DEFINING SUBSTRATE REQUIREMENTS FOR CLEAVAGE OF PRELAMIN A BY THE ZINC METALLOPROTEASE ZMPSTE24

By

Timothy David Babatz

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
May 2019

ABSTRACT:

Proteases play important roles in diverse biological processes relevant to human health and disease and defining their substrate specificity is an essential component to understanding molecular mechanisms of proteolysis. ZMPSTE24 is a zinc metalloprotease that is critical to the proteolytic maturation of lamin A, an intermediate filament protein and component of the nuclear lamina. When ZMPSTE24-dependent processing of prelamin A is disrupted by mutations in either the substrate or the protease, accumulation of uncleaved prelamin A causes a spectrum of genetic disorders including the premature aging disease Hutchinson Gilford Progeria Syndrome (HGPS) and related progeroid disorders mandibuloacral dysplasia (MAD-B) and restrictive dermopathy (RD).

Prelamin A is the only known mammalian substrate for ZMPSTE24, however, the basis of this specificity remains unclear. To begin to define the sequence requirements for ZMPSTE24 recognition, I have performed a comprehensive mutagenesis scan of the eight residues flanking the cleavage site, from amino acid residue T643 to N650 (TRSY↓LLGN). Mutants were tested in an *in vivo* humanized yeast assay that provides a sensitive measure of ZMPSTE24 processing efficiency.

Substitutions C-terminal to the cleavage site were generally more disruptive than N-terminal substitutions, although R644 shows a modest bias for positively charged residues. Of particular note is L647 in the P1’ position C-terminal to the cleavage site, of which mutation to arginine has long been known to be “uncleavable.” Here I show that charged residues, aromatics, and proline

disrupt cleavage, while hydrophobic residues are well-tolerated. L648 at P2' shows a similar pattern with several key differences.

I have generated a heat map of cleavage efficacy for ZMPSTE24 and positions P1' and P2' have emerged as critical for ZMPSTE24 substrate recognition, with a strong preference for hydrophobic residues. Hydrophobic residues at these two positions are evolutionarily conserved. I have shown that the 8 residues flanking the deduced ZMPSTE24 cleavage site in prelamin A sequences from amphibians, birds, fish, and mammals can promote cleavage when placed in the context of human lamin A, albeit with varying efficiency. Taken together, my results provide mechanistic insights into the recognition of prelamin A by ZMPSTE24 and begin to define a consensus motif with potential predictive value to identify other ZMPSTE24 substrates.

Dissertation Advisor: Dr. Susan Michaelis

Dissertation Reader: Dr. Peter Espenshade

ACKNOWLEDGEMENTS

I would like to thank Dr. Susan Michaelis for being a generous, encouraging, and challenging mentor. From our first meeting at PhD recruitment to joining her lab several years later, Susan was always an inspiring and motivating role model.

For the years prior to joining the Michaelis Lab, I am grateful for the mentorship of Drs. Kathleen Burns and Jef Boeke. I would also like to the thank the faculty and staff of the Human Genetics program and McKusick-Nathans Institute of Genetic Medicine, in particular Dr. Kirby Smith, Dr. David Valle, and Sandy Muscelli. And I would like to thank the additional members of my thesis committee, Dr. Peter Espenshade and Dr. Sin Urban.

The following people also had a significant influence on my scientific training and my time as a PhD candidate: Dr. Marcelo Nobrega, Dr. William Dobyns, Dr. Susan Christian, Dr. Ravinesh Kumar, Dr. Jennifer Poitras, Nina Rajpurohit, Dr. Lauren Zeitels, Jose Llongueras, Dr. Catherine Bennet, Dr. Matt Bender, Dr. Logan Weygandt, Dr. John Nichols, Dr. Philip Storey, Dr. Marty Taylor, Dr. Eric Spear, Kaitlin Wood, Otto Mossberg, and Dr. Khurts Shilagardi.

I’m also grateful for being awarded the Isaac Morris Hay and Lucille Elizabeth Hay Graduate Fellowship Award by the Department of Cell Biology, which funded one year of my PhD studies.

Finally, I would like to thank my parents, grandparents, and other family members who have provided endless support over the years, whether while growing up in Minnesota or as a graduate student on the East Coast.

Dedicated to Richard Joseph Kelly

September 20, 1932 – April 4, 2019

An exemplary scholar, role model, and Grandfather

Table of Contents

List of Figures

Chapter I

Background and Introduction

The work presented in this dissertation is a contribution to a field that has traced a path from basic yeast biology, to mammalian cell biology, to human genetics and clinical phenotypes. In this chapter, I review key advances that led to the identification of the protease ZMPSTE24, its substrate prelamin A, and their roles in human disease.

Yeast a-factor Biogenesis

The yeast *Saccharomyces cerevisiae* can exist in either a haploid or diploid state. Mating can occur between haploid cells, of which there are two mating types, a-type or α-type determined by the alternative mating type cassettes *MAT*a or *MAT*α (Nasmyth, 1982). These haploid mating types secrete mating pheromones, a-factor and α-factor, respectively (reviewed in Michaelis and Barrowman, 2012). When the pheromones are detected by surface receptors on the opposite mating type, a signaling cascade triggers a decision to mate, the cells grow distinctive projections toward the cell of the opposing mating type, the two haploid cells fuse, and subsequently their nuclei fuse, to form a stable diploid cell.

Since these mating pheromones are secreted molecules, they represented an attractive target for studying the secretory pathway, or the means by which a protein is translated and exported out of the cell. The a-factor pheromone is a CAAX protein, typified by a C-terminal CAAX motif (cysteine, aliphatic, aliphatic, any residue) (Zhang and Casey, 1996). This CAAX motif directs several post-translational modifications, including farnesylation of the cysteine residue, endoproteolytic cleavage of the C-terminal tripeptide (-AAX), and carboxymethylation of the farnesylated cysteine (reviewed in Michaelis and Barrowman, 2012) (Fig. 1).

Previously, our lab conducted a mutant hunt and identified *ste24*, a yeast mutant with a mating defect rendering it sterile (Fujimura-Kamada, Nouvet, & Michaelis, 1997). The *ste24* mutant accumulated unprocessed a-factor

precursor, failed to export mature a-factor, and was sterile in a highly sensitive mating assay. This defect was complemented by transformation with a yeast genomic library, the gene was identified, and the sequence of its gene product, Ste24p, was inferred.

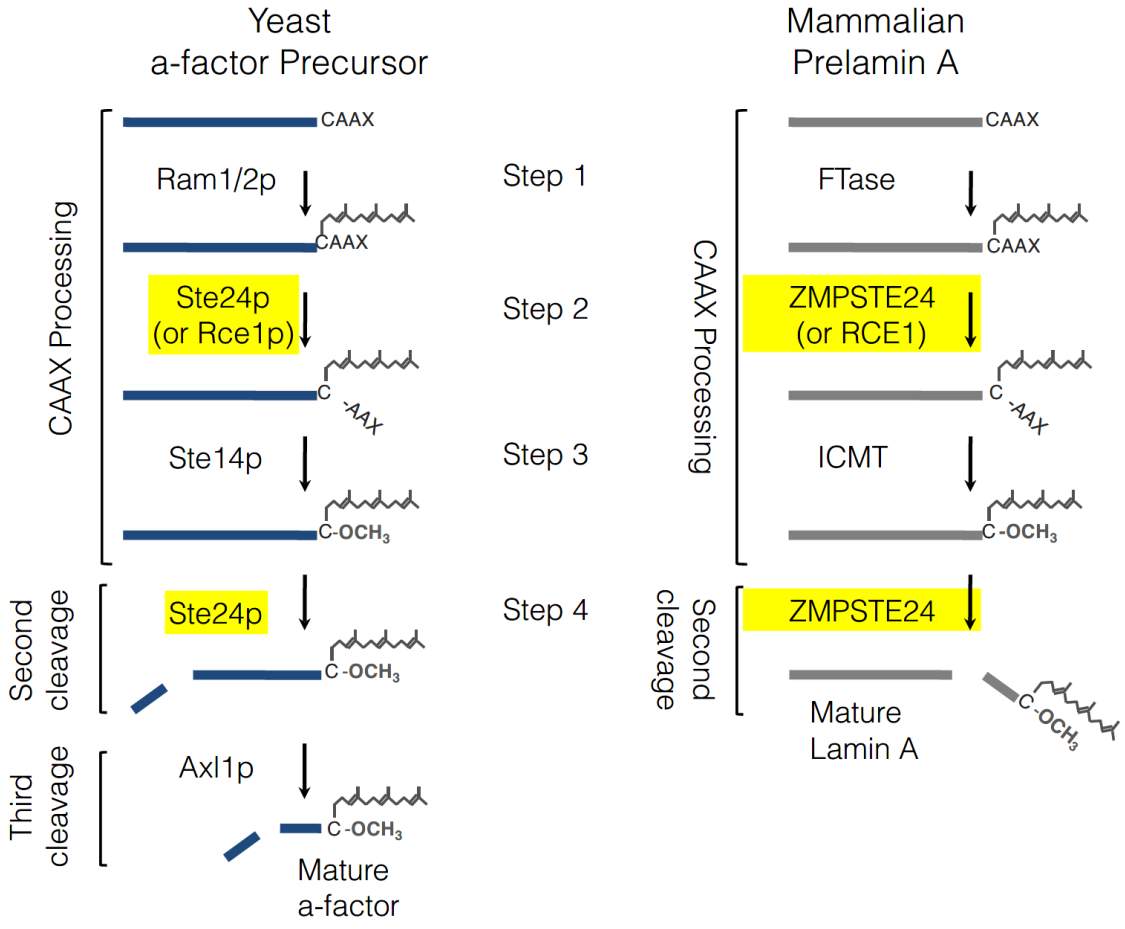

Figure 1: Biogenesis pathways of yeast a-factor and mammalian prelamin A. The nearly identical post-translational processing steps of yeast a-factor and mammalian prelamin A are shown, as well as the respective enzymes that perform these steps. CAAX processing (farnesylation, -AAXing, and carboxymethylation) is followed by upstream cleavage by yeast Ste24p or mammalian ZMPSTE24. Two distinctions between the pathways are that a-factor undergoes an additional N-terminal cleavage by Axl1p, and mature a-factor is the farnesylated secreted cleavage product while mature lamin A loses its farnesylated C-terminus after upstream cleavage.

The amino acid sequence of Ste24p suggested an integral membrane protein with seven transmembrane spans, a canonical HEXXH motif characteristic of zinc metalloproteases and a modified dilysine ER retrieval motif (Tam et al., 1998). A search of EST databases identified mouse and human homologs of Ste24p which had almost identical hydropathy plots. Remarkably, expression of the human Ste24p homolog in yeast rescued the mating defect, despite having only 36% identity and 51% similarity to the yeast Ste24p amino acid sequence.

Further characterization of the human Ste24p homolog, known as ZMPSTE24, revealed a zinc metalloprotease with seven transmembrane spans, the HEXXH motif, and a dilysine ER retrieval signal. ZMPSTE24 localizes to both the inner nuclear membrane and the ER membrane, as does yeast Ste24p (Barrowman, Hamblet, George, & Michaelis, 2008; Barrowman & Michaelis, 2009).

More recently, the crystal structures for both yeast Ste24 (Pryor et al., 2013) and human ZMPSTE24 (Quigley et al., 2013; Clark et al., 2017) were solved, revealing a novel and fascinating structure, distinct from other known integral membrane zinc proteins. Both proteins show remarkable structural similarity, with their seven transmembrane spans forming a large water-filled barrel structure with several side portals, presumably one of which allows for substrate access to the chamber. The zinc-coordinating active site sits inside the chamber at the cytoplasmic-facing end of the structure. The active site is inside a

groove lined with hydrophobic residues, very similar to the active site of thermolysin, a well-characterized bacterial soluble zinc metalloprotease.

Finding the Mammalian Substrate of ZMPSTE24

Returning to the functional characterization of mammalian ZMPSTE24, it was known that ZMPSTE24 was a zinc metalloprotease that could complement the mating defect of Ste24-deficient yeast. However, mammalian cells lack any known homolog of yeast a-factor, and the mammalian substrate of ZMPSTE24 remained unknown. To address this question, two groups generated *Zmpste24*$^{-/-}$ mice (Bergo et al., 2002; Pendás et al., 2002). These mice were viable and healthy at birth, however, within a few weeks they began to show lipodystrophy, skeletal abnormalities, low bodyweight, loss of hair, and had a shortened lifespan.

It was known that like a-factor, the lamin A precursor, prelamin A, was a CAAX protein. The CAAX motif directs several post-translational modifications including endoproteolytic cleavage of its farnesylated prelamin A C-terminus (Weber et al., 1989) (Fig. 1). However, the protease responsible for the cleavage step was unknown. The putative cleavage site 15 amino acids upstream of the farnesylated cysteine had been identified by protein sequencing of the C-terminus of mature lamin A (Weber et al., 1989). Our lab had suggested prelamin A as a possible substrate for mammalian ZMPSTE24 (Tam et al., 1998) based on these observations that it was both farnesylated and cleaved similarly to the a-factor biogenesis pathway.

Mouse embryonic fibroblasts generated from the *Zmpste24*$^{-/-}$ animals were shown to accumulate uncleaved persistently farnesylated prelamin A, providing compelling evidence implicating ZMPSTE24 as the protease responsible for endoproteolytic cleavage of prelamin A (Bergo et al., 2002; Pendás et al., 2002). Furthermore, *Zmpste24*$^{-/-}$ mice showed phenotypic overlap with both the *Lmna*$^{-/-}$ mouse model (Sullivan et al., 1999) and the increasingly well-characterized group of human disease known as laminopathies (Capell & Collins, 2006). The mapping of the Hutchinson-Gilford Progeria Syndrome (HGPS) mutation in 2003 (De Sandre-Giovannoli et al., 2003; Eriksson et al., 2003) to the ZMPSTE24 cleavage site, discussed in further detail below, revealed that a lack of prelamin A processing promotes progeria.

Finally, both -AAXing and upstream cleavage of the prelamin A substrate by recombinant ZMPSTE24 were reconstituted *in vitro*, providing further substantial support for the role of ZMPSTE24 in prelamin A processing (Corrigan et al., 2005).

Function and Biogenesis Pathway of Lamin A

Lamins are intermediate filament proteins that polymerize to form the nuclear lamina, a structure critical to the function the nucleus (Aebi, Cohn, Buhle, & Gerace, 1986). The nuclear lamina plays a role in nuclear structure and mechanoresponsiveness, nuclear pore complex organization, and heterochromatin organization and gene expression (Dechat et al., 2008).

Lamins A and C are transcribed as isoforms from the same gene, *LMNA*, while B-type lamins are encoded by two distinct genes, *LMNB1* and *LMNB2*, (Worman, Fong, Muchir, & Young, 2009). Lamin A and B-type lamins have a C-terminal CAAX motif, which directs the post-translational modifications of the nascent polypeptide prelamin A. However, B-type lamins are persistently farnesylated while the C-terminus of prelamin A is cleaved. Lamin C, a splice variant of *LMNA*, lacks the C-terminus of lamin A, does not have a CAAX motif, and is therefore not cleaved.

As discussed above and illustrated in Figure 1, elucidation of the lamin A biogenesis pathway was largely informed by studies of yeast a-factor biogenesis. Prelamin A is also a CAAX-modified protein and undergoes post-translational processing by an almost identical pathway (Michaelis & Barrowman, 2012) (Fig. 1). First, prelamin A is farnesylated at the cysteine of the CAAX motif by farnesyltransferase, the C-terminal tripeptide is cleaved by RCE1 or ZMPSTE24 (-AAXing), and the now C-terminal cysteine is carboxymethylated by isoprenyl cysteine methyltransferase (ICMT). Finally, a C-terminal 15 amino acid farnesylated tail is cleaved by the protease ZMPSTE24.

Human Phenotypes

Mutations of the gene encoding the substrate, prelamin A, or the protease, ZMPSTE24, lead to accumulation in cells of a persistently farnesylated form of uncleaved prelamin A and causes a spectrum of genetic disorders characterized by myopathy, lipodystrophy, and premature aging (Worman et al., 2009). Three

distinctive diseases are due to accumulation of farnesylated uncleaved prelamin A.

The best studied of these disorders is Hutchinson-Gilford Progeria syndrome (HGPS), an extremely rare, severe, premature aging syndrome. HGPS is caused by an autosomal dominant splicing mutation in *LMNA* that activates a cryptic splice donor site and affects 1 in approximately 4 million live births. This aberrant splicing results in a 50 amino acid deletion in prelamin A that includes the ZMPSTE24 processing site (De Sandre-Giovannoli et al., 2003; Eriksson et al., 2003). A truncated, persistently farnesylated form of prelamin A called progerin accumulates in cells of HGPS patients and causes a characteristic aberrant nuclear morphology phenotype at the cellular level. Patients show symptoms around one year of age including alopecia, lipodystrophy, thin skin, characteristic small jaw, and small stature. Patients usually succumb to heart attack or stroke in their teens due to progressive atherosclerosis (Capell, Collins, & Nabel, 2007).

The causal *LMNA* mutation was identified in 2003 (De Sandre-Giovannoli et al., 2003; Eriksson et al., 2003), which sparked a rapid search for therapies to ameliorate the cardiovascular phenotypes of HGPS patients and extend their lifespan. Farnesyl transferase inhibitors (FTIs) were a class of compounds that already existed as an experimental anti-cancer drug, and it was hypothesized that treatment with FTIs may inhibit the formation of progerin and reduce the toxic effects of its accumulation in cells. Indeed, FTIs reduced the characteristic nuclear blebbing phenotype of HGPS cells and wild-type cells transiently

transfected to express progerin (Capell et al., 2005). Treatment with Lonafarnib also extended the lifespan of the mouse model of HGPS (Yang et al., 2006).

In patients, clinical trials of FTIs are inherently underpowered because of the incredibly small patient population. Despite that, trials of Lonafarnib have been promising, and the treatment has been shown to improve the patients' cardiovascular phenotype and increase lifespan (Gordon et al., 2012, 2018).

Restrictive dermopathy (RD) is a severe autosomal recessive disorder caused by loss or complete inactivation of both *ZMPSTE24* alleles by frameshift or nonsense mutations. RD usually results in death *in utero* or shortly after birth. The phenotype includes tight thin skin, abnormal joint contractures, and severe dysmorphology. RD is caused by a complete loss of ZMPSTE24 activity and toxic accumulation of prelamin A (Navarro et al., 2005).

Mandibuloacral Dysplasia Type B (MAD-B) is a milder form of disease due to heterozygous loss of function mutations of *ZMPSTE24*. Generally one allele is a frameshift or nonsense mutation and the other is a substitution mutation. Residual ZMPSTE24 activity from the remaining allele results in a milder phenotype with onset in adulthood. Symptoms include lipodystophy, skeletal abnormalities, and a characteristic facies (Agarwal & Garg, 2006).

Additionally, there is evidence that cleavage of prelamin A by ZMPSTE24 has a role in normal human aging (Ragnauth et al., 2010). Aged vascular smooth muscle cells (VSMCs) have been observed to accumulate uncleaved prelamin A in culture, and this accumulation was ameliorated by treatment with farnesyl transferase inhibitors (FTIs). The authors also observed accumulation of

prelamin A in medial VSMCs from arteries of aged individuals with atherosclerotic lesions, while prelamin A was absent from vessels of young healthy individuals. Prelamin A has emerged as a novel biomarker of aging in VSMCs and may play a role in the association of cardiovascular disease with normal human aging.

Cardiovascular disease is associated with advanced age and age-related vascular dysfunction is a significant public health concern. Understanding the molecular mechanisms of prelamin A processing by ZMPSTE24, and the ways that the efficiency of this process may decline in aging VSMCs (and possibly endothelial cells), will point toward therapeutic interventions that may be designed to treat age-related vascular decline.

ZMPSTE24 Substrate Specificity

Prelamin A is the only substrate currently known for mammalian ZMPSTE24. Human ZMPSTE24 expressed in yeast can complement the mating defect seen in the *Δste24 Δrce1* double mutant, indicating that ZMPSTE24 can perform both the -AAXing and N-terminal cleavage steps of a-factor biogenesis (Barrowman & Michaelis, 2009; Tam et al., 1998) despite the a-factor N-terminal cleavage site showing little similarity to the prelamin A processing site. In addition, recent reports have highlighted a novel function for ZMPSTE24 as a translocon declogger in the ER, a novel function dependent on its proteolytic activity (Ast, Michaelis, & Schuldiner, 2016; Kayatekin et al., 2018). Thus, ZMPSTE24 has at least three suitable specific substrate amino acid sequences

and a more generalized proteolytic function at the translocon. But any rules for substrate recognition or a substrate consensus sequence remain unknown.

Several pieces of evidence suggest that recognition of prelamin A by ZMPSTE24 is to some extent sensitive to mutation at the prelamin A cleavage site (TRSY↓LLGN). First, mutation of leucine to arginine at residue 647, the P1' position adjacent to the cleavage site, has long been recognized as uncleavable. First described for chicken prelamin A expressed in mammalian cells (Hennekes & Nigg, 1994), this mutant has since been used as a benchmark in the field to represent farnesylated unprocessed prelamin A. Notably, a patient was recently reported harboring this mutation, presenting with a progeroid disorder including short stature, micrognathia, and skin and bone abnormalities (Wang et al., 2016). This patient's cells showed an accumulation of farnesylated prelamin A and aberrant nuclear morphology, indicating the L647R mutation abolishes ZMPSTE24 processing in humans.

Second, a genetic variant causing an R644C substitution at the P3 position has been reported segregating in families with dilated cardiomyopathy (DCM), skeletal abnormalities, and atypical progeria (Csoka et al., 2004; Genschel & Schmidt, 2000; Mercuri et al., 2005; Rankin et al., 2008). However, this variant shows extreme phenotypic heterogeneity and incomplete penetrance. Whether the diverse phenotypes associated with R644C are due to a ZMPSTE24 processing defect, another pathogenic mechanism, or are unrelated to the R644C variant remains unknown, since patient cells have generally not been examined. R644C has also appeared at an appreciable frequency (~0.12%) in

the ExAC database (Lek et al., 2016), a collection of exome sequences from ~60,000 unrelated adults with no history of severe pediatric disease. This suggests that R644C is indeed polymorphic but does not clarify its possible role in associated phenotypes.

Finally, previous work by our group has defined several substrate requirements for efficient proteolytic cleavage of prelamin A by ZMPSTE24. In cultured mammalian cells, exogenously expressing a minimum C-terminal segment of 41 amino acids N-terminally tagged with GFP is sufficient to direct cleavage between residues Y646 and L647 (Barrowman, Hamblet, Kane, & Michaelis, 2012). Limited mutagenesis of several amino acids flanking the cleavage site disrupted cleavage, as did doubling the length the segment between the cleavage site and the farnesylated CAAX motif.

Taken together these data suggest some specific requirements for ZMPSTE24 to recognize and process its substrate prelamin A. However, we currently have an incomplete picture of the landscape surrounding the cleavage site and the features of prelamin A that are specifically recognized by ZMPSTE24.

Chapter II

Amino Acid Sequence Requirements for ZMPSTE24 Cleavage of Prelamin A

In the present study, I sought to exhaustively characterize the rules of allowable substitutions at the residues flanking the prelamin A cleavage site. I utilized a humanized yeast assay developed by our group (Spear et al., 2018) and saturation mutagenesis to test a comprehensive collection of prelamin A cleavage site mutants *in vivo*.

This approach has several distinct advantages. Yeast are very amenable to genetic manipulation, allowing for rapid, low cost screening of large numbers of mutants in an isogenic background. Farnesylation is absolutely required for prelamin A processing, and it is currently not feasible to produce large numbers of synthetic farnesylated substrates *in vitro*. This *in vivo* assay lets yeast perform this post-translational modification and bypasses some of the technical challenges of studying a membrane-associated protease and farnesylated substrate *in vitro*.

The following chapter includes results of this work, discussion and conclusions, and materials and methods used in the work.

Prelamin A Processing Modeled in Yeast

Our humanized yeast system successfully models lamin A biogenesis *in vivo*. Similar to the system described in our previous work (Spear et al., 2018), a C-terminal portion of human prelamin A is co-expressed with full-length human ZMPSTE24 in a *ste24Δ* background. The preceding steps of lamin A biogenesis are accomplished by the yeast homologs of the human CAAX processing machinery, normally involved in a-factor biogenesis. The result is a farnesylated, carboxymethylated ZMPSTE24 substrate (Fig. 2a). Only residues 431-664 of prelamin A are expressed, to avoid the coiled-coil domain and polymerization of mature lamin A monomers (Fig. 2b).

In a modification of the system described in our previous work (Spear et al., 2018), here the full full-length N-terminally HA-tagged ZMPSTE24 ORF is chromosomally integrated at the *TRP1* locus to produce strain SM6303, and a collection of prelamin A mutants are expressed from a centromeric plasmid (Fig 2b). We found this arrangement ideal for maintaining consistent ZMPSTE24 expression while assaying large numbers of prelamin A mutants. This modification is discussed further in Chapter III.

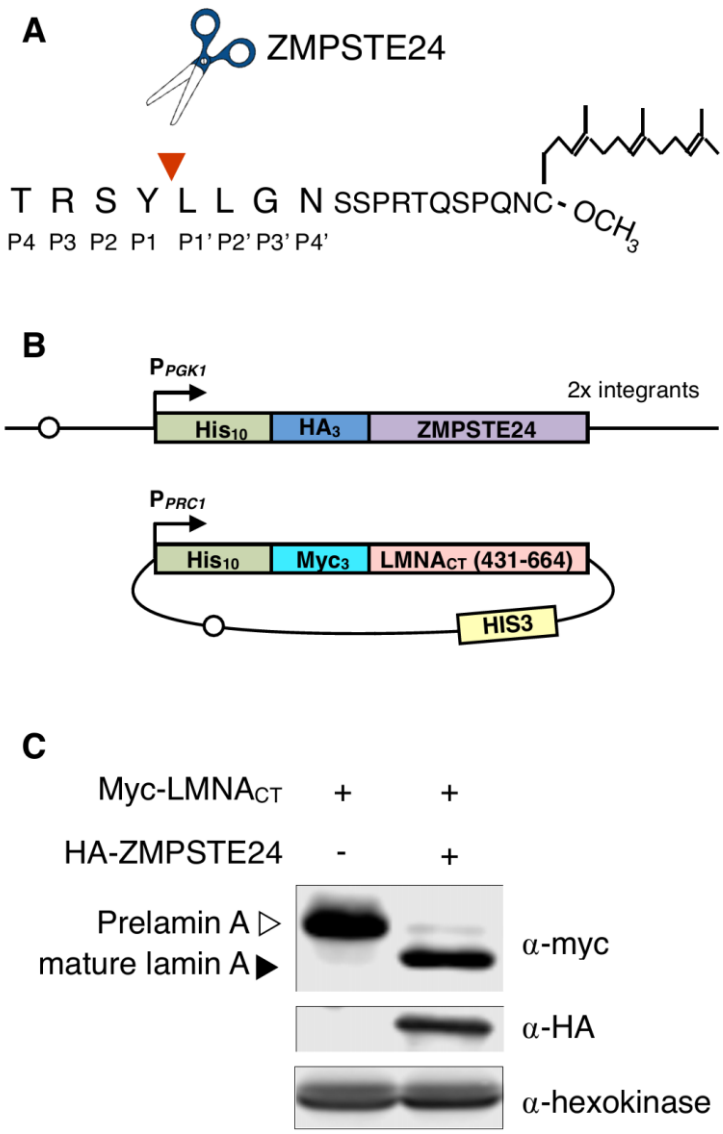

Figure 2: The humanized yeast system used to assess human ZMPSTE24 cleavage of human prelamin A. (A) Schematic of the C-terminus of prelamin A after farnesylation, -AAXing, and carboxymethylation. The ZMPSTE24 cleavage site between prelamin A residues Y646 and L647 is indicated by a red arrowhead, and substrate residues mutagenized in this study at positions P4 through P4‘ are labeled. (B) Schematic of the humanized yeast system to assess processing of mutant forms of prelamin A. Full-length human ZMPSTE24 with an N-terminal 10His-3HA epitope tag is expressed from the PGK1 promoter (P_{PGK1}) and is chromosomally integrated into a *ste24Δ* strain background at the *TRP1* locus, resulting in strain SM6303, which contains 2 tandem copies of the construct shown. The human prelamin A model substrate contains amino acids 431-664 from the C terminus of prelamin A, encoded by *LMNA*. In plasmid pSM3393 this substrate is referred to as $LMNA_{CT}$ and is N-

terminally tagged with a 10His-3myc epitope. Wild-type and mutant versions of $LMNA_{CT}$ are expressed from the *PRC1* promoter (P_{PRC1}) on a *CEN* HIS3 plasmid in strain SM6303. Open circles represent centromeric sequences on the chromosome and plasmid. (C) Lysates from a *ste24Δ* strain (SM4826) with plasmid-expressed prelamin A with and without chromosomally-integrated ZMPSTE24 (strains SM6303 and SM4826 respectively) were analyzed for prelamin A processing by 10% SDS-PAGE and western blotting. Prelamin A and mature lamin A were detected with anti-myc antibodies; ZMPSTE24 was detected with anti-HA antibodies. Hexokinase is the loading control.

An immunoblot probing for myc-tagged lamin A and HA-tagged ZMPSTE24 shows successful prelamin A maturation in yeast that is ZMPSTE24-dependent (Fig. 2c). In the presence of ZMPSTE24, bands representing uncleaved prelamin A and cleaved mature lamin A are clearly resolved.

I note that there is a residual pool of unprocessed prelamin A that persists in steady-state in this assay, unlike the complete processing that is typically observed in the native context. This could be due to the exogenous expression of lamin A exceeding the capacity of native yeast RAM1 and RAM2, orthologs of farnesytransferase. Final cleavage by ZMPSTE24 is dependent on farnesylation, and we have not ruled out the possibility that cleavage site mutations may affect the required CAAX modifications preceding upstream cleavage. However, this possibility seems unlikely, since transplanted CAAX motifs attached to a variety of passenger proteins are fully modified, suggesting upstream sequences do not affect farnesylation (Hildebrandt et al., 2016). Alternatively, the residual unprocessed prelamin A may be inaccessible to ZMPSTE24 if it is not reaching the cytosolic face of the ER membrane, where this processing step occurs.

This assay is sensitive to the stoichiometry of protease and substrate, as increasing ZMPSTE24 abundance increased prelamin A cleavage efficiency

(Spear et al., 2019), (see also Chapter III, Figure 12). However, the highest cleavage efficiency we have seen is ~80%, with a pool of unprocessed prelamin A persisting. Additionally, the assay yields a measure of cleavage efficiency for processing of prelamin A by ZMPSTE24. However, I note that the assay reflects steady-state levels of prelamin A processing and does not indicate rates of processing or enzyme kinetics.

Comprehensive Scanning Mutagenesis of Prelamin A Cleavage Site

I used a combination of degenerate oligonucleotide mutagenesis and Gibson assembly to comprehensively mutagenize the eight residues flanking the prelamin A cleavage site, positions P4—P4', yielding 152 mutated variants of the prelamin A expression construct described in Figure 2b. These plasmids were individually transformed into ZMPSTE24-expressing strain SM6303 and assayed by immunoblot as described in Figure 2c. For each residue, wild-type and 19 mutants were assayed in parallel, and cleavage efficiency relative to wild-type was plotted in Figure 3a-d and Figure 4a-d. Figure 5 shows the summation of these results as a heatmap of cleavage efficiency across the cleavage site.

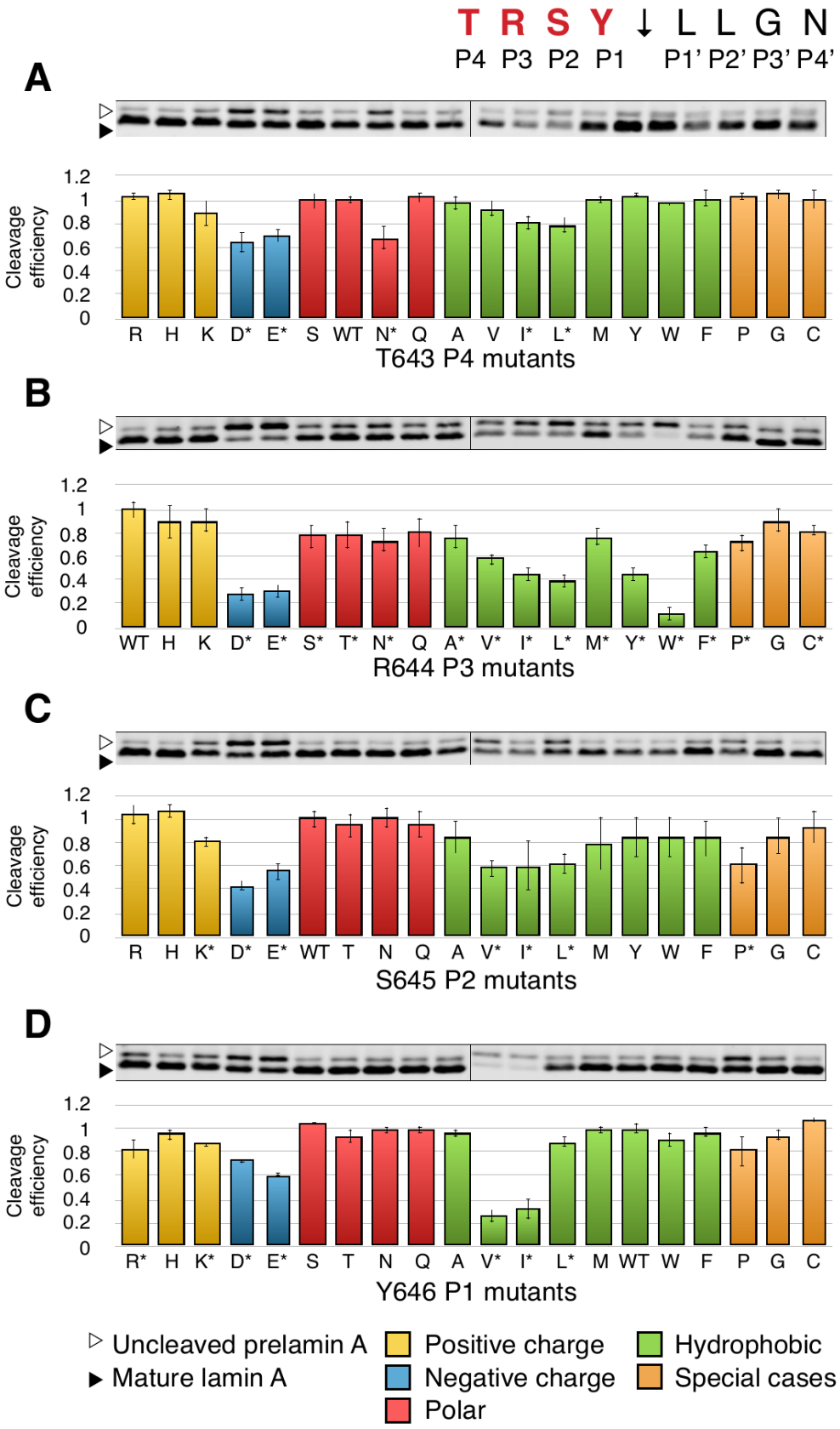
T R S Y ↓ L L G N
P4 P3 P2 P1 P1' P2' P3' P4'
A
Cleavage efficiency
1.2
1
0.8
0.6
0.4
0.2
0
R H K D* E* S WT N* Q A V I* L* M Y W F P G C
T643 P4 mutants
B
Cleavage efficiency
WT H K D* E* S* T* N* Q A* V* I* L* M* Y* W* F* P* G C*
R644 P3 mutants
C
Cleavage efficiency
R H K* D* E* WT T N Q A V* I* L* M Y W F P* G C
S645 P2 mutants
D
Cleavage efficiency
R* H K* D* E* S T N Q A V* I* L* M WT W F P G C
Y646 P1 mutants
Uncleaved prelamin A
Mature lamin A
Positive charge
Negative charge
Polar
Hydrophobic
Special cases

Figure 3: Comprehensive scanning mutagenesis of positions P4—P1 in prelamin A to all other 19 amino acids. Western blots for all possible variants of residues T643 (A), R644 (B), S645 (C), and Y646 (D) are shown. Uncleaved prelamin A (empty arrowhead) and mature lamin A (filled arrowhead) are indicated. For each residue, wild-type is indicated (WT; red). Residues are color-coded as indicated. Cleavage efficiency is calculated by dividing mature lamin A by total lamin A abundance (cleaved + uncleaved). Wild-type for each residue is normalized to 1 and all mutants are plotted proportionate to 1. Three biological replicates were performed for each mutant. Asterisks by the residue name indicate mutations in which processing is statistically significantly different than wild-type, Student's T-test, two-tailed, unpaired, N=3.

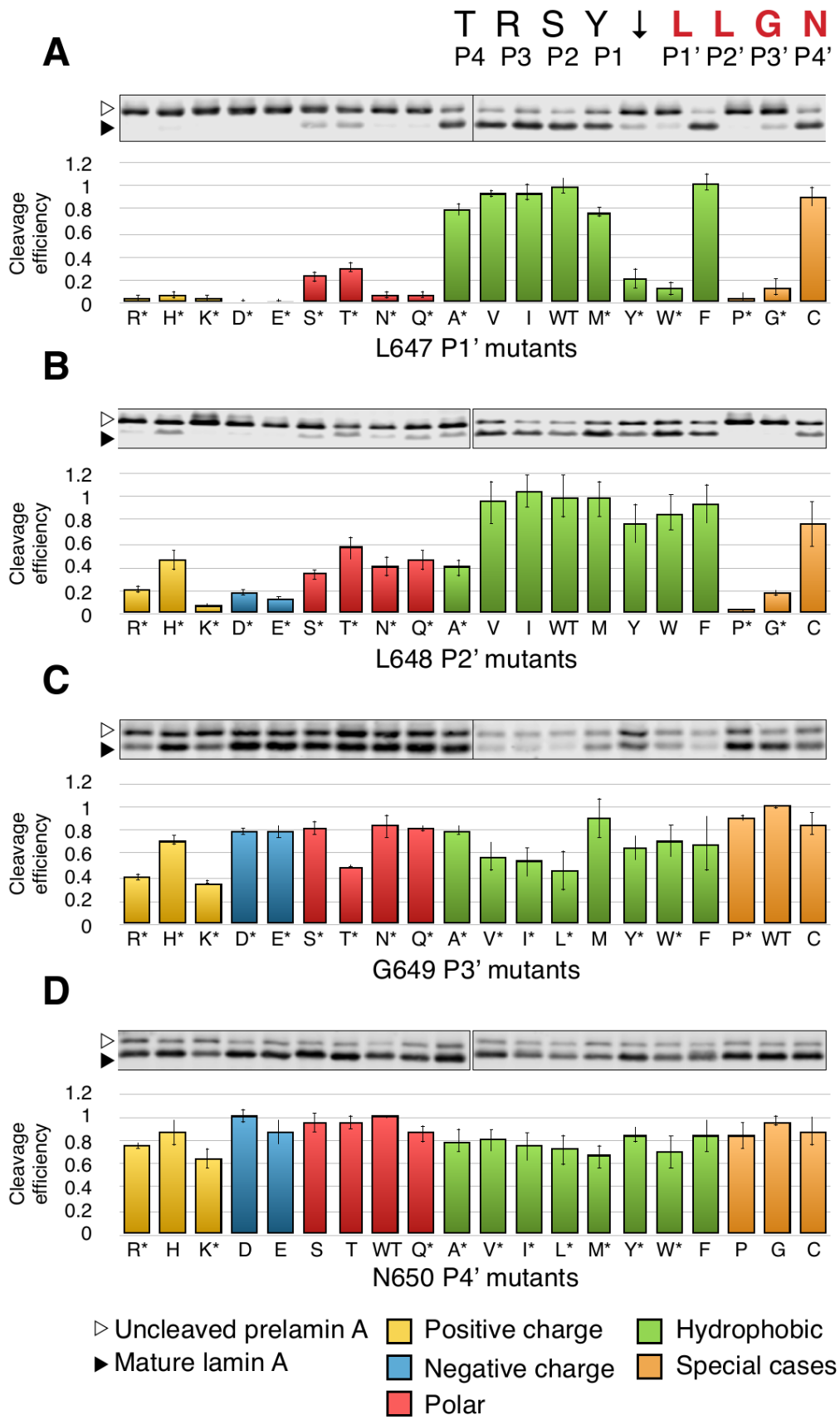
A
T R S Y ↓ L L G N
P4 P3 P2 P1 P1' P2' P3' P4'
Cleavage efficiency
1.2
1
0.8
0.6
0.4
0.2
0
R* H* K* D* E* S* T* N* Q* A* V I WT M* Y* W* F P* G* C
L647 P1' mutants
B
R* H* K* D* E* S* T* N* Q* A* V I WT M Y W F P* G* C
L648 P2' mutants
C
R* H* K* D* E* S* T* N* Q* A* V* I* L* M Y* W* F P* WT C
G649 P3' mutants
D
R* H K* D E S T WT Q* A* V* I* L* M* Y* W* F P G C
N650 P4' mutants
▷ Uncleaved prelamin A
▶ Mature lamin A
Positive charge
Negative charge
Polar
Hydrophobic
Special cases

Figure 4: Comprehensive scanning mutagenesis of positions P1'—P4' in prelamin A to all other 19 amino acids. Western blots for all possible variants of residues L647 (A), L648 (B), G649 (C), and N650 (D), are shown. Uncleaved prelamin A (empty arrowhead) and mature lamin A (filled arrowhead) are indicated. For each residue, wild-type is indicated. Quantification and statistical analysis are identical to those performed for the plots in Figure 2.

Several novel insights can be gathered from these data. First, the cleavage site recognition motif is generally defined by positions P3-P3'. T643 at position P4 and N650 at P4' are rather resistant to amino acid substitution, even radical substitutions, *e.g.* T643Y and N650F (polar to aromatic hydrophobic changes).

Generally, the C-terminal side of the cleavage site (P1'-P3') is more sensitive to mutation than the N-terminal side (P3-P1). No mutations to the N-terminal side of the cleavage site completely abolish processing, while many C-terminal mutations do.

For several mutations, we observed a decrease in total lamin A abundance (sum of cleaved and uncleaved bands) relative to loading control and ZMPSTE24, e.g. Y646V and Y646I (Fig. 3d). To test whether codon usage in yeast was affecting translation of the exogenously expressed human prelamin A, I optimized Y646V and Y646I for expression in yeast but this did not restore total lamin A abundance to wild-type levels. It is possible that stability is affected in these mutants, which could be tested using proteasomal and lysosomal yeast mutants.

P1’ and P2’ Are Critical for Processing

The leucines at positions P1’ and P2’ emerged as critical to ZMPSTE24 processing of prelamin A. Hydrophobic residues are preferred at these positions. As discussed above, L647R has long been known to disrupt processing and has been reported in a patient with a progeroid disorder (Hennekes & Nigg, 1994; Wang et al., 2016). I confirmed this result using the humanized yeast assay, with the L647R mutation having 5% cleavage efficiency relative to wild-type (Fig. 4a and Fig. 5). Comprehensive mutagenesis shows that other charged residues, polar residues, proline, and glycine also nearly completely abolish processing at L647. L648 shows a nearly identical pattern (Fig. 4b and Fig. 5) with several notable differences. The bulky hydrophobic residues tyrosine and tryptophan are disruptive at L647 (P1’) but tolerated at L648 (P2’). Substitution with alanine is tolerated at L647 but disruptive at L648, which had been seen previously in a limited mutagenesis scan in mammalian cells (Barrowman et al., 2012). Substitution of proline and glycine at both L647 and L648 nearly abolish ZMPSTE24 processing while these substitutions are tolerated at all other positions flanking the cleavage site.

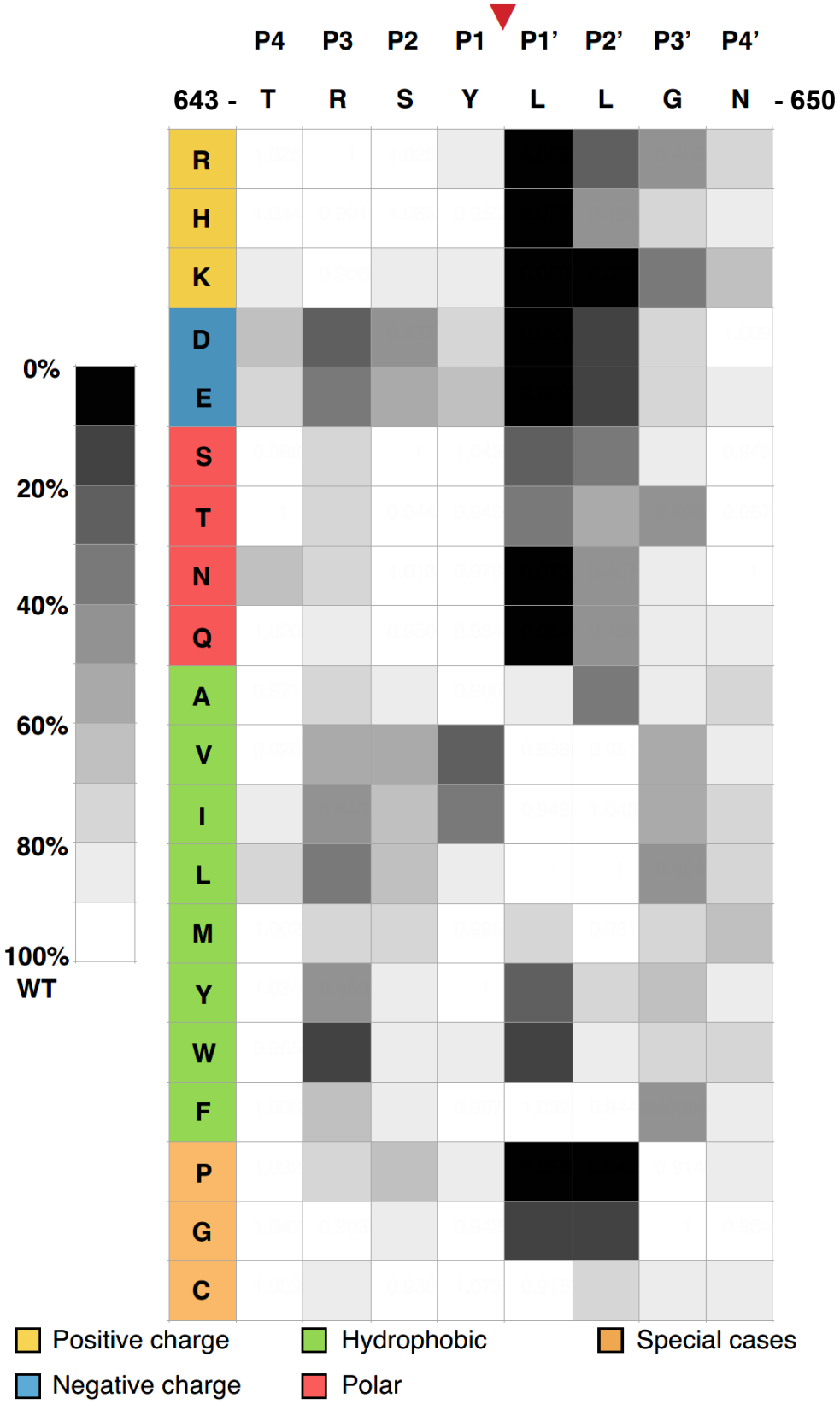
P4
P3
P2
P1
P1'
P2'
P3'
P4'
643 -
T
R
S
Y
L
L
G
N
- 650
R
H
K
D
E
S
T
N
Q
A
V
I
L
M
Y
W
F
P
G
C
0%
20%
40%
60%
80%
100%
WT
Positive charge
Hydrophobic
Special cases
Negative charge
Polar

Figure 5: Heatmap showing the ZMPSTE24 cleavage efficiency of prelamin A for every individual amino acid substitution in residues P4-P4'. The P4-P4' region surrounding the prelamin A cleavage site was comprehensively interrogated by positional scanning mutagenesis, as shown in Figs. 2 and 3. The effects of individual amino acid substitutions at each site were scored relative to wild-type at 100% cleavage efficiency and assigned into ten bins. Residues are color-coded as indicated. L647 (P1') and L648 (P2') are particularly sensitive to mutation, with charged and polar substitutions, proline, and glycine nearly abolishing processing. L647 is also sensitive to the large aromatic residues tyrosine and tryptophan.

Multiple Evolutionarily Conserved Substitutions at the Prelamin A Cleavage Site are Tolerated by ZMPSTE24

We were interested in whether multiple combinatorial amino acid substitutions at the prelamin A cleavage site would be tolerated by ZMPSTE24. *LMNA* is an evolutionary conserved gene and the prelamin A polypeptide shows evolutionary conservation through vertebrates. We performed a protein sequence alignment for several mammals, birds, fish, and amphibians, demonstrating evolutionary conservation of the C-terminus of prelamin A (Fig. 6a). Notably the CAAX motif is nearly perfectly conserved through species, with the farnesylated cysteine invariantly 4 residues from the C-terminus for all species.

Next, to ascertain the extent to which human ZMPSTE24 can cleave these divergent putative cleavage sites, we introduced amino acid substitutions in positions P4 through P4′ in our prelamin A expression construct to recapitulate the putative cleavage sites observed in our multi-species alignment. For the eight residues flanking the cleavage site these ranged from three to seven amino acid substitutions (Fig. 6b).

From the heatmap data of Figure 4, each single amino acid substitution from the multispecies alignment achieved a cleavage efficiency of at least 70%, with a majority of the substitutions exceeding 90% (Fig. 6b), indicated by the shaded boxes underneath the amino acid substitutions indicated in figure 6b. In general, the putative cleavage sites all conformed to the rules defined by the heatmap data, particularly the requirement of having hydrophobic residues at positions P1′ and P2′. No cleavage-blocking substitutions, *e.g.* R644W, L647R, L647P, were present in the divergent putative cleavage sites.

In the cleavage assay, all of the putative prelamin A cleavage sites were recognized and processed by ZMPSTE24, with only modest defects relative to the wild-type human cleavage site (Fig. 5). The zebrafish sequence, with seven of the eight cleavage site residues substituted, was the only divergent cleavage site with a statistically significant ($p<0.05$, Student's T-test, two-tailed, unpaired, N=3) cleavage defect, although even in this case a substantial amount of cleavage did occur. In general, each putative cleavage site performed similarly in the cleavage assay to the weakest of any single substitution within the divergent site. ZMPSTE24 can tolerate multiple substitutions at the cleavage site as long as any changes conform to the rules defined by the heatmap data of Figure 5.

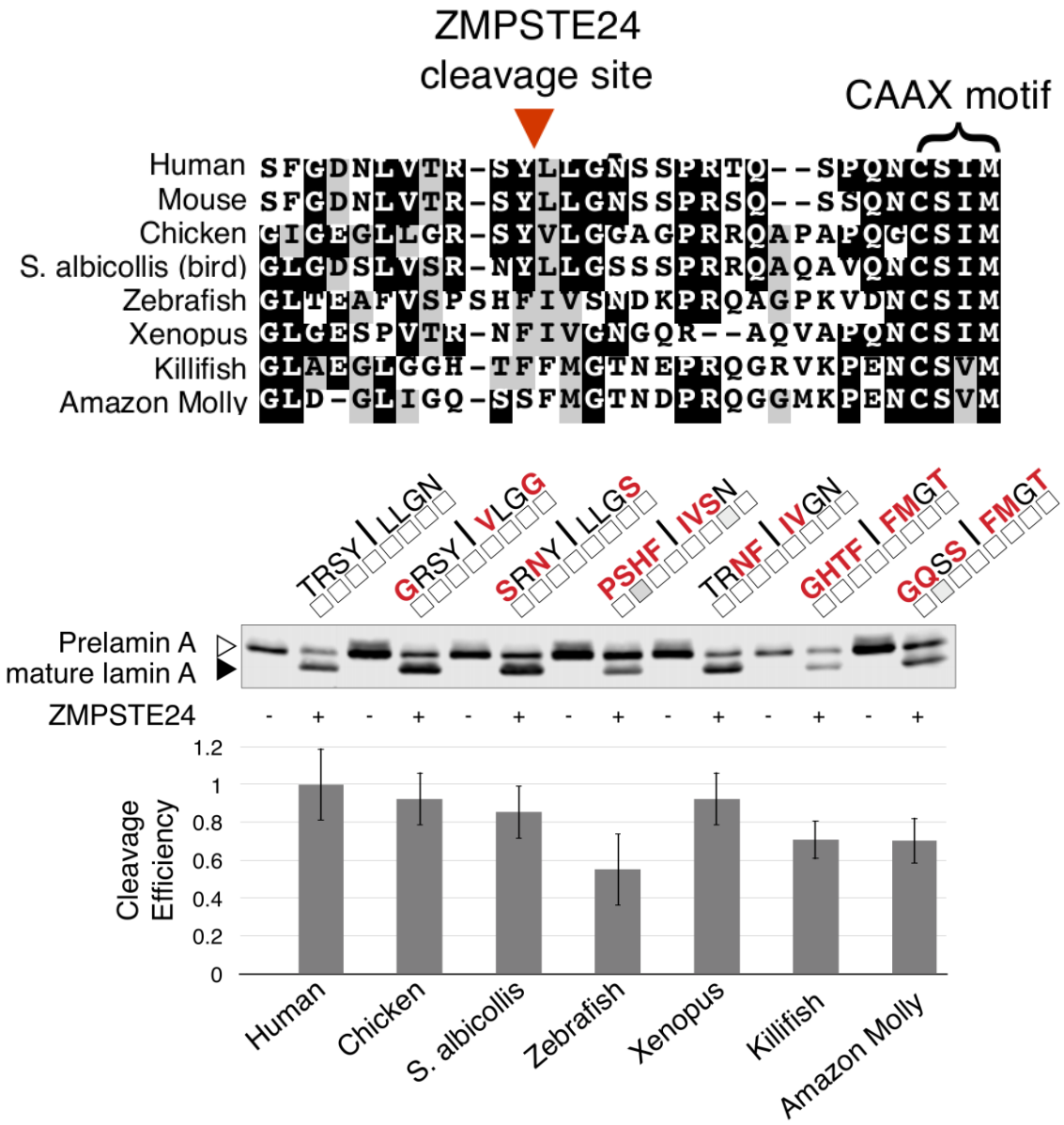

Figure 6: Multi-species analysis of ZMPSTE24-dependent cleavage of prelamin A. (A) Multi-species alignment of the amino acid sequence of the C-terminus of prelamin A. The inferred prelamin A cleavage site and CAAX motif are indicated. Black shading indicates identity and gray shading indicates similarity. (B) Western blots for inferred multi-species variants of the prelamin A cleavage site. Amino acid substitutions made for each species in plasmid pSM3393 are indicated above the western blot, and changes are highlighted in red. The human sequence is that used in the previous figures in plasmid pSM3393. Uncleaved prelamin A (empty arrowhead) and mature lamin A (filled arrowhead) are indicated. Each construct was assayed with (+) and without (-) ZMPSTE24 co-expression. Quantification was performed as in previous figures.

Discussion and Conclusions

Here we have used a humanized yeast assay to ascertain amino acid sequence requirements for cleavage of prelamin by ZMPSTE24. Several high-throughput methods employing degenerate synthetic peptide substrate libraries or phage display libraries have been used to ascertain cleavage site specificity for many diverse proteases (Vidmar et al., 2017). However, ZMPSTE24 presents a unique challenge to these approaches. ZMPSTE24 processing *in vivo* is dependent on farnesylation, a technical hurdle to generating synthetic substrates. *In vitro* farnesylation is not required but improves fidelity. Farnesylation may be needed to improve access to ZMPST24 or to engage a hydrophobic exosite binding site on chamber or interior. In our humanized yeast assay, the first steps of prelamin A biogenesis—farnesylation, -AAXing, and carboxymethylation—are performed by native yeast enzymes, yielding a suitable farnesylated substrate for ZMPSTE24.

Screening our collection of 152 cleavage site mutants in this *in vivo* assay yielded a heatmap of cleavage efficiency (Fig. 5). The mutation of leucine to arginine at the P1’ position serves as a benchmark for the assay, as this mutation has long been recognized as uncleavable. In our assay, we observe approximately 5.6% cleavage compared to wild-type (Figs. 4a and 5).

The two leucines at positions P1’ and P2’ are critical to recognition of prelamin A by ZMPSTE24. Only hydrophobic substitutions were tolerated at both of these positions. Charged and polar residues were very disruptive as were proline and glycine. Hydrophobic and aromatic residues are required in these

positions likely because of interactions with the hydrophobic groove of the ZMPSTE24 active site. Notably, bulky hydrophobic aromatic residues tyrosine (Y) and tryptophan (W) are disruptive at position P1' but not at P2'. This is in contrast to thermolysin, which requires aliphatic or aromatic residues at P1' but has no preference at position P2'.

The arginine at position P3 is also sensitive to substitution, with negatively charged residues, hydrophobic residues, and particularly tryptophan disrupting cleavage. This residue is of particular interest to us because of the association of variant R644C with diverse phenotypes including cardiovascular disease, muscular dystrophy, and skeletal abnormalities (Rankin et al., 2008). The variant shows incomplete penetrance in families and has been detected as a particularly common variant in the ExAC database (Lek et al., 2016). Previously, we observed a significant cleavage defect for this variant in a mammalian expression system (Barrowman et al., 2012). However, in the present study using our humanized yeast assay, R644C shows approximately 80% cleavage efficiency compared to wild-type. The impact of R644C on prelamin A processing, and whether the associated phenotypes are due to an accumulation of prelamin A in cells of patients harboring this mutation, remains unclear. R644H has also been reported to be associated with disease. But disease causality has not been proven (Mercuri E, Bonne G PMID: 15770669), and in our study this amino acid change does not affect processing. Overall, the findings for changes in residue R644 (position P3) are ambiguous in contrast to the L647R (position P1')

mutations, where we see a severe processing defect in this assay and the patient with this mutation showed an accumulation of prelamin A (Wang et al., 2016).

ZMPSTE24 is able to recognize and cleave evolutionarily divergent prelamin A cleavage sites from several species (Fig. 6). None of the substitutions of these putative cleavage sites violate the rules suggested by our heatmap, and indeed all of the sites we introduced were cleaved to a level 50-100% that of wild-type. The rules established here are not specific enough to be predictive of possible substrates, but it is notable that in this assay ZMPSTE24 is recognizing and cleaving several divergent substrate sequences that conform to the rules indicated by our heatmap.

We note that this assay reports steady-state cleavage efficiency in whole-cell lysates and does not measure the kinetics or rates of processing of prelamin A by ZMPSTE24. Also, this assay does not ascertain the exact cleavage site, although gel migration of the cleaved product is consistent with cleavage occurring at or near the predicted cleavage site. The scissile bond that is cleavage by ZMPSTE24 is between residues Y646 and L647. Mutation of protease cleavage recognition sequences can result in shifting the cleavage site (Vidmar, 2017). We cannot exclude this possibility for our mutants that show successful cleavage. Proof that the canonical cleavage site is being used would require sequencing N-terminal fragments of processed prelamin A.

A final caveat for the interpretation of our cleavage site mutant data is that these prelamin A mutants expressed in our assay are presumed to be farnesylated. ZMPSTE24 cleavage of the 15-amino acid C-terminus is dependent

on farnesylation. We have not excluded the possibility that the cleavage site mutants tested here are affecting farnesylation and that the cleavage defects we see are due to non-farnesylated prelamin A substrate. To our knowledge, no cleavage site mutations of positions P4 through P4′ have been reported to affect the preceding post-translational processing steps (-AAXing, farnesylation, and carboxymethylation) and the L647R “uncleavable” prelamin A mutant protein maintains its membrane localization (Hennekes and Nigg), providing indirect evidence that it is fully farnesylated. Furthermore, numerous mutations in yeast a-factor and Ras proteins affecting residues nearby and upstream of the CAAX motif do not appear to affect farnesylation, making it unlikely that the mutations studied here (all of at least 12 residues upstream of the CAAX motif) impact farnesylation. The C661S mutant does block farnesylation and is therefore uncleaved and has been used as a standard in the field for non-farnesylated substrate (Spear et al., 2018). Farnesylation and carboxymethylation could in principle be confirmed for select mutants by probing an immunoblot with a farnesyl-specific prelamin A antibody (Dominici et al., 2009) in the absence of ZMPSTE24. The farnesyl-specific antibody is said to detect the -AAXed, farnesylated, and carboxymethylated C-terminus of prelamin A and would confirm that the requisite post-translational modifications are proceeding normally in the context of our cleavage site mutants. However, unfortunately, attempts to successfully use this antibody by our lab and others have been unsuccessful. It also may be possible ultimately that analysis of gel mobility shifts, using

specialized SDS-PAGE conditions and/or analysis by mass spectrometry could be used to confirm farnesylation of our constructs.

ZMPSTE24 forms a chamber in the membrane, with an active site within the chamber that resembles that of the soluble M48 protease thermolysin. Substrate specificity of ZMPSTE24 could be regulated by access to the chamber, positioning within the chamber and active site, amino acid recognition sequence requirements for catalysis, or combinations thereof. In this study I specifically examined the requirement for the eight residues flanking cleavage site/scissile bond. I was particularly interested in how, despite its being in the membrane, ZMPSTE24 compared to thermolysin. Thermolysin is a thermostable protease that is generally non-specific, with only a requirement of aliphatic and aromatic amino acid residues at position P1' (Vidmar, Vizovi Sek, San Turk, Turk, & Fonovi, 2017). Interestingly, we found that like thermolysisin, ZMPSTE24 is also relatively non-specific in terms of a consensus recognition site, but nevertheless has amino acid preferences, with non-allowable residues at several positions, as discussed above.

Prelamin A is the only known mammalian substrate for ZMPSTE24, which is somewhat unexpected for several reasons. Proteases range in substrate specificity from generalized non-specific protein degradation to targeted catalysis of specific substrates. But most proteases have more than one substrate. ZMPSTE24 can perform both the -AAXing step and N-terminal cleavage of yeast a-factor, a farnesylated substrate with a cleavage site that is different than the prelamin A cleavage site. This suggests that there is some amount of tolerance

for ZMPSTE24 substrate selection and raises the possibility of other prenylated proteins being candidate substrates for ZMPSTE24. But what causes ZMPSTE24 to cleave prelamin A and avoid other farnesylated substrates, for example B-type lamins, remains unknown.

ZMPSTE24 has additional functions in the cells beyond its role in lamin A biogenesis. In the ER, ZMPSTE24 acts as a non-specific translocon declogger, cleaving folded substrates that disrupt translocation to the ER lumen (Ast et al., 2016; Kayatekin et al., 2018). This function is dependent on its proteolytic activity. In contrast, ZMPSTE24 also has an emerging role in restriction of several viruses (Fu, Wang, Li, & Dorf, 2017; Li, Fu, Wang, & Dorf, 2017), a function independent of its protease activity. My work here, defining amino acid substrate requirements of ZMPSTE24, will inform future studies of the diverse roles of this protease in human biology and disease.

I conclude that ZMPSTE24 is a zinc metalloprotease that requires a farnsylated substrate with a suitable cleavage site, possibly at an allowable distance from the farnesyl group. It requires hydrophobic residues in positions P1’ and P2’ and also has preferences at position P3, and beyond that has tolerance to recognize and process varying cleavage sites that follow these rules. Other factors such as portal-restricted access to the chamber, orientation of the substrate within the chamber, and conformational interaction with the hydrophobic groove of the active site may also contribute to its substrate specificity.

Materials and Methods:

LMNA Plasmids and Mutagenesis

All *LMNA* mutants in this study are variants of plasmid pSM3393, a *CEN6 HIS3* plasmid derived from pRS313 (Sikorski & Hieter, 1989) which contains the *PRC1* promoter, human $LMNA_{CT}$ (corresponding to amino acids 431-664) that is N-terminally *His_{10}-myc3* tagged, and the 3' non-coding region from *CYC1* (Spear et al., 2019). The collection of 152 *LMNA* substitution mutants and 6 multispecies variants assayed in this study were generated by QuikChange™ mutagenesis (Stratagene, San Diego, CA) with degenerate oligonucleotides or, when necessary, by individual mutagenic PCR reactions combined with NEBuilder® HiFi Assembly (New England Biolabs, Ipswich, MA) Plasmid pSM3393 was the template for all reactions.

For QuikChange™ mutagenesis, antiparallel oligonucleotides were designed to contain a degenerate codon, NNK (N represents G, T, A, or C and K represents G or T), at the target codon. Mutagenic QuikChange™ reactions were performed on template plasmids that had been deleted for the target codon, to reduce the incidence of wild-type amplification products. QuikChange™ reactions were *Dpn*I-treated, transformed into NEB® 5-alpha competent *E. coli* cells, and selected on carbenicillin-containing LB media. Plasmid DNA was prepared from individual colonies and individual mutations were identified and confirmed by Sanger sequencing.

For each position, (P4-P4'; lamin A residues 643-650) nearly all amino acid substitutions were generated by the degenerate oligonucleotide approach. However, when a particular mutation was missing after degenerate oligonucleotide mutagenesis, antiparallel QuikChange™ primers were individually designed and used with outside primers to generate mutation-containing PCR products for *in vitro* assembly with the pSM3393 backbone. The combination of degenerate oligonucleotide mutagenesis and individual PCR-based mutagenesis yielded the complete collection of 152 lamin A cleavage site mutants. The multispecies variants analyzed in Fig. 6 were made by QuikChange mutagenesis, using antiparallel oligonucleotides designed to contain the multiple substitutions of each species' putative lamin A cleavage site.

The *LMNA* variants on pSM3393 were individually transformed into the yeast strain SM6303, or in some cases SM4826, by standard lithium acetate protocols (Kaiser *et al.*, 1994), selected on minimal SC-His plates, and transformants were re-streaked to isolate single colonies.

Yeast Strains Used in this Study

Yeast strain SM4826 (*ste24Δ::kanMX his3 leu2 met15 ura3*) is from the BY4741 deletion collection and is deleted for the yeast *ZMPSTE24* homolog *STE24*. It was used to generate the ZMPSTE24-expressing yeast strain SM6303 in the following way: pSM3428 is an integrating plasmid derived from the pRS304 vector (Sikorski & Hieter, 1989) that contains a nourseothricin resistance marker and full-length human ZMPSTE24 that is codon-optimized for expression

in yeast (kindly provided by Mark Dumont (ref Clark and Dumont paper) , has an N-terminal *$His_{10}HA_3$*-tag, and is driven by the *PGK1* promoter. To chromosomally integrate this *ZMPSTE24* plasmid at the *TRP1* locus, plasmid DNA was linearized by *Eco*RV digestion, transformed into SM4826 cells and transformants were selected for growth on yeast extract peptone dextrose (YPD) plates containing 100 μg/mL nourseothricin. The resulting strain SM6303 has 2 integrated copies of pSM3428 and thus 2 copies of ZMPSTE24.

Prelamin A Cleavage Assay

Overnight cultures were grown at 30°C to saturation in minimal medium (0.67% yeast nitrogen base, 0.5% ammonium sulfate, 2% glucose, supplemented with amino acids), back diluted 1:1000, and cells were grown for 14-16 hours to an OD_{600} of approximately 1.0. Cells (1.0 OD_{600} cell equivalents) were pelleted, washed with water, treated with 0.1 N NaOH, and lysed in SDS-PAGE sample buffer at 65° C for 10 minutes. Lysates were vortexed, pelleted, and supernatants (0.24 OD_{600} cell equivalents per lane) were resolved on 10% SDS polyacrylamide gels for 90 minutes at 150V. Proteins were transferred to nitrocellulose membranes using the Trans-Blot® Turbo™ system (Bio-Rad Laboratories, Hercules, CA) and blocked with Western Blocking Reagent (Roche, Indianapolis, IN).

Lamin A proteins were detected by probing with mouse anti-myc antibodies (clone 4A6, Millipore cat #05-724; 1:10,000 dilution) decorated with goat anti-mouse secondary IRDye 680RD antibodies (LI-COR, Lincoln, NE). The

loading control hexokinase was detected using and rabbit anti-hexokinase (1:200,000 dilution) antibodies decorated with goat anti-rabbit secondary IRDye 800CW (LI-COR) antibodies were used as a loading control. Western blots were visualized using the Odyssey® imaging platform (LI-COR) and bands were quantified using Image Studio™ software (LI-COR). Cleavage efficiency was calculating by dividing mature lamin A signal by total lamin A signal (processed + unprocessed) and normalized to WT processing efficiency which was set to 1.0. Blots were reprobed using rat anti-HA (clone 3F10, Roche cat #11867423001; 1:10,000 dilution) to detect ZMPSTE24 and visualized using goat anti-rat IRDye 680RD.

Multispecies Alignment

Protein sequences for the species studied were acquired from UniProt (UniProt Consortium, 2017) and aligned using ClustalX (Larkin et al., 2007). Putative cleavage sites for divergent species were inferred from conservation and substitution similarity and introduced into pSM3393 via mutagenic PCR and assembly as described above. Resulting constructs were transformed into yeast strains SM6303 and SM4826 and assayed as described above.

Chapter III

Technical Challenges to Developing the Yeast Model System

Our humanized yeast assay, methods and results of which are discussed in detail in chapter II, is a rapid, cost-effective, and sensitive method to examine human prelamin A cleavage by human ZMPSTE24 in yeast. In my thesis work, I was interested in comprehensively interrogating the effects of amino acid substitutions at the residues flanking the cleavage site, to define amino acid sequence requirements for recognition of the prelamin A substrate by the protease, ZMPSTE24. When I began this work, the lab had adapted the humanized yeast assay for modeling human *ZMPSTE24* disease alleles but had done little work studying mutations of *LMNA* using this assay. The earliest stages of my work included optimizing the assay for interrogating *LMNA* mutations, in particular, cleavage site mutations. The goals of my work also included generating many more *LMNA* mutants than the collection of *ZMPSTE24* mutants that had been studied in the lab. The following chapter discusses technical challenges I encountered, methods used to overcome these challenges, and informative lessons for using this assay to study *LMNA* mutations.

Saturation Mutagenesis of the Prelamin A Cleavage Site

I sought to comprehensively mutagenize the residues flanking the scissile bond between the tyrosine and leucine residues at positions 646 and 647, respectively, of prelamin A. The Schechter-Berger nomenclature of protease-substrate interaction (Schechter & Berger, 1967) prescribes the amino acid positions of the substrate C-terminal to the scissile bond as P1', P2', and so forth, and the positions N-terminal to the scissile bond as P1, P2, and so forth. Likewise, the positions of the protease corresponding to the substrate residues are referred to as S1', S2', and S1, S2, respectively.

The decision of how far C-terminal and N-terminal from the scissile bond to study is, to some extent, arbitrary. It is an empirical question: mutations of which residues flanking the scissile bond will have an impact on cleavage efficiency and contribute to an understanding of substrate requirements of the protease? We did have some guidance from other proteases and previous studies of ZMPSTE24 and LMNA to inform our experimental design.

The active site of ZMPSTE24 on the interior of the water-filled chamber is defined by a hydrophobic groove, strikingly similar in structure to the active site of thermolysin, a well-characterized bacterial zinc metalloprotease (Quigley et al., 2013). The P1' position is essential for substrate recognition by thermolysin, with a rather strict requirement of an aliphatic residue at this position, particularly valine, leucine, or isoleucine (Vidmar et al., 2017).

Rhomboid proteases are a family of intramembrane proteases ubiquitous across almost all species (Urban & Dickey, 2011). Rhomboids are unique among

proteases in that their active sites reside within the lipid bilayer of membranes and they cleave transmembrane substrates within their transmembrane spans. Rhomboid functions have been studied extensively (Baker & Urban, 2012), and a comprehensive mutagenesis study revealed that substrate positions P4, P1, and P2' are critical to the function of several rhomboid proteases (Strisovsky, Sharpe, & Freeman, 2009).

For prelamin processing by ZMPSTE24, the leucine residue at position 647, *i.e.* substrate position P1', has long been known to be essential. In chicken, mutation of the homologous leucine residue to arginine was shown to be uncleavable (Hennekes & Nigg, 1994). This L647R substitution has been used in the lamin A field as a marker for uncleavable prelamin A, and our group previously demonstrated this in a mammalian expression system (Barrowman et al., 2012). Finally, the L647R mutation was recently reported in a patient with a progeroid disorder, and the cells of this patient were shown to accumulate uncleaved prelamin A, indicating that this mutation abolished prelamin A processing by ZMPSTE24 in human cells (Wang et al., 2016).

The arginine residue at position 644, *i.e.* substrate position P3, also has clinical significance. The R644C mutation has been associated with disease phenotypes including lipodystrophy, cardiomyopathy, and skeletal abnormalities. However, pedigrees for families harboring this mutation show incomplete penetrance, variable expressivity, and striking phenotypic variability (Rankin et al., 2008). Our group observed that R644C caused a modest prelamin A processing defect in a mammalian expression system (Barrowman et al., 2012).

For our decision of which residues to comprehensively mutate and test in our humanized yeast assay, we needed to define a putative ZMPSTE24 recognition site. Taken together, all the indications discussed above, as well as time and cost considerations, led us to narrow our study to positions P4 through P4', *i.e.* positions 643 through 650.

Next we turned to degenerate oligonucleotide mutagenesis to generate a collection of *LMNA* mutants representing all possible variants of residues P4 through P4'. For each position I designed antiparallel QuikChange primers that contained the degenerate codon NNK at the target residue (N represents any nucleotide and K represents either G or T) (Fig. 7). The NNK degenerate codon has the advantage that it encodes all twenty possible amino acids but reduces the occurrence of stop codons to only one.

```
643 644 645 646   647 648 649 650

ACC CGC TCC TAC | CTC CTG GGC AAC
 T   R   S   Y     L   L   G   N

NNK CGC TCC TAC | CTC CTG GGC AAC
ACC NNK TCC TAC | CTC CTG GGC AAC
ACC CGC NNK TAC | CTC CTG GGC AAC
ACC CGC TCC NNK | CTC CTG GGC AAC
ACC CGC TCC TAC | NNK CTG GGC AAC
ACC CGC TCC TAC | CTC NNK GGC AAC
ACC CGC TCC TAC | CTC CTG NNK AAC
ACC CGC TCC TAC | CTC CTG GGC NNK
```

Figure 7: A schematic of degenerate codons introduced at positions P4 through P4' of the ZMPSTE24 cleavage site is shown. The NNK codon was introduced at

each position to generate a collection of every possible amino acid substitution while minimizing the incidence of stop codon mutations.

I started by using these QuikChange oligonucleotides to mutagenize pSM3167, a yeast integrating plasmid encoding the 10His-3myc-LMNA(431-664) expression cassette discussed in Chapter II. QuikChange mutagenesis was performed according to standard conditions discussed in Chapter II.

Of the eight residues I sought to mutagenize and characterize using the humanized yeast assay, I began with the arginine at 644 in the P3 position due to its clinical significance. After the QuikChange reaction, transforming into *E. coli*, picking colonies, preparing plasmid DNA, and sequencing to identify mutant clones, I was disappointed to detect only two mutants out of 48 colonies sequenced. Clone “644_28” was a CGC to CCT mutation resulting in R644P, and clone “644_46” was a CGC to TCG mutation resulting in R644S. The purpose of using degenerate oligonucleotide mutagenesis was to minimize the number of mutagenic reactions performed. Ideally, only one mutagenic reaction for each residue would result in a pool of mutants representing every possible amino acid substitution for each residue. The initial ~4% mutation rate that I observed, however, negates this advantage and would not have been a viable method for generating the extensive collection of mutants that I needed.

We hypothesized that the primers in the degenerate oligonucleotide mix that were complementary to the wild-type sequence were outcompeting the primers the contained mismatched mutations, resulting in an extreme enrichment of wild-type products. My mutagenesis protocol included an overnight *Dpn*I

digestion of the QuikChange reaction to eliminate wild-type template, and my negative control plates typically had zero or only a handful of colonies, compared to hundreds of colonies on the mutagenesis plates. I suspected that the QuikChange reactions were successful, but the wild-type primers in the mix were producing the vast majority of products.

At a lab meeting early in my project, I presented my strategy and these preliminary data. My labmate Khurts Shilagardi suggested deleting the target codon in the plasmid template, so that there would be no wild-type target to produce wild-type products in the QuikChange reaction, thereby increasing the proportion and diversity of mutant products. I found this to be a simple and elegant solution and went ahead with it right away. For each residue, positions 643 through 650, I used QuikChange Mutagenesis on pSM3167 to delete the target codon. These deletions were confirmed by Sanger sequencing and subsequently used as templates for the degenerate oligonucleotide mutagenesis approach described above.

With this modified approach my numbers improved dramatically. For position 644 using the ΔR644 template, I identified 15 of the possible 19 substitutions after sequencing 48 clones. Other residues showed a similarly high numbers of mutations. There were redundancies and repeats and as expected I observed diminishing returns of mutants as I sequenced increasing numbers of colonies. Usually after sequencing 48 colonies, I had achieved 12-18 of the possible substitutions. Because of sequencing costs and these diminishing

returns, remaining mutants were made individually with QuikChange primers designed for those particular mutations.

This strategy of degenerate oligonucleotide mutagenesis on a plasmid template deleted for the target codon allowed me to efficiently generate a collection of 192 mutants to evaluate in the humanized yeast assay. I thank Khurts for the smart advice in that early lab meeting.

Alternate Versions of the Humanized Yeast Assay

There are several approaches available for expressing heterologous human proteins in yeast. Previously in the lab, human prelamin A had been introduced to yeast on an integrating vector, and human ZMPSTE24 was expressed on a yeast centromeric plasmid. I began my comprehensive mutagenesis work on this yeast integrating plasmid, pSM3197 (Fig. 8). Based on pRS304, pSM3197 contained the tagged C-terminal portion of prelamin A 10His-3myc-LMNA(431-664) driven by the TDH3 promoter, the *HIS3* selectable marker, flanked by two homology arms target toward the promoter and terminator sequences of the *HO* locus of the yeast genome. These homology arms mediate a recombination event during yeast transformation and integrate the intervening sequence as a gene replacement. The integrating portion of the plasmid was liberated by digestion with *Eco*RI and *Sph*I, gel purified, and transformed into yeast carrying the ZMPSTE24-expressing centromeric plasmid (pSM2677), and selected on SC-HIS plates.

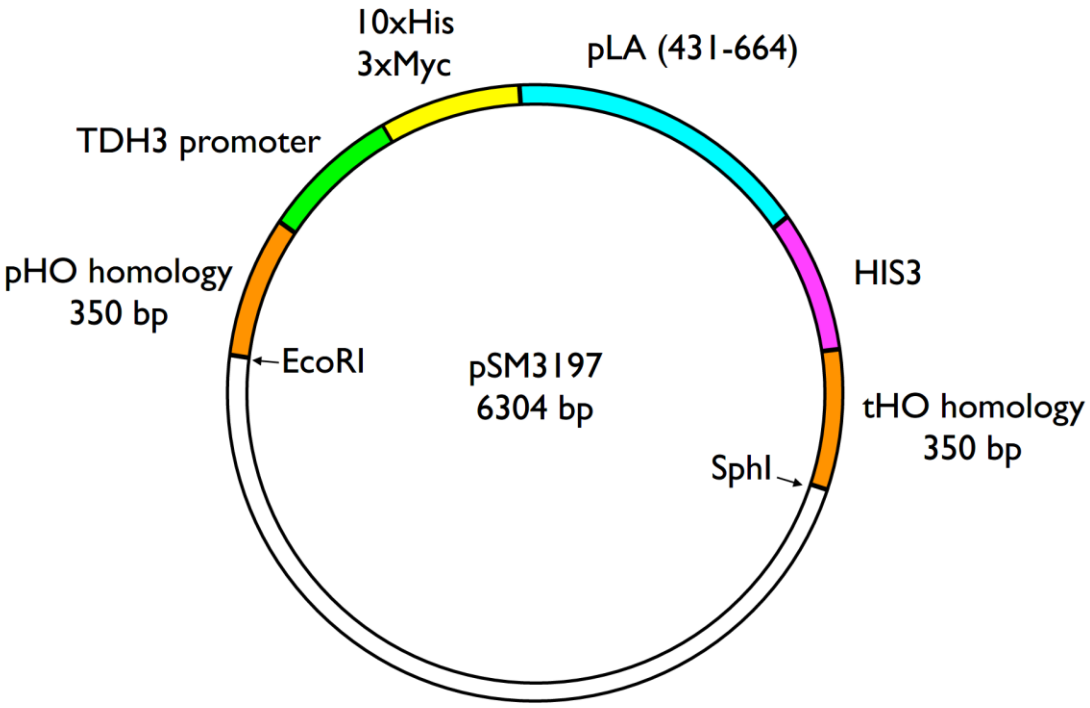

Figure 8: plasmid pSM3197 is a yeast integrating vector that encodes the 10His-3myc-LMNA(431-664) expression cassette and integrates as a single copy at the *HO* locus of the yeast genome, mediated by two ~350 bp homology arms.

Quantifying prelamin A processing by Western blot in this system is the basis of our humanized yeast assay. An accurate measure of cleavage of prelamin by ZMPSTE24 relies on consistent stoichiometry between the protease and the substrate, *i.e.* consistent expression levels for both heterologous proteins in yeast.

As I began to evaluate my growing collection of prelamin A cleavage site mutants, I started to notice dramatic inconsistencies in prelamin A expression among transformants (Fig. 9). Some clones showed little to no tagged prelamin A signal on Western blot, while others showed strong tagged prelamin A signal. I

also observed significant colony size heterogeneity on my selection plates after transformation with the integrating prelamin A DNA (Fig. 10).

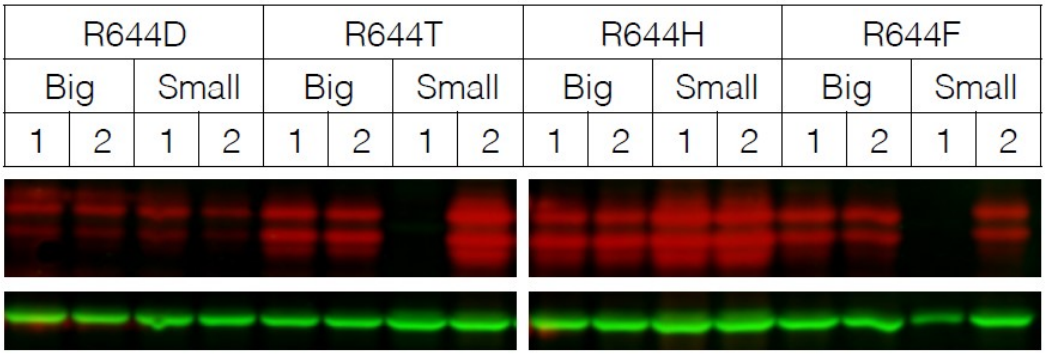

Figure 9: Western blots probed for myc-tagged prelamin A of lysates prepared from cultures of individual transformants are shown for four R644 mutations. For each mutation, two “Big” colonies and two “Small” colonies were selected. For R644T, Small colony #1 shows no prelamin A signal, while Small colony #2 shows higher prelamin A signal than either of the Big colonies. Similar inconsistencies in prelamin A signal were observed for other R644 mutations.

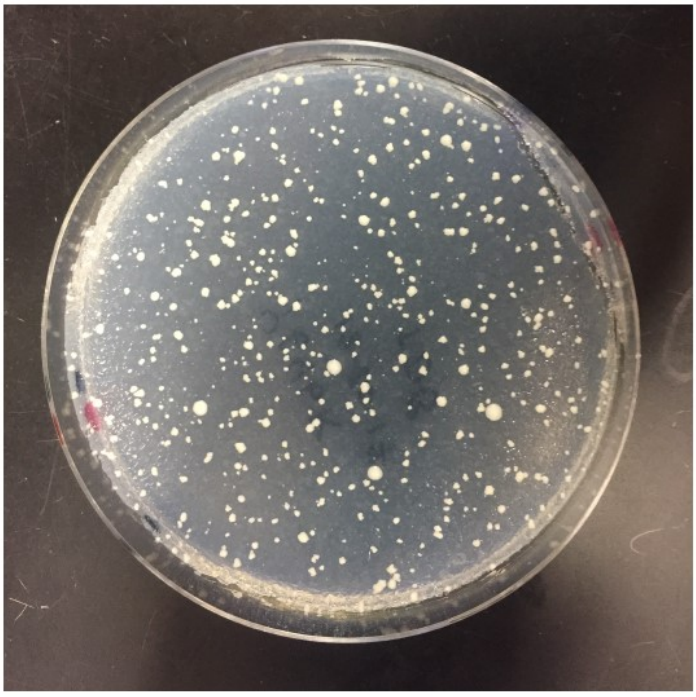

Figure 10: Colonies of $His^{+}Ura^{+}$ transformants expressing the integrated prelamin A expression cassette (pSM3197) and ZMPSTE24 from a centromeric plasmid (pSM2677) are shown. Significant colony size heterogeneity was observed for each transformation plate across several R644 mutations, as pictured here.

The most troubling aspect of this unpredictability of prelamin A expression in my assay was that colony size and prelamin A expression level seemed to be having an impact on the readout of prelamin A cleavage efficiency (Figs. 9 and 11). The advantage, and purpose, of our assay was examining the effect of cleavage site mutations on prelamin A cleavage efficiency in an isogenic background, with any change in prelamin A cleavage efficiency owing directly to the cleavage site mutation introduced to the assay. Having inter-clone variability in the assay was not acceptable, and I needed to either ascertain the source of this variability or change the assay.

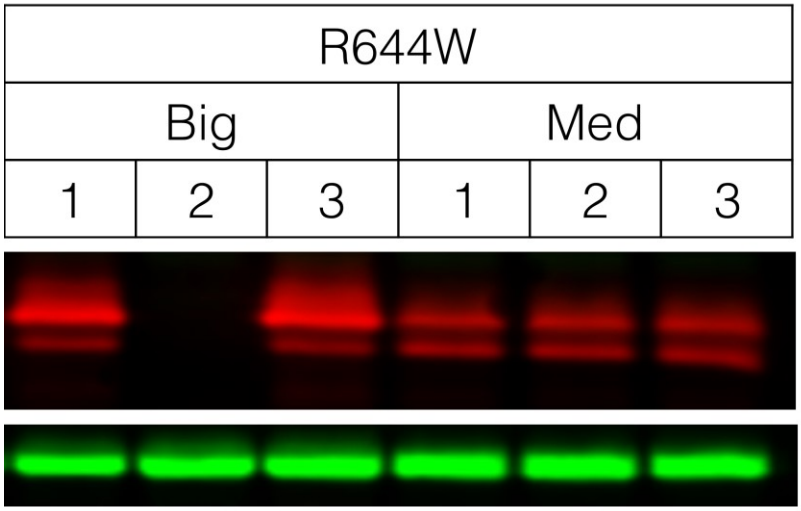

Figure 11: Western blot of six individual transformants expressing prelamin A with an R644W mutation and WT ZMPSTE24. Lysates made from cultures of three “Big” colonies (1-3) and three “Medium” colonies (1-3) are shown. Big colony #2 shows no prelamin A signal, while Big colonies #1 and #3 show strong prelamin A signal with a significant cleavage defect. Medium colonies 1-3 show lower levels of prelamin A expression than the Big colonies, and a higher level of prelamin A processing estimated by dividing cleaved signal over total signal. Similar inconsistencies in perceived prelamin A cleavage efficiency were observed for several R644 mutations.

Because of these intractable inconsistencies in prelamin A expression and some other workflow issues, I decided to modify the assay. The lab had previously had success interrogating *ZMPSTE24* disease alleles by integrating the prelamin A expression cassette and introducing *ZMPSTE24* mutations on a centromeric plasmid. For assaying large number of mutations, this also streamlines the workflow, as plasmid DNA can be transformed directly rather than liberating a linear segment of DNA, gel-purifying, and transforming, as I had been doing with my prelamin A integrations. Accordingly, I decided to integrate ZMPSTE24 once, and introduce my collection of prelamin A mutations on a centromeric plasmid. This approach is illustrated in Fig. 2 and Fig. 12.

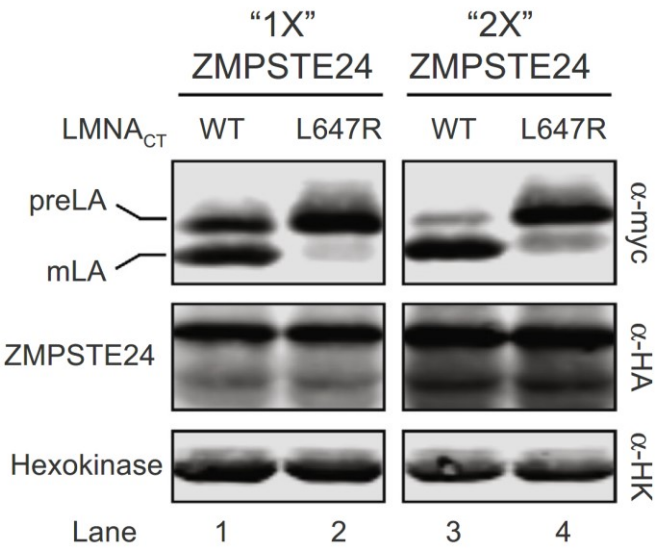

Figure 12: Yeast codon-optimized 10His-3HA- ZMPSTE24 is integrated into a ste24Δ strain at the *TRP1* locus (SM6302/SM6303). LMNA$_{CT}$ variants are then transformed into the yeast strain and selected with SC-his medium. Prelamin A cleavage increases with more ZMPSTE24. SM6302 ("1X ZMP") or SM6303 ("2X ZMP") transformed with WT or L647R LMNA$_{CT}$ were analyzed by western blotting with anti-myc and anti-HA antibodies. Anti-HK (hexokinase) serves as a loading control. Strains with cleavage percentage in parentheses in lanes 1–4 are SM6302/pSM3391 (54.2%), SM6302/pSM3392 (9.9%), SM6303/pSM3391

(79.6%), and SM6303/pSM3392 (15.4%), respectively. Published in Spear *et al.* 2019.

This modification of the assay immediately solved the problems I had been encountering with the previous version of the assay. I observed zero false positives (His^+Ura^+ transformants that showed no prelamin A signal) and prelamin A expression was consistent across multiple transformants for each mutation. Furthermore, the workflow was simplified and allowed for simplified generation and characterization of the 152 mutants that I set out to study.

Changing the assay also revealed a stoichiometric relationship between the amount of protease and substrate. During the *ZMPSTE24* integration at the *TRP1* locus, it was possible to isolate clones with one or multiple integrants. After screening several clones, we saved strains SM6302 and SM6303, with the latter showing higher ZMPSTE24 abundance on Western blot and a corresponding increased in prelamin A cleavage efficiency (Fig. 12). I subsequently performed all my studies of prelamin A cleavage site mutants in strain SM6303.

After working through both of these systems, either integrating prelamin A or ZMPSTE24 and expressing the other from a centromeric plasmid, the latter system was much more reliable for characterizing large numbers of mutants. Both of these systems are described in a Methods publication with contribution from all the members of our lab (Spear et al., 2019).

Chapter IV
Perspectives and Conclusion

In the introduction, I introduced the protease ZMPSTE24, its substrate the intermediate filament protein lamin A, and explained the well-established role of ZMPSTE24 in the proteolytic maturation of lamin A. The successful cleavage of prelamin A by ZMPSTE24 is critical to human health, as when this step fails, the accumulation of uncleaved, farnesylated prelamin A causes severe disease. A significant unknown, however, was the substrate requirements for recognition and cleavage of prelamin A by ZMPSTE24. This thesis makes steps toward a fuller understanding of these requirements. However, there remain unknowns in this protease-substrate relationship and in the field more broadly, representing opportunities for further research.

Prior to the work presented in this thesis, very little was known about substrate requirements for proteolytic processing of prelamin by ZMPSTE24. As discussed in Chapter I, farnesylation was a known substrate requirement, the cleavage site had been identified at the amino acid sequence level, and L647R was known to abolish cleavage. Our group had performed a very limited amount of mutagenesis by arginine/alanine scanning across the cleavage site (Barrowman et al., 2012).

This work offers a detailed and saturated view of amino acid requirements of the prelamin A cleavage site. The heatmap representing the cleavage efficiency data for our collection of cleavage site mutants reveals the rule that ZMPSTE24 requires hydrophobic residues at positions P1′ and P2′ of its farnesylated substrate.

Establishing this rule is a significant contribution to the understanding of the interaction between ZMPSTE24 and its substrate prelamin A. Understanding what is an allowable substrate, and how the protease recognizes that substrate, is essential to understanding the molecular mechanisms of proteolysis. The success or failure of this proteolytic event underlies several severe genetic diseases and may have a role in the cardiovascular decline associated with normal aging. Understanding the mechanisms of this proteolytic event will inform the design and development of therapeutics for these phenotypes.

Several major unknowns remain with respect to ZMPSTE24 and prelamin A and warrant future exploration and study. I was particularly interested in the amino acid requirements for cleavage. In addition to amino acid sequence, other

features of the C-terminus including conformation, flexibility, and distance from the farnesyl to the cleavage site may be requirements for ZMPSTE24 recognition and cleavage. Other members of the Michaelis Lab are making progress toward understanding these other requirements.

One question that has existed in this field for decades is whether there is a role for the cleaved farnesylated C-terminal fragment, whether it persists or is degraded, and where it might go. Why lamin A undergoes several post-translational modifications only to have them cleaved off remains unclear. Lamin C, for example, does not undergo these modifications, and lamin A is apparently dispensable, at least in mice, as evidenced by the healthy lamin C-only mouse (Fong et al., 2006). Investigating the destiny of the cleaved C-terminus is an intriguing idea.

Currently prelamin A is the only known substrate for ZMPSTE24 in this role as a CAAX protease. The work presented here established the rule of hydrophobic residues at P1′ and P2′, but also showed flexibility and non-specificity at residues N-terminal to the cleavage site. In other words, there are rules that ZMPSTE24 will not break, but it does not have a strict consensus sequence. Considering that, along with the emerging role of ZMPSTE24 as a non-specific declogger of translocons in the ER, lead me to suspect that there are other farnesylated substrates that undergo proteolytic processing by ZMPSTE24. Using high-throughput proteomic approaches to identify any additional ZMPSTE24 substrates would be exciting research.

Finally, there are many outstanding questions regarding the pathway from molecular mechanisms to phenotype in patients with HGPS and other progeroid disorders due to prelamin A processing defects. We know that accumulation of farnesylated prelamin A or progerin in cells is toxic and causes disease. We have several molecular markers and cell biological hallmarks of HGPS cells, *e.g.* decreased LAP2α expression and abnormal nuclear morphology. However, there are many breaks in the chain from farnesylated prelamin A, the molecular culprit of these diseases, to phenotype. What exactly is uncleaved prelamin A disrupting in the cell? And how does this cause atherosclerosis in teens with progeria? Does the cardiovascular pathophysiology of progerin in HGPS patients represent acceleration of a normal aging progress or an aberration that merely has an appearance of normal aging? This field is a promising area of research, with therapies and cures as the ultimate goal.

Works Cited

Aebi, U., Cohn, J., Buhle, L., & Gerace, L. (1986). The nuclear lamina is a meshwork of intermediate-type filaments. *Nature*, *323*, 560–564. Retrieved from http://www.nature.com.ezp.welch.jhmi.edu/articles/323560a0.pdf

Agarwal, A. K., & Garg, A. (2006). Genetic Basis of Lipodystrophies and Management of Metabolic Complications. *Annual Review of Medicine*, *57*(1), 297–311. https://doi.org/10.1146/annurev.med.57.022605.114424

Ast, T., Michaelis, S., & Schuldiner, M. (2016). The Protease Ste24 Clears Clogged Translocons. *Cell*, *164*(1–2), 103–114. https://doi.org/10.1016/j.cell.2015.11.053

Baker, R. P., & Urban, S. (2012). Architectural and thermodynamic principles underlying intramembrane protease function. *Nature Chemical Biology*, *8*(9), 759–768. https://doi.org/10.1038/nchembio.1021

Barrowman, J., Hamblet, C., George, C. M., & Michaelis, S. (2008). Analysis of Prelamin A Biogenesis Reveals the Nucleus to be a CaaX Processing Compartment. *Molecular Biology of the Cell*, *19*, 5398–5408. https://doi.org/10.1091/mbc.E08

Barrowman, J., Hamblet, C., Kane, M. S., & Michaelis, S. (2012). Requirements for efficient proteolytic cleavage of prelamin A by ZMPSTE24. *PLoS ONE*, *7*(2), 1–8. https://doi.org/10.1371/journal.pone.0032120

Barrowman, J., & Michaelis, S. (2009). ZMPSTE24, an integral membrane zinc metalloprotease with a connection to progeroid disorders. *Biological Chemistry*, *390*(8), 761–773. https://doi.org/10.1515/BC.2009.080

Bergo, M. O., Gavino, B., Ross, J., Schmidt, W. K., Hong, C., Kendall, L. V, … Young, S. G. (2002). *Zmpste24 deficiency in mice causes spontaneous bone fractures, muscle weakness, and a prelamin A processing defect*. Retrieved from www.pnas.orgcgidoi10.1073pnas.192460799

Capell, B. C., & Collins, F. S. (2006). Human laminopathies: nuclei gone genetically awry. *Nature Reviews Genetics*, *7*(12), 940–952. https://doi.org/10.1038/nrg1906

Capell, B. C., Collins, F. S., & Nabel, E. G. (2007). Mechanisms of cardiovascular disease in accelerated aging syndromes. *Circulation Research*, *101*(1), 13–26. https://doi.org/10.1161/CIRCRESAHA.107.153692

Capell, B. C., Erdos, M. R., Madigan, J. P., Fiordalisi, J. J., Varga, R., Conneely, K. N., … Collins, F. S. (2005). *Inhibiting farnesylation of progerin prevents the characteristic nuclear blebbing of Hutchinson-Gilford progeria syndrome*. Retrieved from www.pnas.orgcgidoi10.1073pnas.0506001102

Clark, K. M., Jenkins, J. L., Fedoriw, N., Dumont, M.E. (2017). Human Caax protease ZMPSTE24 expressed in yeast: Structure and inhibition by HIV protease inhibitors. *Protein Sci. 26*(2):242-257.

Corrigan, D. P., Kuszczak, D., Rusinol, A. E., Thewke, D. P., Hrycyna, C. A., Michaelis, S., Sinesky, M. S. (2005). Prelamin A endoproteolytic processing *in vitro* by recombinant Zmpste24. *Biochem. J.* 387:129-138.

Csoka, A., Cao, H., Sammak, P., Constantinescu, D., Schatten, G., & Hegele, R. (2004). Novel lamin A/C gene (LMNA) mutations in atypical progeroid syndromes. *Journal of Medical Genetics*, *41*(4), 304–308.

https://doi.org/10.1136/jmg.2003.015651

De Sandre-Giovannoli, A., Bernard, R., Cau, P., Navarro, C., Amiel, J., Ne Boccaccio, I., … Lévy, N. (2003). Lamin A Truncation in Hutchinson-Gilford Progeria. *Science (New York, N.Y.)*, *300*(5628), 2055. https://doi.org/10.1126/science.1084125

Dechat, T., Pfleghaar, K., Sengupta, K., Shimi, T., Shumaker, D. K., Solimando, L., & Goldman, R. D. (2008). Nuclear lamins: major factors in the structural organization and function of the nucleus and chromatin. https://doi.org/10.1101/gad.1652708

Dominici, S., Fiori, V., Magnani, M., Schena, E., et al. (2009). Different prelamin A forms accumulate in human fibroblasts: a study in experimental models and progeria. *European Journal of Histochemistry*. 53(1): 43-52.

Eriksson, M., Brown, W. T., Gordon, L. B., Glynn, M. W., Singer, J., Scott, L., … Collins, F. S. (2003). Recurrent de novo point mutations in lamin A cause Hutchinson-Gilford progeria syndrome. *Nature*, *423*, 293–298.

Fong, L. G., Ng, J. K., … Bergo, M. O., Young, S. G. (2006). Prelamin A and lamin A appear to be dispensable in the nuclear lamina. *J. Clin. Invest.* 116(3):743-752.

Fu, B., Wang, L., Li, S., & Dorf, M. E. (2017). ZMPSTE24 defends against influenza and other pathogenic viruses. *The Journal of Experimental Medicine*, *214*(4), 919–929. https://doi.org/10.1084/jem.20161270

Fujimura-Kamada, K., Nouvet, F. J., & Michaelis, S. (1997). A Novel Membrane-associated Metalloprotease, Ste24p, Is Required for the First Step of NH 2-

terminal Processing of the Yeast a-Factor Precursor. *The Journal of Cell Biology*, *136*(2), 271–285. https://doi.org/10.1083/jcb.136.2.271

Genschel, J., & Schmidt, H. H.-J. (2000). Mutations in the LMNA gene encoding lamin A/C. *Human Mutation*, *16*(6), 451–459. https://doi.org/10.1002/1098-1004(200012)16:6<451::AID-HUMU1>3.0.CO;2-9

Gordon, L. B., Kleinman, M. E., Miller, D. T., Neuberg, D. S., Giobbie-Hurder, A., Gerhard-Herman, M., … Kieran, M. W. (2012). Clinical trial of a farnesyltransferase inhibitor in children with Hutchinson-Gilford progeria syndrome. *Proc. Natl. Acad. Sci. U.S.A*, *109*(41), 16666–16671. https://doi.org/10.1073/pnas.1202529109

Gordon, L. B., Shappell, H., Massaro, J., D'Agostino, R. B., Brazier, J., Campbell, S. E., … Kieran, M. W. (2018). Association of Lonafarnib Treatment vs No Treatment With Mortality Rate in Patients With Hutchinson-Gilford Progeria Syndrome. *JAMA*, *319*(16), 1687. https://doi.org/10.1001/jama.2018.3264

Hennekes, H., & Nigg, E. A. (1994). The role of isoprenylation in membrane attachment of nuclear lamins. *Journal of Cell Science*, *107*, 1019–1029. Retrieved from http://jcs.biologists.org.ezp.welch.jhmi.edu/content/joces/107/4/1019.full.pdf

Hildebrandt, E.R., Cheng, M., Zhao, P., Kim, J.H., Wells, L., Schmidt, W.K. (2016). A shunt pathway limits the CaaX processing of Hsp40 Ydj1p and regulates Ydj1p-dependent phenotypes. *eLIFE*. 5:e15899.DOI: 10.7554/eLife.15899.

Kayatekin, C., Amasino, A., Gaglia, G., Boone, C., Florez, J. C., Lindquist

Correspondence, S., … Lindquist, S. (2018). Translocon Declogger Ste24 Protects against IAPP Oligomer-Induced Proteotoxicity In Brief Translocon Declogger Ste24 Protects against IAPP Oligomer-Induced Proteotoxicity. *Cell*, *173*, 62–73.e9. https://doi.org/10.1016/j.cell.2018.02.026

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., Mcgettigan, P. A., Mcwilliam, H., … Bateman, A. (2007). Clustal W and Clustal X version 2.0, *23*(21), 2947–2948. https://doi.org/10.1093/bioinformatics/btm404

Lek, M., Karczewski, K. J., Minikel, E. V, samocha, K. E., Banks, E., Fennell, timothy, … Berghout, J. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature Publishing Group*, *536*(6), 57. https://doi.org/10.1038/nature19057

Li, S., Fu, B., Wang, L., & Dorf, M. E. (2017). ZMPSTE24 Is Downstream Effector of Interferon-Induced Transmembrane Antiviral Activity. *DNA and Cell Biology*, *36*(7), 513–517. https://doi.org/10.1089/dna.2017.3791

Mercuri, E., Brown, S. C., Nihoyannopoulos, P., Poulton, J., Kinali, M., Richard, P., … Muntoni, F. (2005). Extreme variability of skeletal and cardiac muscle involvement in patients with mutations in exon 11 of the lamin A/C gene. *Muscle and Nerve*, *31*(5), 602–609. https://doi.org/10.1002/mus.20293

Michaelis, S., & Barrowman, J. (2012). Biogenesis of the Saccharomyces cerevisiae Pheromone a-Factor, from Yeast Mating to Human Disease. https://doi.org/10.1128/MMBR.00010-12

Nasmyth, K. A. (1982). Molecular Genetics of Yeast Mating Type. *Ann. Rev. Genet.* 16:439-500.

Navarro, C. L., Kleijer, W. J., Badens, C., Boyer, A., Boccaccio, I., Freije, J. M., … Cau, P. (2005). Loss of ZMPSTE24 (FACE-1) causes autosomal recessive restrictive dermopathy and accumulation of Lamin A precursors. *Human Molecular Genetics*, *14*(11), 1503–1513. https://doi.org/10.1093/hmg/ddi159

Pendás, A. M., Zhou, Z., Cadiñanos, J., Freije, J. M. P., Wang, J., Hultenby, K., … López-Otín, C. (2002). Defective prelamin A processing and muscular and adipocyte alterations in Zmpste24 metalloproteinase-deficient mice. *Nature Genetics*, *31*(1), 94–99. https://doi.org/10.1038/ng871

Pryor, E. E., Horanyi, P. S., Clark, K. M., Fedoriw, N., Connelly, S. M., Koszelak-Rosenblum, M., … Dumont, M. E. (2013). Structure of the integral membrane protein CAAX protease Ste24p. *Science (New York, N.Y.)*, *339*(6127), 1600–1604. https://doi.org/10.1126/science.1232048

Quigley, A., Dong, Y. Y., Pike, A. C. W., Dong, L., Shrestha, L., Berridge, G., … Carpenter, E. P. (2013). The structural basis of ZMPSTE24-dependent laminopathies. *Science (New York, N.Y.)*, *339*(6127), 1604–1607. https://doi.org/10.1126/science.1231513

Ragnauth, C. D., Warren, D. T., Liu, Y., Mcnair, R., Tajsic, T., Figg, N., … Shanahan, C. M. (2010). Prelamin A Acts to Accelerate Smooth Muscle Cell Senescence and Is a Novel Biomarker of Human Vascular Aging. https://doi.org/10.1161/CIRCULATIONAHA.109.902056

Rankin, J., Auer-Grumbach, M., Bagg, W., Colclough, K., Duong, N. T., Fenton-May, J., … Ellard, S. (2008). Extreme phenotypic diversity and

nonpenetrance in families with the LMNA gene mutation R644C. *American Journal of Medical Genetics, Part A*, *146*(12), 1530–1542. https://doi.org/10.1002/ajmg.a.32331

Schechter, I., & Berger, A. (1967). On the Size of the Active Site in Proteases. I. Papain. *Biochemical and Biophysical Research Communications*, *27*(2), 157–162. https://doi.org/10.1016/j.bbrc.2012.08.015

Sikorski, R. S., & Hieter, P. (1989). *A System of Shuttle Vectors and Yeast Host Strains Designed for Efficient Manipulation of DNA in Saccharomyces ceratisiae*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.334.9789&rep=rep1&type=pdf

Spear, E. D., Alford, R. F., Wood, K. M., Mossberg, O. W., Odinammadu, K., Babatz, T. D., … Michaelis, S. (2019). A humanized yeast system to analyze cleavage of prelamin A by ZMPSTE24. *Methods*, (December 2018). https://doi.org/10.1016/j.ymeth.2019.01.001

Spear, E. D., Hsu, E.-T., Nie, L., Carpenter, E. P., Hrycyna, C. A., & Michaelis, S. (2018). ZMPSTE24 missense mutations that cause progeroid diseases decrease prelamin A cleavage activity and/or protein stability. https://doi.org/10.1242/dmm.033670

Strisovsky, K., Sharpe, H. J., & Freeman, M. (2009). Sequence-Specific Intramembrane Proteolysis: Identification of a Recognition Motif in Rhomboid Substrates. *Molecular Cell*, *36*(6), 1048–1059. https://doi.org/10.1016/J.MOLCEL.2009.11.006

Sullivan, T., Escalante-Alcalde, D., Bhatt, H., Anver, M., Bhat, N., Nagashima, K., … Burke, B. (1999). Loss of A-type Lamin Expression Compromises Nuclear Envelope Integrity Leading to Muscular Dystrophy. *The Journal of Cell Biology*, *147*(5), 913–919. https://doi.org/10.1083/jcb.147.5.913

Tam, A., Nouvet, F. J., Fujimura-Kamada, K., Slunt, H., Sisodia, S. S., & Michaelis, S. (1998). Dual Roles for Ste24p in Yeast a-Factor Maturation: NH 2-terminal Proteolysis and COOH-terminal CAAX Processing. *The Journal of Cell Biology*, *142*(3), 635–649. https://doi.org/10.1083/jcb.142.3.635

UniProt Consortium. (2017). UniProt: the universal protein knowledgebase The UniProt Consortium. *Nucleic Acids Research*, *45*. https://doi.org/10.1093/nar/gkw1099

Urban, S., & Dickey, S. W. (2011). The rhomboid protease family : a decade of progress on function and mechanism. *Genome Biology*, *12*(231), 1–10. https://doi.org/10.1186/gb-2011-12-10-231

Vidmar, R., Vizovi Sek, M., San Turk, D., Turk, B., & Fonovi, M. (2017). Protease cleavage site fingerprinting by label-free in-gel degradomics reveals pH-dependent specificity switch of legumain. *The EMBO Journal*, *36*, 2455–2465. https://doi.org/10.15252/embj.201796750

Wang, Y., Lichter-Konecki, U., Anyane-Yeboa, K., Shaw, J. E., Lu, J. T., Östlund, C., … Worman, H. J. (2016). A mutation abolishing the ZMPSTE24 cleavage site in prelamin A causes a progeroid disorder. *Journal of Cell Science*, *129*(10), 1975–1980. https://doi.org/10.1242/jcs.187302

Weber, K., Plessmann, U., Traub, P. (1989). Maturation of nuclear lamin A involves a specific carboxy-terminal trimming , which removes the polyisoprenylation site from the precursor ; implications for the structure of the nuclear lamina, *257*(2), 411–414.

Worman, H. J., Fong, L. G., Muchir, A., & Young, S. G. (2009). Laminopathies and the long strange trip from basic cell biology to therapy. *The Journal of Clinical Investigation*, *119*, 1825–1836. https://doi.org/10.1172/JCI37679

Yang, S. H., Meta, M., Qiao, X., Frost, D., Bauch, J., Coffinier, C., … Fong, L. G. (2006). A farnesyltransferase inhibitor improves disease phenotypes in mice with a Hutchinson-Gilford progeria syndrome mutation. *The Journal of Clinical Investigation*, *116*(8). https://doi.org/10.1172/JCI28968

Appendix

From the summer of 2010 to the summer of 2014 I worked on human transposon biology in the laboratory of Dr. Kathleen Burns and was co-advised by Dr. Jef Boeke. I was interested in the functional impact of human retrotransposon insertion polymorphisms (RIPs). During that time, I contributed to developing assays for genome-wide characterization of RIP distribution and frequency and made efforts toward cell-based assays for the functional effects of particular RIPs. I authored a review article with Dr. Burns, which is reproduced in this Appendix (Babatz and Burns, 2013).

Functional Impact of the Human Mobilome

Envisioning the RNA origin of DNA may evoke thoughts of a world teeming with self-replicating ribonucleic acids with functions that blur the boundary between the inanimate and simple life. Though sequence evidence for this world is obscure, the RNA origins of our modern human genome are readily apparent; even the casual genome gazer can take in a vast genomic fossil record. Our 3 billion base pairs are overwhelmingly non-coding, and 50% or more are recognizable as repetitive sequences. Most of the repeats are interspersed, meaning that they occur discontinuously as singularly scattered copies in the genome. These include Long INterspersed Elements (LINEs) and Short INterspersed Elements (SINEs). Most abundant in mammals are the retrotransposons, which self-replicate through the reverse transcription of RNA intermediates.

Genomes are not static strings of bases, but instead are mutable, evolving, and structurally dynamic. Three families of human retrotransposons have been active in relatively recent genome evolution and continue to generate human genomic variation today. First, the autonomous LINE-1 element (L1) is 6 kb in its full-length form and contains an internal promoter, two open reading frames (ORFs), and a polyadenylation signal. After RNA polymerase II transcription, mRNA processing, and export to the cytoplasm, ORF1 and ORF2 encode two proteins (ORF1p and ORF2p) that preferentially associate with their encoding RNA (Hohjoh and Singer, 1996; Kulpa and Moran, 2006). The resulting ribonucleoprotein (RNP) particles return the RNA to the nucleus, where ORF2p

initiates target-primed reverse transcription (TPRT) to insert a copied DNA sequence at a new site in the host cell genome (Feng *et al.*, 1996; Mathias *et al.*, 1991; Cost *et al.*, 2002). There are approximately 500,000 copies of L1-derived sequence in the human genome, most of which have been deactivated by 5′-truncation on integration or accumulated mutations. The youngest human-specific L1 subfamily, L1Hs, accounts for about 1% of these copies, with a much smaller subset of 'hot L1' elements driving most of the retrotransposition events.

The remaining families of active retrotransposons are non-autonomous, relying on L1-encoded machinery for mobilization (Dewannieux *et al.*, 2003; Hancks *et al.*, 2011). *Alu* elements are 300 bp long and there are approximately 1,000,000 *Alu*-derived sequences in the human genome. Nucleotide variations within the *Alu* consensus sequence distinguish several subfamilies including the most active *Alu*Y elements (reviewed in Batzer and Deininger, 2002). SVA elements comprise the third family of active retrotransposons; these are a relatively heterogeneous group of sequences with several recognized subfamilies. SVAs are composite elements, named for the sources of their component parts, SINE-R, VNTR, and *Alu* sequences. The SVA consensus sequence is approximately 2 kb and there are roughly 3700 SVA-derived sequences in the human genome (reviewed in Hancks and Kazazian, 2010).

Functional impact

Host genomes are vulnerable to disruption by retrotransposons. How a gene is affected is dependent on both where the mobile DNA insertion occurs and the internal sequence of the element. Exonic insertions can interrupt the

coding region of a gene. In addition, in vitro cell-based assays have revealed that features of L1 sequence such as the promoter, 5'-UTR antisense promoter, and internal splice sites can affect the transcription of target regions and transcript processing (Speek, 2001; Belancio *et al.*, 2008). Concordantly, intronic L1 insertions can induce aberrant transcript initiation, premature truncation, or attenuate transcript elongation, affecting the structure and abundance of target gene transcripts (Wheelan *et al.*, 2005; Han and Boeke, 2005; Han *et al.*, 2004). Similarly, intronic and gene proximal *Alu* insertions can act as enhancers or promoters, introduce cryptic splice donors or acceptors, or abrogate endogenous exon splicing (Bejerano *et al.*, 2006; Thomson *et al.*, 2009; Romanish *et al.*, 2009; Sorek, 2007; Lev-Maor *et al.*, 2003). In clinical genetics, retrotransposon insertions have been implicated in rare cases of highly penetrant Mendelian disorders including hemophilia, cystic fibrosis, neurofibromatosis, and breast and colon cancers (Cordaux and Batzer, 2009; Hancks and Kazazian, 2012). In each respective example the novel insertion creates a highly damaging allele affecting the coding capacity of a known disease gene and thus is fairly definitively implicated as the cause of the disorder.

In contrast, less deleterious insertions that are inherited and constitutive in an individual are major sources of structural variation in the human genome, genetic variation in populations, and occasional causes of genetic disease. Several studies in 2010 demonstrated that retrotransposon insertion polymorphisms (RIPs) are significantly more polymorphic among individual human genomes than previously recognized.

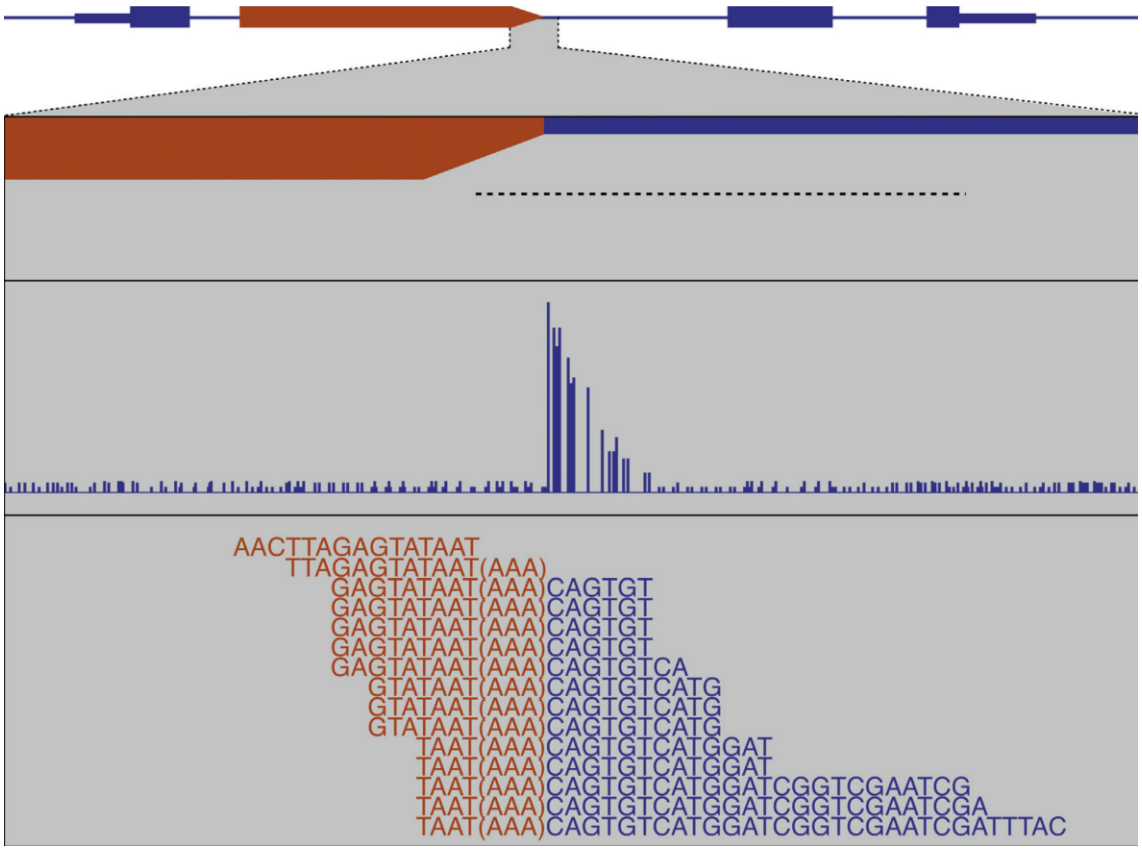

Figure 1: Detecting insertions. A schematized gene containing three exons (blue) is shown with an intronic L1 insertion (orange) in the sense orientation. Zoomed in, primers for repeat detection and the resulting PCR amplicons are represented (black). These assays use a repeat-specific forward primer with a set of non-specific reverse primers or a universal primer that targets adapter sequence ligated to the genomic DNA. Resulting amplicons contain the 3' end of the repeat sequence along with flanking unique genomic sequence. This complex mixture of amplicons can be labeled and hybridized to a microarray or sequenced. Both methods provide the genomic locations of individual insertions. Intensity data from an array platform can be visualized as peaks (vertical bars), which indicate location and orientation of the insertion. In 2nd generation sequencing, short reads show the precise junction between the insertion (orange) and adjacent genomic sequence of the insertion (blue). The variable length poly-A tract (in parentheses) is not templated but results from L1 mRNA processing.

Germline insertions

Because de novo insertions are templated from the relatively few active subfamilies of repeats, each new insertion bears high internal sequence

homology to the subfamily consensus sequence. This feature can be leveraged in targeted approaches for finding mobile DNA variants. These methods can be broadly categorized as either using laboratory-based genomic assays to map repeats or using computational means to recognize insertions in unselected reads in sequencing studies. The first category has included variations on hemi-specific PCRs using one primer within the repeat family of interest and a reaction schema that amplifies adjacent DNA (Huang *et al.*, 2010; Ewing and Kazazian, 2010; Iskow *et al.*, 2010; Witherspoon *et al.* 2010) (Figure 1). The latter has had the advantage of using data from more large-scale collaborative efforts such as the 1000 Genomes project, to identify large numbers of new insertions and phase these with neighboring SNPs (Hormozdiari *et al.*, 2010; Stewart *et al.*, 2011). At the same time, Beck *et al.* reported polymorphism of full-length 'hot' L1Hs elements among populations, suggesting possible inter-individual variation in endogenous L1 activity (Beck *et al.*, 2010). Collectively, new perspectives resulting from these discovery efforts have upheld earlier assertions that mobile DNAs are a major source of human genetic diversity. We currently estimate that between 1/100 and 1/200 individuals harbor a de novo LINE-1 insertion, and that about 1/20 individuals have a de novo *Alu* element insertion. This degree of activity is expected to result in approximately 11,000 L1Hs and 65,000 *Alu*Y common variable insertion alleles (MAF > 0.05) segregating in human additional increasingly rare alleles.

Recently, two studies elucidated the mutational mechanism of inherited insertions in recessive diseases, one of which was a novel insertion discovery. A third study reported a novel mechanism of accumulated Alu RNA cytotoxicity. Taniguchi-Ikeda *et al.* functionally characterized the polymorphic SVA insertion in the 3'-UTR of the *FUKUTIN* gene, which causes Fukuyama muscular dystrophy (FCMD). FCMD is a rare autosomal recessive disease mainly described in the Japanese population (Taniguchi-Ikeda *et al.*, 2011) and currently has no effective treatments. This ancient SVA is on a haplotype shared by >80% of FCMD chromosomes and likely arose as a single ancestral allele (Kobayashi *et al.*, 1998). Long-range RT-PCR and sequence analysis in FCMD lympho- blast mRNA and controls revealed a novel splicing event between a weak alternative donor site in the final *FUKUTIN* coding exon and an acceptor site within the SVA insertion. This aberrant splicing excises the normal stop codon, deletes coding sequences for the 38 C-terminal amino acids of FUKUTIN, and adds 129 amino acids encoded by the SVA sequence. Normal FUKUTIN localizes to the Golgi, while FUKUTIN with SVA-derived C-terminus mislocalizes to the ER.

Remarkably, the authors were able to restore normal mRNA splicing of FCMD alleles to 40% of control levels using a combination of antisense oligonucleotides (AONs) targeted to the alternative splice donor, acceptor, and an exonic splicing enhancer. This strategy increased normal *FUKUTIN* mRNA expression significantly and rescued functional FUKUTIN protein in vivo in knock-in mice homozygous for humanized SVA-containing FUKUTIN. This study elucidated the mutational mechanism of a polymorphic disease-causing

retrotransposon and demonstrated the possibility of splicing modulation therapy as a treatment for FCMD.

Tucker *et al.* performed exome sequencing on a single proband with retinitis pigmentosa (RP), a typically autosomal recessive ciliopathy characterized by loss of photoreceptor cells and a significant amount of genetic heterogeneity (Tucker *et al.*, 2011). After stringent filtering of sequence data from both ABI and Illumina sequencing instruments and prioritization of identified variants, there was a surprising lack of plausible disease-causing mutations verified by Sanger sequencing and present in both datasets. The ABI data showed a compound heterozygous pair of mutations in exon 9 of male germ cell-associated kinase (MAK) that were not in the Illumina data, and Sanger sequencing of the exon surprisingly revealed a homozygous 353-bp Alu insertion. The authors investigated why this important variant was missed by one method and misidentified by the other. The ABI protocol completely removed the *Alu*-MAK junction sequences and therefore sequenced chimeric DNA molecules and identified two heterozygous variants, and the Illumina protocol sequenced the junction sequences but trimmed the *Alu* sequence from the ends of the junction reads during analysis, making it invisible. The authors went on to functionally characterize the homozygous *Alu* variant in induced pluripotent stem cells (iPSCs) derived from the RP patient's dermal fibroblasts and controls, showing that the *Alu* insertion abrogates an essential developmental switch to alternatively include exon 9 and a previously unknown retina-specific exon 12 in retinal lineage differentiated cells. Screening of 1798 additional RP patients

yielded 21 individuals homozygous for the *Alu*, all of whom reported Jewish ancestry, and the *Alu* was absent from controls. This study highlights current strengths and weaknesses of relatively new mutation discovery tools and suggests that there may be significantly more disease-causing insertions than previously appreciated.

Kaneko *et al.* reported a pathogenic role for accumulating *Alu* RNA in retinal pigment epithelial (RPE) degeneration in geographic atrophy (GA), an untreatable advanced form of age-related macular degeneration (Kaneko *et al.*, 2011). This is the first known example of *Alu* RNA cytotoxicity, rather than *Alu*-mediated genomic rearrangement or insertional mutagenesis, causing disease. The authors observed a significant deficit of *DICER1* mRNA and protein and an excess of dsRNA in RPE of GA donor eyes as compared to controls. Subsequent cloning and sequencing of the RNA species showed homology to the human *Alu*Sq subfamily. *DICER1* knockdown increased *Alu* RNA accumulation, and overexpression of exogenous *Alu* in human RPE cells overwhelmed endogenous DICER and reduced cell viability. Finally, RPE cell viability was rescued by antisense oligonucleotides (AONs) targeting *Alu*Sq. The authors concluded that DICER is critical for RPE cell survival by silencing cytotoxic *Alu* transcripts and highlighted a new role for DICER in protecting against dsRNA transcribed from the most abundant element in the human genome. The study raises the intriguing questions of whether upregulation of *Alu*-containing RNA polII-transcribed genes could be pathogenic, or whether *Alu* RNA toxicity plays a role in other tissues and diseases.

Somatic insertions

Perhaps the most exciting developments in the study of human mobile elements relate to somatic retrotransposition, both in terms of normal variability and disease (Figure 2). A series of three papers from the laboratory of Dr. Fred Gage reported the first evidence of somatic retrotransposition in mammalian tissues and suggested the tantalizing hypothesis that ongoing retrotransposition contributes to somatic mosaicism and phenotypic variability in humans. Muotri *et al.* reported in 2005 that an engineered human L1 can retrotranspose in differentiating rat neural progenitor cells (NPCs) in vitro and in the brains of transgenic mice in vivo, however this study did not address insertions from endogenous mobile elements (Muotri *et al.*, 2005). In 2009 Coufal *et al.* went on to show that human fetal brain stem cells and derived neural progenitor cells support retrotransposition of the engineered L1, and used a quantitative multiplexing PCR assay targeting a 30 region of L1 ORF2 to examine endogenous L1 copy number in several tissues (Coufal *et al.*, 2009). Notably, they observed a subtle yet statistically significant increase of genomic L1 copy number in brain tissues as compared to liver and heart. There was variability among individuals and among brain regions, suggesting ongoing somatic retrotransposition during development. Furthermore, in 2010 Muotri *et al.* investigated the relationship between L1 retrotransposition and MeCP2, a methylation-sensitive DNA binding protein frequently mutated in Rett Syndrome (RTT) (Muotri *et al.*, 2010). The authors examined L1 retrotransposition in NPCs derived from induced pluripotent stem cells (iPSCs) from Rett Syndrome patients

and controls. The engineered L1 retrotransposed twice as frequently in RTT NPCs than in controls. Genomic qPCR revealed an increase in L1 ORF2 copy number in RTT brain relative to controls, and in brain relative to heart tissue from the same individual. Taken collectively, this series of papers supported the model of L1 driving genetic heterogeneity in neuronal lineages. The missing piece of evidence was then mapping individual retrotransposition events in the resulting heterogeneous cell populations.

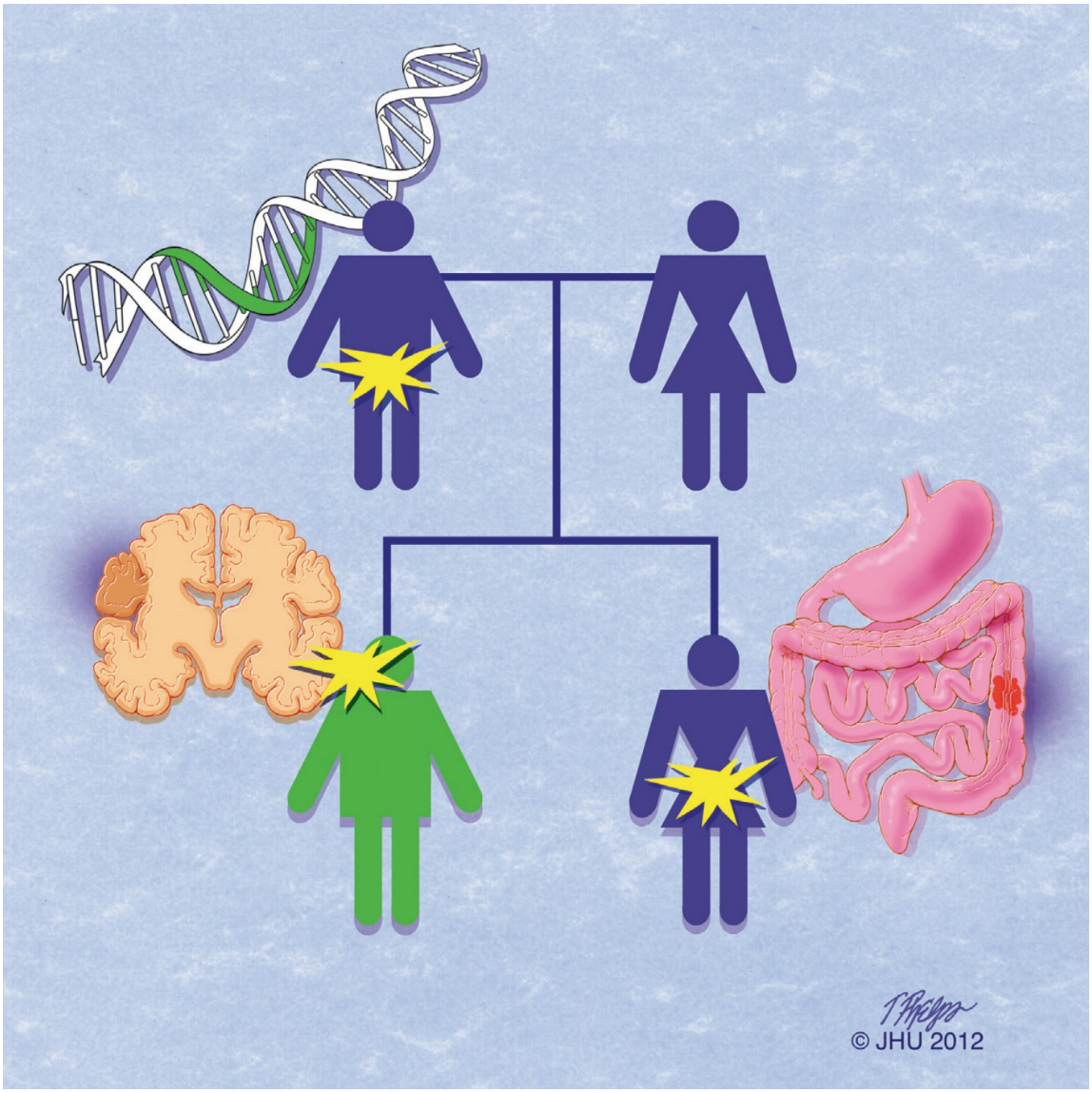

Figure 2: Three contexts for active retrotransposition. This pedigree shows three examples of retrotransposition events (represented by yellow flashes) in the human genome. First, a novel retrotransposon insertion occurs in a subpopulation of germ cells in the father and is transmitted to the male child, who constitutively harbors the insertion as an inherited variant (represented in green). Second, the male child acquires a somatic insertion in his brain during neuronal development, and this leads to a subpopulation of genetically distinct neurons containing the insertion (orange). Finally, the female child acquires a tumorigenic somatic insertion in the epithelial cells of her colon that promotes clonal expansion in the developing tumor (red).

Baillie *et al.* addressed this problem using a novel genomic approach that complemented and validated the work from the Gage group, galvanizing the idea that these elements are mobile in somatic tissues at points beyond early embryonic development and contribute to cellular heterogeneity (Baillie *et al.*, 2011). The authors developed a high-throughput protocol utilizing hybridization-based capture of L1, *Alu*, and SVA sequences followed by deep sequencing of the junctions, conservative filtering, and mapping to the genome. This assay was performed on genomic DNA from hippocampus and caudate nucleus from three individuals. Putative insertions were classified as somatic if they were in one brain subregion and not the other, absent from all known polymorphism catalogs, and absent from similar data from pooled genomic DNA extracted from blood. PCR amplification and sequencing validated many somatic 50 junctions and 80% of somatic L1 insertions corresponded to the youngest subfamilies, L1-Ta and pre-Ta. This study extended somatic mobilization to *Alu* and SVA elements, and analogous to L1, 83% of somatic *Alu* insertions correspond to the youngest *Alu*Y subfamily. The observed somatic insertions disproportionately target protein-coding genes, raising the possibility of phenotypic variability between subpopulations of cells harboring different complements of insertions.

To investigate the frequency of somatic retrotransposition in the brain, Evonry *et al.* applied the L1-seq method of Ewing and Kazazian (2010) to 300 individual neurons from three individuals (Evonry *et al*., 2012). They used single-cell sorting followed by multiple displacement amplification (MDA) of the whole genome before L1-seq, controlling for allelic and locus dropout. They estimate <0.6 somatic L1 insertions per neuron with >80% of neurons having no somatic insertions and conclude that somatic retrotransposition is not an essential source of neuronal diversity in the regions studied. The single somatic L1 whose full length was successfully cloned contained a 5' transduction mapping back to a full length parental L1 that is polymorphic. This study provides a potentially sobering quantification of somatic retrotransposition in the brain, but does revisit the interesting question of polymorphic 'hot' L1s contributing to variation in somatic retrotransposon activity (Beck *et al*., 2010).

Somatic insertions in cancer

Several groups have reported recently that somatic retrotransposition also occurs in various cancers. The first of these, Iskow *et al.*, reported high levels of somatic L1 retrotransposition in lung cancer genomes and detected a hypomethylation signature specific to L1-permissive tumors (Iskow *et al.*, 2010). Whether the identified insertions are a cause or symptom of neoplasia remains unclear, however there is evidence to suggest that tumor-specific insertions are not merely passenger mutations. Lee *et al.* (2012) developed a computational method termed Transposable Element Analyzer (Tea) to detect TEs in whole-genome sequencing data and applied it to data from 43 tumors and matched

normal blood samples. They identified somatic insertions in colorectal, prostate, and ovarian tumors with significant variation in TE activity among tumors, and no insertions were observed in multiple myeloma or glioblastoma. TE target genes were significantly enriched for genes frequently mutated in cancer and target genes showed decreased expression compared to controls.

The great majority of somatic L1s were significantly truncated upon insertion. Solyom *et al.* performed L1-specific amplification followed by resequencing (Ewing and Kazazian, 2010) in 16 colorectal tumors and matched normal colorectal DNA and identified somatic L1 insertions with a high degree of variability among individual tumors (Solyom *et al.*, 2012). Many of the somatic L1s were PCR validated and 50 and 30 junctions were sequenced. Like Lee *et al.* the majority of somatic L1s were significantly truncated. Notably, this group detected an L1 insertion in one gene identified by Lee *et al.*, Roundabout homolog 2 (ROBO2); these findings indicate independently derived insertions in an intron of the gene. This group also observed an interesting positive correlation of number of somatic L1s with patient age. Together, these studies indicate a possible role for somatic retrotransposition in tumorigenesis, however, more specific characterization of target gene insertions sites is required.

Conclusions

The last several years have been exciting times in human transposon biology. New methods such as high throughput sequencing, hybridization-based sequence capture, and repeat-specific computational analysis have allowed for high-resolution detection of RIPs in human genomes and revealed their

extensive polymorphism among diverse human populations. Recent work has validated the accumulating evidence for active retrotransposition in several tissues with the potential to affect gene expression, developmental cell fate, and contribute to tumorigenesis in cancer. Topics for further investigation include the unique properties of various tissues that foster permissive environments for retrotransposon mobilization; improved detection and awareness of insertions in exome and whole-genome sequencing studies; and determined the relative contribution of somatic insertions to functional diversity in heterogeneous cell populations. Alternatively, they may be merely benign rare events with little to no functional impact.

One can also speculate about the role of RIPs in common complex disease. We consider the 'mobilome' as an untapped source of human genomic variation in complex disease. We are reminded of the relatively recent realization of extensive copy number polymorphism (CNP) and subsequent association of many CNPs with risk for common disease (Wain *et al.*, 2009). In contrast to highly deleterious coding insertions, intronic and intergenic insertions can mildly affect transcript structure and abundance, and such mild changes of gene expression are inherently likely to have modest effect sizes and interact with multiple loci and/or environmental factors in complex disease.

Works Cited in Appendix:

Babatz, TD and Burns, KH. Functional impact of the human mobilome. Curr. Opin. Genet. and Dev. 2013, 23:264-270.

Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan PM, Rizzu P, Smith S, Fell M *et al*. Somatic retrotransposition alters the genetic landscape of the human brain. Nature 2011, 479:534-537.

Batzer MA, Deininger PL. Alu repeats and human genomic diversity. Nat Rev Genet 2002, 3:370-379.

Beck CR, Collier P, Macfarlane C, Malig M, Kidd JM, Eichler EE, Badge RM, Moran JV. LINE-1 retrotransposition activity in human genomes. Cell 2010, 141:1159-1170.

Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 2006, 441:87-90.

Belancio VP, Roy-Engel AM, Deininger P. The impact of multiple splice sites in human L1 elements. Gene 2008, 411:38-45.

Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. Nat Rev Genet 2009, 10:691-703.

Cost GJ, Feng Q, Jacquier A, Boeke JD. Human L1 element target-primed reverse transcription in vitro. EMBO J 2002, 21:5899-5910.

Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, Morell M, O'Shea KS, Moran JV, Gage FH. L1 retrotransposition in human neural progenitor cells. Nature 2009, 460:1127-1131.

Dewannieux M, Esnault C, Heidmann T. LINE-mediated retrotransposition of marked Alu sequences. Nat Genet 2003, 35:41-48.

Evrony GD, Cai X, Lee E, Hills LB, Elhosary PC, Lehmann HS, Parker JJ, Atabay KD, Gilmore EC, Poduri A. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell 2012, 151:483-496.

Ewing AD, Kazazian HH Jr. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res 2010, 20:1262-1270.

Feng Q, Moran JV, Kazazian HH Jr, Boeke JD. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 1996, 87:905-916.

Han JS, Szak ST, Boeke JD. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature 2004, 429:268-274.

Han JS, Boeke JD: LINE-1 retrotransposons. modulators of quantity and quality of mammalian gene expression? Bioessays 2005, 27:775-784.

Hancks DC, Kazazian HH Jr. SVA retrotransposons: evolution and genetic instability. Semin Cancer Biol 2010, 20:234-245.

Hancks DC, Goodier JL, Mandal PK, Cheung LE, Kazazian HH Jr. Retrotransposition of marked SVA elements by human L1s in cultured cells. Hum Mol Genet 2011, 20:3386-3400.

Hancks DC, Kazazian HH Jr. Active human retrotransposons: variation and disease. Curr Opin Genet Dev 2012, 22:191-203.

Hohjoh H, Singer MF. Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. EMBO J 1996, 15:630-639.

Hormozdiari F, Alkan C, Ventura M, Hajirasouliha I, Malig M, Hach F, Yorukoglu D, Dao P, Bakhshi M, Sahinalp SC *et al*. Alu repeat discovery and characterization within human genomes. Genome Res 2010, 21:840-849.

Huang CR, Schneider AM, Lu Y, Niranjan T, Shen P, Robinson MA, Steranka JP, Valle D, Civin CI, Wang T *et al*. Mobile interspersed repeats are major structural variants in the human genome. Cell 2010, 141:1171-1182.

Iskow RC, McCabe MT, Mills RE, Torene S, Pittard WS, Neuwald AF, Van Meir EG, Vertino PM, Devine SE. Natural mutagenesis of human genomes by endogenous
retrotransposons. Cell 2010, 141:1253-1261.

Kaneko H, Dridi S, Tarallo V, Gelfand BD, Fowler BJ, Cho WG, Kleinman ME, Ponicsan SL, Hauswirth WW, Chiodo VA *et al*. DICER1 deficit induces Alu RNA toxicity in age-related macular degeneration. Nature 2011, 471:325-330.

Kobayashi K, Nakahori Y, Miyake M, Matsumura K, Kondo-Iida E, Nomura Y, Segawa M, Yoshioka M, Saito K, Osawa M *et al*. An ancient retrotransposal

insertion causes Fukuyama-type congenital muscular dystrophy. Nature 1998, 394:388-392.

Kulpa DA, Moran JV. Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. Nat Struct Mol Biol 2006, 13:655-660.

Lev-Maor G, Sorek R, Shomron N, Ast G. The birth of an alternatively spliced exon: 30 splice-site selection in Alu exons. Science 2003, 300:1288-1291.

Lee E, Iskow R, Yang L, Gokcumen O, Hanseley P, Luquette LJ, Lohr JG, Harris CC, Ding L, Wilson RK *et al*. Landscape of somatic retrotransposition in human cancers. Science 2012, 337:967-971.

Mathias SL, Scott AF, Kazazian HH Jr, Boeke JD, Gabriel A. Reverse transcriptase encoded by a human transposable element. Science 1991, 254:1808-1810.

Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, Gage FH. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature 2005, 435:903-910.

Muotri AR, Marchetto MC, Coufal NG, Oefner R, Yeo G, Nakashima K, Gage FH. L1 retrotransposition in neurons is modulated by MeCP2. Nature 2010, 468:443-446.

Romanish MT, Nakamura H, Lai CB, Wang Y, Mager DL. A novel protein isoform of the multicopy human NAIP gene derives from intragenic Alu SINE promoters. PLoS One 2009, 4:e5761.

Solyom S, Ewing AD, Rahrmann EP, Doucet TT, Nelson HH, Burns MB, Harris RS, Sigmon DF, Casella A, Erlanger B *et al*. Extensive somatic L1 retrotransposition in colorectal tumors. Genome Res 2012, 22:2328-2338.

Sorek R. The birth of new exons: mechanisms and evolutionary consequences. RNA 2007, 13:1603-1608.

Speek M. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol Cell Biol 2001, 21:1973-1985.

Stewart C, Kural D, Stromberg MP, Walker JA, Konkel MK, Stutz AM, Urban AE, Grubert F, Lam HY, Lee WP *et al*. A comprehensive map of mobile element insertion polymorphisms in humans. PLoS Genet 2011, 7:e1002236.

Taniguchi-Ikeda M, Kobayashi K, Kanagawa M, Yu CC, Mori K, Oda T, Kuga A, Kurahashi H, Akman HO, DiMauro S *et al*. Pathogenic exon-trapping by SVA retrotransposon and rescue in Fukuyama muscular dystrophy. Nature 2011, 478:127-131.

Thomson SJ, Goh FG, Banks H, Krausgruber T, Kotenko SV, Foxwell BM, Udalova IA. The role of transposable elements in the regulation of IFN-lambda1 gene expression. Proc Natl Acad Sci U S A 2009, 106:11564-11569.

Tucker BA, Scheetz TE, Mullins RF, Deluca AP, Hoffmann JM, Johnston RM, Jacobson SG, Sheffield VC, Stone EM. Exome sequencing and analysis of induced pluripotent stem cells identify the cilia-related gene male germ cell-associated kinase (MAK) as a cause of retinitis pigmentosa. Proc Natl Acad Sci U S A 2011, 108:E569-E576.

Wain LV, Armour JA, Tobin MD. Genomic copy number variation, human health, and disease. Lancet 2009, 374:340-350.

Wheelan SJ, Aizawa Y, Han JS, Boeke JD. Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution. Genome Res 2005, 15:1073-1078.

Witherspoon DJ, Xing J, Zhang Y, Watkins WS, Batzer MA, Jorde LB. Mobile element scanning (ME-Scan) by targeted high-throughput sequencing. BMC Genomics 2010, 11:410.

BIOGRAPHY

Timothy David Babatz was born in 1986 in Minnesota. Tim did his undergraduate degree in Biology at Wheaton College in Wheaton, Illinois. He was a member of the cross country and track and field teams through his undergraduate years and enjoyed spending a summer in the Black Hills of South Dakota studying ecology, despite eventually being drawn to human genetics.

From 2007 to 2009 Tim worked in the Department of Human Genetics at the University of Chicago in the lab of Dr. William Dobyns gaining valuable research experience and foundational knowledge in the field, which was new to him.

In 2009, Tim was accepted to the Human Genetics program of the Institute of Genetic Medicine at the Johns Hopkins University School of Medicine and moved to Baltimore. From 2010 to 2014 he worked with Dr. Kathleen Burns and Dr. Jef Boeke studying human retrotransposon biology. From 2014 to 2016 he took a leave of absence and worked several jobs having nothing to do with science. In 2016 he returned to Johns Hopkins University and joined Dr. Susan Michaelis' lab to complete the work that comprises the majority of this thesis. While in graduate school Tim maintained his long-standing passion for distance running and began exploring road cycling in the Baltimore County countryside, interests he and Dr. Michaelis share.