lactone containing intermediate or "lariat", which is analogous to the branched intermediate in the standard mechanism (Figure 1.10). In Step 3, the asparagine side chain at position I_{C-1}/G7) cyclizes, producing a succinimide and releasing the extein from the intein. The extein is now cyclized by a lactone bond, which is resolved into the thermodynamically favored lactam bond

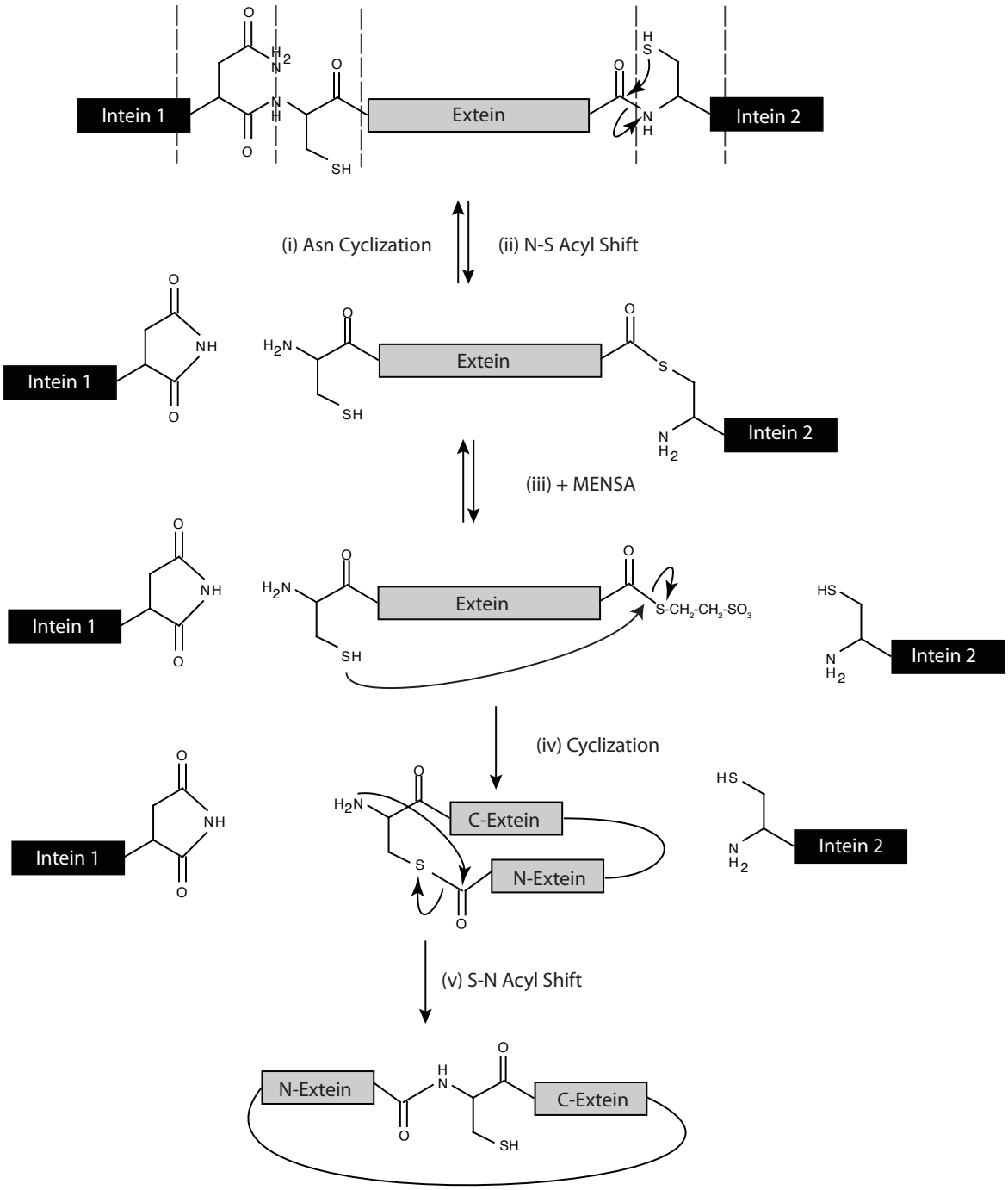


Figure 1.13 | *In vitro* intein-mediated protein cyclization
Two modified full-length inteins are used to cyclize exteins. Intein 1 has its N-terminal splicing reaction blocked. Intein 2 has its C-terminal splicing reaction blocked. **(i)** Intein 1 undergoes C-terminal cleavage through asparagine cyclization. **(ii)** Intein 2 undergoes N-S acyl shift. **(iii)** 2-mercaptoethanesulfonic acid (MENSA) is added to produce the thioester at the C-terminus of the extein. **(iv)** An intramolecular reaction between N-terminal cysteine and the reactive C-terminal thioester results in extein cyclization. **(v)** The mature amide cyclic peptide is produced by an S-N acyl shift,

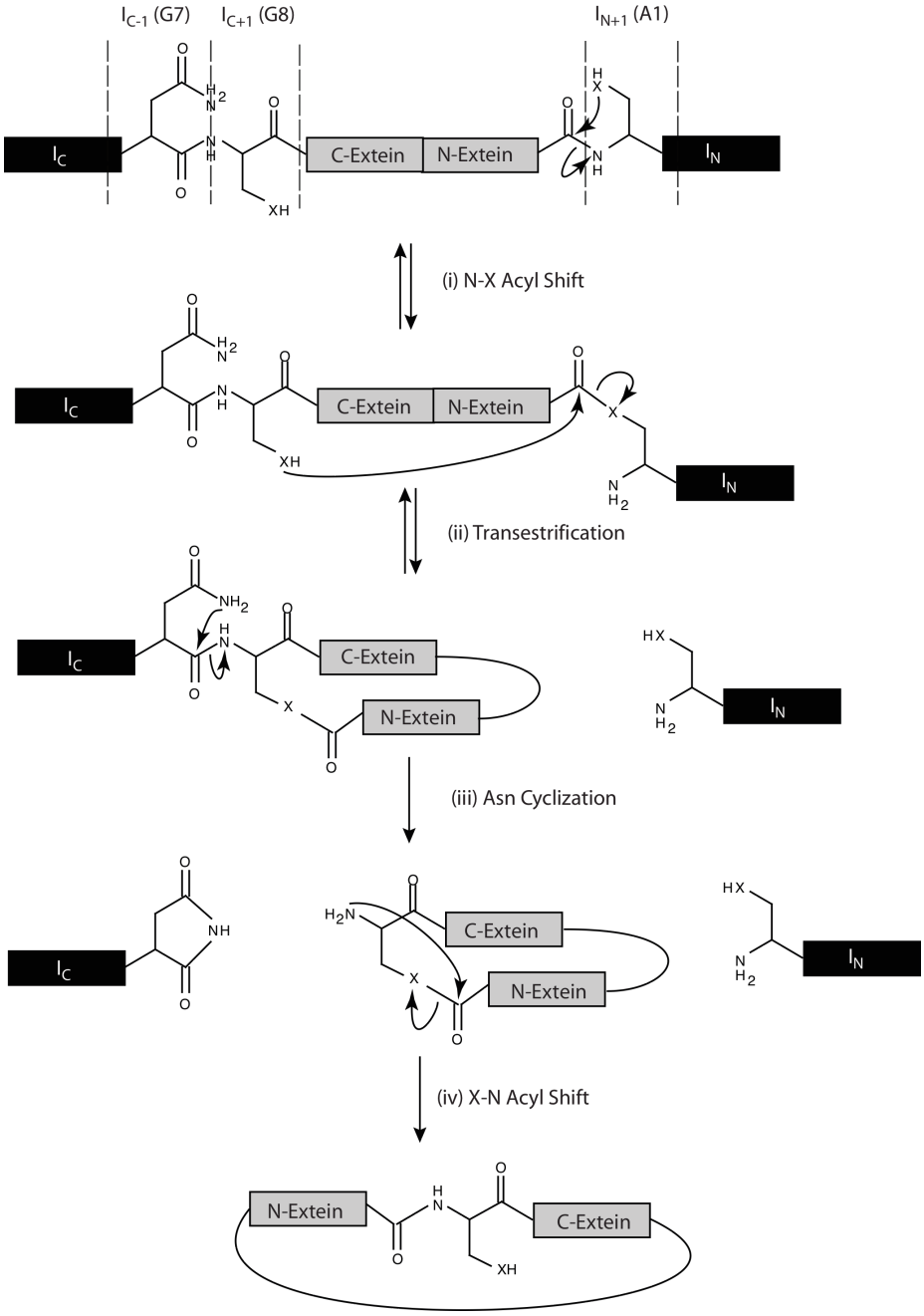


Figure 1.14 | *In vivo* intein-mediated protein cyclization
(i) Cysteine at position I_{N+1} undergoes an N-X acyl shift. **(ii)** Serine at the I_{C+1} positions undergoes a nucleophilic attack on the thioester formed in Step 1. The I_N domain is displaced, resulting in the formation of the lactone product. **(iii)** Asparagine at I_{C+2} position undergoes a cyclization reaction to release the I_C domain and produce the lactone-cyclized peptide. **(iv)** The lactone or thioester bond spontaneously isomerizes to a lactam. Where "X" represents either sulfur or oxygen depending if side chain is cysteine, serine, or threonine.

by a final X-N acyl shift. The result is an amide or lactam cyclized peptide with no free N or C –terminus (Figure 1.14).

1.7 Applications of peptide affinity reagents

1.7.1 Therapeutic target discovery and validation

Peptide affinity reagents can be used to validate the therapeutic potential of protein targets. Initially, siRNA/miRNA can be used to identify proteins that are candidate drug targets. Peptide affinity reagents can then be used to validate these as small molecule drug targets since siRNA/miRNA removes the target protein while peptide affinity reagents do not. Peptide affinity reagents can be co-expressed inside of cells and the effect on the pathways of interest and therapeutic potential of the target can be determined. In addition, peptide affinity reagents can be used to identify protein surface(s) to design small molecule inhibitors against (Bardou *et al.*, 2009). Small molecules that mimic the peptide interaction can be selected using three hybrid assays that displace the peptide affinity reagent. Alternatively, if a peptide/target structure is available, then this information can be used to design small molecule inhibitors.

Peptide affinity reagents can be used to target different protein conformations (Pamonsinlapatham *et al.*, 2008; Davis *et al.*, 2009), which allows specific alleles or post-translationally modified proteins to be inhibited. For example, point mutations that make a protein constitutively active can be specifically targeted either for selective inhibition or targeted destruction (McLean *et al.*, 2009; Colas *et al.*, 2000).

Peptide affinity agents can be used to target protein-protein interactions. Protein interaction surfaces represent a major source of drug targets, which have proven to be difficult to target using conventional small molecule approaches (Cochran, 2000; Toogood, 2002; Berg, 2003). Peptide based therapeutics share the same functional groups as the interactions they are disrupting and therefore should be able to mimic and block any protein-protein interaction.

1.7.2 Therapeutics

Peptides, like antibodies, have the potential to be used as drugs against extracellular targets. Peptides can be injected intravenously to target extracellular targets. Cell specific, tissue specific, or organ specific peptides can be coupled to toxic chemicals, allowing cytotoxic drugs to be delivered in high concentrations to a targeted cell population. This allows lower doses of drugs to be used since they are concentrated by localizing the drug to the target

(Ruoslahti, 2000). As methods for intracellular delivery of peptides progresses peptide affinity reagents will begin to be used as intracellular drugs. A variety of strategies have been developed to deliver peptides and proteins into cells. One strategy is to attach a protein tag that can deliver the peptide to the inside of the cell, also known as cell penetrating peptides (CPP) or protein transduction domains (PTD). The TAT-tag is from the HIV-1-derived Tat-peptide sequence and is used for intracellular peptide delivery (Wadia and Dowdy, 2003). Penetratin, transportan, and MAP (KLAL) peptides have also been used for intracellular delivery (Hällbrink *et al.*, 2001; Thorén *et al.*, 2005). Oligofusion proteins contain basic amino acid repeats are also delivered into cells (Han *et al.*, 2001; Futaki *et al.*, 2007). The intracellular uptake of peptides can also be improved by modifying amino acids. For example, peptide myristoylation has been shown to deliver peptides and proteins into cells (Nelson *et al.*, 2007; Dalton *et al.*, 2005; Ioannides *et al.*, 1990). Peptides localization sequences can also be fused to peptide affinity reagents to localize them to specific tissues or organs within the human body. This offers increased control of peptide delivery. Peptides have been isolated that cross the gastrointestinal mucosal barrier (Duerr *et al.*, 2004), which is crucial for oral delivery of peptides. Peptides have also been isolated that can cross the blood-brain barrier (Pasqualini and Ruoslahti, 1996), offering a unique opportunity to target peptides specifically to the brain.

1.7.3 Disease diagnostics and prognosis

Peptide affinity reagents can be used to visualize cell populations in a living organism using covalently attached fluorescent molecules or radiolabelled dyes (Zehnder-Fjällman *et al.*, 2007; Tolmachev *et al.*, 2006; Schaedel and Reiter, 2006). Tumors can be visualized, allowing for surgical removal or the diagnosis of metastasis (Schaedel and Reiter, 2006). Another potential diagnostic application of peptide affinity reagents is to make peptide microarrays, which through biomarker analysis improves diagnostics and can be used to guide treatment, forming the basis of personalized medicine (Ghadimi and Grade, 2010; Zhang *et al.*, 2009).

1.7.4 Basic research

Peptide affinity reagents can also be used for basic research. Antibodies typically used in Western blot analysis to determine protein abundance, modifications, and isoforms can be replaced with peptide affinity reagents. Western blot analysis is considered semi-quantitative since it involves numerous processing steps and the efficiency of protein transfer to

nitrocellulose varies between different proteins. Ideally, protein quantification inside of cells would be performed since it would skip these processing steps. Peptide affinity reagents unable to cross membranes could be used with standard fixation and permeabilization methods to gain access inside the cells. Peptide affinity reagents able to cross cell membranes would allow in-cell Western blot analysis on living cells. Furthermore, peptide affinity reagents are cheaper to synthesize and are more stable than antibodies, and can be used to replace antibodies in assays including immunoprecipitations, chromatin immunoprecipitations, fluorescent activated cell sorting, and immunohistochemistry.

2 Hypothesis and specific aims

The ease of generating genetically encoded combinatorial peptide libraries provides a readily available source of peptides that can be used to isolate inhibitors against a target of interest. Previously, this approach has been limited to linear peptides or peptides displayed and constrained on the surface of scaffold proteins. These approaches, however, suffer from a variety of limitations including poor stability, non-specific scaffold interactions and large size.

To target intracellular proteins, cyclic peptides must be able to retain their structure inside of cells, which excludes the use of peptides that are cyclized via disulfide bonds. Two methods to screen combinatorial libraries of cyclic peptides not constrained by a disulfide bond, but using genetic techniques have been published (Scott *et al.*, 1999; Millward *et al.*, 2007). In one method, genetically encoded libraries of peptides are chemically crosslinked to produce side-chain to N-terminus cyclized peptides (Millward *et al.*, 2007). In the other method, inteins are used to produce lactam-cyclized "head-to-tail" peptides, which are covalently cyclized through an amide bond (Scott *et al.*, 1999).

Millward *et al.*, have developed a strategy to cyclize peptides via chemical crosslinkers, which form a bond between the N-terminus and an internal lysine side chain forming cyclic peptides that have a free C-terminus (Millward *et al.*, 2007). These peptides are crosslinked at the C-terminus to the mRNA encoding them and selected using mRNA display (Millward *et al.*, 2007). Due to the way the peptides are crosslinked to form cyclic peptides the selection cannot be performed *in vivo*. Therefore this method is limited to *in vitro* selection systems, which require the target protein to be purified and may not be optimized to interact with their target intracellularily.

Lactam peptide cyclization produces peptides with no free N- or C- terminus. This strategy has been used to produce and screen genetically encoded libraries of lactam peptides *in vivo* (Tavassoli and Benkovic, 2007; Scott *et al.*, 1999; Scott *et al.*, 2001; Abel-Santos *et al.*, 2003; Horswill *et al.*, 2004; Naumann *et al.*, 2005; Horswill and Benkovic, 2005; Horswill and Benkovic, 2006). Lactam peptides have been isolated that reduce the toxicity of alpha-synuclein, a major factor in Parkinson's disease (Kritzer *et al.*, 2009), block IL-4 mediated IgE class switching (Kinsella *et al.*, 2002), block AICAR transformylase homodimerization

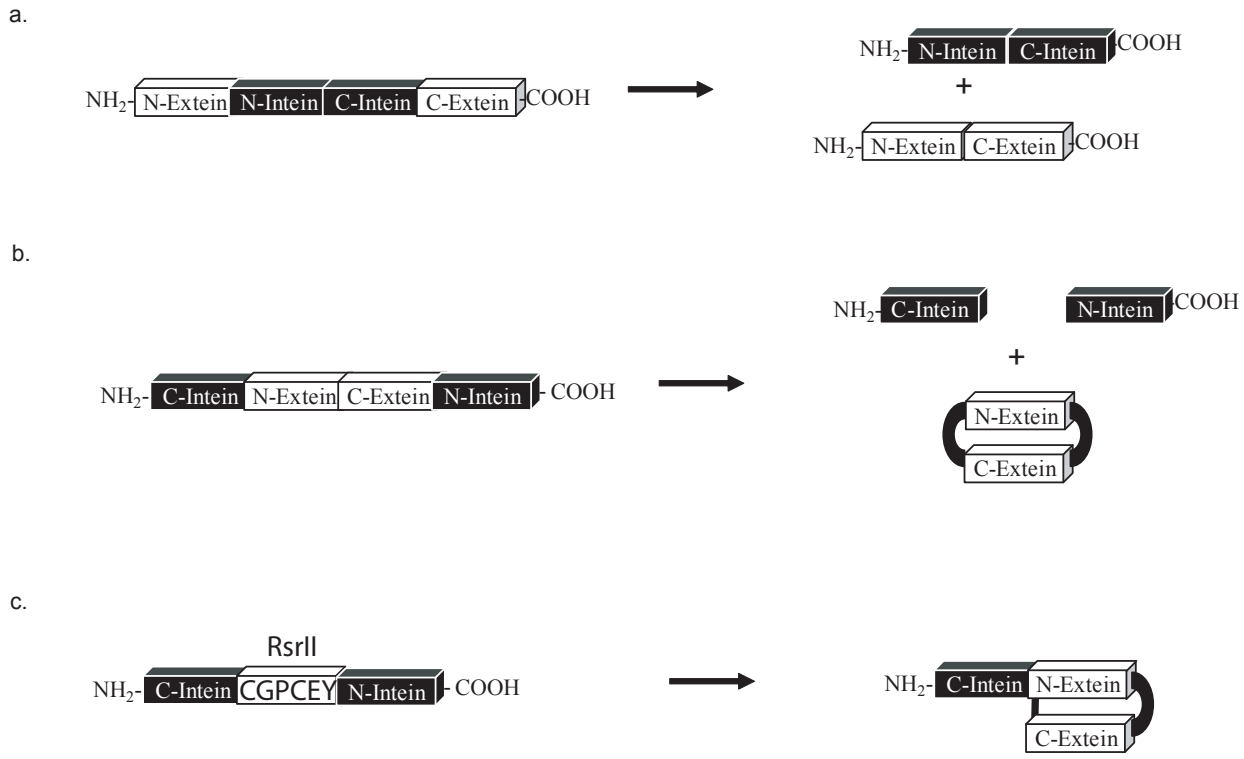


Figure 2.1 | Creating a lactone peptide or lariat
(a) Naturally occurring inteins contain an intein domain flanked on both sides by N- and C-terminal exteins. Intein processing results in the excision of the intein and a joining of the exteins. **(b)** Lactam peptide producing inteins are generated by permutating the order of the inteins and exteins. Intein processing results in the cyclization of the extein and cleavage of the C-and N-intein domains **(c)** Lariat producing inteins are generated by mutating the cyclic peptide intein so that the reaction is blocked at an intermediate step where only the N-intein is cleaved. The lariat intein produces a lactone peptide, which retains the C-intein domain.

(Tavassoli and Benkovic, 2005), inhibit ClpXP protease (Cheng *et al.*, 2007), block HIV budding via the Gag-TSG101 interaction (Tavassoli *et al.*, 2008), and inhibit DAM methyltransferase activity (Naumann *et al.*, 2008). Strategies used to isolate the above lactam peptides are based on one of three screening methodologies: (i) forward screens, which isolate cyclic peptides that inhibit a phenotype, (ii) reverse screens that are only applicable to specific target, or (iii) three-hybrid system where a known protein-protein interaction is disrupted. The first two strategies share the common problem of requiring specialized screens for each target

and are therefore not generally applicable. The third strategy limits the activity of the selected cyclic peptide to disrupting a protein-protein interaction. Even with the limited screening methodologies that can be used to isolate lactam-cyclized peptides, the number of cyclic peptides that have been isolated demonstrates the power of reverse genetic screens that can be performed *in vivo*. Reverse genetic screening strategies that can be used with lactam peptides are limited since protein fusion domains cannot be attached to them. The ability to genetically encode cyclic peptides that are fused to various protein domains would allow a wider variety of reverse genetic selection methods to be used to isolate cyclic peptides. For example, attachment of transcription activation domains and phage coat proteins would allow phage display and Y2H assays to be used to isolate cyclic peptides.

Although libraries of genetically-encoded lactam or cross-linked peptides can be screened they cannot be used in an *in vivo* reverse screen. Lactam peptides have no free N- or C-terminus (Figure 2.1) and thus cannot be used in the Y2H assay. A peptide cyclized through its side-chain to either the N- or C-terminus results in lariat structure with a free end terminus that can be fused to protein domains (Figure 2.1). By permutating the I_N and I_C domains of an intein to flank a continuous extein a lactam peptide can be produced (Scott *et al.*, 1999) (Figure 2.1). We hypothesized that the intein could be engineered to produce a lactone peptide; we refer to this specialized lactone as a lariat. We proposed that lariat structures could be produced by blocking the lactam peptide reaction at an intermediate step where the lariat is formed (Figure 2.1). Lariats have a free N-terminus that can be used for multiple fusion domains, including a transcriptional activation domain, which is required for the Y2H assay. To date, no methods have been developed to screen cyclic peptides against a specific target using a genetic assay such as the Y2H assay. The objective of this thesis was to develop new technology to isolate cyclic peptide inhibitors against proteins targets using the Y2H assay. To achieve this objective, the following aims were pursued.

2.1 Specific Aim 1: Construct lariat peptides

In Aim 1, methods were developed to construct a mutant intein that produced a lactone-cyclized peptide with a covalently attached N-terminus linear peptide "tail", referred to as a "lariat". Combinatorial libraries of lariat peptides were constructed and methods to screen the lariat libraries using the Y2H assay were developed.

2.2 Specific Aim 2: Optimization of lariat technology

In Aim 2, an alternative intein that was more permissive to amino acids at the splice junction was constructed and mutations were introduced to stabilize the lariat.

2.3 Specific Aim 3: Selection and characterization of lariat inhibitors of LexA

In Aim 3, the feasibility of the lariat screening assay was tested by isolating lariat inhibitors of the bacterial repressor protein LexA.

3 Materials and Methods

3.1 General information and protocols

3.1.1 Reagents and suppliers

General lab reagents and supplies were obtained from VWR. The suppliers and their addresses are shown in Table 3.1. Other non-standard reagents used in this study are shown in Table 3.2. Enzymes are listed in Table 3.3 and antibodies are listed in Table 3.4. Specialized lab equipment is listed in Table 3.5. Reagents to make media are listed in Table 3.6. Oligonucleotides are listed in Table 3.7.

Table 3.1 | List of suppliers' addresses

Supplier	Address
Alfa Aesar	Ward Hill, Massachusetts, USA
Amersham (GE)	Baie d'Urfe, Quebec, Canada
BD Biosciences	Mississauga, Ontario, Canada
BiaCore Life Sciences (GE)	Baie d'Urfe, Quebec, Canada
Bio 101 Inc.	Vista, California, USA
Bio-Rad	Hercules, California, USA
EMD	Gibbstown, New Jersey, USA
Fermentas	Burlington, Ontario, Canada
Integrated DNA Technologies (IDT)	Coralville, Iowa, USA
Invitrogen	Carlsbad, California, USA
LI-COR Biosciences	Lincoln, Nebraska, USA
New England Biolabs (NEB)	Ipswich, MA, USA
Q-BIOgene (MP Biomedicals)	Solon, Ohio, USA
Qiagen	Mississauga, Ontario, Canada
Roche	Mississauga, Ontario, Canada
Santa Cruz Biotechnology Inc.	Santa Cruz, California, USA
Sigma-Aldrich	Oakville, Ontario, Canada
Stable Isotopes	Summit,New Jersey, USA
Upstate (Millipore)	Billerica, Massachusetts, USA
V&P Scientific	San Diego, California, USA
VWR	Mississauga, Ontario, Canada
Waters	Milford, Massachusetts, USA

Table 3.2 | Reagents

Reagent	Supplier
Gel Purification Kit	Qiagen
Glass beads (450-600 μm in diameter)	Sigma-Aldrich
$H_2{}^{18}O$ (98%)	Stable Isotopes
HIS-Select® HF Nickel affinity gel (Ni^{2+}-NTA)	Sigma-Aldrich
Mitomycin C	Sigma-Aldrich
Ni^{2+}-NTA Spin column kit	Qiagen
Nitrocellulose	Bio-Rad
Odyssey blocking buffer	LI-COR Biosciences
Oligonucleotides	Integrated DNA Technologies (IDT)
PCR Clean-up Kit	Qiagen
Salmon sperm DNA	Sigma-Aldrich
Sodium (lump in kerosene)	Sigma-Aldrich

Table 3.3 | Enzymes

Enzyme	Supplier
iQ™ SYBR® Green supermix	NEB
Phusion™ Site-Directed Mutagenesis Kit	NEB
Platinum® Taq DNA Polymerase High Fidelity	Invitrogen
Shrimp Alkaline Phosphatase	Fermentas
T4 DNA ligase	NEB
Trypsin (Sequencing grade)	Roche

Table 3.4 | Antibodies

Antibody	Supplier
Goat Anti-Mouse LI-COR IRDye® 680 (680 nm)	LI-COR Biosciences
Goat Anti-Rabbit LI-COR IRDye® 800CW (780 nm)	LI-COR Biosciences
Mouse Anti-HA antibody	Santa Cruz Biotechnology Inc.
Rabbit Anti-LexA antibody	Invitrogen/Sigma-Aldrich
Rabbit Anti-LexA antibody (For ChIP experiments)	Upstate

Table 3.5 | Lab equipment

Equipment	Supplier
Micro-pulser Electroporator	Bio-Rad
VP 384F Replicator	V&P Scientific
VP 408FH Replicator	V&P Scientific
MicroMass LCT (ESI-TOF-LC/MS)	Waters
Fastprep120	Q-BIOgene
Trans-Blot® SD Semi-Dry Electrophoretic transfer cell	Bio-Rad
Mini-PROTEAN® 3 Cell	Bio-Rad
BiaCore3000	BiaCore Life Sciences
NTA Sensor Chip	BiaCore Life Sciences
Miniopticon	Bio-Rad
Symmetry300™ C4 column (3.5 µm x 2.1 µm x 50 mm)	Waters
BioSuite™ C18 PA-A column (3 µm x 2.1 µm x 250 mm)	Waters

Table 3.6 | Suppliers of reagents for media

Reagent	Supplier
Adenine	Alfa Aesar
Agar	BD Biosciences
Ampicillin, sodium salt	EMD
Chloramphenicol	Sigma-Aldrich
CSM, drop out supplements	Biol 101 Inc.
Dextrose/Glucose	BD Biosciences
Galactose	BD Biosciences
Kanamycin sulphate	Sigma-Aldrich
Peptone	BD Biosciences
Sucrose	EMD
Tryptone	BD Biosciences
X-Gal (5-bromo-4chloro-indoyl-£)-D-galactopyranoside)	Invitrogen
Yeast Extract	BD Biosciences
Yeast Nitrogen without Amino Acids	BD Biosciences

Table 3.7 | Oligonucleotides
Oligonucleotides were synthesized by Integrated DNA Technologies. **(a)** Reference number is a unique identifier from IDT and is used to reference the oligonucleotide in the text. **(b)** Oligonucleotide name. P1 and F represent forward primers and P2 and R represent reverse primers. **(c)** Oligonucleotide sequence in 5` to 3` orientation. Oligonucleotides are represented by one letter nucleotide codes, where N = (A, C, T, G) and K = (G, T). /5Phos/ indicates the primer is phosphorylated at the 5`-end.

Reference[a]	Name[b]	Sequence (5` – 3`)[c]
1	Intein Recomb P1	TAC CCT TAT GAT GTG CCA GAT TAT GCC TCT CCC GAA TTC ATG GTT AAG GTT ATT GGT AG
2	N	AAT TTC AGT ACC GAA AGA CAA AGC GTA TTC
3	O	GAA TAC GCT TTG TCT TCG GTA CTG AAA TT
4	J	TGA CCA AAC CTC TGG CGA AGA AGT CCA AAG CTT CTC GAG TTA CTT AAT AGT ACC AGC ATC CA
5	V	CCA ACT CCT AGA ATT GTG AGC AAT
6	P	CCG GAA TTC GGT TAA GGT TAT TGG TAG AA
7	Q	TTA TCT CGA GTT ACT TAA TAG TAC CAG CAT
8	U	ATT GCT CAC AAT TCT AGG AGT TGG
4342615	Intein Oligo 1	ATG GTT AAG GTT ATT GGT AGA AGA TCT TTG GGT GTT CAA AGA ATT TTC GAT ATT GGT TTG CCA CAA GAT CAC AAC TTC TT
4342617	Intein Oligo 3	TGT GAA TAC GCT TTG TCT TTC GGT ACT GAA ATT TTG ACT GTT GAA TAC GGT CCA TTG CCA ATT GGT AAG ATT GTT TCT GA
4342618	Intein Oligo 4	ATA GCT TGA GTG TAA ACT CTA CCT TCT GGA TCA ACA GAG TAA ACA GAA CAG TTA ATT TCT TCA GAA ACA ATC TTA CCA AT
4342619	Intein Oligo 5	AGA GTT TAC ACT CAA GCT ATT GCT CAA TGG CAC GAT AGA GGT GAA CAA GAA GTT TTG GAA TAC GAA TTG GAA GAT GGT TC

Reference[a]	Name[b]	Sequence (5` – 3`)[c]
4342620	Intein Oligo 6	ATA GCC AAC AAT TGG TAA TCA GTA GTC AAG AAT CTG TGA TCA GAA GTA GCT CTA ATA ACA GAA CCA TCT TCC AAT TCG TA
4342621	Intein Oligo 7	GAT TAC CAA TTG TTG GCT ATT GAA GAA ATT TTC GCT AGA CAA TTG GAT TTG TTG ACT TTG GAA AAC ATT AAG CAA ACT GA
4342622	Intein Oligo 8	CTT AAT AGT ACC AGC ATC CAA CAA TGG GAA TGG CAA TCT GTG GTT ATC CAA AGC TTC TTC AGT TTG CTT AAT GTT TTC
6446987	P1 Intein Library Primer	ATA TTG GTT TGC CAC AAG ATC ACA ACT TCT TGT TGG CTA ACG GTG CTA TTG CTC ACG CT
6446988	P2 Intein Library Primer	ATC TTA CCA ATT GGC AAT GGA CCG TAT TCA ACA GTC AAA ATT TCA GTA CCG AAA GAC AA
6446990	Lariat intein Library	GTG CTA TTG CTC ACG CTT CTN NKN NKN NKN NKN NKN NKN NKG AAT ACT GTT TGT CTT TCG GTA CTG AAA T
6513182	pEG202 Chk P1	GGG CTG GCG GTT GGG GTT ATT C
6513183	pEG202 Chk P2	CAT GCC GGT AGA GGT GTG GTC AA
6615380	Intein Recomb P2 (w/ter)	TGA CCA AAC CTC TGG CGA AGA AGT CCA AAG CTT CTC GAG TTA CTT AAT AGT ACC AGC ATC CA
6641791	Oligo 2 Intein Fix	AAA GAC AAA GCG TAT TCA CAC GGA CCG CAC GAC GAG TGA GCA ATA GCA CCG TTA GCC AAC AAG AAG TTG TGA TCT TGT GG
7917928	P2 Intein/XhoI	TTA TCT CGA GTT ACT TAA TAG TAC CAG CAT
8340694	P1 Intein/EcoRI+1	CCG GAA TTC GGT TAA GGT TAT TGG TAG AA
15951127	P2 Npu-DnaE-N/pIL	CCA AAC CTC TGG CGA AGA AGT CCA AAG CTT CTC GAG TTA GTT TGG CAA GTT ATC AAC TC
15951128	Npu-DnaE-N Yeast#1	TGT TTG TCT TAC GAA ACT GAA ATT TTG ACT GTT GAA TAC GGT TTG TTG CCA ATT GGT AAG ATT GTT GAA AAG AGA ATT GA

Reference[a]	Name[b]	Sequence (5` – 3`)[c]
15951129	Npu-DnaE-N Yeast#2	CAT TGA GCA ACT GGT TGA GTG TAA ATG TTA CCG TTG TTA TCA ACA GAG TAA ACA GTA CAT TCA ATT CTC TTT TCA ACA AT
15951130	Npu-DnaE-N Yeast#3	ACT CAA CCA GTT GCT CAA TGG CAC GAT AGA GGT GAA CAA GAAA GTT TTTC GAA TAC TGT TTG GAA GAT GGT TCT TTG ATT AG
15951131	Npu-DnaE-N Yeast#4	ATT TCA TCA ATT GGC AAC ATT TGA CCA TCA ACA GTC ATG AAC TTG TGA TCC TTA GTA GCT CTA ATC AAA GAAA CCA TCT TC
15951132	Npu-DnaE-N Yeast#5	ATG TTG CCA ATT GAT GAA ATT TTC GAA AGA GAA TTG GAT TTG ATG AGA GTT GAT AAC TTG CCA AAC
16415915	Random 10mer (InteinII)	GTG CTA TTG CTC ACG CTT CTN NKN NKN NKN NKN NKN NKN NKN NKN NKN NKT GTT TGT CTT ACG AAA CTG A
16415916	Random 5mer (InteinII)	GTG CTA TTG CTC ACG CTT CTN NKN NKN NKN NKN NKT GTT TGT CTT ACG AAA CTG A
16415917	Random 5mer(12aa) (InteinII)	GTG CTA TTG CTC ACG CTT CTB NTB NTB NTB NTB NTT GTT TGT CTT ACG AAA CTG A
16415918	P2 Library Amp/Npu	ATC TTA CCA ATT GGC AAC AAA CCG TAT TCA ACA GTC AAA ATT TCA GTT TCG TAA GAC AA
16646144	P1 Npu-DnaE-N/NruI/pIL	TTC TTG TTG GCT AAC GGT GCT ATT GCT CAC GCT TCG CGA TGT TTG TCT TAC GAA ACT GA
26911159	L2-Stop-Ssp-Ssp	ATT GCT CAC GCT TCT AGG AGT TGG GAT CTT CCT GGG GAA TAC TAA TAG TTG TCT TTC GGT
27171877	ScrambledL2-Ssp	ATT GCT CAC GCT TCT GAT CCT GGG CTT AGG AGT TGG GAA TAC TGT TTG TCT TTC GGT
27171878	InvertedL2-Ssp	ATT GCT CAC GCT TCT GGG CCT CTT GAT TGG AGT AGG GAA TAC TGT TTG TCT TTC GGT
27431776	L2-R2A	ATT GCT CAC GCT TCT GCT AGT TGG GAT CTT CCT GGG GAA TAC TGT TTG TCT TTC GGT

Reference[a]	Name[b]	Sequence (5` – 3`)[c]
27431777	L2-S3A	ATT GCT CAC GCT TCT AGG GCT TGG GAT CTT CCT GGG GAA TAC TGT TTG TCT TTC GGT
27431778	L2-W4A	ATT GCT CAC GCT TCT AGG AGT GCT GAT CTT CCT GGG GAA TAC TGT TTG TCT TTC GGT
27431779	L2-D5A	ATT GCT CAC GCT TCT AGG AGT TGG GCT CTT CCT GGG GAA TAC TGT TTG TCT TTC GGT
27431780	L2-L6A	ATT GCT CAC GCT TCT AGG AGT TGG GAT GCT CCT GGG GAA TAC TGT TTG TCT TTC GGT
27431781	L2-P7A	ATT GCT CAC GCT TCT AGG AGT TGG GAT CTT GCT GGG GAA TAC TGT TTG TCT TTC GGT
27431782	L2-G8A	ATT GCT CAC GCT TCT AGG AGT TGG GAT CTT CCT GCT GAA TAC TGT TTG TCT TTC GGT
27431783	L2-E9A	ATT GCT CAC GCT TCT AGG AGT TGG GAT CTT CCT GGG GCT TAC TGT TTG TCT TTC GGT
27431784	L2-Y10A	ATT GCT CAC GCT TCT AGG AGT TGG GAT CTT CCT GGG GAA GCT TGT TTG TCT TTC GGT
30088469	pJG4-5 Chk P1	GGA CAG GAG ATG CCG ATG GA
30088470	pJG4-5 Chk P2	GCA AGG TAG ACA AGC CGA CAA C
33903878	L2 HG6A NG7D For	/5Phos/GGT GCT ATT GCT GCT GAT TCT AGG AGT TGG
33903879	L2 HG6A NG7Y For	/5Phos/GGT GCT ATT GCT GCT TAT TCT AGG AGT TGG
33903880	L2 HG6A NG7K For	/5Phos/GGT GCT ATT GCT GCT AAA TCT AGG AGT TGG
33903881	L2 HG6H NG7Q For	/5Phos/GGT GCT ATT GCT CAC CAA TCT AGG AGT TGG
33903882	L2 HG6L NG7N For	/5Phos/GGT GCT ATT GCT TTG AAT TCT AGG AGT TGG
33903883	L2 HG6N NG7N For	/5Phos/GGT GCT ATT GCT AAT AAT TCT AGG AGT TGG
33903884	L2 HG6D NG7N For	/5Phos/GGT GCT ATT GCT GAT AAT TCT AGG AGT TGG
33903885	Intein G2 Rev	/5Phos/GTT AGC CAA CAA GAA GTT GTG ATC TTG
33903886	Ssp-RB11L For	/5Phos/GCT ACT TCT GAT CAC TTG TTC TTG ACT ACT GAT

Reference[a]	Name[b]	Sequence (5` – 3`)[c]
33903887	Ssp-RB11Y For	/5Phos/GCT ACT TCT GAT CAC TAT TTC TTG ACT ACT GAT
33903888	Ssp-RB11E For	/5Phos/GCT ACT TCT GAT CAC GAA TTC TTG ACT ACT GAT
33903889	Ssp-RB11X For	/5Phos/GCT ACT TCT GAT CAC NNK TTC TTG ACT ACT GAT
33903890	Ssp-B11 Rev	/5Phos/TCT AAT AAC AGA ACC ATC TTC CAA TTC
33903891	L2 HG6X NG7X Lib	TTG GCT AAC GGT GCT ATT GCT NNK NNK TCT AGG AGT TGG GAT CTT CCT GGG GAA TAC TGT TTG TCT TTC GGT ACT GAA
33903892	P1 Ssp Intein Lib Amp/G6	ATT TTC GAT ATT GGT TTG CCA CAA GAT CAC AAC TTC TTG TTG GCT AAC GGT GCT ATT GCT
38891611	P1 lexA LexA-Chip	TAT GGT CGC ATT TTG GAT AAC
38891612	P2 lexA LexA-Chip	ATA CCT GTC TGG CTG ATG TG
38891615	P1 yabN (sgrR) LexA-Chip	CTA ATC TGG TCT ATT TCG CTG
38891616	P2 yabN (sgrR) LexA-Chip	TCT CTC CTG TAT GCC ACT G

3.1.2 Synthetic peptides

The biotin-labeled linear L2 peptide was synthesized by GenScript. The biotin-labeled L2 lactone peptide was synthesized by Protein Peptide Research. The lactone peptide was synthesized on 2-chlorotrityl resin using serine(trityl) as a protecting group for the serine involved in the side chain cyclization. The side chain and N-terminal protected peptide was cleaved from the chlorotrityl resin using 20% trifluorethanol/dichloromethane. The trityl group was then removed from the side chain of serine using 1% trifluoroacetic acid/dichloromethane. The peptide was solubilized in a small quantity of dimethylformamide and cyclized using benzotriazol-1-yloxy Tris(dimethylamino)phosphonium hexafluorophosphate/diisopropylethyl-amine in dichloromethane. Once cyclized, the remaining protecting groups were removed using 95% trifluoroacetic acid scavengers.

3.1.3 Strains

Table 3.8 | *S. cerevisiae* strains and genotypes

Strain	Genotype	Reference:
EGY2	MATa, ura3, his3, trp1, leu2	(Cohen *et al.*, 1998)
EGY48	MATα, 6LexA-LEU2, ura3-52, leu2, his3,trp1, GAL+	(Cohen *et al.*, 1998)
EY93	MATa, ura2, his3, trp1, leu2, ade2::URA3	This work
EY111	MATα, his3, trp1, ura3::LexA8op-LacZ, ade2::URA3-LexA8op-ADE2, leu2::LexA6op-LEU2	This work

Table 3.9 | *E. coli* strains and genotypes

Strain	Marker	Genotype	Source
BL21-CodonPlus®(DE3)-RIL	Chloramphenicol	*F- ompT gal dcm Ion* $hsd_B(r_B^- m_B^-)$ *lambda(DE3(lacI lacUV5-T7 gene 1 ind1 sam7 nin5)*	Stratagene
BL21(DE3)		*F- ompT gal dcm Ion* $hsd_B(r_B^- m_B^-)$ *lambda(DE3(lacI lacUV5-T7 gene 1 ind1 sam7 nin5)*	Novagen
MC1061	Streptomycin	*F*$^-$ *araD139 Δ(ara*A-*leu)*7697 *galE15 galK16 Δ(lac)X74 rpsL (Str*^{r}*) hsdR2* $(r_K^- m_K^+)$ *mcrA mcrB1*	(Wertman *et al.*, 1986)
XL1-Blue	Tet^r (Tetracyclin)	*rec*A1 *end*A1 *gyr*A96 *thi*-1 *hsd*R17 *sup*E44 *rel*A1 *lac* (F´ *pro*AB *lac*I^{q}ZΔM15 Tn10 (Tet^r))	Stratagene
ER2925	cam^r (Chloramphenicol)	***E. coli* K12:** *ara-14 leuB6 fhuA31 lacY1 tsx78 glnV44 galK2 galT22 mcrA dcm-6 hisG4 rfbD1 R(zgb210::Tn10)TetS endA1 rpsL136 dam13::Tn9 xylA-5 mtl-1 thi-1 mcrB1 hsdR2*	New England Biolabs

3.1.4 Plasmids

3.1.4.1 Yeast plasmids

pEG202 (Figure 3.1) and pJG4-5 (Figure 3.2) were used a bait and prey expression plasmids, respectively, in the Y2H assay.

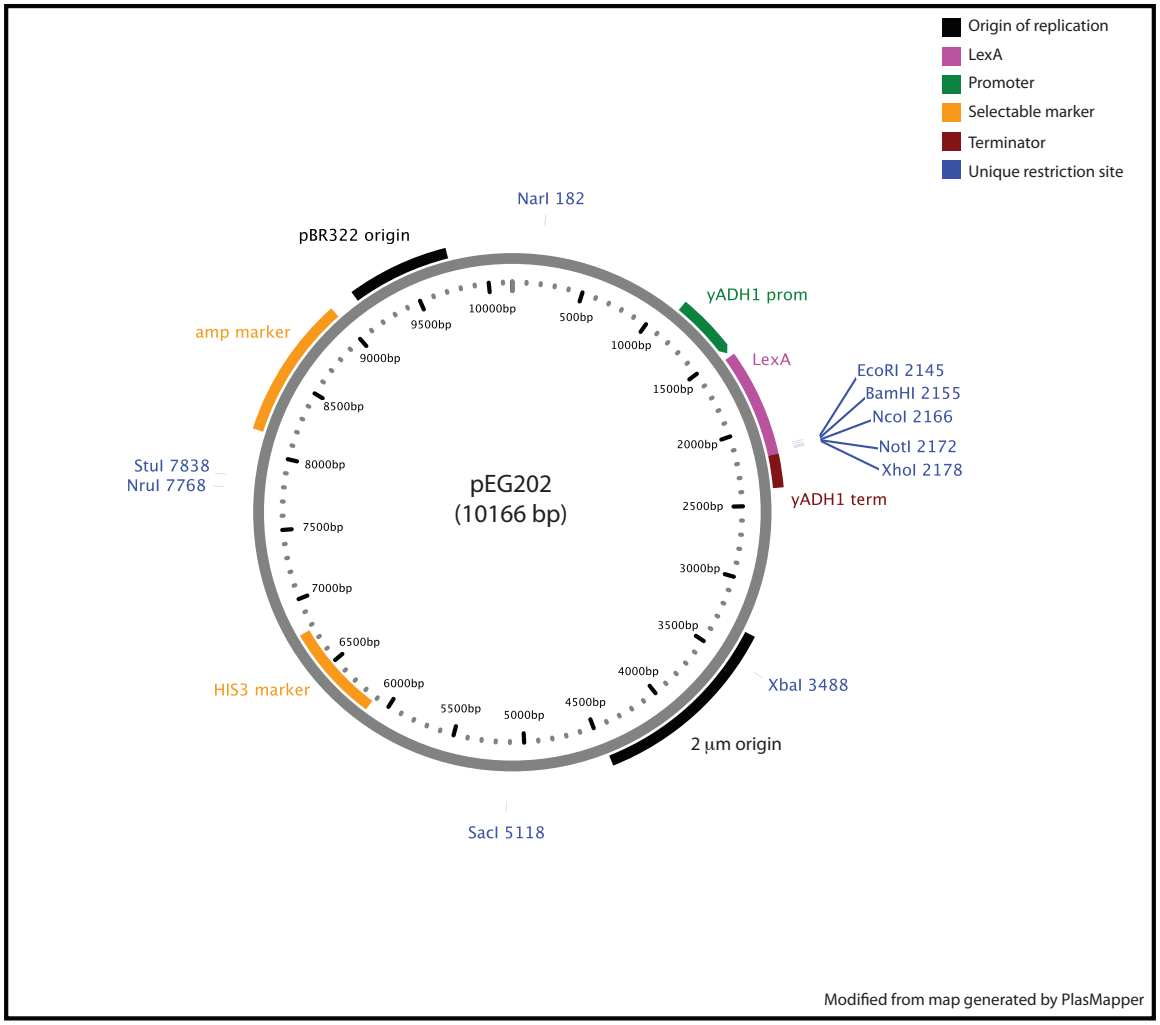


Figure 3.1 | pEG202

pEG202 (GenBank accession U89960) is the bait plasmid for the Y2H assay. Bait protein expression is controlled by the alcohol dehydrogenase promoter (yADH1 Prom). The multiple cloning site contains five unique restriction enzyme sites that allow genes to be cloned as C-terminus fusions to the DNA binding domain LexA. The yeast alcohol dehydrogenase terminator (yADH1 term) is used to terminate transcription. A HIS3 gene is used to maintain the plasmid in yeast. The 2 µm origin of replication is used to maintain the plasmid in a high copy number in yeast. The ampicillin marker (amp marker) is used for selection and the pBR322 origin is used to maintain the plasmid in a high copy number in *E. coli*.

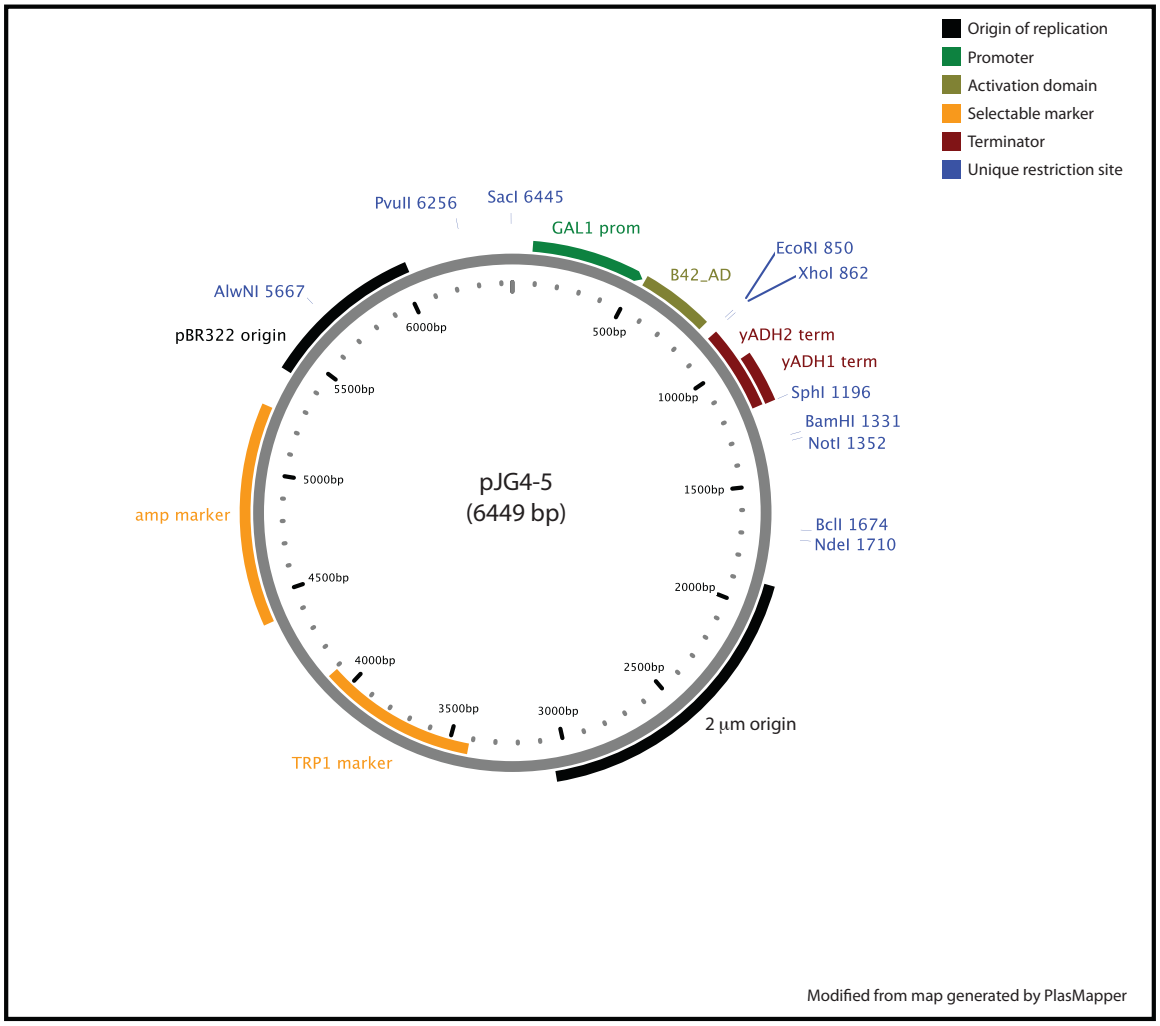


Figure 3.2 | pJG4-5

pJG4-5 (GenBank accession U89961) is the prey plasmid for the Y2H assay. Prey protein expression is controlled by a galactose promoter (GAL1 Prom). E*coR*I and *Xho*I restriction sites are used to clone genes that are C-terminal fusions to the fusion tag consisting of the haemagglutinin tag (HA), nuclear localization sequence (NLS), and B42 activation domain (B42_AD). The yeast alcohol dehydrogenase terminator (yADH1 term) is used to terminate transcription. A TRP1 gene is used to maintain the plasmid in yeast. The 2 µm origin of replication is used to maintain the plasmid in a high copy number in yeast. The ampicillin marker (amp marker) is used to select and the pBR322 origin is used to maintain the plasmid in a high copy number in *E. coli*.

3.1.4.2 *E. coli* plasmid

pET28b was used to express proteins in *E. coli* (Figure 3.3).

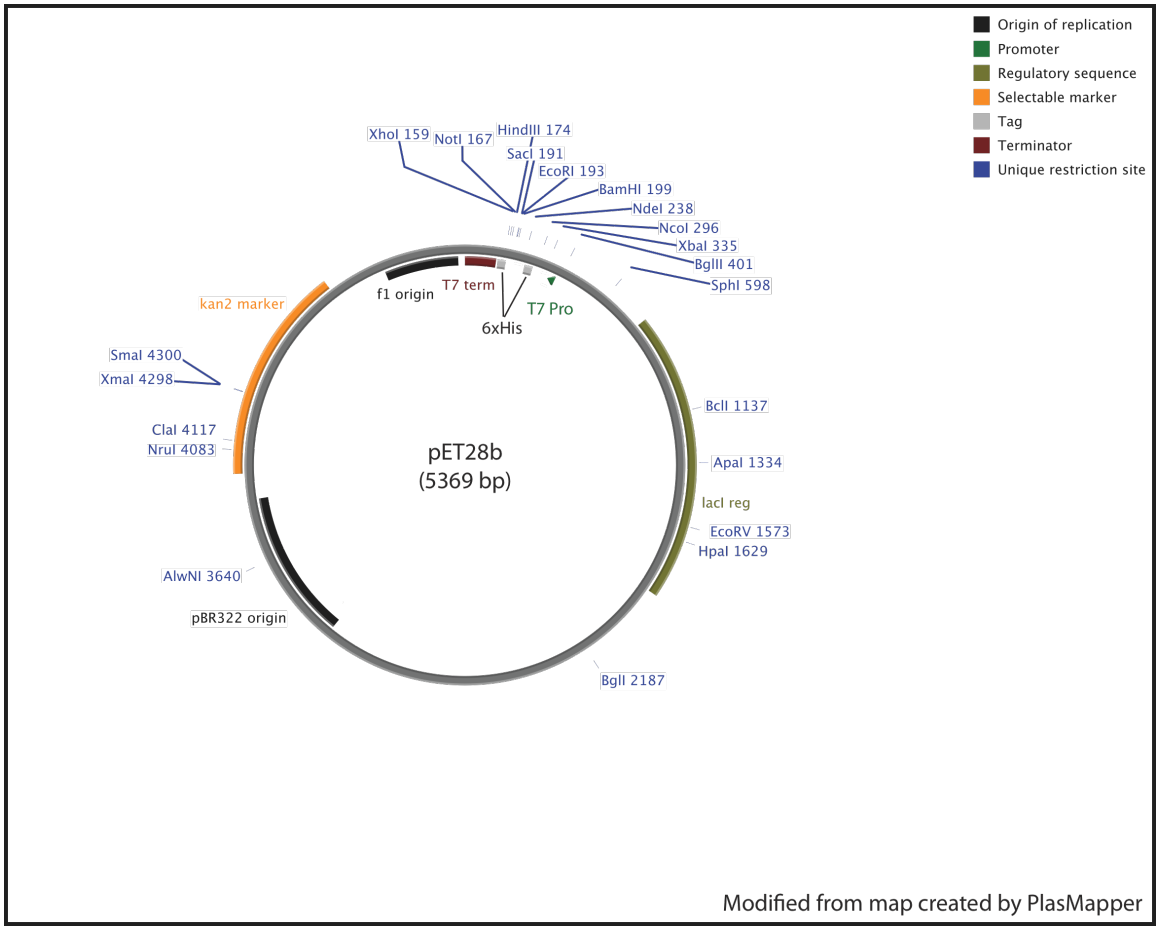


Figure 3.3 | pET28b

pET28b was used to express proteins in *E. coli*. The T7 promoter (T7 pro) is an IPTG inducible promoter. Proteins can be expressed with a C-terminal six-histidine sequence tag (6xHis), an N-terminal 6xHis tag or both. A T7 terminator sequence is used terminate transcription. A kanamycin resistance gene (kan2 marker) is used for plasmid selection in *E. coli*. The pBR322 origin of replication allows for high copy plasmid replication in *E. coli*. The F1 origin of replication (F1 ori) allows single-stranded DNA (ssDNA) production.

3.1.5 Sodium dodecyl sulfate polyacrylamide gel electrophoresis

Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) using the Laemmli buffer system (Laemmli, 1970) was used to separate proteins based on their apparent molecular weight. A 4% stacking gel (0.125 M Tris-Cl pH 8.8, 0.1% SDS, 4% degassed acrylamide/bis-acrylamide (37.5:1 ratio), 0.1% TEMED (v/v), and 0.05% (w/v) ammonium

persulphate) was used for all gels. A resolving gel of appropriate concentration (0.375 M Tris-Cl pH 8.8, 0.1% SDS, 8 – 20% degassed acrylamide/bis-acrylamide 37.5:1 ratio, 0.05% TEMED (v/v) and 0.05% (w/v) ammonium persulphate) was used to separate proteins. Proteins were suspended in 2xSDS loading dye (62.5 mM Tris-HCl pH 6.8, 25% (v/v) glycerol, 2% (w/v) SDS, 0.01% (w/v) bromophenol blue, 5% (v/v)). Beta-mercaptoethanol was added to the loading buffer when proteins were separated under reducing conditions. Samples were heated at 95 °C for five minutes and then cooled on ice prior to loading. Either 1.5 mm (W) or 0.75 mm (W) x 8 cm (L) x 7.3 cm (H) gels were used. Gels were run for fifteen minutes at a constant voltage of 100 V followed by an additional 35 – 45 minutes at 200 V. The running buffer consisted of (25 mM Tris-HCl pH 8.3, 190 mM glycine, and 0.1% (w/v) SDS) approximately 500 mL was used per gel apparatus.

3.1.6 Coomassie-staining

Coomassie-staining was used to visualize proteins after SDS-PAGE. Fixing solution (25% (v/v) isopropanol, 10% (v/v) acetic acid) was applied to SDS-PAGE gels for 15 minutes to immobilize proteins. Gels were then incubated in coomassie blue staining solution consisting of 10% (v/v) acetic acid, and 0.012% (w/v) bromophenol blue for a minimum of four hours. Gels were then destained in 10% (v/v) acetic acid until bands were clearly visible. Gels were stored in water until photographed using a GelDoc imager or using the 680 nm channel on the Licor Odyssey infrared imaging system.

3.1.7 Western blot analysis

Western blot analysis was used to visualize specific proteins after SDS-PAGE. Nitrocellulose (0.45 μm) and blotting paper were incubated in transblot buffer (48 mM Tris-HCl pH 8.3, 39 mM glycine, 20% (v/v) methanol, 0.04% (w/v) SDS) for 10 minutes. A piece of pre-soaked blotting paper was placed on the Trans-Blot® SD Semi-dry electrophoretic transfer cell followed by the nitrocellulose membrane, the gel, and a second piece of pre-soaked blotting paper. Transblotting was performed at 15 V for 45 minutes. The nitrocellulose membrane was blocked for one hour at room temperature with 10 mL Odyssey blocking buffer. Primary antibody was added at the appropriate dilution in 10 mL Odyssey blocking buffer + 0.1% Tween-20, and incubated for one hour at room temperature or overnight at 4 °C. Membranes were washed three times with approximately 20 mL PBS-T (Phosphate buffered saline supplemented with Tween-20) (10 mM $Na_2HPO_4 \cdot 7H_2O$, 17.6 mM KH_2PO_4, pH 7.4, 137

mM NaCl, 26.8 mM KCl, 0.1% (v/v) Tween 20)) and then incubated with secondary antibody. Goat-anti-mouse or goat-anti-rabbit (1:15,000 dilution) secondary antibodies conjugated to near-infrared dyes IRDye 680 nm or IRdye 800CW were incubate with the membrane for one hour at room temperature. Membranes were washed three times with 20 mL PBS-T, then once in PBS without Tween 20. Membranes were stored in PBS without Tween 20 or dried and visualized using the Licor Odyssey infrared imaging system.

3.1.8 Agarose gel electrophoresis

To determine the size or concentration of DNA, samples were mixed with 6x loading buffer (50% glycerol, 0.2 M EDTA pH 8.3, 0.05% bromophenol blue) to a final concentration of 1x. Samples were resolved in an agarose gel consisting of 0.5 to 1.5% (w/v) ultrapure agarose in 1X SB Buffer (10 mM sodium pH 8.5 with boric acid) (Brody *et al.*, 2004) and 0.5 µg/mL ethidium bromide. Gels were run at a constant voltage between 200 - 500 V for seven to thirty minutes depending on the separation required. For rapid analysis of PCR products, the agarose gels were run at 500 V for seven minutes. Plasmids requiring more separation were run at 200 V for thirty minutes. Standard conditions were typically 300 V for ten minutes. Gels were visualized and photographed using a UV light transilluminator.

3.1.9 DNA gel purification and extraction

Gel purification was used to clean-up PCR products and restriction enzyme digested plasmids. Gel purification was performed using a Gel Purification kit as per manufacture's instructions. PCR products were purified using a PCR Clean-up kit as per the manufacture's instructions. DNA was concentrated by alcohol precipitation if necessary (Moore and Dowhan, 2002).

3.1.10 DNA sequencing

Sequencing was either performed using an ABI Prism 310 or sent to the Plant Biotechnology Institute (PBI), National Research Council of Canada (NRC). For sequencing using the ABI Prism 310 Genetic analyzer, the ABI prism big-dye terminator cycle sequencing ready reaction kit v1.1 was used according to manufacture's directions. Unextended primers and the unincorporated dye were removed using a DyeEx 2.0 spin kit (Qiagen). Samples were injected at 4 V for 15 – 30 seconds onto a GA310 capillary column (47 cm length, with internal diameter 50 µm, ABI) the column was filled with POP6 polymer (Performance Optimized Polymer 6, ABI) and heated to 50 °C. Samples were run at 15 V for 36 – 60 minutes depending

on the length of sequence data required. For sequencing outsourced to the Plant Biotechnology Institute (PBI), National Research Council of Canada plasmids were diluted to 0.05 µg/µL and primers were diluted to a concentration of 3.2 pmol/µL. For direct sequencing from *E. coli*, single colonies were grown overnight in a 96-well u-bottom plate at 37 °C in 200 µL of LB supplemented with the appropriate antibiotic and 10% glycerol. The plate was sealed and frozen at – 80 °C. Primers (30088469) and (30088470) were used to sequence inserts in pJG4-5 and lariat plasmids (Section 3.3.8). Primers (6513182) and (6513183) were used to sequence inserts in pEG202.

3.1.11 Statistical analysis

Unless otherwise noted, data was reported as the mean ± standard deviation. P-values were calculated using a two-tailed t-test with Prism 4.0c for Macintosh (Graphpad).

3.2 Polymerase chain reactions (PCR)

3.2.1 High-fidelity PCR cloning reaction

PCR reactions were performed in 50 µL reaction containing 60 mM Tris-SO_4 (pH 8.9), 180 mM $(NH_4)_2SO_4$, 1.5 mM $MgSO_4$, 200 µM dNTPs, 1 µM of forward and reverse primers, 100 ng template DNA, and 1 Unit/50 µL of Platinum® Taq DNA Polymerase High Fidelity. PCR products were amplified with a two minute initial denaturation step at 95 °C followed by 20 cycles of 95 °C for thirty seconds, 50 °C for thirty seconds, 68 °C for one minute per kilobasepair. A final extension step of seven minutes at 68 °C was included for products over one kilobasepair in length.

3.2.2 Low-fidelity PCR

Low-fidelity PCR was used to check plasmids for the correct insert, cloning of small DNA fragments, and for generating libraries. A standard PCR reaction was performed in 50 µL reaction containing 1x *Taq* Buffer (10 mM Tris-HCl pH 8.3, 50 mM KCl, 1.5 mM $MgCl_2$), 200 µM dNTP, 1 µM Primers, 10 - 300 ng template, and 1 µL of *Taq*. DNA template was denatured with a five minute denaturation step at 95 °C, followed by 20 – 30 cycles of amplification. Each cycle contained a denaturation step at 95 °C for thirty seconds, an annealing step at 50 °C for thirty seconds, and an extension step at 72 °C for one minute per kilobasepair of DNA.

3.2.3 *E. coli* colony PCR reaction

Colony PCR from *E. coli* was performed to verify the construction of a clone or insert before subsequent plasmid purification. A single colony was picked from an LB plate containing the appropriate antibiotic with a pipette tip. The colony was first dipped into a 50 µL PCR reaction mix containing 1X PCR buffer (10 mM Tris-HCl pH 8.3, 50 mM KCl, 1.5 mM $MgCl_2$), 200 µM dNTPs, and 1 µM of each primer. An initial denaturation step of ten minutes at 95 °C was used to lyse the cells, followed by 30 cycles. Each cycle consisted of a denaturation step at 95 °C for thirty seconds, an annealing step at 52 °C for thirty seconds, and an extension step at 72 °C for one minute per kilobasepair. The pipette tip, after being dipped in the PCR reaction mix, was placed in 5 mL of media containing the appropriate antibiotic selection for subsequent plasmid purification. Alternatively, 1 – 5 µL of an overnight culture was used as template instead of a colony in the PCR reaction mix.

3.2.4 Yeast colony PCR reaction

A fresh yeast colony was picked from a plate, resuspended in 20 µL of 0.02 N NaOH, heated at 95 °C for five minutes, and centrifuged at 13,000 x g. Alternatively, 1.1 x 10^7 cfu of yeast cells were resuspended in 50 µL of 0.02 N NaOH and heated at 95 °C for five minutes and centrifuged. One-tenth of the PCR reaction volume was used as a template for the PCR reaction.

3.3 General yeast protocols

3.3.1 Yeast media

Media was prepared as described by Geyer and Brent (Geyer and Brent, 2000). Media containing dextrose/glucose was diluted from an autoclaved stock solution of 40% (w/v) dextrose. Synthetic media was defined by an S, followed by the type of sugar, which was followed by any modifications to the media. A negative sign indicates a dropout or removal of an amino acid or nucleotide. A plus sign indicates a media supplement. For example, a typical prey selection plate was SD W-, meaning the synthetic media was supplemented with dextrose (D) and lacked tryptophan (W-). A typical Y2H full selection plate was SGR H-W-L-A- X-gal+, indicating that the synthetic medium contained galactose/raffinose, the chromagenic substrate x-gal and lacked the amino acids histidine, tryptophan, and leucine, and the nucleotide adenine.

YPDA medium was prepared with 1% (w/v) yeast extract, 2% (w/v) peptone, 80 mg/L adenine, and 2% (w/v) dextrose in ddH_2O (Filtered and sterilized water). For solid media, 2% (w/v) agar was added.

Synthetic medium was prepared with 0.67% (w/v) yeast nitrogen base without amino acids and was supplemented with complete supplemental medium (CSM) lacking the appropriate amino acid(s) (Bio 101). 2% (w/v) dextrose or 2% (w/v) galactose was supplemented with 1% (w/v) raffinose or sucrose in ddH_2O was used as the carbon source. For solid medium, 2% (w/v) agar was added.

X-Gal plates were prepared like the synthetic plates except they were supplemented with BU salts and 80 µg/L of 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal). X-gal solution was prepared by dissolving 20 mg/mL of X-gal in dimethyl formamide to a final concentration of 80 µg/mL. 10xBU salts were prepared by dissolving 70 g of $Na_2HPO_4 \cdot 7H_2O$, 30 g of NaH_2PO_4 in 900 mL of ddH_2O and the pH of the 10xBU salts was adjusted to 7.0.

3.3.2 Propagation and manipulation of yeast

Standard techniques were used to culture and propagate yeast (Geyer and Brent, 2000). Liquid cultures were grown at 30 °C with shaking at 250 rpm. Cultures on solid medium were inverted and grown at 30 °C.

3.3.3 Lithium acetate transformation

Transformation procedures were used to transfer plasmids into yeast or to clone DNA fragments into plasmids by *in vivo* homologous recombination (Hua *et al.*, 1997). Yeast frozen competent cells were prepared according to the procedure reported by Gietz and Schiestl, (Gietz and Schiestl, 2007). Transformation of yeast cells in 96-well format was performed according to Gietz and Schiestl (Gietz and Schiestl, 2007). The high efficiency low throughput method (Gietz *et al.*, 1995; Geyer and Brent, 2000) was used for library generation since it gave the highest number of transformants. It was also used for routine cloning, if competent frozen cells were not available. Lithium acetate competent yeast cells were prepared from a 10 mL overnight culture in the appropriate media and grown at 30 °C with shaking. Using the conversion factor of 1.1×10^7 CFU/mLxOD_{600}, the culture was diluted to 5×10^6 CFU/mL in 50 mL of appropriate media, which provided enough cells for ten 50 µL transformations. The cells were grown to an OD_{600} of 0.6 – 0.8. Cells were collected by centrifuging at 4,000 x g at room temperature and washed once in 25 mL ddH_2O. The supernatant was removed from the

yeast pellet and the pellet was resuspended in 1 mL of 100 mM LiOAc. The sample was centrifuged for one minute at 4,000 x g and resuspended in 500 µL of 100 mM LiOAc and aliquoted into ten 50 µL aliquots. The aliquots were centrifuged again and the supernatant was removed. The pellet was overlaid with 240 µL polyethylene glycol (50% (w/v) PEG-3500), 36 µL of 1 M LiOAc, 25 µL of 2 mg/mL single-stranded DNA, plasmid DNA, and PCR insert (optional). The mixture was vortexed rapidly to resuspend the pellet and then incubated for thirty minutes at 30 °C and then heat shocked at 42 °C for fifteen minutes. The sample was centrifuged at 4,000 x g and the supernatant was removed, the pellet was resuspended in ddH_2O and plated on the appropriate media.

3.3.4 Preparation of plasmid DNA from yeast

3.3.4.1 Yeast mini-preparation

Plasmids were isolated according to protocols described previously (Hoffman and Winston, 1987; Geyer and Brent, 2000). Briefly, a single yeast colony was grown overnight in 2 mL of appropriate synthetic amino acid dropout media in a shaking incubator at 30 °C. Cells were harvested by centrifugation at 18,000 x g for thirty seconds. The cell pellet was resuspended in 200 µL of breaking buffer (2% (v/v) Triton X-100, 1% (v/v) SDS, 100 mM NaCl, 10 mM Tris-HCl pH 8.0, 1 mM EDTA) followed by the addition of 300 µg of glass beads and 200 µL of phenol-chloroform-isoamyl alcohol (25:24:1, v/v/v). Yeast cell walls were disrupted by vortexing the mixture for two minutes. The mixture was centrifuged for ten minutes at 18,000 x g and 50 µL of the aqueous layer was stored at – 20 °C. Typically 3 - 5 µL of the aqueous solution was used for transforming *E. coli* cells by electroporation.

3.3.5 Preparation of proteins from yeast

3.3.5.1 Standard protein isolation method

A single yeast colony was grown to saturation overnight in liquid media at 30 °C with shaking. The cells were centrifuged at 3,220 x g for five minutes at room temperature and washed in ddH_2O. The cells were induced in 10 mL of appropriate media at 30 °C for eight hours and collected by centrifugation at 3,220 x g for five minutes. Cells were washed in ddH_2O and resuspended in 300 µL of disruption buffer (20 mM Tris-HCl pH 7.0, 300 mM $(NH_4)_2SO_4$, 10 mM $MgCl_2$, 1 mM EDTA, 5% glycerol), supplemented with 1 mM PMSF (phenylmethylsulphonyl fluoride), 1 mM DTT, and 0.3 g acid-washed glass beads (425-600 µm diameter). Cells were ruptured using a Fastprep120 at setting 6.0 for twenty seconds at 4 °C.

The disruption was repeated three times. The cell lysate was centrifuged twice at 4 °C at 18,000 x g for fifteen minutes and the clarified supernatant was used for subsequent analyses (Dunn and Wobbe, 2001). Twenty microliters was typically used for Western blot analysis.

3.3.5.2 NaOH protein isolation method

Cell lysates were prepared using NaOH (Kushnirov, 2000). A single yeast colony was incubated overnight in 2 mL of the appropriate media in a shaking incubator at 30 °C. The following day, 1 mL of the culture was centrifuged for one minute at 13,000 x g. The pellet was resuspended in 100 µL of ddH_2O followed by the addition of 100 µL of 0.2 M NaOH. The mixture was incubated for five minutes and then centrifuged for one minute at 13,000 x g. The pellet was resuspended in 50 µL of 1xSDS sample buffer (0.06 M Tris-Cl pH 6.8, 5% glycerol, 2% SDS, 4% beta-mercaptoethanol, 0.0025% bromophenol blue) and boiled for three minutes. Ten microliters of cell lysate was used for SDS-PAGE or Western blot analysis.

3.3.6 Library mating

Target proteins were cloned into the pEG202 bait plasmid, which has the auxotrophic selection marker histidine. The bait plasmid was transformed into EY111 and grown in 1 L of SD H-A+ to an OD_{600} of 0.6 – 1.0. The number of bait cells was calculated by the optical density (OD) using the conversion factor of one OD_{600} equal to 1.1×10^7 cfu/mL.

Lariats were cloned into the pIN01 or pIL500 prey plasmids. Prey plasmids had the tryptophan auxotrophic selection marker. Prior to mating, the titer of the lariat library was predetermined by plating serial dilutions of the library onto SD W- plates. Sufficient library cells were diluted in 50 mL of SD W- media to obtain approximately 20 library copies. The library was grown for 2 – 4 hrs at 30 °C and the OD_{600} was determined.

The bait and prey cells were mixed at a bait to prey ratio of 20:1 and spread on 100 x 15 mm YPDA plates using glass beads. Plates were incubated at 30 °C for 24 – 48 hours. The mated cells were scraped off plates using a sterile glass microscope slide into a 50 mL Falcon tube. The cells were washed three times by resuspending them in 50 mL of PBS and then pelleting the cells by centrifugation at 1,000 x g. After washing, the pellet was resuspended in glycerol freeze down solution (25 mM Tris-HCl pH 8.0, 65% glycerol, 0.1 M $MgSO_4$) and stored at – 80 °C. A sample of the library was thawed and serial dilutions were plated onto SD W-, SD H-, and SD H-W-. These plates were used to determine the diploids per microliter and mating efficiency. The diploids per microliter were calculated directly by

measuring the CFU/µL on the SD H-W- plates. The mating efficiency was calculated by dividing the number of CFU/µL on the SD W- plates by the number of diploids or CFU/µL on the SD H-W-. In addition, a trial screen using 1 µL, 10 µL, and 100 µL of library was performed on SGR H-W-L-A- X-gal plates to estimate the number of interactions/µL and to determine the optimal cell density to perform the Y2H screens.

3.3.7 Y2H screening

An aliquot of the mated library cells was thawed and incubated at 30 °C in 50 mL of SGR H-W- to allow the cells to recover and start expressing the bait and prey. Cells were then concentrated to the appropriate dilution and plated using glass beads onto SGR H-W-L- plates and incubated at 30 °C for five days. Typically, cells were plated at a density of 100,000 to 1,000,000 diploids/plate. Five replicates of a serial dilution of the library were also plated onto SD H-W- to determine the number of diploids/µL, which was used to determine the total number of diploids screened.

Colonies that formed on the SGR H-W-L- were replica plated onto SGR H-W-A- X-gal after five days using velvet and a replicator block. Plates were then incubated at 30 °C for an additional five days. During this time, colonies that grew and turned blue, were circled on the plate and numbered according to which day they were visible on the plate. Colonies that had appeared earliest and had the best selection characteristics, were picked off the selection plates and incubated overnight in SGR H-W-L-A- liquid selection medium. These initial positive Y2H assay colonies were mixed 50:50 with glycerol freeze down solution (25 mM Tris-HCl pH 8.0, 65% glycerol, 0.1 M $MgSO_4$) and stored at –80 °C for further analysis.

3.3.8 Rechecking Y2H interactions

Plasmids were isolated from yeast cells showing a positive Y2H interaction using the yeast mini-prep protocol and were transformed into *E. coli* by electroporation. Three to five individual colonies from each *E. coli* transformation were selected and plasmids were extracted using the *E. coli* mini-prep protocol and transformed into EY93 using the lithium acetate transformation protocol. One yeast colony from each transformation was mated to the appropriate target as well as a negative control plasmid containing the empty bait plasmid (pEG202). Yeast were mated on YPDA for 24 hours. Diploids were selected on SD H-W- media and plated on SGR H-W-L-, SGR H-W-L-A-, SGR H-W-L-X-gal, and SGR H-W-L-A-X-gal to estimate and rank the strength of the interactions. Diploids were also plated on SD H-

W-L-X-gal plates to confirm that the Y2H interaction was dependent on prey expression. Plasmids encoding lariats that showed an interaction that was dependent on prey expression were isolated using yeast mini-prep protocol and transformed into *E. coli* by electroporation, amplified, and purified. Plasmids were sequenced to determine the sequence of the interacting lariat.

3.4 General *E. coli* protocols

3.4.1 *E. coli* media

Lysogeny broth (LB) also commonly mistakenly referred to as Luria broth in the literature (Bertani, 2004), was prepared with 1% tryptone, 0.5% yeast extract, 85.6 mM NaCl, and 1 mM NaOH in ddH_2O. The appropriate antibiotic was added after the media cooled to 55 °C. Solid media contained 2% agar (Bertani, 2004; Bertani, 1951).

Ampicillin was used at a final concentration of 100 µg/mL. Kanamycin was used at a final concentration of 50 µg/mL. Chloramphenicol was used for plasmid selection at a final concentration of 25 µg/mL.

3.4.2 Propagation and manipulation of *E. coli*

Standard techniques were used to culture and propagate *E. coli* (Elbing and Brent, 2002). Liquid cultures were grown at 37 °C with shaking at 250 rpm. Cultures on solid medium were inverted and grown overnight at 37 °C.

3.4.3 Preparation of plasmid DNA from *E. coli*

Standard techniques were used to isolate small quantities of plasmid in single or 96-well format from *E. coli* using the alkaline lysis procedure (Engebrecht *et al.*, 2001). Large scale preparations of plasmids were performed using alkaline lysis and PEG precipitation (Heilig et al., 2001). Plasmids were stored in the TE (10 mM Tris-HCl pH 7.5, 1 mM EDTA). Plasmid DNA was also purified using a modified alkaline lysis protocol as described by Qiagen for use of their plasmid mini-preparation kits.

3.4.4 Transformation of plasmids into *E. coli*

3.4.4.1 Preparation of electrocompetent *E. coli*

E. coli cells were prepared for electroporation using standard protocols (Seidman et al., 2001). Briefly, the *E. coli* strain was struck out on the appropriate media to obtain single colonies. A single colony was transferred to 5 mL of LB and grown overnight at 37 °C in a

shaking incubator. The overnight culture was diluted in 500 mL of LB to an OD_{600nm} of 0.6. All subsequent steps were performed at 4 °C on ice using pre-chilled tubes. The cells were centrifuged at 4,000 x g for twenty minutes at 4 °C and washed twice in 250 mL of ddH_2O. The cells were resuspended in 20 mL of ice cold sterile water, transferred to a 50 mL Falcon tube, and centrifuged for ten minutes at 4200 x g at 4 °C. The pellet was resuspended in 50 mL of 10% glycerol and centrifuged at 4,200 x g at 4 °C. The pellet was resuspend in one pellet volume of ice-cold 10% (v/v) glycerol. Fifty-microliter aliquots were stored at −80°C.

3.4.4.2 Electroporation

E. coli cells were transformed using electroporation (Seidman et al., 2001). Briefly, 1 µL of plasmid at a concentration of 50-150 ng/µL was mixed with 50 µL of competent cells. The mixture was transferred to an ice-cold electroporation cuvette. Cells were then electroporated with a field strength of ~ 12.5 kV/cm (Ec2 BioRad). Electroporated cells were mixed with 500 µL of LB and incubated at 37 °C for thirty minutes in a shaking incubator. Cells were then plated onto LB agar plates containing the appropriate antibiotic.

3.5 Lariat construction and characterization

3.5.1 *Ssp-Ssp* intein gene construction

The lariat gene was synthetically constructed and inserted into pJG4-5 resulting in the plasmid pIN01, which expresses an inactive *Synechocystis spp. strain PCC6803* (*Ssp*) *DnaE* intein with a CPGC peptide between the I_C and I_N domains. The inactive intein gene was assembled by mixing 0.1 µg of each oligonucleotide (4342615; 6641791; 4342617; 4342618; 4342619; 4342620; 4342621; 4342622) with 2.5 units of *pfu* polymerase, 200 µM dNTPs, 20 mM Tris-Cl pH 8.8, 10 mM $(NH_4)_2SO_4$, 10 mM KCl, 0.1% (v/v) Triton X-100, 0.1 mg/mL bovine serum albumin (BSA), and 2 mM $MgSO_4$. The assembly reaction was incubated for five minutes at 95 °C, followed by 25 cycles of thirty seconds at 95 °C, thirty seconds at 50 °C and ninety seconds at 72 °C, and a final incubation for ten minutes at 72 °C. The inactive intein gene was amplified using $1/5^{th}$ (10 µl) of the assembly reaction and 1 µM of PCR primers (1 and 6615380) in a 50 µl PCR reaction with the reaction conditions and amplification cycles described above. We used lithium acetate transformation (Schiestl and Gietz, 1989) with 500 ng of *Eco*RI/*Xho*I

digested pJG4-5 (GenBank accession U89961) (Gyuris *et al.*, 1993) and 400 ng of PCR amplified inactive intein to clone the inactive intein into pJG4-5 by *in vivo* homologous recombination in EY93 (Ma *et al.*, 1987).

3.5.2 *Ssp-Ssp* R7 intein library construction

The lariat library (pIL-R7) was constructed by replacing the CPGC linker peptide in pIN01 with a combinatorial seven amino acid peptide using oligonucleotide (6446990). Oligonucleotide (6446990) changed the amino acid at position I_{N+1} from an alanine to cysteine. Oligonucleotide (6446990) was amplified by PCR using primers (6446987 and 6446988). The reaction conditions described above in section 3.5.1 were used with seven amplification cycles consisting of a denaturing step at 95 °C for thirty seconds, an annealing step at 55 °C for thirty seconds, and an extension step at 72 °C for fifteen seconds. Fifty micrograms of pIN01 was digested with *Rsr*II overnight at 37 °C according to manufacturer's directions and the digested plasmid was dephosphorylated with 10 units of shrimp alkaline phosphatase for three hours at 37 °C. The peptide library was cloned into *Rsr*II-digested pIN01 using *in vivo* homologous recombination (Ma *et al.*, 1987) in EY93. One-hundred lithium acetate transformations were performed (Schiestl and Gietz, 1989) with each transformation containing 400 ng of amplified oligonucleotide (6446990) and 1 μg of *Rsr*II-digested pIN01.

3.5.3 Western blot analysis of *Ssp-Ssp* R7 intein library

Expression of L2 lariat and intein constructs in EY93 was monitored using Western blot analysis with an anti-HA antibody. Samples were normalized using their OD_{600} and 20 μL of supernatant was analyzed by Western blot analysis with an anti-HA tag antibody (1:200 dilution). The Licor Odyssey infrared imaging system was used to visualize and quantify blots.

3.5.4 Construction of *E. coli* L2 lariat expression plasmid

pETIL-L2, which expresses L2 lariat, was constructed by PCR amplifying the entire L2 lariat intein gene from pIL-L2 (described in section 3.9) gene including the stop codon using primers (8340694 and 7917928). The PCR product was digested with *Eco*RI and *Xho*I and cloned into pET28b using standard restriction enzyme based cloning procedures.

3.5.5 Expression and purification of L2 lariat for LC/MS analysis

The His-tag L2 lariat construct was purified using a Ni^{2+}-NTA Spin column kit. Briefly, BL21-CP *E. coli* were transformed with the L2 lariat expression plasmid. The L2 lariat was expressed by inducing a 0.4 OD_{600} culture of BL21-CP with 1 mM IPTG (Isopropyl β-D-1-

thiogalactopyranoside) for three hours. The cells were washed, suspended in PBS, 0.05% Triton X-100, 1 mg/mL lysozyme, and lysed using sonication or the Fastprep120 (20 seconds, setting 6.0). The lysate was centrifuged at 10,000 x g for twenty minutes at 4 ^{o}C and the clarified supernatant was passed through a Ni^{2+}-NTA Spin column kit. The column was washed three times with 50 mM NaH_2PO_4 pH 7.0 and 300 mM NaCl. The L2 lariat was eluted using 50 mM NaH_2PO_4 pH 7.0, 250 mM NaCl, and 100 mM EDTA.

The lariat was separated from the I_N and unprocessed inteins and desalted using a C4 reverse phase column (Symmetry300^{TM} C4 3.5 μm 2.1 x 50 mm Column) with a gradient of 5% Buffer A / 95% Buffer B to 25% Buffer A / 75% Buffer B over 20 minutes (Buffer A: H_2O and 0.1% formic acid (v/v), Buffer B: acetonitrile and 0.08% formic acid (v/v)). The fraction containing the lariat peptide was collected, lyophilized, and stored at – 80 °C.

3.5.6 Chemoselective lactone cleavage

To determine the amount of lactone-cyclized lariat in the sample prior to MS analysis, we forced the cleavage of the lactone bound using $Na^{18}OH$. This procedure results in a chemoselective ring opening, where the lactone bond was selectively cleaved by the basic conditions, leaving the amide bonds intact (Hagelin, 2005). 0.5 M $Na^{18}OH$ was prepared by dissolving sodium in 98% $H_2{}^{18}O$. Five-hundred micrograms of lyophilized Ni^{2+}-NTA/reverse phase HPLC purified His-tagged L2 lariat was treated with either 0.5 M $Na^{16}OH$ or 0.5 M $Na^{18}OH$ for 16 hours at room temperature. The reaction was acidified with 0.5 N HCl to give a final pH between 2.0 and 7.0.

3.5.7 Trypsin digestion of hydrolyzed lariat

The NaOH hydrolyzed L2 lariat was purified by HPLC under the same conditions described previously using a C4 reverse phase column (Symmetry300^{TM} C4 3.5 μm 2.1 x 50 mm Column) to remove any residual ^{18}O and salts from previous steps. The purified L2 lariat was lyophilized, hydrolyzed, and resuspended in 6 M urea and 100 mM Tris-HCl pH 8.0. The sample was heated at 80 °C for ten minutes, and then cooled to room temperature and diluted 10-fold in 100 mM Tris-HCl pH 8.0. The sample was treated with 0.68 μg of modified sequencing grade trypsin and incubated overnight (18 hours) at 37 ^{o}C. The tryptic digests were separated from the $Na^{16}OH$ or the $Na^{18}OH$ treated samples with a BioSuite^{TM} C18 PA-A 3 μm 2.1 x 250 mm column using a gradient of 5% Buffer A / 95% Buffer B to 50% Buffer A / 50% Buffer B over twenty minutes (Buffer A: H_2O and 0.1% formic acid (v/v), Buffer B:

acetonitrile and 0.08% formic acid (v/v)). Eluted peptides were analyzed using ESI-TOF(+) MS.

3.5.8 LC/MS analysis of trypsin fragments

The molecular weights of the eluted peptides were determined using ESI(+)-TOF MS. The multi-charged lariat spectrums were resolved using maximum entropy software (MaxEnt) to determine the ratio of hydrolyzed to unhydrolyzed lariat after HPLC/Mass spectrometry analysis. The raw spectra were processed using MATCHING software (Fernández-de-Cossio *et al.*, 2004).

3.6 Lariat stabilization and optimization

3.6.1 *Ssp-Npu* intein gene construction

pIL500 (Figure 3.4) was constructed using pIN01, which was similar to pJG4-5 with the following modifications. pIL500 retained the *Ssp* DnaE I_C domain and had the *Ssp* DnaE I_N domain replaced with the *Npu* DnaE I_N domain. The *Ssp* DnaE I_C domain and the *Npu* DnaE I_N domains were separated by an *Nru*I restriction site. In addition, the *E. coli* ampicillin resistance marker was replaced by a kanamycin resistance marker. pIN01 was digested with *Rsr*II and *Xho*I in NEBuffer 4 (50 mM Tris-Acetate pH 7.9, 50 mM potassium acetate, 10 mM magnesium acetate, 1 mM dithiothreitol) for three hours at 37 °C. A synthetic *Npu* DnaE gene was constructed from five oligonucleotides that contained codons optimized for expression in *S. cerivisiae*. The *Npu* DnaE gene was constructed in three steps: (i) Dimer extension, (ii) Full length construction, and (iii) Full length amplification. In dimer extension step, approximately 1 µg (20 µM) of oligonucleotides (15951128; 15951129; 15951130; 15951131; 15951132; 15951127), were mixed together in pairs in separate PCR tubes with 60 mM Tris-SO_4 (pH 8.9), 18 mM NH_4SO_4, 2 mM $MgSO_4$, 10 mM dNTPs, and 1.0 unit of Platinum® Taq DNA Polymerase High Fidelity. These dimers were extended by an initial five minute denaturation step, followed by 5 rounds of 95 °C for thirty seconds, 55 °C for thirty seconds, and 68 °C for fifteen seconds. In the full length construction step, the full length *Npu* DnaE gene was constructed by mixing the dimers formed in dimer extension step in a single reaction with 60 mM Tris-SO_4 (pH 8.9), 18 mM NH_4SO_4, 2 mM $MgSO_4$, 10 mM dNTPs, and 1.0 Unit of Platinum® Taq DNA Polymerase High Fidelity under the exact same conditions as in the dimer extension. Finally, in the full length amplification step, the gene was selectively amplified from the pool of incomplete dimer extensions, resulting in the full length gene. One-tenth of product

from the full length construction step was mixed with 20 μM (16646144), 20 μM (15951127), 60 mM Tris-SO_4 (pH 8.9), 18 mM NH_4SO_4, 2 mM $MgSO_4$, 200 μM dNTPs, and 1.0 unit of Platinum® Taq DNA Polymerase High Fidelity. The PCR reaction was initially denatured for five minutes at 95 °C, followed by 25 cycles of 95 °C for thirty seconds, 55 °C for thirty seconds, and 68 °C for thirty seconds. The synthetic *Npu* DnaE gene was cloned into pIN01 digested with *Rsr*II and *Xho*I in the yeast strain EY93 using lithium acetate transformation and homologous *in vivo* recombination. The plasmid produced by this transformation was referred to as pIL100.

The kanamycin resistance gene was cloned into pIL100 at the ampicillin resistance gene site. pIL100 was digested with *ScaI* overnight at 37 °C. The kanamycin resistance gene was prepared by PCR amplification using P1 20 μM P1 Kan-switch Primer (#19), 20 μM Kan-switch Primer (#20), 60 mM Tris-SO4 (pH 8.9), 18 mM NH_4SO_4, 2 mM $MgSO_4$, 10 mM dNTPs, and 1.0 unit of Platinum® Taq DNA Polymerase High Fidelity. An initial denaturation of five minutes at 95 °C was used, followed by 25 cycles of 95 °C for thirty seconds, 55 °C for thirty seconds, 68 °C for one minute. The kanamycin resistance gene was cloned into pIL100 using lithium acetate transformation *and in vivo* homologous recombination. The resulting positive clone referred to as pIL500 was confirmed by sequencing. pIL500 expresses the *Ssp DnaE* I_C domain and the *Npu DnaE* I_N domain with an arginine amino acid between the two domains.

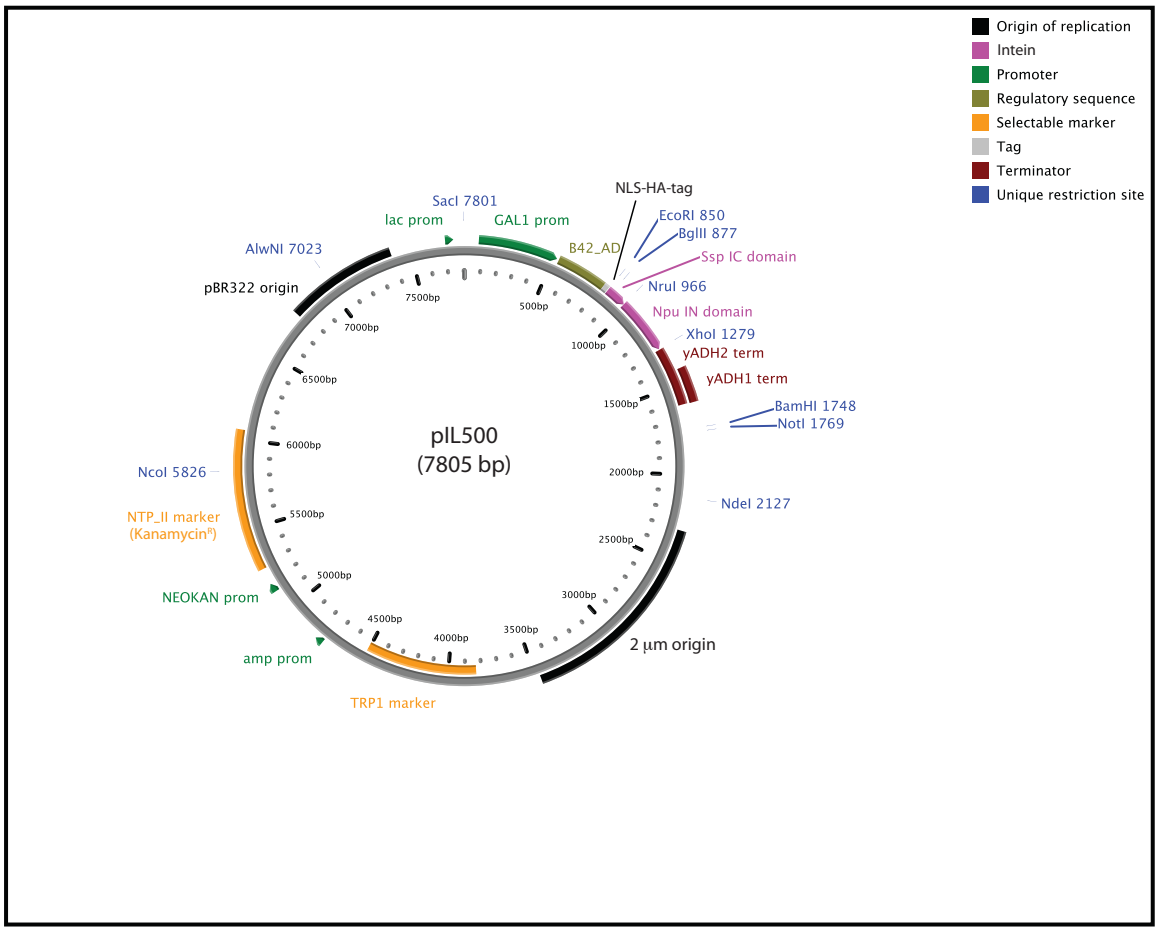


Figure 3.4 | pIL500
The library plasmid pIL500 was used to create lariat libraries and is based on the pIN01/pJG4-5 backbone. pIL500 contains a galactose promoter (GAL1 Prom), which drives expression of the *Ssp-Npu* lariat intein fused to an activation domain (B42_AD), a nuclear localization sequence (NLS), and a haemagglutinin tag (HA). The *Ssp* I_C domain and *Npu* I_N domain are separated by the *Nru*I restriction site. The yeast alcohol dehydrogenase terminator sequence (yADH1/2 ter) is used to terminate transcription. pIL500 contains a TRP1 gene involved in the biosynthesis of tryptophan for auxotrophic selection in yeast and the 2 µm high copy origin of replication for plasmid replication in yeast. pIL500 contains the kanamycin/neomycin resistance marker gene for selection in *E. coli*. The pBR322 origin allows replication in *E. coli*.

3.6.2 *Ssp-Npu* library construction

Three additional pIL500 lariat libraries were constructed: a random five-mer library (pIL-R5), a random 10-mer library (pIL-R10), and a focused five-mer library (pIL-F5). The SR linker peptide in pIL500 was replaced with a combinatorial five or ten amino acid peptide using the library oligonucleotides (16415915; 16415916; 16415917). The library oligonucleotides were amplified using PCR with primers (16646144 and 16415918). We used the reaction conditions described above with seven amplification cycles consisting of a denaturing step at 95 °C for thirty seconds, an annealing step at 55 °C for thirty seconds, and an extension step at 72 °C for fifteen seconds. pIL500 was digested with *Nru*I overnight at 37 °C. *Nru*I-digested pIL500 was dephosphorylated with 10 units of shrimp alkaline phosphatase for three hours at 37 °C. We cloned the library into pIL500 using lithium acetate transformation in EY93 (Ma *et al.*, 1987). Reactions were performed in 100 μL aliquots. For the focused five-mer library (pIL-F5), five reactions were performed. For the random ten-mer library (pIL-R10) fifty reactions were performed. For the random five-mer library (pIL-R5), twenty-five reactions were performed. Homologous *in vivo* recombination was used to create the library. The pIL500 digested plasmid was co-transformed with the PCR reaction product into EY93 using the lithium acetate high efficiency protocol (Schiestl and Gietz, 1989).

3.6.3 Construction of G6/G7/B11 mutants

Twenty-nine mutations were constructed in pET28b using the Phusion™ Site-Directed Mutagenesis Kit. Mutations were introduced at the G6, G7, and B11 positions by following the manufacture's instructions. Mutations were introduced at the G6/G7 positions by using primers (33903878; 33903879; 33903880; 33903881; 33903882; 33903384) with the primer (33903885). Each primer combination resulted in a different mutation at the G6/G7 position. These plasmids were used as templates to introduce the mutations at the B11 position, which was performed using primers (33903886; 33903887; 33903888) with the primer (33903890).

3.6.4 Expression and purification of mutant lariats for LC/MS analysis

The His-tag L2 lariats were purified using a Ni^{2+}-NTA Spin column kit. Briefly, BL21-CP *E. coli* were transformed with L2 lariat mutant expression plasmids by electroporation. The L2 lariat constructs were expressed by inducing a 0.4 OD_{600} culture of BL21-CP with 1 mM

IPTG for three hours. The cells were washed in ddH_2O and resuspended in PBS with 0.05% Triton X-100 and 1 mg/mL lysozyme. Cells were lysed using a Fastprep120 (20 seconds, setting 6.0). The lysate was centrifuged at 10,000 x g for twenty minutes at 4 ^oC and passed the clarified supernatant through a Ni^{2+}-NTA column. The column was washed three times with 50 mM NaH_2PO_4 pH 7.0 and 300 mM NaCl. The lariats were eluted using 50 mM NaH_2PO_4 pH 7.0, 250 mM NaCl, and 100 mM EDTA.

3.6.5 LC/MS analysis of G6/G7/B11 mutants

We analyzed the lariat by LC/MS using a C4 reverse phase column (Symmetry300™ C4 3.5 μm 2.1 x 50 mm Column). A $1/10^{th}$ dilution of the purified protein was diluted in water and injected onto the C4 reverse phase column with a gradient of 5% Buffer A / 95% Buffer to 25% Buffer A / 75% Buffer B over twenty minutes (Buffer A: H_2O and 0.1% formic acid (v/v), Buffer B: acetonitrile and 0.08% formic acid (v/v)) with a five minute salt divert to remove any salts prior to MS analysis. The spectra were collected and analyzed using MaxEnt software to obtain the molecular weights of the proteins in the samples.

3.7 *Ssp-Ssp* G6/G7/B11 library

A library of mutants at the G6/G7/B11 positions was constructed using a modified Phusion™ Site-Directed Mutagenesis Kit protocol (Figure 3.5). The pIL-L2 plasmid was amplified using four primers. The B11 amino acid in the *Ssp*-I_N domain was randomized with the NNK codon using primers (33903885 and 33903889). The G6/G7 amino acids were randomized using the primers (33903891 and 33903890) (Figure 3.5). The resulting PCR product had 20 bp preceding the B11 position that matched the I_N domain, followed by the randomized B11 position, followed by the rest of the *Ssp*-I_N domain and the pIL-L2 backbone sequence including the *Ssp*-I_C domain up to the G5 position. The second PCR product had 20 bp before the G6/G7 position that matched amino acids in the I_C domain, followed by the randomized G6/G7 positions, followed by the anti-LexA L2 sequence through to the B10 amino acid. The two PCR products were co-transformed using the lithium acetate protocol in 1:3 ratio of mutated sequence to L2 sequence (Figure 3.5).

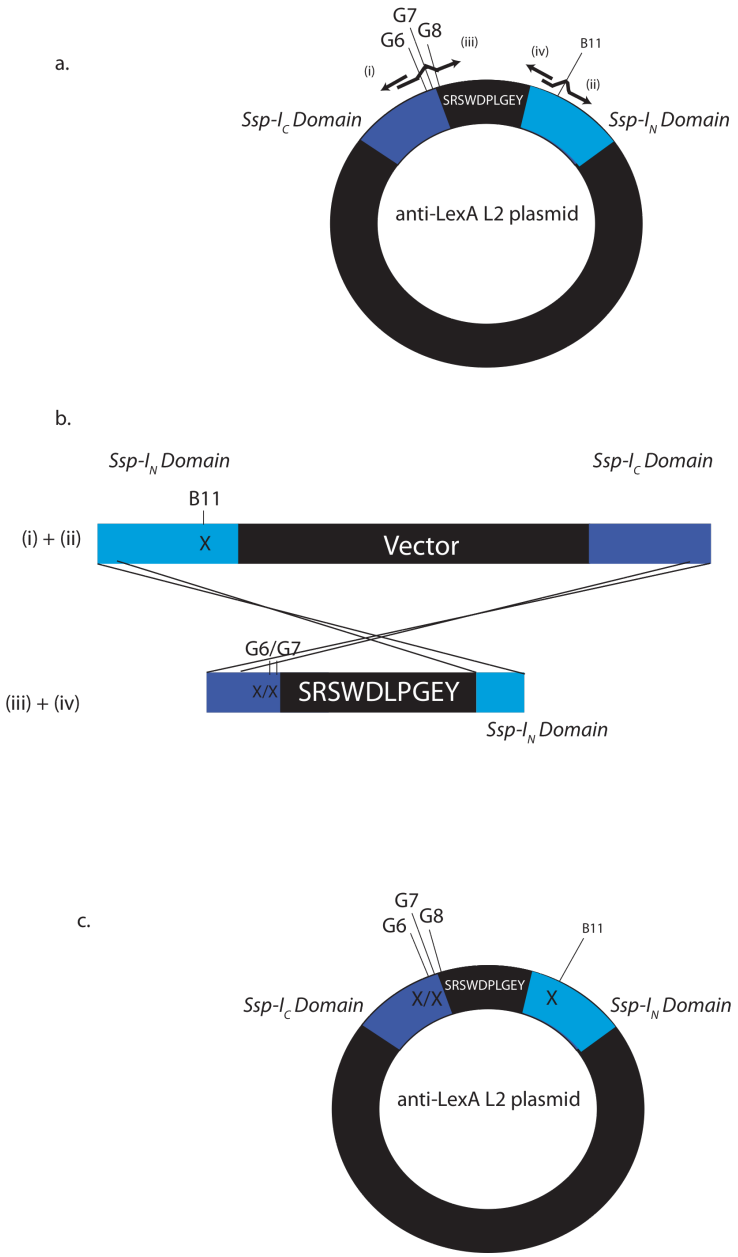


Figure 3.5 | G6/G7/B11 library
The G6/G7/B11 library was constructed in two steps. (a) Step 1: PCR reaction with primer (ii) introduces mutations at the B11 position. Primers (i) and (ii) results in a linear dsDNA product from the B5 amino acid, including the randomized B11 position, to the I_{C-5} amino acid. Primers (iii) and (iv) results in a linear dsDNA product that contains G6/G7 positions and ends at the B10 position. (b) Step 2: *In vivo* homologous recombination. The two PCR products have ends 20 bp in length that are identical. I*n vivo* homologous recombination of the two PCR products results in the formation of the original plasmid with mutations at the G6/G7/B11 amino acid positions. (c) G6/G7/B11 library. Mutations introduced at the G6/G7 and B11 positions represented by "X".

3.8 Y2H screening of the *Ssp-Ssp* G6/G7/B11 library

Yeast interaction mating (Kolonin *et al.*, 2000) was used to test interactions between the *Ssp-Ssp* anti-LexA L2 G6/G7/B11 library and the LexA bait. The transformation frequency of the library creation step was low, yielding approximately 100 cfu/transformation with approximately 10% background (The yield could be improved by extending the region of overlap beyond 20 bp). Since the efficiency was low, only 400 colonies were directly picked and grown for 48 hours at 30 °C in 96-well format in 600 µL of SD W- using 96-well tube blocks. The cells were centrifuged at 4,000 x g and resuspended in 50 µL of ddH_2O in 96-well u-bottom plates. A 96-pin replicator was dipped into the 96-well round-bottom plates containing 50 µL of library yeast cells and used to create four replicates on YPD agar plates, which had already been pinned with cells containing EY111::pEG202, to generate a 384-spot array. The 96-pin replicator was washed between pinnings in 10% bleach for 1 minute, rinsed with water by raising and lowering the replicator 5 – 10 times, sterilized in 95% ethanol, and dried either by flame or by hot air (blow-dryer). Cells were allowed to mate for 48 hours and the resulting colonies were transferred using the 384-pin replicator to SD H-W- and SGR H-W-L-, and SGR H-W-A- X-gal plates.

3.9 Anti-LexA L2 lariat

3.9.1 Isolation of lariats against LexA

The lariat library was screened for interactions with LexA using Y2H interaction mating, which has been described in section 3.3.6 (Kolonin *et al.*, 2000). Briefly, the LexA bait plasmid (pEG202) (Gyuris *et al.*, 1993) was transformed into EY111. EY111::pEG202 was mated to EY93::pIL-R7. EY111::pEG202 was cultured in 500 mL of SD His^- media to an OD_{600} of 0.6 – 0.9. EY111::pEG202 cells were pelleted by centrifugation and resuspended in an equal volume of yeast peptone dextrose (YPD) media. EY93::pIL-R7 cells were mixed with EY111::pEG202 cells at a ratio of 1:20. Yeast cells were mated on YPD plates at 30 °C for 24 hours. Mated yeast cells were pooled and 20 million diploid yeast cells were screened using the Y2H assay to detect lariats that interacted with LexA using the LEU2, ADE2, and LacZ reporter genes. Diploid yeast cells were cultured on SGR $H^-W^-L^-Ade^-$ plates containing X-Gal for approximately seven days. Positive colonies were selected and positive interactions were reconfirmed by isolating pIL-R7 from the positive colonies and repeating the Y2H assay as described above.

3.9.2 Construction of Y2H L2 lariat control plasmids

pIL-L2, which expresses the L2 lariat, was isolated from the pIL-R7 library screen against LexA. The noose amino acid sequence of pIL-L2 was RSWDLPGEY.

pIL-L2-scrambled expresses the L2-scrambled lariat where the noose amino acid sequence is SDPGLRSWEY. pIL-L2-scrambled was constructed using the same strategy as the pIL-R7 library except oligonucleotide (27171877) was used in place of oligonucleotide (6446990) and only a single transformation was performed.

pIL-L2-inverted expresses the L2-inverted lariat where the noose amino acid sequence is SGPLDWSREY. pIL-L2-inverted was constructed using the same strategy as pIL-R7, except oligonucleotide (27171878) was used in place of oligonucleotide (6446990) and only a single transformation was performed.

pIN-L2 expresses the L2 inactive intein. pIN-L2 was constructed by mutating cysteine at I_{N+1} to alanine. Two overlapping PCR fragments were used to introduce the point mutation. Primers (1 and 2) were used to amplify the N-terminus region and primers (3 and 4) were used to amplify the C-terminus region of the L2 lariat. The two PCR products were mixed together and the full-length inactive intein was amplified with primers (1 and 4). The PCR product was cloned into *EcoR*I/*Xho*I-digested pIN01 using *in vivo* homologous recombination in EY93 (Ma *et al.*, 1987).

pLIN-L2 expresses the L2 linear peptide. pLIN-L2 was constructed by introducing two stop codons after the L2 sequence RSWDLPGEY. Oligonucleotide (26911159) was used in place of oligonucleotide (6446990) and only a single transformation was performed.

pACT-L2 expresses the L2 active intein. pACT-L2 was constructed by mutating alanine at I_{C+1} in pIL-L2 to asparagine. Two overlapping PCR fragments were used to introduce the point mutation. Primers (1 and 5) were used to amplify the N-terminus region and primers (8 and 4) were used to amplify the C-terminus region of the L2 lariat. The two PCR products were mixed together and the full-length active intein was amplified with primers (1 and 4). The PCR product was cloned into *EcoR*I/*Xho*I-digested pIN01 using *in vivo* homologous recombination in EY93 (Ma *et al.*, 1987).

Alanine mutant plasmids were constructed using homologous recombination. The lariat sequence SRSWDLPGEY was sequentially mutated to alanine from the second position arginine to the tenth position tyrosine. Oligonucleotides with the appropriate alanine substitution were PCR amplified as above and cloned into pIN01 digested with *Rsr*II.

pALA-R2A expresses the L2 lariat intein with position 2 mutated to alanine (SASWDLPGEY) . The mutation was constructed using the primer (27431776).

pALA-S3A expresses the L2 lariat intein with position 3 mutated to alanine (SRAWDLPGEY). The mutation was constructed using the primer (27431777).

pALA-W4A expresses the L2 lariat intein with position 4 mutated to alanine (SRSADLPGEY). The mutation was constructed using the primer (27431778).

pALA-D5A expresses the L2 lariat intein with position 5 mutated to alanine (SRSWALPGEY). The mutation was constructed using the primer (27431779).

pALA-L6A expresses the L2 lariat intein with position 6 mutated to alanine (SRSWDAPGEY). The mutation was constructed using the primer (27431780).

pALA-P7A expresses the L2 lariat intein with position 7 mutated to alanine (SRSWDLAGEY). The mutation was constructed using the primer (27431781).

pALA-G8A expresses the L2 lariat intein with position 8 mutated to alanine (SRSWDLPAEY). The mutation was constructed using the primer (27431782).

pALA-E9A expresses the L2 lariat intein with position 9 mutated to alanine (SRSWDLPGAY). The mutation was constructed using the primer (27431783).

pALA-Y10A expresses the L2 lariat intein with position 10 mutated to alanine (SRSWDLPGEA). The mutation was constructed using the primer (27431784).

3.9.3 Construction of *E. coli* L2 lariat control plasmids

pETIL-L2 expresses the L2 lariat and was constructed by PCR amplifying the entire L2 lariat gene from pIL-L2, including the stop codon, using primers (6 and 7). The PCR product was digested with *Eco*RI and *Xho*I and cloned into pET28b using standard restriction enzyme based cloning procedures.

pETIL-L2-scrambled expresses the L2-scrambled lariat and was constructed by PCR amplifying the entire scrambled L2 gene including the stop codon, from pIL-L2-scrambled with primers (6 and 7). The PCR product was cloned into pET28b as described above.

pETIN-L2 expresses the L2 inactive intein and was constructed by PCR amplifying the entire inactive L2 gene from pIN-L2 intein gene including the stop codon with primers (6 and 7). The PCR product was cloned into pET28b as described above.

pETLIN-L2 expresses the L2 linear peptide and was constructed by PCR amplifying the entire linear L2 gene from pLIN-L2, including the stop codon, using primers (6 and 7). The PCR product was cloned into pET28b as described above.

pETACT-L2 expresses the L2 active intein and was constructed by PCR amplifying the entire active L2 intein gene from pACT-L2, using primers (6 and 7). The PCR product was cloned into pET28b as described above.

3.9.4 Qualitative Y2H assay

The interaction strength of different constructs was determined by growth and color. The number of diploids cells plated on the selection media was normalized and plated in serial dilutions. Diploid cells were grown overnight in SD H^-W^-. The cells were normalized to 0.5 x 10^6 CFU/μL and serial dilutions starting at one million (2 μL) cells were plated onto SG/R $H^-W^-L^-A^-$X-gal^+ and grown for 5 days at 30 °C.

3.9.5 Surface plasmon resonance assay

Surface plasmon resonance (SPR) was performed using a BiaCore3000 instrument. LexA was purified using HIS-Select® HF Nickel affinity gel (Ni^{2+}-NTA) and diluted in 50 mM Tris-HCl and 150 mM NaCl prior to SPR analysis. Purified His-tagged LexA was immobilized on a NTA Sensor Chip. L2 lariat or L2 linear peptides were dissolved in 50 mM Tris-HCl and 150 mM NaCl. Sequential 2-fold serial dilutions of either synthetic L2 lariat or L2 linear peptides ranging from 125 μM to 7.8 μM were injected over the surface at a flow rate of 10 μL/min. The K_D for the L2 lariat-LexA interaction was calculated by plotting maximum response units versus peptide concentration and fitting the data to a one site binding model using Prism4.0 (GraphPad) statistical software.

3.9.6 Ni^{2+}-NTA chromatin precipitation assay

Ni^{2+}-NTA chromatin precipitation assay was based on previously described procedures (Tamimi *et al.*, 2004). Briefly, BL21-CP cells transformed with lariat expression plasmids and cultured to an OD_{600} of ~ 0.6 in LB with 50 μg/mL of kanamycin. Cultures were either induced for two hours with 0.2 mM IPTG at 30 °C or left un-induced. Forty milliliters of cells were treated with formaldehyde to a final concentration of 1% and incubated for twenty minutes at 25 °C. Glycine was added to a final concentration of 0.5 M and cells were harvested by centrifugation and washed twice with Tris-buffered saline (pH 7.5). Cells were resuspended in 500 μL lysis buffer (10 mM Tris-HCl pH 8.0, 20% sucrose, 500 mM NaCl, 1 mM EDTA, 10

mM imidazole, 4 mg/mL lysozyme) and incubated at 37 °C for thirty minutes. Five-hundred microliters of precipitation buffer (50 mM HEPES-KOH pH 7.5, 500 mM NaCl, 1 mM EDTA, 10 mM imidazole, 1% triton X-100, 0.1% sodium deoxycholate, 0.1% SDS) and PMSF (final concentration 1 mM) were added to the cell extract and the DNA was sheared by sonication to an average size of ~ 500 bp. Insoluble cellular material was removed by centrifugation and 20 μL of the supernatant was kept for use as an "input" sample.

Proteins were precipitated by diluting 250 μL of cross-linked extract to a final volume of 800 μL in precipitation buffer. Fifty microliters of HIS-Select® HF Nickel affinity gel (Ni2+-NTA) were added and the extracts were incubated for ninety minutes at 25 °C with gentle shaking. Extracts were precipitated by centrifugation and the supernatant was removed. Precipitated complexes were resuspended in 750 μL of precipitation buffer, transferred to 0.45 μm membrane microcentrifuge columns, washed once with precipitation buffer, once with precipitation buffer with 500 mM NaCl, once with wash buffer (10 mM Tris-HCl pH 8.0, 250 mM LiCl, 1 mM EDTA, 0.5% Nonidet-P40, 0.5% sodium deoxycholate) and once with Tris-HCl (pH 7.5). Precipitated complexes were eluted by incubation in 100 μL of elution buffer (50 mM Tris-HCl pH 7.5, 500 mM NaCl, 250 mM imidazole, 1% SDS) at 65 °C for ten minutes.

Precipitated complexes and corresponding "input" samples were decross-linked by incubating them at 42 °C for two hours and 65 °C overnight in 0.5X elution buffer with 0.8 mg/mL Proteinase K. Eluted LexA promoter DNA was quantified by real time PCR and normalized to the corresponding input DNA.

3.9.7 LexA chromatin immunoprecipitation assay

The LexA ChIP assay was based on a previously described procedure (Wade and Struhl, 2004). Briefly, BL21-CP cells transformed with *E. coli* lariat expression plasmids were grown to an OD_{600} of ~ 0.6 in LB with 50 μg/mL of kanamycin and then induced for two hours with 0.2 mM IPTG at 30 °C. The culture was split into two samples; one sample was treated with mitomycin C (0.25 μg/mL) and the other left untreated. Both samples were incubated for one hour at 30 °C. Forty milliliters of cells from each sample were treated with formaldehyde to a final concentration of 1% and incubated for 20 minutes at 25 °C. Glycine was added to a final concentration of 0.5 M and cells were harvested by centrifugation and washed twice with Tris-buffered saline (pH 7.5). Cells were resuspended in 500 μL lysis buffer (10 mM Tris-HCl

pH 8.0, 20% sucrose, 50 mM NaCl, 10 mM EDTA, 4 mg/mL lysozyme) and incubated at 37 °C for thirty minutes. Five-hundred microliters of immunoprecipitation (IP) buffer (50 mM HEPES-KOH pH 7.5, 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% SDS) and PMSF (final concentration 1 mM) were added to the cell extract and the DNA was sheared by sonication to an average size of ~ 500 bp. Insoluble cellular material was removed by centrifugation and 20 μL of the supernatant was kept for use as an "input" sample.

Proteins were immunoprecipitated by diluting 250 μL of cross-linked extract to a final volume of 800 μL in IP buffer. Twenty-microliters of protein A Fastflow Sepharose beads with 5 μg of anti-LexA rabbit polyclonal antibody (Upstate) or a no antibody control to test for non-specific protein A bead binding, were added to the extracts and the mixture was incubated for ninety minutes at 25 °C with gentle shaking. Extracts were immunoprecipitated by centrifugation and the supernatant was removed. Immunoprecipitated complexes were resuspended in 750 μL of IP buffer, transferred to 0.45 μm membrane microcentrifuge columns, washed once with IP Buffer, once with IP buffer + 500 mM NaCl, once with wash buffer (10 mM Tris-HCl pH 8.0, 250 mM LiCl, 1 mM EDTA, 0.5% Nonidet-P40, 0.5% sodium deoxycholate), and once with TE (10 mM Tris and 1 mM EDTA (pH 7.5)). Immunoprecipitated complexes were eluted by incubation in 100 μL of elution buffer (50 mM Tris-HCl at pH 7.5, 10 mM EDTA, 1% SDS) at 65 °C for ten minutes.

Immunoprecipitated complexes and corresponding "input" samples were decross-linked by incubating them at 42 °C for two hours and then at 65 °C overnight in 0.5X elution buffer with 0.8 mg/mL Proteinase K. Eluted DNA was further purified using a PCR clean-up kit and eluted in 200 μL.

The percentage of LexA at the operator was calculated by first calculating the occupancy units, which is the ratio of LexA bound to the *lexa* promoter compared to a non-specific control region in the *sgrR (yabN)* gene (Wade and Struhl, 2004). The percent LexA remaining bound to the operator after treatment was calculated by taking the ratio of the occupancy units of the mitomycin C treated sample to the untreated sample and multiplied by 100.

3.9.8 Quantitative real-time PCR

Quantitative real time PCR data was collected with a Miniopticon thermocycler using iQ™ SYBR® Green supermix. Primers (38891611 and 38891612) were used to amplify a region of the *lexa* gene. Primers (38891615 and 38891616) were used to amplify a region in the *sgrR (yabN)* gene. A two-step PCR cycle was used with an initial denaturation of 95 °C for five minutes followed by 95 °C for thirty seconds, and 61 ^{o}C (*lexa*) or 59 °C (*sgrR*) for one minute. The product was amplified for 40 cycles.

3.9.9 Analysis of LexA autoproteolysis

We monitored the effect of the L2 lariat constructs on mitomycin C-induced LexA autoproteolysis using an anti-LexA antibody. Cultures of BL21-CP cells transformed with *E. coli* lariat expression plasmids were grown overnight in 10 mL of LB with 50 μg/mL of kanamycin. Cultures were diluted to an OD_{600} of 0.1 in LB with 50 μg/mL of kanamycin with 1 mM IPTG and cultured at 30 °C to an OD_{600} ~ 0.4 – 0.6. Cells were treated with 100 μg/mL chloramphenicol, incubated for 10 minutes, and split in two. One culture was treated with 0.1 μg/mL of mitomycin C and the second culture was left untreated. Four-milliliters were removed for analysis at the indicated time points. The cells were washed with ddH_2O, and stored at – 80 °C until all samples were collected. Cells were resuspended in 250 μL of PBS with 0.05% Triton X-100 and 0.3 g of acid-washed glass beads, then homogenized using a Fastprep120. Cell lysates were centrifuged at 13,000 x g at 4 ^{o}C, and the cleared supernatants were characterized by Western blot analysis with an anti-LexA antibody (Invitrogen) (1:5000 dilution).

3.9.10 *E. coli* viability assay

Cell viability assays were performed as described by Lin and Little (Lin and Little, 1988). Briefly, BL21-CP cells transformed with *E. coli* lariat expression plasmids were cultured in LB-KAN at 37 °C to an OD_{600} of 0.4. Protein expression was induced with 1 mM IPTG for one hour. The samples were then diluted 100-fold in 5 mL 0.85% NaCl with or without 0.1 μg/mL of mitomycin C. For a zero time point control, we removed 10 μL and diluted it 1000-fold in ice cold LB (1 mL). The remaining sample was incubated at 37 °C for thirty minutes. We then removed 10 μL and diluted it 1000-fold into 1 mL ice-cold LB. We plated 60 μL aliquots from the zero and thirty minute samples on LB plates and incubated the plates inverted at 37 °C overnight. Normalized percent cell survival was calculated by dividing

the number of colony forming units (cfu) after thirty minutes by the number of cfu at the zero time point.

4 Results and Discussion

4.1 Aim 1: Construction of lariat peptides

Reverse genetic screening strategies usually require a specialized fusion protein. For example, phage display and Y2H assays require the peptide library to be fused to a coat protein or transcription activation domain respectively. These protein domains link peptides that interact with the target of interest indirectly to the genes that encode them. To generate a cyclic peptide library that can be screened within a cell using the Y2H assay, a method was needed to genetically encode cyclic peptides that were fused to a transcriptional activation domain. This posed a problem for cyclic peptides, as they do not have a free N- or C- terminus to fuse protein domains. To solve this problem, the intein-mediated cyclic peptide reaction was modified to generate a lactone-cyclized peptide that contained a covalently attached linear peptide, referred to as a lariat. The term lariat was particularly suited to our application since the combinatorial peptide region functioned as a "noose" to capture protein targets. The lactone bond functioned as the "honda" of a lariat, constraining the peptide to form the noose. The linear peptide functioned like the "spoke" or rope that passes through the honda and attaches the target/noose complex to the Y2H transcription machinery, allowing us to select combinatorial peptides that interact with a given target (Figure 4.1).

Inteins have previously been modified to produce lactam peptides that have no free N- or C-termini (Scott *et al.*, 1999). The lactam peptide producing intein reaction has four catalytic steps (Figure 4.2b). In the first step, the intein undergoes an N-S acyl shift. The thiol group on the cysteine side chain of the I_{N+1} amino acid replaces the amide bond with a thioester linkage between the region to be cyclized and the I_N domain. This thioester bond is more susceptible to nucleophilic attack and makes the I_N domain a better leaving group. In the second step, a transesterification reaction occurs where the thioester bond generated in the first step is attacked by the hydroxyl group of the serine side chain at position I_{C+1}. This results in the release of the I_N domain and the formation of the lariat. In the third step, the asparagine at position I_{C-1} undergoes cyclization, which releases the I_C domain from the lariat. In the fourth step, the lactone peptide undergoes an O-N acyl shift, converting the lactone to the more energetically favorable lactam.

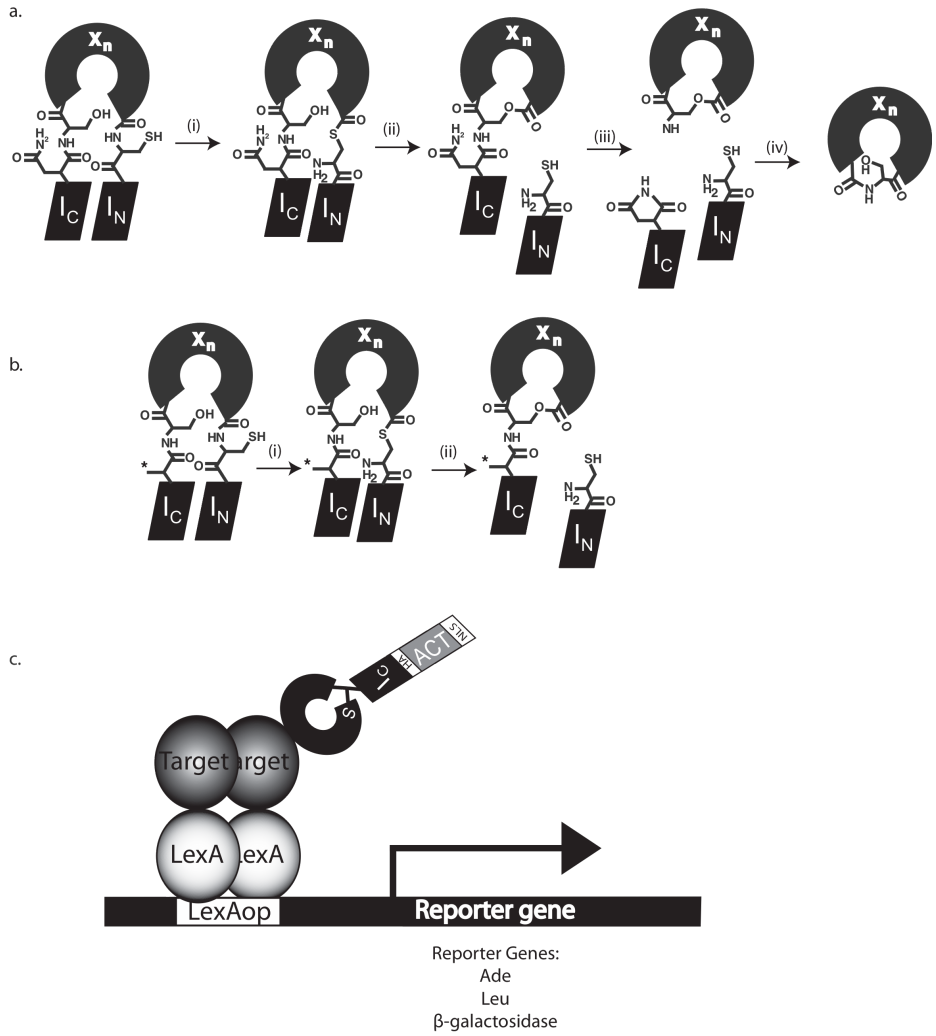


Figure 4.1 | Production and isolation of lariats using the Y2H assay
(a) Intein-mediated lactam peptide production. (i) Intein undergoes an N-S acyl shift using the I_{N+1} cysteine at the peptide-I_N junction. (ii) Transesterification reaction involving I_{C+1} serine at the I_C-peptide junction and the thioester formed in step (i), which releases the I_N domain and produces the lariat intermediate. (iii) I_{C-1} asparagine undergoes a side chain cyclization, which releases the I_C domain and generates a lactone peptide. (iv) Lactone peptide undergoes a thermodynamically favored O-N acyl shift to produce a lactam peptide. **(b)** Intein-mediated production of lariat peptides. The lariat is produced by mutating the asparagine at position I_{C-1} to alanine (*), which inhibits asparagine cyclization and stops the reaction at the lariat intermediate. **(c)** Lariat Y2H assay. Illustration of the lariat peptide used in the Y2H assay. In this example, lariats that interact with the target protein LexA activate transcription of reporter genes Ade, Leu2, and ß-galactosidase. LexAop = LexA operon, I_C = Intein C-terminal domain, I_N = Intein N-terminal domain, HA = haemagglutinin tag, ACT = activation domain, NLS = nuclear localization signal, X_n = peptide noose.

a.

	I_C	I_{C-1}	I_{C+1}	Region to be cyclized	I_{N+1}	I_N
Active	I_C	Asn	Ser	X_n	Cys	I_N
Lariat	I_C	*Ala	Ser	X_n	Cys	I_N
Inactive	I_C	Asn	Ser	X_n	**Ala	I_N

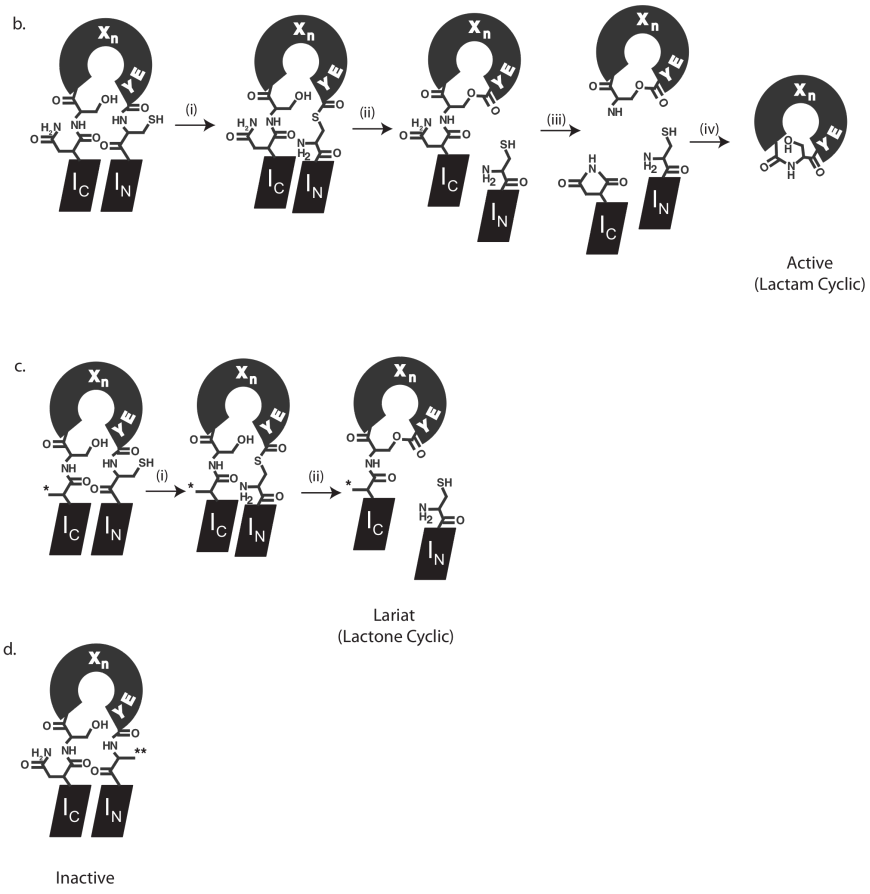


Figure 4.2 | Intein mutations used to generate lariats and inactive inteins
(a) Mutations used to generate the lariat, and inactive inteins. * Indicates the asparagine to alanine mutation at position I_{C-1}/G8 used to generate the lariat intein. ** indicates the cysteine to alanine mutation at position I_{N+1}/A1 used to generate the unprocessed intein. **(b)** The mechanism used by the active intein to generate lactam peptides**.** (i) The unprocessed intein undergoes an N-S acyl shift using the I_{N+1} cysteine at the peptide-I_N junction. (ii) The transesterification reaction, which involves the I_{C+1} serine at the I_C-peptide junction and the thioester formed in step (i), releases the I_N domain and produces the lariat intermediate. (iii) The I_{C-1} asparagine undergoes a side chain cyclization, which releases the I_C domain and generates a lactone peptide. (iv) The lactone peptide undergoes a thermodynamically favored O-N acyl shift to produce a lactam peptide. **(c)** The mechanism used by the lariat intein to produce the lariat. The lariat is produced by mutating the asparagine at position I_{C-1} to alanine (*), which inhibits step (iii), asparagine cyclization, and stops the reaction at the lariat intermediate. **(d)** The inactive intein is produced by a cysteine to alanine mutation (**), which blocks step (i) transesterification.

Three distinct intein structures have been highlighted in this pathway: (i) the unprocessed intein, which has not undergone any steps in the intein reaction, (ii) the lariat, which is produced by the first two steps in the intein reaction, and (iii) the lactam peptide, which is produced when all four steps in the intein reaction occur. Mutations can be introduced into the intein to block processing and produce lariat or unprocessed products (Figure 4.2).

An inactive intein that cannot undergo any processing steps can be generated by blocking Step 1 (N-S Acyl shift). Step 1 can be blocked by mutating cysteine to alanine at position I_{N+1}/A1. The inactive intein can be used to constrain peptides through the interaction between the I_C and I_N domains (Figure 4.2).

We proposed the lariat could be produced by blocking Step 3 (Asparagine cyclization). Step 3 can be blocked by mutating asparagine to alanine at position I_{C-1}/G7. The lariat structure constrains peptides through the lactone bond (Figure 4.2).

4.1.1 *Ssp-Ssp* intein gene construction

Inteins do not naturally occur in an orientation that allows protein cyclization (Figure 2.1). To cyclize peptides (exteins), intein domains must be permutated so that they flank a continuous extein (Figure 2.1). The first step in generating a lariat producing intein was to construct a synthetic intein gene with permutated I_N and I_C domains. The second step was to mutate asparagine at position I_{C-1}/G7 to alanine, which blocks the cyclic peptide producing intein reaction at the lariat intermediate. The third step was to insert an *Rsr*II restriction site between the intein domains, which would allow DNA encoding combinatorial peptides to be cloned in between the intein domains (Figure 2.1).

The *Synechocystis sp*. P*CC6803* (*Ssp)* DnaE gene was used as a template to construct the lariat intein since a permutated version of this intein has previously been used to cyclize proteins *in vivo* (Scott *et al.*, 1999). The *Ssp* DnaE gene is a naturally occurring split intein, which is transcribed and translated as two separate proteins. The entire 3,573,421 bp genome of *Ssp* is available from GenBank (BA000022) (Benson *et al.*, 2009). The two *Ssp* DnaE intein proteins are located on different coding strands separated by 750,000 bp in the genome. The N-terminus of the DnaE gene fused to the *Ssp* I_N domain is located at position 3,561,946 – 3,564,639 (BAA18870.1) in the genome and is expressed as one protein. The *Ssp* I_C domain of the intein fused to the C-terminus of the DnaE gene is expressed as a second protein, which is located on the complementary strand approximately 750,000 bp away at position 736,435 –

737,811 (BAA17242.1). The split intein fragments then associate with each other inside the cell via the interaction between the I_C and I_N domains. The *Ssp* DnaE intein was selected since it was the most well characterized cyclic protein producing intein at the time that this study was initiated (Iwai and Plückthun, 1999; Evans *et al.*, 1999; Scott *et al.*, 1999; Scott *et al.*, 2001; Iwai *et al.*, 2001; Xu and Evans, 2001; Williams *et al.*, 2002; Kinsella *et al.*, 2002).

Codons in the synthetic lariat producing intein were optimized to facilitate expression in the Y2H assay. Codons were changed to represent the most frequently used yeast codons determined from two different sources: (i) The codon usage table from the *Saccharomyces* genome databank (SGD) (ftp://genome-ftp.stanford.edu/pub/codon/ysc.orf.cod), which was produced by J. Michael Cherry using the GCG program CodonFrequency in 1999, from 6,216 open reading frames in yeast; or (ii) from 14,411 complete protein coding genes curated from the NCBI-GenBank Flat file release 160.0 [June 15th 2007] (Benson *et al.*, 2004; Nakamura *et al.*, 2000). Optimizing codon usage of foreign genes for the host organism has previously been shown to increase protein expression (Bennetzen and Hall, 1982; Ikemura, 1985; Mechold *et al.*, 2005; Barrett *et al.*, 2006; Brat *et al.*, 2009; Sastalla *et al.*, 2009). The wildtype *Ssp* DnaE intein sequence and the yeast-optimized sequence are shown in Figure 4.3. Codons that are underlined in Figure 4.3 were changed from the wildtype codon in the *Ssp* DnaE intein to a yeast-optimized codon. The majority of codons 56% (89/159) were changed to optimize the intein for expression in yeast (Figure 4.3).

Mutations, represented by stars (*) (Figure 4.3) in the amino acid sequence at the I_C/I_N junction (I_{C-2}/G6)-HSSCGPCEYA-(I_{N+1}/A1) were introduced to create a generic intein gene that could be used to construct different intein constructs. Histidine (H) at position (I_{C-2}/G6) was included in the lariat intein design since it was shown to facilitate intein processing in the lactam peptide producing intein (Scott *et al.*, 2001). Mutation of the I_{C-1}/G7 amino acid from asparagine to alanine was required to block intein processing and produce the lariat. Our initial lariat intein was designed to have an alanine (A) mutation at position I_{C-1}/G7, however, an error in the construction of the gene resulted in serine (S) at position (I_{C-1}/G7). This error was corrected when the combinatorial oligonucleotide library was cloned into the lariat intein expression plasmid. The peptide noose was formed by the amino acids between the I_{C+1}/G8 and I_{N-1} amino acids (I_{C+1}/G8)-SCGPCEY-(I_{N-1}). The I_{C+1}/G8 amino acid was changed from cysteine (C) to serine (S) to produce the less labile ester bond instead of a thioester bond.

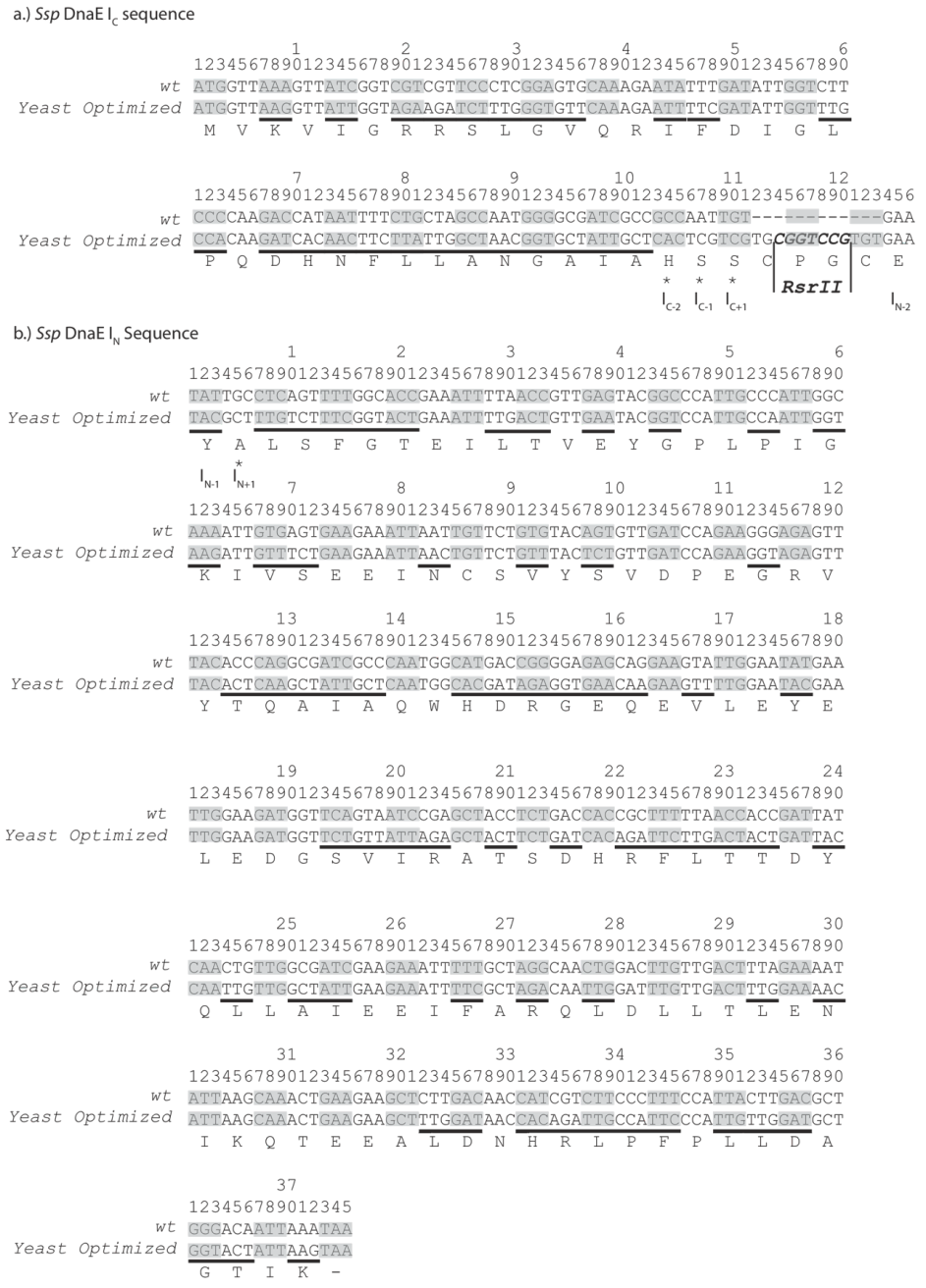


Figure 4.3 | Modified *Ssp-Ssp* DnaE intein gene sequence
(a) *Ssp*-I_C domain. Sequences of the wildtype (*wt)* and yeast codon optimized *Ssp*-I_C intein genes are shown. **(b)** *Ssp*-I_N domain. Sequences of the *wt* and yeast codon optimized *Ssp*-I_N intein genes are shown. Codons are highlighted as alternating grey/white boxes to assist in visualizing codons. Below the nucleotide sequences is the amino acid sequence of the yeast-optimized intein. Underlined codons represent silent mutations that were used to convert the *wt* codon to a yeast-optimized codon. * Indicates non-silent mutations that were introduced.

The I_{C+2}-CGPC-I_{N-3} motif was included since it encoded an *Rsr*II restriction site, which allowed combinatorial peptide libraries or specific peptide sequences to be cloned between the I_C and I_N domains. Glutamate (E) at position (I_{N-2}) and tyrosine (Y) at position (I_{N-1}) were included since they were previously shown to be important for lactam peptide cyclization (Scott *et al.*, 2001). Finally, alanine (A) at position (I_{N+1}/A1) was included to block intein processing at Step 1 (N-X acyl shift). This mutation was necessary to create inactive inteins. The alanine (I_{N+1}/A1) was mutated to cysteine during construction of lariat libraries, which allowed the N-X acyl shift to proceed.

The modified lariat intein gene was synthetically constructed using overlapping oligonucleotides (Figure 4.4) (Stemmer *et al.*, 1995). The lariat gene was amplified using primers that contained 39-nucleotides complementary to a region in the pJG4-5 plasmid where the lariat intein was to be inserted (Figure 4.5). The PCR amplified lariat gene was cloned into pJG4-5 by homologous recombination to generate the pIN01 plasmid (Figure 4.6).

The pIN01 plasmid was used to clone oligonucleotides encoding a combinatorial peptide library between the permutated *Ssp* intein domains. Homologous recombination allowed us to use the pIN01 plasmid to construct the inactive intein, the lariat, and the active intein (Figure 4.2). pIN01 was based on pJG4-5 and has a tryptophan marker for selection in yeast and an ampicillin marker for selection in *E. coli* (Figure 4.6). Protein expression was controlled by the inducible yeast GAL1 promoter. The intein was expressed as a C-terminal fusion protein to the following three domains: (i) Nuclear localization signal (NLS), which imports proteins into the nucleus where the Y2H assay reporters were located. (ii) B42 activation domain, which is a general transcription factor and activates transcription when localized to a promoter. (iii) *Influenza haemagglutinin* tag (HA-tag) [YPYDVPDYA], which allowed detection of the prey using anti-HA antibodies.

Lariat libraries displaying seven random amino acids from their "noose" region were constructed using a degenerate oligonucleotide (Figure 4.7). The random peptide library consisted of the peptide sequence I_{C+1}/G8-[SXXXXXXXEY]-I_{N-1}, where X represents any amino acid encoded by the degenerate NNK codon (N = A, C, G, or T: K = G, or T). The degenerate NNK codon was chosen to maximize amino acid diversity and minimize stop codons (Table 4.1). The degenerate NNK codon encodes 32 different codons (Table 4.1), which code for all twenty amino acids and one stop codon (TAG). At each NNK codon

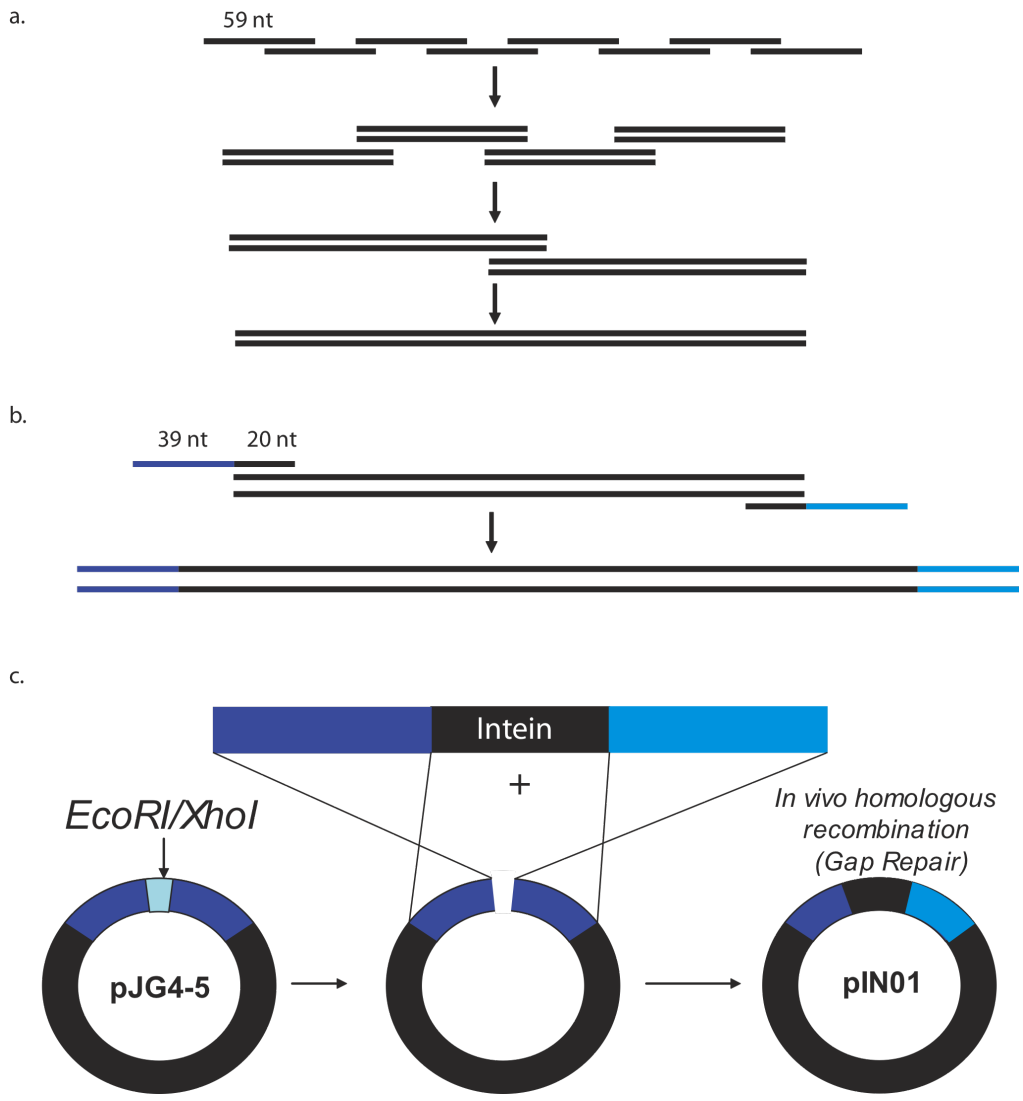


Figure 4.4 | Intein gene construction
The synthetic intein gene was constructed from eight overlapping oligonucleotides encoding the modified lariat intein gene. **(a)** Gene assembly. PCR was used to create the full-length synthetic intein gene from overlapping synthetic oligonucleotides. **(b)** Gene amplification. A second round of PCR with oligonucleotides containing 5`-ends complementary to the pJG4-5 vector (shown in light and dark blue) were used to amplify the full-length intein gene. **(c)** Recombination. The pJG4-5 vector was digested with *EcoR*I and *Xho*I and co-transformed with the amplified intein gene into yeast, which produced the pIN01 plasmid. nt represents nucleotide(s)

a.) pJG4-5 Sequence

```
          1         2         3         4         5         6
1234567890123456789012345678901234567890123456789012345678901234567890
ATGGGTGCTCCTCCAAAAAAGAAGAGAAAGGTAGCTGGTATCAATAAAGATATCGAGGAG
 M  G  A  P  P  K  K  K  R  K  V  A  G  I  N  K  D  I  E  E
          7         8         9        10        11        12
1234567890123456789012345678901234567890123456789012345678901234567890
TGCAATGCCATCATTGAGCAGTTTATCGACTACCTGCGCACCGGACAGGAGATGCCGATG
 C  N  A  I  I  E  Q  F  I  D  Y  L  R  T  G  Q  E  M  P  M
         13        14        15        16        17        18
1234567890123456789012345678901234567890123456789012345678901234567890
GAAATGGCGGATCAGGCGATTAACGTGGTGCCGGGCATGACGCCGAAAACCATTCTTCAC
 E  M  A  D  Q  A  I  N  V  V  P  G  M  T  P  K  T  I  L  H
         19        20        21        22        23        24
1234567890123456789012345678901234567890123456789012345678901234567890
GCCGGGCCGCCGATCCAGCCTGACTGGCTGAAATCGAATGGTTTTCATGAAATTGAAGCG
 A  G  P  P  I  Q  P  D  W  L  K  S  N  G  F  H  E  I  E  A
         25        26        27        28        29        30
1234567890123456789012345678901234567890123456789012345678901234567890
GATGTTAACGATACCAGCCTCTTGCTGAGTGGAGATGCCTCCTACCCTTATGATGTGCCA
 D  V  N  D  T  S  L  L  L  S  G  D  A  S  Y  P  Y  D  V  P
         31        32        33        34        35        36
1234567890123456789012345678901234567890123456789012345678901234567890
GATTATGCCTCTCCCGAATTCGGCCGACTCGAGAAGCTTTGGACTTCTTCGCCAGAGGTT
 D  Y  A  S  P  E  F  G  R  L  E  K  L  W  T  S  S  P  E  V
         37        38        39        40        41        42
1234567890123456789012345678901234567890123456789012345678901234567890
TGGTCAAGTCTCCAATCAAGGTTGTCGGCTTGTCTACCTTGCCAGAAATTTACGAAAAGA
 W  S  S  L  Q  S  R  L  S  A  C  L  P  C  Q  K  F  T  K  R
         43
123456789012
TGGAAAAGGGTC
 W  K  R  V
```

b.) Primers to amplify lariat gene

```
P1:5′ -TAC CCT TAT GAT GTG CCA GAT TAT GCC TCT CCC GAA TTC ATG GTT AAG GTT ATT GGT AG-3′
P2:5′ -TGA CCA AAC CTC TGG CGA AGA AGT CCA AAG CTT CTC GAG TTA CTT AAT AGT ACC AGC ATC CA-3′
```

Figure 4.5 | pJG4-5 fusion cassette region

(a) pJG4-5 fusion protein sequence. A 432 nucleotide segment of pJG4-5 (GenBank accession U89961) (Gyuris *et al.*, 1993) containing the entire fusion cassette including: the SV40 T-antigen nuclear localization signal (NLS) (green) [PKKKRKV] (Kalderon *et al.*, 1984), the bacterial B42 activation domain, the *Haemagglutinin* epitope tag (blue) (HA-tag) [YPYDVPDYA], and restriction sites (*EcoR*I and *Xho*I) (underlined). Alternating codons are highlighted in light grey to assist in visualization. Nucleotide numbering is shown above the sequence. The first ATG represents the start of the open reading frame for the fusion proteins. **(b)** Primers used to amplify the lariat gene. The underlined 39-nucleotide region is complementary to pJG4-5 plasmid at the region where the lariat intein was inserted. P1 is the forward primer and matches the pJG4-5 plasmid at position 283 – 321. P2 is the reverse primer and matches the pJG4-5 plasmid at position 328 – 366. The P2 primer also contains the TAA stop codon. The nucleotide segments that are not underlined are complementary to the lariat intein gene.

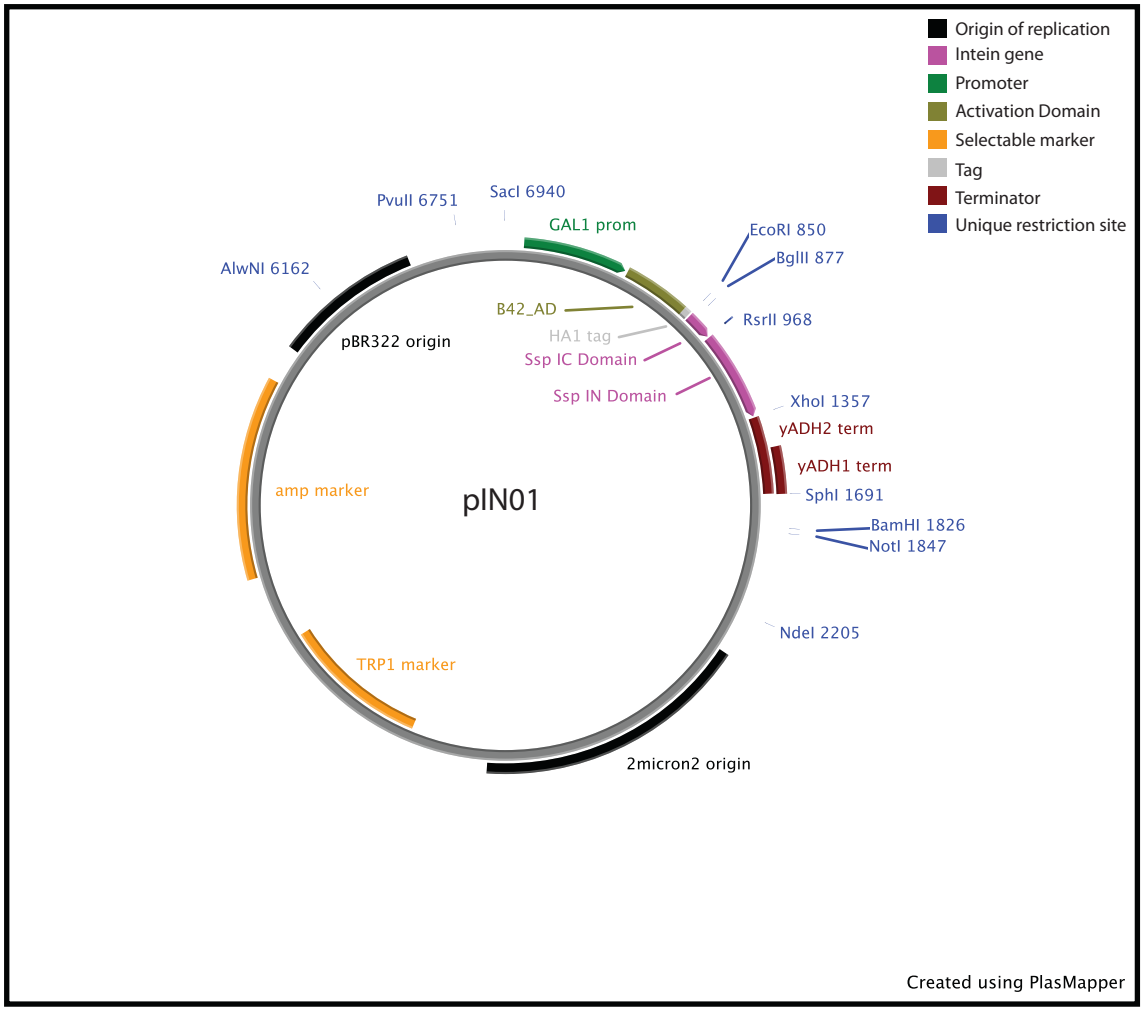


Figure 4.6 | pIN01 plasmid
Map of pIN01. pIN01 is based on the pJG4-5 backbone with the [*Ssp* DnaE I_C - *Rsr*II – *Ssp* DnaE I_N] gene construct inserted between *EcoR*I/*Xho*I restriction sites. pIN01 is a high copy yeast plasmid (2μ origin). pIN01 contains the TRP1 gene that maintains the plasmid in yeast auxotrophic for tryptophan biosynthesis and the ampicillin resistance gene β-lactamase (amp marker), which allows the plasmid to be amplified in *E. coli*. Expression of the intein construct in yeast is controlled by the galactose inducible promoter (GAL1 prom), and is terminated by the yeast alcohol dehydrogenase terminator sequences (yADH1, yADH2 ter). The I_C domain is preceded by the NLS-B42 activation domain-HA tag fusion tag. The I_C domain is flanked by unique *EcoR*I and *Rsr*II sites and the I_N domain was flanked by unique *Rsr*II and *Xho*I restriction sites.

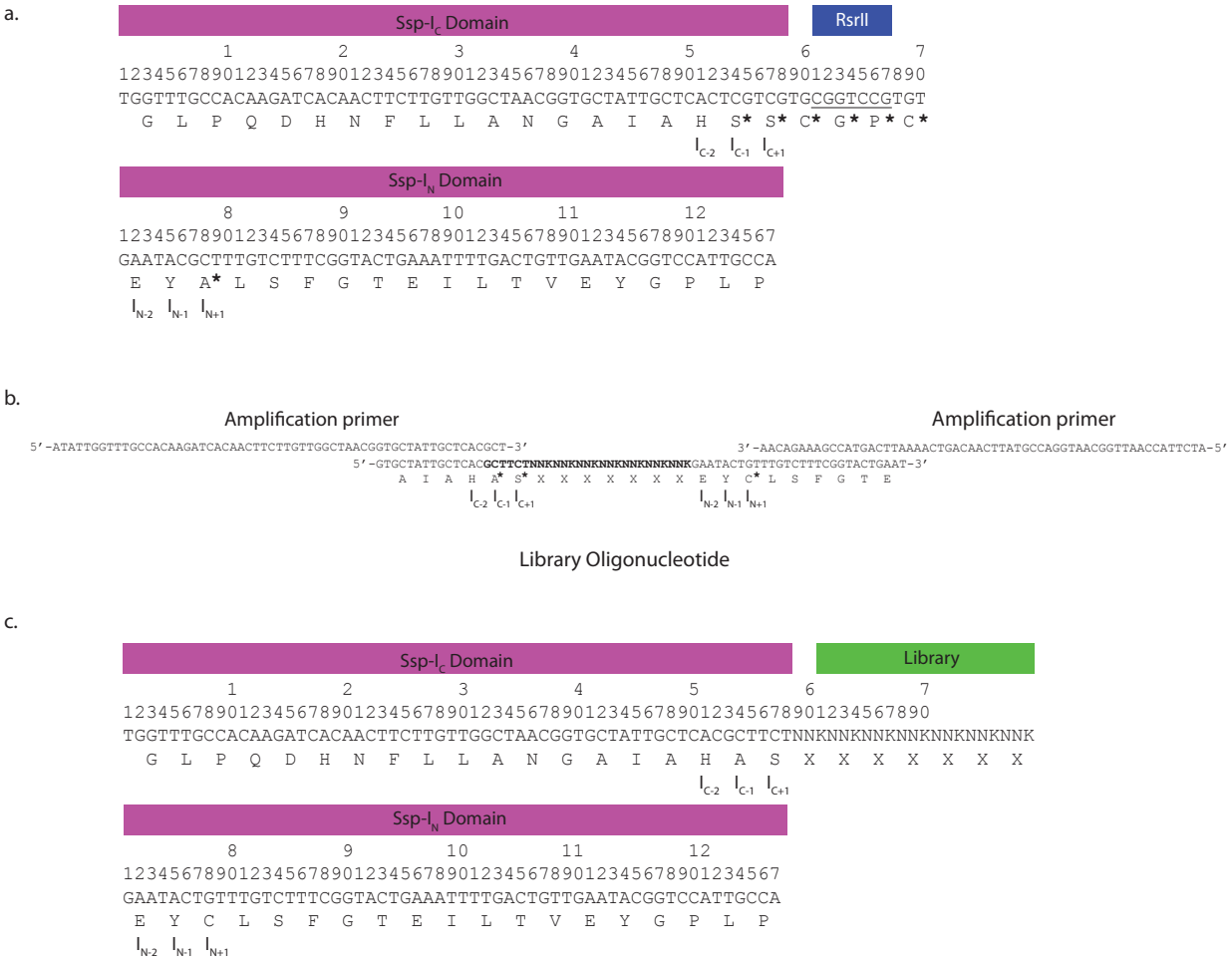


Figure 4.7 | *Ssp-Ssp* R7 lariat library
(a) The lariat intein gene with I_C and I_N domains in pink and the *Rsr*II restriction site highlighted in blue. **(b)** The oligonucleotides encoding the combinatorial peptide library (Library oligonucleotide) were amplified by PCR using amplification primers. The library oligonucleotide dictates the nucleophile at the I_{N+1} position as well as the length of the random region and any fixed amino acids. Stars (*) denote amino acids codons, which are different between pIN01 and the resulting library. **(c)** Combinatorial peptide lariat library with the combinatorial peptide location shown in green and the I_C and I_N domains in pink. K = {G,T}, N = {A, C, G, T}.

position there is a 1/32 or 3.1% chance that the codon would be a stop codon. In comparison, the NNN codon encodes 64 different codons, three of which are stop codons (TAA, TAG, TGA). At each NNN codon there is a 3/64 or 4.7% chance the codon will be a stop codon. Using the NNK codon reduces the frequency of a stop codon occurring at each position by 1.6%. This translates to a reduction in stop codon frequency of 8.5% for a peptide seven amino acids long. In a peptide with seven random amino acids, approximately 20% of sequences (1 –

$31^7/32^7$) will have a stop codon when the NNK codon is used versus 28.5% of sequences (1 – $61^7/64^7$) when the NNN codon is used. The difference in the percentage of sequences containing stop codons increases as the peptide length increases to a maximum difference of 15%, which occurs when the peptide is 25 amino acids long (Figure 4.8).

Another problem with using degenerate codons to create combinatorial peptide libraries is that they generally introduce amino acid biases. Amino acids are overdefined, meaning that a single amino acid is encoded by multiple codons. Depending on the degenerate codon, certain amino acids will be over represented and others under represented. Peptide libraries encoded by the NNK codon have a bias towards the amino acids leucine (12.5%), serine/arginine (9.4%), and alanine, glycine, proline, threonine, and valine (6.3%). The remaining amino acids cysteine, aspartate, glutamate, phenylalanine, histidine, isoleucine, lysine, methionine, asparagine, glutamine, tryptophan, tyrosine, and the stop codon occur at a frequency of 3.1% (Table 4.1).

The combinatorial oligonucleotide library was cloned into the *Rsr*II restriction site of pIN01 using the following strategy. First, the oligonucleotide library was amplified using primers that hybridized to either side of the random region (Figure 4.7). The amplification primers extended the length of the oligonucleotide library by increasing the region that was complementary to the ends of the *Rsr*II-digested pIN01 plasmid from 39-nucleotides to 59-nucleotides. Hua *et al.* showed previously that a 20-nucleotide overlap results in recombination efficiencies of 3.4 ± 0.4% (Hua *et al.*, 1997) and increasing the overlap length to 40-nucleotides gave recombination efficiencies of 93.3 ± 1.9% (Hua *et al.*, 1997). The 59-nucleotide overlap between the combinatorial oligonucleotide and pIN01 was chosen to maximize recombination frequency (> 90%) (Ma *et al.*, 1987) and minimize primer length. A longer primer could have been used to increase efficiency modestly, however longer synthetic DNA oligonucleotides contain more errors and are more expensive. The PCR amplified combinatorial oligonucleotide was cloned into *Rsr*II–digested pIN01 using homologous recombination, which produced a library of lariat intein expression plasmids, referred to as pIL-R7. The homologous recombination cloning strategy also removed the *Rsr*II insertion site in pIL-R7 (Figure 4.7). A library of approximately 20 x 10^6 members was created by performing 100 yeast transformations. Optimization of this procedure resulted in efficiencies up to 5 x 10^5 members per transformation.

Table 4.1 | Amino acids encoded by the NNK degenerate codon
The NNK codon codes for 32 codons. The codon along with the single amino acid code and the amino acid it codes for are included. The 21 different amino acids encoded by the NNK codon are shown along with the number of codons for each amino acid or stop codon (#) and their frequency (%) in the library.

	N N K	Amino Acid	
1	T A G	*	Stop
2	G C G	A	Alanine
3	G C T	A	Alanine
4	T G T	C	Cysteine
5	G A T	D	Aspartic Acid
6	G A G	E	Glutamic Acid
7	T T T	F	Pheylalanine
8	G G G	G	Glycine
9	G G T	G	Glycine
10	C A T	H	Histidine
11	A T T	I	Isoleucine
12	A A G	K	Lysine
13	C T G	L	Leucine
14	C T T	L	Leucine
15	T T G	L	Leucine
16	A T G	M	Methionine
17	A A T	N	Asparagine
18	C C G	P	Proline
19	C C T	P	Proline
20	C A G	Q	Glutamine
21	A G G	R	Arginine
22	C G G	R	Arginine
23	C G T	R	Arginine
24	A G T	S	Serine
25	T C G	S	Serine
26	T C T	S	Serine
27	A C G	T	Threonine
28	A C T	T	Threonine
29	G T G	V	Valine
30	G T T	V	Valine
31	T G G	W	Tryptophan
32	T A T	Y	Tyrosine

	Amino Acid	#	Percent (%)
1	A	2	6.3
2	C	1	3.1
3	D	1	3.1
4	E	1	3.1
5	F	1	3.1
6	G	2	6.3
7	H	1	3.1
8	I	1	3.1
9	K	1	3.1
10	L	3	9.4
11	M	1	3.1
12	N	1	3.1
13	P	2	6.3
14	Q	1	3.1
15	R	3	9.4
16	S	3	9.4
17	T	2	6.3
18	V	2	6.3
19	W	1	3.1
20	Y	1	3.1
21	*	1	3.1

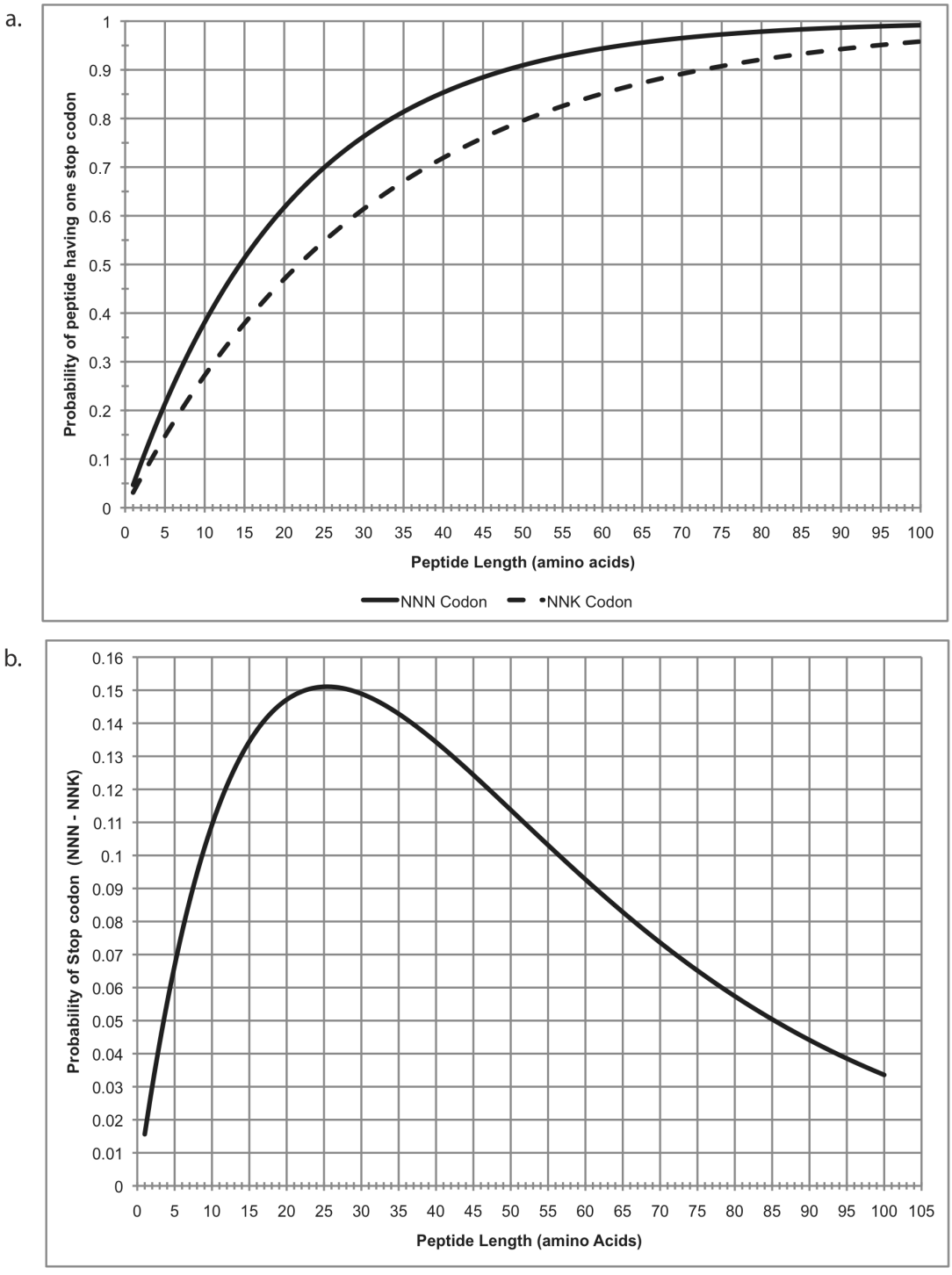


Figure 4.8 | Stop codon frequency in NNK versus NNN codons
(a) The probability of a random sequence containing at least one stop codon versus its peptide length for the NNN (solid line) or NNK (dashed line) codons. **(b)** Differences in the probability of a stop codon occurring between the NNK and NNN codons versus the peptide length. The difference in probability of a stop codon occurring for a peptide of given length between the NNN and NNK codons is calculated by subtracting the NNK curve from the NNN curve in (a).

4.1.2 Characterization of *Ssp-Ssp* R7 intein library

4.1.2.1 *Ssp-Ssp* R7 intein library sequencing

Eighty-four randomly chosen plasmids from the naïve *Ssp-Ssp* R7 library were sequenced to determine the diversity and quality of the library (Figure 4.9). Approximately 32% of the sequences did not contain an insert either due to self-ligation or due to incomplete digestion and purification of the pIN01 plasmid. The no-insert plasmid background did not cause a significant problem when screening the library and was used as a control to detect non-specific interactions between the lariat and target protein or to identify self-activating baits. Even though 32% of the lariat library did not contain a combinatorial peptide insert, empty plasmid sequences were never isolated in screens using this library, indicating that the intein scaffold was inert and did not interact with protein targets.

The remaining 68% of sequences underwent homologous recombination, which resulted in either in-frame or out-of-frame sequences. Out-of-frame sequences made up approximately 18% of the total library. Out-of-frame sequences displayed linear peptides consisting of the I_C domain and a random peptide sequence that varied in length depending on the frame-shift location. Out-of-frame sequences arise when the combinatorial oligonucleotide insert is longer or shorter than expected. This can be due to errors during oligonucleotide synthesis or errors occurring during PCR amplification of the oligonucleotide. Alternatively, an error during recombination can also introduce a frameshift. Out-of-frame sequences can be minimized by gel purifying the degenerate oligonucleotide or by purifying the resulting library PCR product to ensure only products of the correct length are used to construct the lariat libraries.

Approximately 50% of the total library underwent in-frame homologous recombination. In-frame sequences consisted of: lariat inteins, with sequences expected to form lariats; inactive inteins that contained alanine at the I_{N+1} (A1) position, or linear peptide sequences that contained stop codons. Sequences that encoded lariats represented 31% of the total library. Inactive inteins containing an alanine at position (I_{N+1}/A1) represented 6% of the total library. The presence of alanine at (I_{N+1}/A1) was most likely due to a recombination artifact. pIN01 has alanine at this position and the library insert has cysteine (Figure 4.7). If recombination occurs after the I_{N+1}/A1 position then alanine from the pIN01 plasmid is retained. This problem can be addressed by modifying pIN01 to contain cysteine at position I_{N+1}/A1.

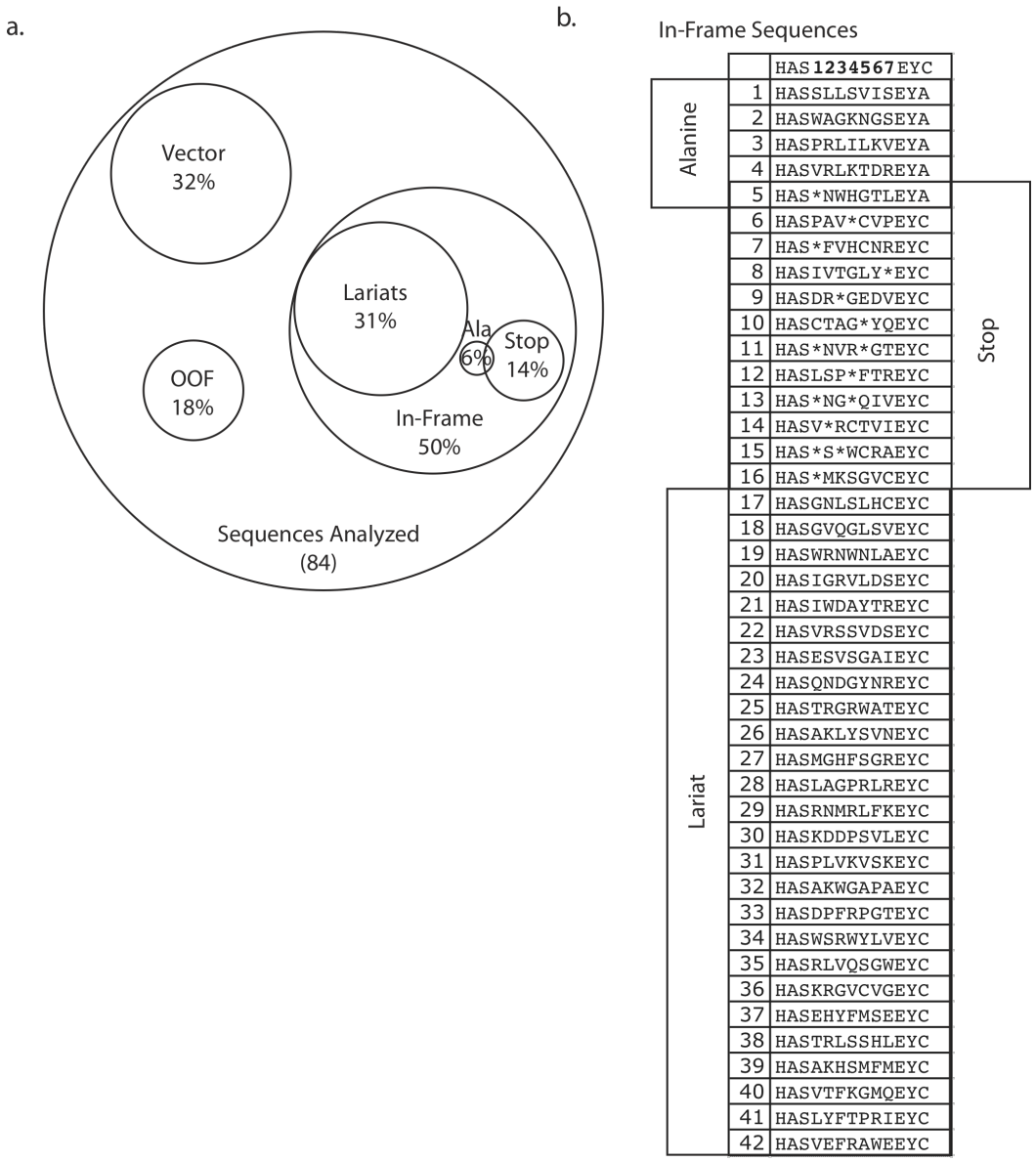


Figure 4.9 | Sequences of naïve *Ssp-Ssp* R7 library members
(a) Venn diagram showing the different classes of sequences present in the lariat library. OOF: represents out-of-frame sequences. Vector: represents no-insert sequences. Lariats: represent sequences capable of forming lariats. Ala: represents sequences containing alanine at I_{N+1}/A1. Stop: represents sequences containing a stop codon. In-frame: represents sequences that are in frame. **(b)** Sequences of naïve library members. The combinatorial noose regions were sorted into lariats, peptides that contained stop codons (Stop), or peptides that contained alanine at the I_{N+1} (A1) position resulting in an unprocessed intein (Alanine). Single letter amino acid codes are used to represent amino acids. * indicates a stop codon. "1234567" represents the positions of the seven amino acids in the randomized region.

Linear peptides containing stop codons represented 21% of the library. This was very close to the theoretical predicted value of 20% for the NNK codon. Stop codons and codon biases can be eliminated by using trinucleotide phosphoramidites instead of degenerate nucleotides (Sondek and Shortle, 1992; Gaytán *et al.*, 1998; Kayushin *et al.*, 2000; Mauriala *et al.*, 2004).

In summary, we constructed a lariat peptide expression plasmid library (pIL-R7) in the MATa yeast strain EY93 using homologous recombination. A total of 20 million library members were constructed. To estimate the number of library members that correctly produced the lariat in yeast, we sequenced eighty-four randomly chosen library members and showed that ~ 31% had the correct sequence with no stop codons and the proper nucleophile at the I_{N+1} (A1) position. Therefore, the final library contained approximately six million different lariats (20 x 10^6 x 0.31 = 6.2 x 10^6).

4.1.2.2 Characterization of lariat formation in the *Ssp-Ssp* R7 intein library

The lariat library contained an N-terminal HA-tag, which allowed lariat processing to be monitored using Western blot analysis with an anti-HA antibody. Expression of lariats in yeast cells (EY93) resulted in the production of the unprocessed lariat intein (~32 kDa) and the lariat (~18 kDa) (Figure 4.10). If processing did not occur then only the unprocessed lariat intein was observed (Figure 4.10). Twenty library members that were predicted to produce a lariat based on their sequences were characterized by Western blot analysis. Approximately 70% of these constructs actually produced the lariat (Figure 4.10). Based on this analysis, it was estimated that approximately 4.4 million (6.2 x 10^6 x 0.71 = 4.4 x 10^6) library members were capable of producing lariats.

Several groups reported that amino acids at positions, I_{C+2}, I_{C+3}, I_{N-1}, and I_{N-2} in the extein affect splicing (Scott *et al.*, 2001; Naumann *et al.*, 2005; Iwai *et al.*, 2006; Amitai *et al.*, 2009), which in our case is in the noose or combinatorial region of the lariat. Since we only analyzed twenty library members for the ability to process we were not able to determine if there was a preference for specific amino acids at these positions. Interestingly, the presence of tryptophan at position I_{C+2} in two sequences blocked processing, suggesting that tryptophan may not be tolerated at this position.

An analysis of more library members would provide further insight into which amino acids or combinations of amino acids block lariat formation. This information would be useful

for future library design strategies as libraries could be designed that exclude amino acids that block splicing at the I_{C+2}, I_{C+3}, I_{N-1}, and I_{N-2} positions. Alternatively, this sequence information could be used to directly identify library members isolated in screens that are capable of forming lariats allowing us to focus our analysis on these members.

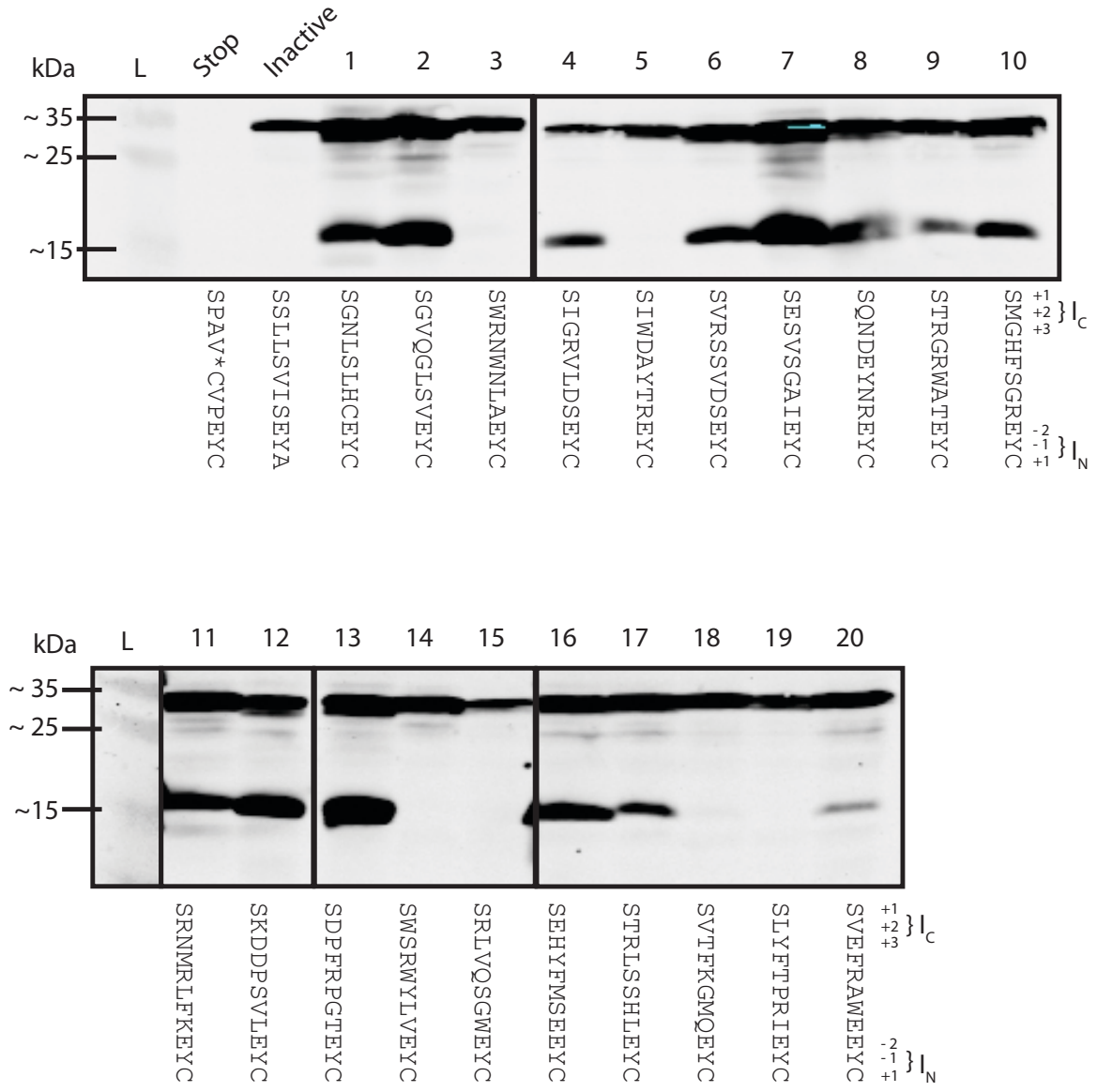


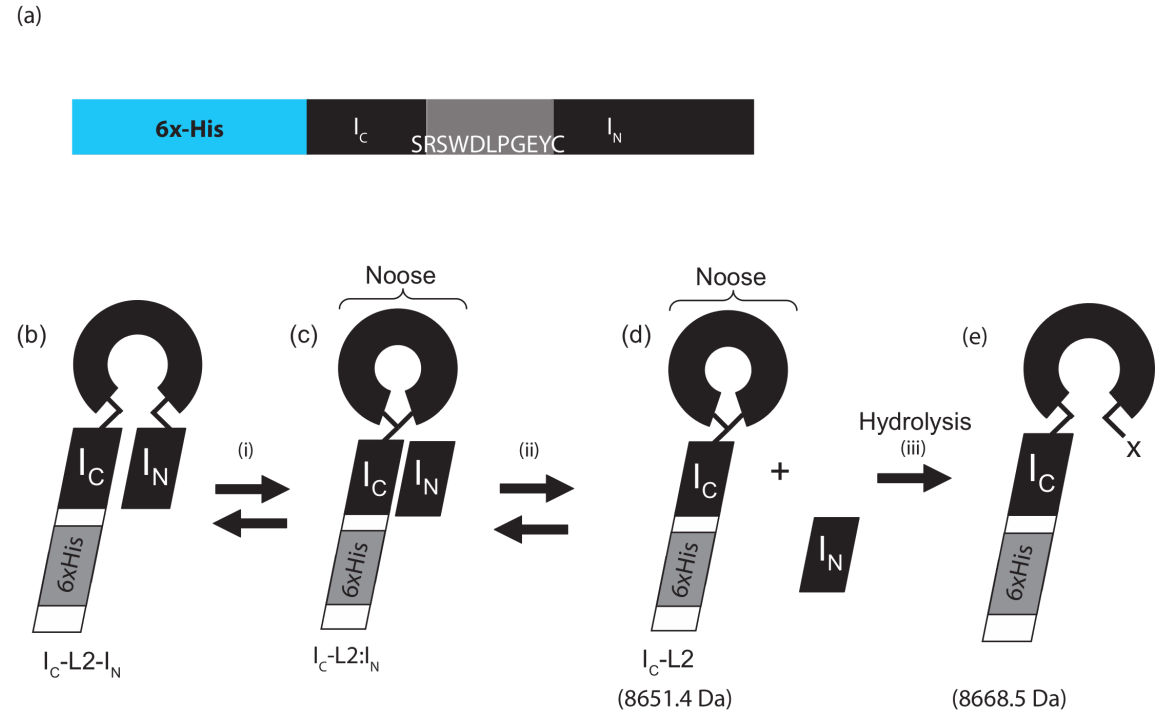
Figure 4.10 | Analysis of *Ssp-Ssp* R7 library processing by Western blot analysis
Naïve library members (1-20) expressed in yeast from the *Ssp-Ssp R7* library were sequenced and analyzed by Western blot analysis using the anti-HA antibody. L = Ladder. Stop and inactive represent two control sequences containing a stop codon or a cysteine to alanine mutation at position (I_{N+1}/A1), respectively. The sequences from positions (I_{C+1}/G8) to (I_{N+1}/A1) for each member are shown below the Western blot. The unprocessed product runs at ~ 32 kDa and the lariat runs at ~ 18 kDa.

4.1.2.3 Characterization of *Ssp-Ssp* R7 intein library member "L2" by LC/MS

The presence and importance of the lariat lactone bond was analyzed using an anti-LexA lariat, referred to as L2. The anti-LexA L2 lariat was isolated using the Y2H assay and is described in detail in section 4.3. Western blot analysis showed that the anti-LexA L2 lariat of the correct size was produced, however it did not confirm if the lariat lactone bond was intact within the cell. Two experiments were performed to confirm the existence of the lariat lactone bond. One experiment directly analyzed whether the lactone bond was formed using liquid chromatography followed by electrospray ionization mass spectrometry (LC/ESI-MS). The other experiment evaluated whether the lactone bond was required for the lariat to bind its target in the Y2H assay.

4.1.2.3.1 Direct LC/MS analysis

To verify whether the L2-lariat was produced in *E. coli*, we used (LC/ESI-MS). The lariat was analyzed in *E. coli* since the lariat could be expressed at higher levels relative to yeast cells. The L2 lariat was subcloned from the plasmid pIN01 to the *E. coli* expression plasmid (pETIL-L2). The pETIL-L2 expression plasmid produced the L2 lariat with an N-terminal histidine tag followed by the I_C domain, the L2 peptide, and the I_N domain (Figure 4.11). The His-tag L2 lariat was purified using Ni^{2+}-NTA affinity chromatography. The His-tag purified lariat products were further purified and desalted using reverse phase LC. Optimal protein expression conditions were determined and the protein was purified, lyophilized, and stored at – 80 °C. Lyophilization did not affect the stability of the lariat. MS analysis of the purified L2 lariat showed two products near the calculated molecular weight of the L2 lariat. The smaller product (8651 Da) corresponded to the L2 lariat, which represented 26% of the total product. The larger product (8669 Da) corresponded to a hydrolyzed lariat product, which represented 74% of the total product (Figure 4.11). Previous MS analysis on the cyclic peptide producing intein reaction reported the lariat intermediate as being present mainly in the hydrolyzed form (Scott *et al.*, 1999; Scott *et al.*, 2001). Potentially, hydrolysis of the lactone bond may have been caused by the high temperatures and acidic conditions used in the MS analysis.



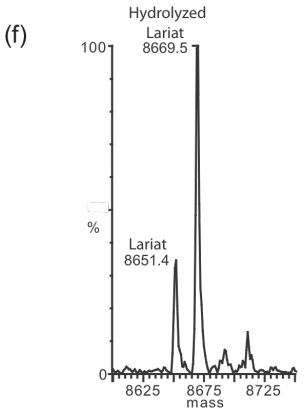


Figure 4.11 | Direct LC/MS analysis of the L2 lariat
(a) Schematic of the L2 lariat construct used for LC/MS analysis. 6xHis represents a six-histidine tag. I_N and I_C are the N-terminal and C-terminal intein domains, respectively. The anti-LexA L2 peptide sequence is shown in the noose region. **(b)** Unprocessed intein represented by I_C-L2-I_N. The I_C domain was covalently attached to the L2 peptide sequence and the I_N domain. **(c)** The unprocessed intein processes into the I_C-L2:I_N where the colon indicates that the I_N domain is not covalently attached. (**d**) The I_N domain is released, which results in the lariat product (I_C-L2). **(e)** Hydrolysis of the intein results in cleavage of the lactone bond. **(f)** LC/MS analysis of the I_C-L2 construct showing hydrolyzed and non-hydrolyzed lariat.

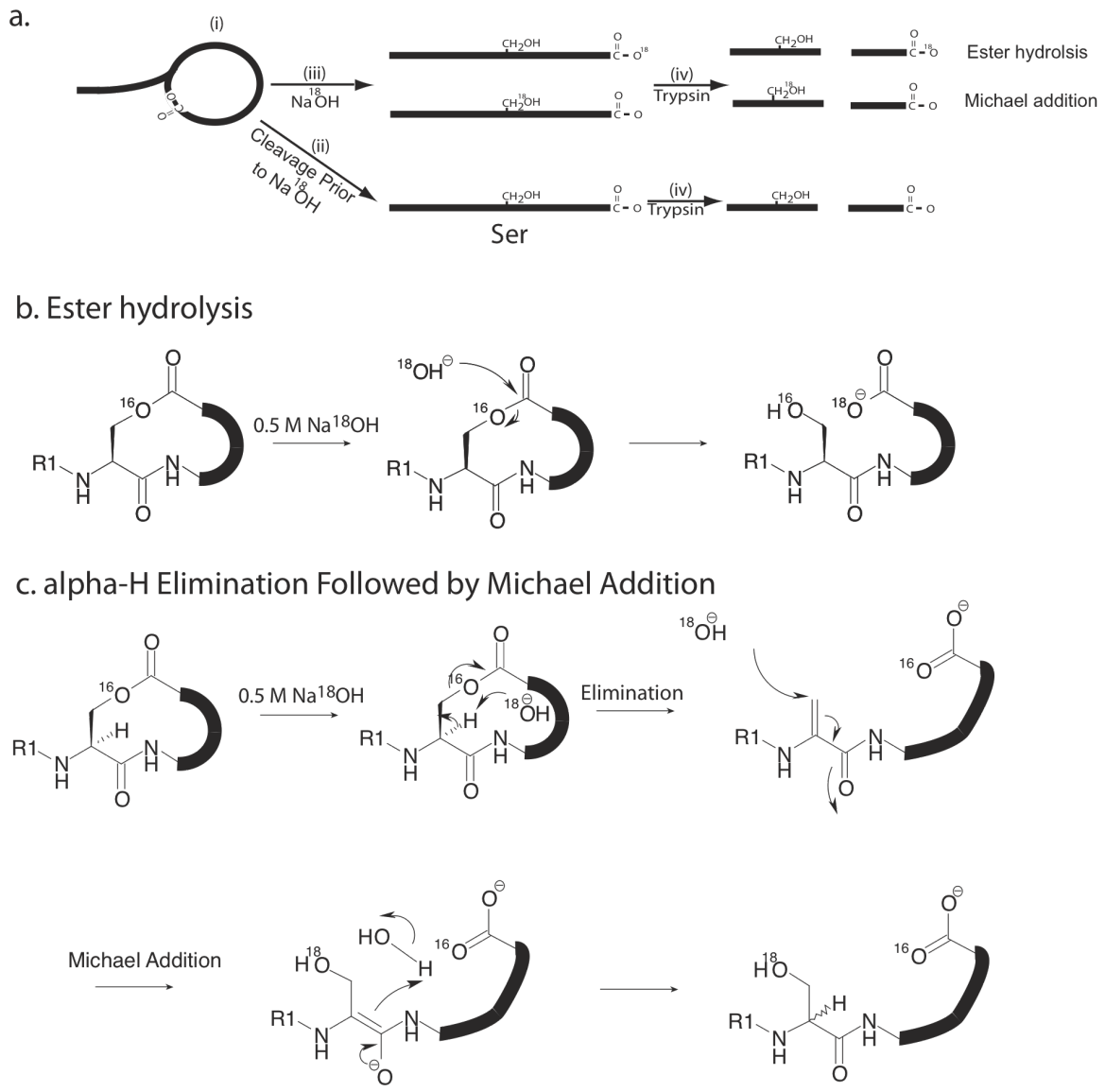


Figure 4.12 | Isotopic analysis used to quantitate the lactone bond
(a) Overview of lariat hydrolysis experiment. (i) Purified lariat before forced hydrolysis. (ii) Lariat hydrolysis occurring inside of the cell results in the incorporation of ^{16}O into the linearized lariat. (ii) Forced lariat hydrolysis using $Na^{18}OH$, results in the incorporation of ^{18}O into the linearized lariat. (iv) Trypsin cleavage results in the two potential ^{18}O incorporation sites being separated onto two different peptides. **(b)** $Na^{18}OH$ induced ester hydrolysis results in the ^{18}O incorporation at the carboxyl terminus of the hydrolyzed lariat. **(c)** $Na^{18}OH$ induced α-H elimination followed by Michael addition results in the ^{18}O incorporation on the serine side chain.

4.1.2.3.2 Indirect LC/MS analysis of the lariat

To determine the amount of lariat present prior to MS analysis, we forced the hydrolysis of the lariat with $Na^{18}OH$. The intact lariat was hydrolyzed by $Na^{18}OH$, which results in the incorporation of ^{18}O. The ^{18}O labeled hydrolyzed lariat can be distinguished from the lariat hydrolyzed by $H_2{}^{16}O$ prior to $Na^{18}OH$ treatment by MS analysis. The L2 lariat was expressed in *E. coli* and purified with a nickel affinity column followed by HPLC and then lyophilized. The purified lariat (I_C-L2) was treated with $Na^{18}OH$, neutralized, desalted by HPLC, lyophilized, and digested with trypsin. The resulting peptide fragments were characterized by LC/MS (Hagelin, 2005). Treatment with NaOH results in a chemoselective lactone ring opening, which leaves amide bonds intact. Heavy oxygen (^{18}O) was incorporated into peptides that contained the lactone bond, but not lariats that had been hydrolyzed prior to the treatment. The lactone bond was selectively cleaved by either base-catalyzed hydrolysis or by α-H elimination generating dehydroalanine, which was then modified by a Michael addition (Figure 4.12). Base-catalyzed hydrolysis resulted in an incorporation of ^{18}O into the carboxyl group at the C-terminus of the lariat. α-H elimination, followed by Michael addition resulted in an incorporation of ^{18}O into the hydroxyl group of the serine side chain. Trypsin digestion was necessary to simplify the analysis and separate these two modifications, allowing the two amino acids involved in the lactone bond to be analyzed separately. Lariats that were hydrolyzed prior to $Na^{18}OH$ treatment had the natural isotopic distribution of oxygen (primarily ^{16}O) incorporated. Lariats that were hydrolyzed by the $Na^{18}OH$ treatment would be selectively labeled with ^{18}O and have a pronounced change from the natural isotopic profile. The ratio of ^{16}O:^{18}O in the $Na^{18}OH$ treated peptides was used to determine the amount of L2 lariat with a lactone bond prior to MS analysis (Figure 4.12).

Trypsin digestion resulted in the peptide SWDLPGEY (966.42 m/z), which contained the amino acids 73-80 from the intein construct and IFDIGLPQDHNFLLANGAIAHASR (2590.352 m/z), which contained amino acids 49-72 of the intein construct (Figure 4.13). The SWDLPGEY (73-80) peptide fragment contained the tyrosine carboxyl group involved in the lactone bond on the last tyrosine residue. The IFDIGLPQDHNFLLANGAIAHASR (49-72) peptide fragment contained the serine hydroxyl involved in the lactone bond.

As a control, the peptides were treated with $Na^{16}OH$ and analyzed by MS. Their spectrums were compared to the theoretical isotopic distribution patterns calculated with MS-ISOTOPE software (Clauser *et al.*, 1999). Peptides treated with $Na^{18}OH$ contained a mixture of ^{18}O and ^{16}O labeled peptides, were deconvoluted to determine the ratio of ^{18}O:^{16}O peptides using MATCHING software (Fernández-de-Cossio *et al.*, 2004). Matching software deconvolutes mixtures of peptides that have been isotopically labeled, resulting in small differences in molecular weight by comparing their observed isotopic patterns to their predicted isotopic patterns. We next fitted the observed isotopic distribution of the $Na^{18}OH$ treated peptides to the theoretical distribution based peptides containing a mixture of ^{16}O and ^{18}O.

For the 73-80 amino acid fragment (966 m/z, which corresponds to the singly charged peptide), the $Na^{16}OH$ treated profile closely matched the theoretical isotopic distribution as expected (Figure 4.14). In the $Na^{18}OH$ treated sample there was a large deviation from the theoretical isotopic distribution at 968 m/z (Figure 4.14). This deviation, which was two daltons heavier than the 966 m/z product, indicating that an ^{18}O was incorporated and showed there was a mixture of two different peptides with different isotopic profiles. One peptide contained the natural isotopic distribution of oxygen (with ^{16}O being the most abundant) and one peptide was specifically labeled with ^{18}O. The percentage of peptides that had either O^{16} or O^{18} incorporated assuming only a single oxygen substitution due to hydrolysis of the lactone bond was determined. MATCHING software calculated the percentage of ^{16}O and ^{18}O containing peptides to be 86% and 14%, respectively (Figure 4.14). When the calculated values were summed together they closely matched the observed spectra (Figure 4.14).

For the 49-72 amino acid fragment, we expected the singly charged peptide to occur at 2590 m/z (Figure 4.13). However, the 2590 m/z fragment was not observed in the $Na^{16}OH$ treated sample. The $Na^{16}OH$ treated 49-72 amino acid fragment had an observed mass of 2591.352 m/z, which corresponded to a product one dalton heavier than that predicted mass for the 49-72 amino acid fragment treated with $Na^{16}O$ (2590.35 m/z). A one dalton shift caused by $Na^{16}OH$ was attributed to deamidation of asparagine. The asparagine at position (I_{C-7}) was susceptible to base-catalyzed deamidation as it was N-terminal to a glycine, which has previously been shown to promote deamidation (Aswad *et al.*, 2000; Takehara and Takahashi, 2003). Deamidation results in the exchange of the amide group on asparagine (NH_2, MW=16) with a hydroxyl group (OH, MW=17), resulting in aspartic acid or isoaspartic acid. We

predicted that treatment of the 49-72 amino acid fragment with $Na^{18}OH$ would result in the incorporation of two ^{18}O molecules; one from the ester hydrolysis and the second one from deamidation. Analysis of the multiply charged (4+) spectrum for the $Na^{16}OH$ treated peptide assuming deamidation closely matched the theoretical spectrum (Figure 4.15). Analysis of the 49-72 amino acid peptide fragment from the lariat treated with $Na^{18}OH$ showed two interesting features. First, there was a shift in the spectra to the left, causing the two first peaks to almost completely disappear compared to the $Na^{16}OH$ treated sample (Figure 4.15). Since treatment with $Na^{16}OH$ also resulted in significant deamidation, this shift was attributed to deamidation, which resulted in the incorporation of one ^{18}O. Second, there was a deviation at the third major peak that corresponded to a two dalton shift, which was due to a second ^{18}O incorporation (Figure 4.15). MATCHING software was used to determine the percentage of O^{16} and O^{18} incorporation assuming two oxygen incorporations. MATCHING software calculated that 8.8% of the [49-72] peptide fragment had two ^{16}O incorporations (2591.35 m/z), 59.0% had one ^{16}O incorporation and one ^{18}O incorporation (2593.35 m/z), and 32.2% had two ^{18}O incorporations (2595.35 m/z) (Figure 4.15). When the calculated values were summed together, they closely matched the observed spectra (Figure 4.15).

The fraction of ^{18}O incorporated into 73-80 and 49-72 peptide fragments indicated that 46% (14% + 32%) of the lariat contained an intact lactone bond prior to MS analysis. This value represented a low estimate of the amount of L2 lariat present in *E. coli* as some of the lariat may have been hydrolyzed during Ni^{2+}-NTA affinity chromatography, and reverse-phase HPLC purification prior to $Na^{18}OH$ hydrolysis. During purification, lariat hydrolysis was expected to occur due to esterases and other contaminating enzymes present with the purified lariat. This data, combined with the fact that many lactone-cyclized peptides exist in nature (Grünewald and Marahiel, 2006), supported the existence of the lariat structure *in vivo*. Most importantly, these results demonstrated that the lariat formed correctly inside of cells.

a. Analysis of ^{18}O incorporation at the tyrosine carboxylic acid at position (I_{N-1})

Treatment	m/z (calc.)	m/z (obs.)			delta
$Na^{16}OH$	966.4203	966.22	±	0.06	0.20
$Na^{18}OH$	968.4203	966.38	±	0.04	2.04

b. Analysis of ^{18}O incorporation at the serine side chain at position (I_{C+1})

		No Deamidation		Deamidation	
Treatment	m/z (obs.)	m/z (calc.)	delta	m/z (calc.)	delta
None	2590.376	2590.352	0.024	NA	NA
$Na^{16}OH$	2591.348	2590.352	0.996	2591.360	-0.012
$Na^{18}OH$	2593.338	2592.352	0.986	2595.360	-2.022

Figure 4.13 | Initial analysis of ^{18}O incorporation by LC/MS
(a) Analysis of ^{18}O incorporation at the tyrosine carboxylic acid at position (I_{N-1}). Trypsin digestion of the $Na^{18}OH$ treated lariat produced a peptide fragment containing tyrosine at position (I_{N-1}) (SWDLPGEY). The ^{16}O product has a calculated mass of 966.420 m/z and the ^{18}O product has a calculated mass of 968.420 m/z. The major peak observed (obs) in both samples was 966 m/z **(b)** Analysis of ^{18}O incorporation at the serine side chain at position (I_{C+1}). Trypsin digestion of the $Na^{18}OH$ treated lariat produced a peptide fragment containing serine at position (I_{C+1}) (IFDIGLPQDHNFLLANGAIAHASR). The mass of this fragment is 2590.352 m/z corresponding to a product 1 Da heavier than the predicted ^{16}O incorporated product. A 1 Da shift was attributed to deamidation of asparagine. The 2+, 3+ 4+ and 5+ charged fragments were analyzed and similar results were obtained. Only the 4+ charged fragment is shown.

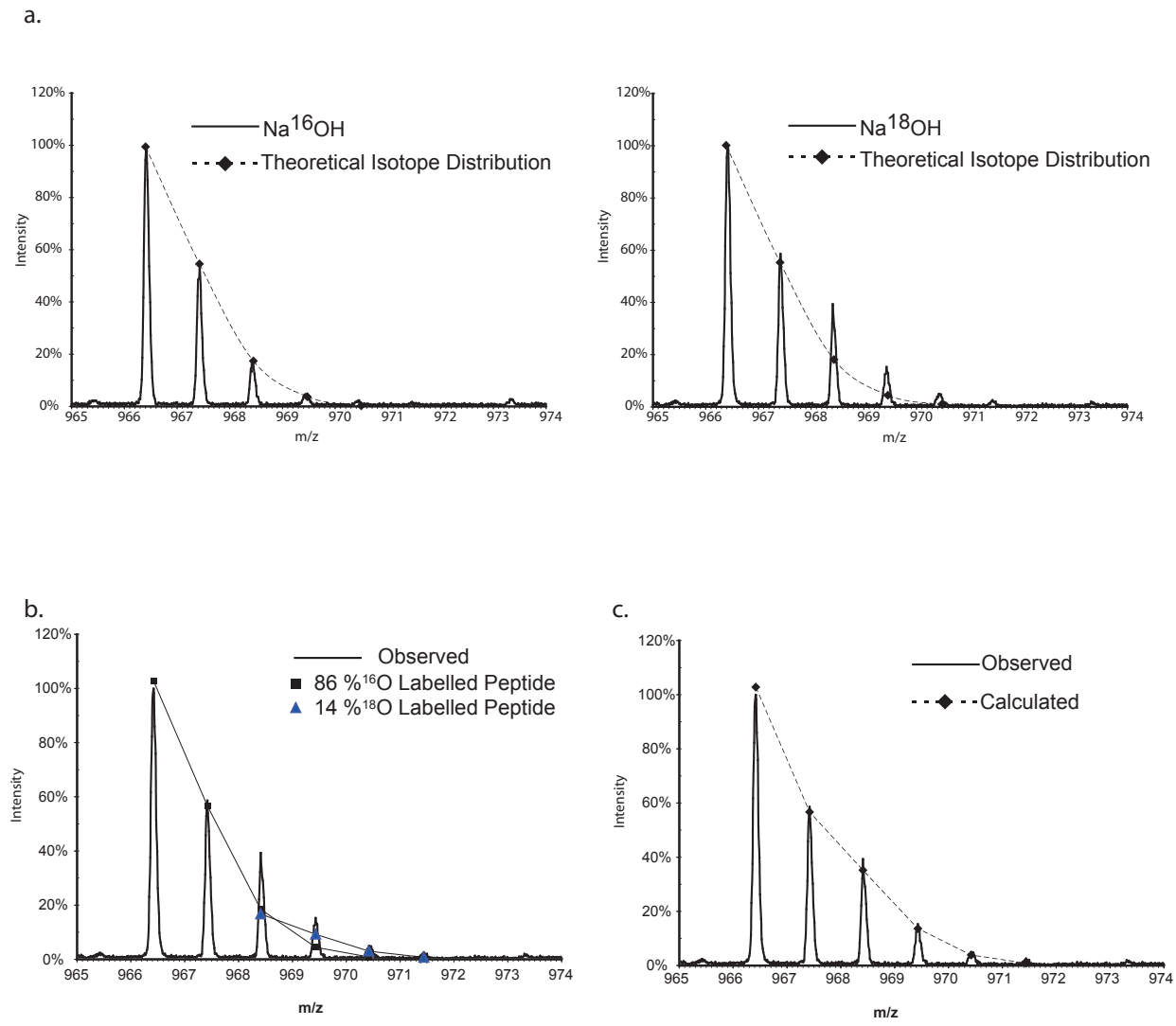


Figure 4.14 | Analysis of incorporation of ^{18}O at the [73-80] peptide fragment
(a) Mass spectrometry analysis of [73-80: SWDLPGEY] peptide fragment from the $Na^{16}OH$ or $Na^{18}OH$ treated samples overlayed with the theoretical isotope distribution calculated using MS-ISOTOPE software. In the $Na^{18}OH$ treated sample there is a large deviation from the theoretical distribution indicating the presence of more than one peptide. **(b)** MATCHING software analysis of the percentages of ^{16}O labeled peptide (966.3 m/z, 86%, squares) and ^{18}O labeled peptide (968.3 m/z, 14%, triangles) in the observed spectrum. The 1+ and 2+ charged fragments were analyzed and similar results were observed. Only the 1+ charge is shown. **(c)** Overlay of the sum of the calculated contributions of the ^{18}O and ^{16}O peptides on the observed SWDLPGEY peptide fragment spectrum.

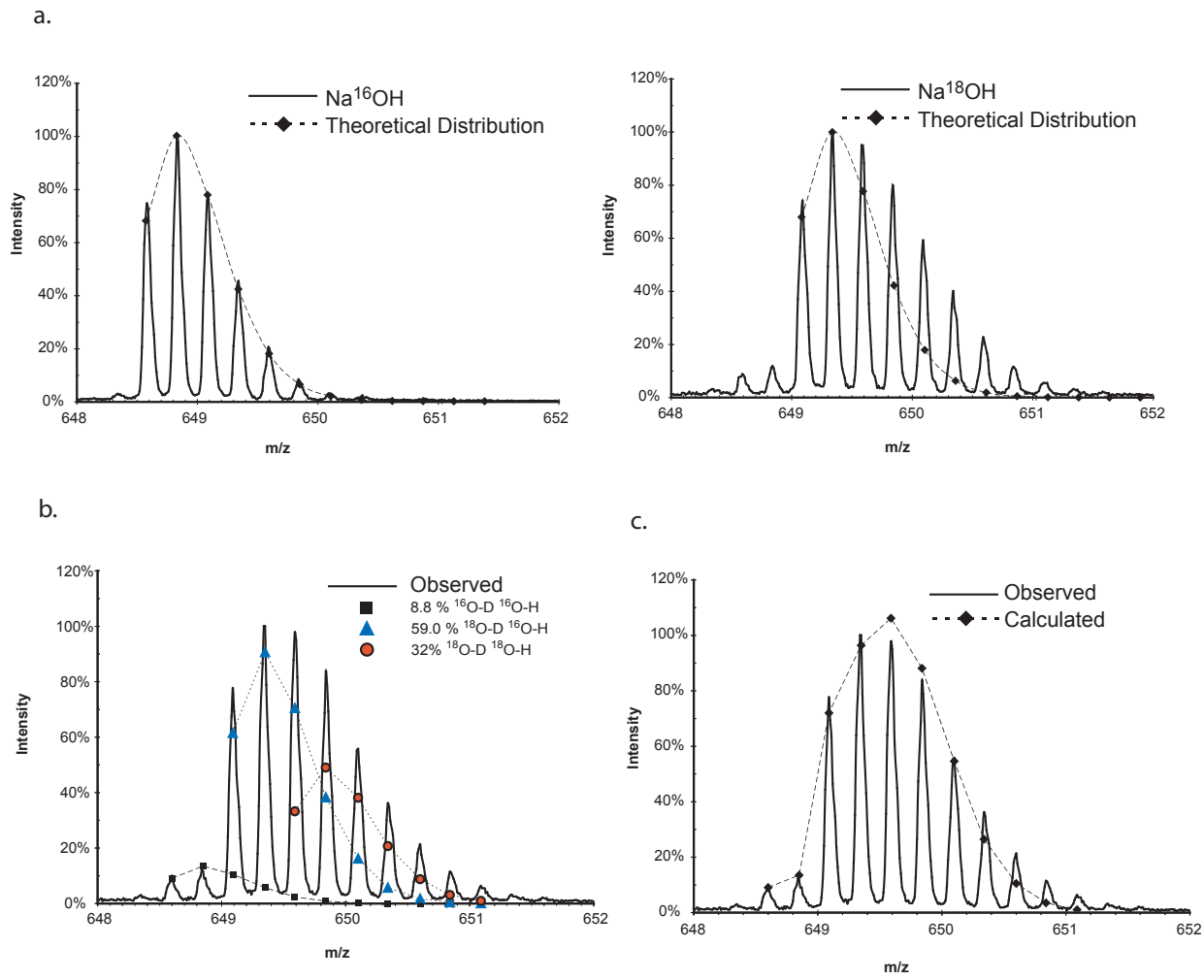


Figure 4.15 | Analysis of the incorporation of ^{18}O at the [49-72] peptide fragment
(a) Mass spectrometry analysis of [49-72: IFDIGLPQDHNFLLANGAIAHASR] peptide fragment from the $Na^{16}OH$ or $Na^{18}OH$ treated samples overlayed with the theoretical isotope distribution (MS-ISOTOPE software). The $Na^{18}OH$ treated sample incorporated two ^{18}O, one from the hydrolysis and the second from deamidation, resulting in a M+H of 2595.36 Da. **(b)** MATCHING software analysis of the percentages of ^{16}O and ^{18}O labeled peptides: (■) a peptide with two ^{16}O substitutions corresponding to deamidation and hydrolysis by ^{16}O (2591.35 m/z, 8.8 %, squares), (▲) a peptide with one ^{16}O and one ^{18}O substitution corresponding to deamidation by ^{18}O and hydrolysis by ^{16}O (2593.35 m/z, 59.0 %, triangles), and (●) a peptide with two ^{18}O substitutions (2595.35 m/z, 32.2%, circles) corresponding to deamidation and hydrolysis by ^{18}O. D = deamidation and H = hydrolysis. **(c)** Overlay of the sum of the calculated contributions of the ^{18}O and ^{16}O peptides on the observed IFDIGLPQDHNFLLANGAIAHASR peptide fragment spectrum.

4.1.3 *Ssp-Ssp* intein processing in the Y2H assay

We demonstrated that the lariat lactone bond was formed within cells using Western blot analysis and LC/MS. We next addressed whether the lactone bond was important for the lariat-target interaction in the Y2H assay. In the Y2H assay, expression of the lariat resulted in a mixture of the unprocessed lariat intein, the lariat, and the hydrolyzed lariat (Figure 4.16). Since expression of the lariat resulted in a mixture of products, we could not determine which species interacted with the target. To address this, we constructed mutant L2 inteins that expressed solely the L2 inactive intein or the L2 linear peptide. The L2 inactive intein produced only unprocessed intein and the L2 linear peptide produced only the hydrolyzed lariat (Figure 4.17). These mutant L2 inteins were tested for their interaction with LexA using the Y2H assay (Figure 4.17). No interaction between the L2 inactive intein and the L2 linear peptide with LexA would indicate that the lariat was the only construct capable of interacting with LexA. This experiment was crucial to determine if the imposed lariat

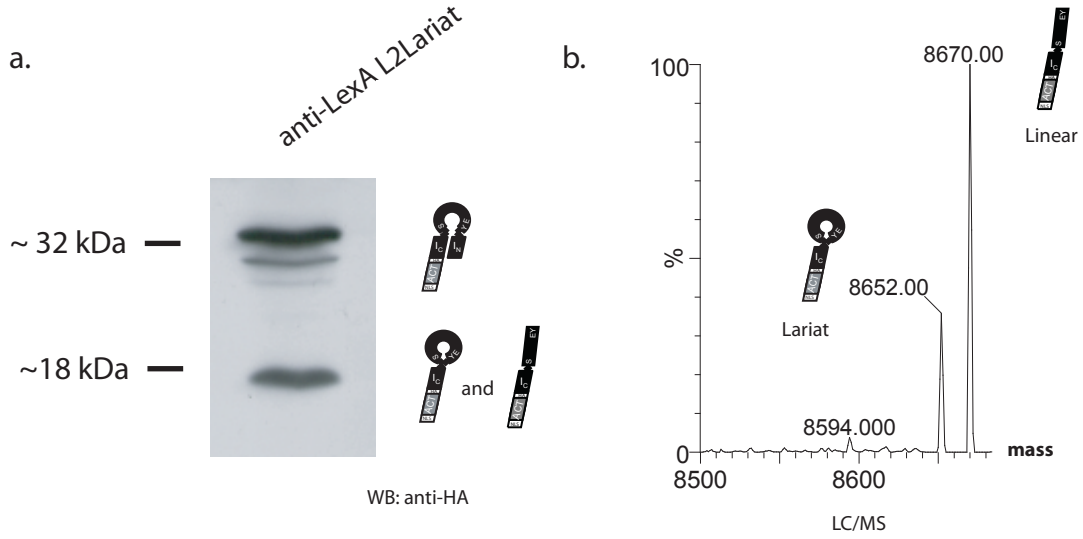


Figure 4.16 | Anti-LexA L2 processing
(a) Western blot analysis of L2 lariat processing in yeast using an anti-HA antibody confirmed the presence of an HA-tagged protein corresponding to the molecular weight of an unprocessed intein ~ 32 kDa and an HA-tagged protein corresponding to the lariat ~ 18 kDa. **(b)** LC/MS analysis of lariat processing in *E. coli* shows two products corresponding to the molecular weight of the lariat and hydrolyzed lariat.

constraint was required for the interaction with its target and if screening cyclized peptides had benefits over linear peptides.

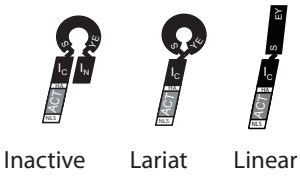
The following inteins were constructed and their expression levels were verified by Western blot analysis: (i) L2 lariat, (ii) L2 inactive intein, and (iii) L2 linear peptide (Figure 4.17). The L2 lariat expression plasmid produced a mixture of products including the unprocessed lariat, the lariat, and the hydrolyzed lariat. The L2 lariat had the noose sequence (I_{C+1}/G8)-**S**RSWDLPGE**Y**-(I_{N-1}), which was constrained by a lactone bond between serine (I_{C+1}/G8) and the carboxyl group of the tyrosine (I_{N-1}). The L2 inactive intein had the sequence (I_{C-2}/G6)-HASRSWDLPGEY**A**-(I_{N+1}/A1). The cysteine at position I_{N+1} (A1) was mutated to alanine, which prevented the L2 inactive intein from undergoing the first step in intein processing and thus it could not produce the lariat. The L2 linear peptide had the sequence (I_{C-2}/G6)-HASRSWDLPGEY**-(I_{N+2}/A2), where * represents a stop codon. The L2 linear peptide was identical to the hydrolyzed lariat. The pIN01 plasmid described previously (Figure 4.6) was included as a negative control. pIN01 produced an inactive intein with the sequence (I_{C-2}/G6)-HSSCGPCEYA-(I_{N+1}/A1).

In the Y2H assay, we did not detect any interactions between LexA and L2 inactive or L2 linear peptides (Figure 4.17). Only the L2 lariat interacted with LexA in the Y2H assay, which highlighted the importance of the lariat conformation in this interaction.

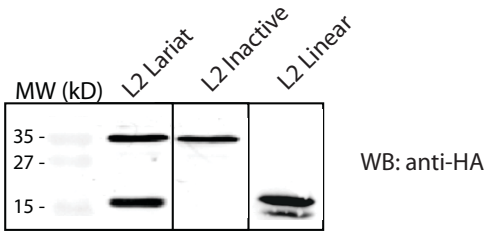
It was also important to show that the requirement for a lactone constraint was a general feature for lariat interactions with other target proteins and that it was not unique to the anti-LexA L2 lariat/LexA interaction. To show this, we randomly tested one lariat isolated against the PR domain of Riz1 and one lariat isolated against the JH1 domain of Jak2. Neither of these lariats interacted with the target when expressed as a linear peptide (Figure 4.18). This result demonstrated that the lariat constraint was required for the interaction of more than one type of lariat. We expect that some lariats will not require the lariat constraint. Kinsella *et al.* have previously shown that cyclization is important for the function of many lactam selected peptides, but some of the cyclic peptides isolated in their screen were still capable of interacting when expressed in an inactive intein construct (Kinsella *et al.*, 2002). Further analysis will be required to determine the percentage of lariats isolated that can also interact in the inactive or linear conformations.

a.

L2 Constructs	I_C	I_C-1	I_C+1	Random Library	I_N+1	I_N
(i) Lariat	I_C	Ala	Ser	RSWDLPGEY	Cys	I_N
(ii) Inactive	I_C	Ala	Ser	RSWDLPGEY	Ala	I_N
(iii) Linear	I_C	Ala	Ser	RSWDLPGEY	Stop	I_N
(iv) Vector (pIN01)	I_C	Ser	Ser	CPGCEY	Ala	I_N



b.



c.

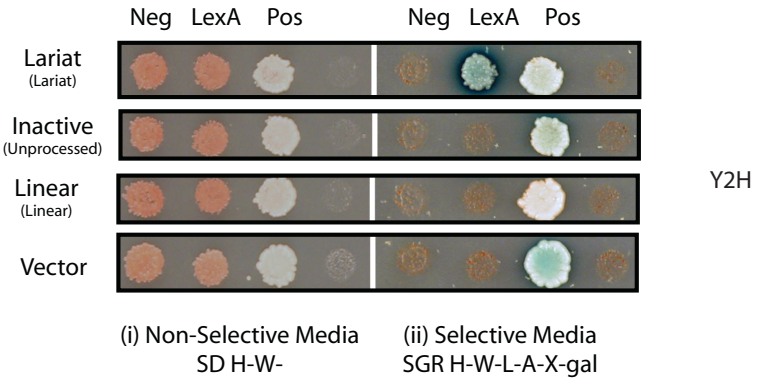


Figure 4.17 | Linear and lariat analysis

(a) Constructs used in the Y2H. Four different constructs were analyzed: (i) Lariat, (ii) Inactive, (iii) Linear, and (iv) vector. The mutations required to generate these constructs are shown. **(b)** Western blot analysis of the L2 lariat inteins. HA-tagged L2 lariat, L2 inactive, and L2 linear expression were constructed and analyzed by Western blot blots using and anti-HA antibody. **(c)** Y2H assay. Y2H analysis of the interaction of LexA with L2 lariat and intein constructs. pIN01 is an inactive lariat expression plasmid that expresses an inactive intein with a CGPC peptide noose. (i) Yeast growth on non-selective synthetic dextrose (SD) His^-, Trp^- media. (ii) Yeast growth on SGR His^-, Trp^-, Leu^-, Ade^-, X-Gal media, which selects for the activation of LEU2, ADE2, and LacZ Y2H reporter genes. Neg: prey was mated to His+ plasmid with no bait. LexA: prey was mated to pEG202 containing LexA bait. Pos: prey was mated to pEG202::B42AD self-activating bait.

a.

Constructs	I_C	I_C-1	I_C+1	Random Library	I_N+1	I_N
(i) anti-LexA Lariat	I_C (*Ssp*)	Ala	Ser	RSWDLPGEY	Cys	(*Ssp*)
(ii) anti-LexA Linear	I_C (*Ssp*)	Ala	Ser	RSWDLPGEY	Stop	(*Ssp*)
(iii) anti-Jak2 Lariat	I_C (*Ssp*)	Ala	Ser	HAFQIMIIHWYIP	Cys	(*Npu*)
(iv) anti-Jak2 Linear	I_C (*Ssp*)	Ala	Ser	HAFQIMIIHWYIP	Stop	(*Npu*)
(v) anti-Riz1 Lariat	I_C (*Ssp*)	Ala	Ser	CPIYPRFSVL	Cys	(*Npu*)
(vi) anti-Riz1 Linear	I_C (*Ssp*)	Ala	Ser	CPIYPRFSVL	Stop	(*Npu*)
(vii) pIN01	I_C (*Ssp*)	Ser	Ser	CGPCEY	Ala	(*Ssp*)
(viii) pIL500	I_C (*Ssp*)	Ala	Ser	R	Cys	(*Npu*)

b.

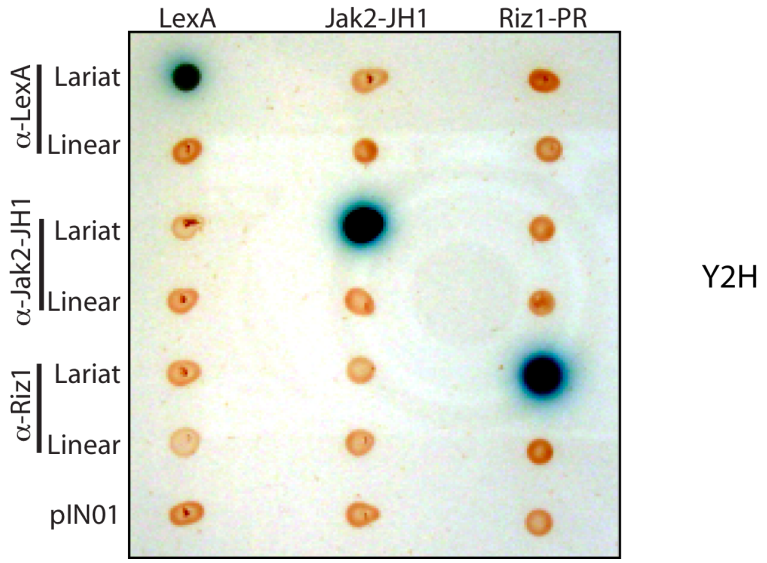


Figure 4.18 | Y2H assay of lariat and linear constructs
Lariats that interact with LexA, Jak2-JH1 domain, and Riz1-PR domain were isolated using the Y2H assay. The anti (α)-LexA, α-Jak2-JH1, and α-Riz1-PR interacting lariats were expressed as lariats or linear peptides and assayed for interactions against LexA, Jak2-JH1, and Riz1-PR. **(a)** Lariat and linear constructs used in the Y2H assay. The species the I_C domain and the I_N domain are from is shown as well as the amino acids at the I_{C-1}/G7, I_{C+1}/G8 and I_{N+1}/A1 positions and the amino acids in the noose region. **(b)** Y2H assay of the lariat and linear peptides. Yeast were grown on SGR His^-, Trp^-, Leu^-, Ade^-, X-Gal media that selects for the activation of LEU2, ADE2, and LacZ Y2H reporter genes. pIN01 was an inactive intein expression plasmid that produced an inactive intein with a CPGC noose region.

4.1.4 Conclusions

We constructed a modified *Ssp* DnaE intein that was capable of producing lariat peptides. Production of the lariat was verified by Western blot analysis and LC/MS. We determined that the lariat was stable inside of cells in significant quantities (46%) by MS. We demonstrated the lariat bond was essential for the interaction of three different lariats with their targets in the Y2H assay.

4.2 Aim 2: Optimization of the lariat technology

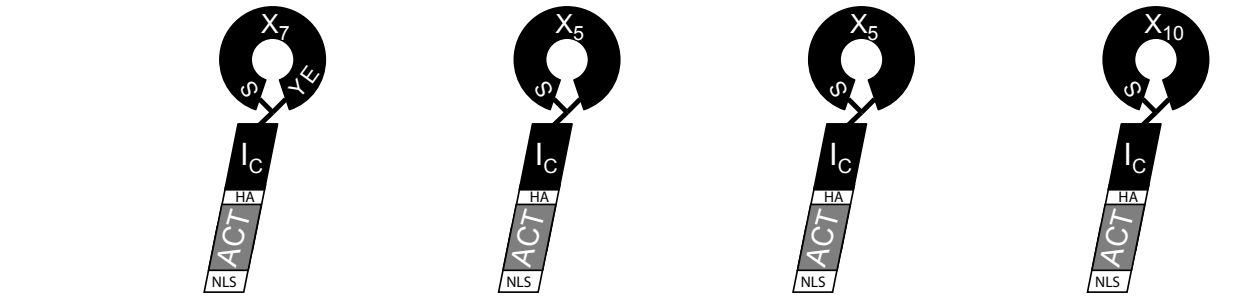
To improve the lariat technology developed in Section 4.1, we developed strategies to enhance the stability of the lariat and to increase the percentage of the lariat library that correctly processed. Previously, the *Npu* DnaE intein was shown to be more tolerant of amino acid substitutions at the I_{C+2} position in the extein (Iwai *et al.*, 2006). To improve the percentage of combinatorial peptides that correctly processed, a hybrid intein containing the I_C domain from *Ssp* DnaE intein and the I_N domain from *Npu* DnaE intein was constructed. To improve the stability of the lariat, mutations were introduced at the G6/G7/B11 positions in the *Ssp-Ssp* DnaE intein.

4.2.1 *Ssp-Npu* hybrid intein

A second-generation lariat library was created using the *Npu* DnaE I_N domain instead of the *Ssp* DnaE I_N domain. The ability of the *Ssp* DnaE to process is dependent on amino acids that flank the cleavage site (Evans *et al.*, 2000; Iwai *et al.*, 2006; Scott *et al.*, 2001; Amitai *et al.*, 2009; Hiraga *et al.*, 2009). Amino acids at both the I_C – N-extein junction and the C-extein – I_N junction influence processivity. This was the rational for fixing the last two residues at the C-extein – I_N junction as glutamic acid and tyrosine in the *Ssp* DnaE intein lariat library (pIL-R7). Iwai *et al.* found that amino acids at the I_C – N-extein junction were important for the *Ssp* DnaE intein processing. When the I_N domain from the *Npu* DnaE intein was combined with the I_C domain of the *Ssp* DnaE intein, the hybrid intein was tolerant to substitution of a greater variety of amino acids at the I_{C+2} position (Iwai *et al.*, 2006). To exploit the promiscuity of the *Npu* DnaE I_N domain for improving lariat peptide libraries, we constructed an *Ssp-Npu* intein gene with the I_C domain from *Ssp* DnaE (Scott *et al.*, 1999) and the I_N domain from *Npu* DnaE (Iwai *et al.*, 2006). The I_N domain of the *Npu* DnaE gene was constructed using the same strategy used previously to construct the *Ssp-Ssp* intein gene

(Section 4.1.1). Several modifications were made in the construction of the *Ssp-Npu* libraries. First, in our *Ssp-Ssp* R7 library, 9% of the library contained non-functional lariats because the plasmid (pIN01) contained alanine at position I_{N+1}/A1. In the *Ssp-Npu* lariat plasmid the I_{N+1}/A1 amino acid was cysteine, which is required at I_{N+1}/A1 to allow the intein to undergo an N-S acyl shift and form the lariat. Second, a kanamycin marker was added to the lariat library plasmid. Both the bait and prey plasmids contained the ampicillin marker and switching the library plasmid to a kanamycin marker facilitated separating the target bait and library plasmids in *E. coli*. Third, we changed the *Rsr*II restriction site between the I_C and I_N domains to *Nru*I. The change was required since the kanamycin gene contains an *Rsr*II restriction site. Cleavage of the *Nru*I restriction site resulted in a blunt end product, which we predicted would reduce the percentage of self-ligating plasmids in the library and decrease the no-insert products. The resulting plasmid expressing the *Ssp* I_C and *Npu* I_N domains separated by an *Nru*I restriction site was referred to as pIL500.

Using the *Ssp-Npu* lariat intein, we explored the effect of peptide size and amino acid diversity on isolating lariats. To test these variables, we created three additional lariat libraries (Figure 4.19). Two of the lariat libraries varied in their peptide size. These libraries were encoded by NNK nucleotide repeats with either five or ten amino acids in the noose region. The third library contained five amino acids encoded by BNT codons in the noose region.



Library	*Ssp-Ssp* R7	*Ssp-Npu* R5	*Ssp-Npu* F5	*Ssp-Npu* R10
Length	7	5	5	10
I_C	Ssp DnaE	Ssp DnaE	Ssp DnaE	Ssp DnaE
I_N	Ssp DnaE	Npu DnaE	Npu DnaE	Npu DnaE
X	Any AA + Stop	Any AA + Stop	{ACDFGHLPRSVY}	Any AA + Stop

Figure 4.19 | Summary of lariat libraries constructed
Four different libraries were constructed in this work with the I_C or I_N domain from either the *Ssp* DnaE or the *Npu* DnaE intein. The *Ssp-Ssp* R7 library contains a fixed EY at the Extein-I_N junction and a combinatorial noose encoding all twenty amino acids (NNK codon) seven amino acids in length. The *Ssp-Npu* R5 library contains no fixed amino acids and encodes all twenty amino acids (NNK codon) in a combinatorial noose five amino acids long. The *Ssp-Npu* F5 library contains no fixed amino acids and encodes the amino acids [ACDFGHLPRSVY] (BNT codon) in a combinatorial noose five amino acids in length. The *Ssp-Npu* R10 library, which has a combinatorial, noose ten amino acids long with all twenty amino acids represented (NNK codon).

4.2.1.1 *Ssp-Npu* R10 intein library

The *Ssp-Npu* R10 library contained ten random amino acids inserted between the *Ssp* DnaE I_C domain and the *Npu* DnaE I_N domain at the *Nru*I restriction site. The *Nru*I restriction endonuclease recognizes the sequence TCG^CGA and produces blunt end DNA products after cleavage. The oligonucleotides encoding the combinatorial peptide library were cloned into pIL500. The theoretical complexity of the library was 2×10^{13} (21^{10}). We obtained a library diversity of $6.5 \pm 0.6 \times 10^{6}$, which represents a very small fraction (3.9×10^{-5} %) of the theoretical amino acid sequence space.

We sequenced ninety-one naïve library members to characterize the *Ssp-Npu* R10 library (Figure 4.20). Four types of sequences were identified: no-insert plasmids, plasmids with out-of-frame sequences, plasmids with sequences containing stop codons, and plasmids

containing sequences in the correct reading frame. Only 1% of the lariat library contained plasmids with a no-insert sequence. This was a significant decrease from the 32% no-insert plasmids observed previously in the *Ssp*-*Ssp* R7 library. This improvement was most likely a result of gel purifying the R10 vector after *Nru*I digestion combined with the blunt ends produced by the *Nru*I restriction enzyme. Only 5% of the lariat library contained plasmids with out-of-frame peptides. Eighty-five sequences underwent in-frame homologous recombination. Twenty-seven percent (23/85) of these sequences contained at least one stop codon. This correlates with the number of sequences theoretically predicted to contain stop codons (27% = $(1 - 31^{10}/32^{10})$). The remaining 68% (62/91) of the library contained sequences expected to form lariats out of a possible 73% (100% - 27% of sequences with stop codons). This was an improvement over the *Ssp*-*Ssp* R7 library where 31% of sequences were expected to form lariats out of a potential 80%. In addition, by modifying the *Ssp*-*Npu* plasmid to contain I_{C+1}/G8-[SRC]-I_{N+1}/A1 at the insertion site instead of the *Ssp*-*Ssp* R7 insert I_{C+1}/G8-[SCGPCEYA]-I_{N+1}/A1, we eliminated the erroneous alanine substitution at position I_{N+1}/A1 found in the *Ssp*-*Ssp* R7 library.

After cleavage of pIL500 by *Nru*I, three nucleotides from the *Nru*I recognition site remain at both ends of the cleaved plasmid. Therefore, these nucleotides could have affected homologous recombination and favored recombination of library sequences that had sequences that matched the restriction site. This would bias the amino acids present in the library, especially at the I_{N-1} position. At the 5`-end of the recombination site (I_{C+1})-SR-(I_{N-1}), pIL500

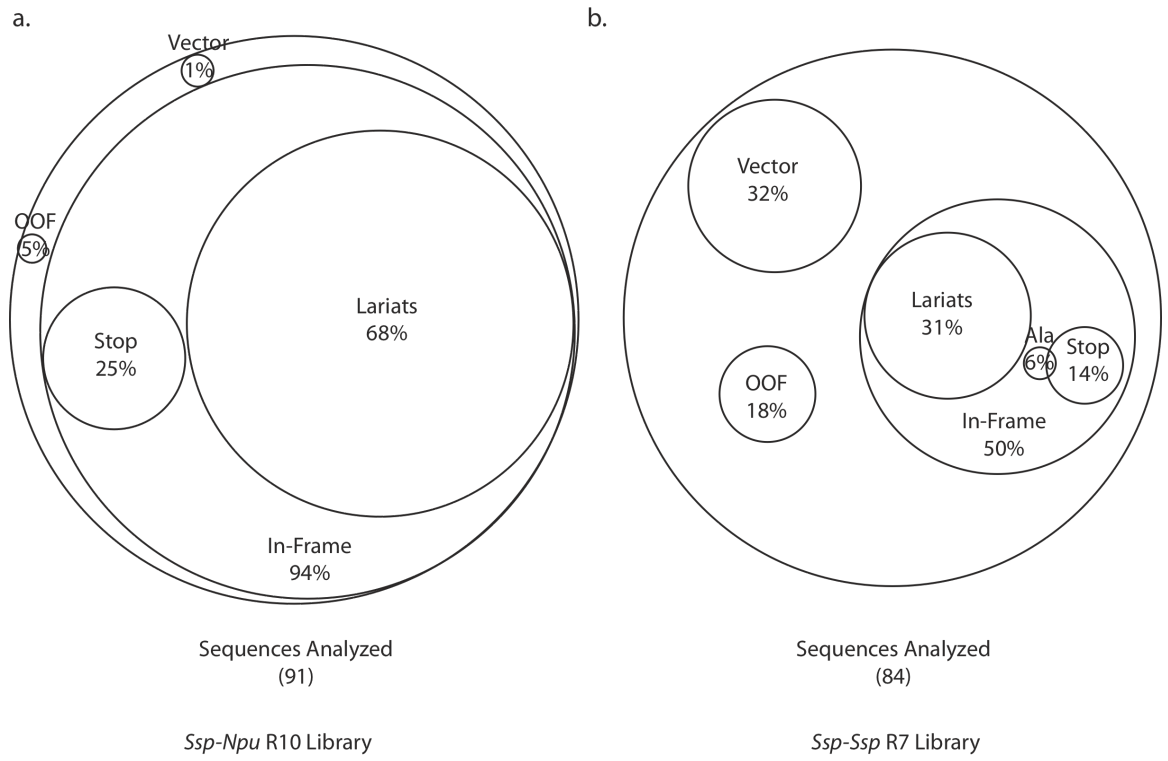


Figure 4.20 | *Ssp-Npu* R10 naïve library sequence analysis
Venn diagrams showing naïve lariat library sequence information. **(a)** The *Ssp-Npu* R10 library. **(b)** The *Ssp-Ssp* R7 library. Vector indicates no-insert plasmids. OOF represents out-of-frame sequences. Ala represents sequences containing an alanine at I_{N+1} (A1). Stop indicates sequences with a stop codon. In-frame represents sequences that are in frame. Lariats represent sequences predicted to form lariats.

had the nucleotide sequence [CAC GCT TCG], which corresponded to the amino acid sequence (I_{C-2}/G6)-HAS-(I_{C+1}/G8). The oligonucleotide encoding the peptide library used the serine TGT codon instead of the serine TCG codon, which was used in pIL500. Only one of eighty-six sequences analyzed contained the serine TCG codon. The low frequency of the TCG codon in the library indicated that homologous recombination favored the nucleotides present in the oligonucleotide encoding the random peptide rather than the nucleotides from the *Nru*I restriction site present in pIL500. At the 3`-end of the recombination site, pIL500 had the nucleotide sequence [CGA TGT], which corresponded to the amino acids ($I_{N-1}$)-RC-($I_{N+1}$/A1). At the 3`-end of the combinatorial peptide, we did not observe a significant bias for the amino acid arginine (I_{N-1}). We may have expected arginine (I_{N-1}) to appear more frequently since the 3`-end of pIL500 after digestion with *Nru*I contains an arginine codon. We analyzed the sequences of randomly chosen library members to determine whether the CGA arginine codon at the 3`-end of pIL500 significantly affected peptide library diversity. The library oligonucleotide was designed to replace the CGA codon with the NNG/T degenerate codon. Only 1/86 sequences contained an A at the third position of the codon instead of G/T. At the second position of the codon, G was observed in 30% of the sequences, which is within the tolerance of the expected value of 23 ± 7%. At the first position of the codon, C was present in 19% of the sequences, which is close to the predicted value of 20 ± 7%. Together, these results indicate that using the *Nru*I restriction enzyme did not substantially bias the sequence diversity of the peptide library.

4.2.1.2 *Ssp-Npu* R5 intein library

The *Ssp-Npu* R5 library contained five random amino acids and was constructed using the NNK codon. The *Ssp-Npu* R5 library has a theoretical diversity of 21^5 (4×10^6) peptides. A library with $8 \pm 2 \times 10^7$ members was created. Eight randomly chosen members of the R5 library were sequenced as a preliminary characterization of the R5 library. The library contained approximately 13% (1/8) members with stop codons. The predicted number of sequences containing one stop codon was 15% = ($1 - 31^5/32^5$). The remaining 87% (7/8) were expected to form lariats.

To compare the frequency that we obtained positive Y2H interactions between the R5 and R10 libraries we screened these libraries against Jak2. We found the R5 library typically

had fewer interactions per diploid screened compared to the R10 library. In screens against Jak2 domains, the R5 library had approximately one interaction for every million diploids screened whereas the R10 library had approximately one interaction for every 200,000 diploids screened. Therefore the R10 library was favored in subsequent screens.

4.2.1.3 *Ssp-Npu* F5 intein library

Genetic screens are limited to libraries containing around 10^9 members. A library expressing all 20 amino acids at each position that is seven amino acids long contains ~ 10^9 (20^7) library members, which is the upper limit of a genetic screen. Typically, genetic screens will only sample a small fraction of all the possible sequences. This random sampling can result in certain functional groups or amino acids never being tested and others being tested multiple times. One solution to improve the ability of a genetic screen to more comprehensively cover a library is to reduce the length of the peptides screened. Decreasing peptide length reduces the diversity of a library, however, shorter peptide libraries may not bind the target optimally.

A second solution is to reduce the number of amino acids substituted at each position. Some amino acids are more frequently involved in forming peptide-protein interactions than others. Tyrosine, for example, dominates antibody:antigen interactions (Koide and Sidhu, 2009). Fab libraries consisting of only two amino acids (tyrosine and serine) have been screened, resulting in high affinity and specific Fabs (Fellouse *et al.*, 2004). Many amino acids share similar functionalities and can be grouped by their chemical properties, such as polar, non-polar, charged, and hydrophobic. For example, aspartate or glutamate are both negatively charged amino acids that differ only by a methylene. We hypothesized a library containing only a single amino acid from each group would allow all the functional possibilities to be screened with a limited set of amino acids. Reducing the number of amino acids allows longer peptide libraries to be screened with greater library coverage. Since the optimal solution may not be presented in a minimal library, subsequent rounds of mutation can be used to improve binding.

The number of different sequences of a five-mer peptide containing all twenty amino acids is around four million. To reduce the diversity of the library, we used the degenerate BNT codon, which codes for 12 (3x4x1) different amino acids. The BNT codon encodes the amino acids: histidine, proline, arginine, leucine, aspartic acid, alanine, glycine, valine,

tyrosine, serine, cysteine, and phenylalanine, which represents hydrophobic, polar, charged, and aromatic amino acids. The library diversity of a five-mer peptide using the BNT codon is reduced substantially from four million to two-hundred and fifty thousand.

The F5 library was constructed using the same protocol as the other lariat libraries described previously. The theoretical diversity of the F5 lariat library was 2×10^5 peptides. The F5 library was constructed with a diversity of $3.6 \pm 0.5 \times 10^5$.

We screened the F5 library against several Jak2 domains and the PR domain of Riz1 using the Y2H assay to compare the ability of the F5 and R5 libraries to interact with protein targets. No positive Y2H interactions were isolated from the F5 library screens against Jak2 domains or the Riz1 PR domain. This preliminary evidence suggests that reducing amino acid diversity may not be an applicable strategy for the selection of cyclic peptides using the Y2H assay.

4.2.2 Lariat lactone bond stabilization

The lactone bond in the lariat was susceptible to hydrolysis, which produced a linear unconstrained peptide. The fraction of L2 lariat that contained an intact lactone bond following purification from *E. coli* was calculated to be 46%. Lariat formation was sufficient for the Y2H assay but improving lariat stability would increase the concentration of lariat within cells and allow the detection of weaker lariat interactions. It would also increase the probability that peptides isolated with the Y2H assay depend on the lariat for their interaction. Stabilization of the lactone-cyclized lariat is important primarily in downstream applications where it would make it easier to purify and store lariats. Enhanced lariat stability would also facilitate lariat purification and storage.

4.2.2.1 Rationale for constructing mutations

Rational and combinatorial approaches were used to generate mutant lariats with enhanced lactone bond stability. For both approaches, we chose the amino acid positions to mutate based on previous studies, which identified amino acids that stabilized the branched intermediate or on structural studies that revealed amino acid positions that are close to the catalytic site. Both structural and mechanistic studies on intein-mediated protein splicing have identified a variety of mutations that result in the accumulation of a branched intermediate, which is analogous to the lariat intermediate in cyclic peptide intein processing (Figure 4.21).

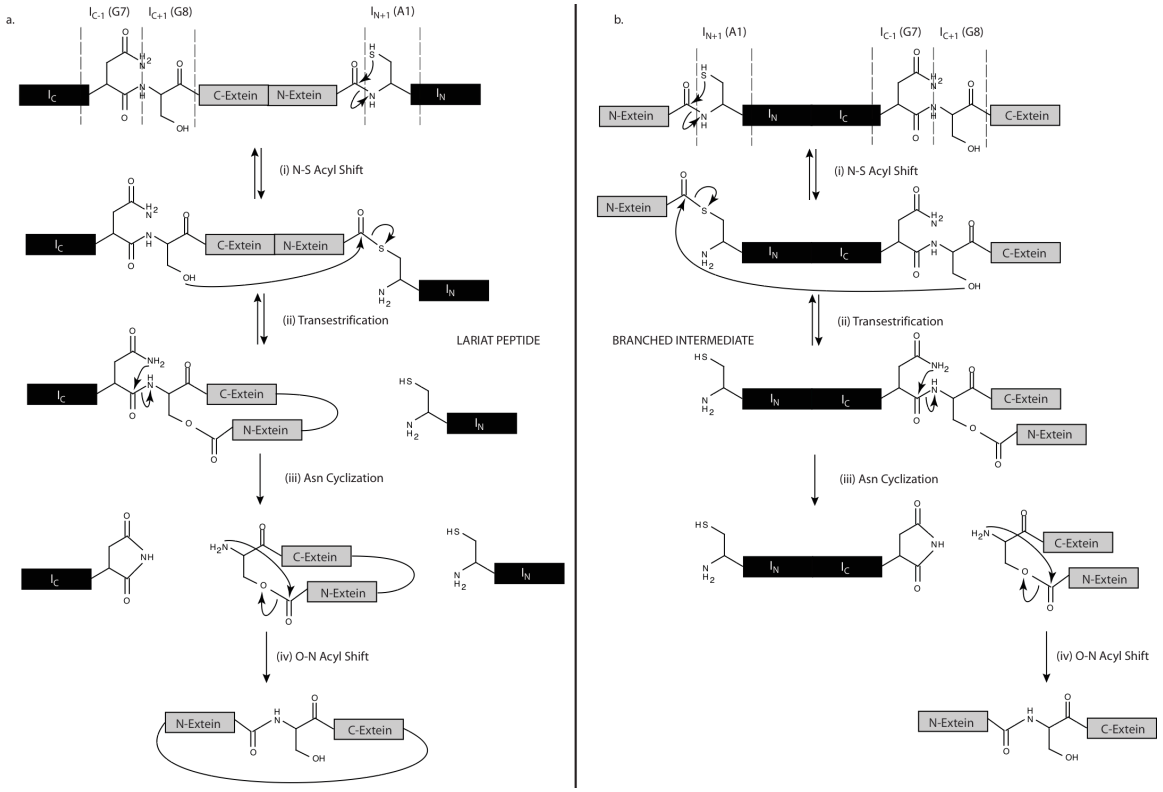


Figure 4.21 | Intein-mediated peptide splicing
(a) Intein-mediated cyclic peptide production. **(b)** Intein-mediated ligation. Note the lariat and branched intermediate structures are analogous structures for the two splicing mechanisms.

Mutational studies have been performed on the following three amino acids identified to be essential for intein processing: cysteine at position I_{N+1}/A1, which is essential for the N-S acyl shift; serine at position I_{C+1}/G8, which is essential for tranesterification; and asparagine at position I_{C-1}/G7, which is essential for asparagine cyclization. Mutational studies have also been performed on many amino acids within the conserved blocks (A, B, F, and G) along with several amino acids outside of these conserved blocks (Kawasaki *et al.*, 1997; Chen *et al.*, 2002; Ding *et al.*, 2003; Sun *et al.*, 2005). All possible mutations within the intein would result in millions of different combinations, which represent a sequence space larger than we were able to study. Thus, we focused our study on amino acids at positions I_{C-2}/G6, I_{C-1}/G7, and B11.

An accumulation of branched intermediate is observed when asparagine at position I_{C-1}/G7 amino acid is not mutated. This is surprising since asparagine is the key residue for the third step in intein processing: asparagine cyclization. If branched intermediate is accumulating then asparagine cyclization must be being blocked by mutations at an alternative site. I_{C-2}/G6 was selected as an alternative site to mutate since histidine at this position has previously been shown to be important for lactam peptide cyclization (Scott *et al.*, 1999; Scott *et al.*, 2001), suggesting I_{C-2}/G6 has a role in allowing the intein to process completely. Since the I_{C-2}/G6 is important for cyclization, mutating this amino acid could block asparagine cyclization. So histidine at position I_{C-2}/G6 was mutated to aspartic acid, asparagine, and leucine while maintaining asparagine at I_{C-1}/G7 to determine if the lariat could be stabilized by I_{C-2}/G6 mutations.

The I_{C-1}/G7 position was selected since it has been shown previously to contribute to the stability of the branched intermediate (Kawasaki *et al.*, 1997). Lysine was selected since previous studies had identified that lysine mutations at I_{C-1}/G7 resulted in the accumulation of a branched intermediate (Kawasaki *et al.*, 1997). Tyrosine was substituted at position I_{C-1}/G7 since it is large and aromatic and we predicted it would have some effect on splicing. Aspartic acid and glutamine are found at I_{C-1}/G7 in a very small percentage of naturally occurring inteins (Perler, 2002) and were also substituted at position I_{C-1}/G7 in place of asparagine. Aspartic acid and glutamine are potentially capable of undergoing cyclization just like asparagine but are likely to undergo cyclization at a slower rate, thus they are interesting substitutions since they will mimic asparagine but potentially halt the intein at the lariat.

Crystal structures of several inteins including the *Ssp* DnaE intein have been solved (Chen *et al.*, 2002; Ding *et al.*, 2003; Sun *et al.*, 2005). Using these structures, we identified an uncharacterized amino acid at position B11 whose role in intein processing has not been examined. B11 is located in the I_N domain and is normally arginine. The B11 amino acid was selected based on the proposed role in asparagine cyclization (Sun *et al.*, 2005) (Figure 4.22). By disrupting this charge relay system we can potentially block asparagine cyclization without mutating the asparagine at (I_{C-1}/G7), which has previously been shown to stabilize the branched intermediate (Kawasaki *et al.*, 1997). We mutated arginine at position B11 to glutamic acid, tyrosine, and leucine, which represent a range of chemical functionality. Table 4.2 lists the complete set of mutations analyzed.

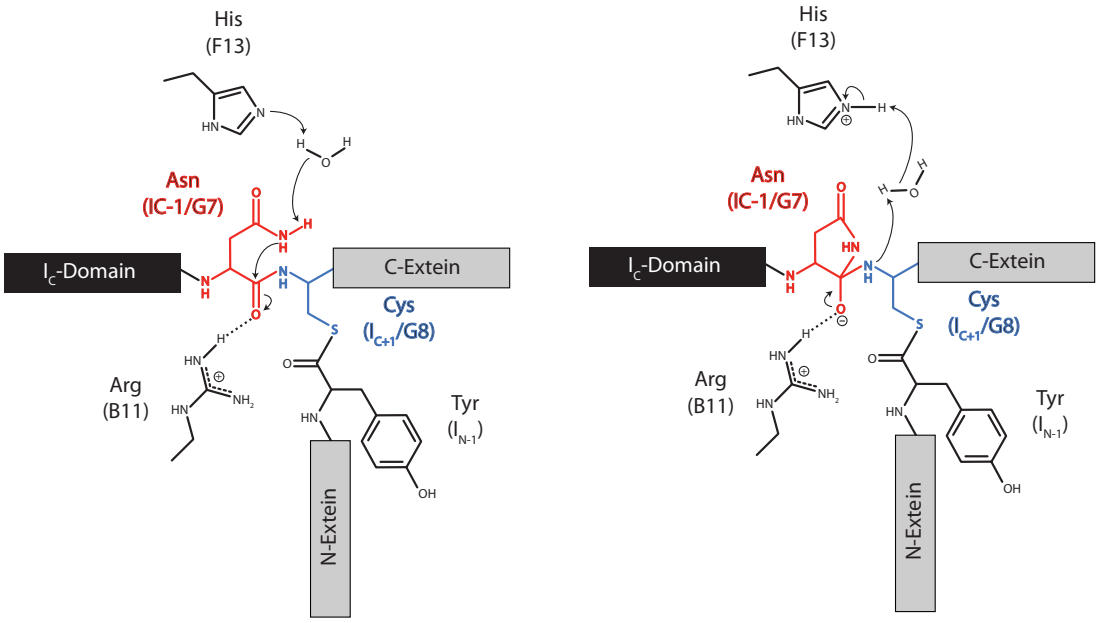


Figure 4.22 | Role of the B11 position in asparagine cyclization
The proposed role of the B11 arginine in the charge relay system of asparagine cyclization, which is resolved by a tetrahedral intermediate and oxyanion binding site, based on the crystal structure of the *Ssp* DnaE intein (Sun *et al.*, 2005). I_C represents the C-terminal domain of the intein. Arrows represent the proposed nucleophilic attacks.

Table 4.2 | Rational mutations tested to stabilize the lariat
The original mutation used to create the lariat in this thesis is represented by Mutant '0'. An additional 29 mutations were constructed to analyze their effect on lariat processing and stability. Amino acids are represented by the single letter amino acid code.

Mutant	G6	G7	B11	Mutant	G6	G7	B11	Mutant	G6	G7	B11
0	H	A	R	**10**	H	Q	L	**20**	A	Y	L
1	N	N	Y	**11**	H	Q	E	**21**	A	Y	E
2	N	N	R	**12**	H	A	Y	**22**	A	K	Y
3	N	N	L	**13**	H	A	L	**23**	A	K	R
4	N	N	F	**14**	H	A	E	**24**	A	K	L
5	L	N	Y	**15**	D	N	Y	**25**	A	K	E
6	L	N	R	**16**	D	N	R	**26**	A	D	Y
7	L	N	L	**17**	D	N	L	**27**	A	D	R
8	L	N	E	**18**	D	N	E	**28**	A	D	L
9	H	Q	R	**19**	A	Y	R	**29**	A	D	E

4.2.2.2 Analysis of lariat stability

Initially, we optimized lariat expression and examined the effects of induction time and the time between collection and analysis on lariat stability. Lariat stability was analyzed using liquid chromatography directly coupled to electrospray ionization time-of-flight mass spectrometry (LC/MS). The LC step was used to desalt and separate the intein products based on hydrophobicity. The lariats were quantified using total ion count (TIC). The peaks eluted from the LC represented different intein fragments and were first integrated to generate a mass/charge (m/z) versus peak intensity spectrum. The charged spectrum was processed using maximum entropy calculations (MaxEnt) to deconvolute the spectrum into its component molecular weight proteins and their abundance. The MaxEnt data was used to calculate '% lariat', '% processed', and '% total lariat'.

% lariat

The "% lariat" represents hydrolysis or the stability of the lactone bond and was the percentage of intact lariat in the sample compared to the hydrolyzed product. The % lariat was calculated by dividing the area under the lariat peak by the sum of the areas under the lariat and hydrolyzed lariat peaks. The hydrolyzed lariat differs from the lariat by the addition of H_2O, which increased its mass by 18 Da.

$$\%Lariat = \frac{Lariat}{Lariat + Hydrolyzed} x100\%$$

% processed

"% processed" represents intein processing and was the percentage of lariat that was produced in the sample. It does not differentiate between lariat and hydrolyzed lariat. Percent processed is defined as the ratio of the area under the peak representing the I_N domain to the total amount of intein represented by the sum of the area under the peak representing the full length unprocessed protein and the I_N domain.

$$\%\mathrm{Pr}ocessed = \frac{I_N Domain}{I_N Domain + Unprocessed} x100\%$$

% total lariat

The "% total lariat" represents the total amount of lariat produced in a cell and is a function of % lariat and % processed. It is defined by the % processed multiplied by the % lariat that is not hydrolyzed.

$$\%TotalLariat = \%Lariat \times \%\mathrm{Pr}ocessed$$

We first established the optimal conditions for expression and the effect of lariat storage on stability on the L2 lariat. We induced cultures expressing the lariat for 2, 4, 7, or 18 hours. The samples were stored at 4 °C and then analyzed again 20 hours later. The induction time

effected the lariat stability with ~ 10% loss of lariat between the samples induced for 2 hours and the samples induced for 18 hours (Figure 4.23). However, the amount of time a sample spent at 4 ^{o}C before processing had little effect. After 20 hours, ~ 1 - 3% of the lariat was hydrolyzed (Figure 4.24). Together, these results suggest that the lariat was being hydrolyzed within the cell and that the lariat was relatively stable after purification.

Once we established the optimal conditions for expression and storage, the other mutants were expressed and analyzed under the same conditions. The mutants were sorted by % lariat, % processed, or % total lariat (Figure 4.25- Figure 4.27).

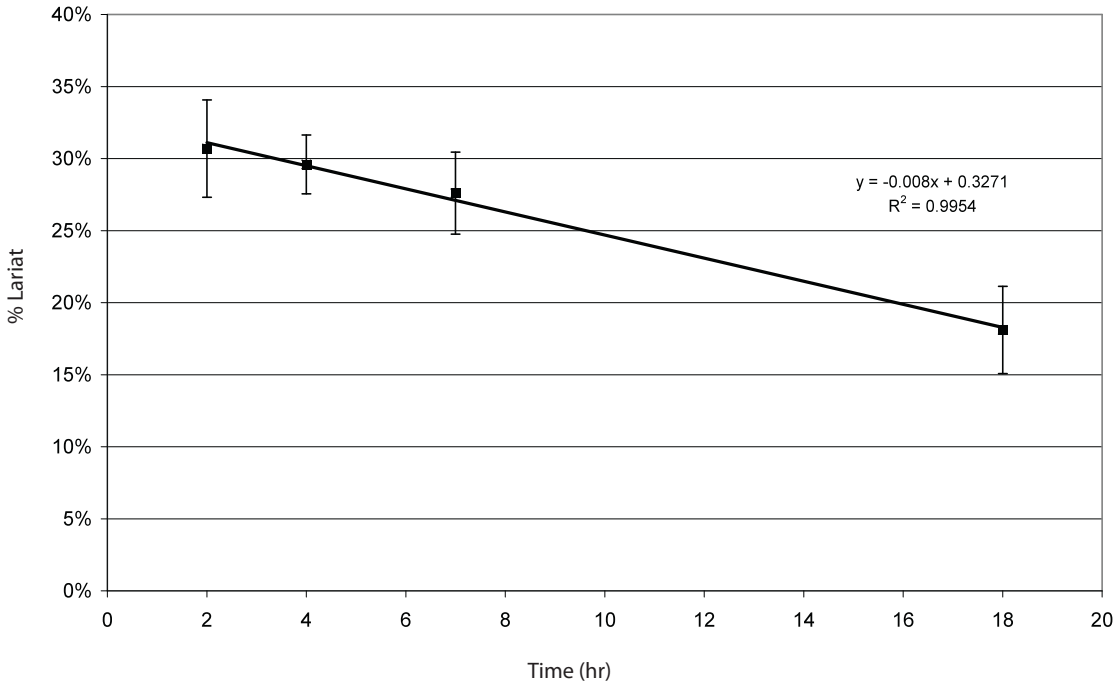


Figure 4.23 | Effect of induction time on lariat hydrolysis
Lariat expression was induced with IPTG for the time indicated and then purified using Ni^{2+}-NTA beads and analyzed immediately after purification by LC/MS. The % lariat was calculated by dividing the area under the lariat peak by the sum of the areas under the lariat and hydrolyzed lariat peaks and plotted versus induction time. Error bars represent the standard deviation of three independent experiments.

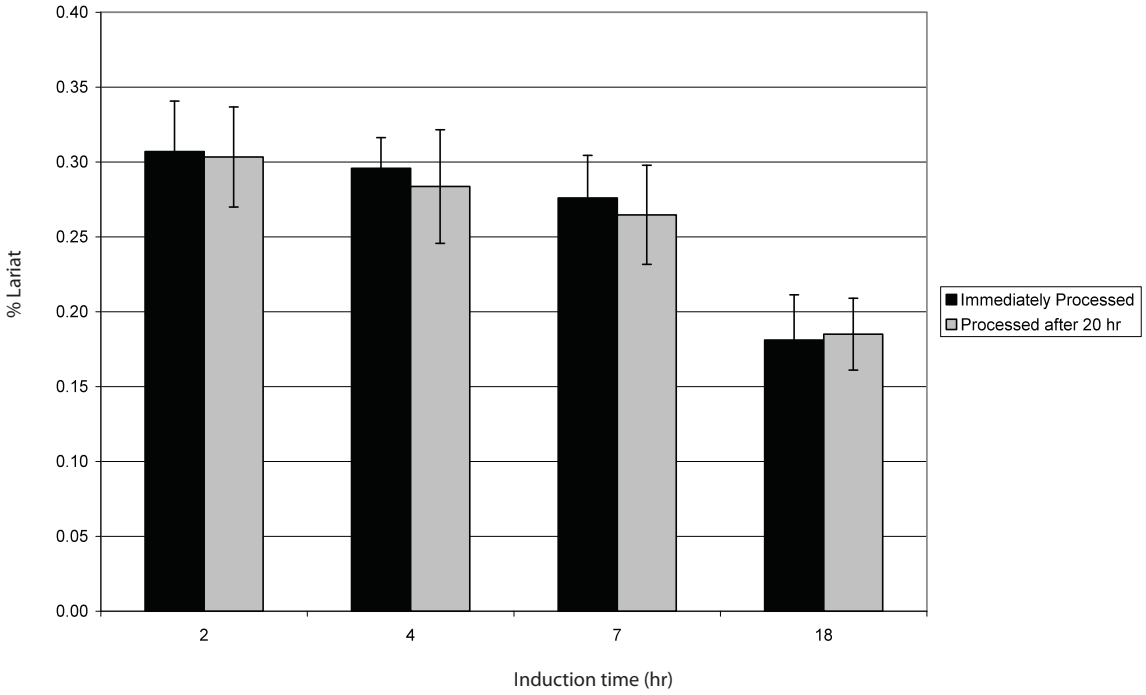


Figure 4.24 | Effect of storage time on lariat hydrolysis
Lariat expression was induced for 2, 4, 7, or 18 hours and purified using Ni^{2+}-NTA beads and analyzed immediately after purification by LC/MS (Black bars) and again after sitting for 20 hours at 4 °C (Grey bars). The % lariat was calculated by dividing the area under the lariat peak by the sum of the areas under the lariat and hydrolyzed lariat peaks. Error bars represent the standard deviation of three independent experiments.

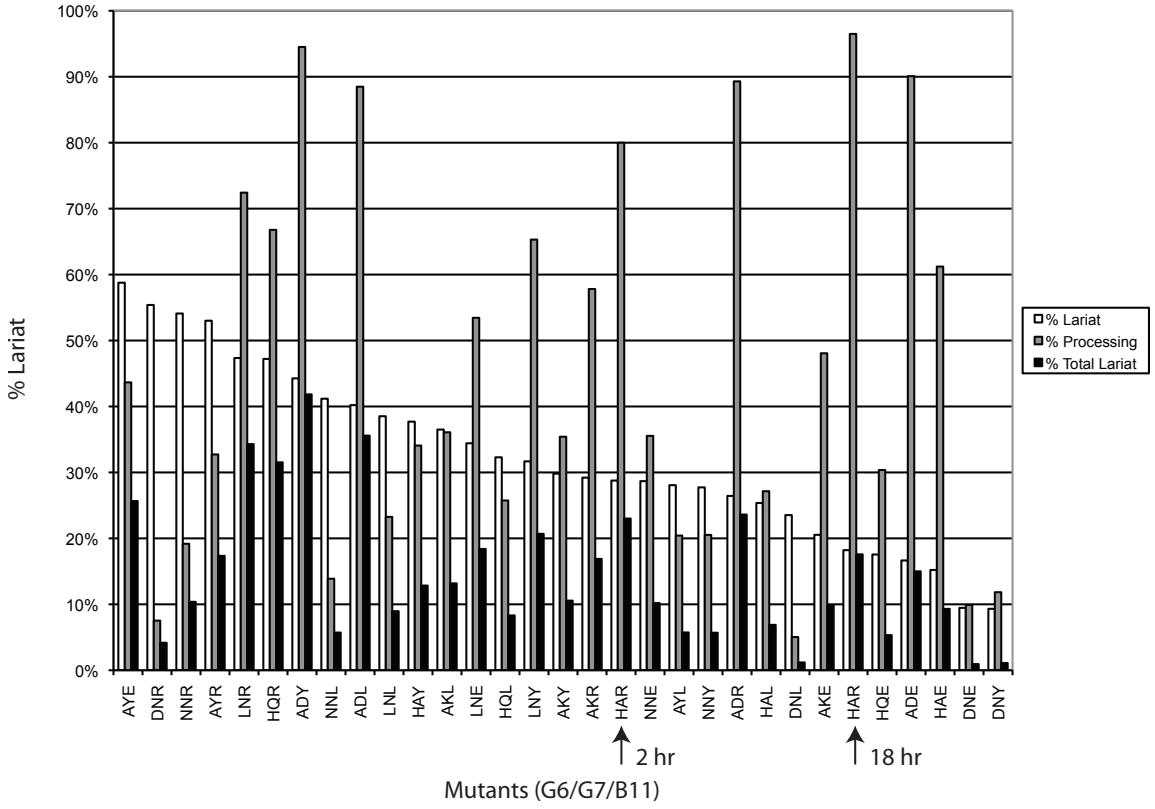


Figure 4.25 | Lariat mutations sorted by % lariat

Lariat expression was induced with IPTG for 2 hours and then purified using Ni^{2+}-NTA beads and analyzed by LC/MS. Mutant lariats are defined by the three amino acid mutations at positions (G6/G7/B11). The mutations are compared to the HAR lariat, which is shown twice; one point representing a 2 hr induction and a second point representing an 18 hr induction. The % lariat (White bars) was calculated by dividing the area under the lariat peak by the sum of the areas under the lariat and hydrolyzed lariat peaks. The % processed (Grey bars) was calculated by dividing the area under the peak representing the I_N domain by the sum of the area under the peak representing the full length unprocessed protein and the I_N domain. The % total lariat (Black bars) was calculated by multiplying the % processed by the % lariat. Mutations are sorted in order of highest % lariat to lowest % lariat.

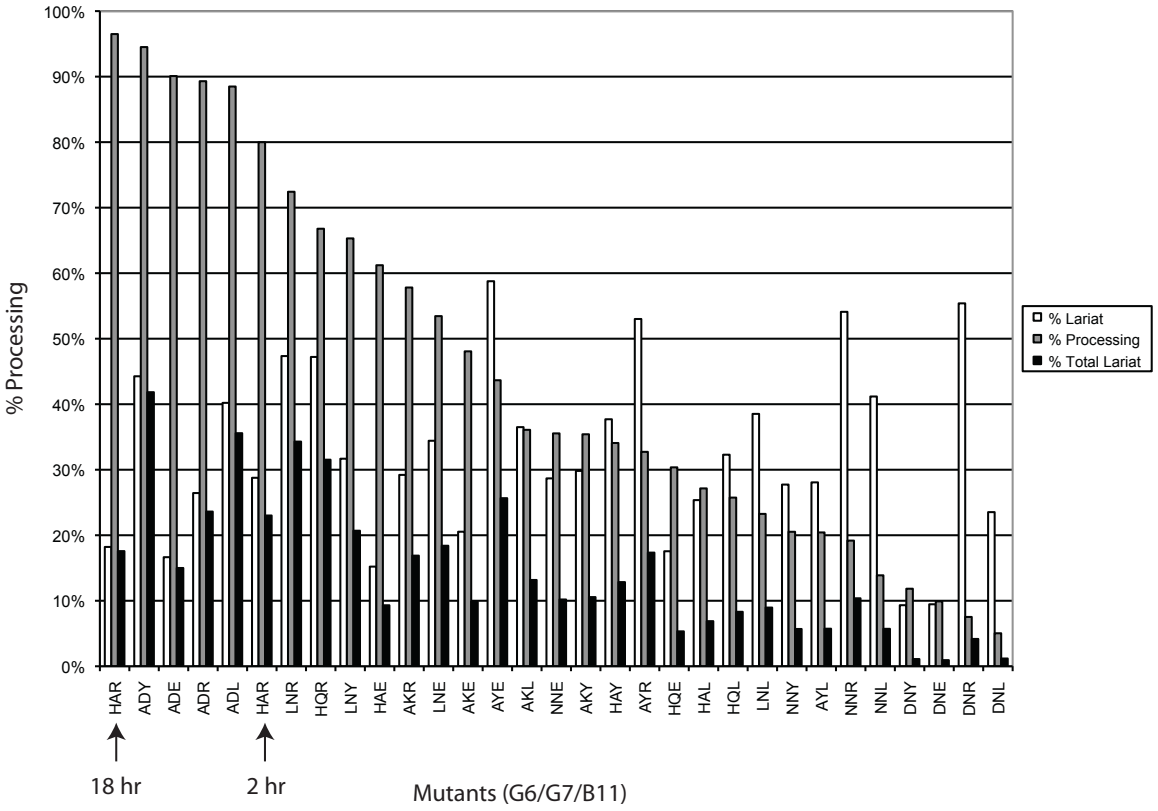


Figure 4.26 | Lariat mutations sorted by % processing
Lariat expression was induced with IPTG for 2 hours and then purified using Ni^{2+}-NTA beads and analyzed by LC/MS. Mutant lariats are defined by the three amino acid mutations at positions (G6/G7/B11). The mutant lariats are compared to the HAR lariat, which is shown twice; one point representing a 2 hr induction and a second point representing an 18 hr induction. The % lariat (White bars) was calculated by dividing the area under the lariat peak by the sum of the areas under the lariat and hydrolyzed lariat peaks. The % processed (Grey bars) was calculated by dividing the area under the peak representing the I_N domain by the sum of the area under the peak representing the full length unprocessed protein and the I_N domain. The % total lariat (Black bars) was calculated by multiplying the % processed by the % lariat. Mutations are sorted in order of highest % processing to lowest % processing.

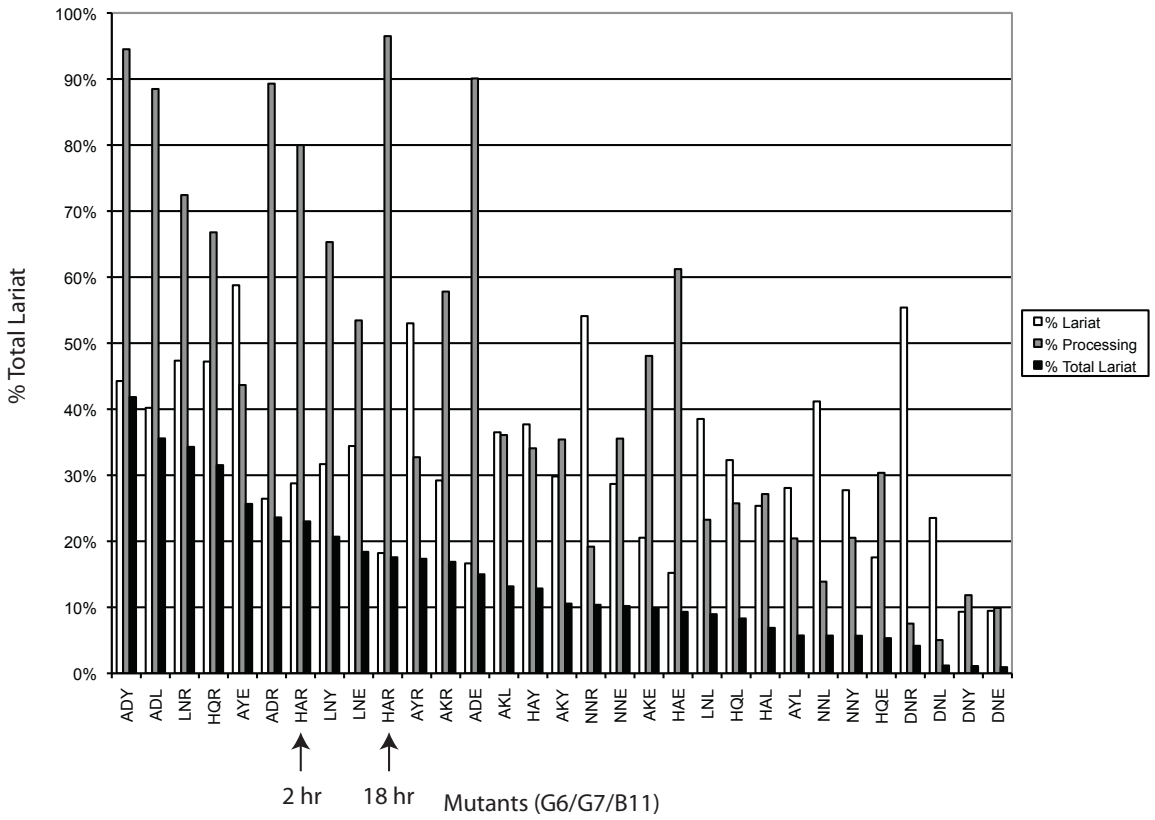


Figure 4.27 | Lariat mutations sorted by % total lariat
Lariat expression was induced with IPTG for 2 hours and then purified using Ni^{2+}-NTA beads and analyzed by LC/MS. Mutant lariats are defined by the three amino acid mutations at positions (G6/G7/B11). The mutant lariats are compared to the HAR lariat, which is shown twice; one point representing a 2 hr induction and a second point representing an 18 hr induction. The % lariat (White bars) was calculated by dividing the area under the lariat peak by the sum of the areas under the lariat and hydrolyzed lariat peaks. The % processed (Grey bars) was calculated by dividing the area under the peak representing the I_N domain by the sum of the area under the peak representing the full length unprocessed protein and the I_N domain. The % total lariat (Black bars) was calculated by multiplying the % processed by the % lariat. Mutations are sorted in order of highest % total lariat to lowest % total lariat.

Highly stable mutants were isolated that showed > 60% lariat by direct MS analysis. The actual amount of lariat *in vivo* was likely higher, since MS processing has been shown to convert lariat into hydrolyzed lariat. Previously, we found the wild-type lariat was present at ~ 25% of the total lariat by direct MS analysis whereas the amount of lariat present prior to MS analysis was 46%.

The data was analyzed by plotting the % processing against the % lariat on a double log scale to allow all the mutations to be visualized (Figure 4.28). Since the percent total lariat was

a function of both % lariat and % processing, diagonal lines on this plot indicate the threshold for 20, 30, 40, and 50% total lariat. Thus, mutants sharing a diagonal line will have the same percentage of total lariat. Mutants that have better stability and higher processing were found in the upper right of the plot and those with poor processing and stability were found in the lower left of the plot. This plot was used to create vector maps, where mutants were connected by lines from the lowest stability (% lariat) to highest stability (% lariat). These vector maps can be compared directly with other mutations to identify maps where mutations result in similar patterns. The direction of the vector, not the magnitude of change, is important in these maps since they were plotted on a double log scale. For example, in Figure 4.29**,** the mutant series (DNX) where the B11 position is varied (X) and the I_{C-2}/G6 (D) and I_{C-1}/G7 (N) positions are fixed was plotted. The resulting vector map can be visually compared to other vector maps to identify mutants that behave similarly.

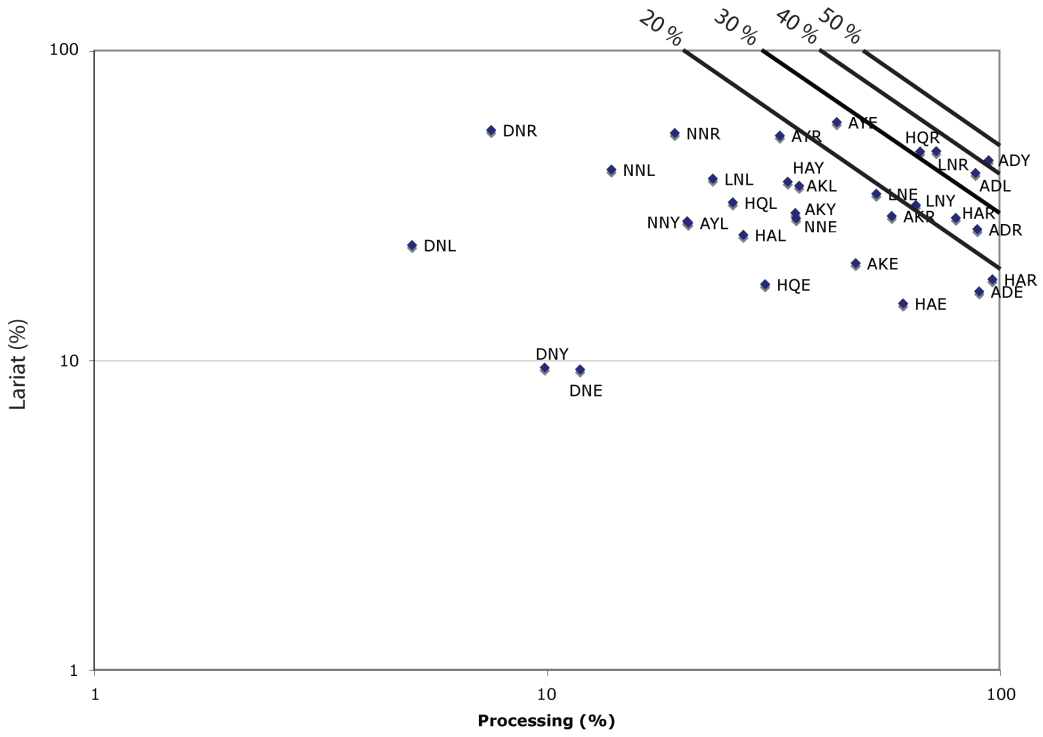


Figure 4.28 | Mutants summary map
Comparison of the percent processing versus percent lariat for mutants, plotted on log scales. The total lariat for 20%, 30%, 40% and 50%, which is calculated by multiplying the percent lariat by the percent processing is shown to assist in comparing mutants. Mutations are named by their position (G6/G7/B11).

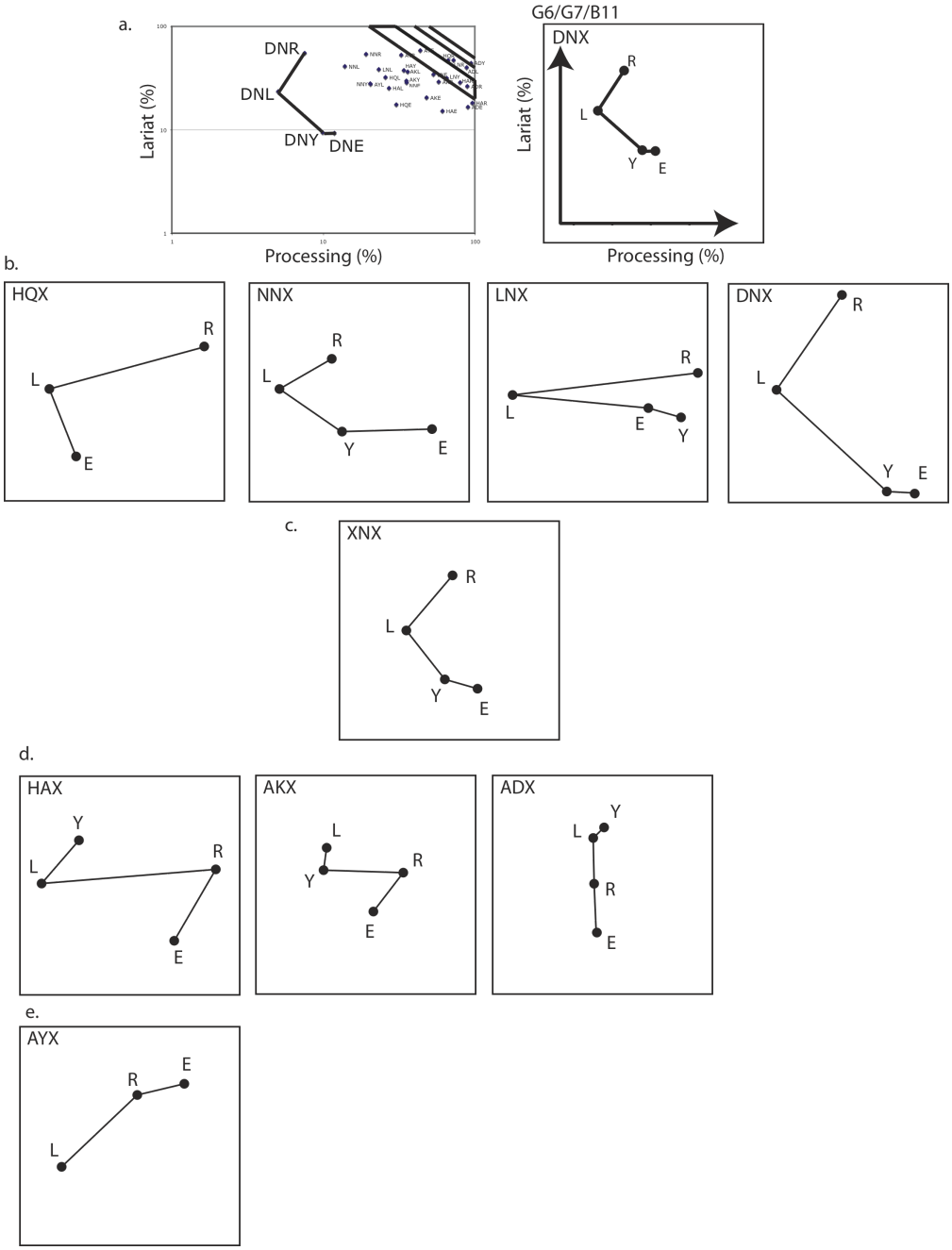


Figure 4.29 | Vector maps showing the G7-B11 interaction
(a) Vector map showing % lariat versus % processing for the G6/G7/B11 mutants, where G6 is aspartic acid (D), G7 is asparagine (N), and B11 is arginine (R), leucine (L), tyrosine (Y), or glutamic acid (D). Mutants are connected with lines from the lowest to the highest percent lariat to produce a vector map. Each mutant is abbreviated as three letters according to its G6/G7/B11 position. "X" represents the amino acid that is varied and is described by the letter next to the dot in the map. **(b)** Vector maps of XNX and XQX mutants. **(c)** The conserved pattern of the vector maps shown in (b). **(d)** Vector maps of HAX, AKX, and ADX mutants. **(e)** Vector map of the AYX mutant.

4.2.2.3 Mutation of the I_{C-2}/G6 amino acid

XNR, XNL, XNY, and XNE mutations were used to determine the effect of mutations at position I_{C-2}/G6 on lariat stability and processing, where X is leucine, asparagine, or aspartic acid (Figure 4.26). In these mutants, the I_{C-2}/G6 amino acid had a drastic effect on processing. Leucine had the best processing, followed by asparagine. Aspartic acid almost completely blocked processing in all cases. As suggested by previous studies, the I_{C-2}/G6 amino acid had the greatest effect on processing rather than on stability, as noted by the grouping of the I_{C-2}/G6 amino acids when sorted by % processing (Figure 4.26). Construction of more mutants where the I_{C-2}/G6 amino acid is paired with alternative I_{C-1}/G7 amino acids and B11 amino acids will be required to determine how the I_{C-1}/G6 amino acid influences processing and stability when amino acids other than asparagine are substituted at the I_{C-1}/G7 position.

4.2.2.4 Effect of asparagine to lysine mutation at position I_{C-1}/G7 on lariat processing

Asparagine at position I_{C-1}/G7 is essential for asparagine cyclization in the intein-mediated cyclization reaction. The asparagine side chain undergoes cyclization to cleave the I_C domain from the lariat and produce a lactone peptide. It is essential to block asparagine cyclization either by directly mutating the I_{C-1}/G7 amino acid or indirectly by mutating another amino acid to produce the lariat. In the initial lariat library (pIL-R7), we mutated the I_{C-1}/G7 asparagine to alanine to block asparagine cyclization. In the standard intein reaction, the branched intermediate accumulates when asparagine at position I_{C-1}/G7 is mutated to lysine (Kawasaki *et al.*, 1997). Not all mutations at position I_{C-1}/G7 result in an accumulation of branched intermediates, for example serine or alanine do not (Chong *et al.*, 1996). Interestingly, mutation of asparagine at position I_{C-1}/G7 to alanine leads to the accumulation of branched intermediate if cysteine at position I_{C+1}/G8 is also mutated to serine (Chong *et al.*, 1996), which is the case for the pIL-R7 lariat intein library. Based on these results, we tested the effect of mutating asparagine to lysine at position I_{C-1}/G7. Lysine was tested in combination with tyrosine, arginine, leucine, and aspartic acid at the B11 position, but only with alanine at the I_{C-2}/G6 position. Surprisingly, the lysine mutation at position I_{C-1}/G7 in all mutation combinations tested was similar to our initial construct and did not result in a significant improvement in stability or processing. This suggested that mutations that stabilized the branched intermediate may not apply to the lariat intermediate or that the lysine mutation at I_{C-1}/G7 had no effect when I_C-2/G6 was alanine.

4.2.2.5 The G7 position interacts with the B11 position

In the absence of histidine at I_{C-2}/G6, it has been suggested that arginine at position B11 can assist in asparagine cyclization by hydrogen bonding to the asparagine carbonyl oxygen at position I_{C-1}/G7 (Ding *et al.*, 2003). B11 is predominately lysine or arginine when I_{C-2} (G6) is not histidine (Perler, 2002). Currently, there are no mutagenic studies on the role of arginine at position B11.

When the vector maps of mutations with I_{C-1}/G7 as either asparagine or glutamine were compared, they showed a similar trend in how they responded to B11 mutations (glutamic acid, tyrosine, leucine, and arginine) (Figure 4.29) (note the HQY and AYY mutations are absent since they were not successfully cloned). Regardless of the I_{C-2}/G6 amino acid (histidine, asparagine, leucine, or aspartic acid), when B11 was glutamic acid or tyrosine, the lariat showed the lowest stability. When B11 was mutated to leucine, the lariat stability increased but the lariat processing decreased compared to when B11 was glutamic acid or tyrosine (L is always to the left and up of Y or E on the vector maps in Figure 4.29). When B11 was arginine lariat stability was increased compared to when B11 was leucine (R is always up and to the right of L on the vector maps in Figure 4.29). This data indicates that there was an interaction between the I_{C-1}/G7 and B11 amino acid, but less of an interaction between the I_{C-2}/G6 and B11 amino acids since the general shape of the vector maps did not change despite radical changes in the I_{C-2}/G6 amino acid.

The vector maps also indicate that the I_{C-1}/G7-B11 interaction could be further optimized by mutating the B11 amino acid from arginine to another positively charged amino acid such as lysine or histidine. Hydrophobic amino acids (leucine) reduced processing, and negatively charged (glutamic acid) and bulky (tyrosine) amino acids generally had good processing but poor stability. Mutating the arginine to another positively charged amino acid may improve both processing and stability since there was only a decrease in processing when B11 was mutated to leucine, suggesting that hydrophobic residues at the B11 position have a negative effect on processing when I_{C-1}/G7 was asparagine or glutamine.

This trend was not observed when I_{C-1}/G7 was alanine, lysine, or aspartic acid. When the I_{C-1}/G7 amino acid was alanine (HAX), lysine (AKX) or aspartic acid (ADX) the vector maps had a similar pattern (Figure 4.29d). In these cases, leucine or tyrosine at B11 had higher stability compared to arginine or glutamic acid. This pattern was opposite of the pattern

observed for the I_{C-1}/G6 (HQX, NNX, LNX, DNX) mutations. In the HAX, AKX, and ADX vector maps, glutamic acid had the lowest lariat stability and was found at the bottom of the vector map. Arginine (HAR, AKR, ADR) was always located above the glutamic acid mutation, indicating that arginine improved the lariat stability. In the cases when I_{C-1}/G7 was alanine (HAX) or lysine (AKX) processing was improved. Leucine (L) and tyrosine (Y) were found together on the vector map and generally resulted in a further improvement in lariat stability. The exception was for the (ADX) series, where stability was improved at the expense of processing (Figure 4.29). The ADY mutation had particularly good processing and stability, and further mutations to improve this construct should be explored. If the amino acids tested were representative of other amino acids in their class, then these mutations demonstrated that in general a negative charge (aspartic acid) resulted in good processing but reduced stability, a hydrophobic amino acid (leucine) improved lariat stability at the cost of processing, and an aromatic (tyrosine) amino acid improved stability and effected processing the least, representing the best improvement. Based on this analysis, other aromatic and hydrophobic amino acids should be substituted for the B11 amino acid (X) in the ADX construct.

When the I_{C-1}/G7 amino acid was tyrosine, a third distinct pattern was observed (Figure 4.29e). Again, highlighting the sensitivity of the I_{C-1}/G7 amino acid to changes in the B11 mutation and emphasizing that the optimal amino acids at the I_{C-1}/G7 and B11 positions depend on the amino acids at other positions.

4.2.2.6 Mutation of asparagine to a similar amino acid at position I_{C-1}/G7,

Asparagine, glutamine, and aspartic acid have all been shown to allow processing when at the I_{C-1}/G7 position. Asparagine at this position has been associated with a build up of the branched intermediate, but mutations that block asparagine cyclization must be introduced at other positions to create the lariat. Histidine at position I_{C-2}/G6 assists in asparagine cyclization by hydrogen bonding to the asparagine carbonyl oxygen at position I_{C-1}/G7. Branched intermediate accumulates when histidine at position I_{C-2}/G6 is mutated to leucine, asparagine, or glutamine. This also depends on the amino acid at position I_{C-1}/G7 since when asparagine at this position is mutated to alanine no branched intermediate is observed (Xu and Perler, 1996). This observation suggests that asparagine at position I_{C-1}/G7 is important for branched intermediate accumulation caused by I_{C-2}/G6 mutations.

To demonstrate that mutations at position I_{C-1}/G7, besides the previously tested asparagine to alanine mutation, can enhance the stability of the lactone bond, we mutated asparagine to glutamine at position I_{C-1}/G7. Mutation of the I_{C-1}/G7 amino acid to glutamine resulted in the stabilization of the lactone bond from the 29% lactone observed with alanine at I_{C-1}/G7 to 47% lactone observed with glutamine at I_{C-1}/G7. Glutamine at position I_{C-1}/G7 still maintained good lariat processing (67%). This result was interesting since it was theoretically possible that this mutation would still undergo processing to the lactam peptide based on the results with other inteins, where substitution of asparagine at I_{C-1}/G7 with glutamine still allowed processing.

4.2.2.7 Combinations of G6/G7/B11 mutations

By looking at the vector maps, where the I_{C-1}/G7 position was varied, we can make the following observations (Figure 4.30). The best processing (ADL, ADY) and most stable mutants (AYR, AYE) all have I_{C-2}/G6 as alanine. If B11 was charged (arginine, glutamic acid), then the optimal I_{C-1}/G7 amino acid for lariat stability was tyrosine followed by lysine followed by aspartic acid (Figure 4.30). Under the opposite conditions, when B11 was polar/hydrophobic (tyrosine, leucine), the optimal I_{C-1}/G7 amino acid was charged aspartic acid followed by lysine followed by tyrosine (Figure 4.30). The charged I_{C-1}/G7 and hydrophobic B11 combination (ADL or ADY) was better at processing than the hydrophobic I_{C-1}/G7 and charged B11 combination (AYR, AYE), which was better at processing. Interesting mutations to try based on this observation would be to fix I_{C-2}/G6 as alanine, mutate I_{C-1}/G7 to glutamic acid, and mutate B11 to tyrosine, phenylalanine, or tryptophan. In the alternative case, where I_{C-1}/G7 was tyrosine mutation to tryptophan, phenylalanine, serine, or leucine in combination with mutation of B11 to arginine, lysine, glutamic acid, or aspartic acid, may result in lariats with increased processing and stability.

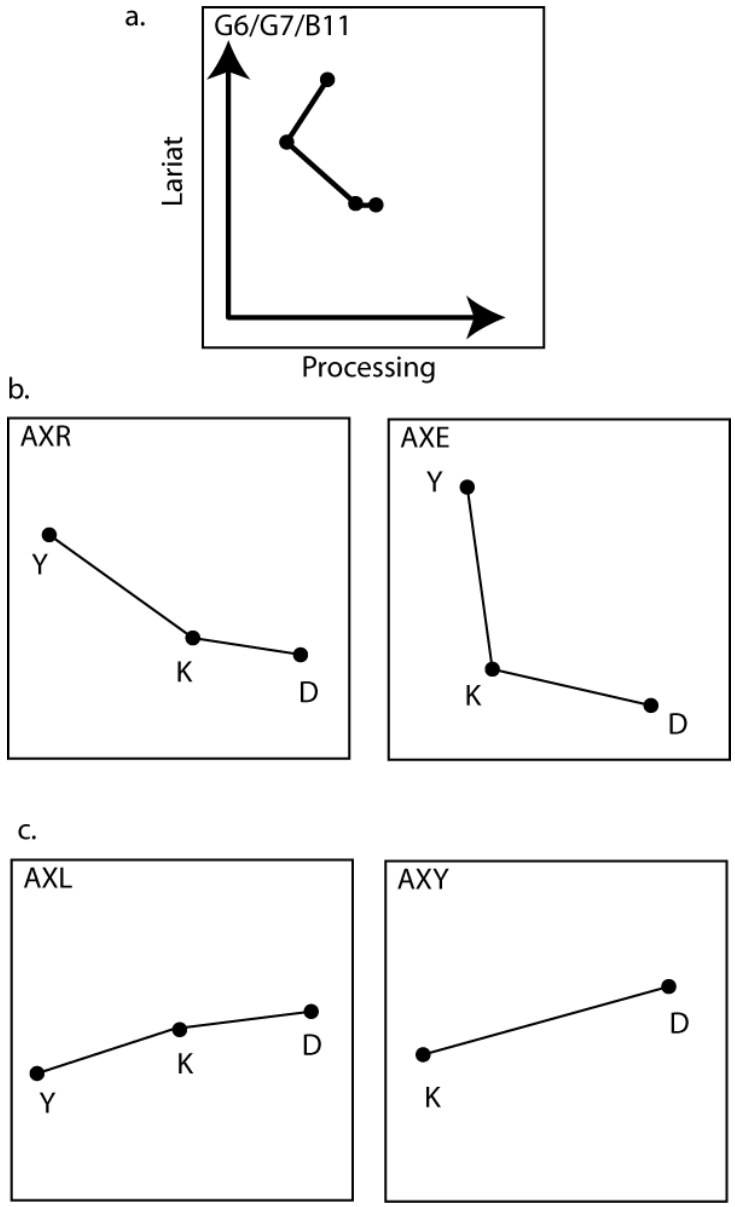


Figure 4.30 | Vector maps for the G7 amino acid
(a) Vector maps are created from Figure 4.28 by connecting mutants with lines from lowest to highest percent lariat, vector maps with similar patterns are shown together. **(b)** Vector maps of AXR and AXE mutants. Each mutant is abbreviated as three letters according to its G6/G7/B11 position. **(c)** Vector maps of AXY and AXY mutants. "X" represents variable amino acids shown in the panel.

4.2.2.8 Conclusions

From this short list of 29 mutations, we identified a number of mutants that were substantially more stable than then our initial lariat construct (HAR). In addition, this study identified a number of potential mutations to optimize lariat stability and processing.

The mutant with the highest processing and stability had alanine at I_{C-2}/G6, aspartic acid at I_{C-1}/G7 and tyrosine at position B11 amino acid was tyrosine. We found negatively charged amino acids at the B11 position had poor stability, but hydrophobic and aromatic amino acids at B11 increased stability. Other aromatic amino acids should be tried at the B11 position including tryptophan and phenylalanine. Since hydrophobic amino acids were also found to increase stability, other hydrophobic amino acids such as isoleucine should also be tried to find the optimal amino acid. In addition, a more positively charged amino acid such as lysine or histidine should be tested to fully determine the effect of a positively charged amino acid on stability and processing.

When the I_{C-2}/G6 amino acid was alanine and the B11 amino acid was charged (arginine, glutamic acid), the optimal amino acids at the I_{C-1}/G7 position were tyrosine followed by lysine. Other aromatic amino acids at the I_{C-1}/G7 position should be tried in this combination.

The I_{C-2}/G6 position was not optimized and amino acids having hydrophobic side chains may also be used to stabilize the lactone bond. The following amino acids may also be substituted at position I_{C-2}/G6: tryptophan, phenylalanine, leucine, isoleucine, methionine, and tyrosine. These amino acids have hydrophobic side chains, which could exclude water from the reactive site and reduce hydrolysis while still permitting processing.

4.2.3 Combinatorial mutagenesis of the G6/G7/B11 positions in the anti-LexA lariat

The anti-LexA L2 lariat (described in Section 4.3) was found to interact with LexA only when the lactone bond was present. This specificity was exploited to develop a combinatorial Y2H screen to select lariat mutants that retained the lariat conformation, since a loss of the lariat formation would result in no interaction in the Y2H assay. A combinatorial library was designed, where all 20 amino acids were potentially substituted at the G6/G7/B11 positions. A library of approximately 400 different constructs was created with mutations at the G6/G7/B11 positions. A library with all the combinations of twenty amino acids at all three positions would contain 8000 (20^3) different proteins. We only sampled 5% (400/8000) of this protein

sequence space. This library of 400 members was tested for mutants that could interact with LexA using the Y2H assay, where the 400 library members were arrayed in triplicate. Eleven unique colonies were selected after two days of growth on full selective media. Three colonies were detected in all three replicates, five colonies were detected on two of the replicates, and three colonies were detected in only one replicate. Six of the eleven mutant lariats were successfully sequenced (Figure 4.31). The rational screen demonstrated that the optimal amino acids for stability could not be determined independently and were highly dependent on the identity of the other amino acids at positions (G6/G7/B11). This combinatorial screen was important since it can sample all of the possible combinations of the G6/G7/B11 amino acids. Mutations were found at all three positions (G6/G7/B11), indicating that the L2 lariat was tolerant to mutations at these positions and that they were likely not involved in the interaction with the target. The processing of these mutants has not yet been characterized.

Lariats with alanine, asparagine, serine, and glycine at the I_{C-2}/G6 position were isolated. Asparagine and glycine at position I_{C-2}/G6 occurred in combination with serine at position I_{C-1}/G7. Our rational mutagenesis analysis indicated the I_{C-2}/G6 amino acid primarily played a role in processing and negatively charged amino acids (aspartic acid) were associated with poor processing. This was supported by the lack of negatively charged amino acids selected at this position in the random screen.

Mutations at the I_{C-1}/G7 position are essential for blocking the intein processing to create the lariat. Asparagine at the I_{C-1}/G7 position allows processing to the lactam peptide in the absence of other mutations. We found serine, alanine, and phenylalanine at the I_{C-1}/G7 position. We did not isolate asparagine at this position, even though in the rational screen we demonstrated that lariat formation was possible with the I_{C-1}/G7 as the wildtype asparagine as long as the I_{C-2}/G6 amino acid was also mutated. Alanine was isolated at I_{C-2}/G6, which was the original mutation that we used to block intein processing. Serine and phenylalanine were also isolated at the I_{C-1}/G7 position. The rational screen did not use serine at the I_{C-1}/G7 position, so it is not possible to compare the two experiments. However, phenylalanine was interesting since in the rational mutagenesis, we found that tyrosine at the I_{C-1}/G7 position was associated with higher stability when B11 was a charged amino acid (arginine, glutamic acid) and hypothesized that phenylalanine at position I_{C-1}/G7 would improve lariat stability and

processing. The random screen isolated this proposed mutation; phenylalanine at the I_{C-1}/G7 position and arginine at the B11 position, supporting this hypothesis.

Times Isolated	G6	G7	G8	R7EY	A1	B11
wildtype	H	A	S	RSWDLPGEY	C	R
2	A	A	S	RSWDLPGEY	C	R
2	N	S	S	RSWDLPGEY	C	G
1	G	S	S	RSWDLPGEY	C	R
1	S	F	S	RSWDLPGEY	C	R

Figure 4.31 | Random mutagenesis Y2H screen
Mutant anti-LexA L2 lariats isolated in the random mutagenesis screen.

4.2.3.1 Conclusions

Sequencing of the naïve library should be performed to ensure that B11 and the other positions were mutagenized and had the proper theoretical distribution of amino acids at each position. Processing of these mutants in yeast and *E. coli* should be tested along with their interaction strength with LexA. Based on these limited results, a large scale screen of all 8000 mutants of the anti-LexA L2 lariat and one or more other lariats isolated against different targets that also depend on lariat formation should be performed. Each screen would isolate a number of viable mutations at these sites. By overlaying these datasets, the optimal G6/G7/B11 amino acids could be chosen. Mutations that occur in all constructs are unlikely to be involved in the lariat-target interaction and would be ideal candidates to be used in the design of neutral scaffold.

4.2.4 Mutations in the extein

Mutations in the extein were also found to affect the lariats stability (Figure 4.32). The anti-LexA L2 lariat had the extein sequence (SRSWDLPGEY). When glutamic acid in the extein region was mutated to alanine it resulted in ~ 61% lariat as analyzed by direct LC/MS (Figure 4.32). This was significant improvement over the original anti-LexA L2 lariat, which had only ~ 26% lariat (Figure 4.32). These values represent a low estimate of the amount of lariat present since we did not correct them for the lariat that was hydrolyzed during MS analysis. This mutant demonstrated that the extein region was involved in stability. Mutations that stabilize the lariat in the extein are less useful for short peptides displayed from the lariat since they interfere with the random region and restrict the amino acid space. However, mutations to the extein region that stabilize the lariat are still of interest for stabilizing proteins where the amino acids at the intein-extein junction can be fixed.

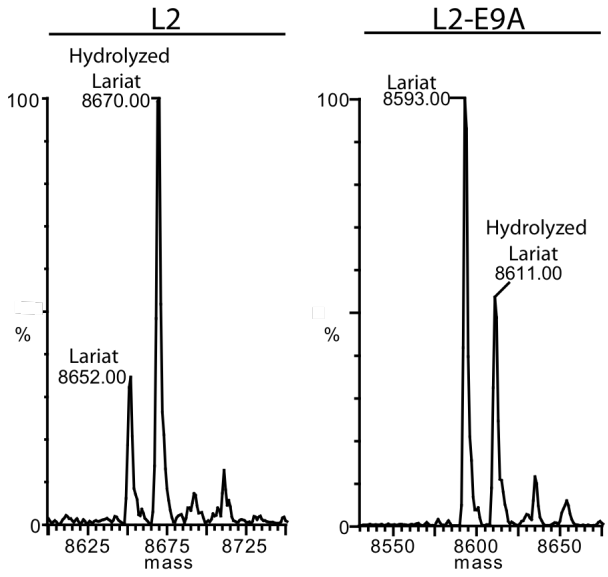


Figure 4.32 | LC/MS analysis of the L2 E9A lariat
ESI-TOF MS analysis of His-tag purified L2 and L2 E9A lariats produced in BL21-CP were purified by reverse phase HPLC and the combined m/z spectra were deconvoluted using MaxEnt (Waters) software to give a chromatogram of mass versus percent abundance. The L2 lariat (8651.7 calc; 8652 obs) and hydrolyzed L2 lariat (8669.7 calc; 8670 obs) His-tagged products and the L2 E9A lariat (8593.7 calc 8593; obs) and hydrolyzed L2 lariat 8611.7 calc; 8611 obs) His-tagged products are shown.

4.2.5 Conclusions

In summary, we have shown that specific mutations at sites G6/G7/B11 can stabilize the lariat. We have used this information to developed hypotheses for possible optimal mutations combinations.

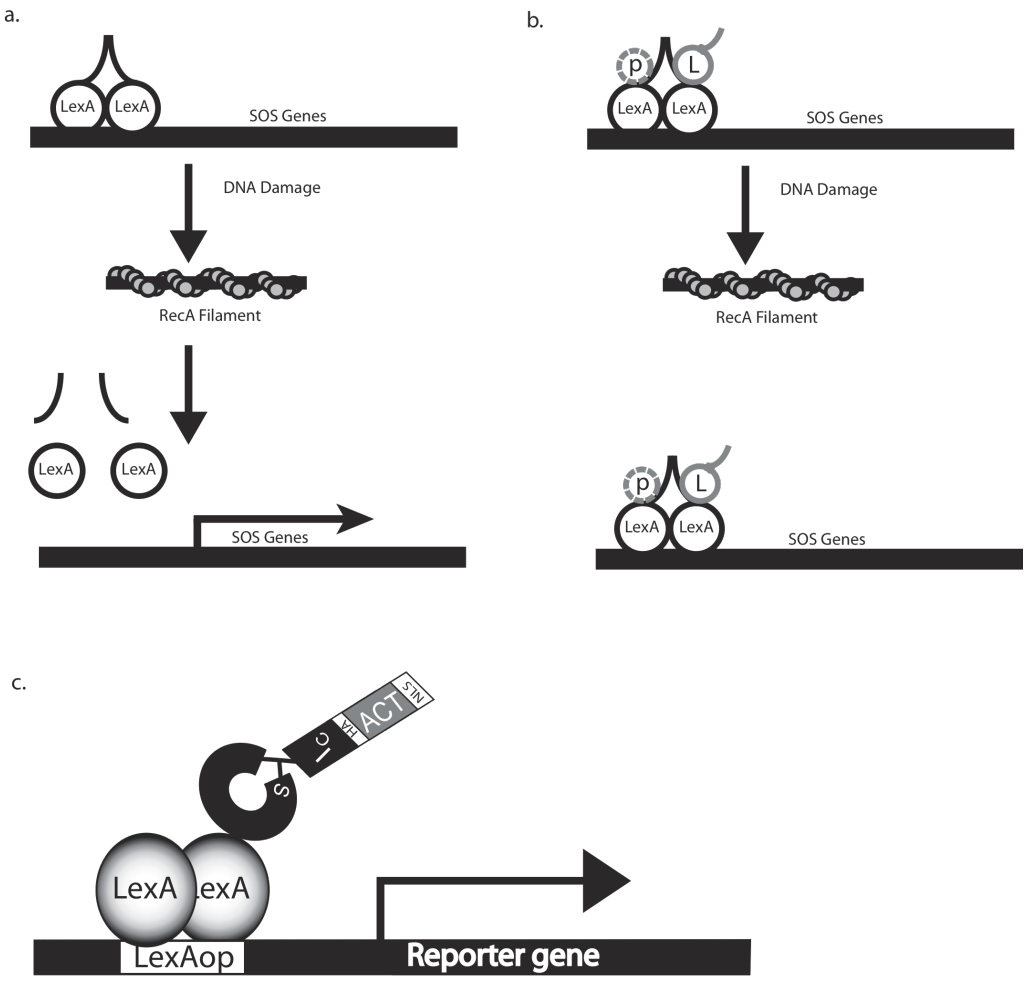
We have recently demonstrated that lariats can be synthetically produced, which eliminates the need to develop extremely stable lariats for applications that are compatible with synthetically produced lariats. After a lariat is isolated from the Y2H assay (or alternative assay), the corresponding synthetic lariat peptide can be chemically synthesized for *in vitro* studies. However, a lariat with enhanced stability would still be beneficial for *in vivo* assays, where lariats are expressed within a cell.

4.3 Aim 3: Screening lariat peptide libraries using the Y2H assay

4.3.1 LexA as a therapeutic target

Discovery of novel antibiotics are of general concern as 'superbugs' or bacteria that are resistant to antibiotics pose significant health risks and have associated costs (Woodford *et al.*, 2009). However, antibiotic resistance is an inevitable result of the use of antibiotics and bacterial evolution. Interestingly, antibiotics may even contribute to their own resistance by activating the SOS response pathway and inducing a hypermutagenic state (Miller *et al.*, 2004; Cirz *et al.*, 2005). Several SOS response genes have now been linked to bacterial resistance through a variety of mechanisms including inhibiting cell division, inducing recombination processes resulting in increased horizontal transfer of antibiotic resistance, and induction of error prone polymerases (Miller *et al.*, 2004; Cirz *et al.*, 2005; Guerin *et al.*, 2009). Antibiotics typically target a limited number of proteins and pathways (Miesel *et al.*, 2003; Brown and Nathwani, 2005). The average antibiotic is in its third or fourth generation of modification and use; it is unclear how much longer they will be effective (Miesel *et al.*, 2003). It is generally accepted that novel strategies to isolate antibiotics targeting new proteins and pathways are required (Davies, 1994; Barrett and Barrett, 2003).

We proposed to isolate lariats that potentiated the activity of existing cytotoxic reagents by blocking the SOS response pathway and inhibiting LexA. LexA is involved in the SOS response pathway of bacteria (Figure 4.33a). LexA is a repressor protein that regulates the expression of between twenty to forty genes, depending on the bacteria studied (Lin and Little,



d.

```
Lib: SXXXXXXXEY
 L1: SSWDLPQGEY
 L2: SRSWDLPGEY
```

Figure 4.33 | LexA and anti-LexA lariats
(a) LexA induction of SOS response. LexA binds as a dimer to DNA and blocks transcription of genes controlled by the LexA regulon. RecA is activated by forming filaments on ssDNA, which is created by DNA damage. Activated RecA binds to LexA and induces autoproteolysis, releasing LexA from the DNA and allowing transcription of the genes repressed by LexA. **(b)** LexA as an antimicrobial target. Synthetic cyclic peptides or lariat peptides, which bind to LexA, can prevent cleavage by RecA preventing induction of SOS response genes. (**c)** Isolation of lariats using Y2H assay. Lariats are selected that interact with LexA dimers using the Y2H assay. **(d)** Anti-LexA lariat sequences. Two unique sequences were isolated from the *Ssp-Ssp* R7 library (L1 and L2). "X" represents the combinatorial region of the peptide.

1988; Courcelle *et al.*, 2001; Au *et al.*, 2005; Jin *et al.*, 2007; Cirz *et al.*, 2007). LexA represses the translation of these genes by blocking transcription. LexA binds to DNA operators in the SOS regulon with the consensus sequence ($TACTG(TA)_5CAGTA$) (Walker, 1984). When LexA is cleaved, it no longer binds to DNA, alleviating transcriptional repression. LexA undergoes autoproteolytic cleavage at a very slow rate, allowing some turn over at LexA operators, but the primary mechanism of LexA degradation and thus activation of genes regulated by LexA is through RecA mediated co-proteolysis (Lin and Little, 1988). RecA is activated by the presence of single stranded DNA (ssDNA) that is generated when DNA is damaged. RecA forms filaments on ssDNA, converting RecA into a conformation that catalyzes the proteolysis of LexA. It has not been determined whether LexA cleavage occurs in solution or while bound to the operator. However evidence suggests LexA undergoes a conformational change when bound to DNA (Roland *et al.*, 1992). RecA has also been proposed to produce a change in LexA structure between a cleavable form and a non-cleavable form (Luo *et al.*, 2001). The cleavable and noncleavable conformation depends on whether the active site lysine is buried or exposed to the solvent.

We hypothesized that cyclic peptides that block the cleavage of LexA will affect the ability of the bacteria to respond to an antibiotic (Walker, 1984; Lin and Little, 1988) and increase the antibiotics potency. Inhibitors of LexA will also prevent bacteria from entering a hypermutagenic state (Cirz *et al.*, 2005; Miller *et al.*, 2004) and reduce antibiotic resistance. Since LexA is not present in humans, it would have no effect on host DNA damage repair systems.

4.3.2 Calculating the number of lariat library copies to screen

Using statistical analysis described by Patrick *et al.*, the optimal number of library copies to screen was determined by the size of the library (Figure 4.34) (Patrick *et al.*, 2003). The probability that a library is complete (P_C) and contains every distinct member (V) is a function of the number of distinct members (V) and the total number of clones (L).

$$P_c = (1 - e^{-L/V})^V \qquad \text{(eq: 1)}$$

The number of library copies is given by the total number of clones (L) divided by the number of distinct members (V). Figure 4.34 shows the graph of the probability of

completeness (P_c) values versus number of library copies (L/V) for several different library sizes (V). Equation 1 can also be rearranged to solve directly for the total number of clones required (L) (eq: 2) and the simplified equation (eq: 3) if $V >> -\ln P_c$, which is generally true for $P_C > 90\%$ (Patrick *et al.*, 2003).

$$L = -V \ln\left[1 - \exp\left(\frac{\ln P_c}{V}\right)\right] \qquad \textit{(eq: 2)}$$

$$L = -V \ln\left(-\frac{\ln P_c}{V}\right) \qquad \textit{where } V >> -\ln P_c \qquad \text{(eq: 3)}$$

This equation is useful to determine the exact number of clones required to reach the desired probability of completeness (P_C) given a known library size (V). Generally, we screen libraries containing 20 – 100 million members. Figure 4.34 shows us that 20 library copies is sufficient to ensure a 95% probability of screening a library with even 100 million different members. Therefore we screened 20 library copies in subsequent screens.

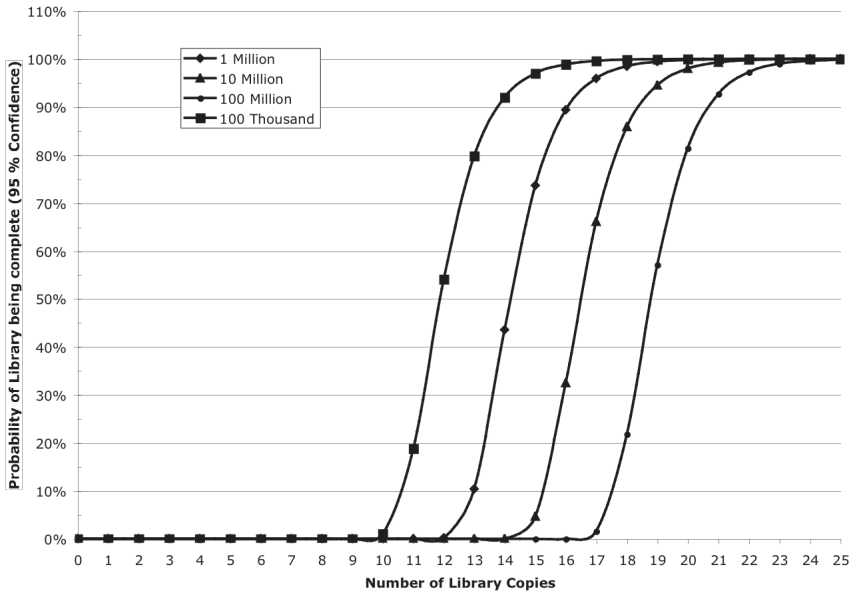


Figure 4.34 | Determining the number of library copies to screen
Graph of number of library copies (L/V) versus probability of the library being complete (P_C) for the following library sizes: 100,000 (squares), 1,000,000 (diamonds), 10,000,000 (triangles), and 100,000,000 (circles).

4.3.3 Screening the *Ssp-Ssp* R7 lariat library against LexA

We mated the *Ssp-Ssp* R7 library of approximately four million lariats in the MATa strain EY93 to the MATα strain EY111 containing the LexA target plasmid and Y2H reporter genes. Using the Y2H interaction trap in (Figure 4.33c), we isolated fourteen colonies encoding two unique lariats that interacted with LexA (Figure 4.33d): L1: (**S**SWDLPQG**EY**) and L2: (**S**RSWDLPG**EY**). The region in bold was fixed and the underlined region was the randomized region in the lariat library. Within the randomized region, both sequences contained the identical (SWDLP) motif. Both L1 and L2 randomized regions end in glycine. The difference between the two sequences was that the L1 sequence contained a glutamine after the SWDLP motif, whereas the L2 sequence had an arginine before the SWDLP motif.

Interestingly, both sequences contained glycine and proline that are required to make sharp bends in amino acid structure. These amino acids might allow the cyclic peptide to adopt a tightly folded structure by folding back on itself. We chose the L2 lariat for further analysis as it contained more charged amino acids. The grand average of hydropathicity calculated by EXPASY for the L2 peptide was -1.350 compared to -1.250 for the L1 peptide (Kyte and Doolittle, 1982), which we predicted would enhance its solubility.

4.3.4 Binding to LexA

4.3.4.1 Y2H mutants

To confirm the importance of the lariat structure for the L2 lariat-LexA interaction, we cloned the noose region of the L2 lariat into four modified intein expression plasmids along with three controls (Figure 4.37): (i) L2 lariat, (ii) L2 inactive, (iii) L2 linear, (iv) L2 active, (v) L2 scrambled, (vi) L2 inverted, and (vii) vector control (pIN01). The expression of these constructs was confirmed by Western blot analysis with an antibody against the N-terminal haemagglutinin (HA) tag and LC/MS analysis in *E. coli*.

The L2 lariat plasmid (Figure 4.37) produced two HA-tagged products, the unprocessed intein (~ 32 kDa) and the lariat (~ 18 kDa) (Figure 4.16). Processing of the L2 lariat was also confirmed and quantitated by LC/MS (Figure 4.35).

The L2 inactive intein plasmid (pIN-L2) (Figure 4.37) expressed a lariat precursor that did not undergo any steps in the intein-mediated cyclization reaction. The inactive intein contains the insert sequence (I_{C-2}/G6)-HASRSWDLPGEYA-(I_{N+1}/A1). In the inactive intein, the cysteine at position I_{N+1} was mutated to an alanine. This mutation blocked the first step in

the lariat producing intein reaction. The noose sequence remains identical to the L2 lariat and was constrained by the affinity of the I_C and I_N domains. The L2 inactive intein produced only unprocessed intein (~ 32 kDa).

The L2 linear intein plasmid (pLIN-L2) (Figure 4.37) expressed a linear version of the L2 peptide that was not cyclized by a lactone bond. The L2 linear intein had the noose sequence (I_C-2/G6)-HASRSWDLPGEY**-(I_{N+2}/A2) where (*) indicates a stop codon. Otherwise the sequence was identical to the L2 lariat. This plasmid produced a single (~ 18 kDa) HA-tagged product analogous to the hydrolyzed lariat.

The L2 active intein plasmid (pACT-L2) (Figure 4.37) expressed the lactam L2 peptide. It may also contain the unprocessed, the lariat, the linear, and the lactam L2. The L2 active has the noose sequence (I_{C-2}/G6)-HNSRSWDLPGEYC-(I_{N+1}/A1). The asparagine at position I_{C+2} replaces the alanine present in the L2 lariat. The asparagine allows the third step in intein processing to occur. The L2 active intein plasmid produced unprocessed and I_C domain (~ 17 kDa) HA-tagged products. In yeast, the L2 active intein processed poorly relative to the L2 lariat and the unprocessed product migrated slightly slower in a SDS-PAGE gel. Similar differences in active L2 intein and L2 lariat processing were also observed in *E. coli*, where after IPTG induction ~ 65% of the L2 lariat was correctly processed, whereas only ~ 19% of the active L2 intein underwent processing to the lariat intermediate (Figure 4.35-4.37).

The L2 scrambled (pIL-L2-scrambled) expressed a lariat with the same amino acids as the L2 lariat but in a random order. The L2 scrambled construct had the insert (I_{C-2}/G6)-HASDPGLRSWEYC-(I_{N+1}/A1). The L2 inverted plasmid (pIL-L2-inverted) expressed a lariat with the same amino acids as the L2 lariat in an inverted order. The L2 inverted construct had the insert (I_{C-2}/G6)-HASGPLDWSREYC-(I_{N+1}/A1). The control vector plasmid (pIN01) expresses an inactive intein with a CGPC noose. The control vector had the insert sequence (I_{C-2}/G6)-HSSCGPCEYA-(I_{N+1}/A1) and was the template that the other vectors were cloned into.

In the Y2H assay, we did not detect any interactions between LexA and the L2 inactive intein, the L2 active intein, the L2 linear peptide, or between L2 scrambled and L2 inverted controls (Figure 4.37). Under less stringent conditions where only the LEU2 reporter gene was used, a weak interaction with the L2 inactive and active constructs was observed (Figure 4.38). These results demonstrate that the lactone bond was important for the interaction in the Y2H assay and that the interaction was dependent on the peptide sequence in the noose.

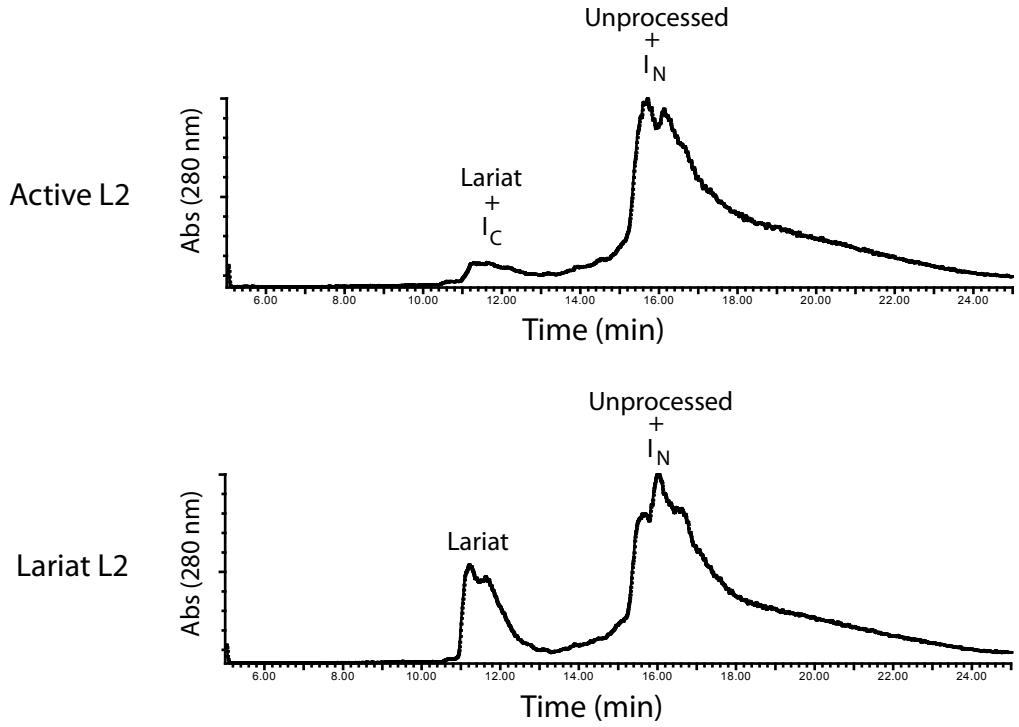
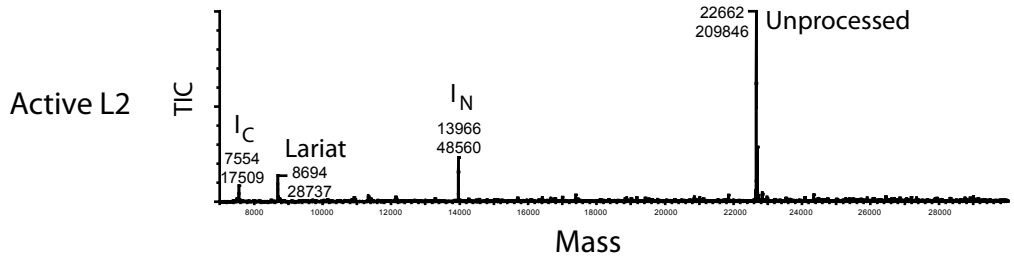
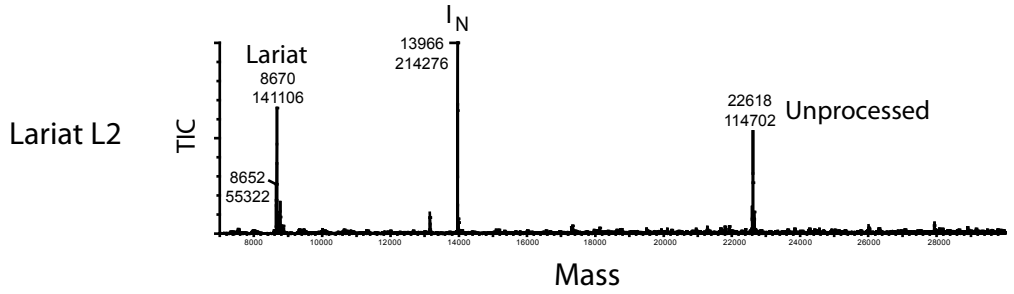


Figure 4.35 | HPLC analysis of L2 lariat and active L2 intein processing
Absorbance at 280 nm versus elution time of Histidine-tag purified L2 lariat and L2 active intein produced in BL21-CP. Peaks containing Lariat, Lariat + I_C, and I_N + unprocessed intein are shown.



$$\text{\% Lariat Processing} = \frac{I_N}{I_N + \text{Unprocessed}} = \frac{48560}{48560 + 209846} = 19\%$$



$$\text{\% Lariat Processing} = \frac{I_N}{I_N + \text{Unprocessed}} = \frac{214276}{214276 + 114702} = 65\%$$

Figure 4.36 | ESI-TOF MS analysis of L2 lariat and active intein processing
His-tag purified L2 lariat and L2 active intein produced in BL21-CP. The m/z spectra of the lariat products eluted from reverse phase HPLC column between 10 and 18 minutes were combined and deconvoluted using MaxEnt (Waters) software to give a chromatogram showing mass versus total ion count (TIC). For active L2, the I_C (7555.17 calc; 7554 obs) hydrolyzed lariat (8694.7 calc; 8694 obs), I_N (13966.7 calc; 13966 obs) and unprocessed (22661.4 calc; 22662 obs) His-tagged products were observed. For the L2 lariat, the lariat (8651.7 calc; 8652 obs), hydrolyzed lariat (8669.7 calc; 8670 obs), I_N fragment (13966.7 calc: 13966 obs), and unprocessed (22618.4 calc; 22618 obs) His-tagged products were observed. TIC for the area under each peak is shown below the mass. The percent processing to the lariat or lariat intermediate for the L2 lariat and L2 active intein are shown below each chromatogram, respectively.

a.

b.

Constructs	I_C	I_C-1	I_C+1	Random Library	I_N+1
(i) L2 Lariat	I_C	Ala	Ser	RSWDLPGEY	Cys
(ii) L2 Inactive	I_C	Ala	Ser	RSWDLPGEY	Ala
(iii) L2 Linear	I_C	Ala	Ser	RSWDLPGEY	Stop
(iv) L2 Active	I_C	Asn	Ser	RSWDLPGEY	Cys
(v) L2 Scrambled	I_C	Ala	Ser	DPGLRSWEY	Cys
(vi) L2 Inverted	I_C	Ser	Ser	GPLDWSREY	Ala
(vii) Vector	I_C	Ser	Ser	CPGCEY	Ala

c.

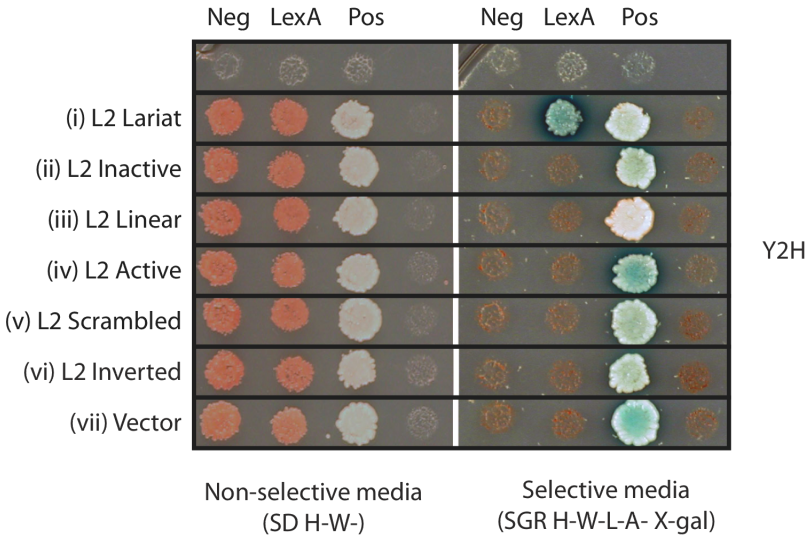


Figure 4.37 | Y2H LexA constructs

(a) Experimental overview of the Y2H assay used to determine which constructs interacted with LexA and LexA mutants. **(b)** Description of constructs, the constructs and their sequences are shown. I_C = Intein C-terminal domain, I_N = Intein N-terminal domain, amino acids are designated by single letter amino acid code, or three letter amino acid code**.** **(c)** Y2H analysis of the interaction of LexA with L2 lariat and intein constructs. Yeast growth on non-selective synthetic dextrose (SD) His^-, Trp^- media. Yeast growth on SGR His^-, Trp^-, Leu^-, Ade^-, X-Gal media, which selects for the activation of LEU2, ADE2, and LacZ Y2H reporter genes. Neg: prey was mated to His+ plasmid with no bait. LexA: prey was mated to pEG202 containing LexA bait. Pos: prey was mated to pEG202::B42AD self-activating bait.

4.3.4.2 Y2H alanine scan

To further explore how the L2 lariat interacted with LexA, we performed a series of alanine mutations in the noose region of the L2 lariat. An alanine scan involved sequentially mutating each amino acid to alanine (Figure 4.38). In this way, each amino acid side chain can be tested independently for its contribution to the interaction (Gregoret and Sauer, 1993; Morrison and Weiss, 2001). The L2 lariat sequence was tested against nine different alanine mutants. These alanine mutations were tested along with the anti-LexA L2 lariat for their ability to interact with LexA and two mutant forms of LexA, S119A and K156A (Figure 4.38). LexA exists in either a noncleavable (NC) or cleavable (C) conformation, where the C conformation accounts for < 0.1% of the LexA population at pH 7 (Roland *et al.*, 1992). Luo *et al.* proposed that RecA promotes LexA autoproteolysis by shifting the equilibrium towards the C conformation (Luo *et al.*, 2001). LexA S119A mutant has a change in the active site serine nucleophile that prevents cleavage (Slilaty and Little, 1987), but does not alter the NC conformation (Luo *et al.*, 2001). LexA K156A mutant has a change in the active site lysine general base that prevents cleavage and shifts the equilibrium towards the C conformation (Luo *et al.*, 2001). Previously, it has been shown using Aptaprint® technology (Baines and Colas, 2006) that the location of an interaction between a peptide affinity reagent and its target can be identified by making mutations in the target protein. If the mutation results in a loss of interaction with the peptide affinity reagent, then that site can be presumed to be either directly or indirectly involved in the interaction.

By analyzing the interactions between the L2 alanine mutants and the LexA mutants we could infer information on how the L2 lariat interacted with LexA. In the alanine scan with LexA, all of the amino acids were involved in the peptide LexA interaction with the exception of the E9A mutant, which interacted with similar strength to the L2 lariat sequence (Figure 4.38). When the less stringent Leu2 reporter gene was used, the G8 amino acid could be replaced with alanine without completely abolishing the LexA interaction (Figure 4.38). The G8 amino acid was likely not directly involved in the interaction and contributed to proper folding of the cyclic peptide since it is a highly flexible amino acid with no side chain.

The similarities and differences between the L1 and L2 sequences provide some insight into the interaction of the L2 lariat with LexA. The difference between the L1 and L2

sequences were that the L1 had a glutamine after the SWDLP motif and the L2 sequence had an arginine before the SWDLP motif. The arginine was found to be required for the interaction of the L2 lariat with LexA, which could indicate one of several things. Either the arginine was required for proper folding or the arginine was involved in the interaction. There may be a weak interaction between the R2A mutant and LexA as there seems to be some growth at the maximum density plated on the leucine minus selection media. Plating cells at a higher density or by using less stringent reporters like the 8-LexAop-Leu plasmid could be employed to examine this further.

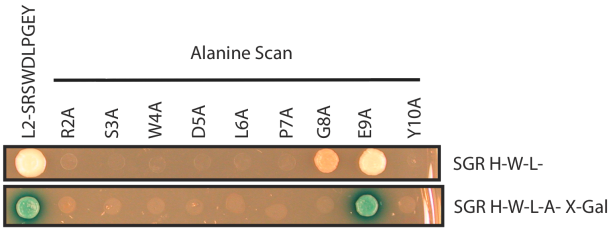
We also used the L2 lariat alanine scan to determine which amino acids were important for the interaction of the L2 lariat and two LexA mutants. The L2 lariat interacted with LexA and to a lesser extent LexA K156A but not with LexA S119A (Figure 4.38). This result suggested that the L2 lariat interacted near the LexA autoproteolysis active site possibly by forming a direct interaction with the S119 amino acid. In addition, the weaker interaction with the K156A mutant, which is proposed to shift the equilibrium to the C form of LexA, indicated the L2 lariat may interact with the NC form of LexA since the interaction with the K156A mutant was reduced. Further evidence for the interaction of the L2 lariat directly with LexA S119 was shown by the ability of the L2 E9A lariat mutant to interact with LexA. This result suggested that the E9 amino acid of the L2 lariat may interact with LexA S119.

a.

b.

Constructs	I_C	I_C-1	I_C +1	Random Library	I_N+1
(i) L2 Lariat	I_C	Ala	Ser	RSWDLPGEY	Cys
(ii) L2 R2A	I_C	Ala	Ser	ASWDLPGEY	Cys
(iii) L2 S3A	I_C	Ala	Ser	RAWDLPGEY	Cys
(iv) L2 W4A	I_C	Ala	Ser	RSADLPGEY	Cys
(v) L2 D5A	I_C	Ala	Ser	RSWALPGEY	Cys
(vi) L2 L6A	I_C	Ala	Ser	RSWDAPGEY	Cys
(vii) L2 P7A	I_C	Ala	Ser	RSWDLAGEY	Cys
(viii) L2 G8A	I_C	Ala	Ser	RSWDLPAEY	Cys
(ix) L2 E9A	I_C	Ala	Ser	RSWDLPGAY	Cys
(x) L2 Y10A	I_C	Ala	Ser	RSWDLPGEA	Cys

c.



d.

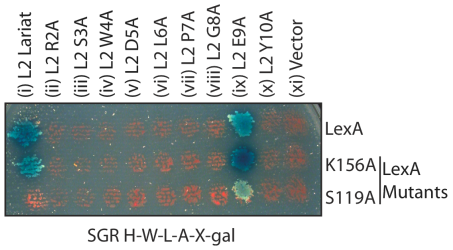


Figure 4.38 | Anti-LexA-L2 alanine scanning mutagenesis

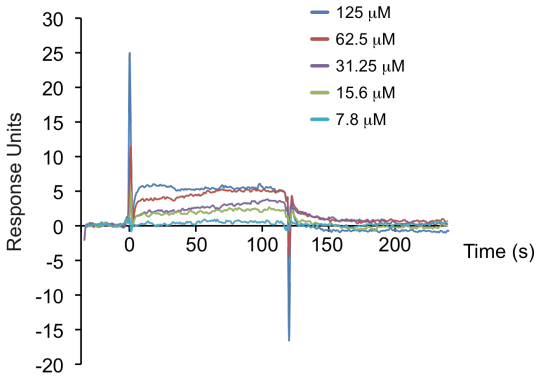
(a) Experimental overview of the Y2H assay used to determine which constructs interacted with LexA and LexA mutants. **(b)** Description of constructs, the constructs and their sequences are shown. I_C = Intein C-terminal domain, I_N = Intein N-terminal domain, amino acids are designated by single letter amino acid code, or three letter amino acid code. Alanine mutants are described by the noose position that has been changed to alanine. The I_{C+1} (G8) position is considered the first amino acid of the loop; it was not mutated since this would prevent the lariat forming and is not expected to interact. **(c)** Y2H alanine scan, showing the interaction of the alanine mutants with LexA on different selective media. Generally SD H-W-L- is considered less stringent than SGR H-W-L-A- X-gal media. **(d)** Y2H analysis with LexA mutants. Analysis of the interaction of LexA and LexA mutants, K156A, S119A, with L2 lariat and L2 lariat alanine mutants. Yeast were grown on SGR H^- W^- L^- A^- X-gal media that selects for the activation of LEU2, ADE2, and LacZ Y2H reporter genes.

4.3.4.3 Surface plasmon resonance analysis of binding

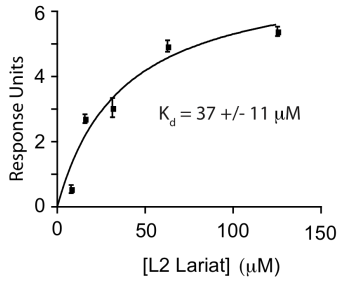
We used surface plasmon resonance (SPR) to determine the dissociation constant (K_d) between LexA and a synthetic L2 lariat peptide (Figure 4.39). We constructed a 6xHis-LexA fusion protein and bound it to a Biacore Ni^{2+}-NTA sensor chip (Nieba *et al.*, 1997). The synthetic L2 lariat peptide was synthesized with an alanine and biotin moiety at the N-terminus, replacing the I_C domain in the L2 lariat. The synthetic L2 lariat peptide interacted with LexA with a K_d of 37 μM. In contrast, we were unable to detect an interaction between LexA and the synthetic biotinylated linear L2 peptide at concentrations up to 125 μM (Figure 4.39).

Since SPR is not a label free system, the orientation and native conformation of LexA might be affected, which may alter its affinity with the L2 lariat peptide. When LexA was first bound to the Ni^{2+}-NTA chip, and then a synthetic LexA DNA operator was passed over the immobilized LexA there was a large decrease in signal. This decrease in signal could be a result of LexA dimerizing on the DNA and dissociating from the Ni^{2+}-NTA chip or possibly undergoing a conformational change when bound to DNA that prevented it from binding the sensor chip. The *in vitro* binding constant between L2 lariat peptide and LexA may differ from the *in vivo* binding constant since the L2 lariat may interact with a specific chromosome associated conformation of LexA. In *E. coli*, ~ 20% of LexA is free in solution, with the majority of LexA associated with DNA (Sassanfar and Roberts, 1990). This hypothesis is further supported by structural modeling and biochemical studies suggesting that LexA undergoes a conformational change when it binds its operator (Butala *et al.*, 2007; Chattopadhyaya and Pal, 2004; Groban *et al.*, 2005).

a.



b.



c.

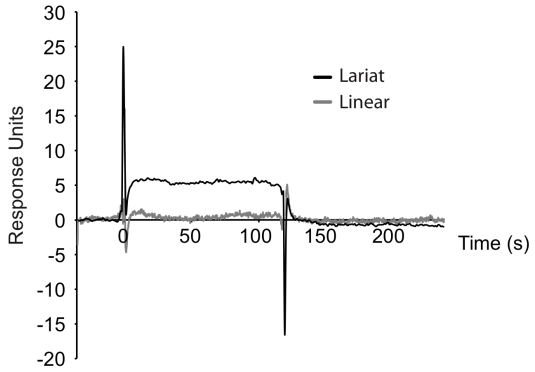


Figure 4.39 | Surface plasmon resonance (SPR) analysis L2 peptides with LexA
(a) SPR sensograms of L2 Lariat-LexA interaction. Overlay plot of the association and dissociation of synthetic L2 lariat peptide ranging in concentration from 7.8 μM to 125 μM with LexA immobilized on a Ni^{2+}-NTA SPR Chip. **(b)** Determination of dissociation constant. Maximum SPR response units were calculated for L2 lariat concentrations ranging from 7.8 μM to 125 μM. The binding isotherm was fit to a one-site binding model to calculate the K_d +/- standard error using Prism4.0 (GraphPad) statistical software. Error bars represent standard deviation for two experiments. **(c)** Linear versus lariat. Overlay plot comparing the association and dissociation of L2 lariat peptide and L2 linear peptide at 125 μM.

4.3.4.4 Intracellular interaction of LexA with L2 lariat

The weak binding of the LexA proteins with the anti-LexA L2 lariat might suggest that the anti-LexA L2 lariat interacted with the native LexA bound to DNA complex. We reasoned instead that the L2 lariat may interact preferentially with a specific chromosome associated conformation of LexA.

To investigate whether the L2 lariat was associated with the LexA promoter inside the cell, we used Ni^{2+}-nitrilotriacetic acid (Ni^{2+}-NTA) chromatin precipitation and quantitative PCR assay with His-tagged L2 lariat and Ni^{2+}-NTA affinity chromatography (Figure 4.40) (Tamimi *et al.*, 2004). We expressed the lariat in *E. coli* cells and then crosslinked protein-DNA and protein-protein-DNA complexes using formaldehyde. The DNA was sheared into small fragments using sonication and the DNA associated with the lariat was purified on a Ni^{2+}-NTA column. The crosslinks between the DNA associated with the His-tagged lariat were removed and the DNA was analyzed to see if there was an enrichment of the LexA specific genes with the L2 lariat compared to the non-specific control expressing the L2 scrambled lariat. Expression of the L2 lariat resulted in a fifty-fold enhancement in the amount of LexA promoter DNA precipitated in the induced versus uninduced samples compared to a ten-fold enhancement for the L2 scrambled control lariat (Figure 4.40). This indicated that in *E. coli*, there is an increased association of the L2 lariat with LexA than of the scrambled control lariat with LexA. Combined with the Y2H data, this supported the conclusion that there was a specific *in vivo* anti-LexA L2 lariat interaction with LexA.

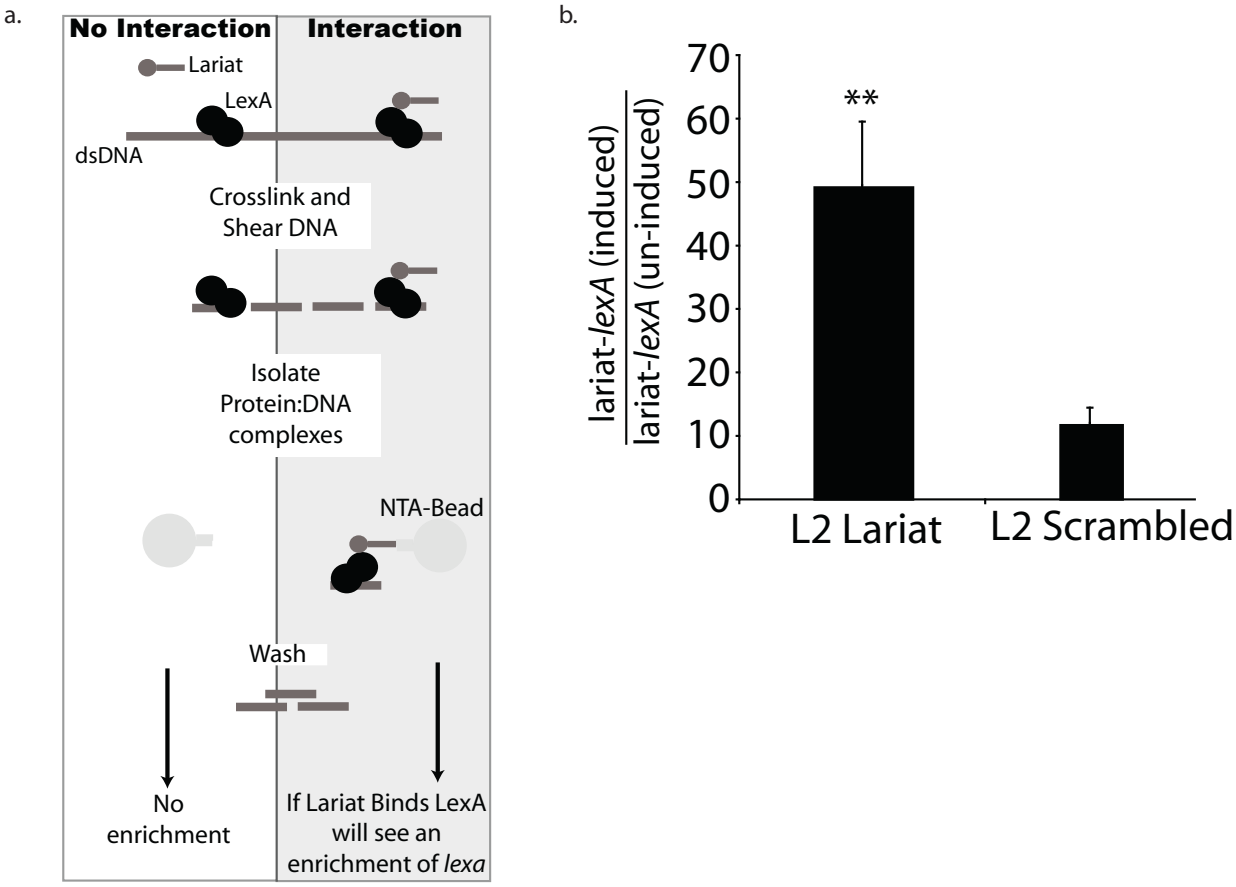


Figure 4.40 | Association of L2 lariat with LexA at *lexA* promoters
(a) Analysis of the association of the L2 and L2 scrambled lariats with the *lexA* promoter using Ni^{2+}-NTA chromatin precipitation assay. Precipitated *lexA* promoter DNA was quantified by real time PCR and normalized to the corresponding input DNA. **(b)** Histograms show the ratio of Ni^{2+}-NTA precipitated *lexA* promoter DNA (lariat-*lexA* complex) between IPTG-induced and un-induced His-tagged L2 and L2 scrambled lariats. Error bars represent the standard deviation from three independent experiments. (**) represents p-value < 0.01 between L2 lariat and L2 scrambled lariat.

4.3.5 L2 lariat potentiation of mitomycin C activity

We hypothesized that if LexA cleavage was inhibited that it would result in the potentiation of the action of cytotoxic drugs such as mitomycin C (MMC). Protecting LexA from cleavage prevents the SOS response genes from being activated, which would kill bacteria at a lower concentration of MMC than would be required if the SOS pathway was activated.

We tested the ability of L2 lariat to inhibit *E. coli* growth in the presence of MMC using the survival assay described by Lin and Little (Lin and Little, 1988). We expressed L2 scrambled lariat, L2 lariat, L2 linear peptide, L2 inactive intein, and L2 active intein in *E. coli*, exposed the bacteria to MMC in 0.85% NaCl for one hour, and assayed their survival (Figure 4.41). Expression of the L2 lariat enhanced the activity of MMC and reduced cell viability to ~ 30% relative to the L2 scrambled lariat. Expression of L2 linear peptide, L2 active intein, and L2 inactive intein enhanced the activity of MMC relative to the L2 scrambled control by ~ 68%, 65%, and 53%, respectively. Interestingly, this differed from the results observed in the Y2H assay where the L2 linear and L2 inactive had no effect. The effect in bacteria might be attributed to the high concentration of intein present in the cell.

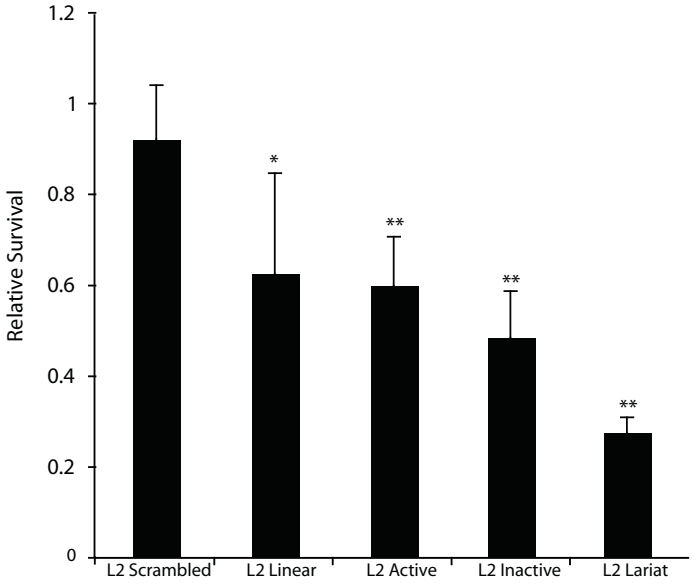


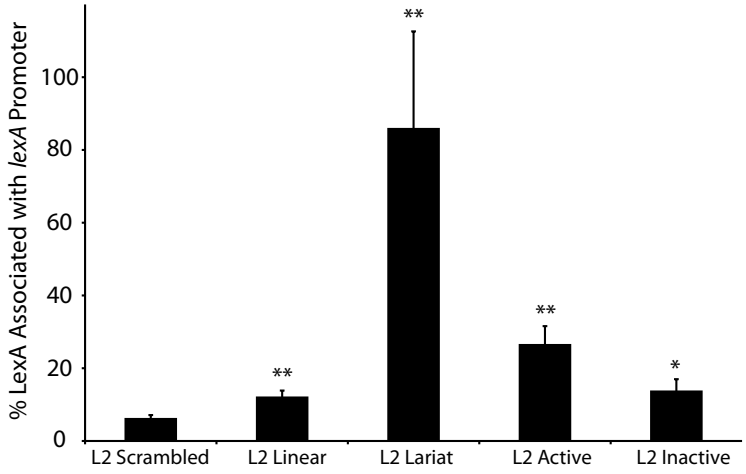
Figure 4.41 | Anti-LexA L2 potentiation of mitomycin C
Normalized relative cell survival was calculated by dividing the number of colony forming units (cfu) after 30 minutes MMC treatment by the number of cfu at the zero hour time point. Error bars represent the standard deviation of five independent experiments. (**) represents p-value < 0.01 and (*) represents p-value < 0.05 between L2 lariat and intein constructs and the L2 scrambled lariat.

4.3.6 LexA protection and stability

Since the L2 lariat potentiated the activity of MMC, we tested if the toxicity was mediated through LexA. We monitored the ability of the L2 lariat to inhibit MMC induced depletion of LexA at the *lexA* promoter using chromatin immunoprecipitation (ChIP) and quantitative PCR (Figure 4.42) (Wade *et al.*, 2005). MMC is a potent inducer of the bacterial SOS response that causes the release of LexA from its promoter (Lin and Little, 1988). We observed that MMC treatment depleted 94% of LexA from the *lexA* promoter in cells expressing the L2 scrambled lariat control, whereas only 14% of LexA was depleted in cells expressing the L2 lariat. Expression of L2 active intein, L2 inactive intein, and L2 linear peptide resulted in 73%, 86%, and 88% of LexA being depleted from *lexA* promoter after MMC treatment, respectively.

We monitored the ability of the L2 lariat to block MMC induced LexA autoproteolysis using Western blot analysis with an anti-LexA antibody. MMC activates RecA co-protease activity and induces cleavage of LexA (Lin and Little, 1988). We monitored degradation of LexA after exposure to MMC (Yasuda *et al.*, 1998) in the presence and absence of L2 lariat, L2 scrambled lariat, L2 linear peptide, L2 inactive intein, and L2 active intein (Figure 4.42). We observed that LexA was significantly degraded (98%) after 1 hr treatment with MMC in cells expressing L2 scrambled control. In contrast, the L2 lariat completely blocked MMC-induced LexA degradation. L2 linear peptide, L2 active intein, and L2 inactive intein partially blocked MMC-induced LexA proteolysis with 87%, 65%, and 67% of LexA remaining, respectively. A possible reason for the low activity of the L2 active intein in the activity assays was the low processing of L2 active intein in *E. coli* relative to the L2 lariat (Figure 4.35 - 4.37). The L2 lariat was identical to the L2 active intein except for an asparagine to alanine mutation at position I_{C-1}.

a.



b.

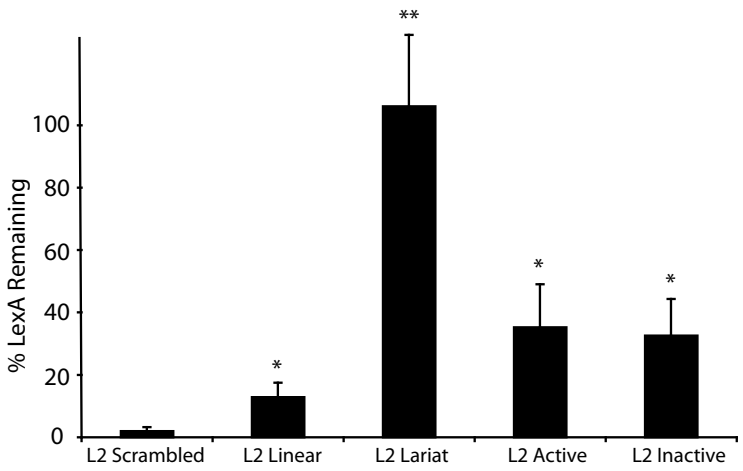


Figure 4.42 | anti-LexA L2 effect on LexA stability
L2 lariat inhibition of LexA autoproteolysis. **(a)** ChIP analysis of the influence of L2 lariat and intein constructs on MMC-induced depletion of LexA at the *lexA* promoter. The percentage of LexA remaining bound to the *lexA* promoter after MMC treatment was measured as a ratio of the occupancy units of MMC treated to untreated samples multiplied by 100. Error bars represent the standard deviation from three independent experiments. (**) represents p-value < 0.01 and (*) represents p-value < 0.05 between L2 lariat and intein constructs and the L2 scrambled lariat. **(b)** Western blot analysis of LexA cleavage induced by MMC treatment. BL21-CP cells expressing L2 lariat and intein constructs were treated with MMC and chloramphenicol. Cell extracts were analyzed by Western blot analysis using Anti-LexA antibody at 0 and 1 hour after MMC addition. The percent LexA that does not undergo proteolysis after 1 hour treatment with MMC is reported. Error bars represent the standard deviation from three independent experiments. (**) represents a p-value < 0.01 and (*) represents a p-value < 0.05 between L2 lariat and intein constructs and the L2 scrambled lariat.

4.3.7 Conclusions

We developed a new strategy to display lactone peptides with a covalently attached linear peptide. The N-terminus of the lariat was fused to a transcription activation domain, allowing lariats to be genetically screened using the Y2H assay. Alternative protein moieties such as localization sequences, fluorescent proteins, and membrane permeable peptides, *etc* could also be attached via the N-terminus. Lariats can be produced in prokaryotic and eukaryotic environments and the noose region can be used to display combinatorial peptide libraries. Lariats are constrained by a lactone bond, allowing a variety of peptide sequences to be displayed. Lactone bonds are stable under reducing conditions, which allows lariat libraries to be screened using intracellular assays such as the Y2H assay. This is in contrast to disulfide-bond constrained peptide libraries, which are not stable under reducing conditions and must be screened using *in vitro* selection strategies such as phage display (McLafferty *et al.*, 1993).

Lariats have advantages over intein-produced lactam-cyclized peptides for analyzing intracellular protein function. First, for *Ssp* DnaE intein-based constructs, the lariat is produced at higher levels than its corresponding L2 cyclic peptide. Second, lariat libraries can be screened using the Y2H assay to isolate lariats against a target protein. Combinatorial libraries of cyclic peptides generated by intein-mediated cyclization have been successfully screened to isolate cyclic peptides that inhibit protein interactions and activities (Naumann *et al.*, 2008; Horswill *et al.*, 2004; Tavassoli and Benkovic, 2005; Tavassoli *et al.*, 2008) and cell phenotypes (Kinsella *et al.*, 2002; Nilsson *et al.*, 2005). These selection strategies are more difficult to use than a Y2H assay to target a specific protein. In principle, it should be possible to generate cyclic peptides based on the noose sequence of the lariat that still retains the same activity as the lariat. Kinsella *et al.* showed that cyclic peptides that inhibit IL-4 signaling in B-cells still retain their activity when they were expressed in a mutated intein construct that did not produce cyclic peptides (Kinsella *et al.*, 2002).

Therefore, it is conceivable that in cases where the lariat cyclic peptide structure is not dependent on the geometry of the lactone bond, that it should be possible to generate peptides that are cyclized through an amide backbone bond, which retain the activity of the lariat peptide. Alternatively, cyclic peptides can also be synthesized that are cyclized through the side chain serine and C-terminal carboxylic acid (Sewald and Jakubke, 2002), which are identical to the noose region of the lariat.

In summary, we have presented a new method to isolate lariat peptides against a given target protein using the Y2H system. We used lariat peptide Y2H assay to generate lariat inhibitors of LexA and validate LexA as a therapeutic target for potentiating the antimicrobial effects of reagents that activate the SOS response pathway. The lariat technology provides a rapid high throughput assay for isolating peptide inhibitors that can be used for the reverse analysis of protein function, as drugs or pseudo-drugs for validating therapeutic targets.

4.4 Thesis summary

We demonstrated that inteins can be engineered to stably produce lariat peptides. Combinatorial libraries of lariat peptides can be constructed and screened, resulting in lariats that have biologically relevant activity against their intended targets. In addition, we have shown that the stability of the lariat can be improved and we have outlined several strategies to further improve lariat stability. We developed a genetic screen for isolating "lariat" peptides that function as *trans* dominant inhibitors of protein function. Lariats consist of a lactone peptide with a covalently attached transcription activation domain, which allows combinatorial lariat libraries to be screened for protein interactions using the yeast two-hybrid assay.

Lariat peptides possess desired traits for characterizing the function and therapeutic potential of proteins. Lariat peptides are constrained by a lactone bond, which allows them to be used in reducing intracellular environments. Lariats function as *trans* dominant inhibitors, which unlike genetic approaches that inactivate genes by deletion or creation of loss of function mutations, can be easily used to inhibit protein function in diploid organisms. Lariats inhibit their protein target directly and have the potential to block specific interactions with a protein while leaving other interactions unperturbed. This is in contrast to *trans* dominant agents like anti-sense RNA/DNA, ribozymes, and RNAi that block transcription and translation of their target. Thus, results obtained using lariat inhibitors can be used directly to evaluate the therapeutic potential of inhibiting targets with small molecule drugs. Together, these properties make lariats an ideal reagent for assessing whether small molecules inhibitors can be generated against a protein target. Further, the small size of lariats makes them amenable to chemical synthesis and allows their structure to be easily solved, which makes them useful as potential drugs or drug leads.

4.5 Future directions

4.5.1 Aim 1: Future directions

Additional studies are needed to determine the amino acids or amino acid combinations that block and allow lariat processing at the I_{C+1}, I_{C+2} positions. This information could be used to design new libraries, where a higher percentage of lariats undergo processing. Additionally, a new library should be constructed with variable I_{N-1} and I_{N-2} amino acids to determine their effect on processing and to compare the efficiency of processing of the *Ssp-Ssp* and *Ssp-Npu* libraries.

Lariats produced using inteins are not limited to displaying short peptide libraries in the Y2H assay. Inteins have previously been used to cyclize the full-length proteins beta-lactamase, and GFP (Iwai and Plückthun, 1999; Iwai *et al.*, 2001), indicating that lactone cyclization of proteins and not just peptides should also be possible.

Finally, the lariat technology should be applied to other selection strategies. This is necessary since to generate highly specific and functional antibodies or peptides, multiple rounds of selection in different organisms is often required. For example, phage display, Y2H assays, and mammalian two-hybrid assays have all been combined to isolate functional intracellular scFvs (Nam *et al.*, 2008). Some selection strategies, which could be performed with lariats are: (i) mRNA display, which would allow the incorporation of non-natural amino acids, which may be required to increase bioavailability; (ii) phage display, which would allow more library members to be screened; (iii) mammalian two-hybrid assay, which would allow the isolation of lariats that function directly in mammalian cells where ultimately therapeutic lariats must function. However, mammalian screens are limited to even smaller libraries than yeast screens. Ideally, retroviral vectors would be constructed that allow subcloning of lariats isolated in the Y2H assay. These lariats could then be tested for their appropriate phenotype in the relevant cellular background.

4.5.2 Aim 2: Future directions

Lariat production has been enhanced by mutagenesis of amino acids in the vicinity of the lactone bond. To further optimize the lariat's stability, mutations isolated from our rational and random mutagenic screens should be characterized in greater detail. Characterization of these mutants would allow us to better understand the mechanism of stabilization. Specifically determining the rate of cleavage for various mutants at different pH's could be used to

determine if cleavage of the lactone bond is through acid or base catalyzed reactions, which would allow us to better understand the mechanism of lactone bond cleavage and how these amino acids stabilize the lactone bond.

In addition to the G6/G7/B11 positions, the following positions were also identified and would be of potential interest. First, aspartic acid at position F4 coordinates water near the lactone bond and participates in N-S acyl transfer and transesterification by polarizing the carbonyl to assist in nucleophilic attack by cysteine or serine at positions I_{N+1}/A1 and I_{C+1}/G8. Mutation of F4 from aspartic acid to glutamic acid or glutamine may allow N-S acyl transfer and transesterification to occur, while stabilizing the lactone bond.

Second, mutation of histidine at position F13 to alanine does not block asparagine cyclization, but substitution of a bulky hydrophobic amino acid at F13 may be used to stabilize the lactone bond. Therefore, substitution with bulky hydrophobic amino acids like phenylalanine, leucine, or isoleucine could be tried at the F13 position.

Third, bulky or charged amino acids such as: tryptophan, phenylalanine, tyrosine, leucine, lysine, or arginine, substituted at the F14 position may be used to disrupt the correct positioning of histidine at position F13 and thus block asparagine cyclization and stabilize the lactone bond.

Fourth, mutation of phenylalanine at position F15 to alanine blocks asparagine cyclization, while mutation to tyrosine slightly inhibits asparagine cyclization. Accordingly, mutation of F15 to a bulky hydrophobic amino acid may be used to block asparagine cyclization and stabilize the lactone bond by excluding water. Tryptophan and leucine are good candidates for mutations at F13 to stabilize the lactone bond.

Alternatively, disrupting the charged surface responsible for interaction of the I_N and I_C domains (Dassa *et al.*, 2007) would optimizing this interface for a single turnover reaction that still allows the I_C and I_N domains to interact for a high processing rate, but is not sufficiently strong to remain attached after I_C domain release and would prevent the reverse transesterification reaction, thus stabilizing the lariat.

4.5.3 Aim 3: Future directions

Future advances in the lariat technology will require the development of strategies to transition lariats into valuable biological molecules. One general technique to isolate therapeutic peptide inhibitors is to first screen a combinatorial pool of peptides to isolate

peptide affinity reagents that interact with the target of interest. Then, the peptide affinity reagents isolated in the screen are tested for the desired inhibitory activity. Finally, the selected peptide inhibitor is optimized for cellular uptake and other pharmaceutical properties, or used as a lead molecule in small molecule design (Hruby *et al.*, 2002).

Since lariats are a new class of peptide affinity reagent, this three-step process must be optimized. In the immediate future, rigorous selection techniques that result in high affinity binding lariats must be developed. The low dissociation constant for the anti-LexA L2 lariat isolated using this technology, and studies on antibody fragments isolated in yeast (Tanaka and Rabbitts, 2009), suggest that isolation of therapeutic lariats or any peptide affinity reagent using only the Y2H assay will not be sufficient for most therapeutic targets. Therefore, combinations of different screening/selection strategies can be employed to isolate high affinity binders, including phage display, mRNA display, Y2H assays, and mammalian two-hybrid assays. These high affinity lariats can initially be used to replace antibodies and can be used directly in diagnostic and research applications that simply require the lariat to bind the target of interest. The advantage of these peptides over antibodies is their small size, which allows them to be synthesized, reducing their cost.

Rigorous selections, followed by secondary selection techniques that identify lariats with inhibitory activity, must also be performed. One strategy to isolate inhibitors for human tyrosine kinases would be to select peptides that could overcome the growth suppression phenotype in yeast (Takashima *et al.*, 2003; Trible *et al.*, 2006). Another strategy to isolate inhibitors that block protein-protein interactions would be to use a Y3H assay such as the reverse transactivator assay (RTA) (Hirst *et al.*, 2001; Joshi *et al.*, 2007) to select peptides that disrupt a known interaction. Alternatively, lariat peptides selected in the Y2H assay could be optimized for inhibitory activity using chemically synthesized lariats.

Lariats with inhibitory activity can then be transitioned into small molecules using yeast-based assays, which screen libraries of small molecules for ones that are able to disrupt a lariat-protein interaction (Joshi *et al.*, 2007). Alternatively, lariats can be modeled *in silico* (Yongye *et al.*, 2009; Maupetit *et al.*, 2009). Structures generated in *in silico* can then be used in docking programs to predict where and how the peptide interacts with its target (Maupetit *et al.*, 2009). This information can then used to generate small molecule drugs (Rubinstein and Niv, 2009). Since protein:protein interactions are becoming interesting drug targets, some

companies are pursuing small molecules with peptide-like motifs called peptidic modulators (Rubinstein and Niv, 2009). The fusion of the lariat peptide technology with this type of small molecule could potentially allow the creation of a small molecule with the same specificity as the peptide with improved drug qualities.

Even if lariats do not have an inhibitory effect, and only bind their target, their free N-terminus makes them well-suited to attach toxic chemicals, drugs and peptides, allowing them to target specific proteins or cell types (Ruoslahti, 2000).

5 References

Abed, N., Bickle, M., Mari, B., Schapira, M., Sanjuan-España, R., Robbe Sermesant, K., Moncorgé, O., Mouradian-Garcia, S., Barbry, P., *et al.* (2007). A comparative analysis of perturbations caused by a gene knock-out, a dominant negative allele, and a set of peptide aptamers. Mol Cell Proteomics *6*, 2110-2121.

Abedi, M., Caponigro, G., Shen, J., Hansen, S., Sandrock, T., and Kamb, A. (2001). Transcriptional transactivation by selected short random peptides attached to lexA-GFP fusion proteins. BMC Mol Biol *2*, 10.

Abedi, M.R., Caponigro, G., and Kamb, A. (1998). Green fluorescent protein as a scaffold for intracellular presentation of peptides. Nucleic Acids Res *26*, 623-630.

Abel-Santos, E., Scott, C.P., and Benkovic, S.J. (2003). Use of inteins for the in vivo production of stable cyclic peptide libraries in E. coli. Methods Mol Biol *205*, 281-294.

Agrawal, S., and Kandimalla, E.R. (2004). Role of Toll-like receptors in antisense and siRNA [corrected]. Nat Biotechnol *22*, 1533-1537.

Aires da Silva, F., Corte-Real, S., and Goncalves, J. (2008). Recombinant antibodies as therapeutic agents: pathways for modeling new biodrugs. BioDrugs *22*, 301-314.

Altamura, M., Dragoni, E., Infantino, A.S., Legnani, L., Ludbrook, S.B., Menchi, G., Toma, L., and Nativi, C. (2009). Cyclic glycopeptidomimetics through a versatile sugar-based scaffold. Bioorg Med Chem Lett *19*, 3841-3844.

Amitai, G., Callahan, B.P., Stanger, M.J., Belfort, G., and Belfort, M. (2009). Modulation of intein activity by its neighboring extein substrates. Proc Natl Acad Sci U S A *106*, 11005-11010.

Amitai, G., Dassa, B., and Pietrokovski, S. (2004). Protein splicing of inteins with atypical glutamine and aspartate C-terminal residues. J Biol Chem *279*, 3121-3131.

Amstutz, P., Koch, H., Binz, H.K., Deuber, S.A., and Plückthun, A. (2006). Rapid selection of specific MAP kinase-binders from designed ankyrin repeat protein libraries. Protein Eng Des Sel *19*, 219-229.

Andrade, M.A., Perez-Iratxeta, C., and Ponting, C.P. (2001). Protein repeats: structures, functions, and evolution. J Struct Biol *134*, 117-131.

Antos, J.M., Popp, M.W., Ernst, R., Chew, G.L., Spooner, E., and Ploegh, H.L. (2009). A straight path to circular proteins. J Biol Chem *284*, 16028-16036.

Appert, A., Nam, C.H., Lobato, N., Priego, E., Miguel, R.N., Blundell, T., Drynan, L., Sewell, H., Tanaka, T., and Rabbitts, T. (2009). Targeting LMO2 with a peptide aptamer establishes a necessary function in overt T-cell neoplasia. Cancer Res *69*, 4784-4790.

Aranko, A.S., Züger, S., Buchinger, E., and Iwaï, H. (2009). In vivo and in vitro protein ligation by naturally occurring and engineered split DnaE inteins. PLoS ONE *4*, e5185.

Armon-Omer, A., Levin, A., Hayouka, Z., Butz, K., Hoppe-Seyler, F., Loya, S., Hizi, A., Friedler, A., and Loyter, A. (2008). Correlation between shiftide activity and HIV-1 integrase inhibition by a peptide selected from a combinatorial library. J Mol Biol *376*, 971-982.

Aswad, D.W., Paranandi, M.V., and Schurter, B.T. (2000). Isoaspartate in peptides and proteins: formation, significance, and analysis. J Pharm Biomed Anal *21*, 1129-1136.

Au, N., Kuester-Schoeck, E., Mandava, V., Bothwell, L.E., Canny, S.P., Chachu, K., Colavito, S.A., Fuller, S.N., Groban, E.S., *et al.* (2005). Genetic composition of the Bacillus subtilis SOS system. J Bacteriol *187*, 7655-7666.

Baek, D., Villén, J., Shin, C., Camargo, F.D., Gygi, S.P., and Bartel, D.P. (2008). The impact of microRNAs on protein output. Nature *455*, 64-71.

Baines, I.C., and Colas, P. (2006). Peptide aptamers as guides for small-molecule drug discovery. Drug Discov Today *11*, 334-341.

Baker, K., Bleczinski, C., Lin, H., Salazar-Jimenez, G., Sengupta, D., Krane, S., and Cornish, V.W. (2002). Chemical complementation: a reaction-independent genetic assay for enzyme catalysis. Proc Natl Acad Sci U S A *99*, 16537-16542.

Baldwin, J.J. (1996). Design, synthesis and use of binary encoded synthetic chemical libraries. Mol Divers *2*, 81-88.

Bardou, C., Borie, C., Bickle, M., Rudkin, B.B., and Colas, P. (2009). Peptide aptamers for small molecule drug discovery. Methods Mol Biol *535*, 373-388.

Barrett, C.T., and Barrett, J.F. (2003). Antibacterials: are the new entries enough to deal with the emerging resistance problems? Curr Opin Biotechnol *14*, 621-626.

Barrett, J.W., Sun, Y., Nazarian, S.H., Belsito, T.A., Brunetti, C.R., and McFadden, G. (2006). Optimization of codon usage of poxvirus genes allows for improved transient expression in mammalian cells. Virus Genes *33*, 15-26.

Becker, F., Murthi, K., Smith, C., Come, J., Costa-Roldán, N., Kaufmann, C., Hanke, U., Degenhart, C., Baumann, S., *et al.* (2004). A three-hybrid approach to scanning the proteome for targets of small molecule kinase inhibitors. Chem Biol *11*, 211-223.

Bennetzen, J.L., and Hall, B.D. (1982). Codon selection in yeast. J Biol Chem *257*, 3026-3031.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2009). GenBank. Nucleic Acids Res *37*, D26-D31.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. (2004). GenBank: update. Nucleic Acids Res *32*, D23-D26.

Berer, N., Rudi, A., Goldberg, I., Benayahu, Y., and Kashman, Y. (2004). Callynormine A, a new marine cyclic peptide of a novel class. Org Lett *6*, 2543-2545.

Berg, T. (2003). Modulation of protein-protein interactions with small organic molecules. Angew Chem Int Ed Engl *42*, 2462-2481.

Bernstein, D.S., Buter, N., Stumpf, C., and Wickens, M. (2002). Analyzing mRNA-protein complexes using a yeast three-hybrid system. Methods *26*, 123-141.

Bertani, G. (1951). Studies on lysogenesis. I. The mode of phage liberation by lysogenic Escherichia coli. J Bacteriol *62*, 293-300.

Bertani, G. (2004). Lysogeny at mid-twentieth century: P1, P2, and other experimental systems. J Bacteriol *186*, 595-600.

Binz, H.K., Amstutz, P., Kohl, A., Stumpp, M.T., Briand, C., Forrer, P., Grütter, M.G., and Plückthun, A. (2004). High-affinity binders selected from designed ankyrin repeat protein libraries. Nat Biotechnol *22*, 575-582.

Bird, R.E., Hardman, K.D., Jacobson, J.W., Johnson, S., Kaufman, B.M., Lee, S.M., Lee, T., Pope, S.H., Riordan, G.S., and Whitlow, M. (1988). Single-chain antigen-binding proteins. Science *242*, 423-426.

Birmingham, A., Anderson, E.M., Reynolds, A., Ilsley-Tyree, D., Leake, D., Fedorov, Y., Baskerville, S., Maksimova, E., Robinson, K., *et al.* (2006). 3' UTR seed matches, but not overall identity, are associated with RNAi off-targets. Nat Methods *3*, 199-204.

Blum, J.H., Dove, S.L., Hochschild, A., and Mekalanos, J.J. (2000). Isolation of peptide aptamers that inhibit intracellular processes. Proc Natl Acad Sci U S A *97*, 2241-2246.

Boder, E.T., and Wittrup, K.D. (1997). Yeast surface display for screening combinatorial polypeptide libraries. Nat Biotechnol *15*, 553-557.

Borel, J.F. (2002). History of the discovery of cyclosporin and of its early pharmacological development. Wien Klin Wochenschr *114*, 433-437.

Borghouts, C., Kunz, C., Delis, N., and Groner, B. (2008). Monomeric recombinant peptide aptamers are required for efficient intracellular uptake and target inhibition. Mol Cancer Res *6*, 267-281.

Bork, P. (1993). Hundreds of ankyrin-like repeats in functionally diverse proteins: mobile modules that cross phyla horizontally? Proteins *17*, 363-374.

Boulianne, G.L., Hozumi, N., and Shulman, M.J. (1984). Production of functional chimaeric mouse/human antibody. Nature *312*, 643-646.

Böttger, A., Böttger, V., Sparks, A., Liu, W.L., Howard, S.F., and Lane, D.P. (1997). Design of a synthetic Mdm2-binding mini protein that activates the p53 response in vivo. Curr Biol *7*, 860-869.

Brat, D., Boles, E., and Wiedemann, B. (2009). Functional expression of a bacterial xylose isomerase in Saccharomyces cerevisiae. Appl Environ Microbiol *75*, 2304-2311.

Broder, Y.C., Katz, S., and Aronheim, A. (1998). The ras recruitment system, a novel approach to the study of protein-protein interactions. Curr Biol *8*, 1121-1124.

Brody, J.R., Calhoun, E.S., Gallmeier, E., Creavalle, T.D., and Kern, S.E. (2004). Ultra-fast high-resolution agarose electrophoresis of DNA and RNA using low-molarity conductive media. Biotechniques *37*, 598-602.

Brown, E.M., and Nathwani, D. (2005). Antibiotic cycling or rotation: a systematic review of the evidence of efficacy. J Antimicrob Chemother *55*, 6-9.

Bruno, W.J., Knill, E., Balding, D.J., Bruce, D.C., Doggett, N.A., Sawhill, W.W., Stallings, R.L., Whittaker, C.C., and Torney, D.C. (1995). Efficient pooling designs for library screening. Genomics *26*, 21-30.

Buerger, C., Nagel-Wolfrum, K., Kunz, C., Wittig, I., Butz, K., Hoppe-Seyler, F., and Groner, B. (2003). Sequence-specific peptide aptamers, interacting with the intracellular domain of the epidermal growth factor receptor, interfere with Stat3 activation and inhibit the growth of tumor cells. J Biol Chem *278*, 37610-37621.

Busche, A.E., Aranko, A.S., Talebzadeh-Farooji, M., Bernhard, F., Dötsch, V., and Iwaï, H. (2009). Segmental isotopic labeling of a central domain in a multidomain protein by protein trans-splicing using only one robust DnaE intein. Angew Chem Int Ed Engl *48*, 6128-6131.

Butala, M., Hodoscek, M., Anderluh, G., Podlesek, Z., and Zgur-Bertok, D. (2007). Intradomain LexA rotation is a prerequisite for DNA binding specificity. FEBS Lett *581*, 4816-4820.

Butz, K., Denk, C., Fitscher, B., Crnkovic-Mertens, I., Ullmann, A., Schröder, C.H., and Hoppe-Seyler, F. (2001). Peptide aptamers targeting the hepatitis B virus core protein: a new class of molecules with antiviral activity. Oncogene *20*, 6579-6586.

Butz, K., Denk, C., Ullmann, A., Scheffner, M., and Hoppe-Seyler, F. (2000). Induction of apoptosis in human papillomaviruspositive cancer cells by peptide aptamers targeting the viral E6 oncoprotein. Proc Natl Acad Sci U S A *97*, 6693-6697.

Caponigro, G., Abedi, M.R., Hurlburt, A.P., Maxfield, A., Judd, W., and Kamb, A. (1998). Transdominant genetic analysis of a growth control pathway. Proc Natl Acad Sci U S A *95*, 7508-7513.

Carthew, R.W., and Sontheimer, E.J. (2009). Origins and Mechanisms of miRNAs and siRNAs. Cell *136*, 642-655.

Castanotto, D., Sakurai, K., Lingeman, R., Li, H., Shively, L., Aagaard, L., Soifer, H., Gatignol, A., Riggs, A., and Rossi, J.J. (2007). Combinatorial delivery of small interfering RNAs reduces RNAi efficacy by selective incorporation into RISC. Nucleic Acids Res *35*, 5154-5164.

Cerantola, V., Guillas, I., Roubaty, C., Vionnet, C., Uldry, D., Knudsen, J., and Conzelmann, A. (2009). Aureobasidin A arrests growth of yeast cells through both ceramide intoxication and deprivation of essential inositolphosphorylceramides. Mol Microbiol *71*, 1523-1537.

Chang, H.W., Nguyen, D.D., Washington, E., Walker, I.D., and Holloway, S.A. (2006). Phage display antibodies to allelic determinants of canine blood cells. J Immunol Methods *311*, 1-11.

de Chassey, B., Mikaelian, I., Mathieu, A.L., Bickle, M., Olivier, D., Nègre, D., Cosset, F.L., Rudkin, B.B., and Colas, P. (2007). An antiproliferative genetic screening identifies a peptide aptamer that targets calcineurin and up-regulates its activity. Mol Cell Proteomics *6*, 451-459.

Chattopadhyay, A., Tate, S.A., Beswick, R.W., Wagner, S.D., and Ko Ferrigno, P. (2006). A peptide aptamer to antagonize BCL-6 function. Oncogene *25*, 2223-2233.

Chattopadhyaya, R., and Pal, A. (2004). Improved model of a LexA repressor dimer bound to recA operator. J Biomol Struct Dyn *21*, 681-689.

Chen, X., Xu, M.Q., Ding, Y., Ferrandon, S., and Rao, Z. (2002). Crystallographic study of a naturally occurring trans-splicing intein from Synechocystis sp. PCC 6803. Acta Crystallogr D Biol Crystallogr *58*, 1204-1206.

Cheng, L., Naumann, T.A., Horswill, A.R., Hong, S.J., Venters, B.J., Tomsho, J.W., Benkovic, S.J., and Keiler, K.C. (2007). Discovery of antibacterial cyclic peptides that inhibit the ClpXP protease. Protein Sci *16*, 1535-1542.

Cheng, Y.Q., and Walton, J.D. (2000). A eukaryotic alanine racemase gene involved in cyclic peptide biosynthesis. J Biol Chem *275*, 4906-4911.

Chong, S., Mersha, F.B., Comb, D.G., Scott, M.E., Landry, D., Vence, L.M., Perler, F.B., Benner, J., Kucera, R.B., *et al.* (1997). Single-column purification of free recombinant proteins using a self-cleavable affinity tag derived from a protein splicing element. Gene *192*, 271-281.

Chong, S., Montello, G.E., Zhang, A., Cantor, E.J., Liao, W., Xu, M.Q., and Benner, J. (1998). Utilizing the C-terminal cleavage activity of a protein splicing element to purify recombinant proteins in a single chromatographic step. Nucleic Acids Res *26*, 5109-5115.

Chong, S., Shao, Y., Paulus, H., Benner, J., Perler, F.B., and Xu, M.Q. (1996). Protein splicing involving the Saccharomyces cerevisiae VMA intein. The steps in the splicing

pathway, side reactions leading to protein cleavage, and establishment of an in vitro splicing system. J Biol Chem *271*, 22159-22168.

Chong, S., Williams, K.S., Wotkowicz, C., and Xu, M.Q. (1998). Modulation of protein splicing of the Saccharomyces cerevisiae vacuolar membrane ATPase intein. J Biol Chem *273*, 10567-10577.

Cirz, R.T., Chin, J.K., Andes, D.R., de Crécy-Lagard, V., Craig, W.A., and Romesberg, F.E. (2005). Inhibition of mutation and combating the evolution of antibiotic resistance. PLoS Biol *3*, e176.

Cirz, R.T., Jones, M.B., Gingles, N.A., Minogue, T.D., Jarrahi, B., Peterson, S.N., and Romesberg, F.E. (2007). Complete and SOS-mediated response of Staphylococcus aureus to the antibiotic ciprofloxacin. J Bacteriol *189*, 531-539.

Clauser, K.R., Baker, P., and Burlingame, A.L. (1999). Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. Anal Chem *71*, 2871-2882.

Cochran, A.G. (2000). Antagonists of protein-protein interactions. Chem Biol *7*, R85-R94.

Cohen, B.A., Colas, P., and Brent, R. (1998). An artificial cell-cycle inhibitor isolated from a combinatorial library. Proc Natl Acad Sci U S A *95*, 14272-14277.

Colas, P., Cohen, B., Jessen, T., Grishina, I., McCoy, J., and Brent, R. (1996). Genetic selection of peptide aptamers that recognize and inhibit cyclin-dependent kinase 2. Nature *380*, 548-550.

Colas, P., Cohen, B., Ko Ferrigno, P., Silver, P.A., and Brent, R. (2000). Targeted modification and transportation of cellular proteins. Proc Natl Acad Sci U S A *97*, 13720-13725.

Collet, J.F., Peisach, D., Bardwell, J.C., and Xu, Z. (2005). The crystal structure of TrxA(CACA): Insights into the formation of a [2Fe-2S] iron-sulfur cluster in an Escherichia coli thioredoxin mutant. Protein Sci *14*, 1863-1869.

Cotton, F.A., Hazen, E.E., and Legg, M.J. (1979). Staphylococcal nuclease: proposed mechanism of action based on structure of enzyme-thymidine 3',5'-bisphosphate-calcium ion complex at 1.5-A resolution. Proc Natl Acad Sci U S A *76*, 2551-2555.

Courcelle, J., Khodursky, A., Peter, B., Brown, P.O., and Hanawalt, P.C. (2001). Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient Escherichia coli. Genetics *158*, 41-64.

Craik, D.J. (2006). Chemistry. Seamless proteins tie up their loose ends. Science *311*, 1563-1564.

Craik, D.J., Daly, N.L., Bond, T., and Waine, C. (1999). Plant cyclotides: A unique family of cyclic and knotted proteins that defines the cyclic cystine knot structural motif. J Mol Biol *294*, 1327-1336.

Craik, D.J., Simonsen, S., and Daly, N.L. (2002). The cyclotides: novel macrocyclic peptides as scaffolds in drug design. Curr Opin Drug Discov Devel *5*, 251-260.

Crnković-Mertens, I., Bulkescher, J., Mensger, C., Hoppe-Seyler, F., and Hoppe-Seyler, K. (2010). Isolation of peptides blocking the function of anti-apoptotic Livin protein. Cell Mol Life Sci epub ahead of print.

D'Ursi, A.M., Giusti, L., Albrizio, S., Porchia, F., Esposito, C., Caliendo, G., Gargini, C., Novellino, E., Lucacchini, A., *et al.* (2006). A membrane-permeable peptide containing the last 21 residues of the G alpha(s) carboxyl terminus inhibits G(s)-coupled receptor signaling in intact cells: correlations between peptide structure and biological activity. Mol Pharmacol *69*, 727-736.

Dalton, G.D., Smith, F.L., Smith, P.A., and Dewey, W.L. (2005). Alterations in brain Protein Kinase A activity and reversal of morphine tolerance by two fragments of native Protein Kinase A inhibitor peptide (PKI). Neuropharmacology *48*, 648-657.

Dassa, B., Amitai, G., Caspi, J., Schueler-Furman, O., and Pietrokovski, S. (2007). Trans protein splicing of cyanobacterial split inteins in endogenous and exogenous combinations. Biochemistry *46*, 322-330.

Davies, J. (1994). New pathogens and old resistance genes. Microbiologia *10*, 9-12.

Davis, J.J., Tkac, J., Humphreys, R., Buxton, A.T., Lee, T.A., and Ko Ferrigno, P. (2009). Peptide aptamers in label-free protein detection: 2. Chemical optimization and detection of distinct protein isoforms. Anal Chem *81*, 3314-3320.

Davis, J.J., Tkac, J., Laurenson, S., and Ko Ferrigno, P. (2007). Peptide aptamers in label-free protein detection: 1. Characterization of the immobilized scaffold. Anal Chem *79*, 1089-1096.

Derossi, D., Calvet, S., Trembleau, A., Brunissen, A., Chassaing, G., and Prochiantz, A. (1996). Cell internalization of the third helix of the Antennapedia homeodomain is receptor-independent. J Biol Chem *271*, 18188-18193.

Desplancq, D., King, D.J., Lawson, A.D., and Mountain, A. (1994). Multimerization behaviour of single chain Fv variants for the tumour-binding antibody B72.3. Protein Eng *7*, 1027-1033.

Dickinson, C.D., Veerapandian, B., Dai, X.P., Hamlin, R.C., Xuong, N.H., Ruoslahti, E., and Ely, K.R. (1994). Crystal structure of the tenth type III cell adhesion module of human fibronectin. J Mol Biol *236*, 1079-1092.

Ding, Y., Xu, M.Q., Ghosh, I., Chen, X., Ferrandon, S., Lesage, G., and Rao, Z. (2003). Crystal structure of a mini-intein reveals a conserved catalytic module involved in side chain cyclization of asparagine during protein splicing. J Biol Chem *278*, 39133-39142.

Du, D.Z., Hwang, F.K., Wu, W., and Znati, T. (2006). New construction for transversal design. J Comput Biol *13*, 990-995.

Duerr, D.M., White, S.J., and Schluesener, H.J. (2004). Identification of peptide sequences that induce the transport of phage across the gastrointestinal mucosal barrier. J Virol Methods *116*, 177-180.

Dunn, B., and Wobbe, C.R. (2001). Preparation of protein extracts from yeast. Curr Protoc Mol Biol *Chapter 13*, Unit13.13.

Dymalla, S., Scheffner, M., Weber, E., Sehr, P., Lohrey, C., Hoppe-Seyler, F., and Hoppe-Seyler, K. (2009). A novel peptide motif binding to and blocking the intracellular activity of the human papillomavirus E6 oncoprotein. J Mol Med *87*, 321-331.

Edwards, P.J., and Morrell, A.I. (2002). Solid-phase compound library synthesis in drug design and development. Curr Opin Drug Discov Devel *5*, 594-605.

Elbing, K., and Brent, R. (2002). Media preparation and bacteriological tools. Curr Protoc Mol Biol *Chapter 1*, Unit 1.1.

Engebrecht, J., Brent, R., and Kaderbhai, M.A. (2001). Minipreps of plasmid DNA. Curr Protoc Mol Biol *Chapter 1*, Unit1.6.

Erlanson, D.A., Chytil, M., and Verdine, G.L. (1996). The leucine zipper domain controls the orientation of AP-1 in the NFAT.AP-1.DNA complex. Chem Biol *3*, 981-991.

Escoubas, P., Quinton, L., and Nicholson, G.M. (2008). Venomics: unravelling the complexity of animal venoms with mass spectrometry. J Mass Spectrom *43*, 279-295.

Evans, D., Johnson, S., Laurenson, S., Davies, A.G., Ko Ferrigno, P., and Wälti, C. (2008). Electrical protein detection in cell lysates using high-density peptide-aptamer microarrays. J Biol *7*, 3.

Evans, T.C., Benner, J., and Xu, M.Q. (1998). Semisynthesis of cytotoxic proteins using a modified protein splicing element. Protein Sci *7*, 2256-2264.

Evans, T.C., Benner, J., and Xu, M.Q. (1999). The cyclization and polymerization of bacterially expressed proteins using modified self-splicing inteins. J Biol Chem *274*, 18359-18363.

Evans, T.C., Martin, D., Kolly, R., Panne, D., Sun, L., Ghosh, I., Chen, L., Benner, J., Liu, X.Q., and Xu, M.Q. (2000). Protein trans-splicing and cyclization by a naturally split intein from the dnaE gene of Synechocystis species PCC6803. J Biol Chem *275*, 9091-9094.

Fabbrizio, E., Le Cam, L., Polanowska, J., Kaczorek, M., Lamb, N., Brent, R., and Sardet, C. (1999). Inhibition of mammalian cell proliferation by genetically selected peptide aptamers that functionally antagonize E2F activity. Oncogene *18*, 4357-4363.

Faghihi, M.A., and Wahlestedt, C. (2009). Regulatory roles of natural antisense transcripts. Nat Rev Mol Cell Biol *10*, 637-643.

Fellouse, F.A., Wiesmann, C., and Sidhu, S.S. (2004). Synthetic antibodies from a four-amino-acid code: a dominant role for tyrosine in antigen recognition. Proc Natl Acad Sci U S A *101*, 12467-12472.

Fernández-de-Cossio, J., Gonzalez, L.J., Satomi, Y., Betancourt, L., Ramos, Y., Huerta, V., Besada, V., Padron, G., Minamino, N., and Takao, T. (2004). Automated interpretation of mass spectra of complex mixtures by matching of isotope peak distributions. Rapid Commun Mass Spectrom *18*, 2465-2472.

Fields, S., and Song, O. (1989). A novel genetic system to detect protein-protein interactions. Nature *340*, 245-246.

Finking, R., and Marahiel, M.A. (2004). Biosynthesis of nonribosomal peptides. Annu Rev Microbiol *58*, 453-488.

Finley, R.L., and Brent, R. (1994). Interaction mating reveals binary and ternary connections between Drosophila cell cycle regulators. Proc Natl Acad Sci U S A *91*, 12980-12984.

Fletcher, J.M., and Hughes, R.A. (2009). Modified low molecular weight cyclic peptides as mimetics of BDNF with improved potency, proteolytic stability and transmembrane passage in vitro. Bioorg Med Chem *17*, 2695-2702.

Forrer, P., Stumpp, M.T., Binz, H.K., and Plückthun, A. (2003). A novel strategy to design binding molecules harnessing the modular nature of repeat proteins. FEBS Lett *539*, 2-6.

Forster, A.C., Tan, Z., Nalam, M.N., Lin, H., Qu, H., Cornish, V.W., and Blacklow, S.C. (2003). Programming peptidomimetic syntheses by translating genetic codes designed de novo. Proc Natl Acad Sci U S A *100*, 6353-6357.

Francisco, J.A., Campbell, R., Iverson, B.L., and Georgiou, G. (1993). Production and fluorescence-activated cell sorting of Escherichia coli expressing a functional antibody fragment on the external surface. Proc Natl Acad Sci U S A *90*, 10444-10448.

Frankel, A., Millward, S.W., and Roberts, R.W. (2003). Encodamers: unnatural peptide oligomers encoded in RNA. Chem Biol *10*, 1043-1050.

Friedman, M., Orlova, A., Johansson, E., Eriksson, T.L., Höidén-Guthenberg, I., Tolmachev, V., Nilsson, F.Y., and Ståhl, S. (2008). Directed evolution to low nanomolar affinity of a tumor-targeting epidermal growth factor receptor-binding affibody molecule. J Mol Biol *376*, 1388-1402.

Futaki, S., Nakase, I., Tadokoro, A., Takeuchi, T., and Jones, A.T. (2007). Arginine-rich peptides and their internalization mechanisms. Biochem Soc Trans *35*, 784-787.

Gao, J., Sidhu, S.S., and Wells, J.A. (2009). Two-state selection of conformation-specific antibodies. Proc Natl Acad Sci U S A *106*, 3071-3076.

Gaytán, P., Yañez, J., Sánchez, F., Mackie, H., and Soberón, X. (1998). Combination of DMT-mononucleotide and Fmoc-trinucleotide phosphoramidites in oligonucleotide synthesis affords an automatable codon-level mutagenesis method. Chem Biol *5*, 519-527.

Geda, P., Patury, S., Ma, J., Bharucha, N., Dobry, C.J., Lawson, S.K., Gestwicki, J.E., and Kumar, A. (2008). A small molecule-directed approach to control protein localization and function. Yeast *25*, 577-594.

De Genst, E., Silence, K., Decanniere, K., Conrath, K., Loris, R., Kinne, J., Muyldermans, S., and Wyns, L. (2006). Molecular basis for the preferential cleft recognition by dromedary heavy-chain antibodies. Proc Natl Acad Sci U S A *103*, 4586-4591.

Geyer, C.R., and Brent, R. (2000). Selection of genetic agents from random peptide aptamer expression libraries. Methods Enzymol *328*, 171-208.

Geyer, C.R., Colman-Lerner, A., and Brent, R. (1999). "Mutagenesis" by peptide aptamers identifies genetic network members and pathway connections. Proc Natl Acad Sci U S A *96*, 8567-8572.

Ghadimi, B.M., and Grade, M. (2010). Cancer gene profiling for response prediction. Methods Mol Biol *576*, 327-339.

Giebel, L.B., Cass, R.T., Milligan, D.L., Young, D.C., Arze, R., and Johnson, C.R. (1995). Screening of cyclic peptide phage libraries identifies ligands that bind streptavidin with high affinities. Biochemistry *34*, 15430-15435.

Gietz, R.D., and Schiestl, R.H. (2007). Frozen competent yeast cells that can be transformed with high efficiency using the LiAc/SS carrier DNA/PEG method. Nat Protoc *2*, 1-4.

Gietz, R.D., Schiestl, R.H., Willems, A.R., and Woods, R.A. (1995). Studies on the transformation of intact yeast cells by the LiAc/SS-DNA/PEG procedure. Yeast *11*, 355-360.

Gilch, S., Kehler, C., and Schätzl, H.M. (2007). Peptide aptamers expressed in the secretory pathway interfere with cellular PrPSc formation. J Mol Biol *371*, 362-373.

Giriat, I., and Muir, T.W. (2003). Protein semi-synthesis in living cells. J Am Chem Soc *125*, 7180-7181.

Gogarten, J.P., and Hilario, E. (2006). Inteins, introns, and homing endonucleases: recent revelations about the life cycle of parasitic genetic elements. BMC Evol Biol *6*, 94.

Gogarten, J.P., Senejani, A.G., Zhaxybayeva, O., Olendzenski, L., and Hilario, E. (2002). Inteins: structure, function, and evolution. Annu Rev Microbiol *56*, 263-287.

Golemis, E.A., and Brent, R. (1992). Fused protein domains inhibit DNA binding by LexA. Mol Cell Biol *12*, 3006-3014.

Golemis, E.A., Serebriiskii, I., Finley, R.L., Kolonin, M.G., Gyuris, J., and Brent, R. (2009). Interaction trap/two-hybrid system to identify interacting proteins. Curr Protoc Protein Sci *Chapter 19*, Unit19.2.

Gregoret, L.M., and Sauer, R.T. (1993). Additivity of mutant effects assessed by binomial mutagenesis. Proc Natl Acad Sci U S A *90*, 4246-4250.

Groban, E.S., Johnson, M.B., Banky, P., Burnett, P.G., Calderon, G.L., Dwyer, E.C., Fuller, S.N., Gebre, B., King, L.M., *et al.* (2005). Binding of the Bacillus subtilis LexA protein to the SOS operator. Nucleic Acids Res *33*, 6287-6295.

Grönwall, C., Jonsson, A., Lindström, S., Gunneriusson, E., Ståhl, S., and Herne, N. (2007). Selection and characterization of Affibody ligands binding to Alzheimer amyloid beta peptides. J Biotechnol *128*, 162-183.

Grönwall, C., Snelders, E., Palm, A.J., Eriksson, F., Herne, N., and Ståhl, S. (2008). Generation of Affibody ligands binding interleukin-2 receptor alpha/CD25. Biotechnol Appl Biochem *50*, 97-112.

Grünewald, J., and Marahiel, M.A. (2006). Chemoenzymatic and template-directed synthesis of bioactive macrocyclic peptides. Microbiol Mol Biol Rev *70*, 121-146.

Grzyb, J., Latowski, D., and Strzałka, K. (2006). Lipocalins - a family portrait. J Plant Physiol *163*, 895-915.

Guerin, E., Cambray, G., Sanchez-Alberola, N., Campoy, S., Erill, I., Da Re, S., Gonzalez-Zorn, B., Barbé, J., Ploy, M.C., and Mazel, D. (2009). The SOS response controls integron recombination. Science *324*, 1034.

Gyuris, J., Golemis, E., Chertkov, H., and Brent, R. (1993). Cdi1, a human G1 and S phase protein phosphatase that associates with Cdk2. Cell *75*, 791-803.

Hackel, B.J., Kapila, A., and Wittrup, K.D. (2008). Picomolar affinity fibronectin domains engineered utilizing loop length diversity, recursive mutagenesis, and loop shuffling. J Mol Biol *381*, 1238-1252.

Hagelin, G. (2005). Mass spectrometric investigation of Maltacines E1a and E1b--two members of the Maltacine family of peptide antibiotics. Rapid Commun Mass Spectrom *19*, 3633-3642.

Han, K., Jeon, M.J., Kim, S.H., Ki, D., Bahn, J.H., Lee, K.S., Park, J., and Choi, S.Y. (2001). Efficient intracellular delivery of an exogenous protein GFP with genetically fused basic oligopeptides. Mol Cells *12*, 267-271.

Hanes, J., and Plückthun, A. (1997). In vitro selection and evolution of functional proteins by using ribosome display. Proc Natl Acad Sci U S A *94*, 4937-4942.

Hällbrink, M., Florén, A., Elmquist, A., Pooga, M., Bartfai, T., and Langel, U. (2001). Cargo delivery kinetics of cell-penetrating peptides. Biochim Biophys Acta *1515*, 101-109.

Heilig, J.S., Elbing, K.L., and Brent, R. (2001). Large-scale preparation of plasmid DNA. Curr Protoc Mol Biol *Chapter 1*, Unit1.7.

Heinicke, L.A., Wong, C.J., Lary, J., Nallagatla, S.R., Diegelman-Parente, A., Zheng, X., Cole, J.L., and Bevilacqua, P.C. (2009). RNA dimerization promotes PKR dimerization and activation. J Mol Biol *390*, 319-338.

Heitz, F., Morris, M.C., and Divita, G. (2009). Twenty years of cell-penetrating peptides: from molecular mechanisms to therapeutics. Br J Pharmacol *157*, 195-206.

Hiraga, K., Soga, I., Dansereau, J.T., Pereira, B., Derbyshire, V., Du, Z., Wang, C., Van Roey, P., Belfort, G., and Belfort, M. (2009). Selection and structure of hyperactive inteins: peripheral changes relayed to the catalytic center. J Mol Biol *393*, 1106-1117.

Hirst, M., Ho, C., Sabourin, L., Rudnicki, M., Penn, L., and Sadowski, I. (2001). A two-hybrid system for transactivator bait proteins. Proc Natl Acad Sci U S A *98*, 8726-8731.

Hitoshi, Y., Gururaja, T., Pearsall, D.M., Lang, W., Sharma, P., Huang, B., Catalano, S.M., McLaughlin, J., Pali, E., *et al.* (2003). Cellular localization and antiproliferative effect of peptides discovered from a functional screen of a retrovirally delivered random peptide library. Chem Biol *10*, 975-987.

Hoffman, C.S., and Winston, F. (1987). A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformation of Escherichia coli. Gene *57*, 267-272.

Hornung, V., Guenthner-Biller, M., Bourquin, C., Ablasser, A., Schlee, M., Uematsu, S., Noronha, A., Manoharan, M., Akira, S., *et al.* (2005). Sequence-specific potent induction of IFN-alpha by short interfering RNA in plasmacytoid dendritic cells through TLR7. Nat Med *11*, 263-270.

Horswill, A.R., and Benkovic, S.J. (2005). Cyclic peptides, a chemical genetics tool for biologists. Cell Cycle *4*, 552-555.

Horswill, A.R., and Benkovic, S.J. (2006). Identifying small-molecule modulators of protein-protein interactions. Curr Protoc Protein Sci *Chapter 19*, Unit 19.15.

Horswill, A.R., Savinov, S.N., and Benkovic, S.J. (2004). A systematic method for identifying small-molecule modulators of protein-protein interactions. Proc Natl Acad Sci U S A *101*, 15591-15596.

Hortin, G.L., and Murthy, J. (2002). Substrate size selectivity of 20S proteasomes: analysis with variable-sized synthetic substrates. J Protein Chem *21*, 333-337.

Hosse, R.J., Rothe, A., and Power, B.E. (2006). A new generation of protein display scaffolds for molecular recognition. Protein Sci *15*, 14-27.

Hruby, V.J., Qui, W., Okayama, T., and Soloshonok, V.A. (2002). Design of nonpeptides from peptide ligands for peptide receptors. Methods Enzymol *343*, 91-123.

Hua, S.B., Qiu, M., Chan, E., Zhu, L., and Luo, Y. (1997). Minimum length of sequence homology required for in vivo cloning by homologous recombination in yeast. Plasmid *38*, 91-96.

Huang, J., and Schreiber, S.L. (1997). A yeast genetic system for selecting small molecule inhibitors of protein-protein interactions in nanodroplets. Proc Natl Acad Sci U S A *94*, 13396-13401.

Huang, J., Koide, A., Nettle, K.W., Greene, G.L., and Koide, S. (2006). Conformation-specific affinity purification of proteins using engineered binding proteins: application to the estrogen receptor. Protein Expr Purif *47*, 348-354.

Huston, J.S., Levinson, D., Mudgett-Hunter, M., Tai, M.S., Novotný, J., Margolies, M.N., Ridge, R.J., Bruccoleri, R.E., Haber, E., and Crea, R. (1988). Protein engineering of antibody binding sites: recovery of specific activity in an anti-digoxin single-chain Fv analogue produced in Escherichia coli. Proc Natl Acad Sci U S A *85*, 5879-5883.

Huwe, C.M. (2006). Synthetic library design. Drug Discov Today *11*, 763-767.

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol *2*, 13-34.

Ioannides, C.G., Freedman, R.S., Liskamp, R.M., Ward, N.E., and O'Brian, C.A. (1990). Inhibition of IL-2 receptor induction and IL-2 production in the human leukemic cell line Jurkat by a novel peptide inhibitor of protein kinase C. Cell Immunol *131*, 242-252.

Iwai, H., and Plückthun, A. (1999). Circular beta-lactamase: stability enhancement by cyclizing the backbone. FEBS Lett *459*, 166-172.

Iwai, H., Lingel, A., and Pluckthun, A. (2001). Cyclic green fluorescent protein produced in vivo using an artificially split PI-PfuI intein from Pyrococcus furiosus. J Biol Chem *276*, 16548-16554.

Iwai, H., Züger, S., Jin, J., and Tam, P.H. (2006). Highly efficient protein trans-splicing by a naturally split DnaE intein from Nostoc punctiforme. FEBS Lett *580*, 1853-1858.

Jackson, A.L., Bartz, S.R., Schelter, J., Kobayashi, S.V., Burchard, J., Mao, M., Li, B., Cavet, G., and Linsley, P.S. (2003). Expression profiling reveals off-target gene regulation by RNAi. Nat Biotechnol *21*, 635-637.

Jacobs, M.D., and Fox, R.O. (1994). Staphylococcal nuclease folding intermediate characterized by hydrogen exchange and NMR spectroscopy. Proc Natl Acad Sci U S A *91*, 449-453.

Jain, R.K. (1990). Physiological barriers to delivery of monoclonal antibodies and other macromolecules in tumors. Cancer Res *50*, 814s-819s.

Jennings, C., West, J., Waine, C., Craik, D., and Anderson, M. (2001). Biosynthesis and insecticidal properties of plant cyclotides: the cyclic knotted proteins from Oldenlandia affinis. Proc Natl Acad Sci U S A *98*, 10614-10619.

Jermutus, L., Ryabova, L.A., and Plückthun, A. (1998). Recent advances in producing and selecting functional proteins by using cell-free translation. Curr Opin Biotechnol *9*, 534-548.

Jin, F., Hazbun, T., Michaud, G.A., Salcius, M., Predki, P.F., Fields, S., and Huang, J. (2006). A pooling-deconvolution strategy for biological network elucidation. Nat Methods *3*, 183-189.

Jin, H., Retallack, D.M., Stelman, S.J., Hershberger, C.D., and Ramseier, T. (2007). Characterization of the SOS response of Pseudomonas fluorescens strain DC206 using whole-genome transcript analysis. FEMS Microbiol Lett *269*, 256-264.

Johnson, S., Evans, D., Laurenson, S., Paul, D., Davies, A.G., Ferrigno, P.K., and Wälti, C. (2008). Surface-immobilized peptide aptamers as probe molecules for protein detection. Anal Chem *80*, 978-983.

Jones, P.T., Dear, P.H., Foote, J., Neuberger, M.S., and Winter, G. (1986). Replacing the complementarity-determining regions in a human antibody with those from a mouse. Nature *321*, 522-525.

Joshi, P.B., Hirst, M., Malcolm, T., Parent, J., Mitchell, D., Lund, K., and Sadowski, I. (2007). Identification of protein interaction antagonists using the repressed transactivator two-hybrid system. Biotechniques *42*, 635-644.

Joung, J.K., Ramm, E.I., and Pabo, C.O. (2000). A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions. Proc Natl Acad Sci U S A *97*, 7382-7387.

Judge, A.D., Sood, V., Shaw, J.R., Fang, D., McClintock, K., and MacLachlan, I. (2005). Sequence-dependent stimulation of the mammalian innate immune response by synthetic siRNA. Nat Biotechnol *23*, 457-462.

Kaiser, N., Lischka, P., Wagenknecht, N., and Stamminger, T. (2009). Inhibition of human cytomegalovirus replication via peptide aptamers directed against the nonconventional nuclear localization signal of the essential viral replication factor pUL84. J Virol *83*, 11902-11913.

Kalderon, D., Roberts, B.L., Richardson, W.D., and Smith, A.E. (1984). A short amino acid sequence able to specify nuclear location. Cell *39*, 499-509.

Kamb, A., and Teng, D.H. (2000). Transdominant genetics, peptide inhibitors and drug targets. Curr Opin Mol Ther *2*, 662-669.

Karatan, E., Merguerian, M., Han, Z., Scholle, M.D., Koide, S., and Kay, B.K. (2004). Molecular recognition properties of FN3 monobodies that bind the Src SH3 domain. Chem Biol *11*, 835-844.

Kato-Stankiewicz, J., Hakimi, I., Zhi, G., Zhang, J., Serebriiskii, I., Guo, L., Edamatsu, H., Koide, H., Menon, S., *et al.* (2002). Inhibitors of Ras/Raf-1 interaction identified by two-hybrid screening revert Ras-dependent transformation phenotypes in human cancer cells. Proc Natl Acad Sci U S A *99*, 14398-14403.

Kawakami, T., Murakami, H., and Suga, H. (2008). Messenger RNA-programmed incorporation of multiple N-methyl-amino acids into linear and cyclic peptides. Chem Biol *15*, 32-42.

Kawasaki, M., Nogami, S., Satow, Y., Ohya, Y., and Anraku, Y. (1997). Identification of three core regions essential for protein splicing of the yeast Vma1 protozyme. A random mutagenesis study of the entire Vma1-derived endonuclease sequence. J Biol Chem *272*, 15668-15674.

Kawe, M., Forrer, P., Amstutz, P., and Plückthun, A. (2006). Isolation of intracellular proteinase inhibitors derived from designed ankyrin repeat proteins by genetic screening. J Biol Chem *281*, 40252-40263.

Kayushin, A., Korosteleva, M., and Miroshnikov, A. (2000). Large-scale solid-phase preparation of 3'-unprotected trinucleotide phosphotriesters--precursors for synthesis of trinucleotide phosphoramidites. Nucleosides Nucleotides Nucleic Acids *19*, 1967-1976.

Kayushin, A.L., Korosteleva, M.D., Miroshnikov, A.I., Kosch, W., Zubov, D., and Piel, N. (1996). A convenient approach to the synthesis of trinucleotide phosphoramidites--synthons for the generation of oligonucleotide/peptide libraries. Nucleic Acids Res *24*, 3748-3755.

Khazaeli, M.B., Conry, R.M., and LoBuglio, A.F. (1994). Human immune response to monoclonal antibodies. J Immunother Emphasis Tumor Immunol *15*, 42-52.

Kim, Y.J., Neelamegam, R., Heo, M.A., Edwardraja, S., Paik, H.J., and Lee, S.G. (2008). Improving the productivity of single-chain Fv antibody against c-Met by rearranging the order of its variable domains. J Microbiol Biotechnol *18*, 1186-1190.

Kinsella, T.M., Ohashi, C.T., Harder, A.G., Yam, G.C., Li, W., Peelle, B., Pali, E.S., Bennett, M.K., Molineaux, S.M., *et al.* (2002). Retrovirally delivered random cyclic Peptide libraries yield inhibitors of interleukin-4 signaling in human B cells. J Biol Chem *277*, 37512-37518.

Kleinkauf, H., and Von Döhren, H. (1996). A nonribosomal system of peptide biosynthesis. Eur J Biochem *236*, 335-351.

Klevenz, B., Butz, K., and Hoppe-Seyler, F. (2002). Peptide aptamers: exchange of the thioredoxin-A scaffold by alternative platform proteins and its influence on target protein binding. Cell Mol Life Sci *59*, 1993-1998.

Kley, N. (2004). Chemical dimerizers and three-hybrid systems: scanning the proteome for targets of organic small molecules. Chem Biol *11*, 599-608.

Koide, A., Abbatiello, S., Rothgery, L., and Koide, S. (2002). Probing protein conformational changes in living cells by using designer binding proteins: application to the estrogen receptor. Proc Natl Acad Sci U S A *99*, 1253-1258.

Koide, A., Bailey, C.W., Huang, X., and Koide, S. (1998). The fibronectin type III domain as a scaffold for novel binding proteins. J Mol Biol *284*, 1141-1151.

Koide, S., and Sidhu, S.S. (2009). The importance of being tyrosine: lessons in molecular recognition from minimalist synthetic binding proteins. ACS Chem Biol *4*, 325-334.

Kolonin, M.G., and Finley, R.L. (1998). Targeting cyclin-dependent kinases in Drosophila with peptide aptamers. Proc Natl Acad Sci U S A *95*, 14266-14271.

Kolonin, M.G., Zhong, J., and Finley, R.L. (2000). Interaction mating methods in two-hybrid systems. Methods Enzymol *328*, 26-46.

Kritzer, J.A., Hamamichi, S., McCaffery, J.M., Santagata, S., Naumann, T.A., Caldwell, K.A., Caldwell, G.A., and Lindquist, S. (2009). Rapid selection of cyclic peptides that reduce alpha-synuclein toxicity in yeast and animal models. Nat Chem Biol *5*, 655-663.

Kunz, C., Borghouts, C., Buerger, C., and Groner, B. (2006). Peptide aptamers with binding specificity for the intracellular domain of the ErbB2 receptor interfere with AKT signaling and sensitize breast cancer cells to Taxol. Mol Cancer Res *4*, 983-998.

Kurtz, S.E., Esposito, K., Tang, W., and Menzel, R. (2003). Inhibition of an activated Ras protein with genetically selected peptide aptamers. Biotechnol Bioeng *82*, 38-46.

Kushnirov, V.V. (2000). Rapid and reliable protein extraction from yeast. Yeast *16*, 857-860.

Kyte, J., and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. J Mol Biol *157*, 105-132.

Ladner, R.C. (1995). Constrained peptides as binding entities. Trends Biotechnol *13*, 426-430.

Laemmli, U.K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature *227*, 680-685.

Laffly, E., and Sodoyer, R. (2005). Monoclonal and recombinant antibodies, 30 years after . Hum Antibodies *14*, 33-55.

Lauwereys, M., Arbabi Ghahroudi, M., Desmyter, A., Kinne, J., Hölzer, W., De Genst, E., Wyns, L., and Muyldermans, S. (1998). Potent enzyme inhibitors derived from dromedary heavy-chain antibodies. EMBO J *17*, 3512-3520.

LaVallie, E.R., DiBlasio, E.A., Kovacic, S., Grant, K.L., Schendel, P.F., and McCoy, J.M. (1993). A thioredoxin gene fusion expression system that circumvents inclusion body formation in the E. coli cytoplasm. Biotechnology (N Y) *11*, 187-193.

Leahy, D.J., Hendrickson, W.A., Aukhil, I., and Erickson, H.P. (1992). Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein. Science *258*, 987-991.

Li, P., and Roller, P.P. (2002). Cyclization strategies in peptide derived drug design. Curr Top Med Chem *2*, 325-341.

Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D., Chesneau, A., *et al.* (2004). A map of the interactome network of the metazoan C. elegans. Science *303*, 540-543.

Liao, H.I., Olson, C.A., Hwang, S., Deng, H., Wong, E., Baric, R.S., Roberts, R.W., and Sun, R. (2009). mRNA display design of fibronectin-based intrabodies that detect and inhibit severe acute respiratory syndrome coronavirus nucleocapsid protein. J Biol Chem *284*, 17512-17520.

Licitra, E.J., and Liu, J.O. (1996). A three-hybrid system for detecting small ligand-protein receptor interactions. Proc Natl Acad Sci U S A *93*, 12817-12821.

Lin, L.L., and Little, J.W. (1988). Isolation and characterization of noncleavable (Ind-) mutants of the LexA repressor of Escherichia coli K-12. J Bacteriol *170*, 2163-2173.

Lobato, M.N., and Rabbitts, T.H. (2004). Intracellular antibodies as specific reagents for functional ablation: future therapeutic molecules. Curr Mol Med *4*, 519-528.

Lopez-Ochoa, L., Ramirez-Prado, J., and Hanley-Bowdoin, L. (2006). Peptide aptamers that bind to a geminivirus replication protein interfere with viral replication in plant cells. J Virol *80*, 5841-5853.

Lu, Z., Murray, K.S., Van Cleave, V., LaVallie, E.R., Stahl, M.L., and McCoy, J.M. (1995). Expression of thioredoxin random peptide libraries on the Escherichia coli cell surface as functional fusions to flagellin: a system designed for exploring protein-protein interactions. Biotechnology (N Y) *13*, 366-372.

Ludwig, C., Schwarzer, D., Zettler, J., Garbe, D., Janning, P., Czeslik, C., and Mootz, H.D. (2009). Semisynthesis of proteins using split inteins. Methods Enzymol *462*, 77-96.

Luo, Y., Pfuetzner, R.A., Mosimann, S., Paetzel, M., Frey, E.A., Cherney, M., Kim, B., Little, J.W., and Strynadka, N.C. (2001). Crystal structure of LexA: a conformational switch for regulation of self-cleavage. Cell *106*, 585-594.

Ma, H., Kunes, S., Schatz, P.J., and Botstein, D. (1987). Plasmid construction by homologous recombination in yeast. Gene *58*, 201-216.

Main, A.L., Harvey, T.S., Baron, M., Boyd, J., and Campbell, I.D. (1992). The three-dimensional structure of the tenth type III module of fibronectin: an insight into RGD-mediated interactions. Cell *71*, 671-678.

Makhnevych, T., Sydorskyy, Y., Xin, X., Srikumar, T., Vizeacoumar, F.J., Jeram, S.M., Li, Z., Bahr, S., Andrews, B.J., *et al.* (2009). Global map of SUMO function revealed by protein-protein interaction and genetic networks. Mol Cell *33*, 124-135.

Martel, V., Filhol, O., Colas, P., and Cochet, C. (2006). p53-dependent inhibition of mammalian cell survival by a genetically selected peptide aptamer that targets the regulatory subunit of protein kinase CK2. Oncogene *25*, 7343-7353.

Mathys, S., Evans, T.C., Chute, I.C., Wu, H., Chong, S., Benner, J., Liu, X.Q., and Xu, M.Q. (1999). Characterization of a self-splicing mini-intein and its conversion into autocatalytic N- and C-terminal cleavage elements: facile production of protein building blocks for protein ligation. Gene *231*, 1-13.

Maupetit, J., Derreumaux, P., and Tuffery, P. (2009). PEP-FOLD: an online resource for de novo peptide structure prediction. Nucleic Acids Res *37*, W498-W503.

Mauriala, T., Auriola, S., Azhayev, A., Kayushin, A., Korosteleva, M., and Miroshnikov, A. (2004). HPLC electrospray mass spectrometric characterization of trimeric building blocks for oligonucleotide synthesis. J Pharm Biomed Anal *34*, 199-206.

Mern, D.S., Hoppe-Seyler, K., Hoppe-Seyler, F., Hasskarl, J., and Burwinkel, B. (2010). Targeting Id1 and Id3 by a specific peptide aptamer induces E-box promoter activity, cell cycle arrest, and apoptosis in breast cancer cells. Breast Cancer Res Treat epub ahead of print.

McLafferty, M.A., Kent, R.B., Ladner, R.C., and Markland, W. (1993). M13 bacteriophage displaying disulfide-constrained microproteins. Gene *128*, 29-36.

McLean, M.A., Rajfur, Z., Chen, Z., Humphrey, D., Yang, B., Sligar, S.G., and Jacobson, K. (2009). Mechanism of Chromophore Assisted Laser Inactivation Employing Fluorescent Proteins. Anal Chem *81*, 1755-1761.

Mechold, U., Gilbert, C., and Ogryzko, V. (2005). Codon optimization of the BirA enzyme gene leads to higher expression and an improved efficiency of biotinylation of target proteins in mammalian cells. J Biotechnol *116*, 245-249.

Meyer, S.C., Gaj, T., and Ghosh, I. (2006). Highly selective cyclic peptide ligands for NeutrAvidin and avidin identified by phage display. Chem Biol Drug Des *68*, 3-10.

Miesel, L., Greene, J., and Black, T.A. (2003). Genetic strategies for antibacterial drug discovery. Nat Rev Genet *4*, 442-456.

Miller, C., Thomsen, L.E., Gaggero, C., Mosseri, R., Ingmer, H., and Cohen, S.N. (2004). SOS response induction by beta-lactams and bacterial defense against antibiotic lethality. Science *305*, 1629-1631.

Millward, S.W., Fiacco, S., Austin, R.J., and Roberts, R.W. (2007). Design of cyclic peptides that bind protein surfaces with antibody-like affinity. ACS Chem Biol *2*, 625-634.

Millward, S.W., Takahashi, T.T., and Roberts, R.W. (2005). A general route for post-translational cyclization of mRNA display libraries. J Am Chem Soc *127*, 14142-14143.

Moore, D., and Dowhan, D. (2002). Purification and concentration of DNA from aqueous solutions. Curr Protoc Mol Biol *Chapter 2*, Unit 2.1A.

Morrison, K.L., and Weiss, G.A. (2001). Combinatorial alanine-scanning. Curr Opin Chem Biol *5*, 302-307.

Morrison, S.L., Johnson, M.J., Herzenberg, L.A., and Oi, V.T. (1984). Chimeric human antibody molecules: mouse antigen-binding domains with human constant region domains. Proc Natl Acad Sci U S A *81*, 6851-6855.

Muona, M., Aranko, A.S., and Iwai, H. (2008). Segmental isotopic labelling of a multidomain protein by protein ligation by protein trans-splicing. Chembiochem *9*, 2958-2961.

Nagel-Wolfrum, K., Buerger, C., Wittig, I., Butz, K., Hoppe-Seyler, F., and Groner, B. (2004). The interaction of specific peptide aptamers with the DNA binding domain and the dimerization domain of the transcription factor Stat3 inhibits transactivation and induces apoptosis in tumor cells. Mol Cancer Res *2*, 170-182.

Nakajima, E., Goto, Y., Sako, Y., Murakami, H., and Suga, H. (2009). Ribosomal synthesis of peptides with C-terminal lactams, thiolactones, and alkylamides. Chembiochem *10*, 1186-1192.

Nakamura, Y., Gojobori, T., and Ikemura, T. (2000). Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucleic Acids Res *28*, 292.

Nam, C.H., Lobato, M.N., Appert, A., Drynan, L.F., Tanaka, T., and Rabbitts, T.H. (2008). An antibody inhibitor of the LMO2-protein complex blocks its normal and tumorigenic functions. Oncogene *27*, 4962-4968.

Nauenburg, S., Zwerschke, W., and Jansen-Durr, P. (2001). Induction of apoptosis in cervical carcinoma cells by peptide aptamers that bind to the HPV-16 E7 oncoprotein. FASEB J *15*, 592-594.

Naumann, T.A., Savinov, S.N., and Benkovic, S.J. (2005). Engineering an affinity tag for genetically encoded cyclic peptides. Biotechnol Bioeng *92*, 820-830.

Naumann, T.A., Tavassoli, A., and Benkovic, S.J. (2008). Genetic selection of cyclic peptide Dam methyltransferase inhibitors. Chembiochem *9*, 194-197.

Nelson, A.R., Borland, L., Allbritton, N.L., and Sims, C.E. (2007). Myristoyl-based transport of peptides into living cells. Biochemistry *46*, 14771-14781.

Nieba, L., Nieba-Axmann, S.E., Persson, A., Hämäläinen, M., Edebratt, F., Hansson, A., Lidholm, J., Magnusson, K., Karlsson, A.F., and Plückthun, A. (1997). BIACORE

analysis of histidine-tagged proteins using a chelating NTA sensor chip. Anal Biochem *252*, 217-228.

Nilsson, L.O., Louassini, M., and Abel-Santos, E. (2005). Using siclopps for the discovery of novel antimicrobial peptides and their targets. Protein Pept Lett *12*, 795-799.

Nord, K., Gunneriusson, E., Ringdahl, J., Ståhl, S., Uhlén, M., and Nygren, P.A. (1997). Binding proteins selected from combinatorial libraries of an alpha-helical bacterial receptor domain. Nat Biotechnol *15*, 772-777.

Noren, C.J., Anthony-Cahill, S.J., Griffith, M.C., and Schultz, P.G. (1989). A general method for site-specific incorporation of unnatural amino acids into proteins. Science *244*, 182-188.

Norman, T.C., Smith, D.L., Sorger, P.K., Drees, B.L., O'Rourke, S.M., Hughes, T.R., Roberts, C.J., Friend, S.H., Fields, S., and Murray, A.W. (1999). Genetic selection of peptide inhibitors of biological pathways. Science *285*, 591-595.

Nouvion, A.L., Thibaut, J., Lohez, O.D., Venet, S., Colas, P., Gillet, G., and Lalle, P. (2007). Modulation of Nr-13 antideath activity by peptide aptamers. Oncogene *26*, 701-710.

Olson, C.A., Liao, H.I., Sun, R., and Roberts, R.W. (2008). mRNA display selection of a high-affinity, modification-specific phospho-IkappaBalpha-binding fibronectin. ACS Chem Biol *3*, 480-485.

Ozawa, T., Nogami, S., Sato, M., Ohya, Y., and Umezawa, Y. (2000). A fluorescent indicator for detecting protein-protein interactions in vivo based on protein splicing. Anal Chem *72*, 5151-5157.

Pamonsinlapatham, P., Hadj-Slimane, R., Raynaud, F., Bickle, M., Corneloup, C., Barthelaix, A., Lepelletier, Y., Mercier, P., Schapira, M., *et al.* (2008). A RasGAP SH3 peptide aptamer inhibits RasGAP-Aurora interaction and induces caspase-independent tumor cell death. PLoS ONE *3*, e2902.

Pasqualini, R., and Ruoslahti, E. (1996). Organ targeting in vivo using phage display peptide libraries. Nature *380*, 364-366.

Patrick, W.M., Firth, A.E., and Blackburn, J.M. (2003). User-friendly algorithms for estimating completeness and diversity in randomized protein-encoding libraries. Protein Eng *16*, 451-457.

Pawson, T., and Nash, P. (2000). Protein-protein interactions define specificity in signal transduction. Genes Dev *14*, 1027-1047.

Peelle, B., Gururaja, T.L., Payan, D.G., and Anderson, D.C. (2001a). Characterization and use of green fluorescent proteins from Renilla mulleri and Ptilosarcus guernyi for the human cell display of functional peptides. J Protein Chem *20*, 507-519.

Peelle, B., Lorens, J., Li, W., Bogenberger, J., Payan, D.G., and Anderson, D.C. (2001b). Intracellular protein scaffold-mediated display of random peptide libraries for phenotypic screens in mammalian cells. Chem Biol *8*, 521-534.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A *96*, 4285-4288.

Perler, F.B. (2002). InBase: the Intein Database. Nucleic Acids Res *30*, 383-384.

Pietrokovski, S. (1994). Conserved sequence features of inteins (protein introns) and their use in identifying new inteins and related proteins. Protein Sci *3*, 2340-2350.

Pietrokovski, S. (1998). Modular organization of inteins and C-terminal autocatalytic domains. Protein Sci *7*, 64-71.

Poland, B.W., Xu, M.Q., and Quiocho, F.A. (2000). Structural insights into the protein splicing mechanism of PI-SceI. J Biol Chem *275*, 16408-16413.

Presta, L.G. (2006). Engineering of therapeutic antibodies to minimize immunogenicity and optimize function. Adv Drug Deliv Rev *58*, 640-656.

Reichert, J.M. (2008). Monoclonal antibodies as innovative therapeutics. Curr Pharm Biotechnol *9*, 423-430.

Reiter, Y., Brinkmann, U., Kreitman, R.J., Jung, S.H., Lee, B., and Pastan, I. (1994). Stabilization of the Fv fragments in recombinant immunotoxins by disulfide bonds engineered into conserved framework regions. Biochemistry *33*, 5451-5459.

Remy, I., and Michnick, S.W. (1999). Clonal selection and in vivo quantitation of protein interactions with protein-fragment complementation assays. Proc Natl Acad Sci U S A *96*, 5394-5399.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., *et al.* (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311-1323.

Roberts, R.W., and Szostak, J.W. (1997). RNA-peptide fusions for the in vitro selection of peptides and proteins. Proc Natl Acad Sci U S A *94*, 12297-12302.

Robinson, C.R., and Sauer, R.T. (1998). Optimizing the stability of single-chain proteins by linker length and composition mutagenesis. Proc Natl Acad Sci U S A *95*, 5929-5934.

Roland, K.L., Smith, M.H., Rupley, J.A., and Little, J.W. (1992). In vitro analysis of mutant LexA proteins with an increased rate of specific cleavage. J Mol Biol *228*, 395-408.

Rosengren, K.J., Clark, R.J., Daly, N.L., Göransson, U., Jones, A., and Craik, D.J. (2003). Microcin J25 has a threaded sidechain-to-backbone ring structure and not a head-to-tail cyclized backbone. J Am Chem Soc *125*, 12464-12474.

Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., *et al.* (2005). Towards a proteome-scale map of the human protein-protein interaction network. Nature *437*, 1173-1178.

Rubinstein, M., and Niv, M.Y. (2009). Peptidic modulators of protein-protein interactions: progress and challenges in computational design. Biopolymers *91*, 505-513.

Ruoslahti, E. (2000). Targeting tumor vasculature with homing peptides from phage display. Semin Cancer Biol *10*, 435-442.

Saccucci, L., Crance, J.M., Colas, P., Bickle, M., Garin, D., and Iseni, F. (2009). Inhibition of vaccinia virus replication by peptide aptamers. Antiviral Res *82*, 134-140.

Saether, O., Craik, D.J., Campbell, I.D., Sletten, K., Juul, J., and Norman, D.G. (1995). Elucidation of the primary and three-dimensional structure of the uterotonic polypeptide kalata B1. Biochemistry *34*, 4147-4158.

Saska, I., Gillon, A.D., Hatsugai, N., Dietzgen, R.G., Hara-Nishimura, I., Anderson, M.A., and Craik, D.J. (2007). An asparaginyl endopeptidase mediates in vivo protein backbone cyclization. J Biol Chem *282*, 29721-29728.

Sassanfar, M., and Roberts, J.W. (1990). Nature of the SOS-inducing signal in Escherichia coli. The involvement of DNA replication. J Mol Biol *212*, 79-96.

Sastalla, I., Chim, K., Cheung, G.Y., Pomerantsev, A.P., and Leppla, S.H. (2009). Codon-optimized fluorescent proteins designed for expression in low GC Gram-positive bacteria. Appl Environ Microbiol *75*, 2099-2110.

Scacheri, P.C., Rozenblatt-Rosen, O., Caplen, N.J., Wolfsberg, T.G., Umayam, L., Lee, J.C., Hughes, C.M., Shanmugam, K.S., Bhattacharjee, A., *et al.* (2004). Short interfering RNAs can induce unexpected and divergent changes in the levels of untargeted proteins in mammalian cells. Proc Natl Acad Sci U S A *101*, 1892-1897.

Schaedel, O., and Reiter, Y. (2006). Antibodies and their fragments as anti-cancer agents. Curr Pharm Des *12*, 363-378.

Schiestl, R.H., and Gietz, R.D. (1989). High efficiency transformation of intact yeast cells using single stranded nucleic acids as a carrier. Curr Genet *16*, 339-346.

Schlehuber, S., and Skerra, A. (2005). Anticalins in drug development. BioDrugs *19*, 279-288.

Schmidt, S., Diriong, S., Méry, J., Fabbrizio, E., and Debant, A. (2002). Identification of the first Rho-GEF inhibitor, TRIPalpha, which targets the RhoA-specific GEF domain of Trio. FEBS Lett *523*, 35-42.

Schwartz, A.S., Yu, J., Gardenour, K.R., Finley, R.L., and Ideker, T. (2009). Cost-effective strategies for completing the interactome. Nat Methods *6*, 55-61.

Schweizer, A., Rusert, P., Berlinger, L., Ruprecht, C.R., Mann, A., Corthésy, S., Turville, S.G., Aravantinou, M., Fischer, M., *et al.* (2008). CD4-specific designed ankyrin repeat proteins are novel potent HIV entry inhibitors with unique characteristics. PLoS Pathog *4*, e1000109.

Scott, C.P., Abel-Santos, E., Jones, A.D., and Benkovic, S.J. (2001). Structural requirements for the biosynthesis of backbone cyclic peptide libraries. Chem Biol *8*, 801-815.

Scott, C.P., Abel-Santos, E., Wall, M., Wahnon, D.C., and Benkovic, S.J. (1999). Production of cyclic peptides and proteins in vivo. Proc Natl Acad Sci U S A *96*, 13638-13643.

Sedgwick, S.G., and Smerdon, S.J. (1999). The ankyrin repeat: a diversity of interactions on a common structural framework. Trends Biochem Sci *24*, 311-316.

Seidman, C.E., Struhl, K., Sheen, J., and Jessen, T. (2001). Introduction of plasmid DNA into cells. Curr Protoc Mol Biol *Chapter 1*, Unit1.8.

Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. Nature *455*, 58-63.

SenGupta, D.J., Zhang, B., Kraemer, B., Pochart, P., Fields, S., and Wickens, M. (1996). A three-hybrid system to detect RNA-protein interactions in vivo. Proc Natl Acad Sci U S A *93*, 8496-8501.

Sewald, N., and Jakubke, H. (2002). Peptides: chemistry and biology. Weinheim: Wiley-VCH.

Shaki-Loewenstein, S., Zfania, R., Hyland, S., Wels, W.S., and Benhar, I. (2005). A universal strategy for stable intracellular antibodies. J Immunol Methods *303*, 19-39.

Shiau, A.K., Massari, M.E., and Ozbal, C.C. (2008). Back to basics: label-free technologies for small molecule screening. Comb Chem High Throughput Screen *11*, 231-237.

Sidhu, S.S., and Koide, S. (2007). Phage display for engineering and analyzing protein interaction interfaces. Curr Opin Struct Biol *17*, 481-487.

Skerra, A. (2007). Alternative non-antibody scaffolds for molecular recognition. Curr Opin Biotechnol *18*, 295-304.

Skerra, A. (2008). Alternative binding proteins: anticalins - harnessing the structural plasticity of the lipocalin ligand pocket to engineer novel binding activities. FEBS J *275*, 2677-2683.

Slilaty, S.N., and Little, J.W. (1987). Lysine-156 and serine-119 are required for LexA repressor cleavage: a possible mechanism. Proc Natl Acad Sci U S A *84*, 3987-3991.

Smith, G.P. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. Science *228*, 1315-1317.

Sondek, J., and Shortle, D. (1992). A general strategy for random insertion and substitution mutagenesis: substoichiometric coupling of trinucleotide phosphoramidites. Proc Natl Acad Sci U S A *89*, 3581-3585.

Southworth, M.W., Amaya, K., Evans, T.C., Xu, M.Q., and Perler, F.B. (1999). Purification of proteins fused to either the amino or carboxy terminus of the Mycobacterium xenopi gyrase A intein. Biotechniques *27*, 110-120.

Southworth, M.W., Benner, J., and Perler, F.B. (2000). An alternative protein splicing mechanism for inteins lacking an N-terminal nucleophile. EMBO J *19*, 5019-5026.

Stagljar, I., Korostensky, C., Johnsson, N., and te Heesen, S. (1998). A genetic system based on split-ubiquitin for the analysis of interactions between membrane proteins in vivo. Proc Natl Acad Sci U S A *95*, 5187-5192.

Stemmer, W.P., Crameri, A., Ha, K.D., Brennan, T.M., and Heyneker, H.L. (1995). Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. Gene *164*, 49-53.

Stijlemans, B., Conrath, K., Cortez-Retamozo, V., Van Xong, H., Wyns, L., Senter, P., Revets, H., De Baetselier, P., Muyldermans, S., and Magez, S. (2004). Efficient targeting of conserved cryptic epitopes of infectious agents by single domain antibodies. African trypanosomes as paradigm. J Biol Chem *279*, 1256-1261.

Sun, P., Ye, S., Ferrandon, S., Evans, T.C., Xu, M.Q., and Rao, Z. (2005). Crystal structures of an intein from the split dnaE gene of Synechocystis sp. PCC6803 reveal the catalytic model without the penultimate histidine and the mechanism of zinc ion inhibition of protein splicing. J Mol Biol *353*, 1093-1105.

Suzuki, D.T., Grigliatti, T., and Williamson, R. (1971). Temperature-sensitive mutations in Drosophila melanogaster. VII. A mutation (para-ts) causing reversible adult paralysis. Proc Natl Acad Sci U S A *68*, 890-893.

Takashima, Y., Delfino, F.J., Engen, J.R., Superti-Furga, G., and Smithgall, T.E. (2003). Regulation of c-Fes tyrosine kinase activity by coiled-coil and SH2 domains: analysis with Saccharomyces cerevisiae. Biochemistry *42*, 3567-3574.

Takehara, T., and Takahashi, H. (2003). Suppression of Bcl-xL deamidation in human hepatocellular carcinomas. Cancer Res *63*, 3054-3057.

Takesako, K., Kuroda, H., Inoue, T., Haruna, F., Yoshikawa, Y., Kato, I., Uchida, K., Hiratani, T., and Yamaguchi, H. (1993). Biological properties of aureobasidin A, a cyclic depsipeptide antifungal antibiotic. J Antibiot (Tokyo) *46*, 1414-1420.

Tamimi, Y., Lines, M., Coca-Prados, M., and Walter, M.A. (2004). Identification of target genes regulated by FOXC1 using nickel agarose-based chromatin enrichment. Invest Ophthalmol Vis Sci *45*, 3904-3913.

Tan, D.S. (2005). Diversity-oriented synthesis: exploring the intersections between chemistry and biology. Nat Chem Biol *1*, 74-84.

Tanaka, T., and Rabbitts, T.H. (2008). Functional intracellular antibody fragments do not require invariant intra-domain disulfide bonds. J Mol Biol *376*, 749-757.

Tanaka, T., and Rabbitts, T.H. (2009). Selection of complementary single-variable domains for building monoclonal antibodies to native proteins. Nucleic Acids Res *37*, e41.

Tang, Y.Q., Yuan, J., Osapay, G., Osapay, K., Tran, D., Miller, C.J., Ouellette, A.J., and Selsted, M.E. (1999). A cyclic antimicrobial peptide produced in primate leukocytes by the ligation of two truncated alpha-defensins. Science *286*, 498-502.

Tavassoli, A., and Benkovic, S.J. (2005). Genetically selected cyclic-peptide inhibitors of AICAR transformylase homodimerization. Angew Chem Int Ed Engl *44*, 2760-2763.

Tavassoli, A., and Benkovic, S.J. (2007). Split-intein mediated circular ligation used in the synthesis of cyclic peptide libraries in E. coli. Nat Protoc *2*, 1126-1133.

Tavassoli, A., Lu, Q., Gam, J., Pan, H., Benkovic, S.J., and Cohen, S.N. (2008). Inhibition of HIV budding by a genetically selected cyclic peptide targeting the Gag-TSG101 interaction. ACS Chem Biol *3*, 757-764.

Telenti, A., Southworth, M., Alcaide, F., Daugelat, S., Jacobs, W.R., and Perler, F.B. (1997). The Mycobacterium xenopi GyrA protein splicing element: characterization of a minimal intein. J Bacteriol *179*, 6378-6382.

Thorén, P.E., Persson, D., Lincoln, P., and Nordén, B. (2005). Membrane destabilizing properties of cell-penetrating peptides. Biophys Chem *114*, 169-179.

Tolmachev, V., Nilsson, F.Y., Widström, C., Andersson, K., Rosik, D., Gedda, L., Wennborg, A., and Orlova, A. (2006). 111In-benzyl-DTPA-ZHER2:342, an affibody-based conjugate for in vivo imaging of HER2 expression in malignant tumors. J Nucl Med *47*, 846-853.

Tomai, E., Butz, K., Lohrey, C., von Weizsäcker, F., Zentgraf, H., and Hoppe-Seyler, F. (2006). Peptide aptamer-mediated inhibition of target proteins by sequestration into aggresomes. J Biol Chem *281*, 21345-21352.

Toogood, P.L. (2002). Inhibition of protein-protein association by small molecules: approaches and progress. J Med Chem *45*, 1543-1558.

Trabi, M., and Craik, D.J. (2002). Circular proteins--no end in sight. Trends Biochem Sci *27*, 132-138.

Trible, R.P., Emert-Sedlak, L., and Smithgall, T.E. (2006). HIV-1 Nef selectively activates Src family kinases Hck, Lyn, and c-Src through direct SH3 domain interaction. J Biol Chem *281*, 27029-27038.

Vagner, J., Qu, H., and Hruby, V.J. (2008). Peptidomimetics, a synthetic tool of drug discovery. Curr Opin Chem Biol *12*, 292-296.

Venkataraman, N., Cole, A.L., Ruchala, P., Waring, A.J., Lehrer, R.I., Stuchlik, O., Pohl, J., and Cole, A.M. (2009). Reawakening retrocyclins: ancestral human defensins active against HIV-1. PLoS Biol *7*, e95.

Venkatesan, K., Rual, J.F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., *et al.* (2009). An empirical framework for binary interactome mapping. Nat Methods *6*, 83-90.

Vidal, M., and Legrain, P. (1999). Yeast forward and reverse 'n'-hybrid systems. Nucleic Acids Res *27*, 919-929.

Vidal, M., Brachmann, R.K., Fattaey, A., Harlow, E., and Boeke, J.D. (1996). Reverse two-hybrid and one-hybrid systems to detect dissociation of protein-protein and DNA-protein interactions. Proc Natl Acad Sci U S A *93*, 10315-10320.

Visintin, M., Settanni, G., Maritan, A., Graziosi, S., Marks, J.D., and Cattaneo, A. (2002). The intracellular antibody capture technology (IACT): towards a consensus sequence for intracellular antibodies. J Mol Biol *317*, 73-83.

Visintin, M., Tse, E., Axelson, H., Rabbitts, T.H., and Cattaneo, A. (1999). Selection of antibodies for intracellular function using a two-hybrid in vivo system. Proc Natl Acad Sci U S A *96*, 11723-11728.

Wade, J.T., and Struhl, K. (2004). Association of RNA polymerase with transcribed regions in Escherichia coli. Proc Natl Acad Sci U S A *101*, 17777-17782.

Wade, J.T., Reppas, N.B., Church, G.M., and Struhl, K. (2005). Genomic analysis of LexA binding reveals the permissive nature of the Escherichia coli genome and identifies unconventional target sites. Genes Dev *19*, 2619-2630.

Wadia, J.S., and Dowdy, S.F. (2003). Modulation of cellular function by TAT mediated transduction of full length proteins. Curr Protein Pept Sci *4*, 97-104.

Walker, G.C. (1984). Mutagenesis and inducible responses to deoxyribonucleic acid damage in Escherichia coli. Microbiol Rev *48*, 60-93.

Watanabe, K., and Oikawa, H. (2007). Robust platform for de novo production of heterologous polyketides and nonribosomal peptides in Escherichia coli. Org Biomol Chem *5*, 593-602.

Webb, T.R. (2005). Current directions in the evolution of compound libraries. Curr Opin Drug Discov Devel *8*, 303-308.

Wertman, K.F., Wyman, A.R., and Botstein, D. (1986). Host/vector interactions which affect the viability of recombinant phage lambda clones. Gene *49*, 253-262.

Wikman, M., Rowcliffe, E., Friedman, M., Henning, P., Lindholm, L., Olofsson, S., and Ståhl, S. (2006). Selection and characterization of an HIV-1 gp120-binding affibody ligand. Biotechnol Appl Biochem *45*, 93-105.

Wikman, M., Steffen, A.C., Gunneriusson, E., Tolmachev, V., Adams, G.P., Carlsson, J., and Ståhl, S. (2004). Selection and characterization of HER2/neu-binding affibody ligands. Protein Eng Des Sel *17*, 455-462.

Williams, N.K., Prosselkov, P., Liepinsh, E., Line, I., Sharipo, A., Littler, D.R., Curmi, P.M., Otting, G., and Dixon, N.E. (2002). In vivo protein cyclization promoted by a circularly permuted Synechocystis sp. PCC6803 DnaB mini-intein. J Biol Chem *277*, 7790-7798.

Wood, R.J., Pascoe, D.D., Brown, Z.K., Medlicott, E.M., Kriek, M., Neylon, C., and Roach, P.L. (2004). Optimized conjugation of a fluorescent label to proteins via intein-mediated activation and ligation. Bioconjug Chem *15*, 366-372.

Woodford, N., Wareham, D.W., and UK Antibacterial Antisense Study Group (2009). Tackling antibiotic resistance: a dose of common antisense? J Antimicrob Chemother *63*, 225-229.

Woodman, R., Yeh, J.T., Laurenson, S., and Ko Ferrigno, P. (2005). Design and validation of a neutral protein scaffold for the presentation of peptide aptamers. J Mol Biol *352*, 1118-1133.

Wörn, A., and Plückthun, A. (1998). An intrinsically stable antibody scFv fragment can tolerate the loss of both disulfide bonds and fold correctly. FEBS Lett *427*, 357-361.

Wörn, A., and Plückthun, A. (2001). Stability engineering of antibody single-chain Fv fragments. J Mol Biol *305*, 989-1010.

Wu, H., Hu, Z., and Liu, X.Q. (1998). Protein trans-splicing by a split intein encoded in a split DnaE gene of Synechocystis sp. PCC6803. Proc Natl Acad Sci U S A *95*, 9226-9231.

Wyler, E., Kaminska, M., Coïc, Y.M., Baleux, F., Véron, M., and Agou, F. (2007). Inhibition of NF-kappaB activation with designed ankyrin-repeat proteins targeting the ubiquitin-binding/oligomerization domain of NEMO. Protein Sci *16*, 2013-2022.

Xia, Z., Xing, Y., So, M.K., Koh, A.L., Sinclair, R., and Rao, J. (2008). Multiplex detection of protease activity with quantum dot nanosensors prepared by intein-mediated specific bioconjugation. Anal Chem *80*, 8649-8655.

Xu, L., Aha, P., Gu, K., Kuimelis, R.G., Kurz, M., Lam, T., Lim, A.C., Liu, H., Lohse, P.A., *et al.* (2002). Directed evolution of high-affinity antibody mimics using mRNA display. Chem Biol *9*, 933-942.

Xu, M.Q., and Evans, T.C. (2001). Intein-mediated ligation and cyclization of expressed proteins. Methods *24*, 257-277.

Xu, M.Q., and Perler, F.B. (1996). The mechanism of protein splicing and its modulation by mutation. EMBO J *15*, 5146-5153.

Yagodkin, A., Azhayev, A., Roivainen, J., Antopolsky, M., Kayushin, A., Korosteleva, M., Miroshnikov, A., Randolph, J., and Mackie, H. (2007). Improved synthesis of trinucleotide phosphoramidites and generation of randomized oligonucleotide libraries. Nucleosides Nucleotides Nucleic Acids *26*, 473-497.

Yasuda, T., Morimatsu, K., Horii, T., Nagata, T., and Ohmori, H. (1998). Inhibition of Escherichia coli RecA coprotease activities by DinI. EMBO J *17*, 3207-3216.

Yin, X., Giap, C., Lazo, J.S., and Prochownik, E.V. (2003). Low molecular weight inhibitors of Myc-Max interaction and function. Oncogene *22*, 6151-6159.

Yokota, T., Milenic, D.E., Whitlow, M., and Schlom, J. (1992). Rapid tumor penetration of a single-chain Fv and comparison with other immunoglobulin forms. Cancer Res *52*, 3402-3408.

Yongye, A.B., Li, Y., Giulianotti, M.A., Yu, Y., Houghten, R.A., and Martínez-Mayorga, K. (2009). Modeling of peptides containing D-amino acids: implications on cyclization. J Comput Aided Mol Des *23*, 677-689.

Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., *et al.* (2008). High-quality binary protein interaction map of the yeast interactome network. Science *322*, 104-110.

Zehnder-Fjällman, A.H., Marty, C., Halin, C., Hohn, A., Schibli, R., Ballmer-Hofer, K., and Schwendener, R.A. (2007). Evaluation of anti-VEGFR-3 specific scFv antibodies as potential therapeutic and diagnostic tools for tumor lymph-angiogenesis. Oncol Rep *18*, 933-941.

Zhang, D.Y., Ye, F., Gao, L., Liu, X., Zhao, X., Che, Y., Wang, H., Wang, L., Wu, J., *et al.* (2009). Proteomics, pathway array and signaling network-based medicine in cancer. Cell Div *4*, 20.

Zhelev, Z., Bakalova, R., Ohba, H., Ewis, A., Ishikawa, M., Shinohara, Y., and Baba, Y. (2004). Suppression of bcr-abl synthesis by siRNAs or tyrosine kinase activity by Glivec alters different oncogenes, apoptotic/antiapoptotic genes and cell proliferation factors (microarray study). FEBS Lett *570*, 195-204.

Züger, S., and Iwai, H. (2005). Intein-based biosynthetic incorporation of unlabeled protein tags into isotopically labeled proteins for NMR studies. Nat Biotechnol *23*, 736-740.