

The Evolutionary Potential of New Protein Domains

Jeremy Hollis

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:

Dr. Melody Campbell, Chair

Dr. Harmit Malik

Dr. Douglas Fowler

Program Authorized to Offer Degree:

Molecular and Cellular Biology

©Copyright 2026
Jeremy Hollis

University of Washington

Abstract

The Evolutionary Potential of New Protein Domains

Jeremy Hollis

Chair of the Supervisory Committee:

Dr. Melody Campbell

Division of Basic Sciences, Fred Hutchinson Cancer Center

Proteins are the molecular working units that build, control, maintain and/or innovate in the biological realm. Proteins power everything from the most fundamental, shared processes across all of life to the highly specialized, niche interactions that generate specificity and distinction between even the closest living relatives. In keeping with their functional breadth, proteins are equally complex in how they are encoded and assembled, and how they divide labor across their physical surface. To understand how proteins function, we often take a reductionist approach, fractionating them from their entirety into subunits that share physical or functional relationships with other protein subunits into what we term “domains.” Often, these subunit boundary delineations are guided by nature: protein domains are typically shared across homologous proteins and have common sequential and structural ancestral features, which help us both understand their evolutionary relationships and their functional similarities.

Work across all scales of biological research has contributed greatly to our understanding of how protein domains contribute to biological complexity. Structural and biochemical work

has given us molecular insight into the atomic mechanisms by which protein domains are built and operate, while work at the cellular, organismal, and population levels has helped us understand how these simple building blocks assemble to form complex machinery that fuels life. Although some proteins are comprised of only a single functional domain, many contain multiple domains which collectively allow complex work by a single protein by uniting multiple specific processes to create a multifunctional protein. For example, the combination of a LIM domain and kinase domain may allow phosphorylation to happen specifically at curved cell surfaces, offering a mechanism for cells to sense and respond to exterior confinement via intracellular signaling. This complexity is widespread; 60% of proteins across the eukaryotic proteome contain multiple functional domains.

I am fascinated by how these discrete subunits can build, stack, and recombine to generate the beautiful diversity that exists in the natural world. I am interested in both how new domains move and diversify across evolutionary timescales, as well as how domains can retain function despite strong evolutionary pressures to innovate. To tackle these large questions, I have primarily focused my studies on the integrin α I domain as a case study. This domain fundamentally changed how integrins interact both with the exterior world and the interior of the cell and shows a remarkable diversity across the chordate tree of life. I have secondarily traced the evolutionary retention of the scm3 domain, a small protein domain responsible for delineating the centromeric region of chromosomes necessary for faithful cell division that has been functionally retained despite near complete sequence turnover in *Metazoa*.

Integrins are heterodimeric cell surface receptor proteins comprised of an alpha (α) and beta (β) subunit which transmit signals from ligand binding across the cytoplasmic membrane to trigger cellular responses via signaling cascades. Integrins are largely structurally homologous

across all of *Metazoa*, but some α subunits contain an additional domain called the α I or “inserted” domain, a von Willebrand Factor A-family domain which fundamentally changed how integrins interact with their ligands. When I began my thesis work, we knew structurally how these domains in isolation could change structural conformation, as well as separately how they could bind to proteinaceous ligands. We did not know, however, where this domain came from and how it was poised for evolutionary success despite its insertion into an exceptionally risky area of the integrin protein. We also lacked a structural sense of how ligand binding was allosterically communicated through this domain to the rest of the integrin molecule and ultimately into the cell.

To understand how this domain mechanistically transmits signal, in my first thesis chapter, I employed single particle cryogenic electron microscopy (cryoEM) to compare two integrins: α E β 7, which contains an I domain, and α 4 β 7 which does not. I generated high-resolution structures of the integrins in both apo and ligand-bound states to directly observe how ligand binding changed upon acquisition of the I domain, as well as how it uses a structurally labile C-terminal helix to link ion coordination to integrin activation. These studies generated the first high-resolution structures of these integrin ectodomains and further revealed how ancestral integrin signaling could broadly be retained despite the steric occlusion of the ancestral ligand binding site by the I domain. To understand the evolutionary origin of the I domain, I combined phylogenetic analyses with structural modeling, ancestral reconstruction, and flow cytometry which allowed me to pinpoint the serendipitous features of its acquisition that permitted this complex ion coordination behavior from its inception. I also found that this domain has a collagen origin, hinting at a possible ancestral function in collagen binding that is still present in extant integrin I domains across chordates.

Integrin I domains vary widely in their ligand binding capacities, with some showing extreme specificity for a single cognate ligand and others having exceptional promiscuity. Thus, the I domain is also a great model for understanding how specificity is generated in the biological world. To dissect the biochemical basis of this dichotomy, in my second chapter I focused on how two closely related integrins – the highly specific $\alpha\text{L}\beta\text{2}$ and the highly promiscuous $\alpha\text{M}\beta\text{2}$ – distinguish their ligand binding profiles at the amino acid level. I used deep mutational scanning to probe how individual mutations affect ligand affinities at a high-throughput scale, cryoEM to mechanistically dissect division-of-labor across the integrin surface, and phylogenetics to link these dual behaviors to broad evolutionary trajectories. These studies not only answered longstanding questions in the field as to how $\alpha\text{M}\beta\text{2}$ can be so promiscuous, but how purifying pressures are equally important in both protein-protein interaction and allosteric regulation as well as how avidity can relax constraint on ligand recognition.

Finally, in my third chapter I explore how protein domain structure and function can be preserved despite extreme divergence at the sequence level using a different model domain. In this case, I used a combination of remote homology detection approaches to trace the scm3 domain, which chaperones the centromeric histone variant CENPA to chromosomes, across the animal tree. Due to rapid sequence divergence, it was thought this domain was missing in most animals, including key model species. By uncovering this domain across these “missing lineages,” my work supports the model that structure is more conserved than sequence, function can be maintained across long evolutionary timescales despite a near complete lack of sequence homology, thus changing our perspective on protein constraint as it relates to function.

Altogether these studies shape our perspective on how evolutionary potential drives protein domain innovation. Tracing domain recombination and retention helps inform the

mechanisms by which proteins build complexity and the subsequent limitations of that complexity, or lack thereof. Domain structural studies in the context of near complete protein provide insight into how novelty is united with ancestry to form functional machinery despite the high risks that domain insertions impose. Finally, integrating mutational data informs us where the limits of protein domains lie, and if we look close enough might even give us a clue as to where they'll go next.

Dedicated to the arts

Acknowledgements

It takes a community to build a dissertation. Although this document is a culmination of “my” work, it represents the dedication of over 100 individuals and countless institutions in my life who have contributed to its success. I’m overjoyed to be able to celebrate them here.

I first acknowledge the land on which I both worked and played. Seattle is a beautiful city seated in a beautiful stretch of land where one can be whoever they want to and do whatever they want to do. The caretaking of this land has been accomplished mostly by the collective efforts of the native Coast Salish people, an effort that has been made harder and harder in recent years. I’m grateful that it’s so easy to spend a night at the theater, then do a backpacking trip the next day somewhere that feels truly remote and respected. The Washington Trails Association further supports this endeavor. I ran a lot during my time working on my dissertation; I never pictured myself a runner, but running with people in Cowgill Trail Collective and the Seattle Frontrunners, particularly outdoors, has kept me more active than ever despite working sedentarily more than ever. My time spent outdoors was a great way to refresh and reset when times in lab were less than ideal.

I acknowledge the incredible Seattle theater and dance communities. The artists I’ve worked with are some of the most dedicated and smart people I’ve ever met, and frankly it’s refreshing to have people around me who don’t know anything about what I do in lab. I’m particularly thankful to Village Theatre, ArtsWest, Dacha Theatre, Washington Ensemble Theatre, M!NX, and Purple Lemonade, all of whom bravely support new work at a time when making art is not prioritized. Artists at these organizations, as well as those who work to fund them, ensure our city doesn’t fall prey to a loss of empathy, compassion, and creativity that can happen with an influx of corporate monotony.

I’ve had a rich social community outside of lab work who kept me (relatively) grounded. Sam, James, Andrew, Miguel, Devan, Christian, Erik, Sean and Steve always reminded me that it wasn’t that serious, and getting out of lab at a reasonable time was probably the best thing I could do for myself and my research. I’ll cherish our annual trips in the winter to various warm destinations and I look forward to continuing that tradition in the future. Sam Nutt, in particular, was my ride-or-die during my thesis work. In addition to agreeing to be my roommate before we even met, he shared more adventures with me than anyone else in my life. It’s been refreshing to have someone to relate to going through this process. My partner Dylan Bray was my rock throughout most of my thesis work. He was my biggest ally and supporter, at times my toughest critic when I needed it, and filled my life with love and laughter. I don’t know how I would’ve done it without these incredible people.

I’m grateful to the funding agencies that fuel scientific discovery. In the United States these are predominantly taxpayer-funded, public entities and as such the work we do as scientists carries a strong sense of duty to the greater good. I have been funded completely through the National Institutes of Health and the National Science Foundation, both of which allowed me the financial freedom to pursue exciting discoveries in a relatively unrestricted way. Research is

equally as good at serving intellectual curiosity and human advancement, and it's not lost on me the duty that we have to ensure our discoveries are made both available and approachable to the public. These funders also support the excellent core facilities at the Hutch, which streamline data collection and processing and genuinely propel scientists to focus on discovery. I hope I can permit the trainees I may have in the future to pursue their ideas in a similar way.

I acknowledge my past scientific mentors: Sadie Wignall, Diana Libuda, Geoff Smith, and Heather Rhodes. I trained and learned in various capacities with these astonishing scientists, all of whom encouraged independence, creativity and exploration first and foremost. I was more prepared than I realized moving into graduate school particularly because Sadie allowed me the freedom to pursue my own scientific project while working as a technician in her lab. This project was done jointly with the Libuda lab under supervision of Diana Libuda, who always made time to give me advice even though she was not my direct mentor. Both she and Sadie tipped the balance for me to pursue my graduate work in the biomedical sciences, ultimately one of the best decisions I've ever made.

I acknowledge the formal mentorship I've had while performing my graduate work. First and foremost, Melody Campbell is the best mentor a student could ask for. She is kind, discerning, smart, curious, accomplished, supportive, dedicated, iconic, and patient beyond compare. She afforded me all of the resources I needed to complete this work, as well as exactly enough space and independence to pursue where my curiosity led, without straying from important goals. She is an exceptional role model in both science and leadership, and if I can carry even a fraction of what she has taught me I think my future will be bright. Similarly, Harmit Malik served as a complementary role model and advisor in my studies. He is so good at seeing the big picture, and he is so caring about the people who work with him in any capacity. He navigates challenges with poise and profession, always has an open ear, and like Melody balances life and work priorities amazingly well. These two are a dream mentorship team, and I hope whoever is reading this is at least a little jealous that I was fortunate to work under them.

I acknowledge the less formal mentorship I've had during this time. My first rotation was in the Malik lab in the fall of 2020, during which I worked under Aida de la Cruz and Janet Young. At a time when rigor and reproducibility are particularly emphasized in research, Janet in particular aptly demonstrated the ways to have confidence and enthusiasm in results all while providing great hiking and backpacking recommendations in the area. Aida is one of the most fun people I've ever met- she makes the Malik lab a party every day in addition to being a great scientist, and it seems to me this was one of the reasons I enjoyed going in every day. Likewise Rachel Werther in the Campbell lab taught me everything I know about biochemistry and solved a mystery like a private investigator right out of a noir film. She always kept the lab interesting and always prioritized celebrating our victories, and I'm so grateful she was a resource throughout the entirety of my dissertation work. My committee Sue Biggins, Justin Kollman, and Doug Fowler, too, were critical to my success. Without fail they saw the forest for the trees when I was too in my head and were sages with career advice. I always looked forward to my committee meetings. Sue Biggins is an exceptional division leader and mentor and despite these commitments still found time to write recommendation letters and read papers for me. She balances relationships, administration, and research better than anyone I've ever seen; I'll miss working under her leadership dearly.

I acknowledge my labmates, the folks I work with every day. My baymate Andrés, fellow first-Campbell-Lab-graduate-student Adam, Eva, Rachel, Matthew, Thamiya, and Cat are the best coworkers in the world. We're an eclectic mix of people loosely joined by our love of

science and art and baked goods, but the concentration of talent and kindness in the Campbell lab is unparalleled. They are quick to help when needed, almost scary patient, funny, and always keen to try something crazy. It's hard to capture how amazing they are in words but here's a photo:



I acknowledge the spectacular scientists I was fortunate to mentor during my thesis work. Kevin, Anthony and Chelsea certainly taught me much more than I taught them, and it's been my great joy to see them move on to big, amazing things in research and beyond. Their dedication and open-mindedness were boundless, and they assured me that the future is in good hands.

I have loved my life in Seattle. Deeply, wholly, unabashedly, unconditionally loved it. I recognize my privilege here, as not many graduate students really get to qualify their experiences as joyful, happy, uplifting, positive, etc. I am so blessed to be surrounded by these supportive individuals and communities here every day make life more and more worth living. Cheers to the amazing people in my life, to you the reader, and to my past, present, and future. Onward!

Table of Contents

List of Figures	xv
List of Tables	xvii
List of Abbreviations	xviii
Chapter 1 Introduction	1
1.1 Abstract	1
1.2 Integrins and their functions	1
1.2.1 A structural perspective on integrin function	1
1.2.2 The many integrin ligands and their duties	4
1.2.3 Ancestral ligand binding and the RGD cleft	6
1.2.4 Structural insights into intracellular integrin signaling	
1.3 Structural features of the integrin αI domain	8
1.3.1 Overview of I domain activation	9
1.3.2 Structural specifications of collagen binding	11
1.3.3 Leukocyte integrins coming into view	13
1.3.4 Pathogenic exploitation of the I domain	16
1.3.5 Ligand recognition beyond the I domain	17
1.4 Protein domains as evolutionary fodder	18
1.5 Thesis overview	20
1.6 List of publications	20
Chapter 2 Molecular exaptation by the integrin αI domain	22
2.1 Abstract	22
2.2 Introduction	22
2.3 Results	23
2.3.1 Integrin $\alpha 4\beta 7$ adopts distinct conformations	23
2.3.2 The inactive $\alpha E\beta 7$ I domain is unexpectedly dynamic	25
2.3.3 Ligand binding stabilizes the open $\alpha E\beta 7$ I domain	27
2.3.4 The integrin I domain was derived from collagen	29
2.3.5 The allosteric internal ligand was a de novo molecular exaptation	31
2.4 Discussion	33
2.5 Materials and methods	34
2.5.1 DNA constructs	34
2.5.2 Protein expression and purification	34
2.5.3 nsEM sample preparation	36
2.5.4 nsEM data acquisition and processing	36
2.5.5 cryoEM sample preparation	37
2.5.6 cryoEM data acquisition and processing	37

2.5.7	<i>Model building</i>	38
2.5.8	<i>System setup for MD simulations</i>	38
2.5.9	<i>MD protocols</i>	39
2.5.10	<i>Biolayer interferometry</i>	39
2.5.11	<i>Sequence curation and HMM searching</i>	40
2.5.12	<i>Sequence alignments</i>	40
2.5.13	<i>Flow cytometry</i>	40
2.5.14	<i>Phylogenetic analysis</i>	41
2.5.15	<i>Ancestral sequence reconstruction</i>	41
2.6	<i>Supplemental information</i>	42
2.7	<i>Afterword</i>	57
Chapter 3	<i>Avidity unlocks promiscuity in a divergent integrin receptor family</i>	58
3.1	<i>Abstract</i>	58
3.2	<i>Introduction</i>	58
3.3	<i>Results</i>	59
3.3.1	<i>Distinct evolutionary trajectories demarcate integrin behavior</i>	59
3.3.2	<i>The specific integrin $\alpha L\beta 2$ occupies a fragile mutational niche</i>	61
3.3.3	<i>Molecular tradeoffs permeate the promiscuous integrin $\alpha M\beta 2$</i>	63
3.3.4	<i>NIF co-opts an essential αM ion coordination site</i>	66
3.3.5	<i>Noncanonical ligand binding by iC3b to $\alpha M\beta 2$ reveals pathways to promiscuity</i>	68
3.4	<i>Discussion</i>	70
3.5	<i>Materials and methods</i>	71
3.5.1	<i>DNA constructs for protein expression</i>	71
3.5.2	<i>Protein expression and purification</i>	71
3.5.3	<i>nsEM sample preparation</i>	72
3.5.4	<i>nsEM data acquisition and processing</i>	72
3.5.5	<i>cryoEM sample preparation</i>	73
3.5.6	<i>cryoEM data acquisition and processing</i>	73
3.5.7	<i>Model building</i>	74
3.5.8	<i>Sequence alignments and phylogenetic analyses</i>	74
3.5.9	<i>Positive selection analyses</i>	74
3.5.10	<i>Flow cytometry</i>	75
3.5.11	<i>Mutant library construction</i>	75
3.5.12	<i>Deep mutational scanning for integrin-ligand binding affinity</i>	77
3.5.13	<i>Structure and data visualization</i>	78
3.6	<i>Supplemental information</i>	79
3.7	<i>Afterword</i>	88
Chapter 4	<i>Remote homology and functional genetics unmask deeply preserved Scm3/HJURP orthologs in metazoans</i>	89
4.1	<i>Abstract</i>	89
4.2	<i>Introduction</i>	89
4.3	<i>Results</i>	91
4.3.1	<i>HJURP orthologs are broadly conserved across deuterostomes</i>	91

4.3.2	<i>The zebrafish gene si:dkeyp-117h8.4 encodes HJURP</i>	93
4.3.3	<i>Drosophila CAL1 is a true HJURP ortholog</i>	95
4.3.4	<i>Nematodes encode HJURP proteins</i>	98
4.3.5	<i>C. elegans nefr-1 encodes a bona fide HJURP ortholog</i>	100
4.3.6	<i>Scm3/HJURP orthologs are broadly conserved across Metazoa</i>	102
4.4	<i>Discussion</i>	105
4.5	<i>Materials and methods</i>	106
4.5.1	<i>HJURP candidate searching and selection criteria</i>	107
4.5.2	<i>Structure analyses</i>	108
4.5.3	<i>Sequence alignments</i>	108
4.5.4	<i>Phylogenetic and dN/dS analyses</i>	108
4.5.5	<i>Recombinant protein expression and purification</i>	108
4.5.6	<i>Amylose pulldown</i>	109
4.5.7	<i>Zebrafish care and maintenance</i>	109
4.5.8	<i>Zebrafish CRISPR generation</i>	110
4.5.9	<i>Zebrafish fluorescence microscopy</i>	110
4.5.10	<i>C. elegans maintenance and genetics</i>	110
4.5.11	<i>C. elegans CRISPR/Cas9 genome modification</i>	111
4.5.12	<i>C. elegans fluorescence microscopy</i>	111
4.5.13	<i>C. elegans NEFR-1 immunoprecipitation</i>	111
4.6	<i>Supplemental information</i>	113
4.7	<i>Afterword</i>	125
Chapter 5	<i>Concluding remarks</i>	126
5.1	<i>Abstract</i>	126
5.2	<i>From building blocks to functional machines</i>	126
5.3	<i>Everything in its context</i>	127
Appendix I	<i>How do integrin subunits find their partners?</i>	128
I.1	<i>Abstract</i>	128
I.2	<i>Main text</i>	128
I.3	<i>Methods</i>	129
Appendix II	<i>The structural basis of orthomyxovirus restriction by MxA</i>	131
II.1	<i>Abstract</i>	131
II.2	<i>Main text</i>	131
II.3	<i>Methods</i>	132
Appendix III	<i>Rapid evolution of the sperm-egg fertilization complex</i>	134
III.1	<i>Abstract</i>	134
III.2	<i>Main text</i>	134
III.3	<i>Methods</i>	135
References		138

List of Figures

Figure 1.1	<i>Integrin nomenclature and conformational background</i>	2
Figure 1.2	<i>Integrin I domain nomenclature and activation</i>	10
Figure 1.3	<i>Overview of published I domain structures</i>	12
Figure 1.4	<i>I domain integrins beyond canonical interactions</i>	17
Figure 2.1	<i>Integrin $\alpha4\beta7$ adopts distinct conformations</i>	24
Figure 2.2	<i>The half-bent $\alpha E\beta7$-inactive structure is dynamic</i>	26
Figure 2.3	<i>The molecular interface between integrin $\alpha E\beta7$ and E-cadherin</i>	28
Figure 2.4	<i>The integrin I domain was co-opted from an ancient collagen gene</i>	30
Figure 2.5	<i>The internal ligand was exapted for integrin activation</i>	32
Supplemental Figure 2.1	<i>Data processing schematic for the $\alpha4\beta7$:MAdCAM-1 complex</i>	42
Supplemental Figure 2.2	<i>The molecular interface between integrin $\alpha4\beta7$ and MAdCAM-1</i>	43
Supplemental Figure 2.3	<i>Integrin $\alpha4\beta7$ has a novel compact state</i>	44
Supplemental Figure 2.4	<i>Integrin $\alpha E\beta7$ occupies canonical conformations</i>	45
Supplemental Figure 2.5	<i>Data processing schematic for the $\alpha E\beta7$:LF61 fab complex</i>	46
Supplemental Figure 2.6	<i>Large-scale and small-scale movement characterization of apo $\alpha E\beta7$</i>	47
Supplemental Figure 2.7	<i>Data processing schematic for the $\alpha E\beta7$:EC12 complex</i>	48
Supplemental Figure 2.8	<i>Molecular features of the open integrin $\alpha E\beta7$</i>	49
Supplemental Figure 2.9	<i>Evolutionary origins of the integrin I domain</i>	50
Supplemental Figure 2.10	<i>Molecular dynamics simulations reveal conserved dynamics in vWFA domains</i>	51
Supplemental Figure 2.11	<i>Phylogenetic and structural analyses of the internal ligand origin</i>	52
Supplemental Figure 2.12	<i>Conformational landscape of the $\alpha L\beta2$-ICAM-1 and $\alpha E\beta7$-E-cadherin headpiece domains</i>	53
Supplemental Figure 2.13	<i>Map quality and validation metrics</i>	54
Figure 3.1	<i>Diverging evolutionary trajectories of the integrin αI domain</i>	60
Figure 3.2	<i>The αL:ICAM-1 interaction is mutationally fragile</i>	62
Figure 3.3	<i>Molecular tradeoffs permeate the promiscuous αM I domain</i>	64
Figure 3.4	<i>NIF closes the αM I domain</i>	67
Figure 3.5	<i>Avidity generates novelty in ligand binding by $\alpha M\beta2$</i>	69
Supplemental Figure 3.1	<i>Rapid evolution of a subset of integrin α subunits</i>	79
Supplemental Figure 3.2	<i>A high throughput yeast display assay to quantify the effects of mutations across the $\alpha L\beta2$ I domain</i>	80
Supplemental Figure 3.3	<i>The mutational constraints of the $\alpha M\beta2$:C3d interaction</i>	81

Supplemental Figure 3.4	<i>The mutational constrains of the $\alpha M\beta 2$:LGH interaction</i>	82
Supplemental Figure 3.5	<i>The mutational constrains of the $\alpha M\beta 2$:NIF interaction</i>	83
Supplemental Figure 3.6	<i>Structural determination of the $\alpha M\beta 2$:NIF interaction</i>	84
Supplemental Figure 3.7	<i>Structural determination of the $\alpha M\beta 2$ homodimer</i>	85
Supplemental Figure 3.8	<i>Structural determination of the $\alpha M\beta 2$:iC3b interaction</i>	86
Figure 4.1	<i>The centromeric chaperone HJURP predates tetrapods</i>	92
Figure 4.2	<i>Danio rerio encodes an HJURP ortholog</i>	95
Figure 4.3	<i>CAL1 is a true Scm3/HJURP ortholog</i>	97
Figure 4.4	<i>Nematode genomes encode a canonical HJURP protein</i>	99
Figure 4.5	<i>C. elegans NEFR-1 is an essential cell division regulator</i>	101
Figure 4.6	<i>HJURP is present in basal metazoa</i>	103
Supplemental Figure 4.1	<i>Structure-function relationships of the human HJURP protein</i>	113
Supplemental Figure 4.2	<i>HJURP orthologs had not previously been found in most metazoan clades</i>	113
Supplemental Figure 4.3	<i>A schematized approach for detecting remote HJURP orthologs</i>	114
Supplemental Figure 4.4	<i>Shared and diverged features of deuterostome HJURPs</i>	115
Supplemental Figure 4.5	<i>Related to in vivo analysis of HJURP</i>	116
Supplemental Figure 4.6	<i>Shared and diverged sequence features of CAL1</i>	117
Supplemental Figure 4.7	<i>An expanded phylogeny of panarthropod HJURP orthologs</i>	118
Supplemental Figure 4.8	<i>NEFR-1 is a nematode cell cycle protein</i>	119
Supplemental Figure 4.9	<i>nefr-1 is an essential gene</i>	120
Supplemental Figure 4.10	<i>NEFR-1 physically associates with CENPA^{HCP-3}</i>	121
Supplemental Figure 4.11	<i>scm3 domain-containing proteins are present in most metazoan clades</i>	122
Supplemental Figure 4.12	<i>AlphaFold predictions colored by pLDDT score</i>	122
Figure I.1	<i>SEL1L may contribute to integrin heterodimer formation</i>	130
Figure II.1	<i>nsEM of Thogotovirus NP and MxA</i>	132
Figure III.1	<i>Positive selection across the fertilization complex</i>	136

List of Tables

Table 1.1	<i>A list of integrins and their endogenous ligands</i>	5
Table 1.2	<i>A list of notable published I domain integrin structures</i>	13
Supplemental Table 2.1	<i>Collection, refinement and validation statistics for cryoEM density maps and models</i>	55
Supplemental Table 2.2	<i>Molecular dynamics systems simulated in this study</i>	56
Supplemental Table 3.1	<i>Collection, refinement and validation statistics for cryoEM density maps and models</i>	87
Supplemental Table 4.1	<i>A representative list of scm3 domain-containing proteins across the opisthokont lineage</i>	123
Supplemental Table 4.2	<i>Nucleic acid constructs used for this study</i>	124
Supplemental Table 4.3	<i>A list of C. elegans strains generated and/or used in this study</i>	124
Table III.1	<i>Positive selection analysis of the Murinae fertilization complex</i>	137

List of Abbreviations

$\alpha(x)\beta(x)$	Reference to integrin name; each x specifies the subunit
ACT	<i>Bordetella</i> adenylate cyclase toxin
ADMIDAS	Adjacent to the MIDAS ion coordination site
BC	Bent closed
BLI	Bio-layer interferometry
Ca^{2+}	Calcium divalent cation
Cal1	Chromosome alignment defect 1; <i>D. melanogaster</i> HJURP ortholog
cd11(x)	Alternative name to alpha integrins associating with $\beta 2$
cd18	Alternative name for integrin $\beta 2$
cd103	Alternative name for integrin αE
CENPA	Centromeric Protein A
CR3/4	Complement receptor 3/4, alternative names for $\alpha M\beta 2$ and $\alpha X\beta 2$
cryoEM	Cryogenic electron microscopy
DMS	Deep mutational scan
dpf/hpf	Days/hours post-fertilization
EC	Extended closed
EC12	E-Cadherin truncation protein with only the first 2 domains
ECM	Extracellular matrix
EM	Electron microscopy
EO	Extended open
ER	Endoplasmic reticulum
ERAD	ER-associated degradation
FRET	Fluorescence resonance energy transfer
GFOGER	Gly-Phe-HydroxyPro-Gly-Glu-Arg peptide
GO	Gene ontology analysis
HJURP	Holliday Junction Recognition Protein
HMM	Hidden Markov Model
ICAM-1	Intercellular Adhesion Molecule 1
IP-MS	Immunoprecipitation-Mass Spectrometry
itga(x)	Genetic nomenclature for integrin alpha subunits
itgb(x)	Genetic nomenclature for integrin beta subunits
kDa	Kilodalton
LFA-1	Lymphocyte Function-Associated Antigen 1, alternative to $\alpha L\beta 2$
LGH	Leukocidin GH
LPAM-1	Lymphocyte Peyer's Patch Adhesion Molecule 1, alternative to $\alpha 4\beta 7$
Mac1	Macrophage 1 antigen, alternative to $\alpha M\beta 2$
MAdCAM-1	Mucosal Addressin Cell Adhesion Molecule 1
Mg^{2+}	Magnesium divalent cation
MIDAS	Metal ion-dependent adhesion site
Mn^{2+}	Manganese divalent cation

<i>NEFR-1</i>	<i>Neurofilament-related 1, C. elegans HJURP ortholog</i>
<i>NIF</i>	<i>Neutrophil inhibitory factor</i>
<i>nsEM</i>	<i>Negative stain electron microscopy</i>
<i>PAMP</i>	<i>Pathogen-associated molecular pattern</i>
<i>RGD</i>	<i>Arg-Gly-Asp peptide</i>
<i>RGLD</i>	<i>Arg-Gly-Leu-Asp</i>
<i>RMSD</i>	<i>Root mean square deviation</i>
<i>scm3</i>	<i>Suppressor of chromosome missegregation 3</i>
<i>SDL</i>	<i>Specificity determining loop</i>
<i>SEC</i>	<i>Size exclusion chromatography</i>
<i>SEL1L</i>	<i>Suppressor of lin-12-like</i>
<i>SortSeq</i>	<i>Method for yeast display-linked deep mutational scanning</i>
<i>SPARK</i>	<i>Sperm protein assembly and receptor-binding key</i>
<i>SyMBS</i>	<i>Synergistic metal binding site</i>
<i>TGF-β</i>	<i>Transforming growth factor beta</i>
<i>T_{rm}</i>	<i>Tissue resident memory cell</i>
<i>UMAP</i>	<i>Uniform manifold approximation and projection</i>
<i>vWFA</i>	<i>von Willebrand factor A</i>

Chapter 1 Introduction

**sections 1.2 and 1.3 are written to be adapted as a book chapter focusing on structural mechanisms of I domain-containing integrin function*

1.1 Abstract

Protein domains serve as individual functional units, the recombination of which contributes to the great and beautiful diversity of life forms we see today. In this thesis I focus on two such domains – the integrin α I domain and the histone chaperone scm3 domain – as models for understanding how protein domains evolve. The integrin α I domain fundamentally changed the ligand binding capacities of the integrin cell surface receptor family, while still maintaining ancestral signaling capabilities despite its risky placement. Although α I domains share many core features, they have diversified greatly in their ligand recognition profiles and biochemical properties. Conversely scm3 domains, completely essential for faithful chromosome segregation, have evolved beyond homology recognition at the sequence level despite retaining the exact same core function across *Metazoa* evolution. Studying these two domains thus provides flip sides of the same coin by revealing the many pathways to adaptation while illuminating a surprising mutational tolerance for a conserved function. This chapter introduces the fundamental background information on these domains specifically, as well as the evolvability of protein domains in general, necessary for synthesizing the proceeding chapters in this thesis.

1.2 Integrins and their structure-function relationships

The cell surface membrane, in addition to creating a physically isolated incubator to facilitate biochemical reactions, faces the daunting task of communicating conditions of the exterior environment to the cell interior. To do this, cells express cell surface receptors: membrane-anchored proteins that can discriminate the vast physiochemical world and subsequently transmit signals to the cell interior in response to these conditions(1). Integrins are α/β heterodimeric receptors that bind to extrinsic ligands and signal that ligand binding to the cell interior through a series of coordinated conformational changes(2–4). Integrins can further signal bi-directionally, possessing the ability to communicate both “inside-out” and “outside-in”(3, 5). The dynamic properties and adaptable nature of the integrin machinery outlined here make them exceptionally versatile receptors that have been tuned for a wide range of cellular and organismal functions.

1.2.1 A structural perspective on integrin function

Each integrin subunit is comprised of a large (~125-200 kDa), glycosylated extracellular domain, a single pass transmembrane helix, and a short cytoplasmic tail that serves as a hub for intracellular signaling(6) (Fig. 1.1A). The alpha (α) subunit ectodomain contains a seven-bladed beta-propellor domain which mediates contacts with both ligands and its associated beta (β) subunit, an adjacent thigh domain, and two calf domains leading to the transmembrane helix. Additionally, nine (α 1, α 2, α 10, α 11, α E, α L, α M, α D, and α X) of the eighteen human α integrins contain an additional domain called the “inserted” or “I” domain, inserted between blades 2 and 3 of the beta-propellor, that sits distal to the cell surface and mediates ligand

binding in integrins which contain it(7). The β subunit is comprised of a β I domain, which serves an analogous capacity in ligand binding and heterodimer formation as the α propeller domain as well as coordinating conformational changes leading to integrin activation. It also contains a hybrid or “thigh” domain followed by a PSI domain, four EGF-like domains and a tail domain proximal to the transmembrane helix(6). Although seven of the eight human β subunits contain a short cytoplasmic tail, β 4 contains an extended, 1100-amino acid intracellular region comprised of a calx-beta domain and four fibronectin-like repeats(8, 9). Collectively integrins pair in 24 combinations in the human body, far from the number of possible heterodimers. It is generally understood that the α subunit contributes primarily to integrin-ligand specificity(10), but the β I domain contains a flexible specificity-determining loop (SDL) which further generates molecular precision(11–13).

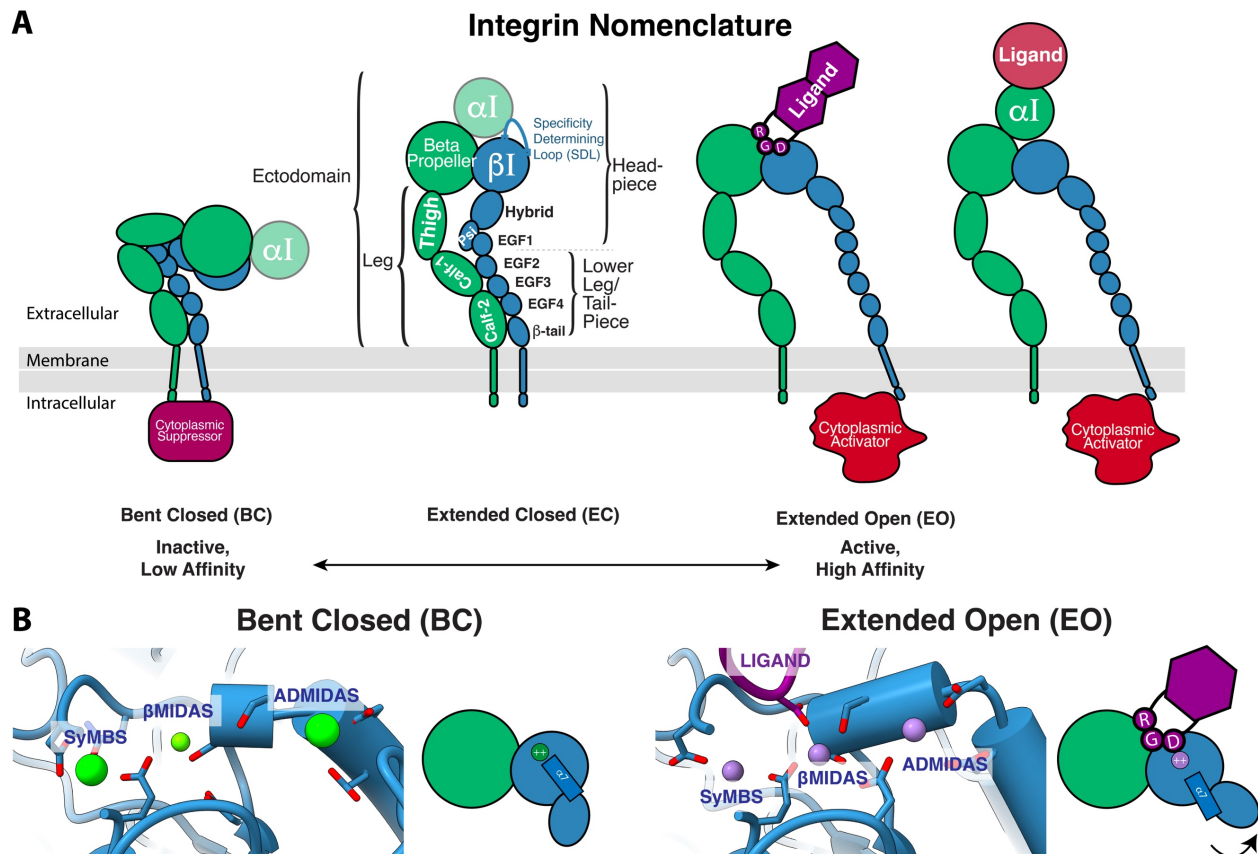


Figure 1.1. Integrin nomenclature and conformational background. (A) A schematic representation of integrin organization and conformational changes. Integrin conformations range from a compact, inactive bent closed state with low ligand affinity (left) to an extended-open, active state with high ligand affinity (right). Approximate hypothesized I domain locations are shown as semi-transparent. These states may be further modified and/or stabilized by both external ligand and cytoplasmic effector binding. (B) Ion coordination controls the structural transition between integrin conformations. Integrins tend to prefer the bent closed conformation, but upon ligand binding the β I β MIDAS is coordinated by the acidic residue contained in the canonical “RGD” motif, causing a pistol-like swingout motion of the β calf domain and propagating an open integrin ectodomain. Structures are adapted from (14).

An extensive amount of structural characterization over the last 3 decades has provided a wealth of mechanistic insight into how integrins operate at the single-molecule level. The first apo integrin ectodomain structure revealed a “closed” or “bent” conformation (“bent closed”,

BC), in which the subunits folded onto themselves to form a compact, clamshell-like shape(15). The α subunit folds between the thigh and calf-1, while the β subunit folds at the intersection of the PSI and EGF-1 domains. This compact shape has been confirmed across the integrin lineage via crystallographic, electron microscopic, and FRET-based structural analyses(6). These conformations also hold true for the integrins which contain the α I domain, suggesting a conserved inactive conformation is critical for integrin function(16). Cellular assays leveraging conformation-specific antibodies have shown that this closed conformation accounts for the vast majority of integrins on the cell surface(17) and is considered to be the inactive, or non-signaling, form of integrin. Subsequent single-particle electron microscopy (EM) studies revealed an alternate open, or active, conformation (“extended open”, EO), in which the integrin legs unfold and open in an extended shape with the beta-propellor and β I domains occupying a position far distal to the transmembrane region(18–20). These studies were further refined into a model in which integrins dynamically fluctuate between three main conformations, with characterization of an extended state with legs closed (“extended closed”, EC) serving as an intermediate between the closed and open conformations(21) (Fig. 1.1A). These conformations have all been natively validated with multiple cell lines and multiple unique integrin heterodimers, suggesting this model is relatively generalizable(22). Still, there are exceptions. For example, α V β 8 has lost the ability to change conformations and seems to signal within the confines of the EC conformation(23, 24). EM studies have also suggested the closed conformation exists with variation, as some integrins prefer an acute bend while others demonstrate a right-angle (“half”) bend(14, 16).

Structural characterization at high resolution has further pinpointed the mechanism by which integrins occupy these conformations. Integrins contain a plethora of divalent cation binding sites which contribute to function(25–27). The beta-propellor contains at least 3 Ca^{2+} -coordinating EF hand repeats, the thigh domain a single Ca^{2+} site, and the β I domain three canonical ion sites: the central MIDAS (Metal Ion-Dependent Adhesion Site), the ADMIDAS (Adjacent to the MIDAS), and the SyMBS (Synergistic Metal Binding Site, also known as the LIMBS). The contribution of the beta-propellor sites remains incompletely understood, although they appear to play important roles in both integrin biogenesis and ligand binding(10, 28–30). Conversely, the three β I sites appear to be the critical ions for integrin conformational change, as mutating any of them has profound impacts on ligand binding and cellular signaling *in vitro*, in cell culture, and *in vivo*. Mutating the MIDAS invariably abolishes ligand binding(31, 32) and generates an entirely inactive, closed integrin, while mutating either the SyMBS or ADMIDAS has varying effects on integrin function(33–36). It is generally thought, however, that the SyMBS serves as a positive regulator of ligand binding(37, 38) whereas the ADMIDAS acts as a negative regulator(39).

Upon complete coordination of the ion MIDAS sphere by an acidic residue contained within an integrin-binding ligand, the β I domain executes a stereotypical piston-like movement, in which the α 1 helix within the β I domain rigidifies and the α 7 helix is shifted into a downward position. This leads to a 90° shift in the positioning of the hybrid domain from the inward-facing closed position to the outward open position, thus serving as the structural basis of integrin activation(40) (Fig. 1.1B). This generalized mechanism can be supplemented by ligand-specific interactions that provide additional synergy to the repositioning of the β I ion cluster, allowing tuning of the integrin activation threshold(26, 27, 32). This swingout propagates through the rest of the integrin leg, leading to separation of the transmembrane helices and allowing subsequent signaling through the exposed individual cytoplasmic tails. This is referred to as “outside-in”

signaling and appears to constitute most integrin activation(41, 42). “Inside-out” signaling further contributes with cytosolic factors sequestering individual subunit tails, leading to destabilization of the transmembrane helix interface and encouraging the integrin molecule to unfold and occupy the EO conformation(18, 43–45). Recent cryoEM studies support an intermediate transmembrane region with transient separation of the membrane helices creating a pathway allowing for this inside-out signaling state(46, 47).

1.2.2 *The many integrin ligands and their duties*

Integrins are particularly notable in terms of the exceptionally vast, structurally diverse ligands they recognize(11, 48). Integrins most canonically recognize various components of the extracellular matrix (ECM)(49), ultimately leading to the recruitment of components that collectively constitute a focal adhesion(50, 51). Although many integrins recognize multiple proteinaceous ECM constituents, studies have revealed preferences for different integrin classes. For example, $\alpha 5 \beta 1$ has become colloquially known as the fibronectin receptor(52), $\alpha V \beta 3$ as the vitronectin receptor(53), and $\alpha 3 \beta 1$, $\alpha 6 \beta 1$, $\alpha 7 \beta 1$, and $\alpha 6 \beta 4$ as the laminin receptors(54, 55).

The ECM-binding integrins, owing to their vast ligand recognition capabilities and unique tissue expression profiles, have varying functions within the whole organism (Table 1.1). As the vitronectin receptor, $\alpha V \beta 3$ contributes greatly to angiogenesis and vascular cell migration, although its secondary recognition of fibronectin, thrombospondin, tenascin and others also implicates it in processes as diverse as leukocyte cell migration and inflammation to osteoclast adhesion(11, 56–58). Although $\alpha V \beta 5$ shares the α subunit with $\alpha V \beta 3$, its differing expression profiles implicate it more heavily in wound healing, axon formation, and phagocytosis of apoptotic cells(59–63). Integrins $\alpha V \beta 6$ and $\alpha V \beta 8$ serve as important activating receptors of the pleiotropic immune cytokine TGF- β which variably lead to immune cell activation and innate immune(13, 64, 65) surveillance, stem cell differentiation, and regulation of inflammation. The last αV integrin, $\alpha V \beta 1$, seems most critical for ECM remodeling and connective tissue formation with overlapping binding to osteopontin, TGF- β , and fibronectin(66–68). $\alpha 8 \beta 1$ has similar overlap with these profiles, playing key roles in lung and kidney growth and homeostasis(69). $\alpha IIb \beta 3$ is strictly limited to the blood platelet lineage, and binds to several ECM components to control blood adhesion and clotting(70, 71).

The laminin-binding integrins primarily bind to the basement membrane ECM constituent laminin, and tune their respective functions via both cell-type expression preferences and molecular kinetic preferences between the 16 human laminin isoforms(54, 72, 73). $\alpha 3 \beta 1$ plays a central role in fundamental organ development and its knockout leads to early mouse neonatal lethality(74, 75). $\alpha 7 \beta 1$ is primarily restricted to bone and muscle tissue, where it mediates connections between muscle fibers and tendon junctions via interactions with the basement membrane(76, 77). The function of $\alpha 6 \beta 1$ is less clear, although it is implicated in angiogenesis as well as ovarian processes(78–80). Finally, $\alpha 6 \beta 4$ is most notably involved in the formation of hemidesmosomes and mediates skin attachment to the basement membrane(81, 82). The collagen-binding integrins operate under similar divisions of labor, using expression and collagen subtype preferences to distinguish function, although they generally share bone and cartilage development as a theme(83–85). $\alpha 1 \beta 1$ and $\alpha 2 \beta 1$ have generalized binding capability, involved in both immune cell development and bone development and repair(86–88). $\alpha 10 \beta 1$ and $\alpha 11 \beta 1$ play more limited roles in chondrogenesis and bone healing, respectively(89–92). These

collagen-binding integrins canonically recognize the GFOGER motif within the collagen tri-helix(93), although they are implicated in secondary ECM binding to varying degrees independent of this recognition pattern(85).

Table 1.1. A list of integrins and their endogenous ligands.

Integrin Ligands		
Subtype	Integrin	Ligands
RGD Receptors	$\alpha V\beta 1$	TGF- β , fibronectin, osteopontin, L1CAM
	$\alpha V\beta 3$	vitronectin, TGF- β , fibronectin, osteopontin, L1CAM, fibrinogen, von Willebrand factor, thrombospondin, fibrillin, tenascin
	$\alpha V\beta 5$	TGF- β , osteopontin, vitronectin, bone sialic protein, thrombospondin, CCN3
	$\alpha V\beta 6$	TGF- β , fibronectin, osteopontin, ADAM
	$\alpha V\beta 8$	TGF- β
	$\alpha 5\beta 1$	Fibronectin, fibrinogen, fibrillin, osteopontin, thrombospondin
	$\alpha 8\beta 1$	TGF- β , tenascin, fibronectin, osteopontin, vitronectin, nephronectin
Laminin Receptors	$\alpha IIb\beta 3$	fibrinogen, fibronectin, thrombospondin, vitronectin
	$\alpha 3\beta 1$	laminin-332, laminin-511
	$\alpha 6\beta 1$	laminin-111, laminin-511, laminin-332
	$\alpha 6\beta 4$	laminin, BP, BPAG1-e, BPAG2
Collagen Receptors	$\alpha 7\beta 1$	laminin-211, laminin-221
	$\alpha 1\beta 1$	collagens I, III, IV, IX, XIII, XVI
	$\alpha 2\beta 1$	collagens I, III, IV, V, XI, XVI, XXIII
	$\alpha 10\beta 1$	collagens I, II, IV, VI, IX, X, XI
Leukocyte Receptors	$\alpha 11\beta 1$	collagens I, IV
	$\alpha 4\beta 1$	VCAM-1, fibronectin
	$\alpha 4\beta 7$	MAdCAM-1, VCAM-1
	$\alpha 9\beta 1$	OPN, Tenascin-C, VCAM-1, Thrombospondin ADAMs
	$\alpha E\beta 7$	E-cadherin
	$\alpha D\beta 2$	Fibrinogen, plasminogen, Cyr61, fibronectin, vitronectin, VCAM-1
	$\alpha L\beta 2$	ICAMs 1-5, JAM-1, ESM
	$\alpha M\beta 2$	ICAMs 1-4, VCAM-1, JAM-3, Thy-1, DC-SIGN, CD40L, fibrinogen, Factor X, GPbI alpha, heparin, ssDNA, dsRNA, LPS, collagen, fibronectin, vitronectin, plasminogen, iC3b, C3d
	$\alpha X\beta 2$	ICAM-1, ICAM-4, Thy-1, VCAM-1, iC3b

Finally, the leukocyte integrins are a diverse collection of receptors that mediate white blood cell differentiation, adhesion, homing, activity, and regulation(94). Integrins $\alpha 4\beta 7$ and $\alpha E\beta 7$ show the most potent phenotypes in gut leukocyte homing and retention(95), where they bind to MAdCAM-1(96) and E-cadherin(97), respectively. Similarly, $\alpha 4\beta 1$ mediates rolling and extravasation of leukocytes to sites of immune activity by binding primarily to the broadly-expressed VCAM-1(98). In contrast, $\alpha 9\beta 1$ has a more pleiotropic ligand profile that includes VCAM-1 as well as disintegrins, metalloproteases, Tenascin-C and osteopontin, and as such its functions appear to be more universal in preventing autoimmunity(99, 100). Lastly, the $\beta 2$ integrins $\alpha D\beta 2$, $\alpha L\beta 2$, $\alpha M\beta 2$ and $\alpha X\beta 2$ play roles most prominently in mediating leukocyte homing to, recognition, and clearing of pathogenic invaders(101, 102). They most notably bind to ICAM-1, but $\alpha M\beta 2$ and $\alpha X\beta 2$ also serve as complement receptors and have exceptionally promiscuous ligand binding profiles(103, 104).

Although many of these interactions have been described *in vivo* through sledgehammer-style manipulation, there remains a considerable gap in our understanding of the biochemical and structural bases for integrin-ligand communication. This is in part due to the inherent limitations of classical techniques like protein crystallography(105), which are refractory to the mostly large, multivalent, heavily glycosylated, highly flexible exchanges between integrins and their ECM ligands. More recently cryoEM has unleashed new potential in observing integrins in more native contexts, allowing high-resolution mapping of key protein interfaces in combination with their downstream effects on integrin conformation. Landmark studies of $\alpha 5\beta 1$:fibronectin(14) and $\alpha V\beta 8$:TGF- β (24), for example, have shifted paradigms of classical biological models. The details of these interactions will be discussed in subsequent sections, but these works have made it clear that advances in structural techniques will continue to benefit the study of basic integrin biology. A more refined understanding of the idiosyncrasies between integrin-ligand relationships is paramount to successful therapeutic targeting, as many integrin therapeutics fail due to off-target, pleiotropic effects.

1.2.3 Ancestral ligand binding and the RGD cleft

Although integrins vary widely in the ligands they interact with, the classical and most well-studied minimal binding motif is the Arginine-Glycine-Aspartate (RGD) tripeptide(106), which is present in the major integrin-binding ECM components including fibronectin, vitronectin, fibrinogen, osteopontin, and thrombospondin, among others. This motif was first characterized in 1984 through a series of truncation mutant experiments, which ultimately found that a RGDS motif (including the core RGD residues) was both necessary and sufficient to induce cell adhesion to fibronectin(107). This motif was quickly leveraged to identify the responsible cellular receptors, integrins, which were found broadly across the animal tree to recognize this same motif(108–110). The first structural study of the $\alpha V\beta 3$ ectodomain in complex with the RGD peptide beautifully uncovered the mechanism of RGD recognition: the RGD motif linearly spans the interface between the two integrin subunits, with the arginine forming salt bridges with the top of the αV beta propellor, the glycine bridging the gap between the αV and $\beta 3$, and the aspartate coordinating the MIDAS cation within $\beta 3$ (111). Subsequent studies with other integrin ectodomains revealed structural variations on this theme, yet the core architecture remained consistent(40, 112). This also lent substantial support to the prevailing hypothesis that MIDAS coordination triggered integrin opening; this was confirmed in subsequent crystallographic studies showing the $\beta 3$ hybrid domain swung out when bound to an

integrin agonist(34, 40). Negative stain EM studies corroborated these findings in the context of the full integrin ectodomain, in which this swingout was demonstrated to propagate through the entire integrin molecule(6).

Despite this shared minimal recognition motif, variations on the RGD have differing effects depending on the integrin analyzed. Motifs that subtly differ in sequence, such as RGDS, GRGD, cyclic(RGDfV), and the drug Cylengitide (cyclic(RGDf(NMe)v)), have dramatically different affinities for different integrins. In addition, integrins have specificity-determining mechanisms beyond the RGD motif that allow them to fill their physiological niches with differing ligands. Single-particle cryoEM studies have begun to resolve these idiosyncratic relationships by revealing in high resolution the structural basis for RGD ligand specificity(113). The continued study of the immune cytokine TGF- β with integrins α V β 6 and α V β 8 has proven especially fruitful in this endeavor(24, 114); cryoEM analysis showed that both the RGD and adjacent helix-forming LXXI/L motif contribute to ligand binding, with additional support by the β 8 specificity-determining loop in α V β 8(24). Subsequent analyses using global entropy models demonstrated that structural flexibility within the RGD LXXI/L is fundamental to TGF- β activation, suggesting that a labile topology of RGD can shift the conformational constraints between local and distant regions of ligands and lead to differential outcomes(115). Further, structural characterization of the α 5 β 1: fibronectin complex revealed an extensive interface along the top of both subunits(14), suggesting an architecture with vast special constraint contributes to integrin-ligand specificity. These α 5 β 1 studies further revealed the prerequisite of ligand binding to induce the EO integrin conformation, a feature not shared among all integrins; this establishes yet another layer of specificity tuning.

Some integrins also recognize non-RGD motifs; nevertheless, their structures beyond the have also supported shared mechanisms governing integrin activation(116). For example, the structure of α 6 β 1 in complex with laminin-511 demonstrated that an EKP motif within the laminin γ subunit served a structurally analogous role to RGD, with E1607 coordinating the β 1 MIDAS and P1609 interacting closely with α 6. This interaction was further supported by extensive contacts between the laminin α and β subunits with integrin α 6, implicating the integrin alpha subunit in providing specificity to integrin-laminin interactions(117). Likewise structures of α 4 β 7: MAdCAM-1(118) and α 4 β 1: VCAM-1(119) revealed dependencies on RGD-like motifs that bridge the integrin subunits (RGLD in MAdCAM-1 and QID in VCAM-1), yet these structures show structural clefts which run perpendicular to that of fibronectin and α 5 β 1, suggesting malleability within ligand orientation along the integrin subunits. Thus, although idiosyncrasy dictates specificity, emerging general trends link these studies into a cohesive body of work, revealing broad MIDAS-dependent strategies that appear to universally modulate integrin ectodomain conformation.

1.2.4 Structural insights into intracellular integrin signaling

Although dwarfed in molecular weight by orders of magnitude compared to the ectodomain, the integrin transmembrane and cytoplasmic domains serve equally critical functional roles in integrin signaling(120, 121). Unlike the well-structured domains within the integrin extracellular region, the cytoplasmic tails are significantly flexible and can adopt unstructured, helical, or sheet conformations depending on their binding partners(122–124). The first and most well-studied cytoplasmic interactor is talin. Talin interacts with NPxY motifs in β subunit tails to physically link integrins to the actin cytoskeleton, ultimately causing changes in

cellular morphology and behavior(125). Structural studies show this motif adopts secondary structure upon talin FERM3 domain binding, further supported with by an aromatic stabilizing interaction and electrostatic interactions which reinforce the F3 domain's upward-facing conformation proximal to the cell membrane(126–128). This positions talin to form contacts with stabilizing lipids and promotes a more sterically favorable interface with integrin, leading to positive feedback and the generation of focal adhesion machinery.

Outside of talin, kindlin is becoming increasingly better characterized as an integrin tail activator responsible for reinforcing the focal adhesion machinery(129, 130). Structural insights into the kindlin2 FERM domain complex with $\beta 1$ and $\beta 3$ integrin tails showed that a distal NPxY motif mediates their binding in a structurally analogous way to that of talin, while secondary contacts with serine/threonine-rich motifs reinforce their stability(131). Beyond this, both kindlin multimerization and tertiary interactions between kindlin and talin concentrate integrins into clusters, which can subsequently form the higher-order multivalent interactions with ECM ligands essential for focal adhesion maturity(132). The role of both tyrosine and serine/threonine phosphorylation on these integrin tails is also becoming more appreciated, as this provides a quick means of modulating signaling, most notably by preventing proteasome-mediated turnover(133, 134). As such several kinases, such as Src, CaMKII and Cdk5, have been implicated as tail binders(135–137).

In addition to these activating molecules, several intracellular suppressors of integrin function have been characterized. Structurally, the filamin A-integrin complex interacts specifically with the closed form of integrin and stabilizes α/β tail-tail contacts through its Ig repeat 2, preventing leg separation and subsequent integrin activation(138). Dok1 has also been structurally characterized as an integrin inhibitor, binding preferentially to the phosphorylated form of the NPxY β tail motif through its PTB domain and implicating the tyrosine phosphorylation state as a central switch between integrin conformations(139, 140). Furthermore, variation in affinity between integrin tails of different subunits may contextually tip the balance in favor of activation or suppression(134).

Central to understanding integrin activation is the relationship between ectodomain conformation, transmembrane region, intracellular unstructured tails(141). Structural insights into full-length integrins are just beginning to appear. Recent cryoEM structures of the inactive $\alpha IIb\beta 3$ show density consistent with closely-associated transmembrane helices supported by possible specific cholesterol cooperation(46, 47). These studies further characterized a metastable intermediate conformation typified by separation of the lower leg region and transmembrane region in the BC conformation. This insight rationalizes the crystallographic data for cytoplasmic tail interactors, which collectively has concluded that stabilizing the separation of the α and β cytoplasmic tails is essential for maturing active integrin complexes(45). Further, the addition of force pulling on the β subunit tail via the actin cytoskeleton in focal adhesions is structurally consistent with a diagonally oriented transmembrane region, contributing to spatial accessibility for activating cytoplasmic effectors(142, 143). The inherent entropy of the membrane lipid composition itself may further contribute to integrin-mediated signaling, intimately incorporating the membrane as a central player in integrin signaling(115). Collectively these structural insights reveal a surprisingly versatile cytoplasmic region marked by an adaptability for context-specific integrin regulation via a diverse array of physical and chemical modulatory mechanisms.

1.3 Structural features of the integrin αI domain

1.3.1 Overview of I domain activation

While RGD or RGD-like cleft binding appears to be the ancestral form of extracellular integrin-ligand interactions, nine ($\alpha 1$, $\alpha 2$, $\alpha 10$, $\alpha 11$, αE , αL , αM , αD , and αX) of the eighteen alpha subunits encoded in the human genome contain an extra structural domain: the “I” or “inserted” domain(3) sits distal to the cell surface and sterically occupies this ancestral ligand-binding region. Based on sequence homology, it was deduced that this domain shares a structural and evolutionary similarity to the von Willebrand Factor A family of protein folds as well as more distantly the βI domain(144, 145). More recent evolutionary analysis shows this domain has a monophyletic origin at the onset of *Olfactores*, the animal clade encompassing tunicates and vertebrates(146). The ancestral I domain appears to stem from a vWFA-family domain contained within a collagen gene(118), consistent with an intimate link between integrin and ECM components. Subsequent gene duplications and diversification led to the wide array of I domain-containing integrins found across the *Olfactores* phylogeny, suggesting a strong adaptive pressure to not only maintain, but to expand, this class of integrins. In extant vertebrates, there are two general categories of I domain integrins: the collagen-binding $\alpha 1\beta 1$, $\alpha 2\beta 1$, $\alpha 10\beta 1$, and $\alpha 11\beta 1$, and the leukocyte-specific $\alpha E\beta 7$, $\alpha L\beta 2$, $\alpha M\beta 2$, $\alpha D\beta 2$, and $\alpha X\beta 2$ (7).

Early mutational data identified the αI domain as the primary ligand binding site for the integrins which contain it(147–151), eschewing a model of cooperative $\alpha\beta$:ligand binding in favor of a ligand-binding mode completely determined by the α subunit. Owing to its compact, stable nature, early crystallographic studies on the αI domain in isolation were remarkably successful in identifying both the fold and function of this enigmatic domain(152–154). To date, all structures of αI domains have revealed a central buried β sheet sandwiched between two clusters of amphipathic α helices, collectively forming a near spherical topology. Like the many members of the vWFA domain family from which it stems, the top distal end of the I domain contains a single MIDAS which coordinates cations through a DXSXS motif in conjunction with physically close but sequentially distant aspartate and threonine residues. Early crystallographic work also found that, much like with the β MIDAS, completion of the α MIDAS coordination sphere triggers the extension of the structurally labile $\alpha 7$ I domain helix, creating distinct “closed” and “open” conformations(153)(Fig. 1.2A). Given the high conservation of the amino acids responsible for this behavior, it is expected that this is a universal mechanism I domains use to switch conformation.

These crystallographic studies revealed the molecular basis for this conformational switch in detail. Upon external MIDAS ion coordination by an acidic residue contained within a cognate ligand, the ion coordination shifts upward from a conserved aspartate to a conserved threonine, creating space for large structural shifts in multiple secondary features along the I domain. This causes key hydrophobic residues to flip from an upward position to a downward one along the C-terminal region of the I domain, sterically occluding hydrophobic pockets for the closed conformation and leading to a 20Å downward shift in the $\alpha 7$ helix via hydrophobic ratcheting, or laddering, down the neighboring $\alpha 1$ helix (Fig. 1.2B). Subsequent biochemical and cellular work validated each of these conformations in the context of the full-length integrin(155–157).

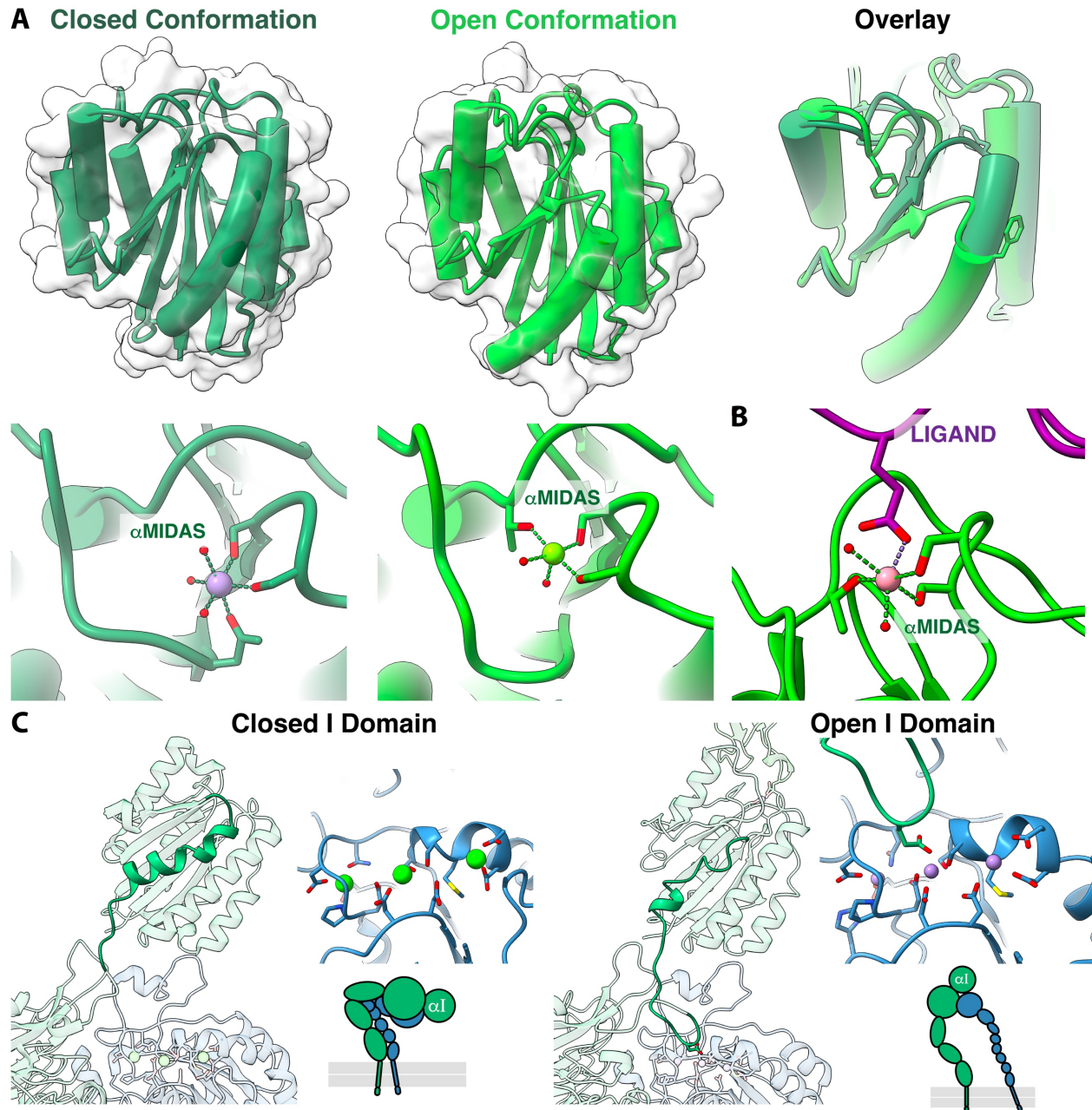


Figure 1.2. Integrin I domain nomenclature and activation. (A) Structures representing the closed (left)(153) and open (middle)(158) conformations of the integrin α I domain from α M. An overlap of the two is shown on the right. The lower panels demonstrate changes in α MIDAS coordination between the two conformations, linking ion coordination to structural change. (B) The α I MIDAS coordination sphere is completed by an extrinsic ligand(159) and, (c) through an allosteric relay via the α I α 7 helix, is communicated to the integrin β subunit through a conserved acidic residue (the “internal ligand”)(118) to induce integrin ectodomain opening and subsequent activation.

Electron microscopy studies, combined with the crystal structure of the entire α X β 2 ectodomain in the BC conformation(160), further found that the α I domain sterically occludes the RGD ligand binding cleft. This suggested that the conformational shifts in the α I domain served as an allosteric relay to connect ligand binding to integrin activation. Importantly, these

studies also found that the “open” I domain conformation was primarily responsible for both ligand binding and subsequent intracellular signaling(161), and identified this conformation as having higher affinity for ligand than that of the closed conformation(5, 162). This work found that, despite the occlusion, I domain-containing integrins occupied the canonical suite of BC, EC and EO ectodomain conformations(16). This data together led to the prevailing molecular model for I domain allostery, in which the opening of the I domain positions a conserved glutamate into the β MIDAS coordination site, leading to the same functional output as an extrinsic ligand for the non-I integrins(163) (Fig. 1.2C). This model was heavily favored but not directly observed until the adoption of cryoEM, which revealed the direct association between ligand binding, the open I domain, and the open integrin ectodomain(118, 164). These studies also found a surprisingly dynamic fluctuation of the I domain between states, suggesting that ligand binding stabilizes, rather than triggers, the open I domain(118). Altogether this work provides a compelling model for how the I domain was able to maintain essential integrin function despite structurally hindering the ancestral activation mechanism and ligand-binding site.

1.3.2 Structural specifications of collagen binding

Owing to their lack of sequence constraint, I domain ligands are exceptionally varied in both structure and function, and remain an active area of investigation in the integrin field (Fig. 1.3A, Table 1.1, Table 1.2). Nevertheless, the collagen-binding integrins $\alpha 1\beta 1$, $\alpha 2\beta 1$, $\alpha 10\beta 1$, and $\alpha 11\beta 1$ seem to have a broadly shared molecular means of interaction with the classical collagen triple helix(85, 165). In general these I domains recognize the shared GFOGER motif(93, 166), where O is a hydroxyproline, through a short helical extension on the top integrin surface unique to the collagen-binding I domains (the “C helix”). The first ligand-bound I domain structure was that of the $\alpha 2$ I domain in complex with a synthetic tripeptide fragment containing GFOGER(159)(Fig. 1.3B). This structure was transformational as it revealed a *bona fide* MIDAS ion coordination by the acidic glutamate ligand residue, establishing the structural basis for I domain recognition and conformational activation. This study also established the open conformation as having high ligand affinity, a rule that has held true for all I domain ligands structurally analyzed to date. Finally, this structural work characterized the spatial intimacy these binding partners share, showing a highly optimized surface for the F, E, and R residues within GFOGER as well as the slot formed for the triple helix by the specialized C helix extension atop the I domain. Complementary mutational studies confirm this specificity(167); even a conservative E to D mutation is detrimental for ligand binding(93).

Despite this conservation, there appears to be plasticity within this GFOGER framework that still allows for tight association(87). Structural characterization of the $\alpha 1$ I domain in complex with a GLOGEN peptide revealed a surprising plasticity in the C helix(168), which was subsequently proposed to constitute a major determinant of idiosyncrasy within the collagen-binding integrins. Similarly, crystallization of the activating mutant E317A in the $\alpha 1$ I domain in the absence of ligand suggested alternate pathways, particularly one involving the movement of Y285 away from the MIDAS region to create more space for unraveling of the C helix unique to $\alpha 1$ (169). More recently, a yeast screening approach has revealed MIDAS-independent binding capabilities of a GFOGMR peptide to the closed $\alpha 2$ I domain, challenging the notion of MIDAS-dependent integrin activation as an essential structural feature(170). Complementary biochemical studies have expanded our understanding of collagen preferences among this integrin family,

revealing the even a single mutation can confer the ability to switch between collagen subtype preferences(171).

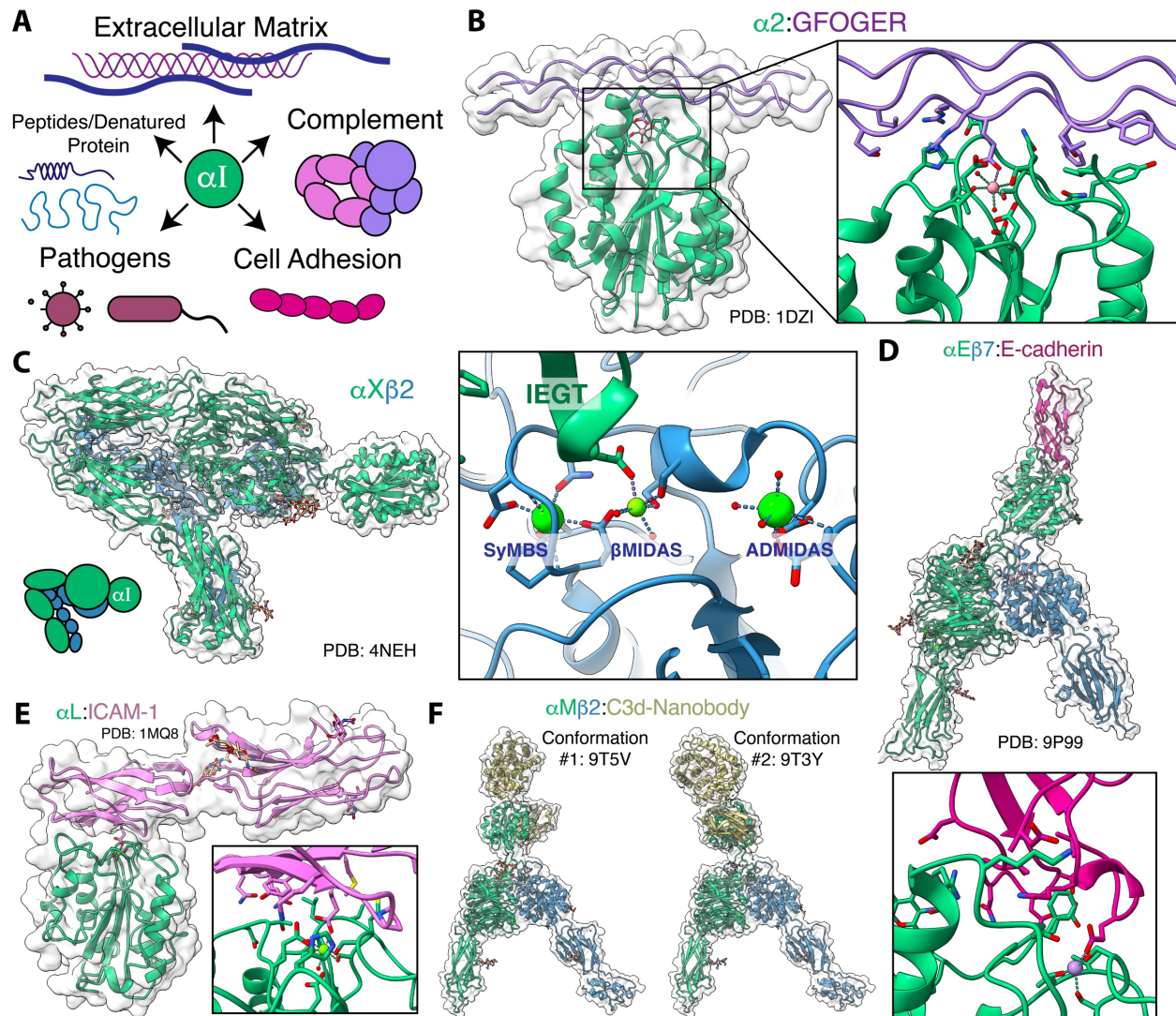


Figure 1.3. Overview of select published I domain structures. (A) The αI domain has an exceptionally vast ligand binding repertoire that doesn't share a common ligand binding sequence or structure motif. (B) The $\alpha 2$ αI domain bound to the collagen GFOGER motif (left)(159) revealed the structural mechanism linking ligand binding to conformational change, with a close-up view of the structural contacts between the integrin and ligands (right) (C) The crystal structure of the BC $\alpha X\beta 2$ ectodomain(172) shows placement of the I domain in relation to the rest of the ectodomain (left) and demonstrates the internal ligand motif linking I domain in contact with the $\beta 2$ subunit (right). (D) The cryoEM structure of the EO $\alpha E\beta 7$ ectodomain bound to E-cadherin(118) demonstrates the entire allosteric relay from ligand binding to integrin activation (top) and the specific contacts mediating this interaction (bottom). (E) The crystal structure of ICAM-1 bound to the αL I domain(173) shows a relatively flat orientation compared to that of αE and E-cadherin. (F) The cryoEM structures of the $\alpha M\beta 2$ headpiece in complex with the complement degradation component C3d(164) reveal a dramatic 150° range of motion in the active αM I domain in relation to the rest of the integrin molecule.

It is intriguing to speculate that the ease of biochemical tuning of collagen-binding preferences has contributed to the great diversity in vertebrate morphology seen in extant species. Still, there is a fundamental knowledge gap in the field between the biochemical

behaviors of collagen receptors and their *in vivo* consequences. This gap is likely due to incomplete biochemical and cell biological understanding of integrin:collagen specificity. Much like with RGD peptide structures, it's almost certain that the reductionist GFOGER minimal motif fails to capture the multivalent interactions that likely occur between integrins and the complex network forming the collagen matrix(174). The continued development of cryoEM techniques will likely contribute to filling this knowledge gap, as this remains the only major integrin family which lacks a high resolution ectodomain structure. Similarly, although whole-organism knockouts classically used to study integrin phenotypes likely have pleiotropic behavior which make pinpointing precise mechanisms difficult(175), recent advances in organoid development may help bridge our understanding of the molecules as they relate to cell biology(176).

Table 1.2. A list of notable published I domain integrin structures.

αI Domain Integrin Structures				
PDB code(s)	Method	Integrin	Ligand	Notes
apo αI domains				
1IDO	X-ray	α M	N/A	Open, Mg ²⁺
1JLM	X-ray	α M	N/A	Closed, Mn ²⁺
1LFA	X-ray	α L	N/A	Closed, Mn ²⁺
1ZOO	X-ray	α L	N/A	Closed, Mg ²⁺
1AOX	X-ray	α 2	N/A	Closed, Mg ²⁺
1QC5	X-ray	α 1	N/A	Closed, Mg ²⁺
1N3Y	X-ray	α X	N/A	Closed
ligand-bound αI domains				
1DZI	X-ray	α 2	collagen	GFOGER peptide helix
1MQ8	X-ray	α L	ICAM-1	ICAM-1 D1 and D2
1T0P	X-ray	α L	ICAM-3	ICAM-3 D1
3BN3	X-ray	α L	ICAM-5	ICAM-5 D1 and D2
4M76	X-ray	α M	C3d	
2M32	X-ray	α 1	collagen	GLOGEN peptide helix
4XW2	X-ray	α M	simvastatin	
6RHV	X-ray	α M	Leukocidin GH	
8VOI	NMR	α M	pleiotrophin	
7AKK	X-ray	α M	iC3b	
3EOA	X-ray	α L	efalizumab	
apo headpieces/ectodomains/full length				
5E6R, 5E6U, 5E6S	X-ray	α L β 2	N/A	Headpiece, 3 different buffers
4NEN, 4NEH	X-ray	α X β 2	N/A	Closed, internal ligand bound
7USM	cryoEM	α M β 2	N/A	Closed ectodomain, I domain too flexible to resolve
8ZJF	cryoEM	α E β 7	N/A	Half-bent, full length, unresolved TM region
9T5W	cryoEM	α M β 2	N/A	Open, R61H mutant
9RM9	cryoEM	α M β 2	N/A	Open headpiece, I domain not resolved
ligand-bound headpieces/ectodomains				
7P2D	X-ray	α M β 2	nanobody	Closed headpiece, Nanobody binds I domain
9TV5, 9T3Y	cryoEM	α M β 2	nanobody-C3d	Open headpiece, I domain rotates 150°
9T5Z	cryoEM	α M β 2	MEM148 fab	Open, I domain not resolved, fab binds β 2
7USL	cryoEM	α M β 2	ACT, M1F5 fab	Closed, I domain not resolved
9P97, 9P98	cryoEM	α E β 7	LF61 fab	Closed and open ectodomains, fab binds α E
9P99	cryoEM	α E β 7	E-cadherin	Open

1.3.3 Leukocyte integrins coming into view

In contrast to the collagen binding integrins, much attention has contributed to our more nuanced understanding of the molecular foundations that define the leukocyte integrins $\alpha E\beta 7$, $\alpha L\beta 2$, $\alpha M\beta 2$, $\alpha D\beta 2$, and $\alpha X\beta 2$. Indeed the first ectodomain structure of an I domain integrin was $\alpha X\beta 2$ in the BC conformation(160)(Fig. 1.3C), which provided a solid structural framework for understanding how this family works. The initial crystal structure as well as its subsequent refinement(172) showed the ion coordinating “internal ligand” motif, a conserved XEGT (where X is any hydrophobic residue) among all I domain α subunits essential for allosteric ligand transduction. The essential glutamate in this domain coordinates the $\beta 2$ MIDAS when the αI domain dons the open conformation, leading to canonical $\beta 2$ MIDAS-triggered hybrid domain swingout and subsequent signaling(118, 164). This open conformation is further stabilized by a conserved hydrophobic pocket supported by each subunit as well as a surprisingly stable, conserved glycan on the top of the beta-propellor(164, 177). This pocket restricts movement of the otherwise flexible internal ligand, creating a structural means of supporting allosteric integrin activation. Given the high conservation of the internal ligand motif, the stable glycan, the C-terminal helical shift and the surrounding hydrophobic pocket of this region, it is very likely this is a generalizable mechanism for all αI domain-containing integrin heterodimers.

Despite the conserved nature of this allosteric mechanism, the direct ligand binding interface appears to be highly adaptable(101, 178). This is particularly represented in the leukocyte integrins which do not share a common ligand or ligand recognition motif, unlike the collagen-binding integrins. The I domains from αE binds uniquely to the first repeat domain from the epithelial signifier protein E-cadherin(97), the αL I domain binds primarily to the first repeat domain of Intercellular Adhesion Molecule 1 (ICAM-1) and its ICAM family homologs(179, 180), and the I domains of αM , αD , and αX have multivalency for a large array of structurally distinct ligands which modulate immune activity pleiotropically(181, 182). Collectively these interactions shape the ability of immune cells to roll, extravasate, migrate, adhere, and retain at sites in need of an immune response(102). This breadth of ligand binding suggests a plasticity in this family of receptors that is particularly well-suited for generating functional immune responses that can keep pace with an ever-evolving extracellular world.

The interaction between $\alpha E\beta 7$ and E-cadherin is critical for effectuating a phenotype termed tissue residence (T_{rm}), especially in gut, skin, and lung tissues, which keeps immune cells stably attached to epithelial tissues to allow for both antigen presentation and continued immune response to foreign bodies and cancer cells(183–187). Structurally this interaction is mediated between the top of the αE I domain and domain 1 of the mature E-cadherin molecule, which coordinates the αE MIDAS via E31(118)(Fig. 1.3D). This interaction is further stabilized by a hydrophobic pocket encasing αE F298 and a salt bridge between E-cadherin R1 and αE K328 which prevents E-cadherin dimerization, allowing immune infiltration of otherwise impenetrable epithelial tissue. This interaction is supported more so by a unique structural “collar” at the intersection of the αE I domain, the αE beta-propellor, and the $\beta 7$ βI domain which restricts αE I domain flexibility and contributes to immune cell stabilization in an otherwise highly dynamic environment. Unlike the other leukocyte integrins, $\alpha E\beta 7$ adopts the half-bent BC state(188) which further poises it to create stable interactions with the epithelium.

Structural studies of the $\alpha L\beta 2$:ICAM-1 interaction have greatly contributed to our understanding of the dynamic, switchable nature of the leukocyte integrins. The crystal structure of the αL I domain:ICAM-1 complex revealed that ICAM-1 domain 1 exclusively binds to the top of the αL I domain, with the αL MIDAS coordinated by ICAM-1 D34(173)(Fig. 1.3E).

Compared to α E:E-cadherin the α L:ICAM-1 surface is relatively flat, although hydrophobic interactions contribute greatly to the specificity of both interfaces. These interactions between the most distal regions of each molecule relative to the cell surface suggest that the significant reach of each domain is critical for function. The α L I domain has served as a great mutational model for understanding allosteric modulation of the I domain α 7 helix, revealing general principles for how hydrophobic and aromatic residues maintain a switchable state that allows rapid attachment and detachment to ICAM-1(189). The dynamic nature of this interaction positions α L β 2 uniquely to mediate a series of complex T cell functions from slow rolling to spreading, endothelial extravasation, and attachment at the immune synapse(190). This is supported by crystal structures of the α L β 2 headpiece, which revealed an extreme flexibility in the α L I domain compared to α E, further contributing to modulating complex motion(177). The magnitude of these behaviors can be amplified or repressed by chemical cues like cytokines as well as biophysical force between molecules, although the structural details of these behaviors remain unresolved.

Unlike the relatively specific niches α E β 7 and α L β 2 occupy, the other leukocyte integrins α M β 2, α X β 2 and α D β 2 appear to have much more pleiotropic function and remain relatively underexplored at both molecular and organismal levels. Perhaps the best studied of the three is α M β 2, which is noteworthy for being particularly promiscuous in its ligand binding behavior; up to 100 structurally distinct ligands have been implicated in binding α M β 2 with functional consequences(103). The first of these ligands to be structurally characterized was C3d(191), a degradation fragment of the C3 complement component which serves as an opsonin for macrophage-mediated phagocytosis. As with other I domains C3d interacts with the top region of the I domain and coordinates the MIDAS, leading to I domain opening and integrin activation. α M β 2 also seems highly flexible at the linker connecting the I domain, which provides a similar amount of structural plasticity as in the α L β 2:ICAM-1 interaction; an extreme near-180° rotation is seen in recent cryoEM data of the α M β 2:C3d complex(164)(Fig. 1.3F). Structural studies have identified other ligands, such as pleiotrophin(192) and the small molecule simvastatin(193), occupying this same region of the α M I domain. The leucine-rich repeat helical cap in the platelet glycoprotein Ib α binds the α M I domain(194) to regulate thrombosis(195), coordinating the MIDAS yet with a slightly different geometry than that of C3d. ICAM-1 also serves as a major ligand for α M β 2, and mutational data suggest the third repeat domain of ICAM-1 is the primary α M β 2 binding region(196, 197) but the structural basis of this complex remains to be determined. Collectively these confirm the MIDAS-containing top region as the primary ligand binding site even across promiscuous receptors, and suggests the I domain surface is constructed such that there are few structural constraints that would inherently impact ligand binding.

Comparatively little is known about the way the α D and α X I domains interact with their respective ligands. α X, like α M, serves as a complement receptor, and low resolution nsEM studies suggest it also interacts with complement at the top of the I domain, yet evolutionary divergence at this region may lead to different patterns of recognition that may tune macrophage behavior. The C3d precursor iC3b may contain two interaction sites outside of the C3d region that interact with the α X I domain, providing hints as to how the two receptors divide labor(198). The α X I domain is also capable of interacting with overlapping ECM components to α M such as fibrinogen and plasminogen, and has emerging functional roles in a variety of autoimmune diseases(104). Similarly, α D has an overlapping ligand repertoire with ICAMs, VCAM-1, ECM

components(199), yet even less is known about the structural mechanisms and functional outcomes of these interactions as there are no structures of $\alpha\text{D}\beta 2$ in any form. The structural determinants of these dynamic, polyvalent receptors thus remain an exciting area for future research.

1.3.4 Pathogenic exploitation of the I domain

Owing to their central role in modulating immunity, integrins are attractive targets for direct pathogenic modulation. This is reflected in the many pathogenic ligands described for these proteins, from viral receptors to pore-forming proteins to PAMPs that may be leveraged for immune recognition and subsequent clearance(200–203). Many such interactions have been described and implicated in pathogenesis, yet few have been characterized at the structural level.

The polyvalent αM I domain has received particularly close attention as a pathogen recognition factor. Multiple pore-forming bacterial toxins have now been shown to bind to αM . The heterodimeric Leukocidin GH (LGH) has been co-crystallized with the αM I domain(204)(Fig. 1.4A). This structure revealed the shared preference for the open I domain conformation with endogenous ligands, suggesting a structural co-option model. LGH attaches at the MIDAS site, supported by an extensive network of electrostatic interactions which comprise a high-affinity surface. The structure further indicated that LGH exploits the inherent integrin conformational landscape to draw itself closer to the cell surface, providing a compelling holistic mechanism for molecular pathogenesis. This is further supported by data suggesting pneumolysin(205), a structurally distinct pore-forming virulence factor, also co-opts the αM I domain. Neutrophil Inhibitory Factor (NIF), a secreted hookworm glycoprotein, shows exceptionally high affinity for the αM I domain and may be the key immune suppresser permitting hookworm pathogenesis(206, 207). Still, recognition of αM I domain by pathogens is a double-edged sword, as αM appears to directly recognize pathogenic patterns, such as LPS(208), zymosan(209) and β -glucan(210), to clear infection, in addition to indirectly enhancing immunity by serving as a complement receptor. These pathogenic interactions have evolutionary impact, as the αM I domain shows signatures of positive (adaptive) selection and has been able to evade binding to LGH in some mammalian systems(211).

$\alpha\text{L}\beta 2$, which is thought to have much higher specificity, also serves as a pathogen recognition factor by binding *Plasmodium* GBP-130 displayed on infected red blood cells(212)(Fig. 1.4B). Structural modeling suggests this interaction is MIDAS-independent, an unexpected deviation from typical pathogenic effectors of the αI domain. This is not the only such example; the I domain from integrin $\alpha 2\beta 1$ has also been heavily implicated as a viral receptor for both rotaviruses(213) and picornaviruses(214). Structural analysis of the $\alpha 2$ I domain complexed with echovirus 1 suggests the side of the $\alpha 2$ I domain, rather than the top, is the primary contact point and is MIDAS-independent; this alternate topology may favor viral entry(215)(Fig. 1.4C). Ross River Virus has been suggested to bind to $\alpha 1\beta 1$ at a site that overlaps with the collagen IV binding region(216), although the structural specifications of this interaction are yet to be determined.

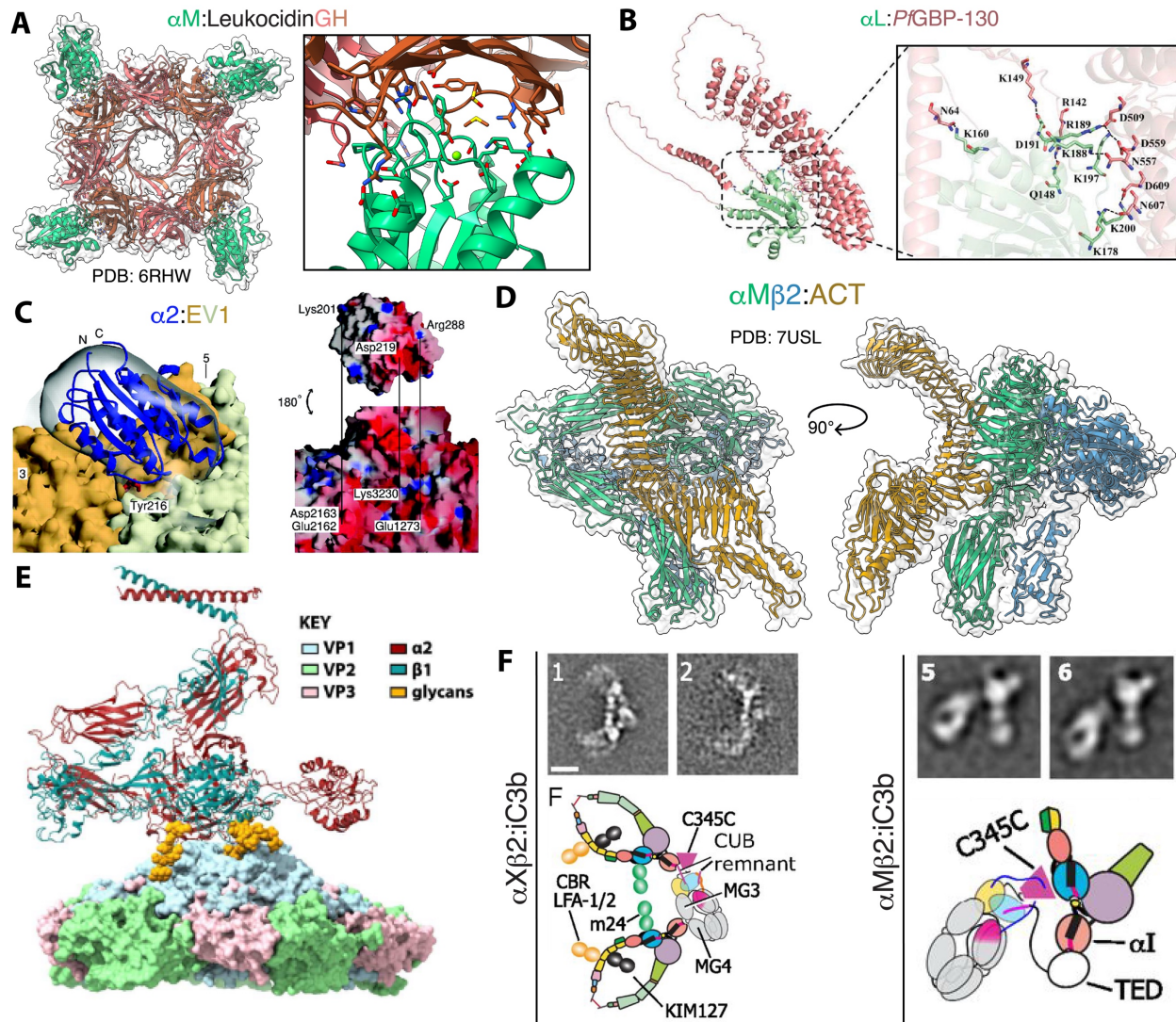


Figure 1.4. I domain integrins beyond canonical interactions. (A) The α M α I domain bound to the pore formed by Leukocidin GH (left)(204) suggests canonical integrin activation pathways are co-opted by pathogenic effectors using an extensive electrostatic surface (right). (B) The predicted α L α I domain bound to *Plasmodium* GBP-130, adapted from Figure 2 of (212), suggests pathogenic binding may universally rely on charged interactions yet does not depend on the MIDAS. (C) A low resolution electron microscopy density of the α 2 α I domain in complex with EV1 virus further supports pathogenic use of the side of the I domain to facilitate infection, adapted from (215). (D) The cryoEM structure of the BC α Mβ2 ectodomain bound to *Bordetella* ACT(217) shows pathogens can take advantage of the predominant BC integrin architecture to facilitate virulence activity, such as pore formation. (E) The cryoEM structure of the α 2β1 in complex with ERAD VLPs demonstrates a reliance on glycosylation to facilitate viral binding, adapted from (218). (F) Low resolution nsEM of the complement component iC3b in complex with either α Xβ2 (left) or α Mβ2 (right) suggests these two closely related receptors may use different structural mechanisms to identify opsonized particles, adapted from (198).

1.3.5 Ligand recognition beyond the I domain

Perhaps the most exciting opportunity provided by advances in cryoEM is the ability to understand integrin structures in the context of the entire molecule, rather than individual domains in isolation. This can have broad impacts, as many integrins have been implicated in binding ligands outside of canonical regions (178). This is particularly true for I-domain

integrins, for which the limitations of crystallography have prevented a nuanced understanding of polyvalency across these complex receptors.

Several recent integrin ectodomain structures have advanced our understanding of ligand binding beyond the I domain. The $\alpha M\beta 2$ BC ectodomain complexed with the *Bordetella* pore-forming virulence factor Adenylate Cyclase Toxin(217) provided the first high-resolution structural glimpse into how effectors may leverage integrin's conformational ensemble for functional output (Fig. 1.4D). This toxin binds primarily the beta-propellor and calves of the bent αM subunit, allowing it to come in proximity with the host leukocyte membrane. This approach is doubly effective for the bacterium as it both prevents integrin-mediated immune activation and provides a direct route for pore-mediated lytic rupture. A similar mode of binding seems to be favored by viruses: a recent cryoEM structure resolved the BC form of $\alpha 2\beta 1$ bound to the picornavirus ERAD(218), finding that recognition is mediated largely by glycans along the $\alpha 2$ subunit with a relatively limited protein-protein interface (Fig. 1.4E). Pathogen exploitation of sugar modifications is particularly surprising yet supports the growing body of work implicating glycans as central, and perhaps mutationally intolerant, players in endogenous integrin function. Pathogenic use of regions distal to the I domain is shared among other bacterial and viral toxins(219–221). However, the structural basis of these interactions and their conservation, or lack thereof, remains to be more fully explored.

Perhaps more tantalizing is the notion that integrins use their exceptionally large ectodomains to perform endogenous functions extending beyond canonical ligand binding sites. As complement receptors recognizing the opsonin iC3b, $\alpha M\beta 2$ and $\alpha X\beta 2$ appear to use surfaces beyond their αI domains for this purpose and in structurally distinct ways(198) (Fig. 1.4F). Although extensive work has identified regions beyond the C3d fragment of iC3b as important for opsonization via $\alpha M\beta 2$, the structural details of the phagocytic cup remain contentious(222–224). Perhaps based on its inherent promiscuity, $\alpha M\beta 2$ appears to be a hotspot for extending ligand binding beyond the borders of the I domain: upwards of 100 characterized ligands for this receptor share overlapping, yet distinct, regions both inside and out of the I domain as well as the “lectin” region which excludes the I domain(103, 209, 210, 225, 226). This division of labor may account for the $\alpha M\beta 2$'s ability to bind so many distinct ligands and may also fuel its context-dependent regulatory effects on immune suppression and activation. This division of binding can also be supplemented by cis-interactions on the same cell surface, both at the αI domain and beyond (227–230). Structural studies exploring the entire integrin molecule are essential to clarify how one receptor balances this overwhelming complexity of chemical cues, and how this can be modulated for therapeutic benefit.

1.4 Protein domains as evolutionary fodder

In the simplest sense, protein domains are the structural and functional building blocks of life. The vast proteome across all living things can be traced back to a handful of core protein folds, thought to be among the most ancient units of molecular function in the biological world. Although there are many ways to classify and cluster protein domains, deep structural homology suggests that as few as 9 protein folds account for the vast majority of natural protein structure variation(231), although a classification of estimates in the 1000s to 10000s appears to be more informative for relating structure to function(232), and perhaps a spectrum is a better representation than discrete families(233). Breakthrough advances in protein structure prediction have helped illuminate a more holistic perspective on the possible protein folds found in nature

and highlighted the limitations in our understanding of protein folding to date(234–236). The virome may be particularly fruitful in expanding our perception of protein fold and function(237).

Because the living proteome contains an intractable number of proteins, our understanding of protein domain function largely relies on homology inference (238). An increasing number of compendia that classify protein domain families based on sequence similarity(239–242) has advanced our understanding of molecular conservation and divergence across evolutionary timescales. While early methods relied primarily on sequence, more recent structure-based searches have exponentially improved our ability to resolve homology across long timescales (243–246). This is particularly informative as sequence tends to diverge an order of magnitude faster than structure(247), so for very distant homologs or rapidly evolving proteins structural comparisons may prove more informative than relying on sequence alone. A valuable tool for understanding evolutionary sequence-structure relationships in protein domains is probing their mutational limits through techniques such as deep mutational scanning (248, 249), which create comprehensive genotype-phenotype landscapes that reveal the limits of protein structure and function. Combining datasets from protein predictions, mutational data, and functional experiments effectively reveals a protein’s potential, a fundamental part of understanding how protein domains can and do evolve, as this thesis will do for selected I-domains in Chapter 3.

The complexity of protein domains scales by orders of magnitude when considering the possibilities of multi-domain proteins (250, 251). This is particularly true for the eukaryotic proteome, which is primarily composed of multi-domain proteins (252, 253). Still, the vast majority of possible protein domain combinations do not exist in the natural world, and protein domain combinations show surprising predictability (254, 255) and indeed seem to arise independently under strong selection (256). Thus there is not only a functional relationship between structure and individual domains, but also between larger domain combinations. Even so, many independent protein domain combinations are unique and relatively rare, suggesting tuned function for biological niches. Despite their molecular size, these patterns of domain combinations have broad-reaching implications in ecological evolutionary processes such as speciation(257–260). Much like in speciation, protein domains have experienced punctuated periods of rapid expansion and diversification(252), suggesting the molecular evolutionary trends mirror the ecological.

How do protein domains acquire novel function? Beyond the aforementioned combinatorial pathways (250), within-domain adaptations can generate new behaviors with surprisingly little change. Some of these adaptations are easily observable at well-established catalytic sites(261) or ligand pockets(262), while other hotspots of evolution may have allosteric impacts rather than direct ones(263) or may contribute to higher-order assembly(264). Recent studies have supported the evolutionary ease of generating entirely new function (265–267), suggesting protein domains have expansive evolutionary potential. Conversely, some protein domains retain an exceptionally conserved function despite extreme sequence divergence(268); this is especially true for proteins involved in immune(269, 270) and mating(271) conflicts.

The scm3 domain is an enigmatic example of conserved domain function amid rapid genetic evolution, which this thesis explores in greater depth in Chapter 4. This domain is the centromeric histone-binding region of the eponymous histone chaperone Scm3(272, 273), which shepherds the centromeric histone variant CENPA to centromeric DNA. Despite having shared function, this protein was identified thrice independently via biochemical methods in yeast(274),

fruit flies(275), and humans, respectively(276, 277). Subsequent work suggested the presence of a shared scm3 domain orthologous between the three species' proteins responsible for directly binding CENPA, yet the sequence identity is much lower than would be expected for a strictly essential protein domain, and it was thought to be missing in most animals(278). This strongly suggests a range of plasticity even within conserved function and directly contrasts with protein domains that make short mutational jumps to functional novelty. Thus, the integrin I domain and the scm3 domain exist at opposite ends of this evolutionary spectrum, and as such will be explored further in this thesis.

1.5 Thesis overview

The remainder of this thesis is presented in three chapters, each of which is associated with a manuscript that is either published or under review, and 3 appendices. Collectively, these chapters detail my studies of two different model systems – the integrin α I domain and the scm3 domain –to understand how molecular divergence either creates novelty or maintains function in the face of strong adaptive pressures.

Chapters 2 and 3 of my thesis focus on the integrin α I domain. As discussed throughout this introduction, this protein domain is an excellent representative example of a structural unit that has moved within the genome, adapted to changing environments, and integrated intimately into an established structural network, all while remaining faithful to its ancestral, core functional properties. In chapter 2, I used single-particle cryoEM to structurally dissect how this domain allosterically links ligand binding to integrin activation via coordinated conformational movement. I compared two integrins, $\alpha 4\beta 7$ and $\alpha E\beta 7$, which either lack or contain an I domain, respectively, to identify adaptive and conserved features that enabled this domain to succeed. I used phylogenetic analyses to trace this domain's evolutionary origins, identifying critical ancestral residues that serendipitously allowed domain co-option. In chapter 3, I used mutational scanning, phylogenetics, and cryoEM, to understand how ligand binding profiles compare between promiscuous and specific integrins, and how avidity permits promiscuity in the highly adaptable integrin $\alpha M\beta 2$.

In chapter 4, I leveraged the rapidly evolving scm3 domain, a histone chaperone, to determine the plasticity of sequence-structure relationships that maintain critical biological function. I used remote homology detection methods to find an evolutionary ancestor for all Scm3 proteins in *Metazoa* and identify it in many species where it was thought to be missing. This chapter provides perspective on using shared domain properties to explore complex protein relationships across scales. I conclude with thoughts on how to contextualize our understanding of these building blocks within the broader scope of a complex living system.

I also include three epithets as appendices. Although these are not fleshed-out narratives with extensive experimental support, they represent concepts that excited me throughout my thesis work and often inspired me when my main lab experiments were going less than smoothly. I hope you enjoy reading about them as much as I enjoyed working on them.

1.6 List of publications

(*) denotes equal authorship

1. **Jeremy A. Hollis**, Melody G. Campbell. Structural determinants of integrin I domain function. Donald Gullberg (ed.), *I domain Integrins*. *In preparation*.

2. **Jeremy A. Hollis**, Zeqi Wang, Rachel Werther, Anthony S. Kemp, Harmit S. Malik, Melody G. Campbell. Avidity unlocks promiscuity in a divergent integrin receptor family. *Under review*.
3. Victoria E Deneke*, Johannes P Suwita*, Haoting Wang, Shingo Tonai, Yonggang Lu, Karin Panser, Alexander Schleiffer, **Jeremy A Hollis**, Maria Novatchkova, Gerhard Duernberger, Karel Stejskal, Gabriela Krssakova, Andreas Blaha, Aleq A R Andresan, Muriel Mirus, Hana Marvanova, Hsin-Yi Chang, Taichi Noda, Alejandro Burga, Elisabeth Roitinger, Masahito Ikawa, Andrea Pauli. The SPARK complex forms the molecular basis of vertebrate fertilization. 2026. bioRxiv 2026.05.13.724128; doi: <https://doi.org/10.64898/2026.05.13.724128>
4. **Jeremy A. Hollis**, Jason A. Stonick*, Irini Topalidou*, Janet M. Young, Cecilia B. Moens, Nicolas J. Lehrbach, Melody G. Campbell, Harmit S. Malik. Remote homology and functional genetics unmask deeply preserved Scm3/HJURP orthologs in metazoans. *Sci Adv*. 2026 July 1. doi: 10.1126/sciadv.aeg5356.
5. **Jeremy A. Hollis**, Matthew C. Chan, Harmit S. Malik, Melody G. Campbell. Molecular exaptation by the integrin α I domain. *Sci Adv*. 2025 Sept 10. doi:10.1126/sciadv.adx9567.

Chapter 2 Molecular exaptation by the integrin α I domain

**This chapter was adapted from Hollis et al. 2025(118).*

2.1 Abstract

Integrins bind ligands between their alpha (α) and beta (β) subunits and transmit signals through conformational changes. Early in chordate evolution, some α subunits acquired an "inserted" (I) domain that expanded integrin's ligand-binding repertoire but obstructed the ancestral ligand pocket, seemingly blocking conventional integrin activation. Here, we compare cryoEM structures of apo and ligand-bound states of the I domain-containing α E β 7 integrin and the I domain-lacking α 4 β 7 integrin to illuminate how the I domain intrinsically mimics an extrinsic ligand to preserve integrin function. We trace the I domain's evolutionary origin to an ancestral collagen-collagen interaction domain, identifying an ancient molecular exaptation that facilitated integrin activation immediately upon I domain insertion. Our analyses reveal the evolutionary and biochemical basis of expanded cellular communication in vertebrates.

2.2 Introduction

Protein evolution often proceeds via small incremental mutational steps. In contrast, protein domain acquisition can spur dramatic novelty, sometimes at the cost of ancestral functions(250). One such dramatic domain acquisition occurred early in chordate evolution in the integrin family of cell-surface receptors, which mediate signaling across the cell membrane and are critical for core processes from embryogenesis to T-cell activation. Integrins are heterodimeric proteins composed of a single alpha (α) and beta (β) subunit (Fig. 1A). Integrins canonically exist in a conformational spectrum from a closed, compact state to an open, extended state, with the closed state considered inactive and the open state considered the major signaling state (Fig. 2.1A). The human genome encodes eight β and eighteen α integrin subunits, nine of which include a 200 amino acid "inserted" von Willebrand factor type-A (vWFA) domain (or I domain)(3, 4, 151, 279, 280). Integrin α -subunits encode seven blades of a beta-propeller fold, with the I domain found between the second and third propeller blades(146) (Fig. 2.1B).

Integrins lacking I domains bind ligands at an interface between α and β subunits. Ligand binding is also directly linked to the conformational changes required to signal integrin activation by a canonical Arg-Gly-Asp (RGD) tripeptide present in many extrinsic ligands(281). However, the insertion of the I domain sterically occluded this ligand-binding interface, potentially disrupting both ligand binding and the allostery required for integrin activation. Prior studies showed that the I domain acquired ligand binding functions, enabling I domain-containing integrins to bind a wider array of ligand moieties and motifs, including collagens, complement components, and other extracellular matrix proteins(178). Moreover, structural characterization of isolated I domains revealed a conformational change in their C-terminal helix that inserts a conserved glutamate into the ancestral ligand-binding pocket of integrin. This shifted glutamate may structurally and functionally substitute for the RGD motif, coordinating a cation essential for integrin extension(172). However, the conformational states of an I domain-containing integrin have never been observed with the required structural detail to resolve ligand binding and integrin activation features. As a result, the mechanism by which domain-triggered integrin activation occurs is still not well understood. Given that half of human α -integrins

contain I domains – including most with critical immune functions – resolving the molecular details of this evolutionary insertion would provide vital insight into a wide array of vertebrate signaling processes and reveal fundamental principles of how protein domains are successfully co-opted.

Here we employ cryoEM, computational heterogeneity, biochemical, and evolutionary analyses to dissect the evolutionary and mechanistic basis of this pivotal protein domain insertion. We directly compare the activation networks of two integrin heterodimers involving $\beta 7$, a β integrin subunit capable of interacting with both types of α -integrin subunits (Fig. 2.1C). The $\beta 7$ integrin family, encompassing $\alpha 4\beta 7$ (LPAM-1) and $\alpha E\beta 7$ (CD103), performs critical roles in facilitating immune cell homing, activation, and retention at sensitive immune barriers(95, 282). Thus, in addition to revealing unprecedented insights into the fundamental biochemical and evolutionary origins of allostery underlying integrin activation, our structures of the $\alpha 4\beta 7$ and $\alpha E\beta 7$ integrins have substantial clinical value(283–285).

2.3 Results

2.3.1 Integrin $\alpha 4\beta 7$ adopts distinct conformations

We first determined a structure of the non-I $\alpha 4\beta 7$ integrin ectodomain bound to its cognate ligand MAdCAM-1 (Fig. 2.1D-E, Figs. S2.1 and S2.13, Table S2.1). This interaction is critical for leukocyte migration to the gut(96), yet no previous structure of the complex has been reported. We find MAdCAM-1 fits snugly within the groove between $\alpha 4$ and $\beta 7$, forming interactions with both integrin subunits (Fig. S2.2A). Contacts primarily occur on the first Ig-like MAdCAM-1 domain, between the MAdCAM-1 FG strands and $\alpha 4$ propeller, and between the MAdCAM-1 CD loop and the $\beta 7$ β I domain (Fig. S2.2A-B). MAdCAM-1 centrally bridges the $\alpha 4$ and $\beta 7$ subunits with an RGLD motif; R39 and L41 span the long axis of the integrin groove, filling the space between $\alpha 4$ and $\beta 7$ (Fig. 2.1F). The D42 residue provides the essential ion-coordination activity within this motif, leading to the open integrin conformation characteristic of an active heterodimer (Fig. 2.1G). These central contacts are further stabilized by the MAdCAM-1 DE beta ribbon loop in the Ig-like domain 2 (D2) contacting the $\beta 7$ specificity-determining loop (SDL) via mutationally intolerant(286) hydrogen bonds (Fig. 2.1H), supported by an additional interface between MAdCAM-1 D1, D2, and the $\beta 7$ SDL (Fig. 2.1I).

Integrins have been proposed to unfold like a switchblade as they transition from their compact to open states(18, 287). Accordingly, we observe a high degree of $\beta 7$ leg flexibility in the $\alpha 4\beta 7$:MAdCAM-1 structure (Movie S1 in online publication) within the open headpiece conformation. However, by resolving the previously(288) undescribed lower leg region of the $\alpha 4$ subunit encompassing the calf-1 and calf-2 domains, we found that the open $\alpha 4\beta 7$ conformation has a slight rotation around a vertical axis (Fig. 2.1D). To accommodate the different orientations of the headpiece relative to the legs between the open and closed $\alpha 4\beta 7$, we propose $\alpha 4\beta 7$ must “untwist” open rather than “unfold” like in the classic switchblade model. We determined an apo $\alpha 4\beta 7$ clasped structure under high calcium buffer conditions previously shown to favor the inactive conformation state(27) (Fig. 2.1J-K, Figs. S2.3A-E and S2.13, Table S2.1). In this apo structure, $\alpha 4\beta 7$ defines a noncanonical bent conformation. The headpiece is skewed, and the $\beta 7$ headpiece region sits below the $\alpha 4$ propeller, mediated by contacts occurring between $\alpha 4$ calf 1 and the PSI/EGF-1 region of $\beta 7$ (Fig. S2.3F). We propose that the simplest means for $\alpha 4\beta 7$ to transition from bent to extended state would be by twisting about the lower leg region. Indeed,

3D flexibility analysis of the $\alpha 4\beta 7$:MAdCAM-1 complex supports this hypothesis (Movie S1 in online publication), revealing a continuous ensemble of conformational twisting. Given our structural data, this “twisted” compact orientation might also explain how $\alpha 4\beta 7$ binds MAdCAM-1 perpendicular to the orientation at which $\alpha 5\beta 1$ binds fibronectin (14, 289) (Fig. S2.3F), despite both integrins having a similar angle of ligand approach or occlusion.

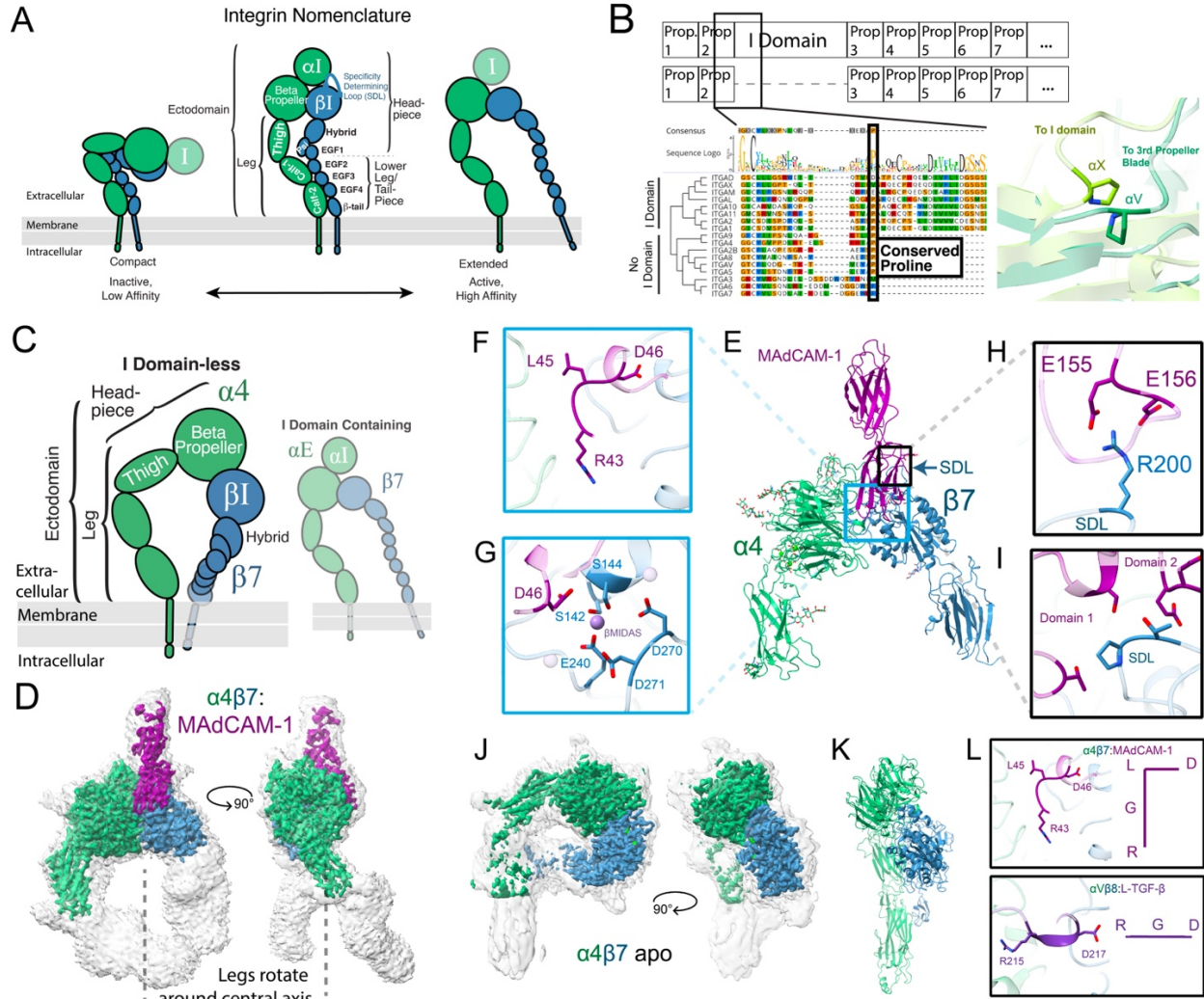


Figure 2.1. Integrin $\alpha 4\beta 7$ adopts novel conformations. (A) A detailed schematic of integrin organization and conformational changes. Integrin conformations range from a compact, inactive state with low ligand affinity (left) to an extended-open, active state with high ligand affinity (right). Approximate hypothesized I domain locations are shown as semi-transparent. (B) An alignment of all human integrin α subunits (except αE) shows insertion of the I domain in the ancestral integrin gene occurred immediately following a conserved proline at the end of the second β -propeller blade. Slightly offset structural overlays of the I domain-less integrin αV (dark green, PDB 1L5G(111)) and I domain-containing αX (lime green, PDB 4NEH(172)) with their respective conserved prolines displayed. (C) A schematic highlighting showing the focus of this figure on the I domain-lacking integrin $\alpha 4\beta 7$. Key domains are labeled as appropriate. (D) CryoEM structure of the integrin $\alpha 4\beta 7$ ectodomain (green/blue) bound to the MAdCAM-1 ectodomain (purple). The low threshold, unsharpened global refinement map is shown in gray, and the inset high threshold, sharpened local refinement map is shown in color. The legs of $\alpha 4$ and $\beta 7$ are not planar; rather they appear to rotate around a vertical axis. (E) Atomic model of the $\alpha 4\beta 7$:MAdCAM-1 complex. Blue box indicates regions of focus in (D-E), black box indicates region of focus in (F). (F) MAdCAM-1 uses an RGLD motif instead of the canonical tripeptide RGD to bridge the $\alpha 4$ and $\beta 7$ subunits. The RGLD residues span the long axis of the cleft

formed between $\alpha 4$ and $\beta 7$. **(G)** MAdCAM-1 D46 completes the $\beta 7$ β MIDAS ion coordination sphere, here modeled as Mn^{2+} . **(H)** R200 in the $\beta 7$ specificity-determining loop (SDL) forms essential salt bridges with a flexible loop in MAdCAM-1. **(I)** Stabilizing contacts intersect between the first and second domains of MAdCAM-1 with the $\beta 7$ SDL. **(J)** The cryoEM structure and **(K)** rigid-fit molecular model of the apo compact $\alpha 4\beta 7$. **(L)** The RGLD motif in MAdCAM-1 adopts a right angle, contrast to the canonical linear RGD motif of $\alpha V\beta 8$ -bound L-TGF- β (24).

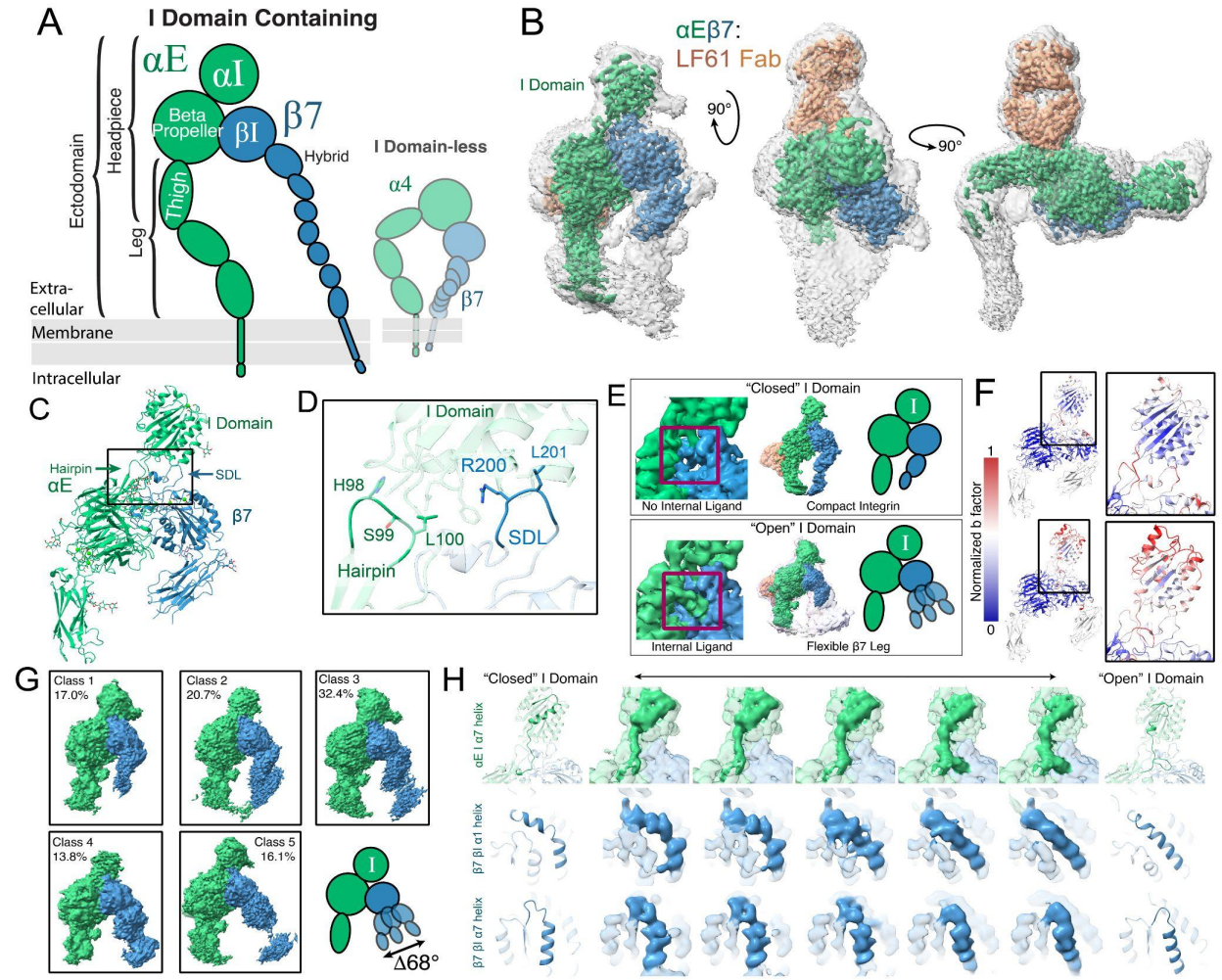
In addition to distinct conformations, we find several other idiosyncrasies that confer specificity to the $\alpha 4\beta 7$:MAdCAM-1 complex. Our open structure reveals that MAdCAM-1 undergoes a conformational shift upon binding integrin. In contrast to existing crystal structures of MAdCAM-1 alone(290, 291), we find that the MAdCAM-1 D beta strand runs antiparallel to the E strand in Ig-like-domain-1 when bound to $\alpha 4\beta 7$ (Fig. S2.2C). Further, in contrast to the well-characterized linear RGD motif within αV integrin ligands(106), we see the RGLD present in MAdCAM-1 is at a right angle to fit within the exceptionally deep perpendicular groove unique to $\alpha 4\beta 7$ (Fig. 2.1L), emphasizing the plasticity of this canonical motif. Prior mutational studies(32, 286, 292) of integrin and MAdCAM-1 have independently shown the importance of $\beta 7$ ion coordination sites and each MAdCAM-1 domain, including a proposed integrin-binding loop in MAdCAM-1 D1 and acidic patch in D2. Our structural data allow us to unite these studies into a cohesive stepwise model of specific ligand binding(32). We propose integrin initially surveys immune sites via the SDL/D2 interaction distal to the cell surface when integrin is bent, then favors cells slowing to roll via tighter dynamic interactions influencing the $\beta 7$ synergistic metal ion-binding site (SyMBS), eventually adhering firmly to the endothelial cell layer via MAdCAM-1 slotting specifically into the $\alpha 4\beta 7$ groove triggering integrin untwisting (Fig. S2.2D). We propose that the “untwisting” of $\alpha 4\beta 7$ we observe in our structural data represents the connection between the intermediate transition state observed in prior low-resolution structural studies of $\alpha 4\beta 7$ (293) and the open state from our study, collectively revealing the pathway towards integrin activation. While our model plausibly unifies our current understanding of $\alpha 4\beta 7$ -mediated adhesion, future studies of high-resolution intermediate states will clarify the timing and sequence of these molecular events. Together, our $\alpha 4\beta 7$ structures represent two unique integrin conformations and reveal specificity mechanisms governing the $\alpha 4\beta 7$:MAdCAM-1 interaction.

2.3.2 *The inactive $\alpha E\beta 7$ I domain is unexpectedly dynamic*

Next, we focused on the structural analysis of the I domain-containing $\alpha E\beta 7$ integrin. Negative stain electron microscopy (nsEM) of the purified $\alpha E\beta 7$ ectodomain in several cation conditions (Fig. S4) revealed both closed-state and open-state conformations. To gain more detailed insight into the compact inactive state of the I domain (Fig. 2.2A), we performed cryoEM on the $\alpha E\beta 7$ ectodomain under non-activating buffer conditions. To facilitate structural analysis, we bound this integrin to LF61, an αE antibody fragment that does not influence ligand binding(294). Surprisingly, 3D analysis revealed two distinct populations of particles. The first, which we had expected to see, was the compact conformation, which we determined to a global resolution of 2.9Å (Fig. 2.2B, Figs. S2.5, S2.6, and S2.13). The intrinsic dynamic variability of the leg results in the overall resolution of the map being quantitatively anisotropic (Figs. S2.5 and S2.6). Nevertheless, most of the headpiece region was rigid, enabling us to confidently model the $\alpha E\beta 7$ headpiece region (Fig. 2.2C, Table S2.1) just as we did for $\alpha 4\beta 7$. The I domain is positioned above the heterodimer interface, which would completely sterically occlude the ancestral RGD pocket of I domain-lacking integrins. We found that the αE I domain makes

additional contacts with the $\beta 7$ SDL and a beta-hairpin on the αE beta-propeller. These supports form a “collar” around the αE I domain (Fig. 2.2D), contributing to the stability necessary for this integrin to perform its biological function in maintaining persistent epithelial contacts (183, 188, 295–297). αE I domain contacts are exceptionally extensive; the αE I domain has a $940\text{\AA}^2$ contact area with the rest of the integrin heterodimer, whereas the αM I domain from PDB:7P2D (298) has only a $493.4\text{\AA}^2$ contact area with the rest of the I domain-containing $\alpha M\beta 2$ integrin.

Figure 2.2. The half-bent $\alpha E\beta 7$ inactive structure is dynamic. (A) A schematic highlighting showing the focus of



this figure on the I domain-containing integrin $\alpha E\beta 7$. **(B)** Three different views of the unsharpened $\alpha E\beta 7$:LF61 Fab complex map. αE (green) $\beta 7$ (blue) and LF61 (orange). The integrin adopts a half-bent conformation in the inactive state, and the I domain rests atop the α/β subunit interface. The colored, sharpened, high-threshold map is shown inset in the unsharpened, grayscale, low-threshold map. **(C)** The refined coordinate model of the compact $\alpha E\beta 7$ headpiece. We model five N-linked glycosylations on αE in addition to two previously reported for $\beta 7$. We do not model LF61 due to a lack of sequence availability. All ions are modeled as Ca^{2+} . Black box indicates regions of focus in (D-E). **(D)** The αE I domain is stabilized by supports from a beta-hairpin on the αE propeller domain as well as the $\beta 7$ specificity-determining loop (SDL). The αE I domain has a $940\text{\AA}^2$ contact area with the rest of the integrin heterodimer, whereas the αM I domain from PDB:7P2D⁶³ has only a $493.4\text{\AA}^2$ contact area with the rest of $\alpha M\beta 2$. **(E)** $\alpha E\beta 7$ assumed two broad states. In the closed I domain state (left) the C-terminal region of the $\alpha 7$ helix, termed the internal ligand, does not engage with $\beta 7$ subunit and the $\beta 7$ leg is closed inward. In the open I domain state (right) the internal ligand engages $\beta 7$, leading to conformational heterogeneity in the $\beta 7$ leg. For both structures

the unsharpened map from a non-uniform refinement is shown in color; for the open I domain representative structures from 3D classification are shown overlaid with low opacity and a schematic is shown to the right. Red boxes indicate internal ligand region. (F) B factor analysis, a quantitative representation of protein motion via thermal energy, shows the I domain becomes broadly more dynamic in the open conformation. (G) 3D classification of 5 open I domain classes suggests a continuous range of $\beta 7$ hybrid domain states upon internal ligand engagement. (H) Representative intermediates from 3D variability analyses show the engagement of the I domain internal ligand is directly linked to opening of the $\beta 7$ β I domain, leading to subsequent hybrid domain swingout.

The I domain contains an internal ligand motif(163) proposed to coordinate the cation site (β MIDAS) in the β subunit via a conserved acidic glutamate. The second population of particles not only showed appreciable heterogeneity in the degree of headpiece opening (Fig. 2.2E, Fig. S2.6, Movie S2 in online publication) but also revealed engagement of the internal ligand. This was unexpected given the high Ca^{2+} levels in this buffer, which have been shown to favor an inactive conformation for many other well-studied integrins(27). Consistent with previous analyses of isolated I domain ectodomains(172), we found that extension of the I domain's C-terminal $\alpha 7$ helix is a prerequisite for internal ligand binding and β MIDAS coordination. This helix in our structure of the internally-liganded $\alpha\text{E}\beta 7$ occupied the “open” conformation previously described in crystal structures of other I domains(153). However, in contrast to the metastable state occupied by $\alpha\text{X}\beta 2$ (172), we found that the $\alpha\text{E}\beta 7$ headpiece becomes broadly dynamic upon internal ligand binding (Fig. 2.2E-F). 3D Classification analysis suggested that the $\beta 7$ hybrid domain adopts a continuous range of motion in this internally-liganded state (Fig. 2.2G). Next, using 3D variability analysis(299), we mapped how the dynamic behavior of the I domain helix-7 is directly linked to the deformation of key β I domain helices in a stepwise manner (Fig. 2.2H). These dynamics are unique to integrins containing the I domain; integrins lacking I domains cannot sample these states in the absence of ligand, as evidenced by our $\alpha 4\beta 7$ apo structure.

2.3.3 Ligand binding stabilizes the open $\alpha\text{E}\beta 7$ I domain

To understand how I domain internal ligand movement is connected to external ligand binding, we next determined a cryoEM structure of the $\alpha\text{E}\beta 7$ ectodomain bound to its primary ligand E-cadherin(97) (Fig. 3A, Figs. S2.7 and S2.13). At a global resolution of 3.4Å, the ligand-binding interface was well-resolved; we modeled the integrin headpiece region as well as E-cadherin EC1 (Fig. 2.3B). Our structure resolves the previous uncertainty about which E-cadherin domains are necessary for integrin binding(300). We find E-cadherin EC1 exclusively binds to the distal surface of the αE I domain, using a short E-cadherin construct (EC12)(301) (Fig. 2.3A). E-cadherin occupies physiological monomeric and dimeric states, with trans-dimerization between adjacent epithelial cells serving as the essential building block of epithelial tissue via higher-order structures termed adherens junctions(302). Despite this, we observed that E-cadherin is only present in a monomeric state when bound to $\alpha\text{E}\beta 7$ (Fig. 2.3C). The $\alpha\text{E}\beta 7$:E-cadherin molecular interface is primarily mediated by a central hydrophobic “lock and key” pocket surrounded by a network of electrostatic bonds(303) (Fig. S2.8A-B, Data S2 in online publication). This interface is further stabilized by the N303 glycosylation on the αE I domain, which makes direct contact with E-cadherin (Fig. S2.8C). The $\alpha\text{E}\beta 7$:E-cadherin complex is highly stable, with a K_D of 51 ± 7.9 nM measured using Bio-Layer Interferometry (Fig. 2.3D). This structure pinpoints molecular interactions essential for sustained tissue residence (T_{m}) that are effectuated by $\alpha\text{E}\beta 7$ (183, 295–297). We found that the αE I domain forms a salt bridge with residue 1 of the mature E-cadherin molecule (Fig. 2.3E), effectively outcompeting adherens

junctions by preventing the “strand-swap” step necessary for E-cadherin’s trans-dimerization(304, 305) due to E-cadherin’s higher affinity to α E β 7 than to itself (Fig. 2.3F).

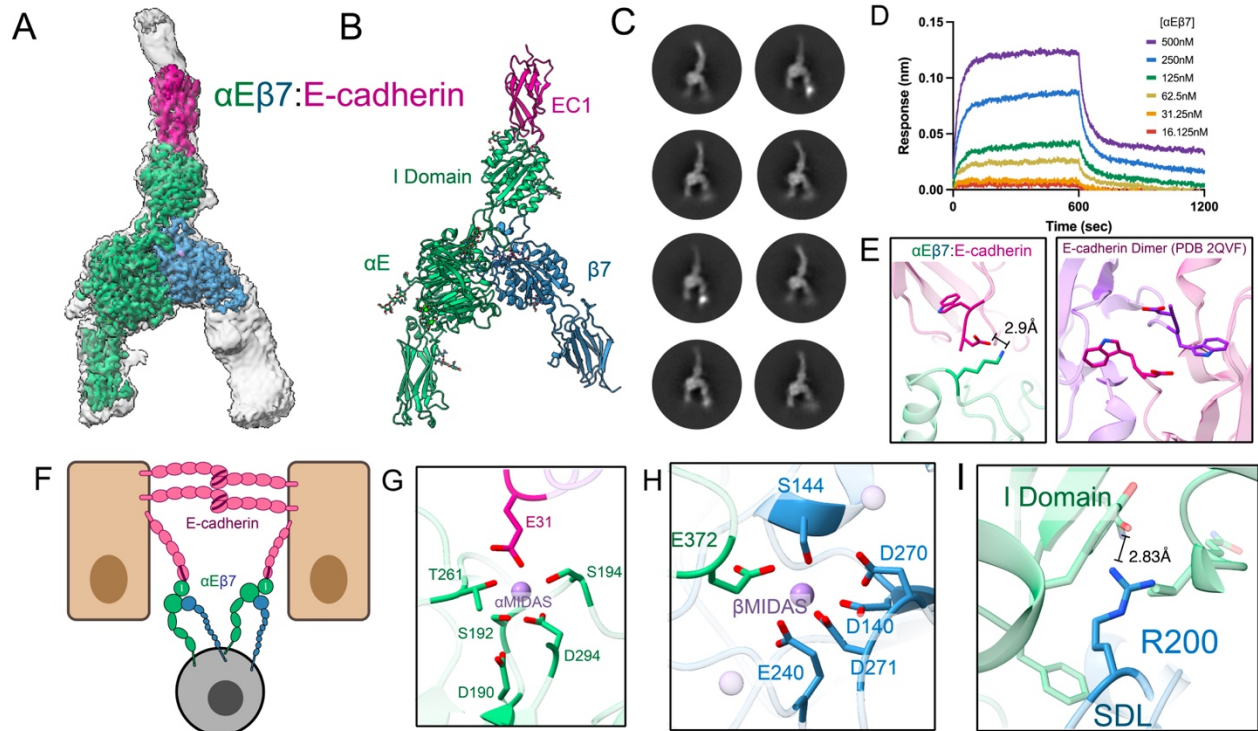


Figure 2.3. The molecular interface between integrin α E β 7 and E-cadherin. (A) CryoEM structure of the integrin α E β 7 ectodomain (green/blue) bound to the first two EC domains of ligand E-cadherin (pink). The low threshold, unsharpened global refinement map is shown in gray, and the inset high threshold, sharpened local refinement map is shown in color. The open integrin legs suggest an active conformation. (B) The atomic model of the α E β 7:E-cadherin complex. Upper black box indicates region of focus in (D). (C) Representative 2D class averages of α E β 7 bound to EC12 suggest α E β 7 binds exclusively to E-cadherin monomers. (D) Bio-Layer Interferometry (BLI) data suggest E-cadherin and α E β 7 form a high-affinity, stable interaction. (E) A salt bridge between the α E I domain and the first residue in the mature E-cadherin prevents the tryptophan stand-swap engagement essential for adherens junction formation(306). (F) A competitive inhibition model for T-cell tissue residence. α E β 7 (CD103) outcompetes E-cadherin trans-dimerization, permitting stable epithelial tissue residence by CD103+ cells. (G) The α E I domain (green) α MIDAS’ ion (purple) coordination sphere is completed by E-cadherin (pink) E31. (H) The β 7 β MIDAS coordination sphere is completed in the ligand bound α E β 7 via the internal ligand E372. (I) The β 7 R200 residue stabilizes the α E I domain with hydrophobic and charge-based interactions.

Integrin I domains contain a cation binding site (α MIDAS) coordinated by a conserved DXSXS motif. The α E I domain α MIDAS is ion-occupied in our structure, and the ion coordination complex is completed by E-cadherin E31 (Fig. 2.3G). Thus, consistent with other integrin I domains(159, 173, 222), α MIDAS coordination also leads to an open α E I domain, with its C-terminal helix assuming an extended conformation (Fig. S2.8D). Despite this dramatic shift, the angle of the I domain relative to the apo structure remains almost entirely unchanged upon ligand binding (Fig. S2.8E). These findings are consistent with our evidence from the compact structure that the α E I domain is exceptionally stable compared to other well-studied integrins. The ligand-bound α E β 7 integrin assumes an open headpiece conformation. However, 3D flexibility analysis of the dataset shows that the β 7 leg still maintains a notable, yet restricted, degree of free movement (27° β 7 hybrid movement compared to 68° in the apo open structure, Movie S3 in online publication). This intrinsic flexibility allows integrin to remain ligand-bound

and resident even while permitting dynamic cell movement and changes in membrane morphology.

In contrast to the unliganded structure, the I domain internal ligand is consistently engaged with $\beta 7$ when bound to E-cadherin. In this open conformation, the I domain interactions shift dramatically. For example, instead of contacting residues E372-D377 of the αE propeller hairpin in the closed structure, the “collar” contacts I domain residues Y366-I369 (Fig. S2.8F). This shift allows the αE I domain to maintain its remarkable rigidity in both inactive and active states and reinforces the helix-7 allosteric structural platform that links ligand binding to integrin activation. Our structures provide an unprecedented opportunity to visualize and compare conformational changes experienced by $\beta 7$ upon binding of an external ligand in the $\alpha 4\beta 7$ non-I domain configuration or by an internal ligand in an I-domain-containing $\alpha E\beta 7$ heterodimer. We find unambiguous support for the structural mimicry of external ligand-mediated integrin activation by the intrinsic I domain at key interaction sites. αE E372 in the $\alpha E\beta 7$ heterodimer coordinates the same β MIDAS ion as MAdCAM-1 D42 within the $\alpha 4\beta 7$:MAdCAM-1 complex (Fig. 2.3H). Moreover, the $\beta 7$ SDL uses the identical R200 residue to support MAdCAM-1 as it does the αE I domain (Fig. 2.3I). This leads to the same open conformation and similar leg flexibility in $\alpha 4\beta 7$ (35° $\beta 7$ hybrid movement) as seen in internally-liganded $\alpha E\beta 7$ (27° $\beta 7$ hybrid movement). The RMSD of the $\beta 7$ β I domain in ligand-bound $\alpha 4\beta 7$ versus $\alpha E\beta 7$ is $0.6\text{\AA}$, indicating that $\beta 7$ ultimately undergoes nearly identical conformational behavior in the context of either $\alpha 4$ or αE when activated. Thus, despite the presence of an entirely new domain, this suggests the core activation mechanism has been preserved between the ancestral and derived integrin states. Collectively, our $\beta 7$ structures reveal an intimate incorporation of the I domain with the ancestral integrin machinery.

2.3.4 The integrin I domain was derived from collagen

Our findings reveal that the evolutionarily ancient I domain insertion expanded the ligand-binding repertoire of integrin proteins without perturbing the ancestral intricate conformational changes required for activation. We speculated that, right from its inception, the I domain’s ancestor might already have encoded all the necessary machinery to facilitate the allosteric communication between ligand binding and integrin activation seen in our structures, including the dynamic helix-7. To test this hypothesis, we aligned a comprehensive selection of integrin I domains across *Olfactores* (tunicates and vertebrates) to find the common ancestor of all I domains. We used this alignment to generate a Hidden Markov Model (HMM) of the protein domain encompassing the I domain and internal ligand, which we used to query the genomes of human, coelacanth, echinoderms, and *Cephalochordata*, a sister group to *Olfactores*. This strategy enabled us to identify homology to the presumed I domain ancestor rather than the multiple other vWFA domains encoded by animal genomes(145) (Fig. S2.9A).

Our analyses reveal that the most likely ancestors of the α -integrin vWFA-like I domains are vWFA domains from extracellular matrix proteins, specifically those of the collagen VI class and its relatives (Fig. 2.4A). Outgroup collagen proteins share high structural homology with the αE I domain (Fig. 2.4B). To understand the origin of the I domain’s dynamic motion, we performed molecular dynamic simulations on ten vWFA domains, which include outgroup collagen domains and extant integrin I domains, from published experimental structures and high-confidence structural predictions (Fig. S2.10A, Table S2.2). We found that the structural dynamics are concentrated in homologous regions among these proteins. Moreover, in

addition to the overall vWFA fold, the dynamic motions of helix-7 are also conserved among these collagen and integrin I domains, supporting a common ancestry of these proteins (Fig 2.4C, Fig. S2.10A). Integrin I domain behavior is ion context-dependent; the presence of calcium, magnesium, or manganese cations differentially influences physiological activation states via structural ion coordination sites(26). To test whether outgroup domains display similar behaviors, we performed these simulations in each ion condition and projected the conformational landscape of each domain as a UMAP. We found that all collagen and integrin I domains sampled unique ion-specific conformations, suggesting that the differential behaviors of the integrin I domain under varying ions are an ancestral trait (Fig 2.4D, Fig. S2.10B).

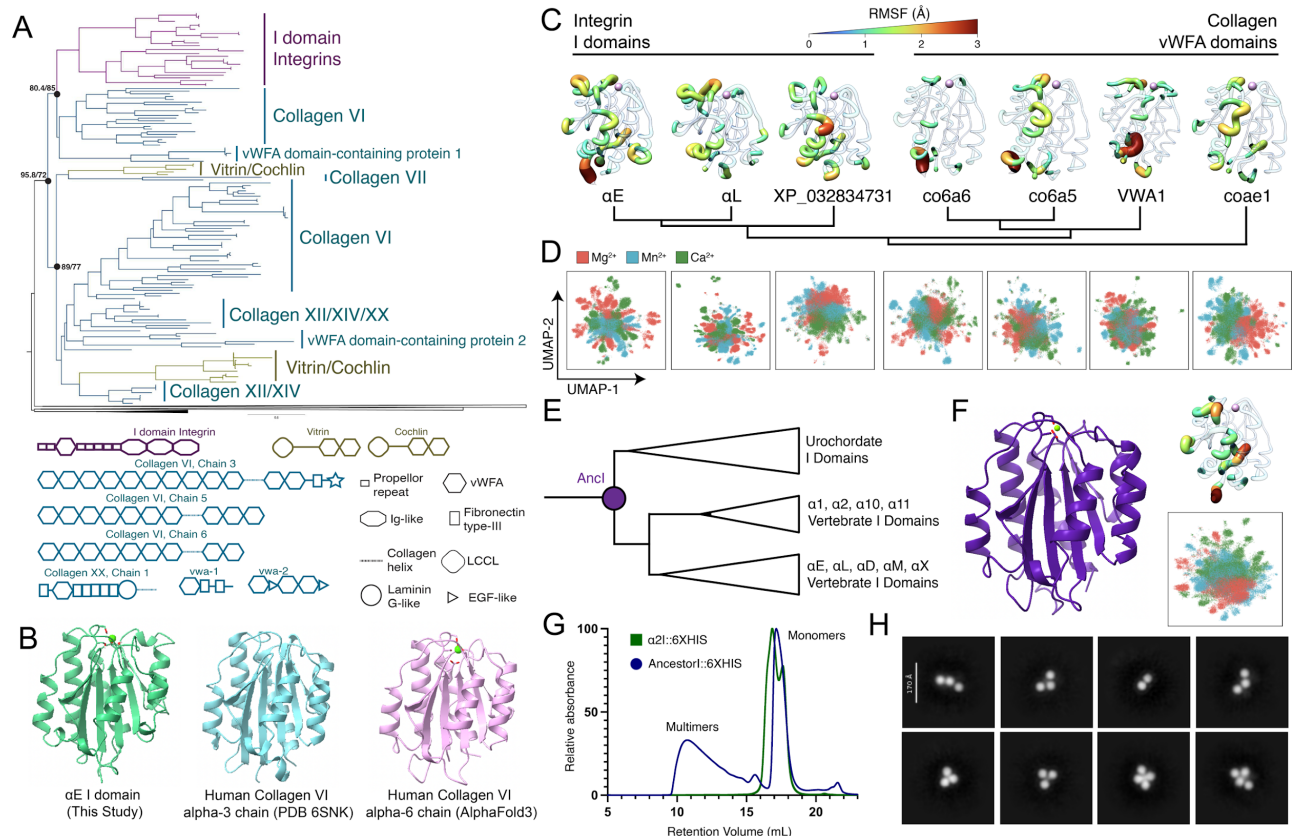


Figure 2.4. The integrin I domain was co-opted from an ancient collagen gene. (A) A phylogenetic tree of the VWF-like domains of the top hits from our HMM search was generated using IQ-TREE. The integrin I domain is most closely related to VWF-like domains from collagen VI proteins. (B) Atomic models of the closed I domain (this study), a human collagen VI alpha-3 (PDB 6SNK) vWFA-like domain, and an AlphaFold3 prediction of a vWFA-like domain of human collagen VI alpha-6 chain (Uniprot C3YQB2) show high visual structural similarities. (C) Molecular dynamics (MD) simulations of representative collagen or integrin vWFA domains bound with Mg²⁺ in the MIDAS site suggest characteristic dynamics were present in the common integrin I domain ancestor. The individual residue flexibility (i.e. root mean square fluctuations, RMSF) is projected on the three-dimensional structure and colored from 0 (blue) to >3 (red). Regions of fluctuations greater than 1.0 Å are highlighted. (D) UMAP representations of conformational landscape of the representative collagen or integrin vWFA domains in (C). 60,000 simulated conformations are projected on the UMAP and colored by the bound ion in the MIDAS site. (E) Phylogenetic tree schematic of sequences used for sequence reconstruction of the ancestral integrin I domain. (F) An AlphaFold3 structural prediction of the ancestral integrin I domain (left) and RMSF calculations and UMAP projection from MD simulations (right). RMSF is colored as in (C). UMAP is colored as in (D). (G) Size exclusion chromatography traces of recombinant 6XHIS-tagged I domains. The trace for the extant collagen-binding α2 I domain is in green, and the trace for the ancestral reconstruction construct is in blue. (H) Representative negative stain class averages of ancestral I domain multimers.

Collagen proteins, especially collagen VI, commonly form oligomeric structures within or between collagen types, which are essential for their homo-interactions(307). To directly test whether the ancestral I domain behaved similarly, we used PAML(308) to computationally reconstruct the common ancestor of all integrin I domains using the I domain alignment and phylogeny (Fig. 2.4E, Fig. S2.9B). AlphaFold2(309) modeling suggested this ancestor structurally resembles extant integrin I domains, and molecular simulations showed this ancestral domain contains the same dynamic behavior as extant integrin I domains (Fig. 2.4F, Fig. S2.10). We expressed and purified this ancestral reconstruction and size exclusion chromatography to reveal it forms multi-order oligomers; this contrasts with the extant $\alpha 2$ I domain, which only ran as a monomer (Fig. 2.4G). Negative stain analysis revealed oligomers highly reminiscent of those formed by extant collagen VI (Fig. 2.4H). Based on our analyses, we infer that the ancestral I domain could bind collagen immediately upon acquisition. Consistent with this inference, extant integrin I domains across the chordate phylogeny bind collagens of varying types, including collagen VI(310). Our ancestral multimer data are phylogenetically most consistent with an ancient avidity-based mode of collagen binding, like that utilized by an extant tunicate I domain(174). This ancestral mechanism is independent of the canonical collagen helical GFOGER motif more recently acquired by vertebrate integrin I domains(93). Furthermore, the ion-coordinating MIDAS, which is completely intact in the ancestral sequence, performs critical functions in extant collagens that contain it(311, 312), just as it does in integrin I domains. Our findings suggest that ion-dependent behavior was retained from the I domain's initial co-option.

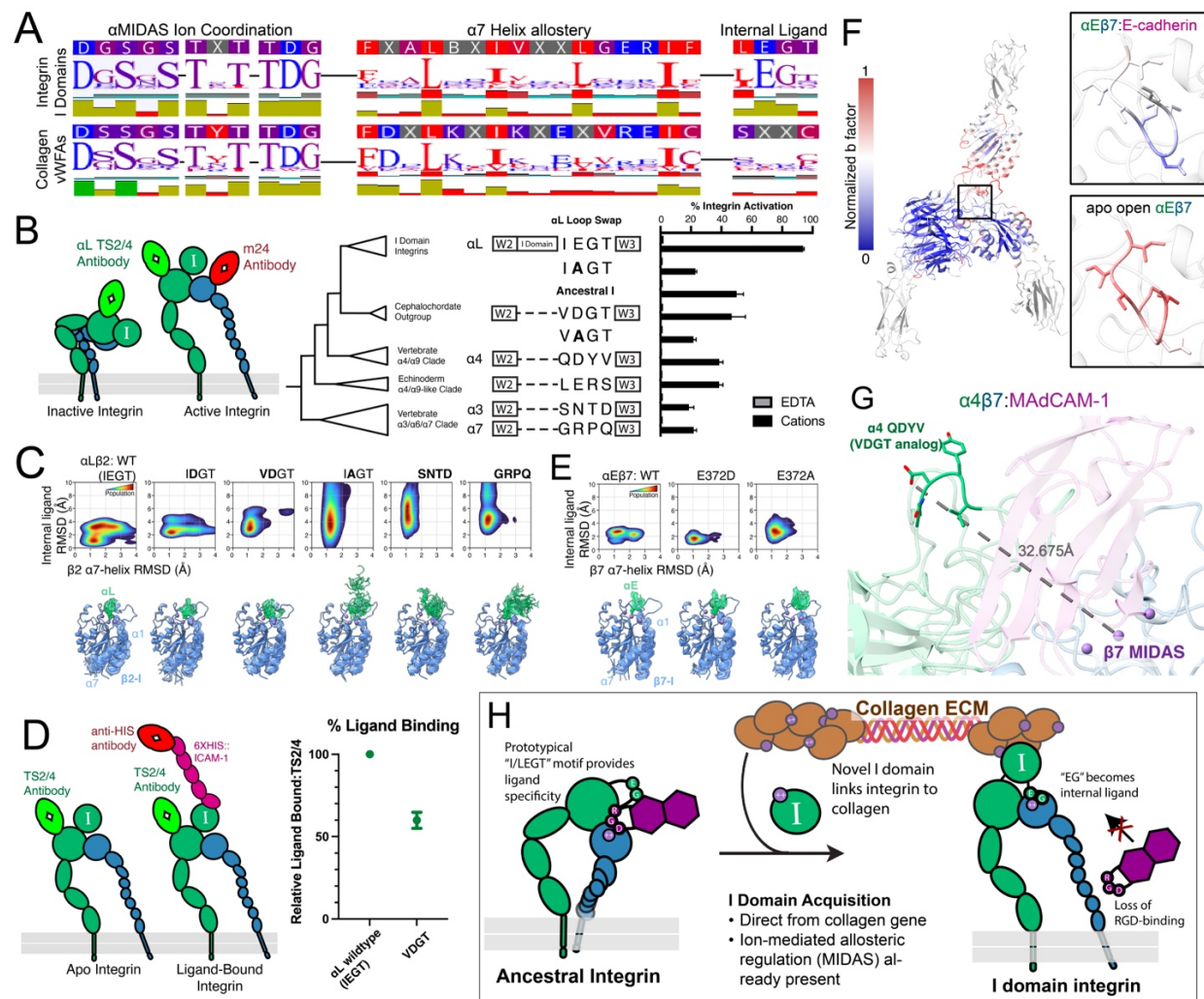
2.3.5 *The allosteric internal ligand was a de novo molecular exaptation*

Our finding of collagen-derived acquisition of the I domain still cannot explain the presence of the highly conserved I/LEGT motif, which contains the β MIDAS-coordinating glutamate required for internal ligand-based activation. We found no conserved matches to this motif either within collagen or within any other vWFA proteins in ProtKB metazoan genomes (Fig. 2.5A). Moreover, all integrin I domains lack the C-terminal cysteine, which is largely conserved in the collagen vWFA domains (Fig. 2.5A).

We thus hypothesized that the I/LEGT motif may have already been present in an ancestral α -integrin gene, but did not incur stringent selective constraints until after it acquired the I domain. To test this possibility, we performed sequence alignments and built phylogenetic trees of the α -integrins within the deuterostome clade containing the I domain α -integrins and the most closely related I-domain-lacking $\alpha 4/\alpha 9$ integrins (Fig. S2.11A). We looked specifically within the region between the second and third domains of the characteristic α -integrin beta-propeller where the I domain was first inserted. In all clades analyzed, we found a conserved acidic residue corresponding to the essential ion-coordinating glutamate within the internal ligand, suggesting this acidic residue was ancestrally present in the lineage of α -integrins that first acquired an I-domain. Moreover, we found a VDGT motif, highly reminiscent of the I/LEGT motif, in a small cephalochordate-specific clade, which is the closest-branching I-domain-lacking outgroup to the I domain α -integrins (Fig. S2.11B).

Using a flow cytometry assay and conformation-specific antibodies, we tested whether these motifs from the non-I domain containing α -integrins are sufficient to activate the extant I domain integrin $\alpha L\beta 2$ (Fig. S2.11C). Indeed, we found that each of these motifs, or ancestral

reconstruction of the entire internal ligand loop, could recapitulate approximately 35-50% of integrin activation. Furthermore, we demonstrated that this activity depended on ion coordination; mutating the acidic D/E residues to alanine in the wild-type I/LEGT or VDGT reduced activation to 20% (Fig. 2.5B). Moreover, adding EDTA to chelate the cation reduced activation in all backgrounds to essentially 0%. These data suggested that this cephalochordate motif was sufficient to activate integrin structurally. Conformational landscapes from molecular simulations revealed that these motif swaps are consistent with the exceptional stabilization conferred by the glutamate within the internal ligand, providing a molecular rationale for our cellular assay (Fig. 2.5C). VDGT is also sufficient to confer ligand binding capability to 60% of levels of wild type, providing a functional and structural rescue (Fig 2.5D, Fig. S2.12). MD simulations suggested this partial rescue is consistent with partial ion coordination occupancy upon loop mutation (Fig. 5C) and behaved similarly for $\alpha\text{E}\beta 7$ (Fig. 2.5E), revealing the dynamic basis of the essential ion-coordinating glutamate in regulating internal ligand binding and β subunit dynamics. This is further supported by b factor analyses from our $\alpha\text{E}\beta 7$ structures showing that ligand binding confers unique stability in this region, particularly E372 (Fig. 2.5F). **Figure 2.5. The internal ligand was exapted for integrin activation.** (A) Logo plots of alignments of either



representative I domains (top) or homologous domains from outgroup human collagens (bottom) show that many important sequence features were likely present in the integrin I domain common ancestor. The β MIDAS ion-

coordinating glutamate (right) was not found in outgroup sequences. The trace directly below the sequence represents hydrophobicity while the bottom trace displays sequence conservation. A 25% consensus identity is displayed above the sequence logo. **(B)** (left) A schematic visualization of the flow cytometry approach used to measure integrin activation in loop-swap experiments within an $\alpha\text{L}\beta 2$ backbone. TS2/4 is a conformation-agnostic αL antibody, while m24 recognizes the active $\beta 2$ subunit. (right) A representative phylogeny for loop-swap experiments. The internal ligand IEGT within the $\alpha\text{L}\beta 2$ backbone was replaced with homologous sequences from across the α integrin tree as indicated and the percent integrin activation was determined via flow cytometry of integrin-expressing HEK239t cells. EDTA reduced activation to essentially 0% in all cases. **(C)** Conformational landscape reveals loop swaps from **(B)** reveal differential coordination of the internal ligand and $\alpha 7$ helix dynamics. Below, representative trajectory snapshots of the β subunit (blue) and positional deviations of the $\alpha 1$ helix, $\alpha 7$ helix, and internal ligand (green). Manganese ions are shown as purple spheres. The MIDAS coordinating glutamate and positional substitution is shown as sticks. Structures were drawn from 25 clusters of the respective conformational landscape. **(D)** (left) A schematic visualization of the flow cytometry approach used to measure ligand binding. An anti-HIS antibody was used as a readout for ICAM-1 binding. The cephalochordate VDGT is sufficient to partially rescue ligand binding in the $\alpha\text{L}\beta 2$ backbone. Points represent mean, bars are standard deviation. **(E)** Conformational landscapes of $\alpha\text{E}\beta 7$ point mutants reveal the dynamic basis of the essential ion-coordinating glutamate in regulating internal ligand binding and β subunit dynamics. Representative snapshots are depicted below similar to **(C)**. **(F)** B factor comparisons of the ligand bound and apo $\alpha\text{E}\beta 7$ structures show unique stabilization of the ion-coordinating E upon ligand binding. **(G)** The molecular structure of $\alpha 4\beta 7$ shows the internal ligand homologous motif is too distant to coordinate the β MIDAS, and thus its function in the I domain background was exapted upon the I domain's insertion. **(H)** A model for the evolutionary origin of the novel integrin I domain.

This I/LEGT-like motif is not preserved in all α -integrins. Our structure of $\alpha 4\beta 7$ revealed that this loop region does not participate in ion coordination or integrin activation in I-domain-lacking integrins (Fig 2.5G) but instead serves to support ligand binding. Indeed, distinct non-VDGT $\alpha 4$ residues are essential for binding to VCAM-1, another $\alpha 4\beta 7$ ligand(313). Homology modeling of the cephalochordate-specific alpha integrin suggested a similar behavior in the direct outgroup (Fig. S2.11D). Thus, the allosteric ion-coordinating internal ligand was a serendipitous molecular exaptation, which could only have occurred in some lineages of α -integrins. This exaptation enabled them to maintain foundational conformation-based signaling while acquiring new ligand-binding capabilities upon I domain insertion. Following this initial exaptation, selection continued to fine-tune alpha integrins to reach their maximum activation threshold. This fine-tuning included the acquisition of the internal ligand “pocket” – a conserved structural series of buried residues capable of finely adjusting integrin kinetics(177) – that are highly diverged in the outgroup cephalochordate integrin (Fig. S2.11B).

2.4 Discussion

Our analyses retrace molecular history to present a structural framework for how a large, 500-million-year-old evolutionary insertion preserved necessary protein functions while serving as a launch pad for adaptation. Despite completely occluding the integrin's ancestral ligand binding site, the I domain remarkably retains the same functional output of an ancestral integrin ligand at multiple levels. In this way, the I domain served two capacities right from its inception: the ability to bind to an ion-coordinating ligand and the coordination of downstream integrin conformational changes. Given the collagen origin of the I domain itself, it is likely that the I domain's co-option into the integrin family catalyzed the ability of integrins to associate with the extracellular matrix directly (Fig. 2.5H). This ancestral feature(174, 314) has expanded and specialized in extant integrins(310).

Most eukaryotic proteins contain multiple distinct structural domains(253), including those that contain other vWFA domains. We anticipate that the form of evolutionary molecular

co-option we have uncovered has provided novelty across a wide array of molecular processes. Our analyses describe the unique and shared structural features of I-domain-mediated integrin activation that were critical for its evolutionary and immunological success.

2.5 Materials and Methods

2.5.1 DNA Constructs.

The regions encoding the wildtype α E ectodomain (M1-H1123, used for integrin:E-cadherin complex) followed by a C-terminal linker, HRV 3C cut site (LEVLFQGP), acidic coil motif (AQCEKELQALEKENAQLWELQALEKELAQ) and Strep-Tag II (WSHPQFEK*) inserted into a pcDNA3.1-Hygro(-)-like backbone was synthesized commercially (GenScript). The R177G/R178G, “RRtoGG,” mutant (clasped construct used for integrin:Fab complex) was generated via site-directed mutagenesis (NEB). The region encoding the β 7 ectodomain (M1-H723) followed by a C-terminal linker (GTGG), HRV 3C cut site (LEVLFQGP), basic coil motif (AQCKKKLQALKKKNAQLKWKLQALKKKLAQ) and 6xHIS tag inserted into a pcDNA3.1-Hygro(-)-like backbone was synthesized commercially (GenScript). The region encoding the α 4 ectodomain (M1-Q970) was amplified from Addgene plasmid #81178 and inserted into a pcDNA3.1-Hygro(-)-like backbone using standard molecular cloning techniques. The region encoding the MAdCAM-1 ectodomain (M1-Q317) was synthesized commercially as a G-block (IDT) and inserted into a pcDNA3.1-Hygro(-) with the same C-terminal linker, 3C cut site, basic coil motif, and 6xHIS tag as β 7 using standard cloning techniques. The region encoding the ICAM-1 ectodomain (M1-E480) was inserted into a pcDNA3.1-Hygro(-) with the same C-terminal linker, 3C cut site, and 6xHIS as for β 7 using standard cloning techniques. For Fc-tagged E-cadherin, the region encoding the E-cadherin ectodomain (D155-A698) with a CD33 signal sequence (GMPLLLLLPLWAGALA) (Gift from Barry Gumbiner) was inserted into a pcDNA3.1-Hygro(-)-like with a C-terminal HRV 3C cut site and human Fc tag amplified from Addgene plasmid 145164 using standard cloning techniques. The bacterial expression plasmid encoding an N-terminal 6xHIS-SUMO-tagged EC12 protein (E-cadherin EC domains 1 and 2, D155-N371) was a gift from Barry Gumbiner. The bacterial expression plasmids for the α 2 (D174-E365) and Ancl I domains were synthesized as G-blocks (IDT) and inserted into a pET24a vector using standard cloning techniques. The full length wildtype α L and β 2 constructs were made in ectodomain pcDNA3.1-Hygro(-) backbones by inserting G-blocks of the transmembrane regions by standard cloning techniques. Loop swaps of the α L IEGT internal ligand region were synthesized as G-blocks and inserted via standard cloning techniques.

2.5.2 Protein expression and purification.

Both integrins α E β 7 and α 4 β 7, as well as MAdCAM-1 and ICAM-1, were expressed in mammalian ExpiCHO cells (ThermoFisher). The basic expression and purification protocol is similar for each protein. Protein expression was performed according to the “Max Titer” manufacturer recommendations. For each integrin, a ratio of 3:2 α : β DNA was transfected for a total of 1 μ g of DNA per mL of cell culture. For MAdCAM-1 and ICAM-1 a total of 1 μ g of DNA per mL of cell culture was used for transfection. Cells were grown to a concentration of 7-10 x 10⁶ cells/mL at 37°C, 8% CO₂, and 90% humidity. The day of transfection, cells were split to 6 x 10⁶ cells/mL and allowed to recover for 2-4 hours. Cells were then transfected according

to manufacturer recommendations using Expifectamine CHO and grown overnight at 37°C, 8% CO₂, and 90% humidity. The next day, cells were given Enhancer and Feed according to manufacturer recommendations and shifted to 32°C, 5% CO₂, and 90% humidity to support protein expression. Five days post-transfection, cells were given a second dose of Feed according to manufacturer recommendations. 8-10 days post-transfection, cells were harvested via centrifugation for 15 minutes at 4°C and 1000xg. The cell supernatant containing secreted protein was further clarified for 20 minutes at 4°C and 30000xg and diluted 2:1 into HisTrap Binding Buffer (20mM NaPO₄, 500mM NaCl, 20mM imidazole, pH 7.4). The clarified supernatant was flowed over a 5mL HisTrap FF Crude column (Cytiva) equilibrated in HisTrap Binding Buffer using an ÄKTA Pure 25 L1 FPLC system. Bound protein was eluted in HisTrap Elution Buffer (20mM NaPO₄, 500mM NaCl, 500mM imidazole, pH 7.4). Fractions containing protein according to A₂₈₀ were pooled, concentrated to 500µL, and buffer-exchanged into integrin storage buffer (20mM Tris-HCl pH 7.4, 150mM NaCl, 1mM MgCl₂, 1mM CaCl₂) using an Amicon Ultra-15 concentrator (Millipore) with a 10kDa cutoff. For clasped integrin proteins and ICAM-1, the concentrate was then purified via gel filtration chromatography using a Superdex 200 Increase 10/300 SEC column (Cytiva) equilibrated with integrin storage buffer. Peak fractions containing integrin according to SDS-PAGE were pooled, concentrated to 2mg/mL, snap frozen in 10% (v/v) glycerol and stored at -80°C. For MAdCAM-1 and unclasped integrins, 1:10 (w/w) of 3C protease was added to the protein concentrate and incubated overnight with end-over-end rotation at 4°C. The next day the cleaved protein was further purified via gel filtration chromatography, concentrated, and stored as was done with the clasped protein.

Human Fc-tagged E-cadherin ectodomain was expressed and purified similarly to integrin and MAdCAM-1 constructs. Ectodomains were expressed in ExpiCHO cells according to the "Max Titer" manufacturer recommendations using a total of 1µg of DNA per mL of cell culture. Cells were fed and harvested and supernatants were clarified as with the integrin constructs. Supernatant was diluted 2:1 in HiTrap binding buffer (50mM Tris-HCl pH 7.4, 150mM NaCl, 3mM CaCl₂) and flowed over a 1mL HiTrap Protein G HP column (Cytiva) equilibrated in HiTrap binding buffer. The column was washed with 10 volumes of HiTrap binding buffer, and protein was eluted with 100mM glycine-HCl, pH 2.7 into 1M Tris-HCl, pH 9.0 to neutralize. Eluate was pooled and concentrated to 500µL and was then further purified via size exclusion chromatography on a Superdex 200 Increase 10/300 SEC column (Cytiva) equilibrated in 50mM Tris-HCl pH 7.4, 150mM NaCl, 3mM CaCl₂. Peak fractions were pooled, concentrated to 2mg/mL, snap frozen with 10% (v/v) glycerol and stored at -80°C.

EC12 was expressed and purified as previously described(306) with slight modifications. Briefly, BL21 DE3 cells containing the expression plasmid were grown to an OD₆₀₀ of 0.6, induced with 0.1mM IPTG, and incubated with shaking overnight at 18°C. The following day cells were harvested at 4000xg for 30 minutes at 4°C and resuspended in 10mL/L of cell culture in EC12 binding buffer (500mM NaCl, 20mM Tris-HCl pH 7.4, 3mM CaCl₂, 20mM imidazole) with EDTA-free cOmplete protease inhibitor cocktail (Roche). The cells were lysed via sonication on ice for 3 minutes and the lysate was clarified for 30 minutes at 4°C, 20000xg. The lysate was flowed over a 1mL HisTrap FF Crude column (Cytiva) equilibrated in EC12 Binding Buffer and washed with 10 CV of EC12 Binding Buffer. Bound protein was eluted in 500mM NaCl, 20mM Tris-HCl pH 7.4, 3mM CaCl₂, 250mM imidazole. Peak fractions were pooled and concentrated using an Amicon Ultra-15 concentrator (Millipore) with a 10kDa cutoff. The concentrate was dialyzed overnight at 4°C into 4L of 50mM Tris-HCl pH 7.4, 150mM NaCl,

3mM CaCl₂ with 250 units of 6xHIS-tagged SUMO protease (ThermoFisher) for scar-free SUMO tag cleavage. The following day the dialysate was run over a 1mL HisTrap FF Crude column to remove the protease and unprocessed EC12. The dialysate was then further purified via gel filtration chromatography on a Superdex 75 Increase 10/300 GL (Cytiva). Peak fractions containing EC12 were pooled, concentrated to 2mg/mL, and snap frozen in 10% v/v glycerol. We attempted to express C-terminally HIS-tagged EC12 in mammalian cell culture but proteins consistently precipitated following HisTrap elution.

Recombinant 6XHIS-tagged integrin I domains ($\alpha 2$ and AncI) were expressed and purified from BL21 DE3 LOBSTR cells. Cells containing the expression plasmid were grown to an OD₆₀₀ of 0.6, induced with 0.1mM IPTG, and incubated with shaking overnight at 18°C. The following day cells were harvested at 4000xg for 30 minutes at 4°C and resuspended in 10mL/L of cell culture in 20mM NaPO₄, 500mM NaCl, 20mM imidazole, pH 7.4, 1% Triton X-100, EDTA-free cOmplete protease inhibitor cocktail (Roche), and 0.5 mM PMSF added immediately before resuspension. Following an hour of 4°C incubation, the lysate was flowed over a 1mL HisTrap Excel column (Cytiva) equilibrated in 20mM NaPO₄, 500mM NaCl, 20mM imidazole, pH 7.4. I domains were eluted in 20mM NaPO₄, 500mM NaCl, 500mM imidazole, pH 7.4 in 1mL fractions. Peak fractions were pooled, concentrated to 500 μ L using an Amicon Ultra-15 concentrator (Millipore) with a 10kDa cutoff, and further purified or analyzed via gel filtration chromatography on a Superdex 200 Increase 10/300 GL (Cytiva) equilibrated in integrin storage buffer. I domain-containing fractions were concentrated to 10mg/mL in integrin storage buffer, snap-frozen in 10%v/v glycerol, and stored at -80°C.

2.5.3 Negative-stain EM sample preparation.

Integrin $\alpha E\beta 7$ was diluted to a final concentration of between 2 and 20 μ g/mL in buffer containing 20mM Tris-HCl pH 7.4, 150mM NaCl, and ion concentrations indicated in Fig. S2.4. 3 μ L of integrin was applied to a glow-discharged 400 mesh copper glider grid (Ted Pella) that had been covered with a thin layer of continuous amorphous carbon. AncI was diluted to 10 μ g/mL in integrin storage buffer and applied to grids in a similar way to $\alpha E\beta 7$. The grids were stained with a solution containing 2% (w/v) uranyl formate as previously described(315).

2.5.4 Negative-stain EM data acquisition and processing.

Data were acquired using a Thermo Fisher Scientific Talos L120C transmission electron microscope operating at 120 kV and recorded on a 4k x 4k Thermo Scientific Ceta camera at a nominal magnification of 92000x with a pixel size of 0.158 nm. Leginon(316) was used to collect 373 ($\alpha E\beta 7$ 5mM CaCl₂), 373 ($\alpha E\beta 7$ 1mM MgCl₂, 1mM CaCl₂), 320 ($\alpha E\beta 7$ 1mM MnCl₂), or 311 (AncI) micrographs at a nominal range of 1.5-2.5 μ m under focus and a dose of approximately 50 e⁻/Å².

Similar processing pipelines were used for all negative stain datasets. Micrographs were processed using cryoSPARC(317). CTF estimation was done using Patch CTF. Initially, 758281 ($\alpha E\beta 7$ 5mM CaCl₂), 736876 ($\alpha E\beta 7$ 1mM MgCl₂, 1mM CaCl₂), 629410 ($\alpha E\beta 7$ 1mM MnCl₂), or 621999 (AncI) particles were picked in a reference-free manner using blob picker in cryoSPARC. The particle picks were subjected to 1 round of particle curation and 1 round of 2D classification followed by 2D class selection for a final particle count of 17971 (5mM CaCl₂), 9092 (1mM MgCl₂, 1mM CaCl₂), 18552 (1mM MnCl₂), or 56638 (AncI) contributing to 2D

classes. The 10 clearest integrin classes for each buffer condition were chosen based on the visibility of the I domain and headpiece region.

2.5.5 *cryoEM sample preparation.*

The apo $\alpha 4\beta 7$ protein was diluted 1:10 in non-activating buffer (20mM Tris-HCl pH 7.4, 150mM NaCl, 5mM CaCl₂) to a final concentration of 0.2mg/mL. The RRtoGG $\alpha E\beta 7$ ectodomain was diluted 1:10 in non-activating buffer and incubated with digested LF61 Fab (Invitrogen) at a 1:4 molar ratio for 45 minutes at room temperature with end-over-end mixing. The integrin:Fab complex was subject to size exclusion chromatography in non-activating buffer and complex peaks were concentrated to 0.25 mg/mL. The wildtype $\alpha E\beta 7$ ectodomain was diluted 1:10 in activating buffer (20mM Tris-HCl pH 7.4, 150mM NaCl, 1mM MnCl₂)(26, 153, 318, 319) and incubated with EC12 at a 1:4 molar ratio for 45 minutes at room temperature with end-over-end mixing. The integrin:ligand complex was subject to size exclusion chromatography in activating buffer and complex peaks were concentrated to 1 mg/mL. Immediately before freezing, stock CHAPS detergent was added to the integrin:ligand complex to a final concentration of 0.05% (v/v). The $\alpha 4\beta 7$:MAdCAM-1 complex was formed and prepared in the same way as $\alpha E\beta 7$:EC12. Following dilution and complex formation steps, all samples were frozen with similar conditions. 3 μ L of protein was applied to Quantifoil grids (EMS) that were glow-discharged for 30s at 15 mA. The grids were blotted with a Vitrobot Mark IV (ThermoFisher) using a blot time of 3-7s and blot force of 4-5 at 100% humidity and 4°C. Grids were plunge-frozen in liquid ethane cooled by liquid nitrogen and stored in liquid nitrogen.

2.5.6 *CryoEM data acquisition and processing.*

The details of datasets and processing pipelines are outlined in Table 2.1, Fig. S2.1 ($\alpha 4\beta 7$:MAdCAM-1), Fig. S2.3 (apo $\alpha 4\beta 7$), Figs. S2.5 and S2.6 ($\alpha E\beta 7$:LF61 Fab) and Fig. S2.7 ($\alpha E\beta 7$:EC12). All datasets were collected on a Glacios cryo-transmission electron microscope (ThermoFisher) operating at 200 kV and recorded with a Gatan K3 Direct Detection Camera. Automated data collection was carried out using SerialEM(320). One hundred-frame movies were recorded in super-resolution counting mode at a nominal magnification of 36000x corresponding to a calibrated super-resolution pixel size of 0.561 Å/px. Each dataset was collected with a nominal defocus range of 1.2-1.8 μ m under focus and a dose of approximately 50 e⁻/Å².

Dose fractionated super-resolution movies were motion-corrected and binned 2x2 by Fourier cropping using MotionCor2(321) within the RELION(322) wrapper. From there, motion-corrected stacks were further processed in cryoSPARC according to each processing pipeline. Briefly, particles were picked using the blob picker, subjected to multiple rounds of 2D and/or 3D classification using *ab initio* and heterogeneous refinement using initial models generated from the data, and further refined with non-uniform and local refinement. Masks for local refinements were generated using UCSF ChimeraX(323) and cryoSPARC. Map sharpening was performed using the Cosmic2(324) server wrapper of DeepEMhancer(325) with the high resolution setting. Local resolution estimation was performed in cryoSPARC. 3D FSC(326) plots were generated from <https://3dfsc.salk.edu/>.

CryoSPARC was used for all motion and variability analyses. We used the cluster setting in 3D Variability Analysis(299) followed by homogeneous and heterogeneous refinement to sort

the $\alpha\text{E}\beta 7$:LF61 Fab particles into open or closed I domain states. Maps showing the conformational range of the internal-liganded $\beta 7$ leg in the Fab-bound structure (Fig 2.2G) were generated using 3D Classification in PCA mode. We generated the volume series showing the continuous motion of the $\beta 7$ in ligand-bound structures using 3DFlex(327). All map visualizations (images and movies) were generated and recorded using UCSF ChimeraX.

2.5.7 Model Building.

For the Fab-bound initial $\alpha\text{E}\beta 7$ model we used an AlphaFold2(309) model of αE and PDB: 3V4P(288), Chain B for $\beta 7$. We predicted the LF61 Fab residues using ModelAngelo(328) and fit them into mouse IgG1 constant residue sequences. The initial models were manually fit into the density using UCSF ChimeraX, followed by dock-in-map in Phenix(329). For the initial $\alpha\text{E}\beta 7$:EC12 model we used the AlphaFold2 αE model, a homology model of $\beta 7$ built in SWISS-MODEL(330) using PDB: 7NWL(14), chain B as the reference, and PDB: 4ZT1(331), Chain A for EC12. We generated the initial $\alpha 4\beta 7$:MAdCAM-1 model using AlphaFold-Multimer(332).

Models were manually adjusted in an iterative way between COOT(333) and ISOLDE(334) within UCSF ChimeraX. The $\beta 7$ hybrid domain was flexibly fit into the ligand-bound models using ISOLDE. The $\alpha\text{E}\beta 7$:LF61 Fab model was refined using the closed I domain map. Glycans were built manually in COOT using the “Carbohydrate” module. Models were built using a combination of the non-uniform and locally-refined, unsharpened and sharpened maps. Rosetta(335) was used to depict secondary structure. All maps used for model building have been deposited.

Electrostatic molecular interactions were determined using the PISA server(336). RMSD calculations for the βI domain were done in ChimeraX using the “matchmaker” function with $\beta 7$ residues V133-S369.

B factor analyses were done as in Jin et al. 2024(115). Briefly, b factors were calculated from Phenix Real-space Refine(329) and normalized to the most stable region of the αE headpiece, the beta-propellor. Graphics were generated by coloring models by b factors in UCSF ChimeraX.

2.5.8 System setup for molecular dynamics simulations.

To understand the ancestral conformational dynamics of the integrin I domain, we simulated a representative sample of integrin I domains (αE , αL , $\alpha 2$, *Petromyzon marinus* XP_032834731), sister clade collagen-derived vWFA domains (co6a6 800-988, co6a5 433-618, coeal 1024-1210, VWA1 29-211), RPN10 to be used as a distant outgroup, and the ancestral reconstruction AncI (described in the following sections). The initial coordinates of the studied I domains were obtained from the cryo-EM model (αE : this study), prior crystal structures (αL : 3F74(337), $\alpha 2$: 5HJ2(338), RPN10: 5LN1(339)), AlphaFold2 (XP_032834731) or AlphaFold3(340) predicted models (co6a6 800-988, co6a5 433-618, coeal 1024-1210, VWA1 29-211). AlphaFold3 predicted models were generated using the online web server (alphafoldserver.com) and AlphaFold2 predictive models were generated in Colabfold(341). The proteins were solvated in a periodic truncated octahedron with OPC water molecules and 150mM NaCl. One magnesium, manganese, or calcium ion was placed in the αMIDAS site. For RPN10, no additional ion was modeled.

To investigate the effect of amino acid substitution on the internal ligand, we modeled truncated integrin systems of α L β 2 bound with ICAM-1 and α E β 7 bound to E-cadherin. In all, the truncated models contain the respective α subunit β -propeller, α subunit I-domain, and β -subunit I-domain. The initial model for α L β 2 was generated by combining SWISS-MODEL homology models of the open β 2 subunit using β 7 as a template (this study), the open α L subunit using PDB: 4NEH(172) as a template, and the ICAM-1-bound α L I domain (PDB 1MQ8(173)). The cryoEM structure α E β 7 bound to E-cadherin was used as the initial model. Both systems were solvated in a periodic truncated octahedron with OPC water molecules and 150mM NaCl. 4 manganese ions were added to the α MIDAS, MIDAS, ADMIDAS, and SyMBS coordinating sites. 3 calcium ions were added to the calcium binding sites of the α subunit.

2.5.9 Molecular dynamics protocols.

All simulations were performed using OpenMM 7.7.0 (I-domain systems)(342) or OpenMM 8.2.0(343) (integrin headpiece systems) employing AMBER19ffSB force field(344), OPC water model, and 12-6-4 Li Merz ion parameters(345). Prior to production, systems were energy minimized for 20,000 steps and equilibrated at 298K (I-domain systems) or 310K (integrin headpiece systems) with backbone atoms harmonically restrained by a 5 kcal/mol*Å² force constant for 10 ns. Langevin dynamics was performed using a Langevin integrator with an integration timestep of 4 fs and collision rate of $\sqrt{2}$ ps⁻¹ and hydrogen mass repartitioning(346). Pressure was maintained using a Monte Carlo barostat with an update frequency of 100 steps. Nonbonded interactions were calculated with a distance cutoff of 10Å. Trajectory snapshots were saved every 100 ps during production simulations. Trajectory lengths and number of replicates of each simulated system are listed in Table S2.2. Simulations were performed on either the Fred Hutch Cancer Center clusters on Nvidia GTX 2080 or in-house workstations on Nvidia GTX 3090.

Trajectories were processed using in-house scripts utilizing CPPTRAJ 6.18.1(347) and MDTraj 1.10.2(348) packages. For Fig. 2.4C and Fig. S2.10, RMSD and RMSF measurements were performed on C α atoms using the initial structure as the reference and considered the last 2.0 μ s of each simulation trajectory. For Fig. 4D, UMAP dimensionality reduction was performed using the UMAP 0.5.6 python package(349). Backbone torsion angles (phi, psi, and omega) were used as features for UMAP with the following parameters: n_neighbors = 200, max_dist = 0.99. For Fig. 5C and Fig. 5E, the last 0.8 μ s of production simulations were considered. To calculate positional displacement (no-fit RMSD) of the internal ligand and helix α 7 of the β subunit, structures were first superpositioned by C α of the β subunit. For the internal ligand, the C α atoms for 2 residues before and after the MIDAS coordinating glutamate residue (or equivalent position, 5 residues total) were considered for RMSD analysis. For helix α 7 of the β subunit, β 2 residues 348 to 363 or β 7 residues 375 to 389 C α atoms were used. 25 clusters from each conformational landscape were generated using a KMeans algorithm implemented in scikit learn 1.5.0 (Fig. S2.12).

2.5.10 Bio-Layer Interferometry (BLI).

Assays were performed on an Octet Red (ForteBio) instrument at 25°C with shaking at 1,000 RPM. Protein A biosensors were hydrated in activating buffer with the addition of 0.1% BSA and 0.02% Tween-20 (BLI buffer) for 15 min. Fc-tagged E-cadherin ectodomain was

loaded at 20 $\mu\text{g/mL}$ in BLI buffer until a threshold of 1.2nm was reached. A baseline equilibration step was performed in BLI buffer for 2 min. Association of $\alpha\text{E}\beta 7$ in BLI buffer at various concentrations in a two-fold dilution series from 500nM to 16.125nM was carried out for 600 s prior to dissociation for 600 s. The data were baseline subtracted prior to fitting performed using a 1:1 binding model and the ForteBio data analysis software. Mean K_D values were determined with a global fit applied to all data from three independent replicates. A representative sensorgram of the three replicates is displayed in the relevant figure.

2.5.11 Sequence curation and Hidden Markov Model (HMM) searching.

Protein sequences for all human integrin subunits were curated in the NCBI database wrapper within Geneious Prime. Representative I domain sequences from tunicates were curated using BLAST with a urochordate taxon restriction and human αL as a search query. Hidden Markov Model (HMM) searching of UniProtKB was performed on the online HMMER server from the European Bioinformatics Institute(350) with the full integrin I domain alignment as an initial query. Since no sequences containing the internal ligand I/LEGT motif were found, we restricted our search to include exclude this region of the I domain alignment and included only results from representative species of human, coelacanth, echinoderms and *Branchiostoma*. Sequences with an individual E-value of less than $1\text{e-}30$ were considered hits and used for subsequent alignment and phylogenetic analyses.

2.5.12 Sequence alignments.

All protein sequence alignments were generated using the default MUSCLE algorithm(351) and visualized using Geneious Prime. Alignments were further curated manually to remove large gaps and sequences with low quality. Sequence alignments are available as supplemental text files in FASTA or PHYLIP format with the following names: Alignment of all human α subunits (Data S1 in online publication) except αE which was excluded due to an extra “X” domain preceding the I domain that is unique to αE , Alignment of integrin I domains across humans and tunicates used for HMM search and ancestor reconstruction (Data S3 in online publication), Alignment of all hits from HMM search (Data S4 in online publication), Alignment of expanded integrin clade encompassing I domain integrins and near neighbors (Data S5 in online publication). Sequence names are provided as protein names, NCBI accession numbers, Uniprot or ENSEMBL identifiers or Aniseed Transcript IDs.

2.5.13 Flow Cytometry

HEK293T cells were passaged in 10cm tissue culture-treated plates every 2-3 days in DMEM+L-glutamine supplemented with 10% FBS and 1% Pen-strep. For integrin activation and ligand binding experiments, HEK293T cells were transfected in 6-well tissue culture-treated plates at 70% confluence using Lipofectamine 2000 according to manufacturer’s instructions. Each well was transfected with 2.5 μg of a 1.5:1 ratio mixture of DNA encoding the αL variants and $\beta 2$ wildtype genes, respectively. Cells were washed 24 hours after transfection and fresh media was applied. Cells were trypsinized and washed 36 hours after transfection and harvested for subsequent experiments.

For integrin activation experiments, each well of cells was centrifuged for 5 minutes at 500xg, washed with warmed 20mM Tris-HCl pH 7.4, 150mM NaCl centrifuged again and resuspended in 500μL of either 20mM Tris-HCl pH 7.4, 150mM NaCl, 5mM EDTA or 20mM Tris-HCl pH 7.4, 150mM NaCl, 1mM MgCl₂ + 1mM MnCl₂. Following a 30 minute rocking incubation at 37°C, 1μL of each FIT-C TS2/4 and APC m24 antibody (BioLegend) was added to each cell suspension. Cells were incubated for one hour, washed with 1mL and resuspended in 100μL of the appropriate buffer, and placed on ice for flow analysis. >10000 cells for each condition were analyzed on either a BD FACSCanto II or FACSsymphony A5. The presence of FITC and APC were measured based on an unstained wildtype integrin control from the same 6-well plate. Each condition was repeated 3 times.

For ligand binding assays, 2x10⁶ cells from each well were centrifuged for 5 minutes at 500xg, washed with warmed 20mM Tris-HCl pH 7.4, 150mM NaCl centrifuged again and resuspended in 500μL of 20mM Tris-HCl pH 7.4, 150mM NaCl, 1mM MgCl₂ + 1mM MnCl₂. Following a 30 minute rocking incubation at 37°C, purified HIS-tagged ICAM-1 ectodomain was added to each cell suspension to a final concentration of 20μg/mL. Cells were incubated another hour at 37°C, washed with ice-cold buffer and resuspended in 100μL buffer with 1μL of each FIT-C TS2/4 and iFluor647 anti-HIS tag antibody. Following a 1 hour incubation at 4°C cells were washed with ice-cold buffer and resuspended on ice in 100μL buffer. >10000 cells for each condition were analyzed on a BD FACSCanto II. The presence of FITC and iFluor647 were measured based on an unstained wildtype integrin control from the same 6-well plate. The percentage of iFluor647-positive cells was normalized to the wildtype control. Each condition was repeated 3 times.

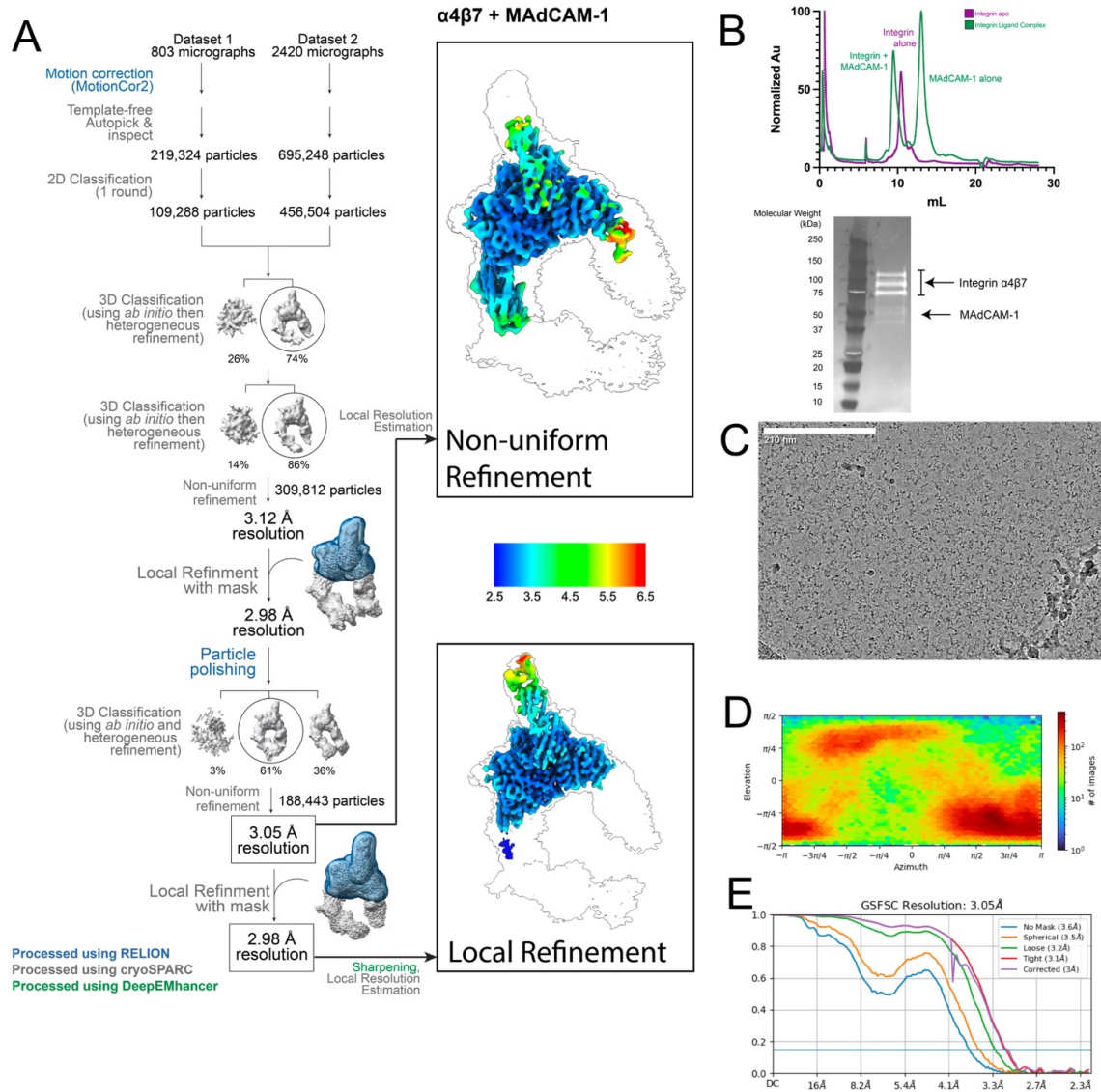
2.5.14 Phylogenetic Analysis.

All phylogenetic trees were generated using a maximum likelihood framework within IQ-TREE(352). Substitution models for each alignment were estimated with IQ-TREE. Both 1000 ultrafast bootstrap and SH-aLRT replicates are displayed as branch support statistics, also generated in IQ-TREE. Trees were visualized in FigTree(353).

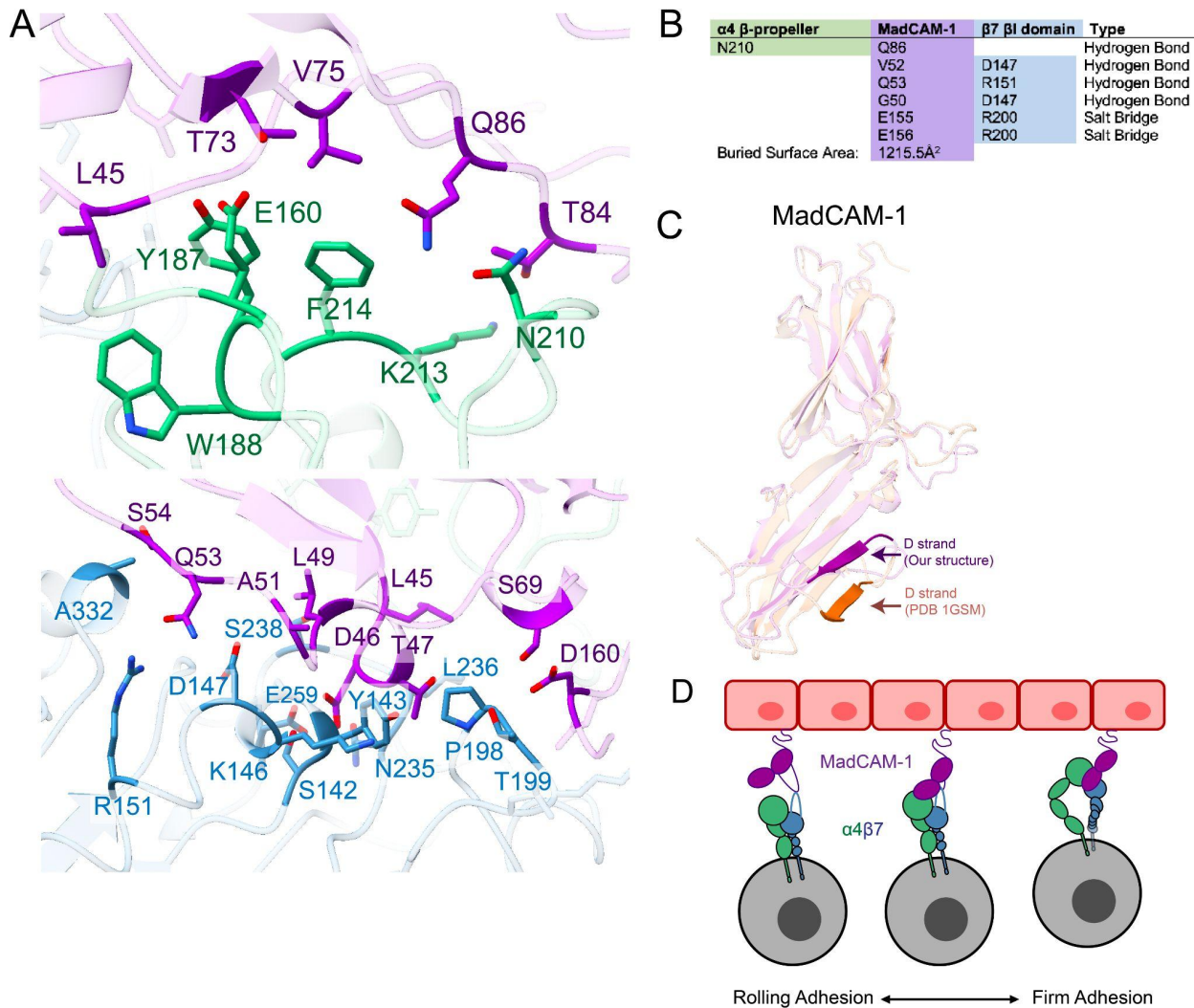
2.5.15 Ancestral Sequence Reconstruction

In order to reconstruct the ancestral I domain sequence, we curated and manually edited an amino acid alignment of I domains from integrins in both vertebrate and urochordate lineages with the closely related cephalochordate collagen vWFA domain as the outgroup. Maximum likelihood-based ancestral reconstruction was carried out using the codeml package in PAML 4.8(308), and the marginal reconstruction at each site was used as the Ancl sequence.

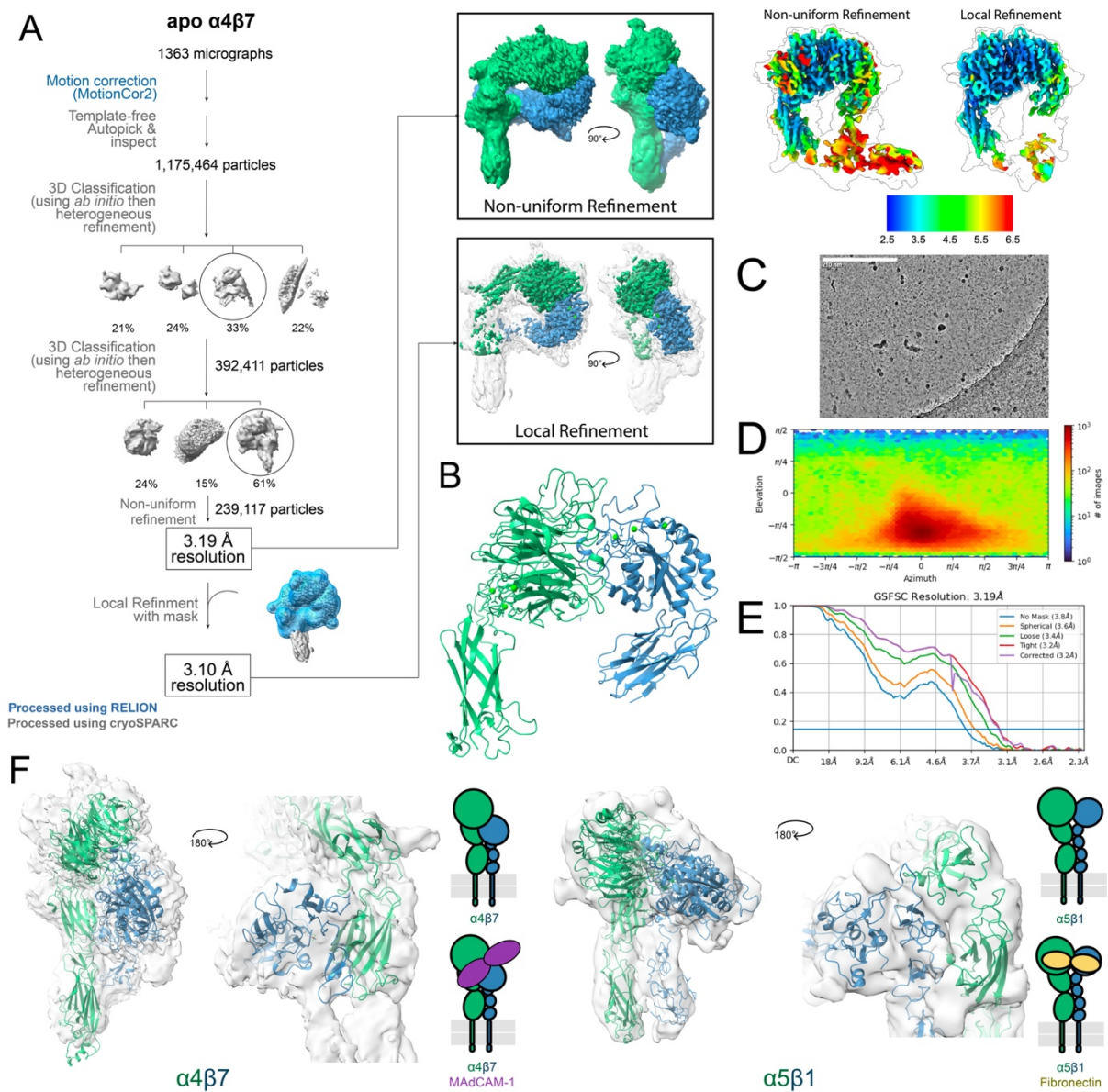
2.6 Supplemental Information



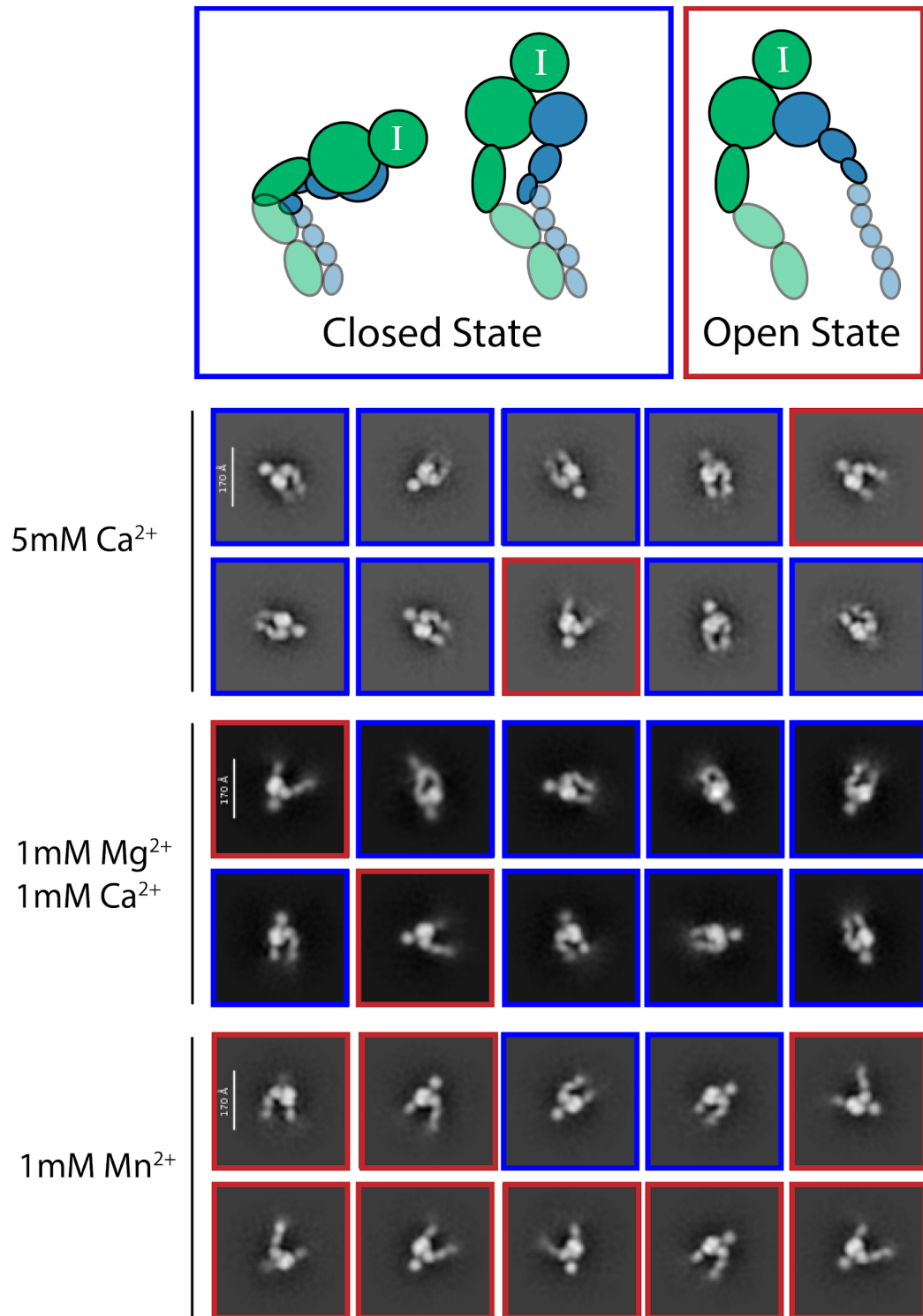
Supplemental Figure 2.1. Data processing schematic for the $\alpha 4\beta 7$:MAdCAM-1 complex. (A) A flowchart for the data processing pipeline of $\alpha 4\beta 7$:MAdCAM-1. Both the global and local refinements were used for model building. (B) Size exclusion chromatography traces showing peak shift for ligand-bound integrin, with SDS-PAGE of the protein complex fraction shown below. (C) Representative micrograph with 210nm scale bar, (D) orientational distribution plot, and (E) gold-standard Fourier Shell Correlation (GSFSC) plot for the globally refined map.



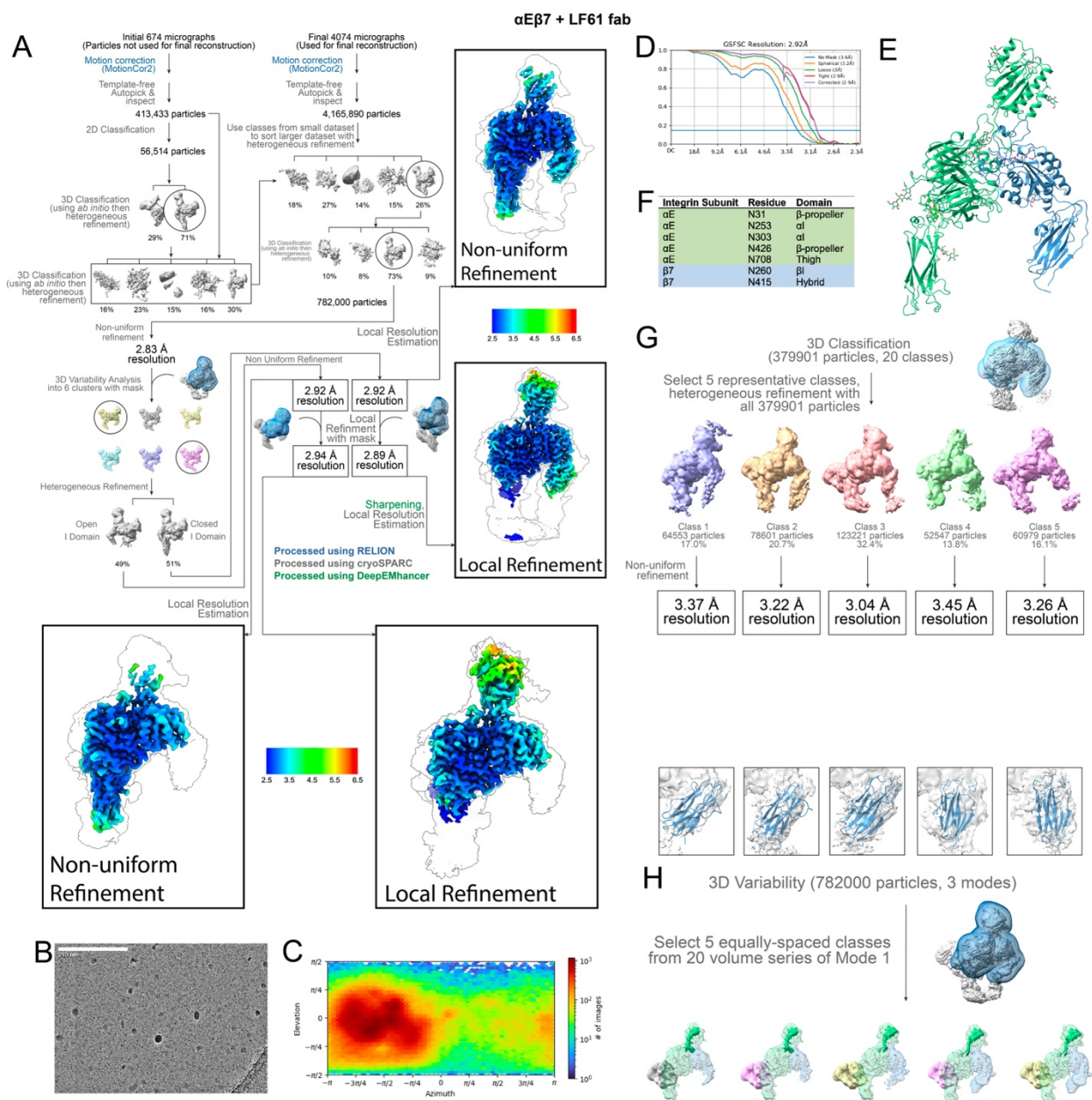
Supplemental Figure 2.2. The molecular interface between integrin $\alpha 4 \beta 7$ and MadCAM-1. (A) MadCAM-1 (purple) binding within the $\alpha 4 \beta 7$ groove is stabilized by contacts with both $\alpha 4$ (green, top) and more extensively with $\beta 7$ (blue, bottom). (B) The PISA server was used to determine electrostatic interactions in the $\alpha 4 \beta 7$:MadCAM-1 complex model (C) MadCAM-1 undergoes a conformational shift upon binding to $\alpha 4 \beta 7$. The D strand in the first Ig-like domain of MadCAM-1 is shown in high opacity for both our structure (purple) and the crystal structure (orange, PDB 1GSM). The conformation of the D strand in the crystal structure of MadCAM-1 alone would sterically clash with our integrin density. (D) An integrative model for how $\alpha 4 \beta 7$ -expressing immune cells structurally identify surveillance sites, slow to roll, and eventually firmly adhere along the gut endothelium.



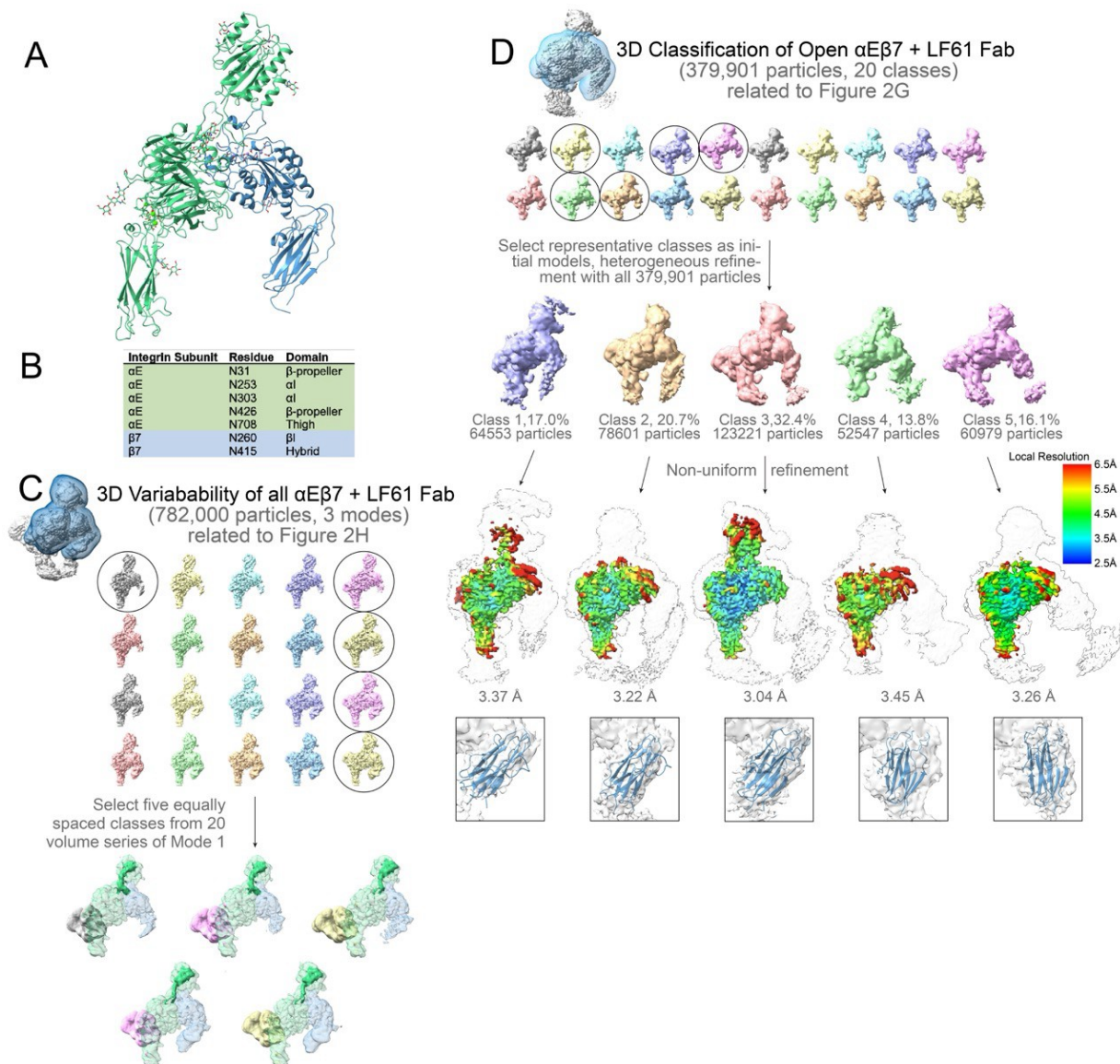
Supplemental Figure 2.3. Integrin $\alpha 4 \beta 7$ has a novel compact state. (A) A flowchart for the data processing pipeline of apo $\alpha 4 \beta 7$. (B) The molecular model of the $\alpha 4 \beta 7$ headpiece. (C) Representative micrograph with 210nm scale bar, (D) orientational distribution plot, and (E) gold-standard Fourier Shell Correlation (GSFSC) plot for the apo $\alpha 4 \beta 7$ structure. (F) Unlike integrin $\alpha 5 \beta 1$, the compact half-bent state of $\alpha 4 \beta 7$ has a significant rotation (40°) at the headpiece mediated by contacts between the two subunits. This may contribute to binding to MAdCAM-1, which binds $\alpha 4 \beta 7$ perpendicular compared to how $\alpha 5 \beta 1$'s ligand fibronectin binds⁸⁴.



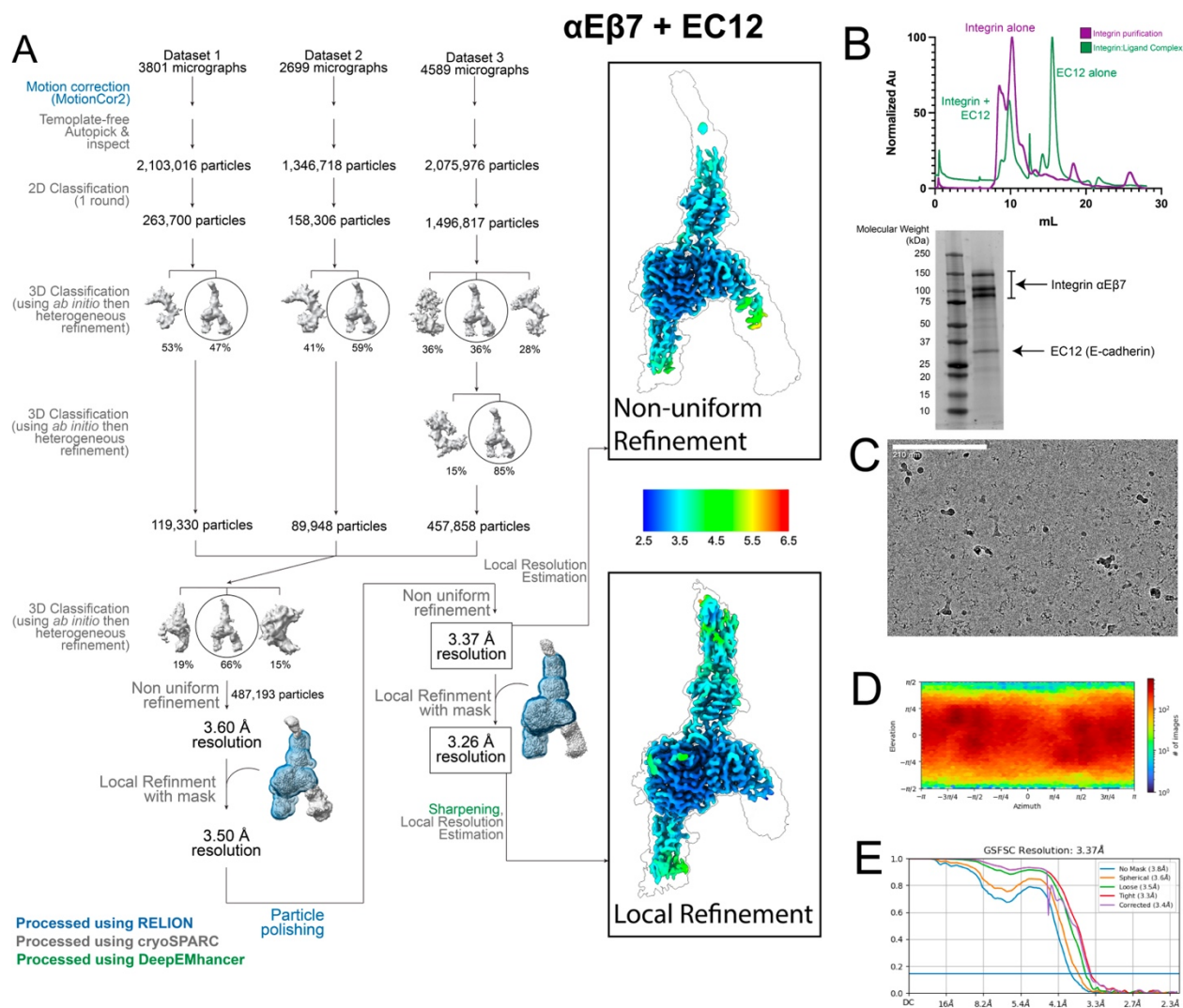
Supplemental Figure 2.4. Integrin $\alpha E\beta 7$ occupies canonical conformations. Negative stain electron microscopy (nsEM) 2D class averages of integrin $\alpha E\beta 7$ in buffers of varying ions show that $\alpha E\beta 7$ samples conformations like those previously described for other integrins. The I domain is clearly resolved in each class, however unlike other integrins the lower leg regions remain unclear in negative stain.



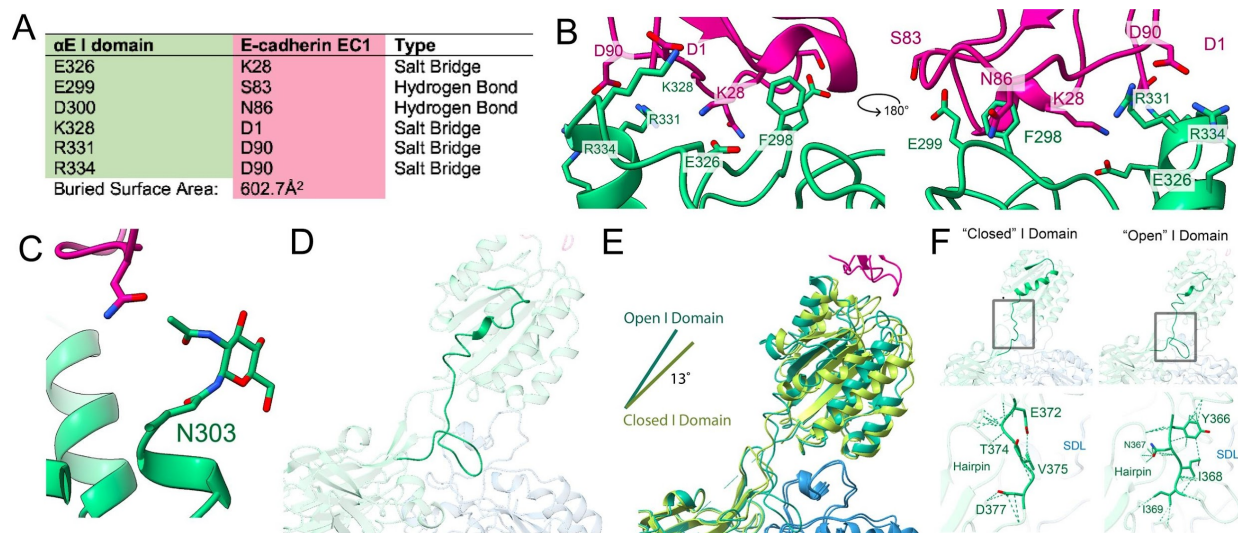
Supplemental Figure 2.5. Data processing schematic for the α E β 7:LF61 Fab complex. (A) A flowchart for the data processing pipeline of α E β 7:LF61 Fab. High-quality particles were sorted into two major classes; those that have an open I domain and those that have a closed I domain. The closed I domain structure, shown on the right with local resolution estimates, was used to model the inactive α E β 7 conformation. The open structure is shown below. (B) Representative micrograph with 210nm scale bar, (C) orientational distribution plot, (D) gold-standard Fourier Shell Correlation (GSFSC) plot for the globally refined closed I domain map. (E) The model of the open I domain structure for α E β 7. (F) Glycan residues and domains within the α E β 7:LF61 model. (G) 3D classification details for the classes presented in Figure 2.5G. The “Open I domain” particles (shown in panel A) were subclassified into 20 classes without alignments. Five representative classes with varying β 7 leg positions from this classification were used as initial models for a subsequent classification using 3D heterogeneous refinement. Finally, each class was refined to high resolution via non-uniform refinement. The β 7 hybrid domain was rigid-fit in each map to calculate leg flexibility. (H) 3D variability processing schematic for the volume series presented in figure 2.5H.



Supplemental Figure 2.6. Large-scale and small scale movement characterization of apo αEβ7. (A) The model of the open I domain structure for αEβ7. (B) Glycan residues and domains within the αEβ7:LF61 model. (C) 3D variability processing schematic for the volume series presented in figure 2.5H. (D) 3D classification to illustrate the large scale movement of the beta-leg relative to the β7 βI domain. The “Open I domain” particles shown in Fig. S2,5 were subclassified into 20 classes without alignments. Five representative classes with varying β7 leg positions from this classification were used as initial models for a subsequent classification using 3D heterogeneous refinement. Finally, each class was refined to high resolution via non-uniform refinement. The β7 hybrid domain was rigid-fit in each map to calculate leg flexibility.

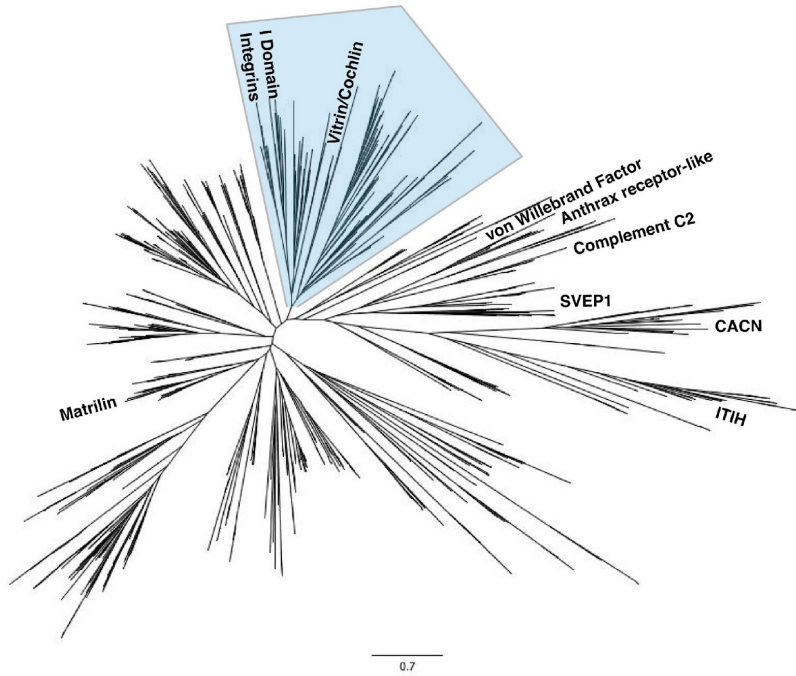


Supplemental Figure 2.7. Data processing schematic for the αEβ7:EC12 complex. (A) Flowchart for the data processing pipeline of αEβ7:EC12. Both the global and local refinements were used for model building. (B) Size exclusion chromatography traces showing peak shift for ligand-bound integrin, with SDS-PAGE of the protein complex fraction shown below. (C) Representative micrograph with 210nm scale bar, (D) orientational distribution plot, and (E) gold-standard Fourier Shell Correlation (GSFSC) plot of the globally refined map.

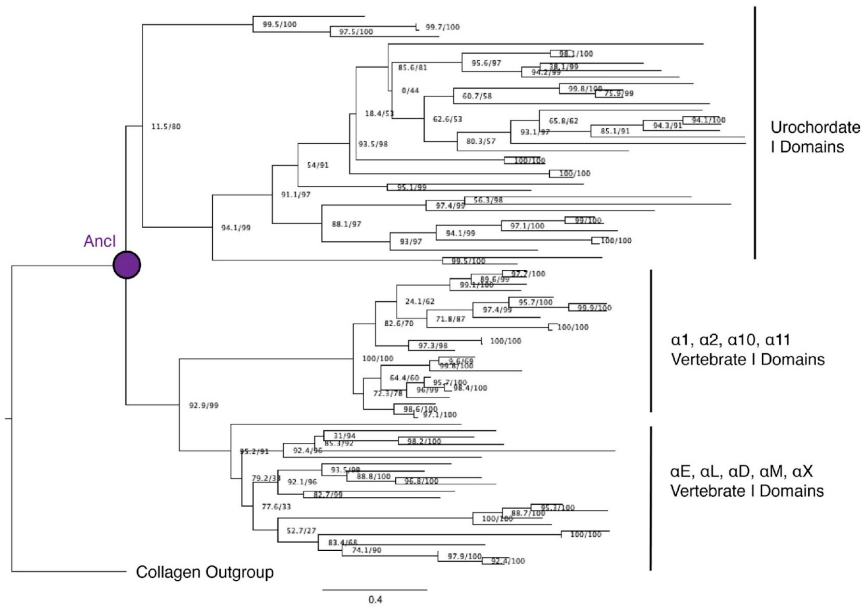


Supplemental Figure 2.8. Molecular features of the open integrin α E β 7. (A) Electrostatic interactions in the α E β 7:EC12 complex model between the α E I domain and EC1. (B) The interface between α E β 7 and E-cadherin is composed primarily of a central hydrophobic residue, F298, that is surrounded by electrostatic interactions. Views are 180° rotations. (C) The N321 glycosylation on the α E I domain contributes to ligand binding. (D) The open structure of the α E I domain α 7 helix. (E) Despite extensive structural changes within the I domain, the location of the I domain relative to the rest of the integrin molecule remains the same when ligand-bound. (F) Extensive low-strength contacts stabilize the I domain's internal ligand in either the apo closed state (left) or ligand-bound open state (right).

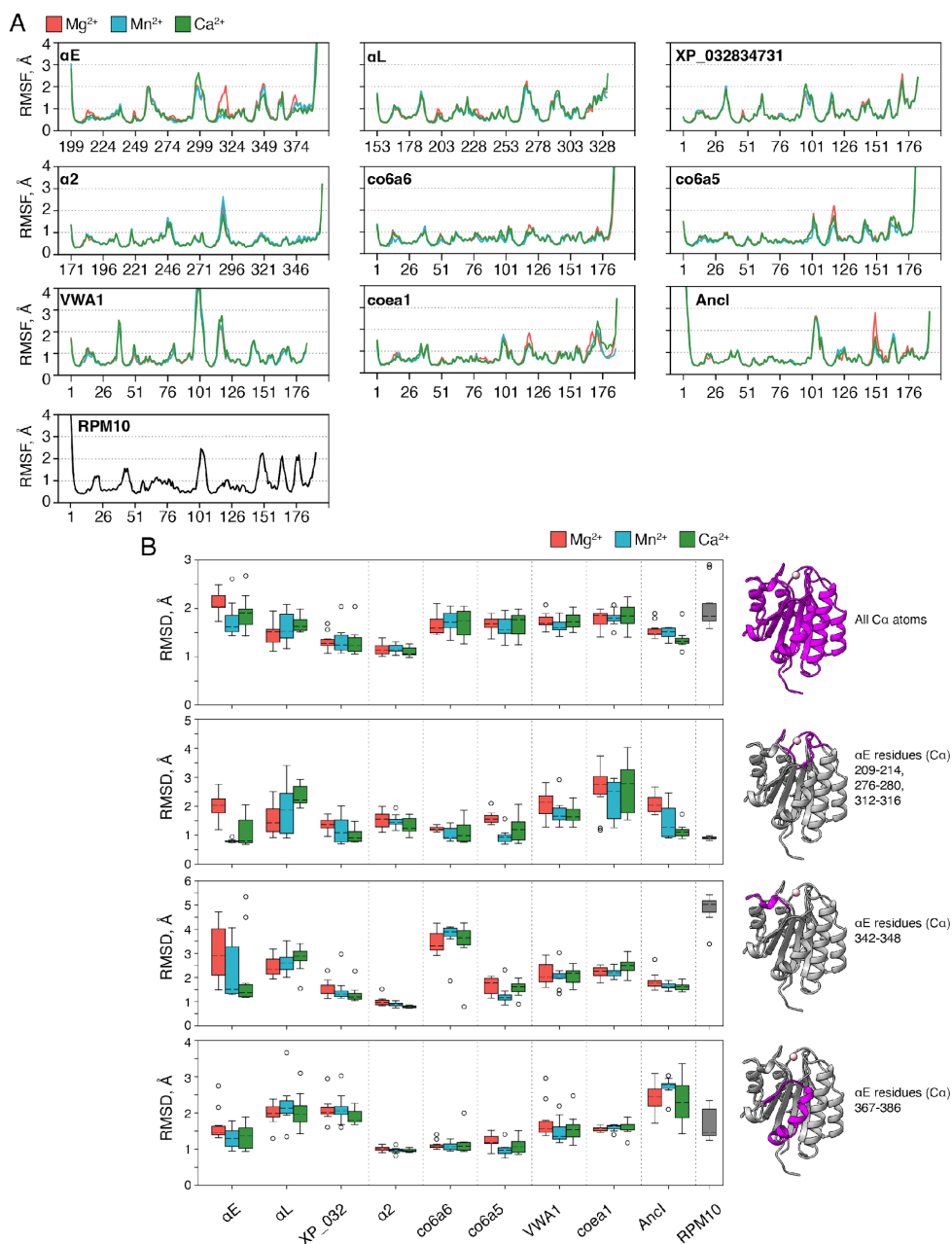
A



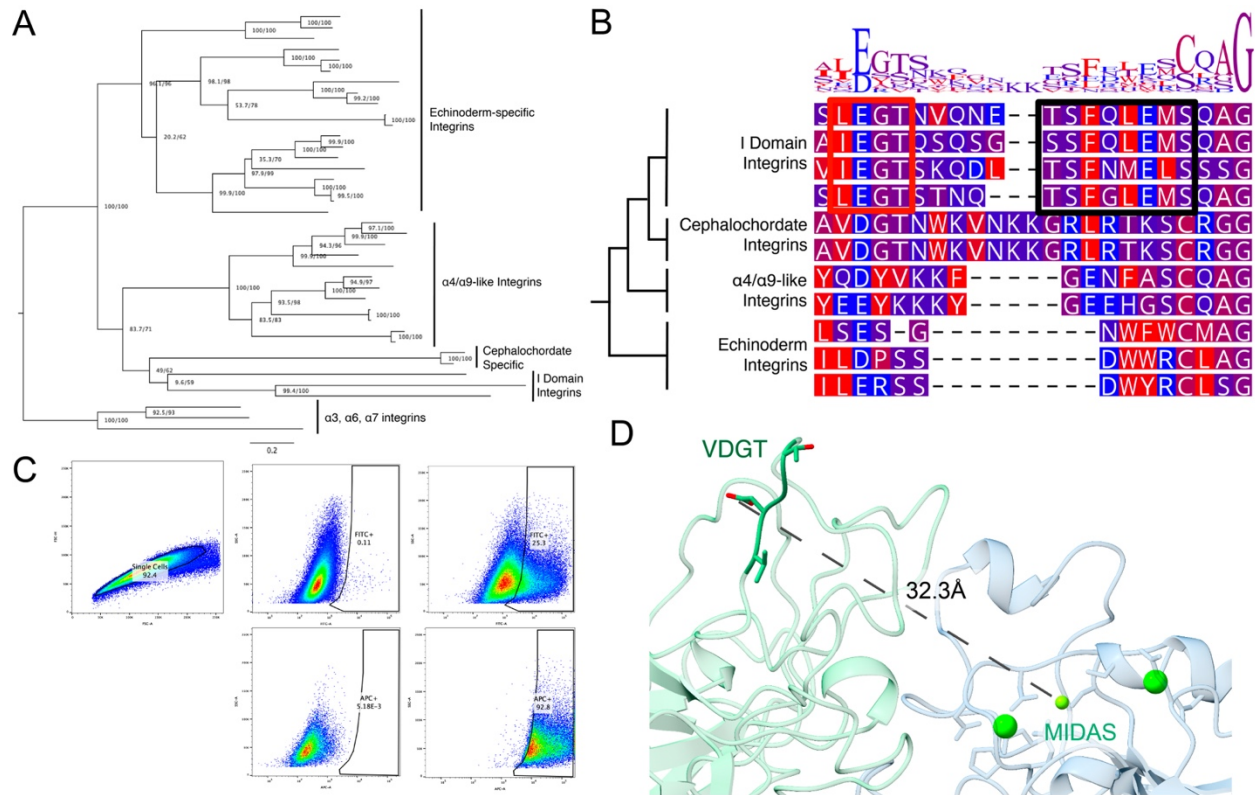
B



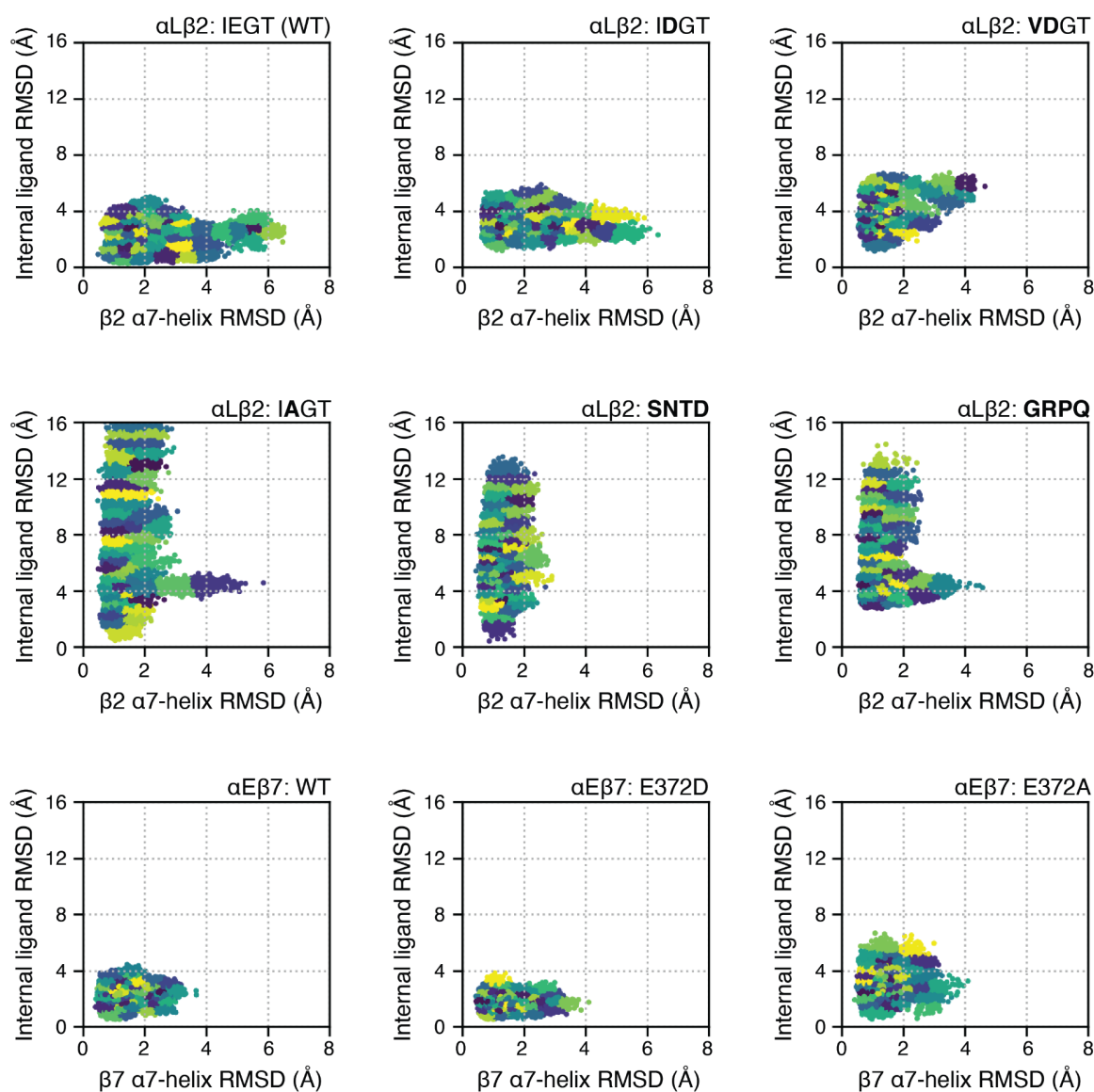
Supplemental Figure 2.9. Evolutionary origins of the integrin I domain. (A) A phylogenetic reconstruction of vWFA domains with close homology to the integrin α I domain was generated in IQ-TREE. Blue box indicates subregion of tree presented in Figure 2.4A. (B) A phylogenetic tree of curated I domain sequences used for ancestral sequence reconstruction. UNIPROT C3YQB2 was used as a representative cephalochordate outgroup sequence. Node values are displayed as “ultrafast bootstrap/SH-aLRT”.



Supplemental Figure 2.10. Molecular dynamics simulations reveal conserved dynamics in vWFA domains. (A) Root mean square fluctuation (RMSF) of 10 simulated vWFA proteins with either Mg^{2+} (red), Mn^{2+} (blue), or Ca^{2+} (green) bound in the MIDAS site. For RPN10, no ion was bound. (B) RMSD calculation of the simulated vWFA proteins. Color represents the cation bound in the MIDAS site (Mg^{2+} , red; Mn^{2+} , blue; Ca^{2+} , green). Regions calculated, from left to right, all Ca atoms, ion coordinated loops, and sites of conformational change and colored magenta in the structure. Error bars represent standard error. For both (A) and (B), values are averaged across 10 independent replicates of 2.4 μs long with only the last 2.0 μs used for analysis and the initial simulation structure used as the reference.

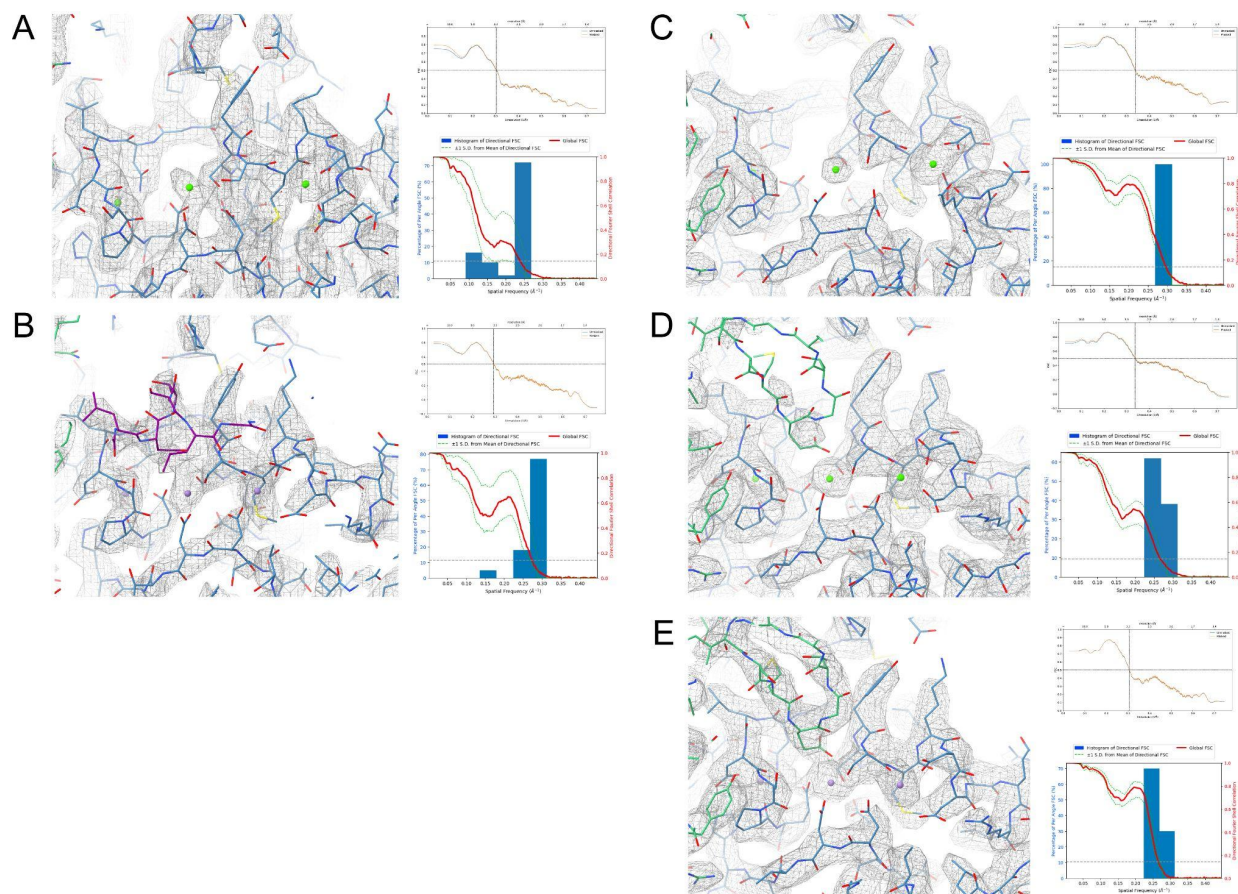


Supplemental Figure 2.11. Phylogenetic and structural analyses of the internal ligand origin. (A) Phylogenetic tree of I domain-containing and related integrins. A small clade of genes unique to cephalochordates are the direct sisters of the I domain integrins. Node values are displayed as “ultrafast bootstrap/SH-aLRT”. (B) Alignment of the loop region between the second and third propellor blade domains of the I domain and related integrins. The direct cephalochordate outgroup contains a VLDGT motif sequentially homologous to the I domain internal ligand, while the broader family also encompassing the $\alpha 4/\alpha 9$ clade contains a conserved acidic residue akin to the essential ion-coordinating glutamate in the internal ligand. Red square indicates the internal ligand in extant integrins. The black square indicates the structural internal ligand pocket that is largely diverged between I domain integrins and their outgroups. (C) Gating strategy used to determine FITC+ and APC+ cells in flow cytometry. Top left: Single cell gating. Top middle: FITC+ gating, negative control. Top right: FITC+ gating, positive control. Bottom middle: APC+ gating, negative control. Bottom right: APC+ gating, positive control. (D) Structural homology model of the cephalochordate sister clade I domain-less integrin. The VLDGT motif is too distant to the β MIDAS to perform ion coordinating activities, and is thus more likely involved in ligand binding.



Supplemental Figure 2.12. Conformational landscape of α L β 2-ICAM-1 and α E β 7-E-cadherin headpiece domains.

Raw conformational landscape projected on the reaction coordinates of the positional deviation (no-fit RMSD) of the internal ligand and α 7 helix of the β subunit clustered in 25 KMeans clusters.



Supplemental Figure 2.13. Map quality and validation metrics. Representative map-model fits of the same $\beta 7$ ion coordination region (left), map-model FSC curves (top right), and 3DFSC curves (bottom right) for **(A)** the apo $\alpha 4\beta 7$, **(B)** the $\alpha 4\beta 7$:MAdCAM-1 complex, **(C)** the $\alpha E\beta 7$ apo closed I domain, **(D)** the $\alpha E\beta 7$ apo open I domain (as expected, the density for the αE internal ligand is relatively weak at this high threshold), and **(E)** the $\alpha E\beta 7$:E-cadherin complex with stronger internal ligand density. In all figures: β subunits are blue; α subunits are green; MAdCAM-1 is purple; E-cadherin is pink.

Table S2.1. Collection, refinement and validation statistics for cryoEM density maps and models. Collection details for all cryoEM maps presented in the text are shown above, and validation statistics for the models subsequently generated from those maps are shown below.

	α E β 7:LF61 Closed I domain (EMDB-xxxx) (PDB xxxx)	α E β 7:LF61 Open I domain (EMDB-xxxx) (PDB xxxx)	α E β 7:EC12 (EMDB- xxxx) (PDB xxxx)	α 4 β 7:MAAdCAM-1 (EMDB-xxxx) (PDB xxxx)	apo α 4 β 7 (EMDB- xxxx) (PDB xxxx)
Data collection and processing					
Magnification	36000x	36000x	36000x	36000x	36000x
Voltage (kV)	200	200	200	200	200
Electron exposure (e ⁻ /Å ²)	50	50	50	50	50
Defocus range (μm, nominal)	1.2-1.8	1.2-1.8	1.2-1.8	1.2-1.8	1.2-1.8
Pixel size (Å)	1.122	1.122	1.122	1.122	1.122
Symmetry imposed	C1	C1	C1	C1	C1
Initial particle images (no.)	4,165,890	4,165,890	5,307,702	914,572	1,175,464
Final particle images (no.)	402,099	379,901	487,193	188,443	202,130
Map resolution (Å)	2.92	2.93	3.37	3.05	3.20
FSC threshold	0.143	0.143	0.143	0.143	0.143
Map resolution range (Å)	2.5-10.2	2.5-16.8	2.8-10.9	2.5-46.5	2.7-47.6
Refinement					
Initial model used (AlphaFold or PDB code)	AlphaFold PDB: 3V4P	α E β 7:EC12	AlphaFold PDB: 7NLW, 4ZT1	AlphaFold- multimer	AlphaFold
Model resolution (Å)	2.92		3.37	3.05	3.20
FSC threshold	0.143	0.143	0.143	0.143	0.143
Model resolution range (Å)	2.5-10.2		2.8-10.9	2.5-46.5	
Model composition					
Non-hydrogen atoms	8933	8986	9792	9150	7447
Protein residues	1135	1147	1250	1162	960
Carbohydrates	22	18	18	22	0
B factors (Å²)					
Protein	113.5	71.4	119.7	87.1	66.14
Ligand	N/A	N/A	N/A	N/A	N/A
R.m.s. deviations					
Bond lengths (Å)	0.012 (11)	0.005 (0)	0.012 (12)	0.012 (10)	0.007 (2)
Bond angles (°)	1.949 (52)	0.662 (4)	2.038 (61)	2.048 (73)	0.748 (9)
Validation					
MolProbity score	0.78	1.69	0.82	0.97	2.13
Clashscore	0.17	4.79	0.05	0.22	14.69
Poor rotamers (%)	0.21	1.35	0.09	0.20	0.61
Ramachandran plot					
Favored (%)	96.63	95.00	95.73	94.37	92.68
Allowed (%)	3.19	5.00	3.95	5.55	7.32
Disallowed (%)	0.18	0	0.32	0.09	0.00

Table S2.2. Molecular dynamics systems simulated in this study. Trajectory lengths and number of replicates of the simulated systems and respectively ligands are shown below.

System	Ligand	Initial Model	Replicates	Trajectory Length (per replicate, μs)
<i>Integrin headpiece</i>				
α E β 7	E-cadherin	This study	4	1.0
α E β 7_E372D	E-cadherin	This study	4	1.0
α E β 7_E372A	E-cadherin	This study	4	1.0
LFA1	ICAM-1	Homology model	6	1.0
LFA1_IDGT	ICAM-1	Homology model	6	1.0
LFA1_VDGT	ICAM-1	Homology model	6	1.0
LFA1_IAGT	ICAM-1	Homology model	6	1.0
LFA1_SNTD	ICAM-1	Homology model	6	1.0
LFA1_GRPQ	ICAM-1	Homology model	6	1.0
<i>I domain</i>				
α E	Mg ²⁺	This study	10	2.4
α E	Mn ²⁺	This study	10	2.4
α E	Ca ²⁺	This study	10	2.4
α L	Mg ²⁺	PDB: 3F74	10	2.4
α L	Mn ²⁺	PDB: 3F74	10	2.4
α L	Ca ²⁺	PDB: 3F74	10	2.4
α 2	Mg ²⁺	PDB: 5HJ2	10	2.4
α 2	Mn ²⁺	PDB: 5HJ2	10	2.4
α 2	Ca ²⁺	PDB: 5HJ2	10	2.4
XP_032834731	Mg ²⁺	AlphaFold2	10	2.4
XP_032834731	Mn ²⁺	AlphaFold2	10	2.4
XP_032834731	Ca ²⁺	AlphaFold2	10	2.4
co6a6	Mg ²⁺	AlphaFold3	10	2.4
co6a6	Mn ²⁺	AlphaFold3	10	2.4
co6a6	Ca ²⁺	AlphaFold3	10	2.4
co6a5	Mg ²⁺	AlphaFold3	10	2.4
co6a5	Mn ²⁺	AlphaFold3	10	2.4
co6a5	Ca ²⁺	AlphaFold3	10	2.4
vwa129	Mg ²⁺	AlphaFold3	10	2.4
vwa129	Mn ²⁺	AlphaFold3	10	2.4
vwa129	Ca ²⁺	AlphaFold3	10	2.4
coea1	Mg ²⁺	AlphaFold3	10	2.4
coea1	Mn ²⁺	AlphaFold3	10	2.4
coea1	Ca ²⁺	AlphaFold3	10	2.4
AncI	Mg ²⁺	AlphaFold3	10	2.4
AncI	Mn ²⁺	AlphaFold3	10	2.4

AncI	Ca ²⁺	AlphaFold3	10	2.4
RPN10	None	PDB: 5LN1	10	2.4

2.7 Afterword

This work was done in collaboration with Dr. Harmit Malik, my co-advisor, and the incredible Dr. Matthew Chan who performed all the molecular dynamics simulations and accompanying analyses that added excellent support to our model of adaptation. The review process for this paper was particularly challenging- probably the toughest part of my entire doctoral experience- but the feedback we got really helped flesh out the story and expand beyond what we initially intended in an exciting way. I'm grateful to everyone who read and responded to drafts of this work, and although it was tough to see its worth immediately after publication, I'm now so proud of what we were able to accomplish.

Chapter 3 Avidity unlocks promiscuity in a divergent integrin receptor family

**This chapter was adapted from Hollis et al. 2026.*

3.1 Abstract

Integrins are a family of cell surface receptors that bind proteinaceous extracellular ligands to signal environmental cues across the plasma membrane. Integrins vary dramatically in their ligand binding profiles, with some displaying exquisite specificity for a single ligand and others showing remarkable promiscuity. Using structural, mutational, cellular and phylogenetic approaches, we define the ligand binding criteria for two closely related integrins: $\alpha\text{L}\beta 2$ (LFA-1, CD11a/CD18), which binds almost exclusively to its cognate ligand ICAM-1, and $\alpha\text{M}\beta 2$ (CR3, Mac-1, CD11b/CD18), which binds a diverse array of structurally disparate ligands. We find that the more discerning $\alpha\text{L}\beta 2$ has strong, entrenched restraints that maintain a highly selective binding surface, whereas the determinants of promiscuous $\alpha\text{M}\beta 2$ binding vary considerably. By obtaining $\alpha\text{M}\beta 2$ structures with complexed endogenous and pathogenic ligands, we identify molecular tradeoffs in integrin-ligand binding at both direct and allosteric sites, distinguish signatures of pathogen mimicry and escape, and identify noncanonical ligand binding pathways that allow for $\alpha\text{M}\beta 2$'s exceptional promiscuity. Our studies uncover the molecular basis of a dramatic divergence in specificity that informs our understanding of how receptor families tune ligand recognition.

3.2 Introduction

Cell surface receptors face the daunting task of discriminating among the exceptionally diverse array of extracellular signals in a complex environment. To do this, cell surface receptors such as chemoreceptors, signaling kinases, and integrins divide their ligand-recognition capacities across paralogs that collectively constitute a larger protein family(3). Receptors within a family often range in their ligand specificities, with some displaying intricate specificity for one ligand and others having a large repertoire of recognized signals. This pattern is exemplified by the αI domain-containing integrins, a family of α/β heterodimeric cell surface receptors which recognize proteinaceous ligands primarily at the distal “I” or “inserted” domain of the corresponding α subunit(3, 7).

The αI domain was a single evolutionary insertion of a vWFA family domain into an ancient integrin gene present in the last common ancestor of *Olfactores*, the clade encompassing chordates and tunicates, which fundamentally changed the integrin family's ligand-binding capabilities. Integrins lacking this domain are restricted to recognizing ligands that contain a tripeptide RGD (Arg-Gly-Asp) or RGD-like motif that bridges the α and β subunits(106), whereas integrins with this αI domain recognize an exceptionally diverse array of structurally and sequentially unrelated ligands that bind directly to the αI domain. Some molecular features appear to be universally shared across all I domains: ligand binding generally propagates conformational changes through a structurally labile C-terminal region via the coordination of a conserved ion site (termed the “MIDAS”), ultimately leading to allosteric downstream integrin activation and subsequent cellular signaling (Fig. 3.1A) (118, 163, 172). These domains

generally adopt “open” conformations with high ligand affinity and mediate signaling, and “closed” conformations with little-to-no ligand binding that prevent signaling. I domains diverge, however, at the tuning of their ligand specificity: some I domain-containing integrins, such as α L, demonstrate high specificity for a single or few cognate protein ligands, while others like α M are extremely promiscuous in their ligand binding profiles despite sharing a recent common ancestor with α L(102, 104). The evolutionary and molecular underpinnings of this dichotomous behavior, as well as the rules that govern ligand binding by promiscuous integrins, are largely uncharacterized.

To understand how this receptor family has diverged to tune these unique ligand recognition profiles, we generated and compared fitness landscapes of ligand binding of two integrins expressed in the vertebrate immune system: α L β 2 (LFA-1, CD11a/CD18), which binds primarily to the cellular adhesion molecule ICAM-1(190), and α M β 2 (CR3, Mac-1, CD11b/CD18), which has a diverse array of proposed innate and pathogenic ligands that modulate immune behavior(103). We used phylogenetics to define the distinct evolutionary trajectories of these two receptors, site-saturating mutagenesis to determine the mutational constraints uniquely imposed by their respective ligands, and cryoEM analyses to uncover noncanonical ligand-binding pathways that rationalize M2’s marked promiscuity. Beyond identifying fundamental features that distinguish these two receptor-ligand profiles, our mutational and structural analyses implicate new pathways for therapeutic action in these promising yet pleiotropic immune adjudicators.

3.3 Results

3.3.1 *Distinct evolutionary trajectories demarcate integrin behavior*

Chordate I domain-containing alpha integrins fall into two subcategories: the collagen-binding alphas α 1, α 2, α 10 and α 11, and the leukocyte-specific subunits α E, α L, α M, α D and α X (Fig. 3.1B)(279). Among these, the collagen-binding integrins as well as α E and α L are thought to have high specificity, while α M, α D and α X are thought to bind many different ligands. To understand the selective constraints that shape these selective and promiscuous integrin behaviors, we performed a phylogenetic analysis of representative I domain-containing integrins across chordates (Fig. 3.1C). We found that the clades encompassing the collagen-binding integrin subunits as well as the genes encoding α E and α L are generally monophyletic and conserved across vertebrates, concordant with the accepted species tree. In contrast, the genes encoding α M, α D and α X are unique to placental mammals, making them much younger than previously accepted. Interestingly, each major vertebrate clade has unique duplications within this α M/D/X group, suggesting multiple independent rapid expansions and contractions of this gene set (Fig. 3.1D). Consistently, we found evidence of duplication within this subclade even in placental mammals (Fig. S3.1A). Unlike most other alpha integrins, the α M, α D, and α X genes are arranged in a tandem array in the mammalian genome, providing a simple genetic recombination mechanism that likely drives this rapid genetic turnover (Fig. S3.1B).

We hypothesized that selective pressure may be driving the distinct trajectories of these two integrin gene classes. We used maximum likelihood methods (PAML) to test for the presence of either positive (diversifying) or negative (purifying) selection across these genes (Fig. 1E). Consistent with their conservation among vertebrates, we found that most genes and codons within genes of the specific integrins evolve under purifying selection in primates (4/6).

In contrast, we found significant evidence of gene-wide and site-specific positive selection in the primate α D/M/X group (2/3). We found a similar pattern of diversifying selective pressures in rodent and fish genes from this clade (Fig. S3.1C), suggesting pervasive and continued pressure to maintain a small population of promiscuous α M/D/X-like genes in all vertebrates. Furthermore, sites of positive selection were concentrated in the ligand-binding I domain of the α D/M/X clade, consistent with ligand binding being the primary selective pressure on these cell surface receptor genes (Fig. 1F). Our findings are consistent with previous work on α M evolution(211) and contrast with the specific integrins, in which we find positive selection largely outside of the primary ligand-binding I domain regions. This genetic turnover and rapid evolution within α D/M/X appear to be primarily driven by the α subunit, as β integrin subunits are generally concordant across the vertebrate tree (Fig. S3.1D). Collectively, these phylogenetic findings draw a distinct correlation between the ligand-binding behaviors of integrin I domains and their evolutionary trajectories.

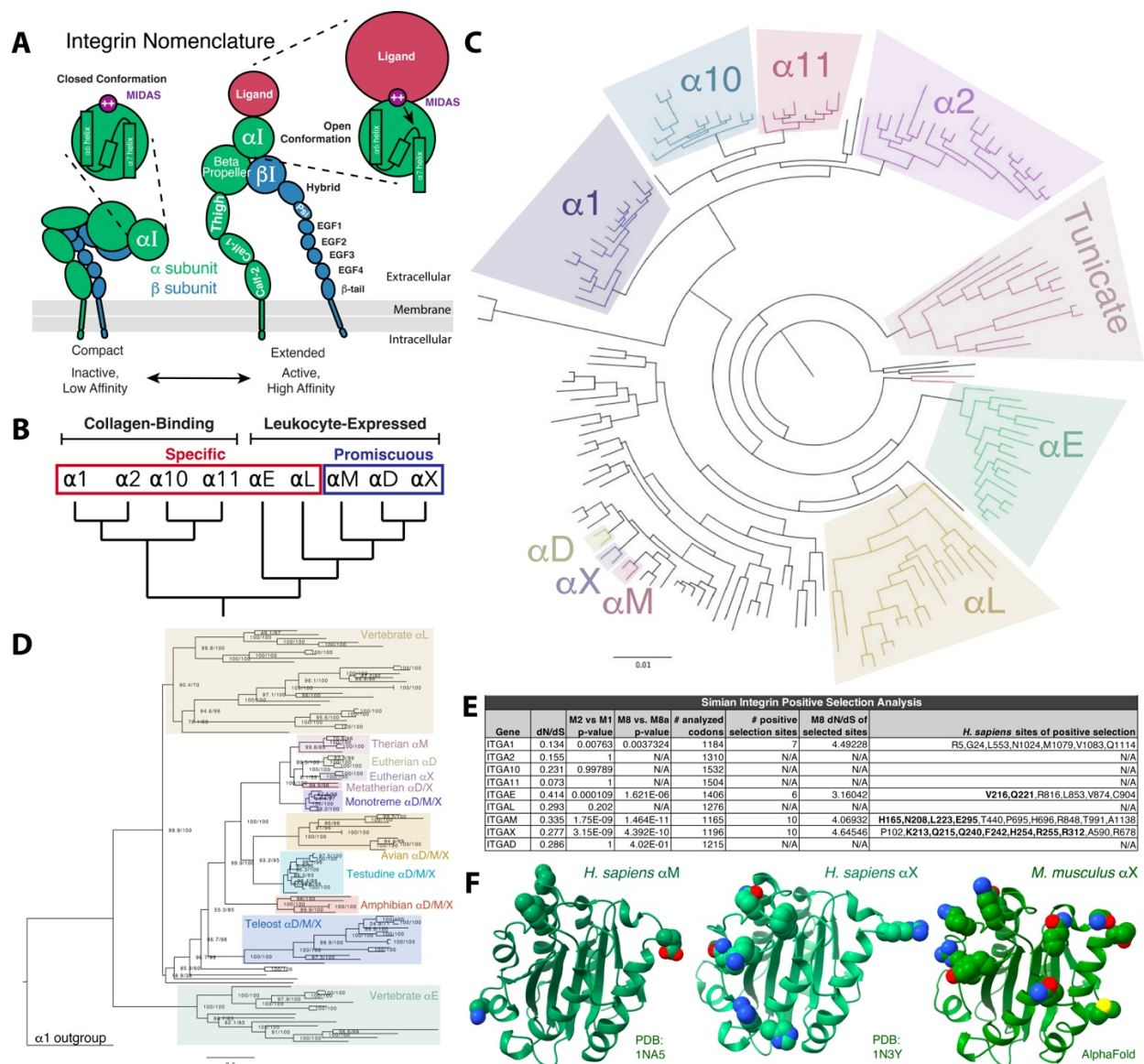


Figure 3.1. Diverging evolutionary trajectories of the integrin α I domain. A. A schematic of the activation

mechanism by which the α I domain allosterically links ligand binding to integrin activation. **B.** Vertebrate I domain-containing integrin α subunits fall into two biochemical categories: most are associated with a single or very few cognate ligands (red), while a subset of leukocyte-expressed subunits have more promiscuous binding activity (blue). **C.** A phylogenetic tree of chordate I domain-containing alpha subunit genes shows α D, α M, and α X are evolutionarily young genes compared to those with high ligand specificity. **D.** A phylogenetic analysis of representative leukocyte I domain-containing integrins shows repeated clade-specific expansions of the α D/M/X clade. **E.** A table summary of positive selection analysis on primate α I domain-containing integrin genes; bolded residues are contained within the I domain. **F.** An atomic structural representation of the residues under positive selection on the primate α M (left), α X (center), and murine α X (right) shows surface-exposed residues largely around the top ligand binding region are under pervasive adaptive pressure.

3.3.2 *The specific integrin α L β 2 occupies a fragile mutational niche*

To understand how these receptors biochemically generate distinct ligand-binding profiles, we adapted a yeast display platform(354) to perform site-saturating mutagenesis(248) of the ligand-binding I domain coupled to ligand titration experiments (Fig. S3.2A). This “TiteSeq” workflow, used commonly in antibody(355) and virus-receptor studies(356), allows us to generate a complete mutational landscape of how each possible amino acid substitution affects I domain-ligand binding affinity by FACS sorting and sequencing yeast based on their relative ligand binding as a measure of fluorescence. We generated duplicate, complete site-saturating libraries (“deep mutational scans”) for both integrins α L and α M (Fig. S3.2B). As integrins are conformationally dynamic and, as a result, heterogeneous in affinity assays(357), we stabilized these domains in ligand-binding open conformations with previously characterized disulfide bonds(358, 359) (Fig. S3.2C). Domains expressed well after induction in isogenic control yeast lines (Fig. S3.2D).

We began by performing the TiteSeq assay with the highly specific integrin α L and its primary cognate ligand ICAM-1. This protein-protein interaction is critical for both leukocyte extravasation and stabilizing the T cell synapse. Thus, its evolutionary conservation reflects its key role in the immune system. We generated a normalized genotype-phenotype map for integrin-ligand binding effects for each possible single missense substitution in the α L I domain, which we present here as a heatmap (Fig. 3.2A, Fig. S3.2E-F). We find a few trends that lend validity to this dataset: disruption of the disulfide bond at either cysteine universally decreases integrin-ligand affinity, mutations to buried secondary structure regions are largely deleterious, consistent with protein expression defects predicted by AlphaMissense(360) (Figure S3.2G), and the data effectively recapitulate previously characterized gain-of-affinity point mutations(189).

We first analyzed the mutational effects at the direct protein-protein interface between α L and ICAM-1 by mapping them onto the published crystal structure(173) (Fig. 3.2B). We expected that the conservation of this interaction, along with the primarily hydrophobic interactions that mediate its specificity, would collectively lead to an entrenched interface that is largely intolerant to mutation. Indeed, among residues involved in direct protein contact, we found very few paths toward increased affinity (Fig. 3.2C). Instead, most mutations markedly reduce integrin-ligand affinity; this is particularly true for mutations at the conserved ion coordination site (MIDAS), which triggers conformational changes in the I domain upon completion of the ion sphere via ligand binding(26), and is thus essential for integrin signaling. In the hydrophobic ring surrounding the MIDAS, even conservative missense substitutions in L204 and L205 have deleterious effects on ligand binding; these mutational trends are reflected in sequence conservation within the α L clade (Fig. 3.2D). Of the more tolerant interface

residues, natural variation in M140 and H264 reflects neutral or higher-affinity substitutions based on our mutational data. These data collectively suggest selective pressures have sculpted an interface that is largely at a high-affinity optimum and reflects a “lock-and-key” interaction.

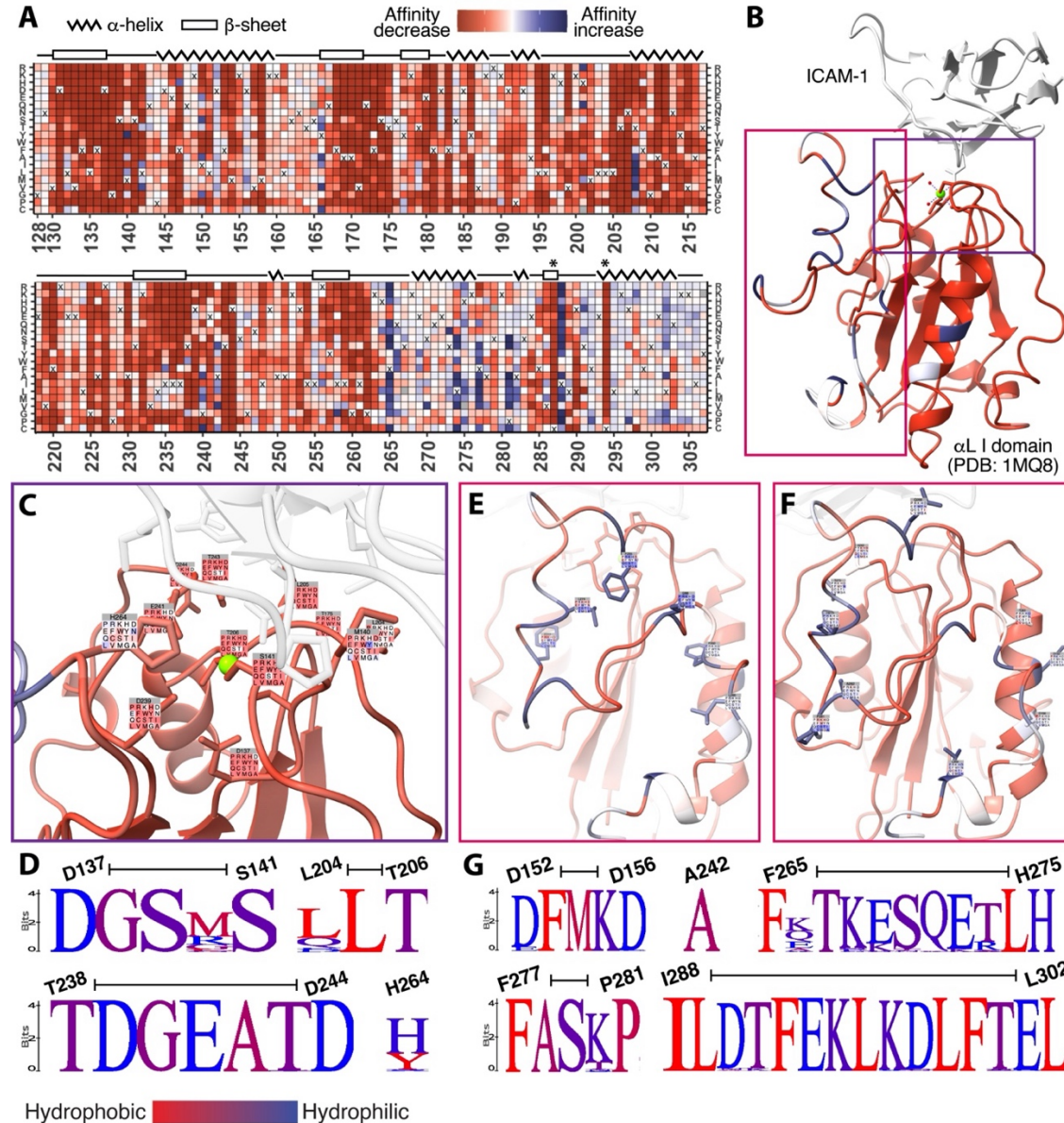


Figure 3.2. The α L:ICAM-1 interaction is mutationally fragile. **A.** A heatmap of the normalized effect of all single nonsynonymous mutations in the α L I domain on binding affinity to ICAM-1. Wildtype residues are indicated with an “x”; red indicates a decrease in affinity and blue an increase in affinity. **B.** The average effects of all mutations mapped to the structure of the α L I domain in complex with ICAM-1. Boxes indicate regions of interest for panel C (purple) and E-F (magenta). **C.** The residues at the α L:ICAM-1 interface are largely intolerant to mutation while maintaining ligand affinity, indicative of a surface optimized for a high affinity interaction. **D.** Logo plots representing the conservation of key residues at the α L:ICAM-1 interface. Residues are largely strictly conserved, and those that aren’t have natural substitutions with relatively benign mutational effects. **E.** Mutations in residues within the ratchet region involved in allosteric signaling are almost universally gain-of-affinity. **F.** Residues beyond the ratchet hydrophobics show similar mutational constraints, indicative of peripheral involvement in allosteric communication. **G.** Logo plots show residues involved in allosteric communication are near strictly conserved in mammals, similar to those at the α L:ICAM-1 interface.

In contrast, we found that the C-terminus of the α L I domain contained a high concentration of neutral and affinity-increasing mutations (Fig. 3.2A). Structurally, this region contains the machinery that links ligand binding to integrin activation via a conformational allosteric relay mediated by adjacent bulky hydrophobic residues in response to MIDAS coordination(153). When we mapped mutational effects onto this region, we found that mutations to these bulky residues universally increased affinity dramatically (Fig. 3.2). Since our TiteSeq analyses assay ligand binding affinity in an open conformation, these findings suggest that these residues have been specifically optimized to allow switching between “open” and “closed” I-domain states. By absolving the need to switch between these states in our affinity-binding assay, not only is high affinity possible; it is inevitable through mutations at certain sites. We also find sites outside of these hydrophobic residues following a similar mutational trajectory (Fig. 3.2F), revealing a larger interaction network mediating the complex conformational positioning of the I domain’s C-terminus. Some of these sites, such as Q266 and A242, are within a single residue of MIDAS and protein-protein interface residues, acting as physical links between ligand binding and allosteric activation. We find that these allosteric sites are equally conserved as those at the protein-protein interface (Fig. 3.2G), implicating selective pressure to maintain a balance between a high-affinity interface and a switchable allosteric signal precariously poised in the low-affinity state through a coordination of nearly half of the I domain. Collectively, these data define a fragile, narrow landscape maintaining the binding and likely the resulting function of the specific α L/ICAM-1 interaction.

3.3.3 *Molecular tradeoffs permeate the promiscuous integrin α M β 2*

We next utilized TiteSeq to determine if the mutational constraints determined in the α L I domain are shared by the promiscuous integrin α M. α M β 2 is a highly pleiotropic immune receptor, contextually either serving as an immune activator or immune suppressor involved in pathogen recognition, immune cell migration, inflammation, and phagocytosis and binding to an equally diverse array of ligands(103, 104) (Fig. S3.3A). The first described ligand for α M β 2 was the complement opsonin iC3b(361), whose C3d fragment (“TED” domain) binds and activates the α M I domain via the MIDAS(164). Structurally, C3d interacts with a similar region of the α M I domain as the ICAM-1- α L interaction. Therefore, we hypothesized this comparison would reveal both shared principles and divergent binding landscapes between these two integrin classes. We performed TiteSeq on the α M I domain with C3d, generating a comprehensive genotype-phenotype map of the interaction (Fig. 3.3A).

As with the α L I domain, we found a few trends in the data that support the approach. Secondary structure showed the most mutational constraint in buried residues, which were largely intolerant to most mutations. These data are consistent with the AlphaMissense predictions for the α M I domain (Fig. S3.3B) and reflect deleterious effects on protein expression. The stabilizing disulfide was also intolerant to mutation, confirming C3d's preference for the open, or active, I domain conformation. Interestingly, a few mutations that are relatively benign in these regions, such as I139F and V264I, are the wild-type residue at the homologous site of the α L I domain (Fig. 3.2A), suggesting epistatic pathways allow for mutational tolerance within buried structure. This explains why buried structural residues are not completely conserved in all I domains (Fig. S3.3C) despite the high fitness cost we observe in our data.

We hypothesized that rapid evolution and ligand promiscuity would lead to a less optimized protein-protein interface in the promiscuous α M compared to α L. We plotted the

average mutational effects across the I domain onto the published co-crystal structure with C3d(191), looking particularly for gain-of-affinity variants at the contact surface (Fig. 3.3B). Surprisingly, we found that this surface was at a relatively high-affinity optimum, although compared to α L/ICAM-1, there was a higher proportion of neutral or gain-of-affinity mutations. This strongly reflects the high fitness advantage of opsonization, an innate response that promotes both pathogen and debris clearance. As with α L, we also found that the allosteric switch region of α M contains bulky aromatic residues which appear to occupy a fitness minimum; some of these sites are strictly conserved with α L, while others are unique to α M (Fig. 3.3C). This further highlights the influence of epistatic effects that permit mutation while still maintaining a functional allosteric machine. Thus, the dual selective demands of high-affinity interfaces with low-affinity allosteric states appear to be a uniting theme for endogenous integrin-ligand complexes (Fig. S3.3D). Our findings are consistent with structural work on integrin activation mechanisms(118, 153, 173), and we find that these regions are the most conserved on the surface of all vertebrate I domains (Fig. 3.3D).

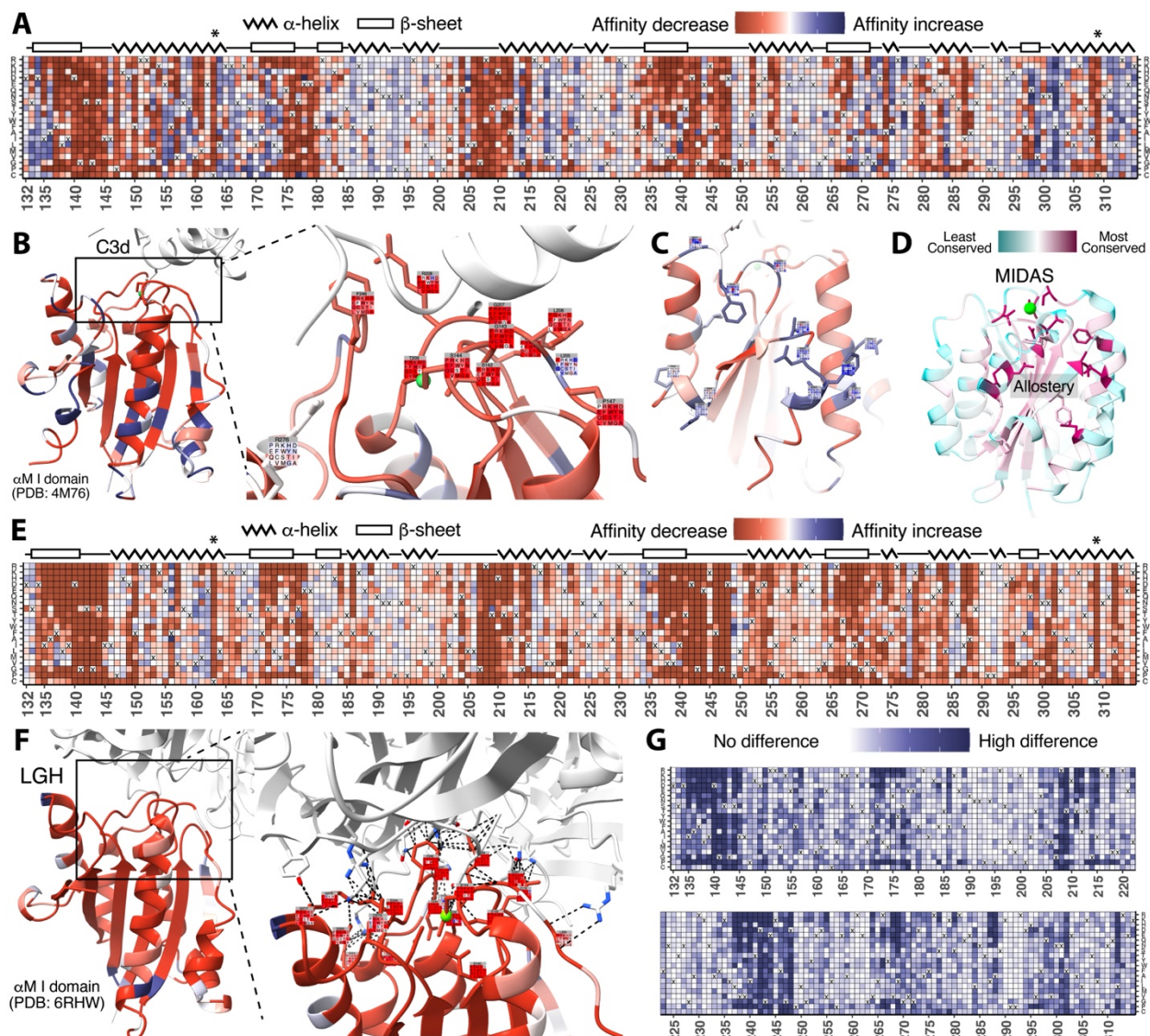


Figure 3.3. Molecular tradeoffs permeate the promiscuous α M I domain. A. A heatmap of the normalized effect

of all single nonsynonymous mutations in the α M I domain on binding affinity to C3d. Wildtype residues are indicated with an “x”; red indicates a decrease in affinity and blue an increase in affinity. **B.** The average effects of all mutations mapped to the structure of the α M I domain in complex with C3d. The left shows the entire I domain, the right focuses on the ligand binding top region and MIDAS. **C.** Like that of α L:ICAM-1, mutations in residues within the ratchet region involved in allosteric signaling are strongly gain-of-affinity. **D.** The conservation of residues in the I domain across vertebrate genes mapped to the structure of the closed α M I domain (PDB: 1NA5). Residues are most conserved in the MIDAS, hydrophobic core, and allosteric switch, while peripheral mutable residues contribute to specificity. **E.** A heatmap of the normalized effect of all single nonsynonymous mutations in the α M I domain on binding affinity to LGH. **F.** Same as (B) with the structure of the α M I domain in complex with LGH. Charged interactions contribute greatly to this interface. **G.** A difference map comparing the relative effects of each mutation between the C3d and LGH mutational backgrounds. Differences are notable in buried residues, some allosteric controls, and charged peripheral residues.

If molecular tradeoffs exist in this promiscuous receptor, we would expect gain-of-affinity variants for some ligands to be loss-of-affinity variants for others (362). To test this, we performed TiteSeq with α M and the well-characterized pathogenic *S. aureus* pore-forming protein Leukocidin GH (LGH) (Fig. 3.3E, S3.4A-B). α M β 2 serves as a surface receptor for this virulence factor, allowing the pore complex to approach the cell membrane and cause lytic rupture(363) (Fig. S3.4C). Previous work has shown species-specific differences in the α M I domain contribute to functional differences in receptor-ligand binding(211), yet the pathways to escape virulence while maintaining functional endogenous interactions remain unknown.

In our TiteSeq dataset, we found many intolerant mutations are explained by the α M/LGH crystal structure (204). Unlike that of α L/ICAM-1, the α M/LGH molecular interface is comprised primarily of charge-based interactions. Accordingly, we see that charge losses and reversals are detrimental to residues in the α M/LGH network (Fig. 3.3F). Many of these deleterious effects are shared with the α M/C3d mutational landscape, suggesting shared dependency of this region. These findings are consistent with the hypothesis that extreme surface charge contributes to α M's promiscuity(364). Although the effects of missense substitutions at the individual amino acid level are largely inseparable between these two ligands, a ‘difference map’ revealed pervasive differential effects of these two ligands (Fig. 3.3G). We see marked differences in mutational tolerance in binding regions unique to each respective ligand, thus defining separation-of-function regions. We also see the mutational tradeoffs are widespread, most commonly appearing in peripheral regions of this top I domain interface (Fig. S3.4D). This is consistent with charge reversals in regions of rapid evolution we and others(211) identify, and indicates the conserved core of the MIDAS region, essential for ligand binding and integrin activation, can be supplemented by either hydrophobic residues that generate specificity (for more specific receptors like α L) or charged residues which allow evolutionary tinkering of promiscuous regions (for more promiscuous receptors like α M).

Outside of the direct interface, we found many shared loss-of-affinity variants in secondary structure regions, consistent with these mutations acting universally to disrupt protein stability. We also found that, like C3d, disrupting the open conformation-stabilizing disulfide bridge invariably disrupts the α M/LGH interaction, confirming LGH's preference for an open conformation(204) (Fig. 3.3E). However, mutations to allosteric activators did not universally increase affinity for LGH in the same way they do for C3d; instead, mutations at these sites had either varying or neutral effects on LGH affinity (Fig. 3.3G). These differences support the model of LGH's proposed mechanism of action, which depends on it first binding to the high-affinity α M I domain then being trafficked to the cell surface via α M β 2 switching to a low-

affinity state where it is released(204) (Fig. S3.4C), whereas C3d is most effect exclusively in the high-affinity α M conformation. Beyond direct protein-protein interactions, molecular tradeoffs shape pathogen recognition and competing interests from the allosteric machinery necessary for α M β 2 function. Collectively, these data identify a shared molecular partitioning of allosteric and contact interfaces among all I domains, yet reveal the paths by which separation-of-function is executed across the promiscuous I domain surface.

3.3.4 Pathogenic NIF co-opts an essential α M ion coordination site

LGH is not the only pathogenic ligand for α M. The hookworm protein Neutrophil Inhibitory Factor (NIF) suppresses host immunity by binding to the I domain to evade detection(206, 207, 365) (Fig. S3.5C). Like many I-domain ligands, including those described I this study, this interaction depends on the MIDAS. Unlike many other I-domain ligands, however, previous work suggests this interaction is independent of I-domain conformation, as NIF binds equally well under both I-domain-activating (Mn^{2+}) and suppressing (Ca^{2+}) cation conditions (206). It is therefore likely that NIF uses biochemically distinct mechanisms to co-opt the essential MIDAS machinery.

To understand how NIF binds to the α M I domain in this unique way, we again performed TiteSeq on the α M I domain but this time with NIF as the ligand (Fig. 3.4A, S3.5A-B). As with previous datasets, we find that buried secondary structure is largely mutationally intolerant. In this set, however, we were surprised to see that disrupting the stabilizing disulfide bridge, as well as the bulky aromatic residues involved in allosteric relay, had anti-correlative effects compared to those of C3d. This suggests that NIF may somehow prefer the closed I domain conformation, despite depending on the presence of the essential MIDAS cation that normally stabilizes the open conformation upon ligand binding. Indeed, compared to C3d, the mutational landscapes of the allosteric region are essentially opposite, most notably the open-conformation-stabilizing disulfide bridge and the bulky hydrophobic switching residues (Fig. 3.4B). This finding was surprising, given that I-domain ligands are almost universally thought to induce open, active conformations (366).

To both contextualize our mutational data and understand how NIF may favor a closed conformation, we performed single-particle electron microscopy analyses of the α M β 2 ectodomain/NIF complex in buffers with varying ions. Integrin heterodimers occupy a spectrum of structural states, ranging from a compact closed conformation, generally thought to have low ligand affinity, to an open conformation with high ligand affinity. We found that, unlike all I-domain integrins analyzed so far, NIF bound equally well to integrins in all conformations in negative-stain averages (Fig. S3.6A). We next performed cryoEM to generate high- resolution information for the α M β 2/NIF interaction (Fig. S3.6B-C). In buffer conditions favoring the active integrin conformation, we determined an open integrin structure to 3.5Å resolution with apparent density for the I domain and NIF (Fig. 3.4C), although this map was anisotropic in resolution and lacked high-resolution features of the NIF epitope, likely due to the I domain's extensive flexibility relative to the rest of the ectodomain(217). We bound the complex to the I domain antibody LM2/1 to increase the mass of the binding region, providing the necessary density to achieve high resolution (Fig. S3.6D-G). We also generated a structure in a more physiological buffer that would allow occupation of open or closed I domain states to resolve NIF's preferred binding modality (Fig. 3.4D-E).

Comparing these two structures, we found that many of the NIF epitope features are shared. Like LGH and C3d, NIF binds to the top region of the I domain, sharing many overlapping contact residues with these other structurally characterized ligands. Our mutational constraint data agree with this structure, and many of the mutations deleterious for ligand binding in other contexts hold true for NIF (Fig. 3.4F), while mutations that are neutral in C3d context, such as for R151 and K200, are deleterious for NIF binding. Most notably, NIF contains an ion-coordinating acidic residue that occupies a virtually identical position in both structures (Fig. 3.4G). Despite this, in the more physiological buffer conditions, the I-domain clearly occupies a closed conformation (Fig. 3.4E). This was unexpected, as it is generally agreed that MIDAS ion coordination by a ligand drives the open I-domain conformation, leading to downstream integrin activation. This structural finding agrees with our mutational data, confirming NIF as an exceptional ion-coordinating, integrin-inactivating ligand.

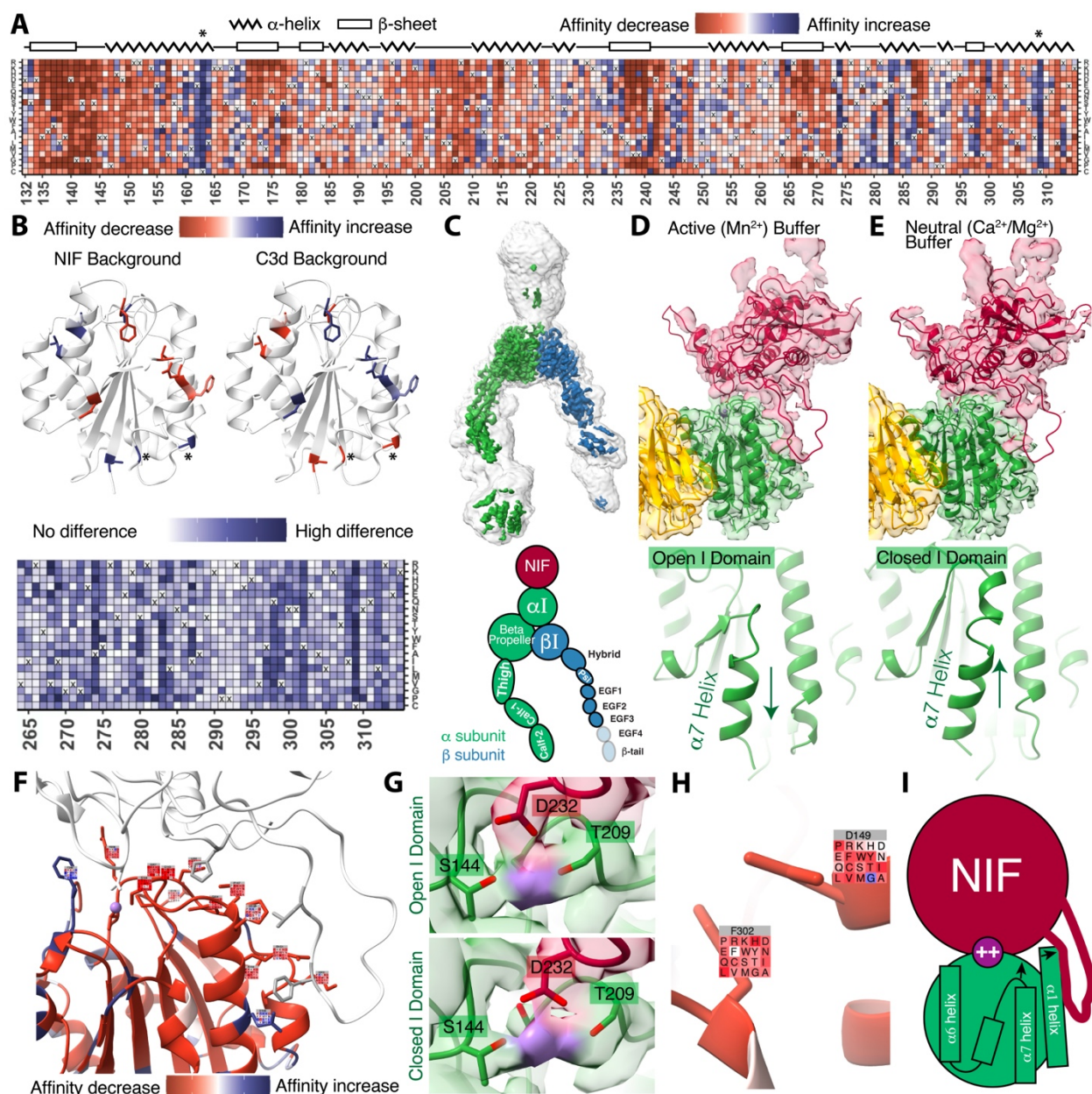


Figure 3.4. NIF closes the α M I domain. **A.** A heatmap of the normalized effect of all single nonsynonymous mutations in the α M I domain on binding affinity to NIF. Wildtype residues are indicated with an “x”; red indicates a decrease in affinity and blue an increase in affinity. **B.** Mutations in residues involved in allosteric signaling have opposite average effects in the NIF (top left) and C3d (top right) backgrounds, indicating a molecular tradeoff in allostery. The difference map (lower) between the two backgrounds indicates these tradeoffs are pervasive in the I domain C-terminus. **C.** The cryoEM structure of α M β 2 bound to NIF shows NIF interacts at the canonical ligand binding region on the α M I domain. **D-E.** Structural models of NIF bound to the I domain in either activating (D) or physiological (E) buffers show NIF is capable of interacting with the I domain in a conformation-agnostic way. **F.** The average mutational effect per residue mapped to the structure of α M β 2:NIF suggests the extra loop cascading down the side of the α M I domain is a critical determinant of ligand binding. **G.** CryoEM data shows NIF coordinates the α M MIDAS in both open (top) and closed (bottom) conformations. **H.** The D149G mutation increases NIF affinity by creating more space for F302 to occupy the closed conformation without affecting hydrophobic interactions. **I.** A model for NIF’s immune evasion. The extension loop inactivates the I domain via interactions with the α 1 helix, favoring the α 7 helix to occupy a closed state.

How does NIF close the α M I domain? Unlike C3d, LGH, and all other structurally characterized I domain ligands so far, NIF contains an unstructured loop which ladders down the side of the I domain, wedging tightly between the α 1 and α 3 helices (Fig. 3.4F). Although conformational shifts occur most dramatically in the α 7 helix to drive I domain opening, this movement depends on reciprocal interactions between α 7 and α 1, which ultimately pull α 1 away from α 3(367). NIF’s unstructured loop disfavors this interaction by serving as a glue to hold α 1 and α 3 together, thus preventing stabilization of the open I domain conformation (Fig. 3.4E). Concordant with this, we find that creating more space in the pocket for F302 of the closed α 7 helix with a D149G mutation increases NIF affinity (Fig. 3.4H); differing mutational constraints at the α 1 helix between NIF and C3d further support this model (Figs. 3.3A, 3.4A). NIF is thus an enigmatic I domain ligand, eschewing rules of integrin activation in a way that both sterically blocks endogenous ligands and favors an inactive integrin conformation, ultimately leading to hookworm’s successful immune evasion.

3.3.5 Noncanonical ligand binding by iC3b to α M β 2 reveals pathways to promiscuity

Our data so far suggest the odds are stacked against the α M I domain, as pathogens take advantage of necessary endogenous processes for personal gain and pathways of escape are limited by α M β 2’s demanding ligand promiscuity. How can this receptor stay competitive, especially given the lack of pathways to increased affinity for C3d? To understand this, we returned to our evolutionary results, which identified regions outside of the I domain under positive selection and possibly important for ligand binding (Fig. 3.1E). This is consistent with prior work implicating non-I regions as essential for α M β 2 function(210, 224, 368). Most notably, C3d is only a degradation fragment of the larger complement opsonin iC3b(369), the first described ligand for α M β 2 and the most essential for α M β 2-mediated phagocytosis. This makes iC3b a major pathway utilized by α M β 2 to best pathogenic invaders. Prior work suggests most of iC3b binds outside of the I domain and dramatically increases its affinity for α M β 2 compared to C3d alone(370, 371), yet the molecular underpinnings of this interaction remain unknown (Fig. S3.8A).

As with prior studies(164, 223), we attempted to perform cryoEM analysis on the α M β 2/iC3b complex but were limited by α M β 2’s strong preference to homodimerize (a dimer of α M β 2 dimers) under activating ion conditions, preventing iC3b binding (Fig. S3.7A). To

circumvent this hurdle, we attempted to reconstruct the α M β 2 dimer structure and identify dimer-breaking mutations, but α M β 2 dimers adopted strong preferred orientations on EM grids, preventing high-resolution structure determination (Fig. S3.7B). We found that the addition of LM2/1 led to an increased range of occupied orientations, leading to a more isotropic map with sufficient information to generate an α M β 2 homodimer model (Fig. S3.7C-F). This structure shows the hybrid and psi domains of β 2' occupy the canonical ligand binding region at the top of the α M I domain, with β 2' D54 coordinating the MIDAS (Fig. S3.7G). Mutating this residue (β 2 D54A) was sufficient to abolish α M β 2 dimerization, allowing structural characterization with iC3b (Fig. S7H).

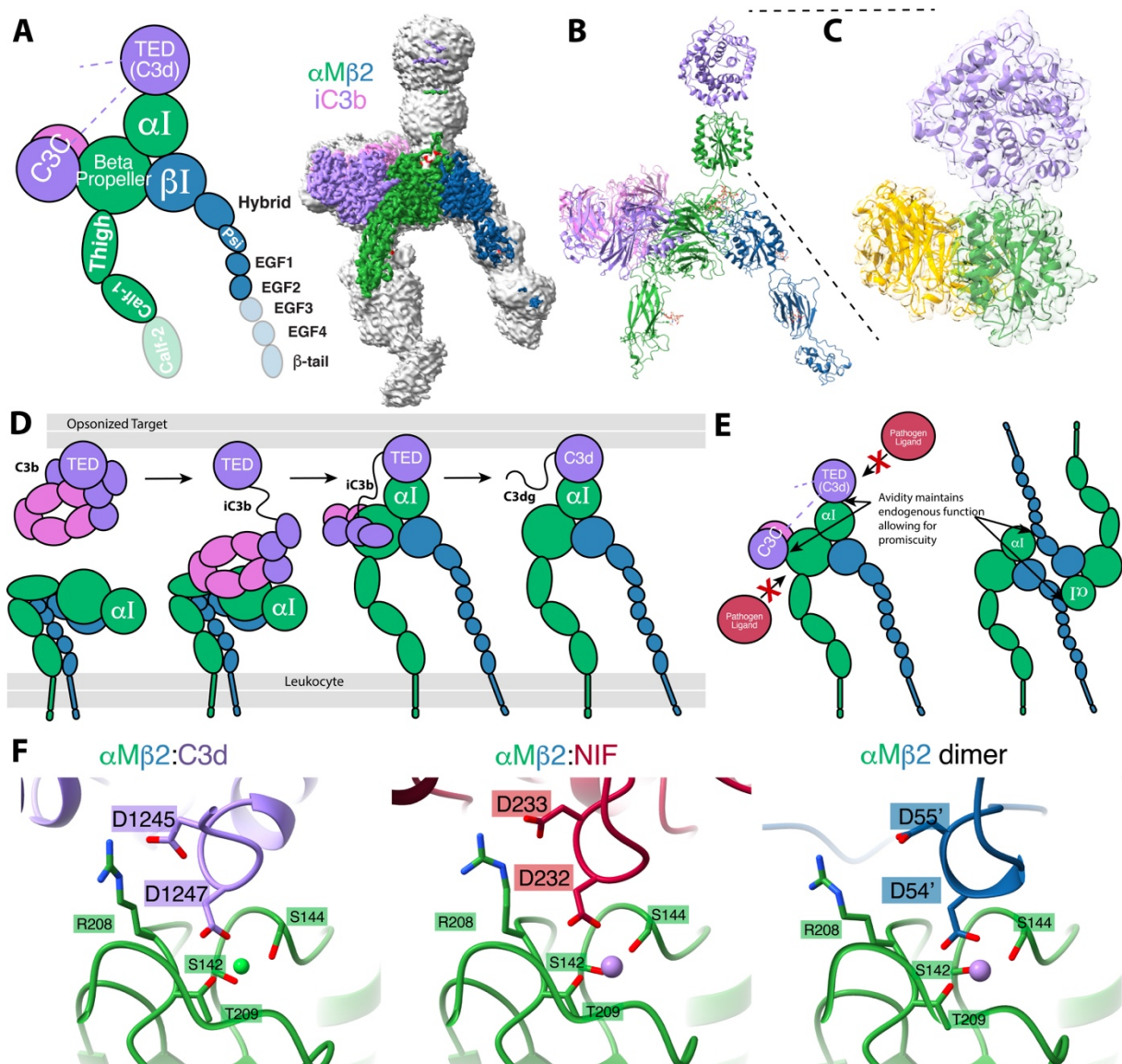


Figure 3.5. Avidity generates novelty in ligand binding by α M β 2. **A.** The cryoEM structure of the α M β 2 D54A mutant bound to iC3b. The C3c region of iC3b wraps tightly around the beta-propeller region of α M. **B.** The molecular model of α M β 2:iC3b in the open, active conformation. **C.** The focused cryoEM structure of the α M I domain bound to C3d and LM2/1 antibody. **D.** A molecular model by which avidity allows for stepwise binding and activation of α M β 2 to opsonized targets. The C3c region

provides a high-affinity surface to bring $\alpha M\beta 2$ close to the target, allowing for subsequent MIDAS-based activation by the TED/C3d region. This activation can be maintained following cleavage into C3dg. **E.** A model for how endogenous ligands may outcompete pathogenic modulators through avidity. **F.** A shared molecular language for ligand binding by the promiscuous αM I domain.

We determined the $\alpha M\beta 2/iC3b$ structure to 3.6Å (Figs. 3.5A, S3.8B-C), although as with NIF this structure was anisotropic in resolution and insufficient to model the I domain region. We took a similar approach as with NIF to increase mass in this region with the C3d fragment (Fig. S3.8D-E), allowing us to create a composite model from high-resolution data for nearly the entirety of iC3b (Fig. 3.5B). Surprisingly, we found that C3d (or the “TED” domain when still attached in iC3b) is the only region of iC3b that binds to the αM I domain; the rest of the protein (the “C3c” region) tightly hugs the αM β -propeller region. This structure further reveals the rules governing M2’s specificity for iC3b. iC3b is a fragment of the complement degradation cascade, precluded by C3b, which rigidly confines the TED domain to the rest of the molecule. Upon proteolytic cleavage of C3b, TED becomes flexible yet still covalently bound to C3c, allowing the dual occupation of both the β propeller region and I domain we see in our $\alpha M\beta 2/iC3b$ structure (Fig. 3.5D). The complex, specific niche this interaction occupies within this cascade adds insurmountable complexity for pathogen mimicry, providing an avidity-based molecular pathway beyond canonical ligand binding that maintains immune effectuation despite M2’s rapid evolution (Fig. 3.5E). In support of this, iC3b’s position also sterically competes with the *Bordetella* adenylate cyclase toxin(217), which binds to the side of $\alpha M\beta 2$ yet cannot extend to the I domain (Fig. S3.8F). This avidity-based model is further consistent with data implicating multiple points of contact for another better-characterized endogenous $\alpha M\beta 2$ ligand, fibrinogen. Our cryoEM data, together with previous work, also reveal an emerging structural motif shared by these remarkably different ligands for the αM I domain (Fig. 3.5F).

3.4 Discussion

Here, we combine phylogenetics, mutational scanning, and structural biology to probe how integrins divide ligand-binding labor. Beyond determining ligand-bound integrin structures that have eluded the field for over two decades, the high-resolution genotype-phenotype datasets we generate provide molecular roadmaps for understanding how a small allosteric machine tunes ligand-binding niches. These maps include targetable separation-of-function regions for integrin therapeutics, which often fail because broadly inhibiting or activating pleiotropic receptors causes off-target effects (372, 373). These maps, combined with phylogenetic analyses, also provide the tools to comprehensively compare epistatic effects of mutations in orthologous domains that open evolutionary pathways for seemingly constrained residues.

Our results paint a highly dynamic portrait of a single ligand-binding domain, one that faces differing pressures to either entrench or innovate based on evolutionary and functional constraints. We find that, in the case of the exceptionally promiscuous receptor $\alpha M\beta 2$, ligand binding battles have extended far beyond the confines of canonical battlefields. In this receptor, avidity has opened otherwise closed pathways to ligand promiscuity. This contrasts heavily with the closely related $\alpha L\beta 2$, for which the ligand binding domain shows mutational trends consistent with an optimized interface that likely work to prevent the spread of such battles. Given that the diverging evolutionary tempos of these two receptors’ behaviors have both been consistent across vertebrate evolution, we expect it is beneficial for species to express both styles

of receptor, specific and promiscuous, despite the cost of pathogenic co-option. Understanding the ligand-binding profiles of ancestral intermediates may shed further light on how these two divergent strategies arose (374).

The structural and mutational data presented here provide only snapshots of these complex receptor-ligand interactions. We introduced stabilizing mutations necessary to generate high-resolution genotype-phenotype maps of α L β 2 and α M β 2's ligand-binding landscapes, yet our data suggest the I-domain's ability to dynamically fluctuate between open and closed states is critical for the immune balance these receptors must strike. Indeed, in *in vivo* models, introducing high- or low-affinity conformations dramatically disrupts this balance(375, 376). Likewise with our structural data, we differ buffer conditions to favor certain stabilized conformations of integrin-ligand interactions, and the stepwise intermediates that likely occur pervasively *in vivo* are not reflected(229). Future work probing these structural interactions in their native contexts using tomographic, force measurement, and high-resolution fluorescence approaches will shape our understanding of the dynamics that tip the molecular balance towards closed, inactive or open, active states.

3.5 Materials and methods

3.5.1 DNA constructs for mammalian and bacterial expression

The regions encoding the wildtype α M ectodomain (M1-N1105, used for integrin:NIF complex) followed by a C-terminal linker (GTGG), HRV 3C cut site (LEVLFQGP), acidic coil motif (AQCEKELQALEKENAQLWELQALEKELAQ) and Strep-Tag II (WSHPQFEK*) or the β 2 ectodomain (M1-N700) followed by a C-terminal linker (DTSG), HRV 3C cut site (LEVLFQGP), basic coil motif (AQCKKKLQALKKKNAQLKWKLQALKKKLAQ) and 6xHIS tag inserted into a pcDNA3.1-Hygro(-)-like backbone were synthesized commercially (GenScript). Coding sequences for the full length α M, α L and β 2 were generated by replacing the C-terminal ectodomain tags with G-blocks (IDT) encoding the respective transmembrane and C-terminal tail regions via standard cloning techniques. The α M β 2 ectodomain dimer mutations (β 2 D54A, D55A) used for structural analyses as well as the full length integrin point mutations (α M G143W, S144A, F246A, F275S and α L L205W, F265S, H275D, F292S, L302E) used in mutational scanning validation were generated by site-directed mutagenesis. The ICAM-1 ectodomain was previously published(118). A G-block (IDT) encoding the postprocessed *A. caninum* NIF ectodomain (N18-L274) was inserted into a pcDNA3.1-Hygro(-)-like backbone with an N-terminal CD33 signal sequence (GMPLLLLLPLWAGALA) and C-terminal linker (DTSG), HRV 3C cut site (LEVLFQGP) and 6xHIS tag using standard cloning techniques. The human C3 complement fragment C3d (G991-D1288) was synthesized as a bacterial codon-optimized G-block (IDT) and inserted into a modified pET vector with a C-terminal 6xHIS tag. Codon-optimized G-blocks encoding *S. aureus* leukocidin G with C-terminal 6xHIS tag and leukocidin H were inserted into a pETDUET plasmid using standard cloning techniques.

3.5.2 Protein expression and purification

Integrin α M β 2, NIF, and ICAM-1 ectodomains were expressed and purified as previously reported for α E β 7(118). Briefly, protein expression was performed according to the "Max Titer" manufacturer recommendations in ExpiCHO cells (ThermoFisher). A ratio of 3:2 α : β integrin

DNA was transfected for a total of 1 µg of DNA per mL of cell culture. For NIF and ICAM-1 a total of 1 µg of DNA per mL of cell culture was used for transfection. Cells were grown to a concentration of $7-10 \times 10^6$ cells/mL at 37°C, 8% CO₂, and 90% humidity then split to 6×10^6 cells/mL and transfected according to manufacturer recommendations. Following this, cells were fed according to manufacturer recommendations and shifted to 32°C for subsequent growth. 7-8 days post-transfection, cell supernatant was harvested via low-speed and medium-speed centrifugation, then diluted 2:1 into HisTrap Binding Buffer (20mM NaPO₄, 500mM NaCl, 20mM imidazole, pH 7.4). The clarified supernatant was flowed over a 5mL HisTrap FF Crude column (Cytiva) equilibrated in HisTrap Binding Buffer and eluted in HisTrap Elution Buffer (20mM NaPO₄, 500mM NaCl, 500mM imidazole, pH 7.4). Integrins were concentrated and incubated overnight in 1:10 (w/w) of 3C protease at 4°C. The next day the cleaved protein was further purified via preparative gel filtration chromatography in integrin storage buffer (20mM Tris-HCl pH 7.4, 150mM NaCl, 1mM MgCl₂, 1mM CaCl₂), concentrated, snap frozen in 10% v/v glycerol and stored at -80°C. ICAM-1 and NIF were further purified via preparative gel filtration chromatography in TBS (20mM Tris-HCl pH 7.4, 150mM NaCl), concentrated, and stored immediately following HisTrap purification.

C3d was expressed and purified as previously described(191) with slight modifications. Briefly, BL21 LOBSTR cells containing the expression plasmid were grown to an OD₆₀₀ of 0.6, induced with 0.1mM IPTG, and incubated with shaking overnight at 18°C. The following day cells were harvested at 4000xg for 30 minutes at 4°C and resuspended in 10mL/L of cell culture in HisTrap binding buffer (20mM NaPO₄, 500mM NaCl, 500mM imidazole, pH 7.4) with EDTA-free cOmplete protease inhibitor cocktail (Roche) and 1mM PMSF. The cells were lysed via sonication on ice for 3 minutes and the lysate was clarified for 30 minutes at 4°C, 30000xg. The lysate was flowed over a 5mL HisTrap FF Crude column (Cytiva) equilibrated in HisTrap binding buffer, washed with 5 CV of binding buffer, and eluted with HisTrap Elution Buffer. Peak fractions were concentrated and further purified via gel filtration chromatography on a Superdex 200 Increase 10/300 GL (Cytiva) in TBS (20mM Tris-HCl pH 7.4, 150mM NaCl), pooled, concentrated, and snap frozen in 10% v/v glycerol. LeukocidinGH was expressed and purified using the same protocol as C3d.

The LM2/1 antibody (eBioscience) and iC3b (CompTech) were purchased commercially and stored according to manufacturer instructions.

3.5.3 *nsEM sample preparation*

Wildtype αMβ2 ectodomain as well as the D54A and D54A/D55A variants were diluted to a final concentration of between 2 and 20 µg/mL in integrin active buffer (20mM Tris-HCl pH 7.4, 150mM NaCl, 1mM MnCl₂). For NIF complexes, the wildtype αMβ2 ectodomain was incubated at a 1:1.5 molar ratio with NIF in integrin storage buffer for 30 minutes prior to grid application. 3 µL of integrin was applied to a glow-discharged 400 mesh copper glider grid with a thin layer of formvar and carbon (Ted Pella). The grids were stained with a solution containing 2% (w/v) uranyl formate as previously described(315).

3.5.4 *nsEM data acquisition and processing*

Data were acquired using a Thermo Fisher Scientific Talos L120C transmission electron microscope operating at 120 kV and recorded on a 4k x 4k Thermo Scientific Ceta camera at a

nominal magnification of 92000x with a pixel size of 0.158 nm. Leginon(316) was used to collect α M β 2 wildtype, α M β 2 D54A, α M β 2 D54A/D55A, or α M β 2:NIF micrographs at a nominal range of 1.5-2.5 μ m under focus and a dose of approximately 50 e⁻/Å².

Similar processing pipelines were used for all negative stain datasets. Micrographs were processed using cryoSPARC(317). CTF estimation was done using Patch CTF. Initially particles were picked in a reference-free manner using blob picker in cryoSPARC. The particle picks were subjected to 1-3 rounds of particle curation and 1 round of 2D classification followed by 2D class selection for a final particles contributing to 2D classes. Three representative classes for each condition were chosen to represent described phenotypes.

3.5.5 cryoEM sample preparation

For the α M β 2:NIF ectodomain structure, wildtype α M β 2 protein with NIF in a 1:1.5 molar ratio was diluted in active buffer (20mM Tris-HCl pH 7.4, 150mM NaCl, 1mM MnCl₂) to a final concentration of 1mg/mL, then incubated for 30 minutes at room temperature prior to dilution in 0.05% (v/v) CHAPS detergent and grid application. For the I domain:NIF structures, wildtype α M β 2 protein with NIF and LM2/1 IgG in a 1:1.5:2 molar ratio was diluted in either active buffer (20mM Tris-HCl pH 7.4, 150mM NaCl, 1mM MnCl₂) or storage buffer (20mM Tris-HCl pH 7.4, 150mM NaCl, 1mM MgCl₂, 1mM MgCl₂) to a final concentration of 0.2mg/mL, then incubated for 30 minutes at room temperature prior to grid application. For the α M β 2 dimer structure, wildtype α M β 2 ectodomain was incubated with LM2/1 IgG in a 1:1.5 molar ratio for 30 minutes in active buffer at 0.2mg/mL for 30 minutes prior to grid application. For the iC3b structure, α M β 2 D54A was incubated with iC3b in a 1:1.5 molar ratio in HEPES active buffer, purified via size exclusion chromatography and concentrated to 1 mg/mL. CHAPS was added to 0.05% (v/v) immediately prior to grid application. For C3d complexes, α M β 2 D54A ectodomain with C3d and LM2/1 IgG in a 1:2:2 molar ratio was diluted in active buffer to 0.2mg/mL and incubated at room temperature for 45 minutes prior to grid application. Following dilution and complex formation steps, all samples were frozen with similar conditions. 3 μ L of protein was applied to Quantifoil grids (EMS) that were glow-discharged for 30s at 15 mA. The grids were blotted with a Vitrobot Mark IV (ThermoFisher) using a blot time of 3-7s and blot force of 4-5 at 100% humidity and 4°C. Grids were plunge-frozen in liquid ethane cooled by liquid nitrogen and stored in liquid nitrogen.

3.5.6 cryoEM data acquisition and processing

Data collection and processing were performed as previous. The details of datasets and processing pipelines are outlined in Table 3.1, Fig. S3.6 (α M β 2:NIF structures), Fig. S3.7 (α M β 2 dimer), and Fig. S3.8 (α M β 2:iC3b and α M β 2:C3d). All datasets were collected on a Glacios cryo-transmission electron microscope (ThermoFisher) operating at 200 kV and recorded with a Gatan K3 Direct Detection Camera. Automated data collection was carried out using SerialEM(320). One hundred-frame movies were recorded in super-resolution counting mode at a nominal magnification of 36000x corresponding to a calibrated super-resolution pixel size of 0.561 Å/px. Each dataset was collected with a nominal defocus range of 1.2-1.8 μ m under focus and a dose of approximately 50 e⁻/Å².

Dose fractionated super-resolution movies were motion-corrected and binned 2x2 by Fourier cropping using MotionCor2(321) within the RELION(322) wrapper. From there, motion-

corrected stacks were further processed in cryoSPARC according to each processing pipeline. Briefly, particles were picked using the blob picker, subjected to multiple rounds of 2D and/or 3D classification using *ab initio* and heterogeneous refinement using initial models generated from the data, and further refined with non-uniform and local refinement. Masks for local refinements were generated using UCSF ChimeraX(323) and cryoSPARC. Map sharpening was performed using either the CryoSPARC or Cosmic2(324) server wrapper of DeepEMhancer(325) with the high resolution setting. Local resolution estimation was performed in cryoSPARC. 3D FSC(326) plots were generated from <https://3dfsc.salk.edu/>. All map visualizations were generated and recorded using UCSF ChimeraX.

3.5.7 Model building

Initial models used for building and refinement are provided in Table S3.1. Models were built and modified in COOT(333), with glycans manually added where necessary using the carbohydrate module. Models were then iteratively fit and refined between ISOLDE and PHENIX(329) using both sharpened and unsharpened maps. A final real space refine job was done for all models. Molecular interactions were determined using the PISA server(336). Model visualizations were generated and recorded using UCSF ChimeraX. PDB identifiers are provided in Table S3.1.

3.5.8 Sequence alignments and phylogenetic analyses

All protein sequence alignments were generated using the default MUSCLE algorithm(351) and visualized using Geneious Prime. Alignments were further curated both automatically via ClipKIT(377) and manually to remove large gaps, sequences with low quality, and variable regions uninformative for phylogenetics. Sequence alignments are available as supplemental text files in FASTA format with the following names: Alignment of I domain-containing integrins across representative Olfactores in Fig. 3.1A (Data S1 in online publication), Alignment of vertebrate leukocyte I domain integrins in Fig. 3.1B (Data S2 in online publication), Alignment of vertebrate beta subunits in Fig. S3.1C (Data S3 in online publication). Sequence names are provided as protein names, NCBI accession numbers, Uniprot or ENSEMBL identifiers or Aniseed Transcript IDs.

All phylogenetic trees were generated via the IQ-Tree(352, 378) web server or local installation, allowing for free-rate heterogeneity and using optimized substitution models, and then visualized in FigTree v1.4.4 (<https://github.com/rambaut/figtree>). Values in tree nodes are presented as “1000 ultrafast bootstrap/SH-aLRT.” Phylogenies were constructed from alignments of the region encompassing the β -propellor, α I domain, and thigh for alpha subunits and the β I domain, hybrid and psi domains for beta subunits.

3.5.9 Positive selection analyses

Whole-gene dN/dS calculations and site-specific selection tests were performed using maximum-likelihood estimation with the codeml algorithm in PAML (308). Model 0 was used to calculate whole-gene dN/dS. Site-specific positive selection was calculated as previously described(379) for all tested genes. Briefly, models that either include (models 2, 8) or exclude (model 1, 8a) a category allowing dN/dS>1, indicating positive selection, were compared.

Codons with Bayes Empirical Bayes scores above 0.9 for model 8 were highlighted as evolving under positive selection. Codon alignments used for analysis can be found in Data S4 in the online publication.

3.5.10 Flow cytometry

To validate our deep mutational analysis in a full-length integrin mammalian system, we leveraged conformation-specific antibodies and ligand titration assays paired to flow cytometry. To determine the activation state of wildtype or mutant integrins, we paired a conformation-agnostic antibody for either α M (M1/70, eBioscience) or α L (TS2/4, Biolegend) as an expression control and a β 2 antibody specific for the active integrin conformation (m24, Biolegend). ExpiCHO cells were transfected with plasmids encoding wildtype or mutant integrin constructs in a 3:2 α : β ratio in a 6-well plate format according to manufacturer instructions for the “standard” expression. After reaching $7\text{--}10 \times 10^6$ cells/mL, cells were split to 6×10^6 cells in a 2mL volume per well, transfected, fed following a 24 hour incubation and harvested for flow cytometry 40–48 hours post transfection. For conformation assays 50 μ L per well of cell volume were distributed in a 96-well plate, pelleted at 500xg for 5 minutes, washed twice with room temperature TBS, resuspended in 100 μ L of storage buffer, and incubated for 30 minutes at 37°C with gentle rotation. Cells were then washed twice with ice-cold integrin storage buffer, resuspended in ice-cold storage buffer with 0.25 μ L of each antibody per well, and incubated for another hour with gentle rocking at 4°C. Cells were washed twice with ice-cold storage buffer, resuspended in 100 μ L of ice-cold storage buffer, and analyzed in the Fred Hutchinson Cancer Center Flow Cytometry shared resource facility.

A similar workflow was performed for ligand-binding assays, but instead of m24 antibody an anti-HIS antibody (Genscript) was used to detect bound ligand. Cells were transfected, grown, and harvested in the same manner as the conformation test. Following plating and TBS washes, cells were resuspended in 100 μ L of room temperature integrin active buffer, incubated for 30 minutes at 37°C with gentle rotation, then transferred to a plate containing a 10-fold dilution series of respective ligand (ICAM-1 for α L β 2 and C3d, NIF, or LGH for α M β 2) in active buffer. Following a 1 hour 37°C incubation, cells were washed twice with ice-cold integrin active buffer, resuspended in ice-cold active buffer with 0.25 μ L of each antibody per well, and incubated for another hour with gentle rocking at 4°C. Cells were washed twice with ice-cold storage buffer, resuspended in 100 μ L of ice-cold active buffer, and analyzed in the Fred Hutchinson Cancer Center Flow Cytometry shared resource facility.

For all conditions a negative transfection control was used to determine single cell gates, followed by single-antibody controls to determine fluorescence. Data were analyzed via FlowJo first gating for single cells, then FITC-positive cells (for protein expression), and finally APC/647-positive cells (for either open conformation or ligand-bound). The ratio of APC/647-positive cells within the parent FITC-positive gate was plotted in PRISM (GraphPad).

3.5.11 Mutant library construction

The complete library construction pipeline was adapted from previous work on Coronavirus spike-ACE2 interactions⁽³⁵⁶⁾. Codon-optimized α I domain constructs were ordered as G-blocks (IDT) and cloned into a galactose-inducible centromeric yeast display plasmid (ADDGENE), replacing the RBD sequence with the integrin domain sequence. The α L I domain

was stabilized in a high-affinity state with the K287C/K294C(380) disulfide bond, and the α M I domain was stabilized with a similar Q163C/Q309C(358) mutation. These were referred to as “wildtype” controls for the purposes of the deep mutational scanning assay. From these, single site saturation mutagenesis libraries for the entire α L and α M I domains were produced by Genscript and delivered as dsDNA fragments, allowing for wildtype and synonymous sequences, with flanking sequences for subsequent library assembly. A sample α L sequence is given below with uppercase letters denoting the mutated region.

tctgcaggctagtgggtggaggaggctctggtggaggcggccgcggaggcggagggtcggctagccatatgGGCAACGTGGA
TCTGGTGTTCCTGTTTGACGGCTCCATGTCTCTGCAGCCTGATGAGTTTCAGAAGATC
CTGGATTTTCATGAAGGACGTGATGAAGAAGCTGTCCAATACCTCTTATCAGTTTGCC
GCCGTGCAGTTCAGCACATCCTACAAGACCGAGTTCGACTTCAGCGACTATGTGAAG
AGAAAGGACCCCGATGCCCTGCTGAAGCACGTGAAGCACATGCTGCTGCTGACCAA
CACATTCGGCGCCATCAATTACGTGGCCACCGAGGTGTTTCGCGAGGAGCTGGGAG
CAAGGCCTGACGCAACAAAGGTGCTGATCATCATCACAGATGGCGAGGCCACCGAC
TCTGGCAACATCGATGCCGCCAAGGACATCATCAGGTATATCATCGGCATCGGCAA
GCACTTCCAGACAAAGGAGAGCCAGGAGACCCTGCACAAGTTTGCCAGCAAGCCCG
CCTCCGAGTTTGTGTGTATCCTGGATACCTTCGAGTGTCTGAAGGACCTGTTTACAG
AGCTGCAGAAGAAGATCTACtctcagggggggcggtccgaacaaaagctatttctgaagaggacttgtaatagagatc
tgataacaacagtgtagatgaacaaaatcgacttgggtccactgtacttttagctcgtacaaaatacaatatacttttcatttctccgtaaacac
atgttttccatgtaataatccttttctattttcgttcggttaccaactttacacatactttatagctattcacttctatacactaaaaactaagacaa
ttttaatttgcctgcctgcatatttcaattgttataaattcctataatttatcctattagtagctaaaaaagatgaatgtgaatcgaatcctaagag
aatt

From these libraries, a randomized N16 barcode and illumina sequencing handles were appended to each fragment via PCR(356). This barcoded fragment sequence for α L is given below.

tctgcaggctagtgggtggaggaggctctggtggaggcggccgcggaggcggagggtcggctagccatatgGGCAACGTGGA
TCTGGTGTTCCTGTTTGACGGCTCCATGTCTCTGCAGCCTGATGAGTTTCAGAAGATC
CTGGATTTTCATGAAGGACGTGATGAAGAAGCTGTCCAATACCTCTTATCAGTTTGCC
GCCGTGCAGTTCAGCACATCCTACAAGACCGAGTTCGACTTCAGCGACTATGTGAAG
AGAAAGGACCCCGATGCCCTGCTGAAGCACGTGAAGCACATGCTGCTGCTGACCAA
CACATTCGGCGCCATCAATTACGTGGCCACCGAGGTGTTTCGCGAGGAGCTGGGAG
CAAGGCCTGACGCAACAAAGGTGCTGATCATCATCACAGATGGCGAGGCCACCGAC
TCTGGCAACATCGATGCCGCCAAGGACATCATCAGGTATATCATCGGCATCGGCAA
GCACTTCCAGACAAAGGAGAGCCAGGAGACCCTGCACAAGTTTGCCAGCAAGCCCG
CCTCCGAGTTTGTGTGTATCCTGGATACCTTCGAGTGTCTGAAGGACCTGTTTACAG
AGCTGCAGAAGAAGATCTACtctcagggggggcggtccgaacaaaagctatttctgaagaggacttgtaatagagatc
tgataacaacagtgtagatgaacaaaatcgacttgggtccactgtacttttagctcgtacaaaatacaatatacttttcatttctccgtaaacac
atgttttccatgtaataatccttttctattttcgttcggttaccaactttacacatactttatagctattcacttctatacactaaaaactaagacaa
ttttaatttgcctgcctgcatatttcaattgttataaattcctataatttatcctattagtagctaaaaaagatgaatgtgaatcgaatcctaagag
aattaatgatacggcgaccaccgagatctacacttttcctacacgacgtcttccgatctNNNNNNNNNNNNNNNNNNgcgg
ccgcgagctccaattcgcctatagtgagtcgtattacaattcactgg

These barcoded fragments were introduced to NotI/SacI-digested parental plasmids via HiFi assembly, leading to a final plasmid library. This plasmid pool was heat-shock transformed in duplicate for each integrin domain into competent DH5 α cells (NEB) and plated at 10-fold dilutions onto ampicillin-containing LB plates. Colonies were scraped at an estimated 50000 cfu bottleneck, plasmid purified, and transformed into the AWY101 surface display *S. cerevisiae* strain(381) at library scale(382). Yeast were grown outgrown, frozen in 18 OD/mL aliquots, and stored at -80°C. To link N16 barcodes to codon variants, we performed PacBio long-read sequencing of plasmids miniprep-prepared from bacteria as described previously(356). Circular consensus sequence reads were processed and quality-controlled using alignparse(383). For each library we obtained at least one barcode per possible variant, with the vast majority sampled at least twice.

3.5.12 Deep mutational scanning for integrin-ligand binding affinity

The complete deep mutational scanning pipeline (TiteSeq) was adapted from previous work on Coronavirus spike-ACE2 interactions(356). The assay couples a titration experiment of the library yeast pool with ligand to FACS, sorting into bins based on ligand fluorescence. All titration experiments were performed in duplicate using the replicate libraries for each integrin domain, and R-squared values for replicates are provided in supplemental figures which tended to show concordance between libraries. For streamlining analysis and presentation only library 2 for each experiment is shown in figures, although the data for all replicates are provided in online repositories for completeness.

To select for yeast clones expressing I domain variants, a “pre-sort” was done for each library in which yeast were sorted based on the C-terminal Myc tag encoded in the I domain expression construct. Frozen yeast libraries were thawed and outgrown for 8-10 hours in SDCAA media, back-diluted to 0.67 OD600 in SGCAA to induce I domain expression, and incubated at room temperature overnight with gentle rotation. Control yeast strains with wildtype domains were grown in a similar manner. Following 16-18 hours of induction, yeast were washed in TBSM-BSA (20mM Tris-HCl pH 7.4, 150mM NaCl, 1mM MnCl₂, 0.2mg/mL BSA), and labeled with 1:100 FITC anti-Myc antibody to label expressed I domains for 1 hour at 4°C. Yeast were then washed twice with ice-cold TBSM-BSA and resuspended in cold TBSM (20mM Tris-HCl pH 7.4, 150mM NaCl, 1mM MnCl₂) for FACS sorting. Cells were binned on singletons then on FITC as in Fig. S3.2. Approximately 20 million FITC+ cells were sorted into collection media (SDCAA+1%BSA), pelleted, and grown overnight in 5mL SDCAA + pen-strep. These outgrowths were frozen in 18OD/mL and stored at -80°C as “FITC+ libraries” for subsequent TiteSeq processing.

TiteSeq was performed as previous(356). Control strains or FITC+ libraries were thawed and outgrown, induced for I domain expression, and washed in the same way as full libraries. Following this, libraries were split into 10-OD aliquots and rotated overnight at room temperature with varying half-log titration series of ligand as well as no-ligand and saturating-ligand concentrations for wildtype control strains. The next morning yeast were washed twice with ice-cold TBSM-BSA and incubated at 4°C for 1 hour with 1:100 FITC anti-Myc and 1:200 iFluor647 anti-HIS antibodies. Cells were then washed twice and resuspended in TBSM for FACS. For each ligand concentration at least 7 million single FITC+ cells were sorted into four different ligand fluorescence bins as shown in Fig. S3.2. These cells were outgrown overnight in 1mL SDCAA + pen-strep in a 96-well plate. These outgrowths were then plasmid minipreped

using a Zymo 96-well yeast miniprep kit per manufacturer instructions, with the exception of an extended (3 hour) zymolase treatment and single -80°C freeze/thaw cycle prior to the kit lysis step. The N16 barcodes were then PCR amplified from the miniprep pools with an additional appended barcode for Illumina multiplexing and submitted for Illumina NovaSeq sequencing.

Demultiplexed N16 barcode reads were matched to library variants using the `dms_variants` tool (https://github.com/jbloomlab/dms_variants), providing a counts value of variant in each FACS bin across all concentrations. These counts were subsequently weighted from the total number of reads in each bin to the number of cells sorted into each bin. The amount of ligand binding for each variant at all ligand concentrations was calculated as a mean bin based on the distribution of counts across all bins. A binding constant (K_D) was calculated using a non-cooperative Hill equation as in(356), with values limited to avoid extreme fits as appropriate. These binding constants were then normalized on a (min,0,max:-1,0,1) scale for presentation and comparison as this study focuses on relative increases or decreases in affinity as opposed to absolute values.

3.5.13 Structure and data visualization

Structures displaying the average effect of all mutations per-site were generated using the `Mutationscores` command in UCSF ChimeraX. Sequence conservation was plotted via “color byattribute.” All cryoEM map and model visualizations were generated in UCSF ChimeraX. Graphs for DMS validation experiments were generated using PRISM.

3.6 Supplemental Information

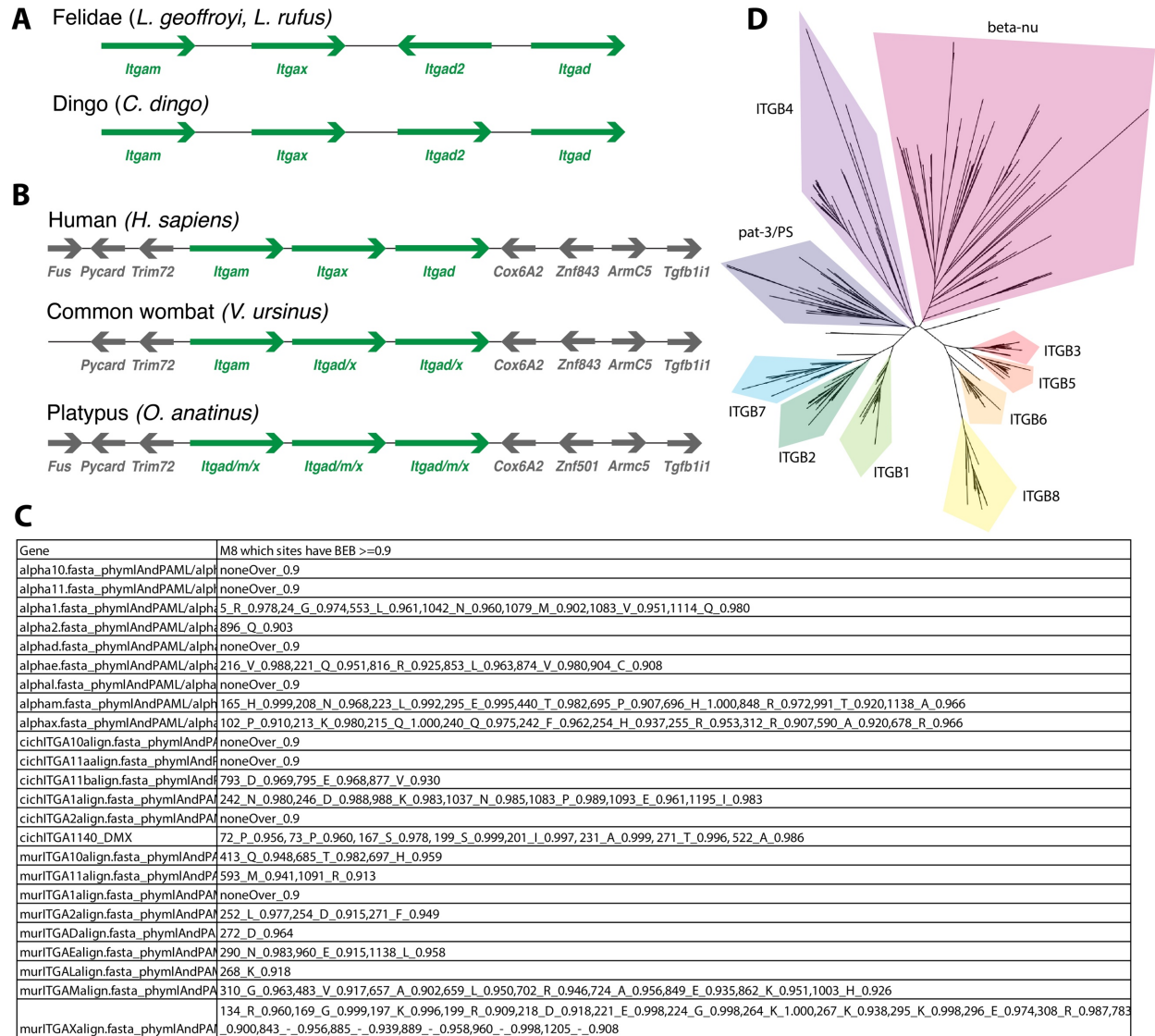


Figure S3.1. Rapid evolution of a subset of integrin α subunits. **A.** Genes in the mammalian α D/M/X clade have duplicated in tandem multiple times. **B.** The genomic neighborhood of the region encompassing the ITGAD/M/X genes in mammals. Genes are arranged in tandem, facilitating ease of genetic turnover. **C.** Expanded PAML positive selection analyses of I domain-containing integrin genes in murine, cichlid, and primate genomes. **D.** Phylogenetic analysis of beta integrin genes indicates concordance across the metazoan phylogeny, unlike that of the alpha subunits.

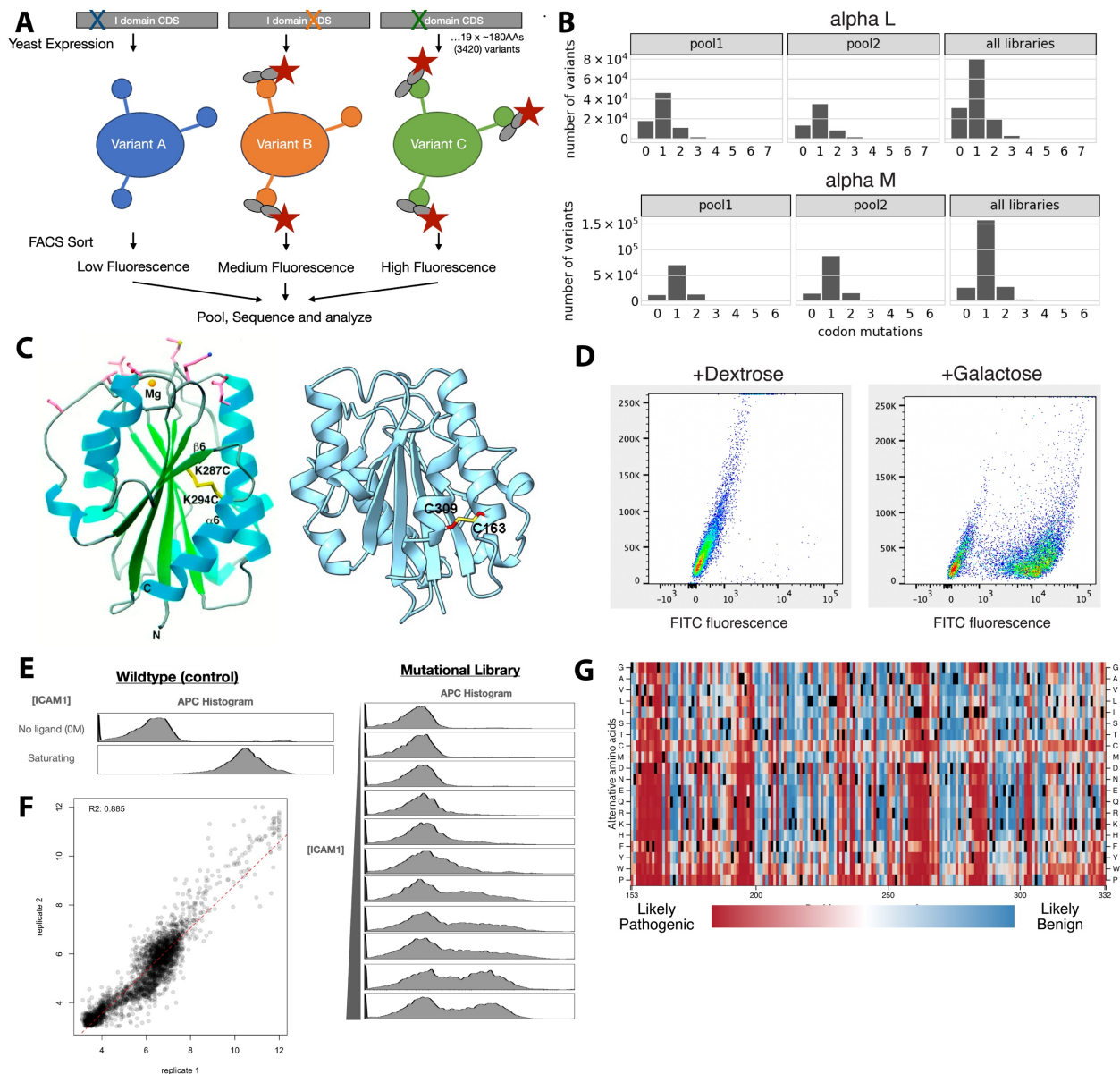


Figure S3.2. A high throughput yeast display assay to quantify the effects of mutations across the α L I domain. **A.** A schematic of the TiteSeq workflow which couples mutational analysis to ligand binding via flow cytometry and Illumina sequencing. **B.** Mutational distributions of each replicate pool (library) for both the α L (top) and α M (bottom) I domains. **C.** Predicted structures indicating the approximate placement of each stabilizing disulfide bond. The α L structure (left) is adapted from (359) and the α M (right) from (358). **D.** A representative flow plot showing inducible surface expression of the α L I domain upon galactose induction. **E.** Representative flow data from the wildtype (left) and α L library (right) showing a shift in APC ligand fluorescence upon increasing titration of ligand. **F.** Correlation plot of ligand affinity values between the two replicate α L libraries. **G.** AlphaMissense data for the α L I domain. Black residues are the wildtype.

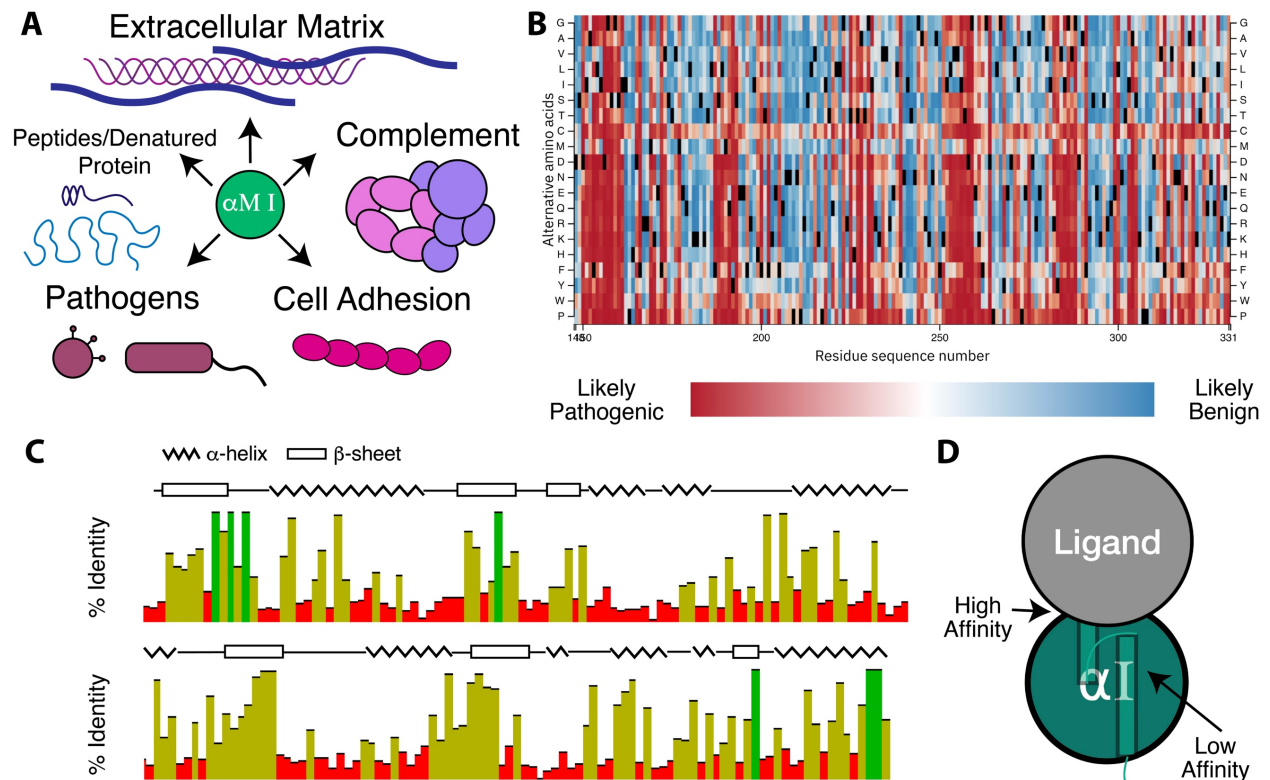


Figure S3.3. The mutational constraints of the α M β 2:C3d interaction. **A.** The α M I domain binds a range of structurally and functionally distinct ligands. **B.** AlphaMissense data for the α M I domain. Black residues are the wildtype. **C.** Conservation values from an alignment of all leukocyte I domains across vertebrates. Secondary structure tends to be more, but not completely, constrained. **D.** A model for how I domains generally optimize high affinity at the ligand interface and low affinity at the allosteric site.

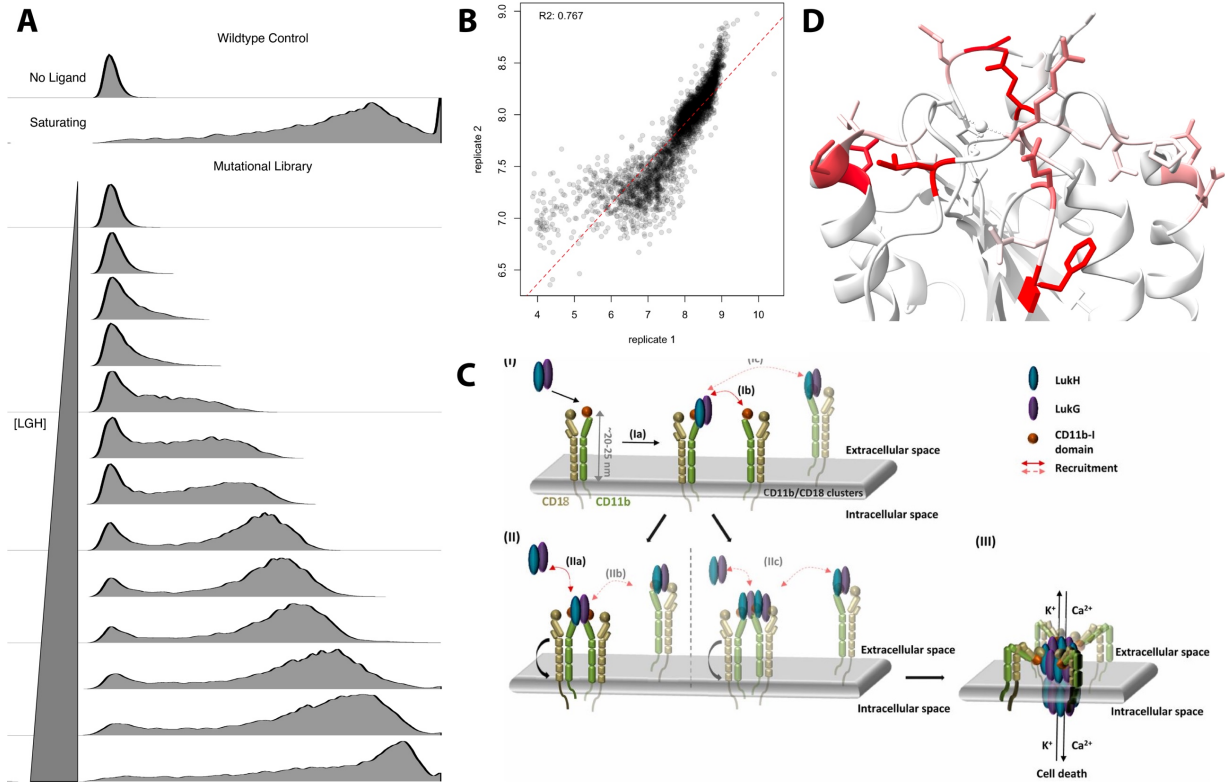


Figure S3.4. The mutational constraints of the α M β 2:LGH interaction. **A.** Representative flow data from the wildtype (top) and α M library (bottom) showing a shift in APC ligand fluorescence upon increasing titration of ligand. **B.** Correlation plot of ligand affinity values between the two replicate α M libraries. **C.** The molecular model for how LGH binds integrin in the open conformation, then forms an active cytolytic pore upon integrin shifting to the closed conformation. Adapted from (204). **D.** Differential mutational values between C3d and LGH mapped onto ligand binding residues in the active α M I domain structure. A redder color indicates a larger sum mutational difference between the two ligands.

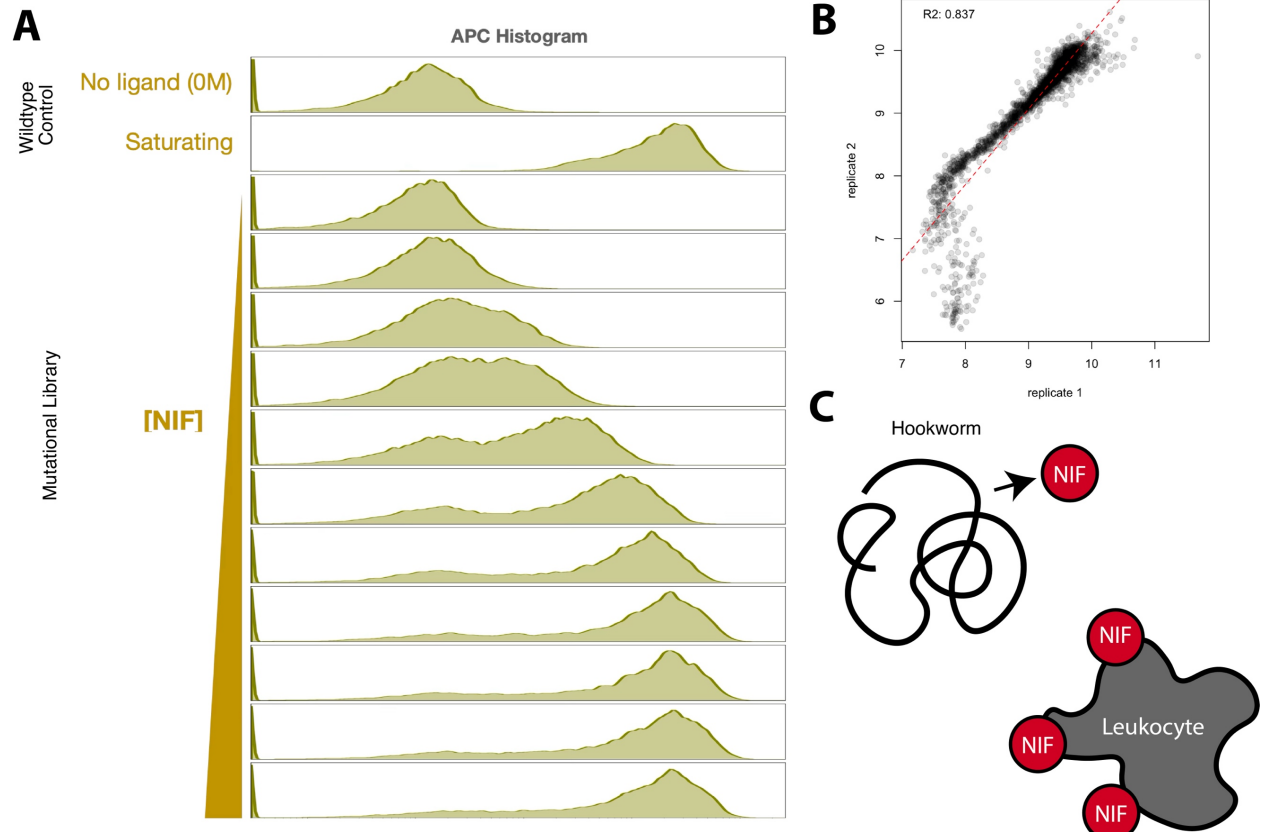


Figure S3.5. The mutational constraints of the α M β 2:NIF interaction. **A.** Representative flow data from the wildtype (top) and α M library (bottom) showing a shift in APC ligand fluorescence upon increasing titration of ligand. **B.** Correlation plot of ligand affinity values between the two replicate α M libraries. **C.** A schematic for how hookworm evades immune detection via steric blockade of host leukocytes.

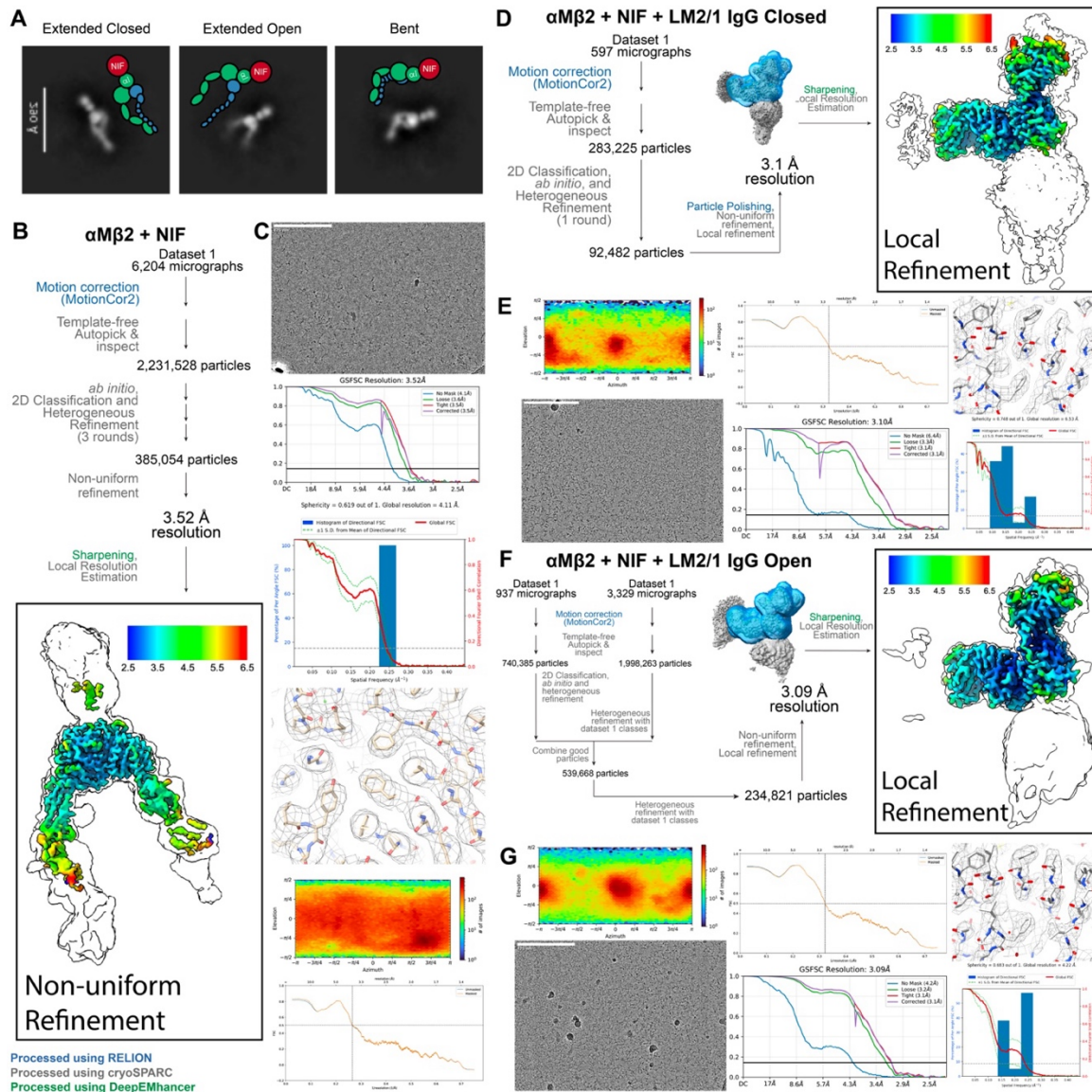


Figure S3.6. Structural determination of the $\alpha M\beta 2$:NIF interaction. **A.** Negative stain averages of the $\alpha M\beta 2$ ectodomain in various conformations bound to NIF. **B.** A data processing workflow for the NIF-bound $\alpha M\beta 2$ ectodomain cryoEM structure. **C.** Validation metrics for the NIF-bound $\alpha M\beta 2$ ectodomain cryoEM structure. Top to bottom: representative micrograph, GSFSC plot, 3DFSC plot, representative map-model fit, orientational distribution plot, map-model correlation plot. **D.** A data processing workflow for the NIF-bound αM I domain cryoEM structure in storage buffer. **E.** Validation metrics for the NIF-bound αM I domain cryoEM structure in storage buffer. Top row: Orientational distribution plot, map-model correlation plot, representative map-model fit. Bottom row: Representative micrograph, GSFSC plot, 3DFSC plot. **F.** A data processing workflow for the NIF-bound αM I domain cryoEM structure in active buffer. **G.** Validation metrics for the NIF-bound αM I domain cryoEM structure in active buffer. Top row: Orientational distribution plot, map-model correlation plot, representative map-model fit. Bottom row: Representative micrograph, GSFSC plot, 3DFSC plot.

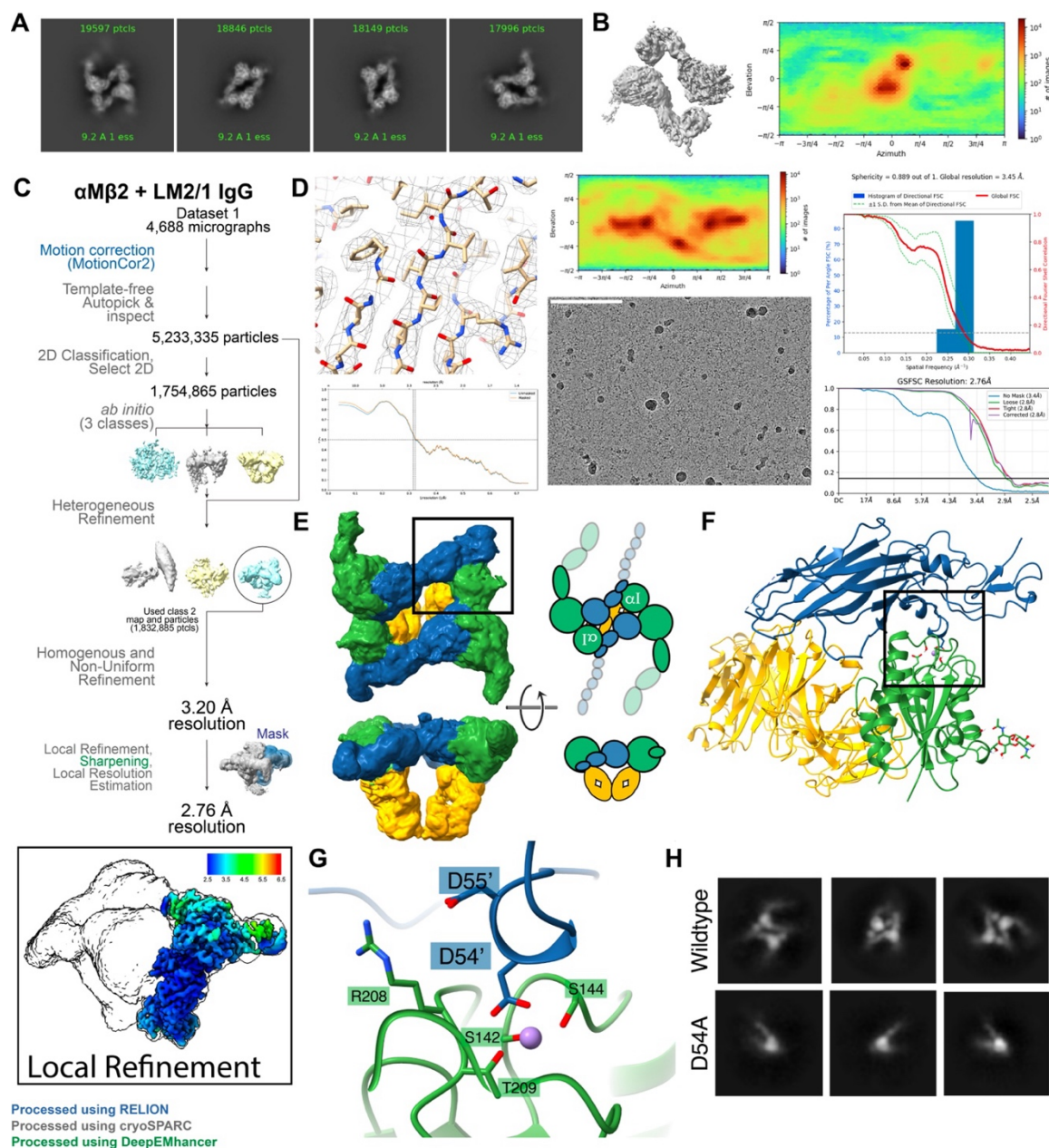


Figure S3.7. Structural determination of the αMβ2 dimer. **A.** 2D class averages of cryoEM data from the αMβ2 ectodomain dimer without added antibody. **B.** Reconstruction (left) and orientational distribution (right) from αMβ2 ectodomain dimer alone lead to a map without near-atomic features. **C.** A data processing workflow for the LM2/1 IgG-bound αMβ2 ectodomain dimer cryoEM structure. **D.** Validation metrics for the LM2/1 IgG-bound αMβ2 ectodomain dimer cryoEM structure. Top row: Representative map-model fit, orientational distribution plot, 3DFSC plot. Bottom row: Map-model correlation plot, representative micrograph, GSFSC plot. **E.** The full, low-threshold LM2/1 IgG-bound αMβ2 ectodomain dimer cryoEM structure; box indicates region used for local refinement. **F.** Molecular data of the local refinement structure; box indicates region of focus for panel g. **G.** The prime β2 subunit coordinates the αM MIDAS in the dimer structure. **H.** 2D averages of the wildtype and D54A αMβ2 ectodomain suggest breaking the MIDAS interaction is sufficient to abolish homodimerization.

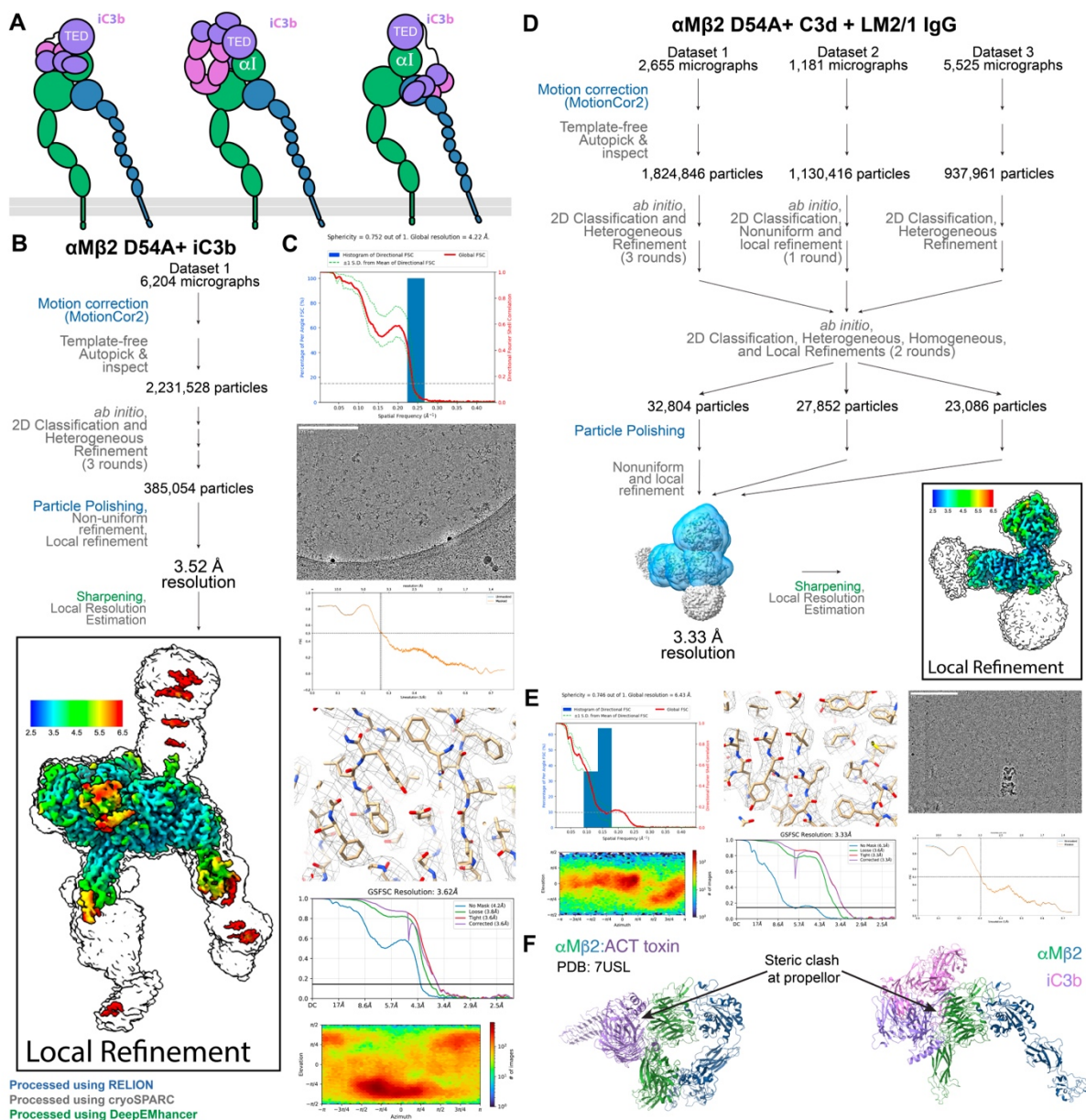


Figure S8. Structural determination of the $\alpha M\beta 2$:iC3b interaction. **A.** Different proposed models for the molecular interface between $\alpha M\beta 2$ and iC3b. **B.** A data processing workflow for the $\alpha M\beta 2$ ectodomain:iC3b cryoEM structure. **C.** Validation metrics for the LM2/1 IgG-bound $\alpha M\beta 2$ ectodomain dimer cryoEM structure. Top to bottom: 3DFSC plot, representative micrograph, map-model correlation plot, representative map-model fit, GSFSC plot, orientational distribution plot. **D.** A data processing workflow for the LM2/1 IgG-bound αM I domain cryoEM structure in complex with C3d. **E.** Validation metrics for the LM2/1: αM I domain:C3d structure. Top row: 3DFSC plot, representative map-model fit, representative micrograph. Bottom row: Orientation distribution plot, GSFSC plot, map-model correlation plot. **F.** A steric clash between ACT toxin and iC3b at the αM beta-propellor.

Table S3.1. Collection, refinement and validation statistics for cryoEM density maps and models. Collection details for all cryoEM maps presented in the text are shown above, and validation statistics for the models subsequently generated from those maps are shown below.

	α M β 2:NIF Ectodomain	α M β 2:LM2/1: NIF open I domain	α M β 2:LM2/1: NIF closed I domain	α M β 2:LM2/1 Dimer	α M β 2:iC3b ectodomain	α M β 2:LM2/1: C3d open I domain
Data collection and processing						
Magnification	36000x	36000x	36000x	36000x	36000x	36000x
Voltage (kV)	200	200	200	200	200	200
Electron exposure (e ⁻ / $\text{\AA}^2$)	50	50	50	50	50	50
Defocus range (μm , nominal)	1.2-1.8	1.2-1.8	1.2-1.8	1.2-1.8	1.2-1.8	1.2-1.8
Pixel size ($\text{\AA}$)	1.122	1.122	1.122	1.122	1.122	1.122
Symmetry imposed	C1	C1	C1	C1	C1	C1
Initial particle images (no.)	2,231,528	2,738,648	283,225	5,233,335	2,455,858	3,893,223
Final particle images (no.)	385,054	234,821	92,482	188,443	261,123	83,742
Map resolution ($\text{\AA}$)	3.52	3.09	3.10	2.76	3.62	3.33
FSC threshold	0.143	0.143	0.143	0.143	0.143	0.143
Map resolution range ($\text{\AA}$)	2.9-53.0	2.6-26.7	2.6-36.3	2.5-7.2	3.1-13.1	3.0-47.2
Refinement						
Initial model used (AlphaFold or PDB code)	AlphaFold PDB: 9RM9	AlphaFold2	α M β 2:LM2/1: NIF open I domain	AlphaFold	α M β 2:NIF Ectodomain, PDB: 7AKK	AlphaFold PDB: 4M76
Model resolution ($\text{\AA}$)	3.52	3.09	3.10	2.76	3.61	3.33
FSC threshold	0.143	0.143	0.143	0.143	0.143	0.143
Model resolution range ($\text{\AA}$)	2.9-53.0	2.6-26.7	2.6-36.3	2.5-7.2	3.1-13.1	3.0-47.2
Model composition						
Non-hydrogen atoms	9302	6878	6932	6459	17224	7177
Protein residues	1165	877	884	829	2164	919
Carbohydrates	23	2	2	2	0	0
B factors ($\text{\AA}^2$)						
Protein	104.52	63.90	60.42	67.55	95.67	58.18
Ligand	117.97	144.73	111.75	115.06	N/A	N/A
R.m.s. deviations						
Bond lengths ($\text{\AA}$)	0.005 (0)	0.003 (0)	0.003 (0)	0.005 (0)	0.003 (0)	0.003 (0)
Bond angles ($^\circ$)	0.736 (3)	0.535 (1)	0.599 (2)	0.636 (3)	0.657 (10)	0.633 (0)
Validation						
MolProbity score	1.72	1.39	1.29	1.31	1.31	1.44
Clashscore	5.45	2.60	2.80	2.99	3.24	3.73
Poor rotamers (%)	1.1	1.71	0.26	0.55	1.11	0.51
Ramachandran plot						
Favored (%)	94.20	97.01	96.58	96.58	97.12	95.94
Allowed (%)	5.63	2.99	3.31	3.42	2.88	4.06
Disallowed (%)	0.17	0	0.11	0.00	0.00	0.00

3.7 Afterword

This chapter is a culmination of the excellent work by trainees whom I supervised in the Campbell lab, and I think there's a lot to be proud of here. Anthony Kemp came out swinging during his summer project; his negative stain skills are second to none, and it was incredibly rewarding to watch him navigate complex literature with poise and excitement. Likewise, Chelsea's independence in both structural biology and developing new assays to validate the mutational data were extremely impressive and made the work in this chapter both fun and achievable. Both fantastic scientists have bright futures ahead of them; the experimental work here is an excellent stepping stone for both, but surely nothing compared to what is to come for both of them.

Chapter 4 Remote homology and functional genetics unmask deeply preserved Scm3/HJURP orthologs in metazoans

**This chapter was adapted from Hollis et al. 2026(384).*

4.1 Abstract

In most animals and fungi, centromere identity and function depend on the Scm3/HJURP chaperone, which deposits CENPA at centromeres. However, Scm3/HJURP orthologs appeared to be missing in insects, nematodes, many vertebrates, and other metazoans, suggesting radical chaperone replacement in these lineages. Here, we combine remote homology detection, AlphaFold-based structural modeling, and functional genetics in zebrafish and *Caenorhabditis elegans* to identify previously unknown Scm3/HJURP orthologs that localize to centromeres and whose loss causes catastrophic mitotic failure. We further show that *Drosophila* CAL1, long considered a functional analog, is instead a highly diverged Scm3/HJURP ortholog. Despite rapid primary-sequence divergence, predicted and known structures reveal a broadly conserved CENPA-H4-binding scm3 fold across fungi, vertebrates, nematodes, insects, and most metazoans. Our work demonstrates how rapid divergence can obscure the broad conservation of essential centromere machinery and provides a generalizable strategy for unmasking missing orthologs.

4.2 Introduction

In most eukaryotes, accurate chromosome inheritance during cell division depends on the precise assembly of centromeric chromatin, a process orchestrated by the deposition of the centromere-specific histone H3 (CenH3) variant, CENPA(385). In fungi and mammals, this deposition is facilitated by a specialized histone chaperone(386) (Fig. S4.1A). *S. cerevisiae* Scm3(272, 274, 387) and human HJURP(276, 277, 388) were independently identified for their roles in CENPA deposition through genetic and biochemical studies. Subsequent evolutionary analyses based on remote homology detection identified fungal Scm3 and animal HJURP proteins as distantly related orthologs(278). Intriguingly, this orthology could not be directly established using traditional methods such as reciprocal BLAST genome queries. Instead, the identification of a choanoflagellate Scm3/HJURP protein, with homology to both fungal and animal proteins, provided a critical evolutionary intermediate that enabled the successful identification of orthology between the fungal Scm3 and animal HJURP proteins. Subsequently, several structural and functional studies demonstrated that Scm3(273, 389) and HJURP(390) share a common ‘scm3 domain’(391) (we refer to the domain as scm3 to distinguish it from the Scm3 protein in yeast and choanoflagellates) that mediates the specific recognition and assembly of CENPA-containing nucleosomes at centromeres, highlighting orthology of these proteins despite divergence of their primary sequences. This functional conservation underscores the significance of Scm3/HJURP-mediated CENPA loading and stabilization(392, 393) in eukaryotic chromosome segregation, dating back to at least the common ancestor of fungi and animals. In contrast, plants appear to utilize a distinct chaperone, nuclear autoantigenic sperm protein (NASP), to deposit CENH3 in *Arabidopsis thaliana* and other land plants(394, 395).

At the molecular level, the Scm3 and HJURP chaperones recognize the CENPA-Targeting Domain (CATD)(396), which encompasses the L1 loop and the $\alpha 2$ helix of the

CENPA histone fold domain (Fig. S4.1B). The unique molecular features of the CENPA CATD are both necessary and sufficient for Scm3/HJURP to distinguish it from other H3 variants and to deposit CENPA strictly at the centromere, thereby making the scm3 domain critical for chaperone specificity(397). The scm3 domain is comprised of a structurally conserved ~25aa α -helix followed by a ~10aa β -strand that runs antiparallel to the CENPA α 2 helix and L1 loop, respectively(273, 390). Most of HJURP lacks experimental structural information and is predicted to be largely disordered. Scm3/HJURP proteins carrying CENPA/H4 dimers are recruited to the centromere via two parallel pathways: the first utilizes the polo-like kinase-licensed Mis18 complex to recruit HJURP(398–400), while the second involves HJURP binding directly to the inner kinetochore protein CENPC(401, 402). These pathways appear to be broadly functionally conserved across the metazoan tree(403–406).

Despite its ancient origins, the evolutionary history of the Scm3/HJURP chaperone remains unclear in many animal lineages. Although sequence profile analyses and sensitive homology searches have identified distant relatives of Scm3 and HJURP in some lineages(278), orthologs appear to be missing in several major animal groups, especially in fish, insects, and nematodes(407) (Fig. S4.2). This is particularly puzzling, as most of these groups still possess canonical CENPA-dependent centromeres(408), indicating that the need for a CENPA chaperone persists, even if the specific chaperone might differ. The absence of detectable Scm3/HJURP homologs in these species has been taken as evidence for the independent evolution of distinct centromeric chaperones or of entirely different centromere assembly processes(278, 409).

The case of insects, especially *Drosophila*, has been central to this puzzle. Genetic and cell biological studies have unambiguously identified the CAL1 protein as the CENPA chaperone in *Drosophila melanogaster*(275, 410, 411) and other dipteran species(412), where it is thought to play a role similar to that of HJURP/Scm3 in CENPA deposition. However, CAL1 shows no obvious sequence similarity to Scm3/HJURP, leading to its classification as a functional analog rather than a true ortholog(409). This distinction is based on the failure to detect homology to the conserved scm3 domain using standard sequence comparison tools. Nonetheless, structural studies have shown that CAL1 does contain an ‘scm3 domain’-like motif, which interacts with CENPA molecules similarly to Scm3/HJURP proteins(413). Therefore, based on current evidence, it remains unclear whether CAL1 is a true ortholog whose orthology is obscured by rapid divergence or an example of convergent evolution, in which a new protein has independently gained the ability to chaperone CENPA.

Unlike insects, nematodes such as *Caenorhabditis elegans* appear to lack a recognizable CENPA chaperone that is either orthologous to Scm3/HJURP or a clear functional equivalent, such as CAL1(403, 414). Instead, recent research suggests that the exceptionally long, helical CENPA N-terminal tail, combined with the centromere protein KNL-2, may have usurped chaperone functions in centromeric chromatin assembly, potentially eliminating the need for a dedicated CENPA chaperone(403, 405). In support of this model, genetic and cytological studies have shown that KNL-2, which lacks homology to Scm3/HJURP, is essential for CENPA loading and kinetochore formation in nematodes(403, 405). This, combined with their holocentric architecture, suggests that *C. elegans* and possibly other nematodes may have evolved a fundamentally different mechanism for centromere specification and maintenance, highlighting the remarkable possibility that the machinery involved in depositing centromeric histones could be highly adaptable across evolution.

Here, we revisit this important question using both classical genetic approaches and recent advances in remote homology detection(415, 416), which incorporate sequence

comparisons and structural predictions, along with recently sequenced genomes (Fig. S4.3). From these analyses, we demonstrate that most metazoan clades have preserved an orthologous centromeric chaperone characterized by the conserved scm3 domain. For example, we identified previously undetected scm3 domain-containing proteins in deuterostomes, including teleost fish and echinoderms. Using genetic perturbations and cytological analyses, we functionally validate our identification of Scm3/HJURP orthologs in zebrafish (*Danio rerio*) and nematodes (*C. elegans*), demonstrating that these putative orthologs localize to centromeres and that their loss leads to catastrophic mitotic defects. We also trace the arthropod ancestry of the scm3 domain, revealing that *Drosophila* CAL1 is a genuine Scm3/HJURP ortholog rather than a functional analog. Structural modeling using AlphaFold(332, 340) predicts that even highly diverged Scm3 domains from species such as zebrafish and nematodes adopt a conserved fold and interact with CENPA/H4 in a manner similar to that of their human and fungal counterparts. Our discovery of scm3 domain-containing proteins in basal arthropods, non-dipteran insects, nematodes, and other early-branching metazoans suggests that the ancestral metazoan and most metazoan lineages possess CENPA chaperones, which include an scm3 domain. Our findings comprehensively revise our understanding of centromeric histone chaperone evolution in animals, highlight the limitations of standard homology detection methods in the face of extreme sequence divergence, and offer general approaches to overcome these limitations.

4.3 Results

4.3.1 HJURP orthologs are broadly conserved across deuterostomes, including teleosts

Since the initial study identifying orthology between human HJURP and yeast Scm3(278), several additional high-quality genome assemblies across the metazoan tree have become available(417). These genomes have provided more power to identify distant homology between clades. We employed an iterative BLAST (PSI-BLAST)(418) approach (Fig. S4.3), which leverages repeated searches with new hits incorporated at each iteration, to identify remote homologs of full-length human HJURP in teleosts (ray-finned fish). Consistent with previous efforts, we were initially unsuccessful. However, we could confidently identify an HJURP-like protein in the genome of *Latimeria chalumnae* (West Indian Ocean coelacanth)(419), a lobe-finned fish, despite this sequence lacking the canonical scm3 domain. Instead, the coelacanth protein was identified by homology with human HJURP residues outside the scm3 domain. Despite lacking an scm3 domain, subsequent iterative searches using the putative coelacanth HJURP sequence successfully identified previously uncharacterized proteins across deuterostomes, including echinoderms, teleost fish, and cephalochordates (Fig. 4.1A). In most deuterostome genomes, only a single strong hit was identified. We generated a MUSCLE(351) alignment of these putative HJURP homologs from representative tetrapods, teleosts, and echinoderms, revealing a conserved scm3 domain across all identified sequences. Our results are supported by the InterPro database(240), which also lists several of these putative deuterostome HJURP proteins as containing a *bona fide* scm3 domain. Phylogenetic analyses of these scm3 domains were generally congruent with the deuterostome species tree(420) (Fig. 4.1B, Fig. S4.4A), consistent with the hypothesis that a monophyletic, broadly retained common HJURP ancestor is present among all deuterostomes.

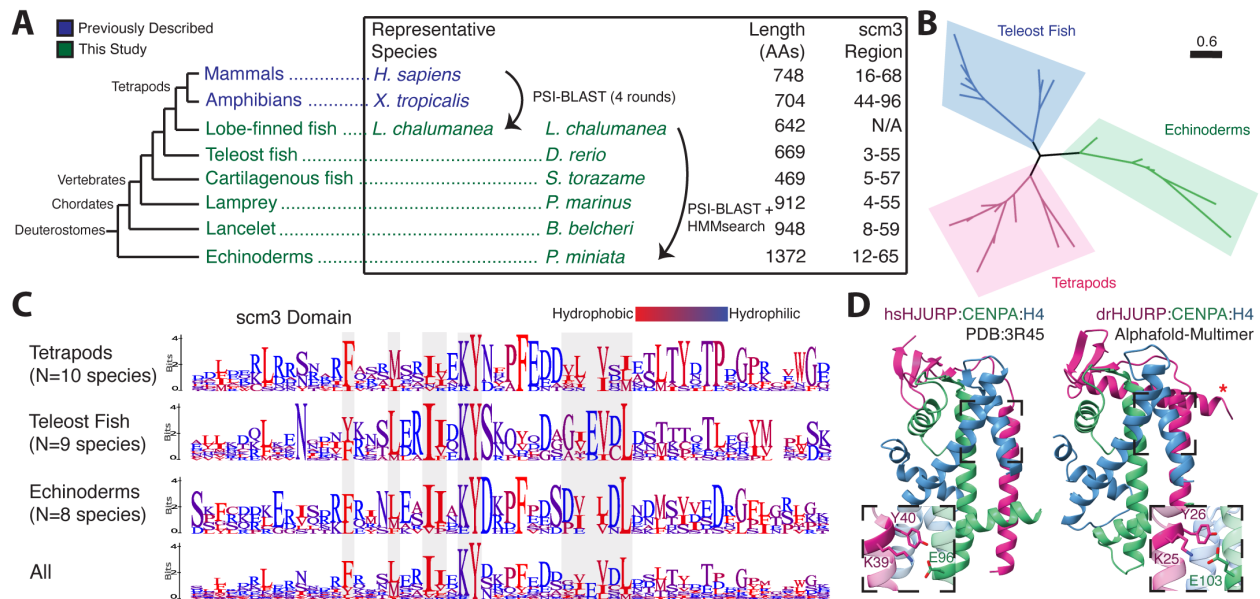


Figure 4.1. The centromeric chaperone HJURP predates tetrapods. **A.** A phylogeny (not to scale) of deuterostome putative HJURP proteins. Deuterostome HJURPs have high variation in length and composition, sharing only a short N-terminal scm3 domain. **B.** Phylogenetic analysis shows that putative deuterostome scm3 domains roughly follow the species tree, consistent with a monophyletic origin of HJURP. **C.** Logo plots of the scm3 domain for representative clades within deuterostomes. scm3 domains were extracted, aligned together, and then separated into distinct clades for the purposes of visualization. **D.** The *Homo sapiens* crystal structure (PDB:3R45, left) and AlphaFold-Multimer prediction of the *Danio rerio* HJURP-CENPA-H4 complex (right) show general structural similarities. The KY motif (insets) interacts with a conserved glutamate on the CENPA $\alpha 2$ helix.

Overall, these sequences showed little homology among deuterostomes, particularly outside the scm3 domain (9.91% average whole-protein identity), making whole-protein alignments intractable and uninformative (Fig. S4.4B). Even within the scm3 domain, there is relatively little sequence conservation (27.26% shared identity) compared to what would be expected of an essential domain in an essential protein(421). Logo plots of scm3 domains from representative clades show only a few broadly conserved amino acids (Fig. 4.1C). Among these is the previously identified KY dipeptide, which is almost universally conserved among deuterostomes, with only select conservative substitutions at the lysine residue. Hydrophobic amino acids spaced approximately 2-4 positions apart also appear to be important for the scm3 domain (Fig. 4.1C); this regular spacing is consistent with the alpha-helical structure of the scm3 domain that runs antiparallel to the $\alpha 2$ helix in the CENPA histone fold domain (Fig. S4.1B). An HHpred(422) E-value of 2.30e-04 supports this finding of structural homology (Table S4.1). HHpred values were also used as an additional validation metric for subsequent searching.

Conversely, while these general patterns define the scm3 domain, it appears to exhibit remarkable sequence plasticity even within deuterostomes. For example, a threonine within the “TLTY” motif was previously thought to be widely conserved across opisthokonts (encompassing fungi, choanoflagellates, and metazoa)(278, 391), but shows only low conservation among these representative deuterostome clades. There is marked diversity in hydrophobic residues helical within and among clades (Fig. 4.1C). Furthermore, some residues are conserved in a lineage-specific manner. For example, the consensus residue immediately following the KY is conserved within, but differs across, representative lineages. Such variation may reflect lineage-specific coevolution between HJURP and the CENPA/H4 histone dimers. Although the scm3 domain appears to be the single shared molecular feature across HJURP

proteins in deuterostomes, tetrapods and teleost fish share an additional motif immediately upstream of the DNA-binding HMD region (Fig. S4.1B) as well as moderate homology within the previously identified(400) Mis18-binding R2 region (Fig. S4.4C). This region of homology is not shared with echinoderms, suggesting that it arose and became entrenched for functional purposes in the common ancestor of tetrapods and teleosts.

Identification of deuterostome HJURPs confirmed that the scm3 domain is the definitive conserved feature of HJURP proteins in deuterostomes. Therefore, we hypothesized that the absence of the scm3 domain in the original coelacanth *L. chalumnae* HJURP protein must have arisen from an assembly error or gap in the coelacanth genome at this site. To further test this hypothesis, we used Logan Search(423) with the *L. chalumnae* gene sequence to query the Sequence Read Archive for homology to this gene. We found a close match to a transcript from the testes of the closely related *L. menadoensis* (Indonesian coelacanth)(424) that, when translated, extended the N-terminus of this protein to include an intact scm3 domain (Fig. S4.4D). Consistent with our assembly junction hypothesis, we could align two different *L. chalumnae* contigs to the bases 1-92 and 157-2691, respectively, of the *L. menadoensis* transcript. However, a contig gap spanning bases 93-156 rendered the sequence of the *L. chalumnae* HJURP homolog and its scm3 domain incomplete (Fig. S4.4D). Overall, our results indicated the scm3 domain is invariant in HJURP orthologs and provides the most conclusive signal for HJURP ortholog identification.

Our remote homology analyses revealed that a previously uncharacterized gene from zebrafish (*Danio rerio*) (ZFIN Gene ID: ZDB-GENE-030804-4; name: *si:dkeyp-117h8.4*(425)) encodes an HJURP ortholog, which we refer to as drHJURP. To gain further insight into the structural constraints that might shape the scm3 domain's evolution in deuterostomes, we performed AlphaFold-multimer analyses of drHJURP in complex with the zebrafish CENPA-H4 dimer. Despite the low sequence homology to scm3 from human HJURP, AlphaFold-multimer predicted with high confidence a protein complex very similar in structure to the *Homo sapiens* crystal structure (Fig. 4.1D). Features broadly conserved in vertebrate scm3 domains include the KY motif (Fig. 4.1D) and a downstream motif comprised of acidic (E/D) and short hydrophobic residues (I/L/V) in a repeating pattern which forms the secondary beta sheet structure. These residues appear to perform orthologous functions in the case of human and zebrafish complexes. However, six CENPA residues essential for specificity in the human HJURP-CENPA interaction(426) are entirely different in *Danio rerio*, as are the complementary HJURP residues (Fig. S4E). Nevertheless, the essential pocket cradling CENPA's specificity-defining CENPA-Targeting Domain (CATD)(396) is considerably similar in shape, with a backbone RMSD of 0.71Å² between the helix-sheet region of the human and predicted zebrafish structures, despite dramatic sequence differences. This AlphaFold prediction reveals the structural basis for sequence plasticity within the scm3 domain: the strongest constraints arise from entrenched hydrophobic interactions (390, 426), which are maintained in the scm3 domain to complement the hydrophobic CENPA residues required for tight nucleosome formation at the centromere. Surprisingly, the AlphaFold prediction also included an additional helix C-terminal to the scm3 domain that also appears to contact the histone dimer (Fig. 4.1D). This feature is reminiscent of the *Drosophila* CAL1-CID-H4(413) crystal structure, which shows that further tertiary structure beyond the central scm3-CENPA parallel helices helps augment CAL1-CENPA interaction specificity.

4.3.2 The zebrafish gene *si:dkeyp-117h8.4* encodes HJURP

To assess the centromeric function of drHJURP *in vivo*, we tested three critical attributes. We first asked whether drHJURP preferentially interacts with drCENPA rather than with canonical H3. To detect a direct interaction between drCENPA and drHJURP, we performed *in vitro* pull-down assays using recombinant proteins. These experiments revealed the hallmark preference of HJURP for CENPA/H4 over canonical H3/H4(427) (Fig. 4.2A). However, we note that this pulldown with drHJURP was qualitatively less efficient than those performed with orthologous tetrapod proteins, including human HJURP(426–428).

A second expectation is that drHJURP should localize to centromeres. To test this expectation, we co-injected zebrafish zygotes with *in vitro*-transcribed mRNAs encoding eGFP-tagged drCENPA (to mark centromeres) and mCherry-tagged drHJURP, then examined the localization of these tagged proteins via live imaging. We observed strong colocalization of drCENPA and drHJURP in embryonic nuclei (Fig. 4.2B). This colocalization was cell-cycle dependent: drHJURP was present at centromeres only in interphase nuclei (16/25 nuclei) and was completely absent in other phases of the cell cycle (Fig. 4.2B, Movie S1 in online publication). In contrast, drCENPA localized to centromeres throughout the cell cycle. This localization pattern mirrors that of human HJURP, which replenishes CENPA at the telophase-to-G1 transition(276, 277, 429). The cell-cycle-dependent localization of drHJURP is also consistent with its expression being most strongly correlated with other cell-cycle components (Fig. S4.5A) in single-cell transcriptomics data(430, 431) and *in situ* hybridization(432).

Finally, we tested the third expectation for drHJURP; given its fundamental role in CENPA deposition, we expected that drHJURP would be essential for early embryonic development(433). To test this possibility, we injected zebrafish embryos with Cas9 protein and CRISPR guide RNAs targeting drHJURP, drCENPA, or a scrambled negative control (Fig. S4.5B). We then examined CRISPR F0 embryos looking for indications of developmental impairment. While the control group was largely unimpaired, drHJURP CRISPR embryos displayed gross morphological defects and extensive signs of necrotic tissue at 24hpf (Fig. 4.2C). By comparison, drCENPA CRISPR embryos displayed a comparatively milder, yet substantial phenotype. We hypothesize that the differences in phenotypic severity between the drHJURP and drCENPA CRISPR embryos can be attributed to the exceptionally long half-life of the CENPA protein(434), together with its considerable maternal deposition. If the primary cause of necrosis in drHJURP and drCENPA CRISPR embryos is centromeric dysregulation, we hypothesized that we would observe signs of catastrophic cell division. Consistent with this expectation, we found extensive evidence of mitotic failure in drHJURP CRISPR embryos, including lagging and bridging chromosomes in anaphase cells as well as fractured nuclei and micronuclei in interphase cells (Fig. 4.2D-E). Furthermore, we find a ~65% reduction of GFP-drCENPA intensity in drHJURP CRISPR embryos compared to scramble controls (Fig. 4.2F). Thus, our *in vivo* data indicate that the *D. rerio* gene *si:dkeyp-117h8.4* indeed encodes the long-missing zebrafish HJURP, validating the conclusions from our remote homology search.

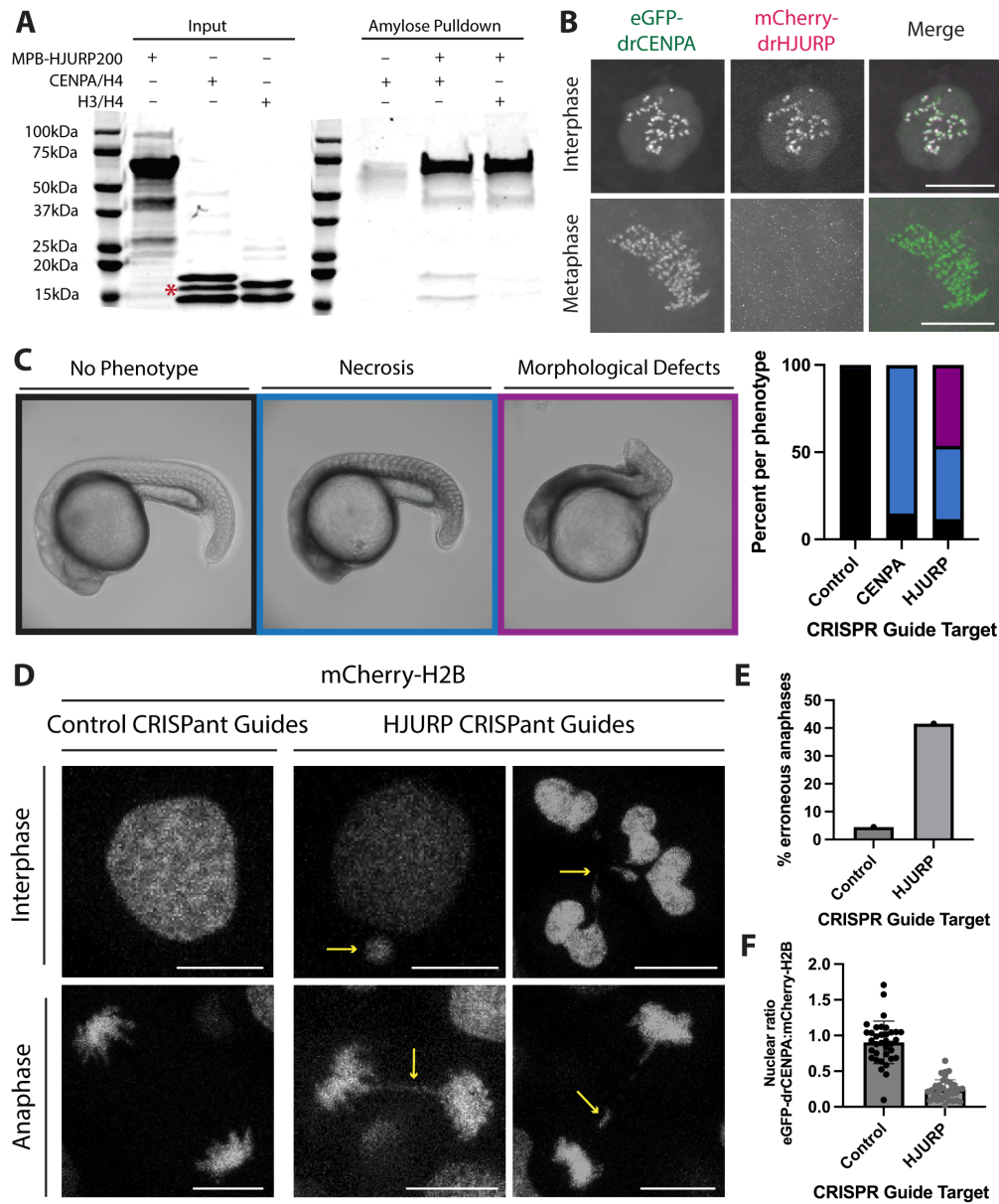


Figure 4.2. *Danio rerio* encodes an HJURP ortholog. **A.** Amylose pulldown assay with purified MBP-drHJURP and drCENPA/H4 or canonical H3/H4 tetramers. MBP-drHJURP interacts directly with drCENPA/H4 tetramers as suggested by co-precipitation of species at drCENPA and H4 sizes. Red asterisk indicates a non-histone contaminant. **B.** Still frames from timelapse series of cell divisions in *D. rerio* embryos expressing GFP-CENPA and mCherry-HJURP show cell cycle-specific colocalization of drHJURP and drCENPA. **C.** Morphological analysis of CRISPR zebrafish larvae with control (left representative image, N=96), CENPA (N=81), or HJURP (center and right representative images, N=138) guides. Severe morphological defects suggest HJURP is an essential gene. **D.** While control guide-injected embryonic cells generally segregate their chromosomes cleanly, HJURP CRISPR embryos have severe chromosome segregation defects, including frequently aberrant anaphases and fragmented nuclei. Scale bars are 5µm. **E.** Quantification of anaphase defects in control or HJURP CRISPR embryos. Numbers of anaphases are 49 for control guides, 105 for HJURP guides. **F.** Relative centromeric CENPA fluorescence in control CRISPR embryos (N=34 nuclei) compared to HJURP CRISPR embryos (N=36 nuclei) suggests a loss of centromeric CENPA in the absence of HJURP.

4.3.3 *Drosophila CAL1* is a true HJURP ortholog

Based on current evidence, it is unclear whether the CAL1 CENPA chaperone identified in *Drosophila* and other *Diptera* represents a divergent Scm3/HJURP ortholog or whether it independently evolved to converge on HJURP/Scm3-like structure and function(407, 409). The success of our remote homology detection strategy in identifying HJURP orthologs in deuterostome lineages led us to investigate whether we could distinguish between these two possibilities by extending our search to arthropods. To this end, we sought to identify broadly distributed arthropod proteins that contain scm3 domains. We generated different Hidden Markov Models (HMMs) via HMMer(435–437) either from deuterostome HJURP scm3 domains, yeast/choanoflagellate scm3 domains, or Dipteran CAL1 proteins, encompassing either just the scm3-like domain or the first 160 residues (Data S1 in online publication). We used these different HMM models to search available arthropod genomes for putative Scm3/HJURP homologs.

To our surprise, we found a non-overlapping set of putative HJURP homologs using these strategies. For example, using the deuterostome or yeast/choanoflagellate HMMs, we were able to successfully identify scm3 domain-encoding genes from basal panarthropods, such as tardigrades and horseshoe crabs, suggesting that the scm3 domain was present in the arthropod common ancestor (Fig. 4.3A). However, this search did not initially reveal any homologs in insects. Conversely, using Dipteran-based HMMs, we identified proteins with scm3-like domains in *Diptera* and other insect clades, including *Hymenoptera* (bees, ants, wasps), *Coleoptera* (beetles), and *Orthoptera* (crickets), but not in other arthropods. However, by searching arthropod genomes using iteratively built HMMs combined with AlphaFold modeling, we identified representative scm3 domain-containing proteins in all major clades of arthropods (Fig. 4.3A). Scm3/HJURP orthologs from *Hymenoptera* were extremely useful evolutionary intermediates because they were found to be homologous to both deuterostome HJURP and insect CAL1 proteins, thus confirming the orthology between dipteran CAL1 and non-insect arthropod Scm3/HJURP proteins. These findings provide the most definitive evidence that *Drosophila* CAL1 is a true, albeit more divergent, ortholog of Scm3/HJURP.

Based on our more comprehensive repertoire of Scm3/HJURP orthologs in arthropods, we sought to understand why earlier efforts might have failed to establish CAL1 as a true HJURP ortholog. We focused on sequence features that make CAL1 appear distant from other proteins with scm3 domains. To our surprise, we found that, unlike deuterostomes, the scm3 domain is not the most conserved region of the CAL1 chaperone in *Diptera* (Fig. S4.6A). Rather, the scm3 domain shows comparably low sequence homology within dipterans compared to other arthropod lineages (Fig. 4.3B). A subset of Dipterans has also lost selective constraint on the definitive KY motif (Fig. 4.3C), although the orthologous DF residues in *Drosophila melanogaster* remain functionally important(413). This loss may be dipteran-specific, as the KY is still largely present in the hymenopteran insect clade (Fig. 4.3B), as is the acidic/short hydrophobic patch. Other features appear to be universal among insects, such as largely invariant tryptophan and arginine residues, which are also widely conserved in the *Daphnia* crustacean lineage. The regular spacing of conserved residues is also consistent with both deuterostome scm3 sequences and the expected parallel helical structure of the chaperone-histone complex(413). Despite these features, insect scm3 domain sequences are considerably diverged from those of early-branching arthropods (Fig. 4.3B). The relaxed sequence constraints on dipteran scm3 domains may explain why initial attempts failed to establish CAL1's orthology

with human HJURP, necessitating the requirement for evolutionary intermediates from other insects and other arthropods described here.

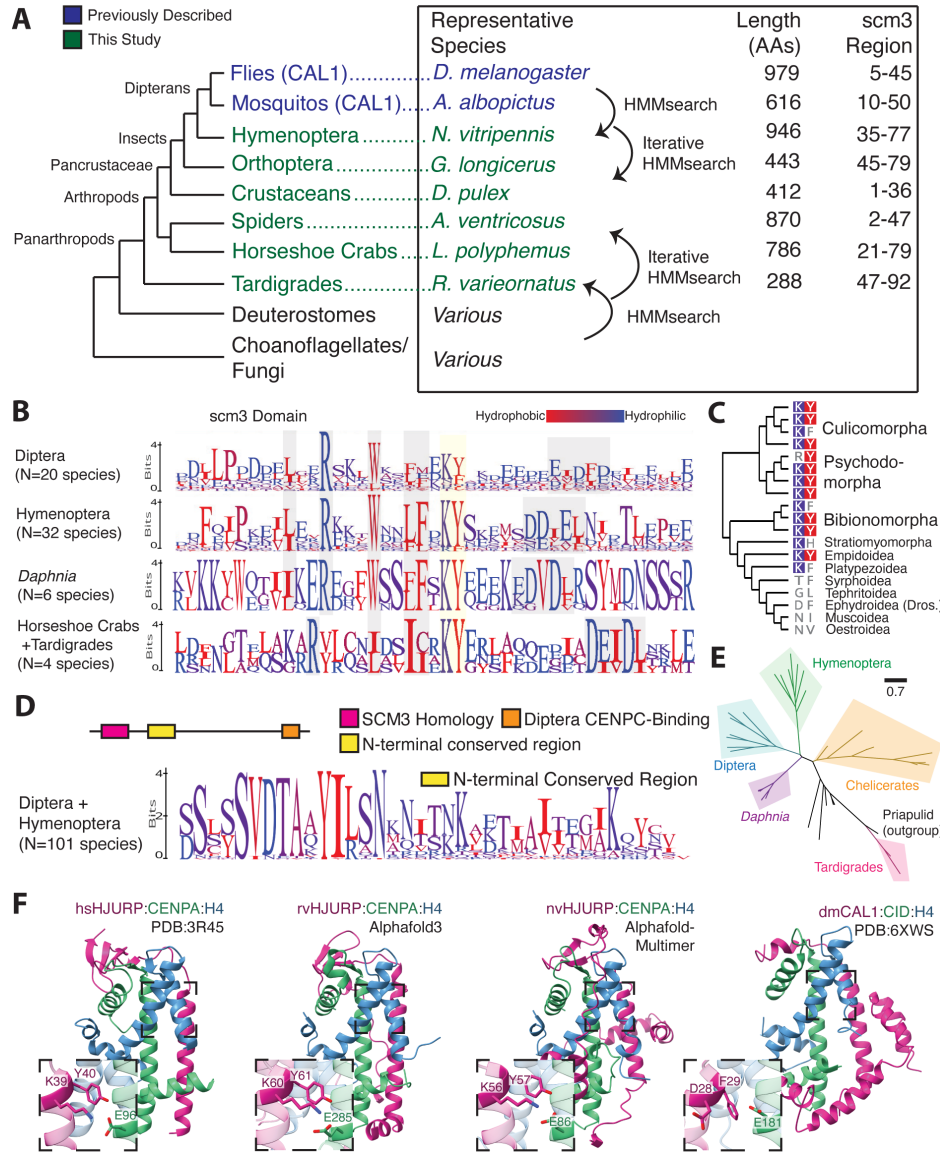


Figure 4.3. CAL1 is a true Scm3/HJURP ortholog. **A.** A schematic of the approach used to establish homology of the scm3 domain between basal arthropods and Dipteran CAL1. We find homologous scm3-like sequences across the arthropod lineage. As with deuterostomes, arthropod sequences are highly variable and share only the scm3 domain. **B.** Logo plots of the scm3 domain for clades within panarthropods suggest broadly conserved residue-level sequence features (gray and yellow highlighting) are present across arthropod scm3 orthologs. **C.** A Dipteran species phylogeny showing amino acids aligned to the conserved KY positions, suggests that the constraints on this motif were lost relatively recently in fly evolution. **D.** Sequence homology is shared beyond the scm3 domain between Hymenoptera and Diptera. A motif flanking the C-terminal side of the scm3 domain shares more homology between the orders than the scm3 domain itself does. **E.** A phylogenetic tree constructed using full length HJURP sequences from representative arthropod species generally follows the species tree, supporting a monophyletic origin of HJURP in arthropods. **F.** Structural predictions of representative tardigrade (rvHJURP) and Hymenopteran (nvHJURP) scm3 domains complexed with their respective species' CENPA and H4 suggest a stepwise progression from the human HJURP structural fold to the *D. melanogaster* CAL1 fold.

As support for the hypothesis of altered constraints on the scm3 domain in *Diptera*, we found that the region immediately C-terminal to the scm3 domain in *Diptera* and *Hymenoptera* shows generally higher sequence homology than the *Diptera* scm3 domain itself (Fig. S4.6A, Fig. 4.3D). This region may form additional contacts beyond the scm3 domain, although it lies just beyond the CAL1 fragment used for the crystal structure(413). These data together suggest the *Hymenoptera* CAL1 is a representative evolutionary intermediate between the ancestral HJURP and the dipteran CAL1 of monophyletic origin; phylogenetic analysis generally supports this placement (Fig. 4.3E, Fig. S4.7).

We employed AlphaFold-multimer(332) to predict conserved features of the putative *Nasonia vitripennis* (*Hymenoptera*) HJURP. We found that the predicted nvHJURP-CENPA-H4 structural topology is highly similar to the *Drosophila melanogaster* CAL1-CENPA-H4 crystal structure (Fig. 4.3F). Secondary and tertiary features of the complex appear to be conserved between insect clades. However, the nvHJURP prediction contains antiparallel beta strands with the CATD region of nvCENPA, which more closely resemble those in the *Homo sapiens* HJURP structure(390) but are absent from the *Drosophila* CAL1 crystal structure. Thus, structural predictions of *Nasonia* HJURP show intermediate features between basal arthropod HJURP and dipteran CAL1 sequences. This supports our hypothesis that constraints on *Drosophila* CAL1 are relaxed or altered relative to other Scm3/HJURP orthologs. These observations suggest a gradual evolutionary trajectory in which the enhanced function of domains C-terminal to the scm3 region in *Diptera* may have relaxed constraints on the scm3 domain in early dipterans. Such a change in constraint would have challenged prior efforts to establish orthology based solely on the scm3 domain. Further, if the hymenopteran prediction represents a structural intermediate, we would expect a more basal arthropod structure to more closely resemble the canonical human and yeast-like fold. Consistent with this expectation, the predicted HJURP-CENPA-H4 structure from the tardigrade *R. varieornatus* (sister clade to all arthropods) also shows features more consistent with the human protein crystal structure, such as a tighter antiparallel helix interface and an extended beta sheet region (Fig. 4.3F).

In addition to an altered scm3 domain and an additional domain immediately C-terminal to scm3 (Fig. 4.3B-D), *Drosophila* CAL1 proteins have additional unique evolutionary constraints not seen in other CENPA chaperones(407, 438), including a C-terminal CENPC-binding domain whose function is thought to replace Mis18 binding present in tetrapod chaperones(413, 439, 440). In contrast to *Diptera*, CAL1/HJURPs in *Hymenoptera* show comparatively low conservation of their C-terminal regions, and there is little likeness between *Diptera* and *Hymenoptera* in this region (Fig. S4.6B). These findings suggest that the CAL1 CENPC-binding domain was acquired after the divergence of *Diptera* and *Hymenoptera*, further highlighting the plasticity of the centromeric machinery across the metazoan tree. Despite these idiosyncrasies, our findings trace the evolutionary trajectory of HJURP within arthropods and strongly support the hypothesis that CAL1 and HJURP proteins are genuine orthologs.

4.3.4 *Nematodes encode HJURP proteins*

Another metazoan lineage that appears to lack Scm3/HJURP orthologs entirely is the nematode lineage. The holocentric nematode *C. elegans* has proven instrumental in advancing our understanding of the fundamental principles that govern the cell cycle, with detailed genetic and cell biological analyses identifying most of the currently known essential kinetochore components in nematodes(441). Nevertheless, neither previous genetic screens nor prior *in silico*

and *in vitro* studies have successfully identified an HJURP ortholog in nematodes(278, 442). Instead, the current model suggests that the exceptionally long N-terminal tail of the *C. elegans* CENPA homolog HCP-3 (hereafter called CENPA^{HCP-3}) interacts uniquely with the centromeric chromatin assembly factor KNL-2, partially replacing HJURP function(403, 405). However, this interaction alone is insufficient to explain how CENPA^{HCP-3} is stably recruited from the cytosol to the centromere.

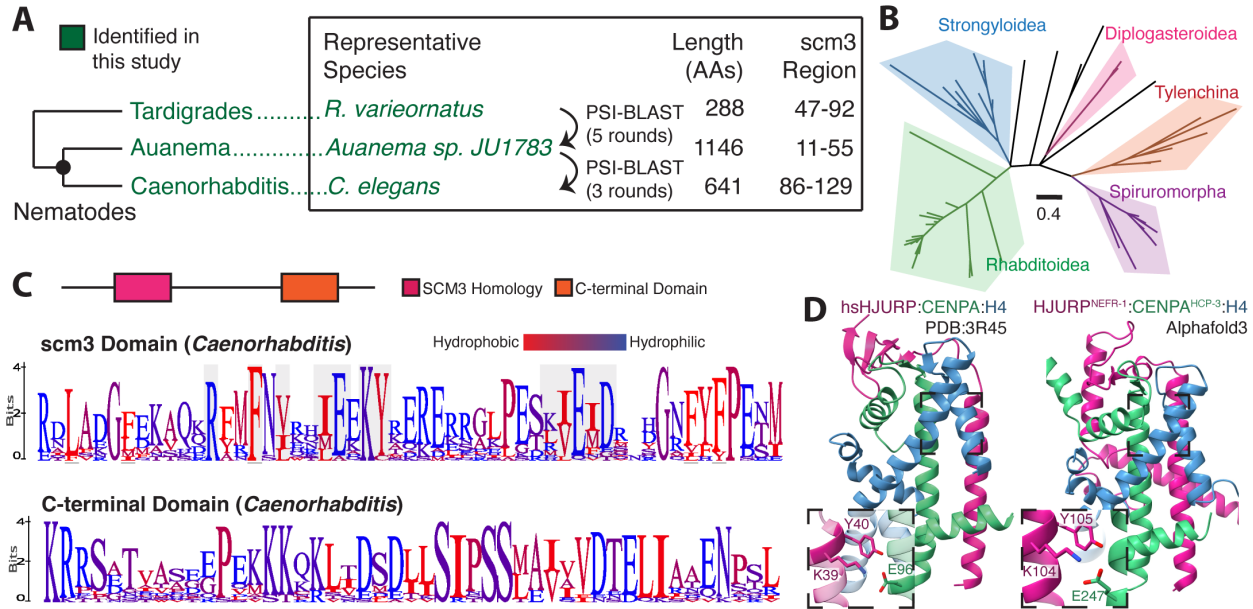


Figure 4.4. Nematode genomes encode a canonical HJURP protein. **A.** A visual schematic of the approach used to identify HJURP orthologs in nematodes. We use an intermediate *Auanema* predicted protein to establish scm3 homology in *Caenorhabditis*. The tardigrade scm3 domain provided the necessary signal for iterative searching of nematode proteomes. **B.** A phylogenetic tree constructed from scm3 domains of diverse nematode species is generally concordant with the species tree. **C.** *Caenorhabditis* HJURP proteins contain the N-terminal scm3 domain as well as a highly conserved C-terminal region. C-terminal logo plot shows a representative subsection of the entire conserved region. Alignments are from 20 species. **D.** A structural prediction of the *C. elegans* scm3 domain in complex with CENPA and H4 suggests structural features are broadly conserved with those of the human ortholog.

Given this functional gap, we sought to determine if our remote homology approach could identify a *bona fide* HJURP in nematodes, including *C. elegans*. Using the scm3/HJURP HMM approach, which was successful in deuterostomes and arthropods, we were still unable to identify an HJURP ortholog in *C. elegans*. This lack of success appeared to confirm current models that nematodes might lack HJURP orthologs. However, our search criteria identified a region of minimal scm3 homology in the protein CAI4225225 from the nematode species *Auanema sp. JU1783*(443), which is only distantly related to *Caenorhabditis*. Using this sequence as a query, we performed iterative BLAST searches of nematode genomes, eventually revealing remote homology to an scm3 domain in *C. elegans* gene *nefr-1*(444) (Fig. 4.4A). As with deuterostomes, we found the phylogenetic placement of putative HJURP orthologs largely follows the nematode species tree(445), suggesting that the common ancestor of nematodes encoded a single HJURP ortholog that has been stably inherited across nematodes (Fig. 4.4B, S4.8A).

NEFR-1's scm3-domain shares key conserved features with that of deuterostomes and fungi, including the KY motif, alternating (E/D) and (I/L/V) residues downstream of the KY, and

a repeating hydrophobic pattern indicative of a helical region that could bind CENPA^{HCP-3} (Fig. 4.4C). This scm3-like region is broadly shared across *Caenorhabditis* despite little homology flanking it (Fig. S4.8B). There is also a conserved C-terminal motif shared across the genus, a feature highly reminiscent of the conserved C-terminal domain in *Drosophila* CAL1, which mediates its binding to CENPC. This motif may mediate a similar molecular interaction in *Caenorhabditis*. We used AlphaFold-multimer(340) to generate a predicted structure of the NEFR-1 scm3 domain in complex with CENPA^{HCP-3} and H4. This prediction shares a very similar global architecture to the human HJURP/CENPA/H4 structure, as well as a conserved interaction of the KY dipeptide and broad helix-sheet interface (Fig. 4.4D), strongly supporting our identification of NEFR-1 as a *bona fide* HJURP protein.

4.3.5 *C. elegans nefr-1* encodes a bona fide HJURP ortholog

C. elegans nefr-1 has been previously studied and described as coding for a neurofilament-like protein involved in neuronal function(444). Thus, our remote homology finding of an scm3 domain in NEFR-1 was initially puzzling. However, in addition to its remote homology and strict conservation in nematodes, including other *Caenorhabditis* species, we identified several attributes that bolster our hypothesis that *nefr-1* plays an important role in centromeric maintenance, a role that may have been overlooked in previous studies. For example, *nefr-1* expression is strongly correlated with genes that perform cell-cycle functions, as would be expected for centromeric function but not necessarily for neuronal function. The *nefr-1* gene is encoded in an operon with the DNA repair gene *msh-6* and shares a transcription factor-binding region with the *C. elegans* separase homolog *sep-1*, both of which are chromatin-associated, cell-cycle-regulated genes. Moreover, single-cell RNA-seq experiments reveal abundant *nefr-1* mRNA expression in the germline, as with CENPA^{HCP-3} (446, 447), but not in neurons (Fig. S4.8C-F).

Buoyed by our success at validating drHJURP, we investigated the centromeric function of *nefr-1* in *C. elegans*. To date, no complete knockout for *nefr-1* has been functionally characterized (Fig. S4.9A). Previous analysis of *nefr-1* relied on an in-frame deletion allele, *ok471* (*C. elegans* Gene Knockout Consortium), which is viable. However, this mutant retains the putative scm3 domain and potentially the C-terminal conserved region, suggesting the possibility that several *nefr-1* function(s) may remain intact in the *ok471* truncation mutant. We used CRISPR-Cas9 to generate a *C. elegans* strain with an N-terminal stop codon and frameshift in the *nefr-1* gene (K15*), providing a near-complete *nefr-1* knockout model (Fig. S4.9A). Unlike the previously characterized deletion allele *ok471*, we were unable to isolate viable *nefr-1*(K15*) homozygotes, suggesting this ‘complete knockout’ mutation confers lethality or sterility. We therefore used a phenotypically marked balancer chromosome to follow the *nefr-1*(K15*) allele (Fig. S4.9B). While heterozygotes progressed with no discernible developmental delay and produced viable progeny, *nefr-1*(K15*) homozygous mutants predominantly arrested development at the L1 stage, were severely uncoordinated, and mostly died within 72 hours of hatching (Fig. S4.9C-D), all consistent with a previously reported aneuploidy phenotype(448). These results show that *nefr-1* is an essential gene.

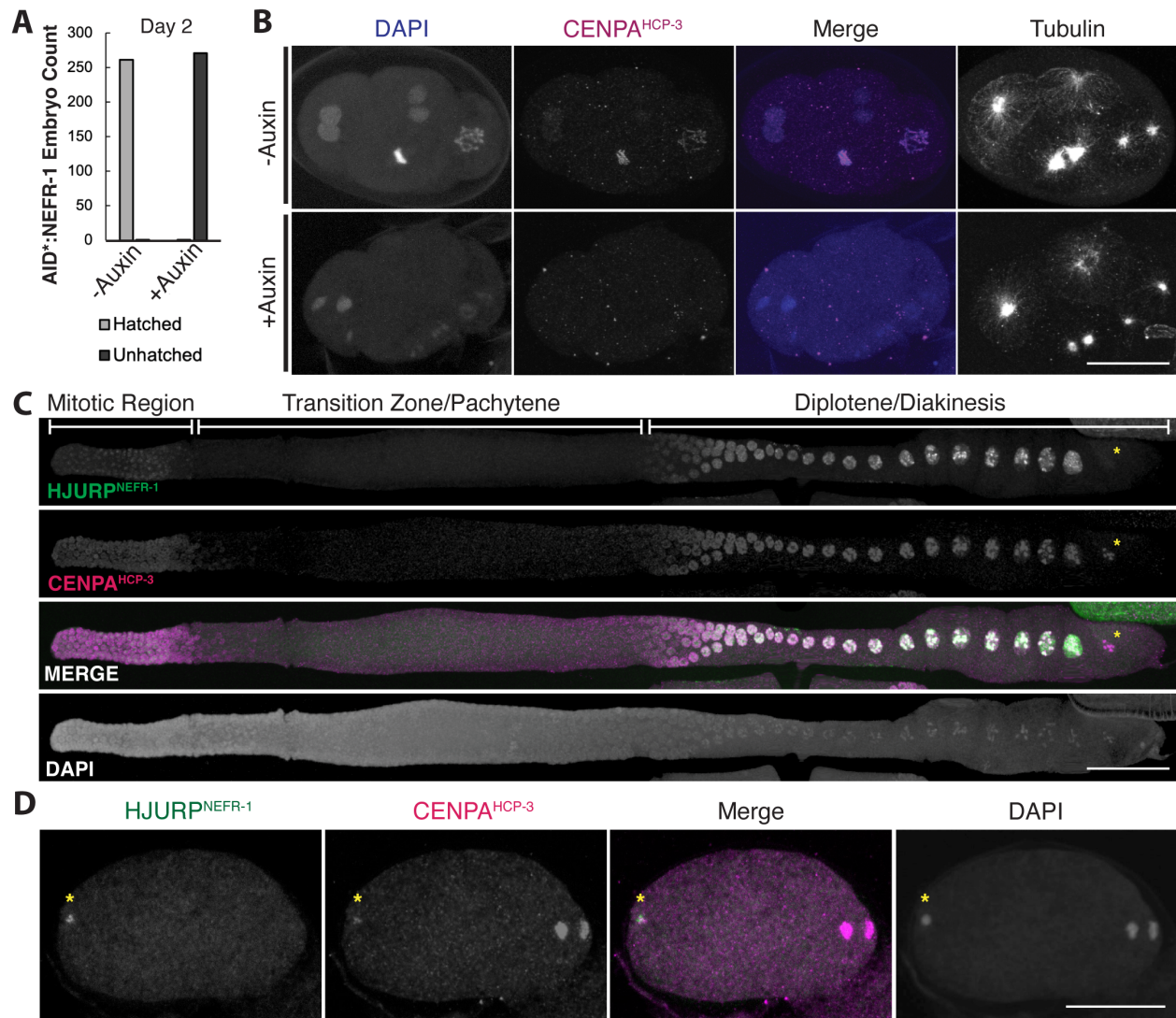


Figure 4.5. *C. elegans* NEFR-1 is an essential cell division regulator. **A.** Embryonic hatching counts from control or auxin-treated worms with an AID-tagged NEFR-1 show that NEFR-1 is essential for early embryonic development. **B.** Representative fixed imaging of control (top row) or auxin-treated (bottom row) 4-cell embryos with an AID-tagged NEFR-1. While control embryos have consistently sized, CENPA-coated chromatin, drug-treated embryos have catastrophic chromosome segregation abnormalities and lack CENPA on their chromatin. Scale bar is 20 μ m. **C.** NEFR-1 co-localizes with CENPA in the *C. elegans* germline, including at the mitotic distal tip (left) and diplotene/diakinesis (right). Yellow asterisk indicates that NEFR-1 leaves chromatin by late diplotene, while CENPA is retained for meiotic divisions. Scale bar is 40 μ m. **D.** A representative fertilized *C. elegans* oocyte undergoing anaphase I. NEFR-1 co-localizes with the paternal genome (yellow asterisk) but is absent from the maternal dividing chromatin (right), suggesting that the maternal machinery reloads CENPA onto the paternal DNA prior to the first zygotic division. Scale bar is 20 μ m.

We were initially puzzled by the ability of K15* homozygotes to progress to the L1 developmental stage, because failure to deposit CENPA^{HCP-3} should cause lethal defects in cell division even earlier during embryogenesis(449, 450). We hypothesized that maternally deposited *nefr-1* mRNA or protein would be sufficient to enable sufficient CENPA^{HCP-3} deposition for faithful mitosis during embryogenesis and delay embryonic lethality. To unambiguously test NEFR-1 function without the confounding effects of maternal deposition, we

generated an *nefr-1* strain with an N-terminal auxin-dependent degron tag (*AID::nefr-1*)(451, 452) and used germline-specific expression of the auxin-dependent ubiquitin ligase TIR1 to achieve acute germline-specific depletion of NEFR-1. Embryos produced following NEFR-1 depletion from the hermaphrodite germline almost invariably failed to hatch (Fig. 4.5A). These findings confirm our hypothesis that maternal deposition of *nefr-1* likely delays its embryonic lethality. To better understand the cause of lethality, we fixed and imaged early embryos after 24 hours of auxin-mediated NEFR-1 depletion. Whereas control embryos showed healthy mitotic features such as a uniform metaphase plate and homogeneous nuclei, auxin-exposed embryos displayed disorganized, highly variable chromosomes that lacked CENPA^{HCP-3} (Fig. 4.5B). These observations are consistent with a kinetochore-null phenotype(449), leading to severe chromosome segregation defects. We conclude that maternally contributed NEFR-1 is essential for embryonic cell division and deposition of CENPA^{HCP-3}.

Previous studies have shown that CENPA^{HCP-3} is loaded onto chromosomes *de novo* during the diplotene phase of *C. elegans* meiosis(453). To observe NEFR-1 localization, we generated an endogenous N-terminal 3xFLAG-tagged *nefr-1* allele (*3xFLAG::nefr-1*) and performed imaging in fixed *C. elegans* animals. We observed strong colocalization between CENPA^{HCP-3} and 3xFLAG::NEFR-1 in the germline, most prominently at the mitotic distal tip and on diplotene and diakinesis chromosomes, consistent with a centromere-defining role for NEFR-1 (Fig. 4.5C). We further observed that NEFR-1 localized to the paternal pronucleus within anaphase oocytes (Fig. 4.5D). This finding suggested a mechanism by which the paternal genome in *C. elegans*, which is devoid of CENPA^{HCP-3} in the spermatocyte(453), can re-establish centromeric chromatin prior to zygotic cell divisions. Like orthologous HJURP proteins(276, 277, 429, 454), NEFR-1 appears to be temporally restricted, being absent from late prophase (Fig. 4.5C) and anaphase (Fig. 4.5D) meiotic chromosomes.

Finally, we performed immunoprecipitation-mass spectrometry (IP-MS) analysis using 3xFLAG::NEFR-1 from worm embryo lysate as bait to identify NEFR-1's *in vivo* interaction partners. We found a unique association with CENPA^{HCP-3}, as well as increased associations with cell cycle proteins, relative to untagged control lysates (Fig. S4.10A-B). As with zebrafish HJURP, we performed pulldown experiments using purified recombinant proteins and found a direct interaction between NEFR-1 and HCP-3/H4 that was stronger than that between NEFR-1 and H3/H4 (Fig. S4.10C). Together, these data confirm that *nefr-1* encodes the long-missing *C. elegans* HJURP ortholog that chaperones CENPA^{HCP-3} to chromatin and defines the holokinetic centromere in *C. elegans*.

4.3.6 *Scm3/HJURP* orthologs are broadly conserved across Metazoa

Finally, we reinvestigated the early ancestry of the scm3 domain in *Metazoa*. In addition to arthropods, we extended our combination of PSI-BLAST and HMM-based searching with HHpred(422) validation (Fig. S4.3) to identify representative scm3 domains in the remainder of the protostome lineage as well as in the early-branching sponges and placozoa (Fig. 4.6A). We successfully identified Scm3/HJURP orthologs in most metazoan lineages we investigated (Fig. 4.6A), including in multiple species in which they were previously not found (Fig. S4.2), entirely filling the evolutionary gap between choanoflagellate Scm3 and human HJURP orthology first established 15 years ago(278) (Fig. S4.11). As in arthropods and deuterostomes, these sequences were highly variable in length and showed very little homology outside the scm3 domain. Structural prediction supports a unified scm3 protein fold across these early-diverging clades (Fig. 4.6B).

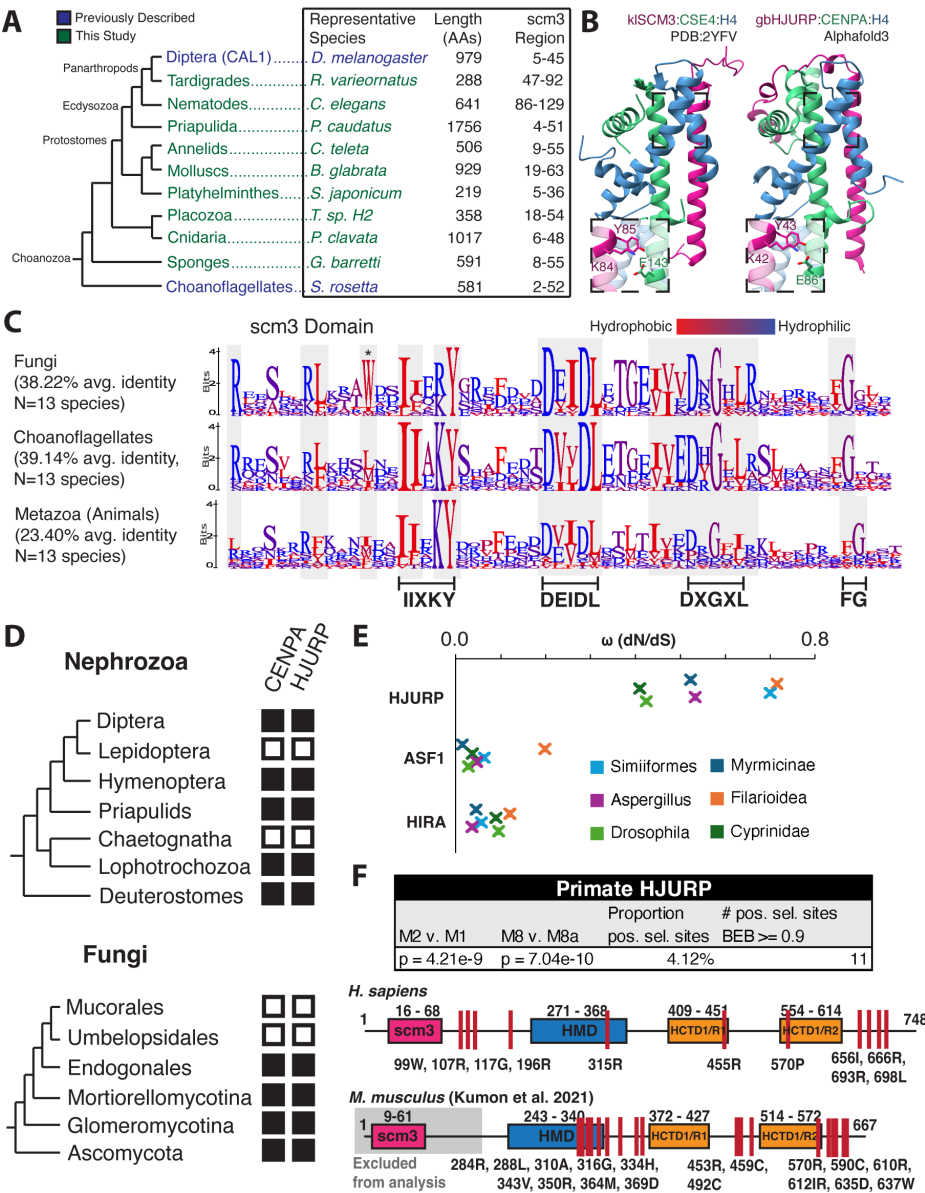


Figure 4.6. HJURP is present in basal metazoa. **A.** An expanded metazoan phylogeny encompassing basally-branching groups and protostomes for which we identified HJURP candidates. **B.** AlphaFold3 predicts a complex of the sponge *G. barretti* candidate HJURP with gbCENPA and H4, which is structurally similar to the published yeast

K. lactis complex. **C.** Logo plots of representative Fungal, Choanoflagellate, and metazoan scm3 domains show select sequence features are conserved across opisthokont evolution. Metazoan scm3 domains show lower identity compared to those of the opisthokont sister clades. **D.** The loss of CENPA is associated with the absence of candidate HJURP proteins across opisthokont evolution. We failed to identify strong HJURP candidates in clades nested between those with clear sister group and outgroup HJURP orthologs. We show only a subset of clades with CENPA losses. **E.** HJURP evolves more rapidly than functionally comparable histone chaperones. Omega (dN/dS) values for HJURP candidates from well-sequenced clades were consistently higher than those for the histone chaperones ASF1 and HIRA. **F.** We identified sites of positive (diversifying) selection in primate HJURP in regions similar to (C-terminus, HMD) those previously identified in rodents (455), indicative of pervasive adaptive pressure.

Successful identification of scm3 domains in these lineages allowed us to investigate whether any features are unique to metazoan HJURP proteins relative to Scm3/HJURP proteins from yeast and choanoflagellates. We generated scm3 domain alignments encompassing major clades of fungi(456), choanoflagellates(457), and animals(458), respectively, of comparable species divergence timescales. We found that metazoan scm3 domains have lower overall sequence identity (Fig. 4.6C). Signatures such as the DXGXL motif, which were likely present in the common ancestor of fungi and animals, appear to evolve under less stringent selective constraints on the metazoan branch. We also found that the acidic/short hydrophobic region immediately following the KY, which interacts specifically with the CENPA CATD beta-strand, faces relaxed constraints in metazoans relative to other opisthokonts (Fig. 4.6C). One reason for this rapid divergence might be the evolutionary acquisition of asymmetric female meiosis in *Metazoa*, which may have led to the rapid evolution of CENPA proteins (459, 460) and, subsequently, of HJURP proteins.

Our analyses indicate that some unusual features of CENPA chaperones and their presence have independently converged in yeast and metazoans. For example, *Dikarya* (a subkingdom of *Fungi*) largely prefer a tryptophan in the same position as CAL1 in *Diptera* (asterisk, Fig. 4.6C). In both *Fungi* and *Metazoa*, the loss of CENPA also coincided with the loss of HJURP; such dual losses have occurred multiple times across the opisthokonts (Fig. 4.6D) (461–463). This is a further indication that HJURP and CENPA are functionally co-dependent, supporting these candidate proteins as *bona fide* functional HJURP orthologs.

Previous work found evidence of positive selection on rodent HJURP (455) and mosquito CAL1(412) orthologs. Given their extreme divergence across clades, we conducted dN/dS analyses of candidate HJURP genes from selected taxa, each of which contains several genomes, to test for rapid evolution. Indeed we found that high nonsynonymous substitution rates are a general feature of this gene: dN/dS values were on average an order of magnitude higher for HJURP genes than for the functionally similar histone chaperones HIRA and ASF1 (464) in multiple metazoan lineages, as well as in fungi (Fig. 4.6E). Additionally, we found evidence of site-specific positive selection across primate evolution, reminiscent of previous observations in rodents (455), identifying HJURP as a gene under constant innovation across long evolutionary timescales (Fig. 4.6F). Much like HJURP in *Murinae*, these regions of selection were concentrated in the C-terminus and around the HMD region, suggesting persistent selective pressure across millions of years of evolution and implicating a functional role in these uncharacterized regions. This remarkably high nonsynonymous substitution rate may have accelerated HJURP's sequence divergence across several metazoan lineages, necessitating the use of remote homology detection methods employed in this study. Consistent with the scm3 domain as the most conserved functional domain, we did not identify residues under positive selection in the N-terminal region of primate HJURP.

4.4 Discussion

HJURP^{Scm3} is a fundamental component of the centromere identity-defining machinery, essential for sustained error-free cell division(385). Despite its cardinal function, HJURP has a paradoxically rapid evolutionary rate – so rapid that it was believed to be missing from many animal genomes(407). By combining PSI-BLAST, HMM-based searches, AlphaFold-multimer predictions, and functional genetic and cytological experiments, we significantly revise this model by demonstrating that *bona fide* Scm3/HJURP orthologs are in fact broadly retained across *Metazoa* but have been obscured by extreme sequence divergence. Our work on HJURP^{Scm3} supports the growing evidence that “losses” of essential genes more often reflect failures of homology detection(465–467) and proposes a structure-informed remote homology detection pipeline to uncover such “missing” genes (Fig. S4.3).

Recent advances in sequencing and structure prediction have made challenges in remote homology detection appreciably more tractable. However, remote homology detection efforts are still often aided by evolutionary intermediates that act as ‘stepping stones’, providing the necessary signal to connect distinct clades. Indeed, the choanoflagellate HJURP served as an intermediate, providing the link to establish orthology between yeast Scm3 and mammalian HJURP proteins. In this study, we found that genes from contentious “living fossil” genomes (468), an informal collection of slowly evolving genomes, such as those of the coelacanth and horseshoe crab, provided similar links for identifying or confirming HJURP orthologs. Even when the genes are unannotated, as is the case for coelacanth, horseshoe crab, *Hymenoptera*, or the nematode *Auanema*, they can provide the ‘stepping stone’ to establish homology. Intriguingly, the *L. chalumnae* hit initially served as such a critical intermediate despite itself lacking an intact scm3 domain.

HJURP is composed predominantly of polar residues and disordered domains, making it a poor candidate for purely structure-based searching. The combination of structure-informed alignments with HHpred and predictions with Alphafold-Multimer, however, greatly reduced the false discovery rate. Continued development of *in silico* screening methods that search for oligomer-specific conformations, such as that of HJURP bound to the CENPA/H4 dimer, will exponentially increase the throughput of this type of remote homology detection.

The use of model organisms to study the cell cycle has been paramount to understanding how and when cell-division perturbations manifest during development, how tissues idiosyncratically employ cell-cycle differences, and how germline tissues fundamentally differ from the soma. By identifying *bona fide* HJURP orthologs in the genetically tractable zebrafish and *C. elegans* models, our work opens avenues for understanding how centromeric identity differentially affects processes in multicellular organisms. For example, our work suggests that the *C. elegans* paternal genome, which lacks CENPA(453) in spermatocytes, reacquires CENPA in the oocyte from maternal machinery prior to the first zygotic division. A broader toolset of model organisms with more complete repertoires of centromeric protein orthologs will allow both rapid and mechanistic dissection of core centromeric functions, while also providing fodder for comparative analyses of the plasticity and evolvability of centromeric apparati across organisms.

Although HJURP sequences are predominantly diverged, the functional principles that govern HJURP are broadly conserved. Across distant lineages, HJURP orthologs preferentially bind CENPA–H4 over H3–H4, localize to centromeres in a cell-cycle-dependent manner, colocalize with CENPA, and are essential for chromosome segregation and development.

AlphaFold-Multimer shows that HJURP–CENPA–H4 complexes from fungi, zebrafish, humans, nematodes, and insects adopt strikingly similar three-dimensional folds, with conserved hydrophobic pockets and KY-centered motifs that maintain recognition of their respective CENPA proteins via their CATD domains despite minimal primary-sequence identity at contact residues. Other features, such as CENPC-binding(402, 413, 427) and DNA-binding(469–473), also appear to be conserved over long evolutionary timescales.

Despite this broad conservation, lineage-specific adaptations can also be found in HJURP proteins. For example, HJURP proteins from vertebrates, tetrapods, and teleost fish share additional motifs immediately upstream of the DNA-binding HMD region and within the Mis18-binding R2 region that are absent in echinoderms. Perhaps the most striking lineage-specific adaptations can be found in insect CAL1 proteins. Although the scm3 domain is the most conserved region across metazoans, it shows relatively low homology in *Diptera* compared with other arthropods. In contrast, a downstream region immediately C-terminal to the scm3 domain exhibits higher sequence conservation. Furthermore, *Diptera* have evolved a structurally distinct CENPC-binding domain at the C-termini of their CAL1 proteins (413). This finding, combined with the variable and incompletely defined contribution of CENPC to HJURP recruitment, even among tetrapods(401, 402, 474), exemplifies how centromeric machinery can be substantially rewired while maintaining functional fidelity.

HJURP genes evolve roughly an order of magnitude faster than other histone chaperones, such as HIRA and ASF1, in *Metazoa*. The previous inability to correctly identify HJURP orthologs in some metazoans is almost certainly a consequence of this more rapid evolution. Although the specific selective pressures driving HJURP’s diversification are still unknown, the accelerated rate of divergence we observe in metazoan scm3 domain-containing proteins, combined with their demonstrated direct role in modulating centromere identity, makes it very likely that these candidate HJURP proteins play an important role in suppressing “selfish” centromeric DNA during ‘centromere drive’(407, 459, 475). The centromere drive model posits that centromeric DNA variants bias their own transmission during asymmetric female meiosis, during which only one of four meiotic products is retained in the oocyte. Conversely, DNA-binding centromeric proteins, such as CENPA(460), diversify to restrict the biased transmission of “selfish centromeres” during female meiosis in both plants and animals.

We hypothesize that HJURP may evolve under dual pressures from centromere drive. First, HJURP’s distinct DNA-binding regions(470, 472, 476) might adapt to centromeric DNA changes to effectively modulate centromere identity. Second, HJURP must keep pace with CENPA’s rapidly evolving CATD region(438) to ensure high fidelity of its preferential interaction with CENPA/H4 over H3/H4. Consistent with this prediction, as recently described in rodents(455), we observe strong signatures of positive selection in primate HJURP, suggesting its role in mitigating this persistent genetic conflict. However, we and others(477) also identify rapid evolution in fungal Scm3, which is not thought to face the same selective pressures from centromere drive as animal HJURP(478), as fungi typically undergo only symmetric meiosis. Our findings suggest that evolutionary forces both during and beyond female meiosis drive the adaptive evolution of Scm3/HJURP chaperones. Our pipeline for identifying extremely remote homology enables an understanding of the essential molecular players in these ancient genetic conflicts and provides a means to connect distant homologies in centromeric biology and beyond.

4.5 Materials and Methods

4.5.1 HJURP Candidate Searching and Selection Criteria

A schematic representation of the approach is shown in Fig. S4.3. Initial HJURP candidates were identified with a combination of iterative BLAST searching (PSI-BLAST, expected threshold 5 and PSI-BLAST threshold 0.5)(418) and Hidden Markov Model searching (HMMsearch, default settings)(436, 437). Proteomes searched include non-redundant protein sequences from NCBI, the UniProtKB/TrEMBL complete database (<https://www.ebi.ac.uk/training/online/courses/uniprot-quick-tour/the-uniprot-databases/uniprotkb/>), Wormbase(479), and InsectBase(480). HMMs were generated from MUSCLE(351) alignments (using default settings) of representative tetrapod, Dipteran, and/or choanoflagellate/yeast sequences. Queried proteomes were searched via the Fred Hutchinson computational cluster implementation of HMMer (using default settings)(436). Initial HMM hits were not limited by any statistical cutoff, although common false-positive signals (including from an insect E4 esterase-like gene) were excluded; the intentional omission of a strict statistical cutoff allowed us to obtain a more holistic interpretation of remote homology data. Anecdotally, we found that strong hits were often within largely unstructured, polar protein sequences, although we did not formally use this criterion for filtering.

Following initial queries, sequences were analyzed in HHpred(422) to confirm regions structurally and sequentially homologous to published scm3 protein domains (PDBs 2L5A(273), 2YFV(389), 3R45(390), 6XWV and 6XWS(413)). HHpred E-values ranged widely (Table S1) but were generally below 1. These hits were subsequently used for trimeric AlphaFold-Multimer (using default setting)(332) predictions of the candidate scm3 domain with a single copy each of the respective species' canonical H4 and CENPA. Only annotated sequences with positive HHpred (default settings)(422) scm3 predictions, successful scm3-like structure predictions, and reciprocal PSI-BLAST or HMMer searches to published Scm3/HJURPs were used for downstream processing. HMMs were updated to include these new sequences, allowing for iterative searching.

The following candidate proteins are presented in the main text figures. **Figure 1:** *L. chalumnae* (XP_006012693.3, PSI-BLAST E-value 6e-04), *D. rerio* (NP_001076347.1, PSI-BLAST E-value 2e-06), *S. torazame* (GCB61218.1, PSI-BLAST E-value 5e-12), *P. marinus* (XP_032806972.1, PSI-BLAST E-value 0.2), *B. belcheri* (XP_019630933.1, PSI-BLAST E-value 0.03), *P. miniata* (XP_038057064.1, PSI-BLAST E-value 0.77). **Figure 3:** *N. vitripennis* (XP_001603895.3, HMMer E-value 0.99), *G. longicercus* (A0AAN9VWU2, HMMer E-value 2.7e-04), *D. pulex* (EFX88527.1, HMMer E-value 1.9e-06) *A. ventricosus* (GBM08647.1, HMMer E-value 0.46), *L. polyphemus* (XP_022237972.1, HMMer E-value 3.3e-03), *R. varieornatus* (GAU99461.1, HMMer E-value 1.3e-12). **Figure 4:** *Auanema* sp. *JU1783* (CAI4225445.1, PSI-BLAST E-value 0.44), *C. elegans* (NP_491161.2, PSI-BLAST E-value 7.0e-19). **Figure 6:** *P. caudatus* (XP_014675948.1, PSI-BLAST E-value 6.0e-03), *C. teleta* (ELT94853.1, PSI-BLAST E-value 4.0e-03), *B. glabrata* (KAI8798300.1, HMMer E-value 3.9e-15), *S. japonicum* (TNN07270.1, HMMer E-value 1.8e-12) *T. sp. H2* (RDD41302.1, HMMer E-value 1.9e-16), *P. clavata* (CAB3982230.1, HMMer E-value 2.0e-05), *G. barretti* (CAI8037030.1, HMMer E-value 3.5e-07).

A table of representative HJURP candidate proteins across Metazoa is provided in Table S4.1, and HMM model alignments are in Data S1 in online publication. Per-residue AlphaFold pLDDT (confidence) scores are mapped onto predicted structures in Supplemental Figure 4.12.

4.5.2 *Structure Analyses*

RMSD backbone calculations and structure alignments, contact determination, and visualizations were performed in UCSF ChimeraX-1.10(323) using the matchmaker tool.

4.5.3 *Sequence Alignments*

Protein sequence alignments were generated with the MUSCLE algorithm (default settings) in Geneious. For HMM model building and scm3 domain visualization, alignments were extensively manually trimmed to remove species-specific insertions. Logo plots(481) were generated from trimmed alignments and colored by hydrophobicity in Geneious. Frequently, the only region successfully aligned across order-level alignments was the N-terminal scm3 region, and it was often the only region used for further analyses. Protein alignments are available in Data S2 in online publication.

Codon alignments were performed using the webPRANK(482) server on whole-gene coding-sequence data from NCBI and were subsequently trimmed manually. Sequences containing large insertions or with exceptionally long phylogenetic branches were excluded from final analyses. Only clades with at least 8 sequences were used for alignments and analysis of evolutionary selective pressures. Codon alignments are available in Data S3 in online publication.

4.5.4 *Phylogenetic and dN/dS Analyses*

All phylogenetic trees were generated via the IQ-Tree(352, 378) web server, allowing for free-rate heterogeneity and using optimized substitution models, and then visualized in FigTree v1.4.4 (<https://github.com/rambaut/figtree>). The phylogenies presented in Figures 1 and 4 were generated from the scm3 domain alignment alone, whereas the tree in Figure 4.3 was generated from whole-gene alignments.

Whole-gene dN/dS calculations and site-specific selection tests were performed using maximum-likelihood estimation with the codeml algorithm in PAML (308). Model 0 was used to calculate whole-gene dN/dS. Site-specific positive selection was calculated as previously described(379) for primate HJURP. Briefly, models that either include (models 2, 8) or exclude (model 1, 8a) a category allowing dN/dS > 1, indicating positive selection, were compared. Codons with Bayes Empirical Bayes scores above 0.9 for model 8 were highlighted as evolving under positive selection.

4.5.5 *Recombinant Protein Expression and Purification*

For MBP-drHJURP200 expression, the first 200 amino acids of zebrafish HJURP were amplified from 2dpf cDNA and cloned into an expression vector with an N-terminal MBP tag and C-terminal 6XHIS tag. Protein expression and purification were carried out largely as in(426) with slight modifications. Briefly, BL21 DE3 LOBSTR cells were transformed with the drHJURP expression vector and grown in Luria Broth with ampicillin at 37°C to an optical density (OD) of 0.6, after which they were induced with 1mM of IPTG and shifted to 18°C with shaking for overnight protein expression. The following day, cells were resuspended in HisTrap

binding buffer (20mM Na₃PO₄, 500mM NaCl, 20mM imidazole pH 7.4) with a cOmplete protease inhibitor tablet (Roche) and lysed via sonication. Following clarification at 30000 × g for 20 minutes at 4°C, the soluble lysate was run over a 5mL HisTrap column equilibrated in binding buffer. The column was washed with 50mL of binding buffer, and protein was eluted with HisTrap elution buffer (20mM Na₃PO₄, 500mM NaCl, 500mM imidazole, pH 7.4). The eluate was then further purified via preparative size exclusion chromatography on a Superdex 200 Increase 10/300 column (Cytiva) in a buffer with 25mM Tris-HCl pH 7.5, 300mM NaCl, 5mM β-mercaptoethanol. Fractions containing pure protein in 10% v/v glycerol were snap-frozen in liquid nitrogen and stored at -80°C. A similar approach was used to prepare the NEFR-1 construct, which consisted of amino acids 68-196 of *C. elegans* NEFR-1, except that the HisTrap eluate was further purified using a Superdex 75 Increase 10/300 column.

Soluble drCENPA/H4 tetramers were purified as an adaptation of Servin and Straight(483). drCENPA was amplified from 2dpf zebrafish embryos and cloned into a bicistronic vector encoding both drCENPA and H4. The vector was transformed into BL21 DE3 cells, grown to an OD of 0.6 at 37°C, then induced to express the tetramer with 1mM IPTG. Following an additional 5 hours of growth, cells were harvested and lysed in histone binding buffer (20mM Tris-HCl pH 7.4, 750mM NaCl, 10mM β-mercaptoethanol, 0.5mM EDTA) with a cOmplete protease inhibitor tablet (Roche). The lysate was cleared at 30000 × g for 20 minutes at 4°C, then bound to a HiTrap SP FastFlow column and washed with binding buffer. The tetramer was eluted with a 20-column-volume gradient into binding buffer containing 2 M NaCl. Fractions containing tetramer were further purified via preparative size exclusion chromatography on a Superdex 200 Increase 10/300 column (Cytiva) in histone binding buffer. Fractions containing pure tetramer were pooled, concentrated, snap-frozen in 10% v/v glycerol, before being stored at -80°C. A similar approach was used to prepare soluble HCP-3/H4 tetramers, except that a truncated form of HCP-3, residues 148-288, was used. Soluble H3/H4 tetramers were purchased commercially (New England Biolabs #M2509).

4.5.6 Amylose Pulldown

Amylose resin pulldowns were performed as in Hori et al.(427) with amylose resin replacing GST resin. 50pmol of MBP-drHJURP or MBP-NEFR-1 was incubated with 30μL of amylose resin equilibrated in pulldown buffer (20 mM HEPES-NaOH, pH 7.4, 750mM NaCl, 0.05% NP-40, 5mM DTT) for 1 hour at 4°C. The resin was then washed three times in pulldown buffer and incubated with 25 pmol of drCENPA/H4, HCP-3/H4, or H3/H4 tetramer, as indicated. The resin was washed three more times in pulldown buffer, then proteins were eluted in 30μL of SDS-PAGE buffer and analyzed via SDS-PAGE stained with GelCode Blue.

4.5.7 Zebrafish Care and Maintenance

Danio rerio animals were maintained at the Fred Hutch Cancer Center in compliance with Institutional Animal Care and Use Committee (IACUC)-approved protocols, and experiments were approved by and conducted in accordance with Fred Hutchinson Cancer Center IACUC standards. Zebrafish were bred and maintained according to an established protocol(484). Sex is not a relevant variable in our experiments, which were performed before the onset of sex determination in zebrafish(485). No transgenic zebrafish lines were used.

4.5.8 Zebrafish CRISPRant Generation

CRISPRant(486) mutant alleles were generated using CRISPR/Cas9 with standard protocols (487). gRNAs were designed using chopchop (<https://chopchop.cbu.uib.no>), ordered as Alt-R sgRNAs (IDT), and pooled in a 1:1 manner for each of three conditions: 1) Control guide(486) with two HJURP scramble guides, 2) three drHJURP guides, and 3) two drCENPA guides. One-cell-stage embryos were injected with 1nL of 10 μ M pooled gRNA and 5 μ M Cas9 protein (IDT, #1081058), and F0 mutant animals were sampled for editing efficiency via PCR. Guide RNAs and genotyping primers used in this study are provided in Table S4.2. Unfertilized embryos were removed after 2hpf, and CRISPRant morphology was scored on 24hpf larvae by eye compared to uninjected control individuals. “Necrotic” individuals had wild-type morphology but noticeable darkened opaque tissue, while “morphological defects” were marked by differences in body shape compared to uninjected references.

4.5.9 Zebrafish Fluorescence Microscopy

For colocalization experiments, drHJURP and drCENPA coding sequences were amplified from 2dpf larval cDNA and cloned into fluorescent tag backbones via standard cloning techniques. mRNAs were generated from template plasmid DNA via *in vitro* transcription (mMessage mMachine, Invitrogen) and purified via gel extraction. 100pg of each purified mRNA was injected into one-cell-stage embryos in a volume of 1nL(488). For CRISPRant analyses, 1nL of 100ng/uL of H2B-mCherry mRNA was injected together with 5 μ M Cas9 protein and 10 μ M of pooled HJURP or control CRISPR guides. For CENPA intensity calculations, these injections were supplemented with 100 ng/uL eGFP-drHJURP. Construct sequences and amplification primers for fluorescent proteins are provided in Table S2.2.

6hpf individuals were embedded in 1.0% low-melt agarose and imaged on a Leica SP8 inverted confocal microscope using a 28°C heated stage and 1.5% laser power at 488nm and 555nm. Stacks with a 1 μ m step size were collected for 5 hours at 6-minute intervals. Representative still frames of different cell cycle stages are shown in Figure 4.2. Images were processed in FIJI(489). For anaphase fidelity quantification, anaphases from time-lapse images were scored as “erroneous” if they included lagging or tethered mCherry-H2B signal between segregating chromatin volumes at the second frame of anaphase onward, while all others were considered wildtype. For drCENPA intensity quantification, 10-image stacks from 10 hpf time-lapse images were max intensity projected, equally thresholded on eGFP intensity to remove background nuclear signal and only isolate centromeric signal, and measured as a ratio of eGFP-drHJURP:mH2B-mCherry per nucleus.

4.5.10 C. elegans Maintenance and Genetics

C. elegans was maintained at 20°C on standard nematode growth media (NGM) and fed *E. coli* OP50, unless otherwise noted. Genotyping primers used in this study are provided in Table S4.2. A list of strains used in this study is provided in Table S4.3.

For *nefr-1* K15STOP embryo viability counts, genotypes were tracked via the presence of the balancer tmC20 which causes GFP expression in the pharynx of heterozygotes and a dumpy phenotype with GFP in homozygotes with wildtype *nefr-1*. Single heterozygotes were allowed to lay eggs overnight and then removed from the plate the following morning. Following a 48-hour

hatching period, worms were genotyped and scored for developmental stage based on standard criteria. Each day following, worms were scored on both developmental stage and survival until 7 days post-laying.

For auxin hatching experiments, three NJL5706 L4 worms were placed on NGM + OP50 plates layered with 500 μ L 500 μ M K-NAA (α -Naphthaleneacetic Acid Potassium Salt) in M9 medium or control plates with 500 μ L of M9 alone. The next afternoon, worms were transferred to fresh plates containing auxin or control solutions and allowed to lay eggs overnight. The next morning, adult worms were removed from the plates, and their eggs were allowed to hatch for 48 hours. Plates were scored on the number of eggs that hatched.

4.5.11 *C. elegans* CRISPR/Cas9 Genome Modification

CRISPR/Cas9 genome editing was performed by microinjection of guide RNA/Cas9 ribonucleoprotein (RNP) complexes (IDT #1072532, #1081058) and custom crRNA along with either ssDNA oligos (IDT) or partially-melted dsDNA oligos as homology-directed repair (HDR) templates(490, 491). See Table S4.2 for guide and repair constructs used for microinjection.

4.5.12 *C. elegans* Fluorescence Microscopy

C. elegans immunofluorescence samples were prepared as described(492, 493) with slight modifications. Gravid adults were dissected into a 15 μ L drop of M9 solution on charged slides (ProbeOn Plus, Fisher) and fixed with a 15 μ L drop of 2% PFA (Sigma) in M9 for 5 minutes. As much liquid as possible was wicked away with blotting paper. The drop was covered with a 1.5 coverslip, and the slide was frozen in liquid nitrogen for 5 minutes. The slide was then removed, the slip quickly cracked off with a razor blade, and the slip was dipped into ice-cold methanol for 1 minute. Slides were then synchronized in PBS-T (PBS with 0.1% Triton X-100). Following synchronization, slides were washed twice in PBS-T, then blocked with a 100 μ L drop of AbDil (PBS pH 7.4, 4% BSA, 0.1% Triton X-100, 0.02% NaN₃) at room temperature for 1 hour. Samples were incubated overnight in 100 μ L of primary antibody diluted in AbDil at 4°C. The next day, slides were washed 3 times in PBS-T, then incubated with 100 μ L of secondary antibody for 1-2 hours at room temperature. Slides were washed 3 times in PBS-T and mounted with 8 μ L of DAPI-containing antifade media. The following antibodies were used: anti-FLAG M2 (1:500, Sigma F1804), anti-HA tag (1:250, Abcam ab236632), anti- α Tubulin (1:250, Abcam ab7291), anti-Rabbit AlexaFluor 555 (1:500, Invitrogen A-21428), anti-Mouse AlexaFluor 488 (1:500, Invitrogen A-11001).

Exposure-standardized images from samples were collected on a Leica STELLARIS8 with Leica Application Suite X software. Stacks of approximately 30 μ m were collected with an oil objective at either 40X (germlines) or 63X (eggs/embryos). Images were processed in FIJI(489), in which color channels were separated, recolored, and merged; DAPI channels were scaled for presentational clarity, and germlines were computationally straightened. All images are maximum-intensity projections of z-stacks that encompass the entire germline or embryo.

4.5.13 *C. elegans* NEFR-1 Immunoprecipitation

Immunoprecipitations from worm embryo lysate were performed as described(494) with slight modifications. Worms from mixed-stage populations were bleach-synchronized and hatched for 24 hours in liquid M9. Synchronized worms were then plated on 20X-concentrated OP50 for large-scale growth and allowed to progress to the gravid adult stage for 3 days. Worms were then bleached, washed, pelleted, and flash-frozen in liquid nitrogen for downstream processing.

Worm pellets were thawed on ice and resuspended 1:0.5 in 1.5X lysis buffer(495) (50mM HEPES pH 7.5, 1mM EGTA, 1mM MgCl₂, 150mM KCl, 15% glycerol, 0.1% NP-40, 0.75mM DTT, cOmplete protease inhibitor tablet, 1X Roche PhosSTOP). Worm suspensions were lysed via sonication for fifteen 30s on/30s off cycles using a Bioruptor Pico sonication device (Diagenode). Lysates were cleared at 19000 × g for 20 minutes at 4°C. Clarified supernatants were transferred to fresh 1.5mL tubes and kept on ice for all subsequent steps. Samples were diluted in lysis buffer to 10mg/mL in 200 µL aliquots and incubated for 2 hours with 100 µL of lysis buffer-equilibrated magnetic anti-FLAG M2 beads (Sigma M8823). Following incubation, beads were isolated with a magnetic stand, washed 3X in wash buffer and 2X in 50mM HEPES pH 7.5, and boiled 1:1 in SDS-PAGE buffer for downstream mass spectrometry analyses.

Mass spectrometry data collection and processing were performed at the Fred Hutchinson Proteomics and Metabolomics core facility. Samples were desalted and run on an Orbitrap Ascend. The raw mass spec data were searched using Proteome Discoverer (PD) version 3.1 against a Uniprot C. elegans database, along with common contaminants. The search results were filtered for high confidence, 1% false discovery rate at the peptide level. The ratio of the sum PEP score was used to determine the relative representation of peptides within the 3xFLAG::NEFR-1 strain compared to an untagged control (Table S4 in online publication). Peptides not found in the negative control were assigned a sum PEP score of 0.5 for comparison. All ratios were normalized to vitellogenins, the most common contaminant in both samples.

4.6 Supplemental Information

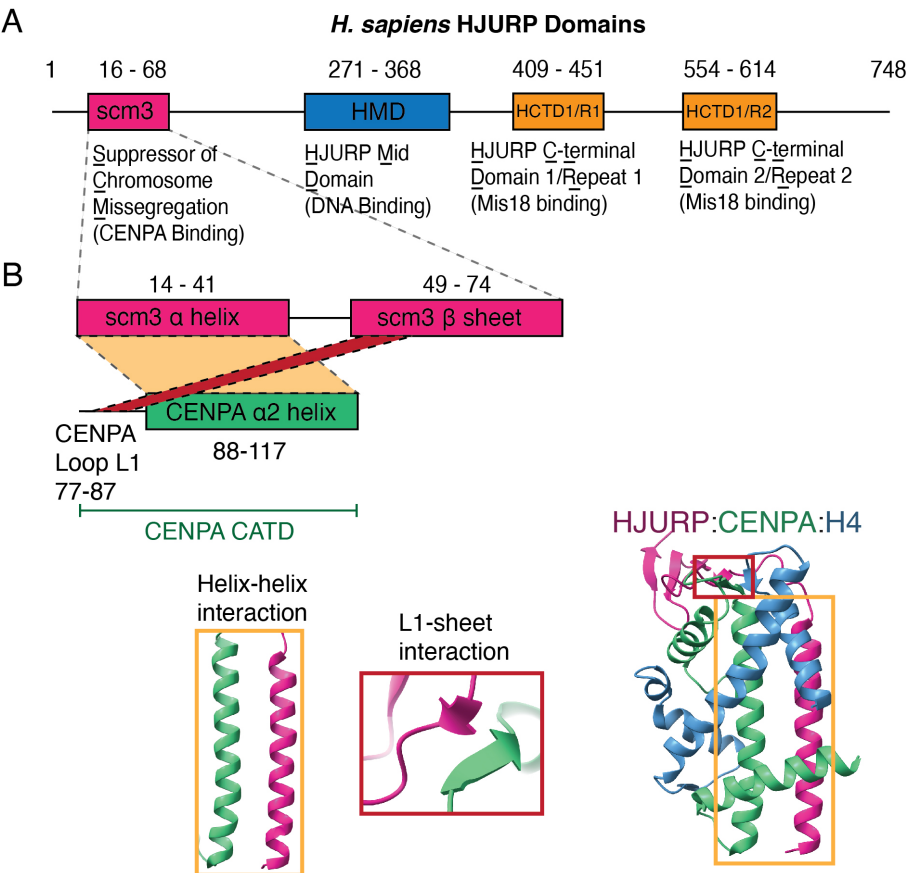


Figure S4.1. Structure-function relationships of the human HJURP protein. **A.** A domain schematic for human HJURP. The suppressor of chromosome missegregation (scm3) domain directly binds CENPA/H4 dimers, the HJURP mid domain (HMD) is implicated in DNA binding, and HJURP C-terminal domains (HCTD) 1 and 2 interact with Mis18 to recruit HJURP to the centromere. **B.** A structural overview of human HJURP-CENPA interactions. CENPA forms an antiparallel helix and strand with HJURP. Structure is PDB 3R45(390).

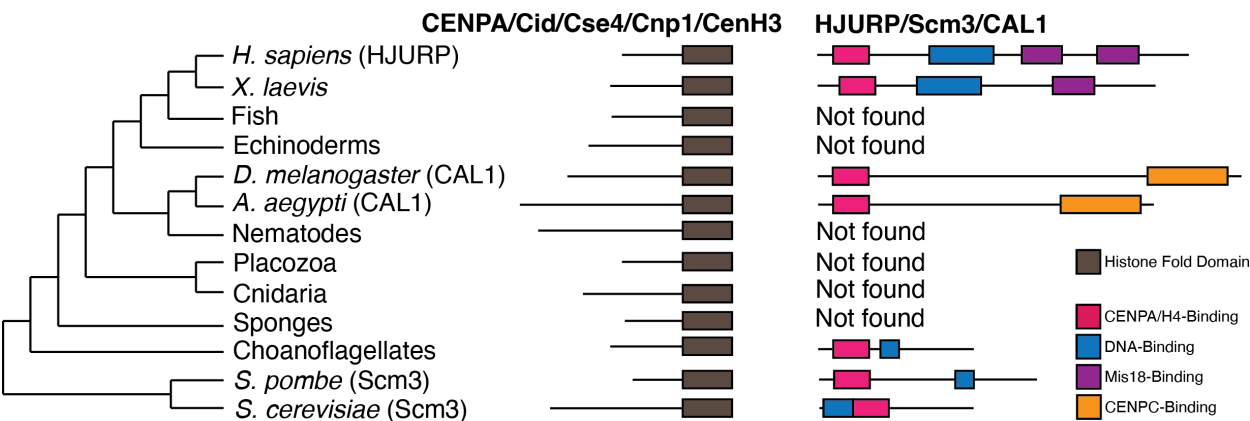


Figure S4.2. HJURP orthologs had not previously been found in most metazoan clades. **A.** An unscaled phylogeny showing relationships among major metazoan groups. Select scm3 domain-containing CENPA chaperones have been independently identified and biochemically characterized across the animal tree, yet many species with CENPA appear to lack a scm3 chaperone.

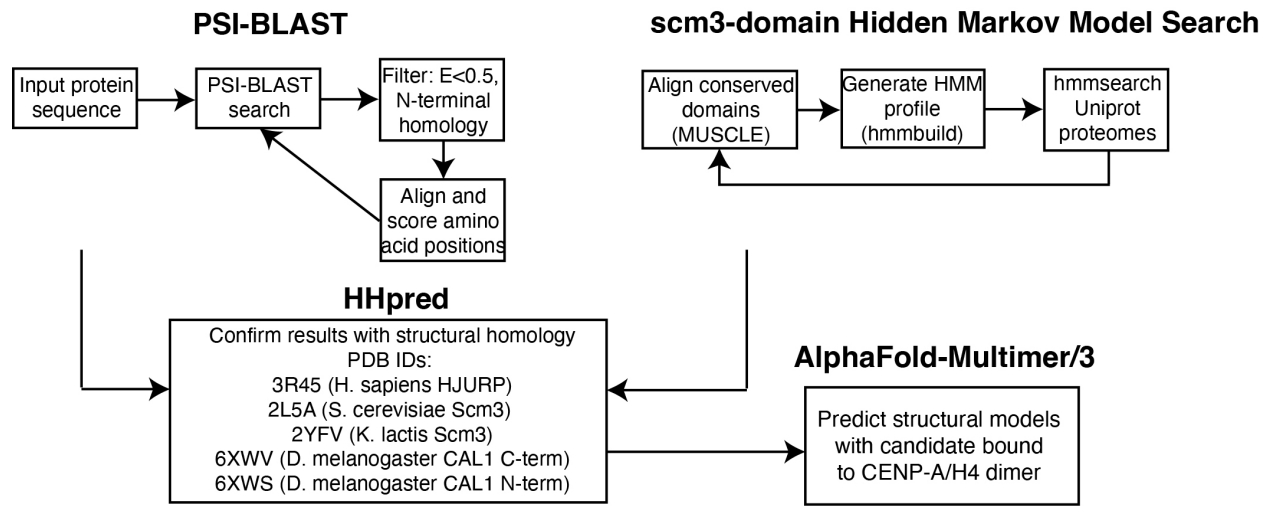


Figure S4.3. A schematized approach for detecting remote HJURP orthologs. We employed a holistic homology search pipeline integrating sequence and structural data to identify and confirm distant Scm3/HJURP orthologs that previously escaped detection.

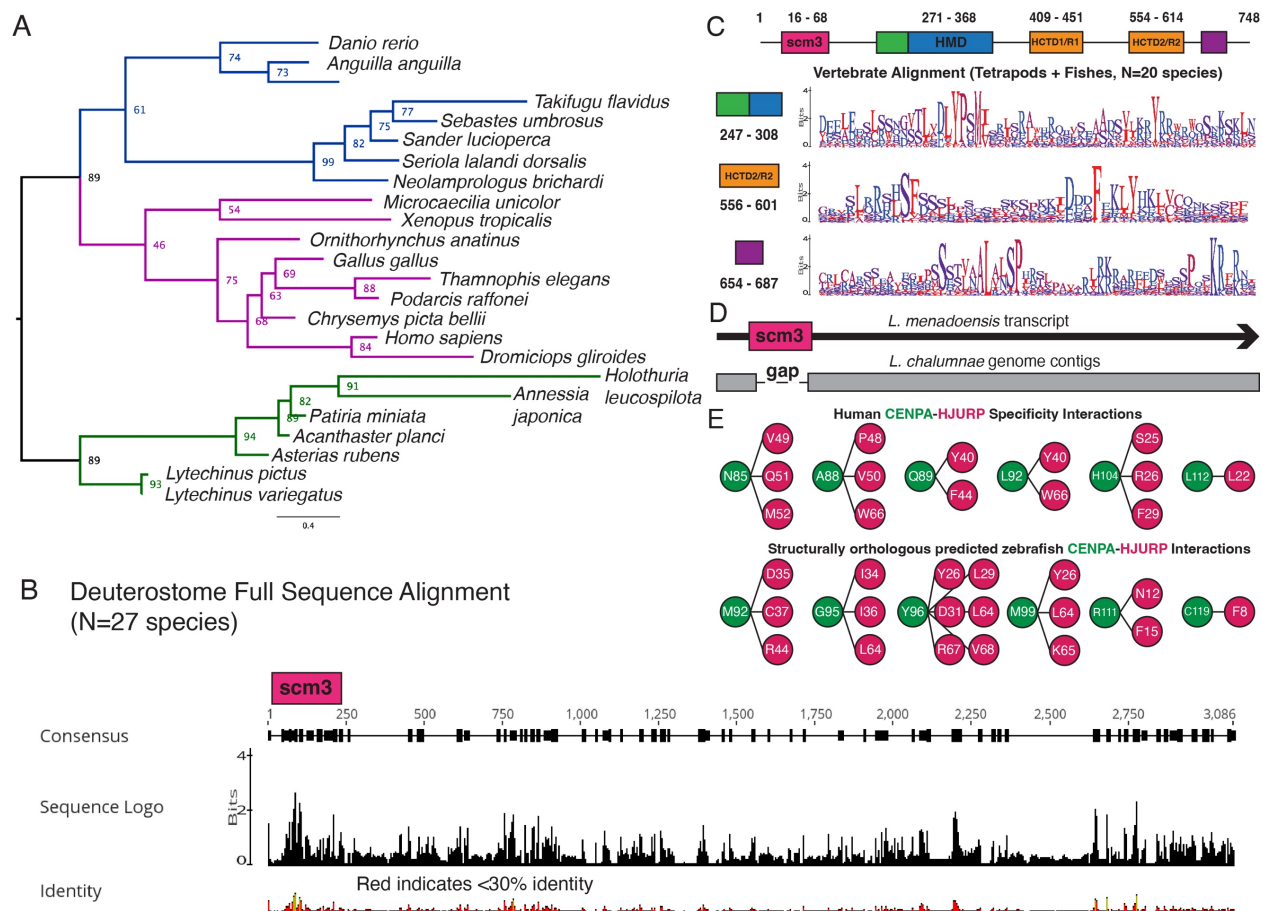


Figure S4.4. Shared and diverged features of deuterostome HJURPs. **A.** An expanded, duplicate phylogeny from Fig. 4.1 of representative deuterostome scm3 domains roughly follows the expected species tree, consistent with a monophyletic origin of HJURP in this clade. Nodes are presented with 1000 ultrafast bootstrap values. **C.** A sequence alignment of full-length putative deuterostome HJURP proteins shows generally poor conservation aside from the N-terminal scm3 domain region. **C.** A domain schematic for human HJURP with additional regions of shared homology between tetrapods and fish displayed as logo plots. Besides the previously identified HMD and HCTD2 domains, we find shared homology N-terminal to the HMD and close to the C terminus. **D.** The *L. chalumnae* genome likely contains a contig gap at the scm3 domain; alignments of translated contig regions to the *L. menadoensis* transcriptome shows coelacanths encode scm3 domain-containing HJURP orthologs. **E.** Structurally orthologous residues that contribute to CENPA-HJURP specificity differ almost entirely between zebrafish and human. Residues are shown that form physical interactions in either the human CENPA/H4/HJURP crystal structure or the predicted zebrafish CENPA/H4/HJURP structure.

A

drHJURP Positively-correlated gene expression	
Gene	r-value
si:ch211-69g19.2	0.291
tpx2	0.289
cdk1	0.285
plk1	0.283
cenpf	0.282
cdc20	0.28
ube2c	0.28
ccnb1	0.279
mki67	0.279
nusap1	0.278
kpna2	0.277
aspm	0.273
top2a	0.273
kifc1	0.271
mad2l1	0.271
kif11	0.268
aurkb	0.265
aurka	0.263

B

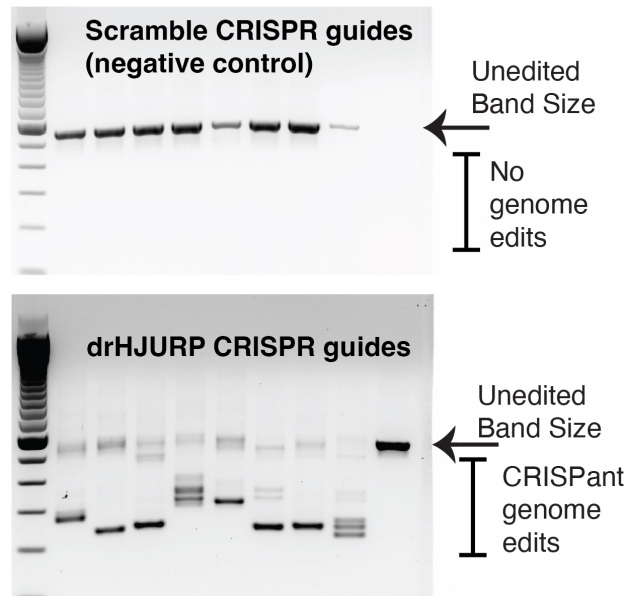


Figure S4.5. Related to *in vivo* analysis of drHJURP. **A.** The expression of the putative zebrafish HJURP is most positively correlated with primarily cell cycle genes such as cell cycle kinases (plk1, aurka/aurkb, cdk1) and spindle microtubule regulators (tpx2, aspm, kif11), among others. Data are provided via Daniocell(430). **B.** Zebrafish CRISPR guides efficiently cause truncation at the HJURP locus (bottom) but non-targeting scrambled control guides fail to do so (top). Bands are PCR products amplified from the HJURP locus of individual CRISPRant embryos; farthest right band is amplified from an uninjected control.

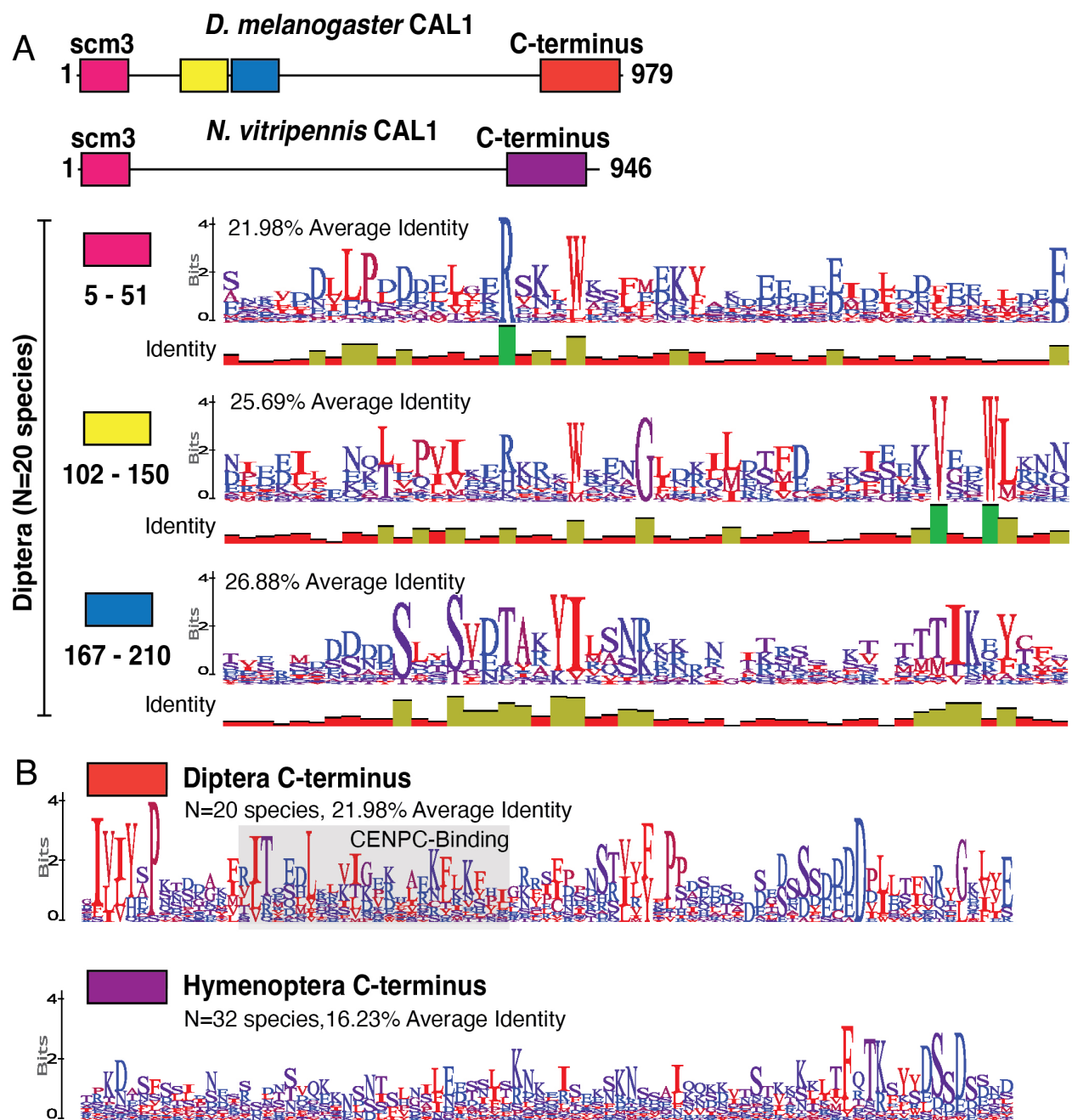


Figure S4.6. Shared and diverged sequence features of CAL1. A. A domain schematic for *Drosophila* and *Nasonia* CAL1 with additional regions of *Drosophila* homology beyond the scm3 domain displayed as logo plots. B. The Hymenopteran CAL1 C-terminus is less constrained than that of Dipteran CAL1.

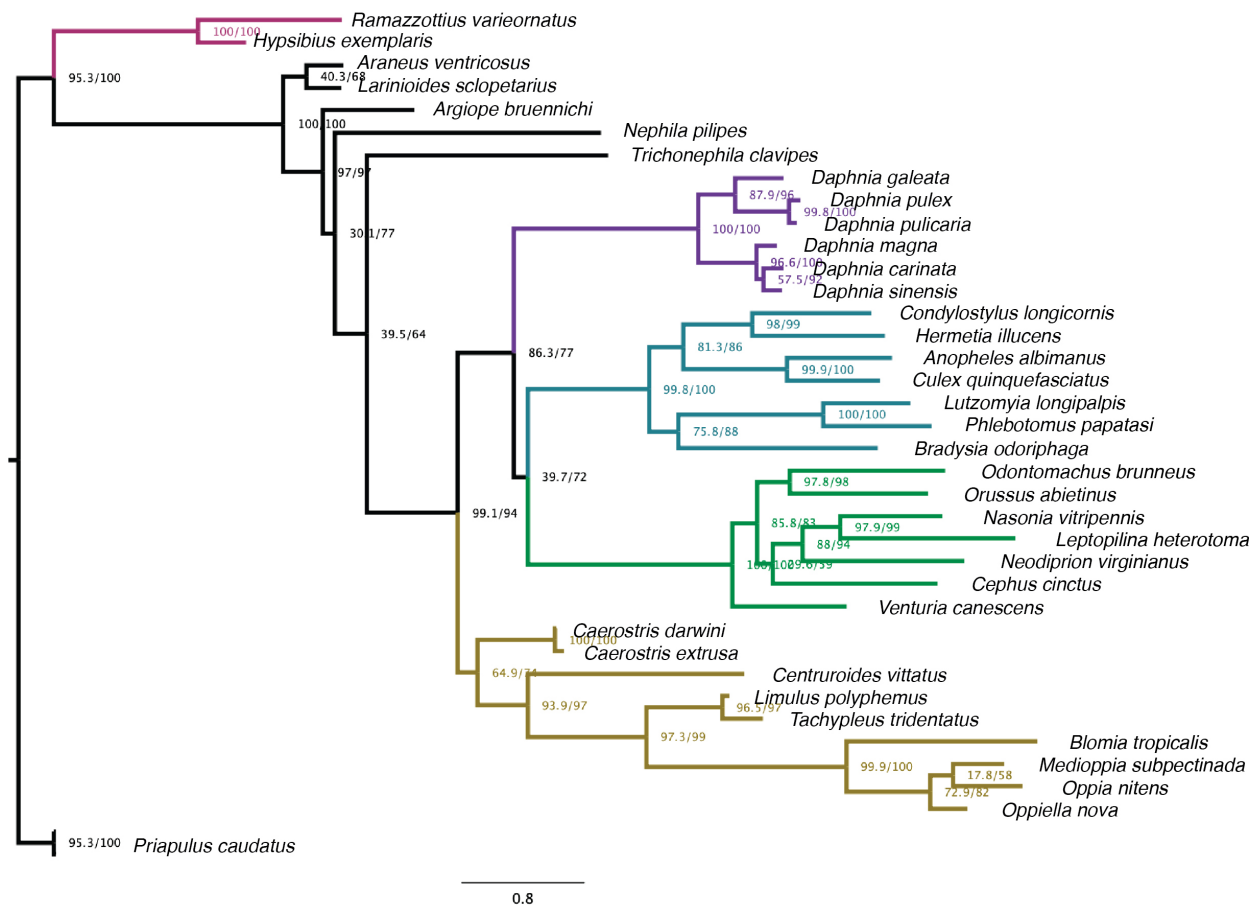


Figure S4.7. An expanded phylogeny of panarthropod HJURP orthologs. An expanded version of the tree in Fig. 4.3E generally follows the accepted species tree, although low homology and sparse sequence availability likely contributes to phylogenetic artefacts. Nodes are presented with 1000 ultrafast bootstrap/SH-aLRT replicate values.

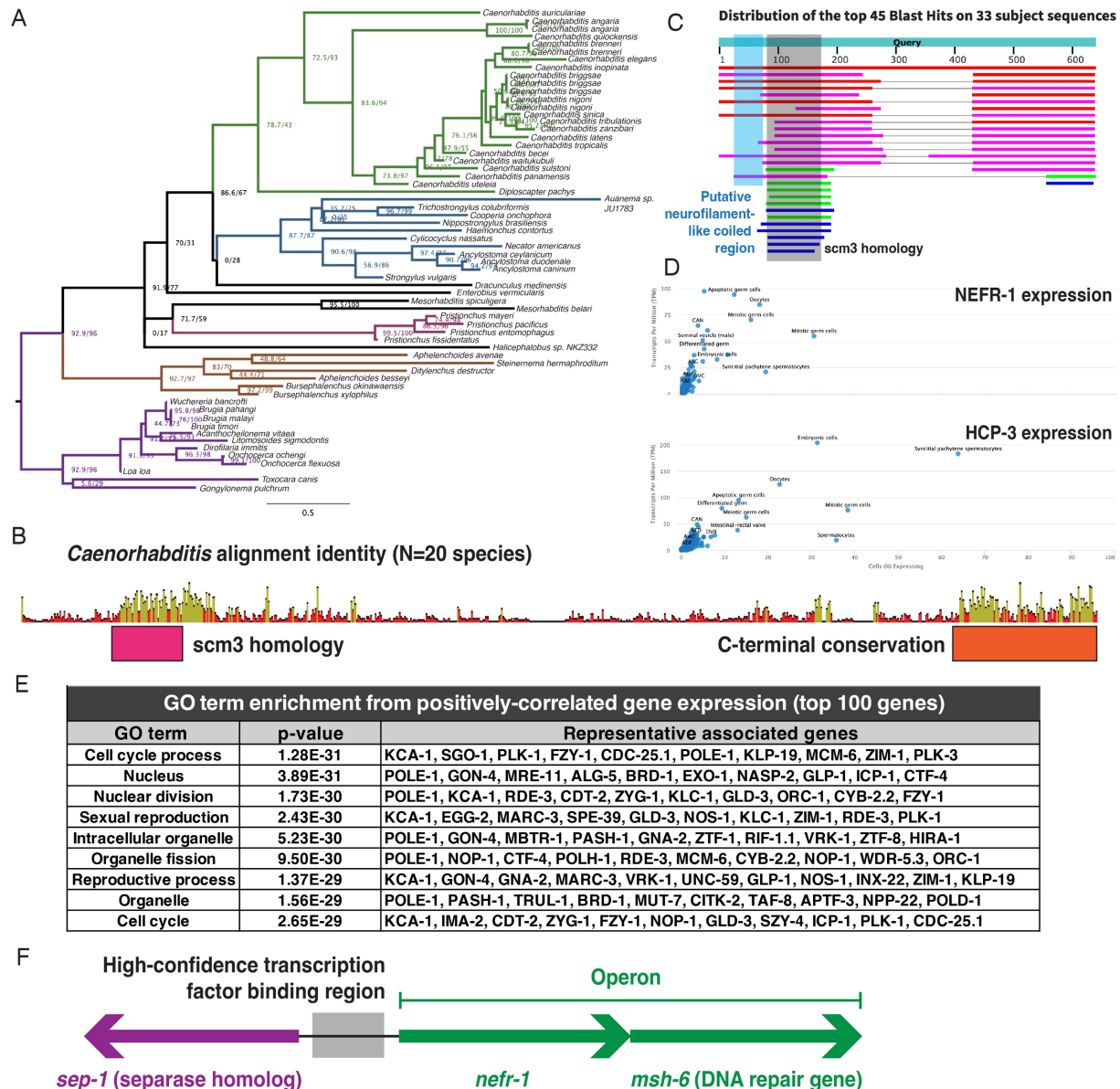


Figure S4.8. NEFR-1 is a nematode cell cycle protein. **A.** An expanded nematode phylogeny presented in Fig. 4.5B roughly follows the species tree, consistent with a monophyletic origin of HJURP^{NEFR-1} in this clade. Nodes are presented with 1000 ultrafast bootstrap/SH-aLRT replicate values. **B.** Conservation pattern of NEFR-1 across *Caenorhabditis* shows there is relatively little sequence conservation outside of the N- and C-terminal regions. **C.** BLAST results show that the scm3 homology and C-terminal regions, rather than the putative neurofilament-like coiled coil region, are the most conserved within nematodes. **D.** NEFR-1, like the worm CENPA homolog HCP-3, is most highly expressed in germline dividing tissue such as oocytes, syncytial spermatocytes, and embryonic cells. Data are from single-cell expression profiles(446). **E.** NEFR-1 has positive expression correlations with predominantly cell cycle genes. Data on the top 100 genes with positive expression correlation to *nefr-1* and their gene ontology (GO) analyses were provided via the WormBase(479) implementation of SPELL(496). **F.** Schematic of the *nefr-1* syntenic neighborhood suggests *nefr-1* is expressed with other chromatin-associated cell cycle genes like *sep-1*, required for anaphase chromosome segregation, and *msh-6*, a DNA mismatch repair gene.

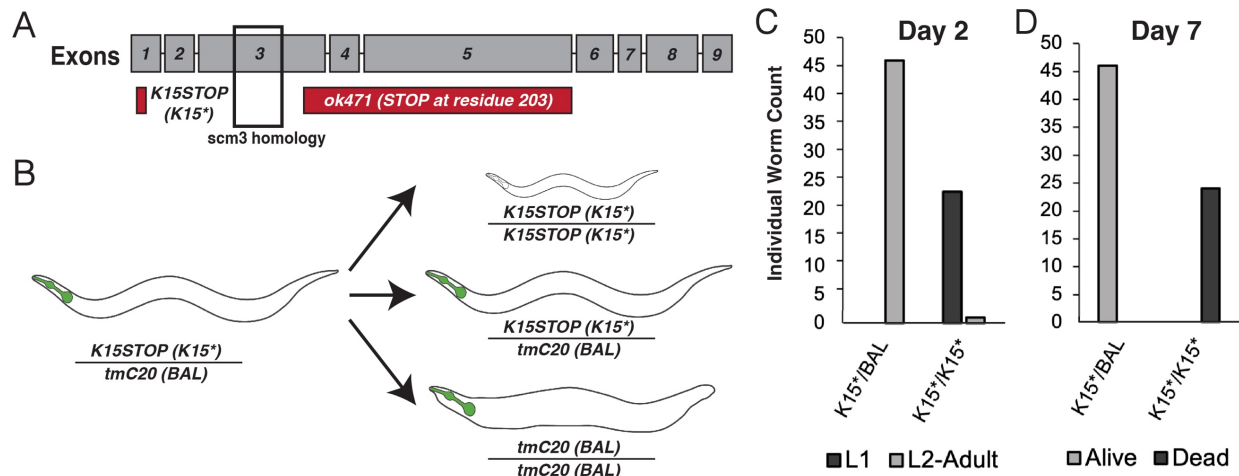


Figure S4.9. NEFR-1 is an essential gene. **A.** Gene structure of *nefr-1* shows that the previously characterized truncation allele *ok471* retains the *scm3* domain, while the *K15** allele generated in this study, which includes a subsequent frameshift, produces a complete knockout. **B.** Balancer crossing scheme to assess viability of *nefr-1*(*K15**) mutants. The *tmC20* balancer confers a GFP pharynx phenotype to heterozygotes and a GFP pharynx + dumpy phenotype to balancer homozygotes. **C.** Worms homozygous for *K15** fail to progress past the L1 stage and **D.** die within a week, while heterozygotes mature and survive into adulthood.

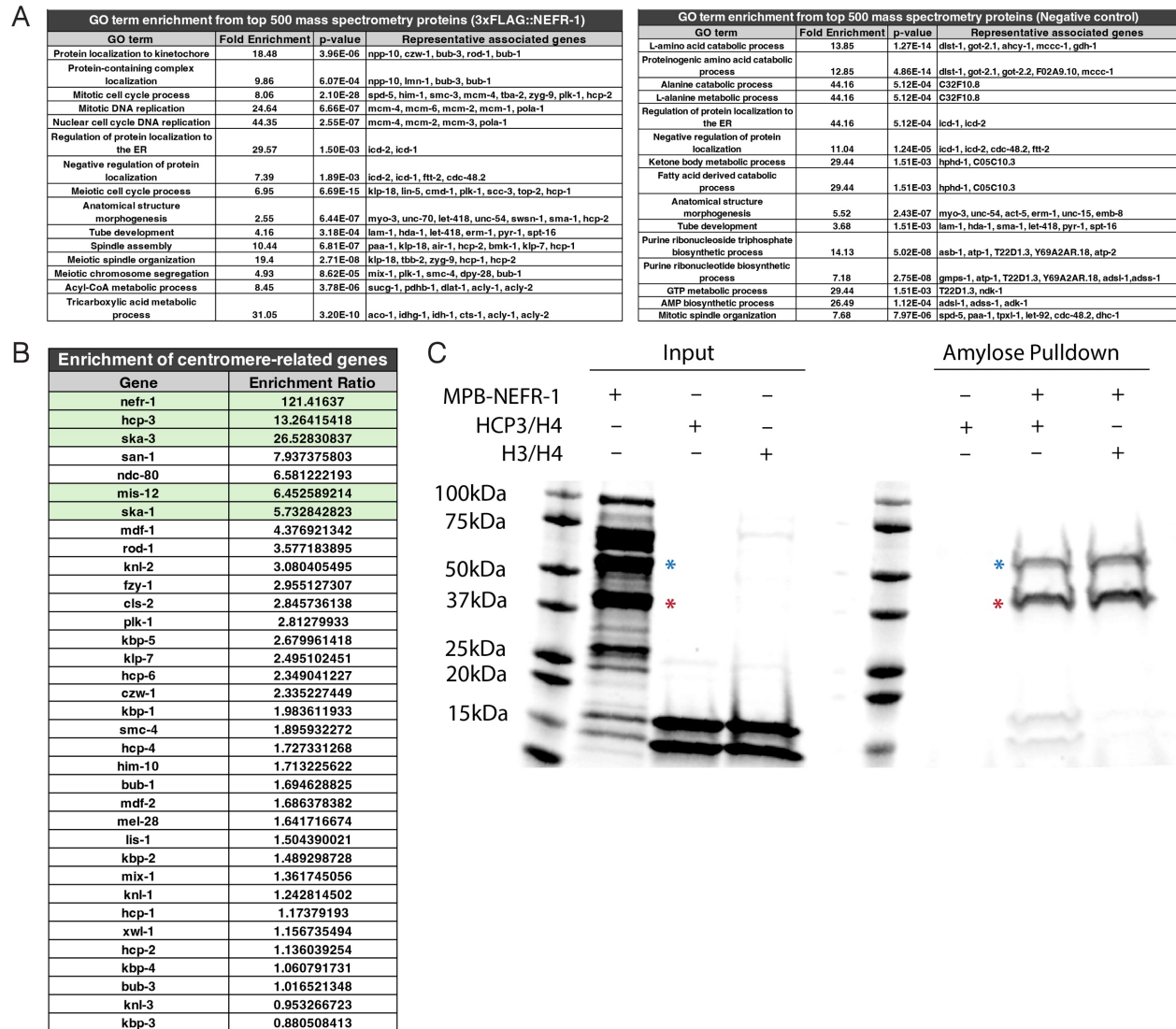


Figure S4.10. NEFR-1 physically associates with CENPA^{HCP-3}. **A.** Top 15 gene ontology (GO) analysis terms of mass spectrometry peptides identified in anti-FLAG immunoprecipitations from worm embryo lysate of either *3xFLAG::nefr-1* (left) or untagged control (right) strains. Tagged strains show specific enrichment of kinetochore and cell cycle-related genes. **B.** Normalized sum PEP score enrichment ratio of centromere-related genes (*441*) in *3xFLAG::nefr-1* immunoprecipitations over untagged controls. Genes highlighted in green, including CENPA^{HCP-3} and NEFR-1, were identified exclusively in *3xFLAG::nefr-1* immunoprecipitations. **C.** Amylose pulldown assay with purified MBP-NEFR-1 and HCP-3/H4 or canonical H3/H4 tetramers. MBP-NEFR-1 interacts directly with HCP-3/H4 tetramers as suggested by co-precipitation of species at HCP-3 and H4 sizes. Blue asterisk indicates MBP-NEFR-1; red asterisk indicates a non-NEFR-1 contaminant.

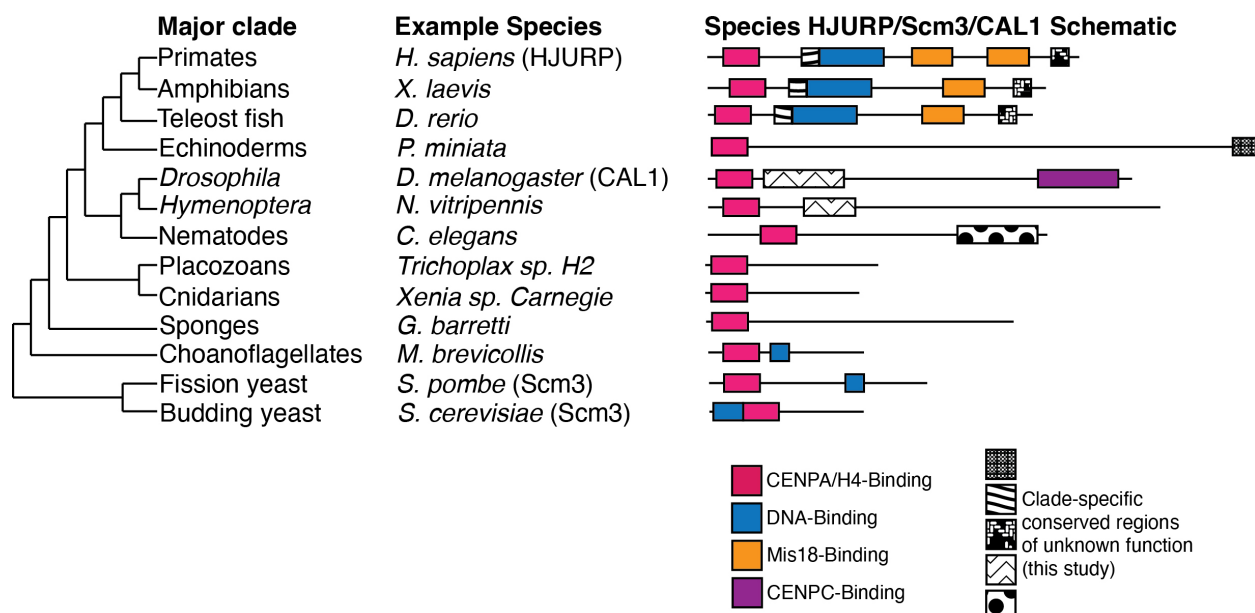


Figure S4.11. scm3 domain-containing proteins are present in most metazoan clades. We identify candidate HJURP orthologs across the metazoan tree and describe conserved domains previously unreported in these proteins that have yet to be functionally characterized.

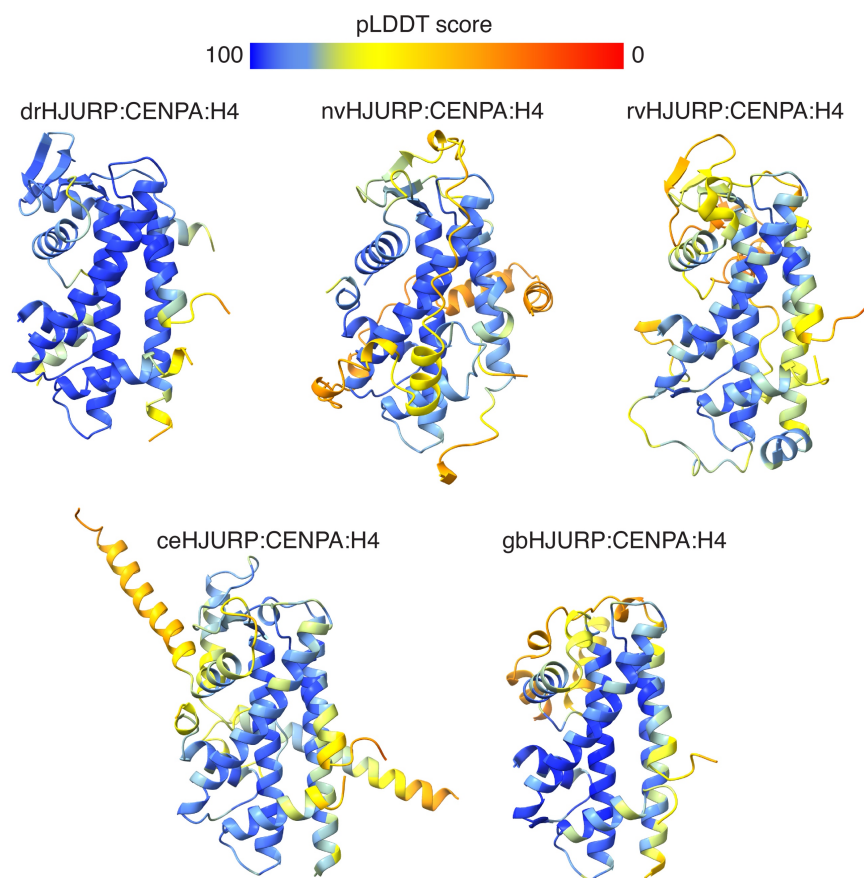


Figure S4.12. AlphaFold predictions colored by pLDDT score. All AlphaFold predictions presented in this study showed moderate to high confidence as measured by pLDDT (>80 at helical regions).

Table S4.1. A representative list of scm3 domain-containing proteins across the opisthokont lineage.

Accession Identifier	Source	Species	Taxonomy	Protein Length (AAs)	scm3 region	HHpred E-value (whole protein)	HHpred E-value (scm3 domain)	
AAH01940.2	NCBI	<i>Homo sapiens</i>	Eukaryota; Metazoa; Chordata; Cra	748	16-68	N/A	N/A	
XP_031752846.1	NCBI	<i>Xenopus tropicalis</i>	Eukaryota; Metazoa; Chordata; Cra	704	24-75	4.00E-08	6.70E-13	
XP_006012693.3	NCBI	<i>Latimeria chalumnae</i>	Eukaryota; Metazoa; Chordata; Cra	642	N/A	N/A	N/A	
NP_001076347.1	NCBI	<i>Danio rerio</i>	Eukaryota; Metazoa; Chordata; Cra	669	3-55	0.21	2.30E-04	
XP_038057064.1	NCBI	<i>Patiria miniata</i>	Eukaryota; Metazoa; Echinodermat	1372	16-64	0.38	6.60E-05	
XP_014675948.1	NCBI	<i>Priapulius caudatus</i>	Eukaryota; Metazoa; Ecdysozoa; Sc	1756	4-50	0.12	6.20E-06	
CAI4225225.1	NCBI	<i>Auanema sp. JU1783</i>	Eukaryota; Metazoa; Ecdysozoa; N	1146	11-55	0.47	6.30E-04	
GAU99461.1	NCBI	<i>Ramazzottius varieornatus</i>	Eukaryota; Metazoa; Ecdysozoa; T	288	47-92	0.61	0.014	
XP_022237972.1	NCBI	<i>Limulus polyphemus</i>	Eukaryota; Metazoa; Ecdysozoa; A	786	21-79	3.8	0.039	
XP_067126489.1	NCBI	<i>Centruroides vittatus</i>	Eukaryota; Metazoa; Ecdysozoa; A	791	17-58	0.047	1.10E-05	
XP_001603895.3	NCBI	<i>Nasonia vitripennis</i>	Eukaryota; Metazoa; Ecdysozoa; A	946	35-77	300	0.22	
NP_650597.2	NCBI	<i>Drosophila melanogaster</i>	Eukaryota; Metazoa; Ecdysozoa; A	979	5-45	N/A	N/A	
KAF5399513.1	NCBI	<i>Paragonimus heterotremus</i>	Eukaryota; Metazoa; Spiralia; Loph	435	4-41	8.9	0.081	
XP_013081743.1	NCBI	<i>Biomphalaria glabrata</i>	Eukaryota; Metazoa; Spiralia; Loph	929	19-63	0.19	6.50E-05	
ELT94853.1	NCBI	<i>Capitella teleta</i>	Eukaryota; Metazoa; Spiralia; Loph	506	9-55	0.0077	1.00E-05	
XP_013387541.1	NCBI	<i>Lingula anatina</i>	Eukaryota; Metazoa; Spiralia; Loph	1823	22-59	0.81	7.30E-05	
XP_067948887.1	NCBI	<i>Watersipora subatra</i>	Eukaryota; Metazoa; Spiralia; Loph	470	18-53	N/A	0.77	
RDD41302.1	NCBI	<i>Trichoplax sp. H2</i>	Eukaryota; Metazoa; Placozoa; Tric	358	18-54	0.37	2.90E-04	
CAB3982230.1	NCBI	<i>Paramuricea clavata</i>	Eukaryota; Metazoa; Cnidaria; Ant	1017	6-48	9	0.01	
CAI8037030.1	NCBI	<i>Geodia barretti</i>	Eukaryota; Metazoa; Porifera; Dem	591	8-55	N/A	2.1	
XP_004998732.1	NCBI	<i>Salpingoeca rosetta</i>	Eukaryota; Choanoflagellata; Crasp	581	2-52	3.90E-05	2.20E-09	
Sarc4_g19985T	Dudin et al. 2019 (128)	<i>Sphaeroforma arctica</i>	Eukaryota; Ichthyosporae; Ichthyop	212	21-68	0.025	2.60E-04	
Unigene2128	Nucleo	Torruella et al. 2015 (129)	<i>Parvularia atlantis</i>	Eukaryota; Rotosphaerida; Nuclear	124	44-83	0.0068	2.70E-04
NP_010142.2	NCBI	<i>Saccharomyces cerevisiae</i>	Eukaryota; Fungi; Dikarya; Ascom	223	92-141	N/A	N/A	

Table S4.2. Nucleic acid constructs used for this study.

Experiment	Construct Type	Purpose	Sequence
Zebrafish CRISPRant	DNA Oligo	HJURP CRISPRant genotyping primer F	CAGAGATCGAGATTTTCACTC
	DNA Oligo	HJURP CRISPRant genotyping primer R	CTTGGGACATAATTCGCTTG
	DNA Oligo	HJURP CRISPRant genotyping primer F	TCACTCGGGAAATTCGAAAG
	DNA Oligo	HJURP CRISPRant genotyping primer R	ATCCAGAAAAGTAGAAATCAGC
	DNA Oligo	HJURP CRISPRant genotyping primer F	TGCGTCAAAGCATATATTGG
	DNA Oligo	HJURP CRISPRant genotyping primer R	GAAAATGACCAGATCATGCTG
	DNA Oligo	CENPA CRISPRant genotyping primer F	CATGGCACACACTCCTAACC
	DNA Oligo	CENPA CRISPRant genotyping primer R	GGAATTGATTGAATCCTGACATG
	DNA Oligo	CENPA CRISPRant genotyping primer F	GAAAGGTTTTAGGTAGGTGC
	DNA Oligo	CENPA CRISPRant genotyping primer R	TCCTTACATCAGCAAGACTG
	Alt-R sgRNA	CRISPRant negative control (published)	GCAGGCAAAGAATCCCTGCC
	Alt-R sgRNA	CRISPRant negative control (HJURP scramble)	ACTAACGCGACTACTAGATA
	Alt-R sgRNA	CRISPRant negative control (HJURP scramble)	ACCTGGTATATCGCTCCGGA
	Alt-R sgRNA	HJURP CRISPRant guide 1	CGATGTCTTCAGAAACACAA
	Alt-R sgRNA	HJURP CRISPRant guide 2	GAGAGAGAGAGAGCCTGTAT
	Alt-R sgRNA	HJURP CRISPRant guide 3	TCTGCACTGTCAATGGGCAC
	Alt-R sgRNA	CENPA CRISPRant guide 1	GGACATGCCTCGCCATACAT
	Alt-R sgRNA	CENPA CRISPRant guide 2	GTGACATACAACCTCGCCCGG
Zebrafish Imaging	DNA Oligo	HJURP amplification from cDNA F	ATGCCTTTTATATACAGATCAGTTCAAG
	DNA Oligo	HJURP amplification from cDNA R	TTACATGTGTTCAACGCCTCT
	DNA Oligo	CENPA amplification from cDNA F	ATGCCTCGCCATACATCG
	DNA Oligo	CENPA amplification from cDNA R	TTACATGTGTTCAACGCCTCT
	Plasmid	eGFP-CENPA for <i>in vitro</i> transcription	GAAGATCCTTTGATCTTTTCTACGGGGT
	Plasmid	mCherry-HJURP for <i>in vitro</i> transcription	CGTAAAAAGGCCGCGTTGCTGGCGTTTT
	Plasmid	H2B-mCherry for <i>in vitro</i> transcription	GAAGATCCTTTGATCTTTTCTACGGGGT
Pulldowns	Plasmid	MBP-drHJURP200-6XHis for recombinant expression	aaaaaaaccaccgctaccagcggtggtgtgttgcggatcaagagct
	Plasmid	drCENPA/H4 for recombinant expression	ggggaattgtgagcggataacaattcccctctagaataattttgttaa
	Plasmid	MBP-NEFR1-6XHis for recombinant expression	TTGAGATCCTTTTTTCTGCGCGTAATCT
	Plasmid	HCP-3/H4 for recombinant expression	GATGGCACCAACGGAGACGGCAATCGC
<i>C. elegans</i> Strain Generation	DNA Oligo	<i>GFP::HA::hcp-3</i> genotyping primer F	CGGGGTTGAGTCCGTCATTTAT
	DNA Oligo	<i>GFP::HA::hcp-3</i> genotyping primer R	ATCAAAATCGTTGACACCTGCG
	DNA Oligo	<i>nefr-1</i> genotyping primer F	GAAGTCCGTGAATGCGCGTG
	DNA Oligo	<i>nefr-1</i> genotyping primer R	TCGAAGAAGGCGTTGCCAAG
	Alt-R crRNA	<i>nefr-1</i> 5' CRISPR guide	GAGACACTCGTTTAAGCGAA
	DNA Oligo	<i>nic1786[K15STOP]</i> repair template	AAGCAAGCGCCGACAAAGAGACTC
	DNA Oligo	<i>nic1802[3XFLAG::nefr-1]</i> repair template	tgaattatttaccatttttcagtcgacATGcactacaagaccatgacg
	dsDNA Oligo	<i>nic1816[AID::nefr-1]</i> repair template	tgaattatttaccatttttcagtcgacATGcctaaagatccagccaaac
	DNA Oligo	<i>nic1803[ALFA::nefr-1]</i> repair template	tgaattatttaccatttttcagtcgacATGCCATCCCGGCT

Table S4.3. A list of *C. elegans* strains generated and/or used in this study.

Strain	Genotype	Source	Notes
N2	<i>Ancestral</i>	Caenorhabditis Genetics Center	
FAS131	<i>hcp-3(uge85[GFP::HA::hcp-3]) III</i>	Prosée et al. 2021	
NJL5575	<i>nefr-1(nic1786[K15STOP] I/ tmc20[unc-14(tmIs1219) dpy-5(tm9715)] I</i>	This study	
NJL5639	<i>nefr-1(nic1786[K15STOP] I/ tmc20[unc-14(tmIs1219) dpy-5(tm9715)] I; his-72(uge30[gfp::his-72]) III</i>	This study	
NJL5662	<i>nefr-1(nic1802[3XFLAG::nefr-1]) I</i>	This study	
NJL5667	<i>nefr-1(nic1803[ALFA::nefr-1]) I; unc-119(kst33) III; nic1744 [rpl-28::NbALFA::linker::GFP::tbb-2] V</i>	This study	
JDW220	<i>wrdsi18[mex-5p::TIR1::F2A::mTagBFP2::AID*::NLS::tbb-2 3'UTR] (I:-5.32)</i>	Ashley et al. 2021	Self-excising cassette was removed to generate NJL5663
NJL5663	<i>wrdsi18[mex-5p::TIR1::F2A::mTagBFP2::AID*::NLS::tbb-2 3'UTR] (I:-5.32)</i>	Ashley et al. 2021/This study	
NJL5706	<i>nefr-1(nic1816[AID*::nefr-1]) I</i>	This study	
NJL5716	<i>nefr-1(nic1802[3XFLAG::nefr-1]) I; hcp-3(uge85[GFP::HA::hcp-3]) III</i>	This study	
NJL5735	<i>wrdsi18[mex-5p::TIR1::F2A::mTagBFP2::AID*::NLS::tbb-2 3'UTR] (I:-5.32); nefr-1(nic1816[AID*::nefr-1]) I; hcp-3(uge85[GFP::HA::hcp-3]) III</i>	This study	

4.7 Afterword

This project represents everything that is beautiful about scientific discovery. It was a delightful collaboration among Fred Hutch Basic Sciences labs involving Jason Stonick and Drs. Janet Young, Harmit Malik, Irini Topalidou, Nicolas Lehrbach, and Cecilia Moens. This project began in the Malik lab long before I moved to Seattle, but I picked it up as a rotation project and had exceptional luck early on. The division's excitability and can-do attitude, along with advances in sequencing that gave us access to essential evolutionary intermediate data, made it possible. It was also an excellent project to work on when the integrin work was at periodic intellectual standstills and often left it alone for a few months; then coming back with fresh eyes was all we needed to find a new ortholog. The work speaks for itself, but I will cherish the joy of working with friends more than any individual finding.

Chapter 5 Concluding remarks.

5.1 Abstract

The biological world contains an awe-inspiring amount of molecular complexity. Although some of this is genuine novelty, most adaptation builds from ancient foundational pieces that are retooled to fill new niches. In this thesis, I've attempted to test the limits of two such building blocks through very specific perturbations, yet this comes at the cost of a systems-level understanding of the consequences of both these domain perturbations specifically and of protein domain adaptation more broadly. During my thesis, I've witnessed an explosion of high-throughput structure-function analyses, usually at the level of individual mutations. I anticipate many of the tools developed during this time will be extrapolated to higher-order macromolecular assembly in the very near future. In this brief chapter I offer a broad-level overview of both the future directions of probing macromolecular assembly as well as insights into how we can take these findings back to organism and population-level biology.

5.2 From building blocks to functional machines

In my thesis work, I've dissected molecular determinants of single-domain structure, function, and plasticity with particularly fine resolution. I've explored how a single domain can either acquire new function with relative simplicity or maintain a conserved function under extreme adaptive pressure. These processes occur across the entire protein domain universe at the level of individual domains, so these findings are likely generalizable to other proteins.

More interesting, however, are the constraints imposed across multi-domain proteins and macromolecular assemblies. Due to limitations in throughput and sequencing coverage, we have often been limited in connecting genotype to phenotype at the level of individual protein domains, or even subsections of these domains. However, recent tool development in massively parallel reporting has expanded the amount of variation we are able to measure by orders of magnitude, suggesting that contextualizing variation within entire biological systems is achievable(249, 267, 497, 498). These methods may be supplemented by the use of natural variation, which provides a rich underutilized source of genotype-phenotype data(499, 500). These approaches will also allow us to contextualize allostery, an understudied yet critical feature of protein biology for which an increasing amount of literature reports strong phenotypic effects(501, 502). Synthesizing all these data will yield a more complete model of the general principles underlying how protein domains can be combined to generate diverse, complex functional machines.

How much domain space have we explored? Early estimates suggest as few as nine simple protein folds account for the vast majority of protein structural diversity present throughout life today(231). Up until the advent of powerful structure prediction, however, methodology relied far too heavily on prior knowledge of protein structure to accurately assess what we don't know. More recently, advances in protein structure prediction have greatly expanded our understanding of structural diversity in the natural world(340). Viral proteomes may be a great place to look next(237, 503); rapid evolution, strong selective pressure, and high mutation rates all suggest that viruses are likely to harbor new folds and domains otherwise inaccessible in cellular organisms. Further, continued advances in protein design and engineering

open new avenues for how protein domains may cooperate, assemble, and function (504, 505). These techniques not only give insight into what is, but what could be.

5.3 Everything in its context

I spent most of my thesis working *in vitro*, in test tubes under controlled conditions where molecular behavior isn't confounded by the dizzying array of physical interactions organic molecules normally experience. Life, however, doesn't happen in a test tube. Life happens through the integration of these interactions into cohesive output, where molecular decisions are contextualized given all the physical and chemical information at hand. Lately I've become interested in moving back into the living organism- rightfully so, as the integrin structure field in particular faces big open questions as to how much of what we find *in vitro* can be recapitulated *in vivo*, as well as how integrins process decision-making challenges in the multicellular organism.

The continued development of novel methodology in cryo-electron tomography (cryoET) in particular holds great promise to answering some of these long-standing questions(506, 507). Briefly, cryoET integrates the high resolution of single-particle cryoEM with the context of the native cellular environment by vitrifying and imaging whole frozen cells or tissues, then mechanically thinning them to a thickness sufficient for transmission EM(508). CryoET studies have yielded fruitful insights into a variety of biological phenomena, from phage assembly to Parkinson's pathology (509, 510). Given the success in observing integrins on vesicular membranes(511), it is theoretically possible to observe them on native cytoplasmic membranes at sufficient resolution to make conclusions about ligand binding, conformational heterogeneity, and clustering that have only been observed at either the single-particle level or at light microscopy resolutions. The biggest challenge facing cryoET is finding the target molecule on the grid, but recent advances in high-contrast tagging or labeling have moved that goal within reach(512, 513). This technology need not be limited to integrins; it has already been applied to studying the centromere *in situ* as a proof of principle (514), and may prove equally exciting for understanding how epigenetic landscapes are shaped and maintained across space and time.

Integrins are inherently force-sensing molecules; although this has been known for over 20 years, the structural changes that occur upon force induction are largely uncharacterized(515). Ligand binding may not be sufficient to induce signaling (516), and emerging evidence suggests that the entropy of integrin-ligand interactions (115), which may change upon force induction, plays a central role in organism-level, integrin-mediated phenotypes. Further, integrin-ligand interactions with strong support on coated surfaces cannot be recapitulated with single molecules in solution(517–519), lending further support to biophysical force as a key modulator of integrin activity. In line with this, some integrin epitopes are exposed only upon force induction(520, 521). Continued development of methods that probe integrin behavior and structure under physiological force will give a more holistic perspective on these dynamic receptors.

Appendix I How do integrin subunits find their partners?

I.1 Abstract

Integrins are trafficked to the cell surface as obligate heterodimers, and their biosynthetic pathways are interdependent on the α and β subunits folding correctly, pairing, and jointly following the secretory pathway. However, little is known about the molecular mechanisms that mediate integrin biosynthetic pairing and secretion. To understand how integrin subunits find each other in the ER, I expressed 3XFLAG-tagged integrin subunits on their own in mammalian cell culture and performed IP-MS to identify intracellular binding partners. The most promising candidate shared among α subunit datasets was the ERAD-associated protein SEL1L, hinting at a molecular quality control mechanism controlling coupling integrin subunit biosynthesis to pairing.

I.2 Main text

Among all possible integrin heterodimer combinations, only a relative few are expressed in the human body(3). One of my early grant aims in the lab proposed a series of experiments attempting to uncover the molecular basis of this puzzling specificity, including expressing and purifying integrin subunits on their own, then pairing them *in vitro*. I quickly ran into a dead end when I found that although I could express and isolate β subunit ectodomains successfully, α ectodomains were refractory to independent cell surface expression. This appears to be universally true, as evolutionarily distant α subunits all behaved similarly in this expression system (Fig. I.1A). From these data, I hypothesized that an early conserved process in integrin was controlling subunit pairing, allowing secretion.

Relatively little is known about the integrin biosynthetic pathway is known. Integrins, like most cell surface receptors, are first translated and folded in the endoplasmic reticulum, followed by glycosylation and signal peptide cleavage(522). They are then trafficked to the Golgi(523), where they are further processed, packaged into secretory vesicles, and transmitted to the cell surface. The expression of integrin α subunits appears to be the rate-limiting step for integrin secretion, as increasing α subunit expression dramatically reduces integrin maturation time through this pathway(524). Supporting this finding, immature integrin β subunits accumulate in the ER until they pair with a cognate α subunit. Further evidence suggests the proper folding of the α subunit is dependent on the presence of stable β subunits(525), suggesting an intimate link between integrin folding and heterodimer formation, assuring stability of these dynamic receptors. In some exceptional cases, high-mannose alpha subunits may skip the Golgi and traffic directly to the cell membrane(522). Once subunits are at the plasma membrane, they may stochastically unpair and repair(526), and the expression, pairing, and secretion of integrin heterodimers may be further controlled by the presence of integrin ligand or other chemical cues(522, 524, 527, 528), as well as divalent cations like Ca^{2+} (529).

To understand which step of the secretory pathway may be limiting α subunit expression, I expressed 3XFLAG-tagged α L and α 5 integrin ectodomain constructs and performed immunoprecipitation coupled with mass spectrometry (IP-MS) and looked for proteins overrepresented compared to a no expression control. As expected, ER/Golgi proteins, stress response components, and the ECM proteins overrepresented in the experimental groups compared to the control (Fig. I.1B), consistent with integrin subunits stalled in the secretory

pathway. Among the differentially represented proteins, the ER-associated degradation (ERAD) component SEL1L(530) stood out as a strong candidate in directly mediating integrin biosynthesis (Fig. I.1C). SEL1L, a member of the Hrd1 E3 ubiquitin ligase complex(531, 532), has been shown in previous studies to bind preferentially with unprocessed integrin β 1 subunits(533), to be essential for integrin α 3 processing and trafficking(534), and to be a substrate of the SEL1L/Hrd1 complex in high-throughput screening(535). Consistent with this, we find evidence of ubiquitinated α L in our IP-MS dataset even in the absence of a proteasome inhibitor (Fig. I.1D). While still preliminary, it is tempting to speculate that SEL1L, and ERAD more broadly, couples integrin chaperoning and dimerization to quality control at the ER and may serve as a limiting factor for which integrin subunits form biologically relevant dimers (Fig. I.1E). Further biochemical and structural probing of this interaction may yield pertinent insights into this process and enable methodology for expressing integrin subunits on their own, which would prove extremely valuable for the field.

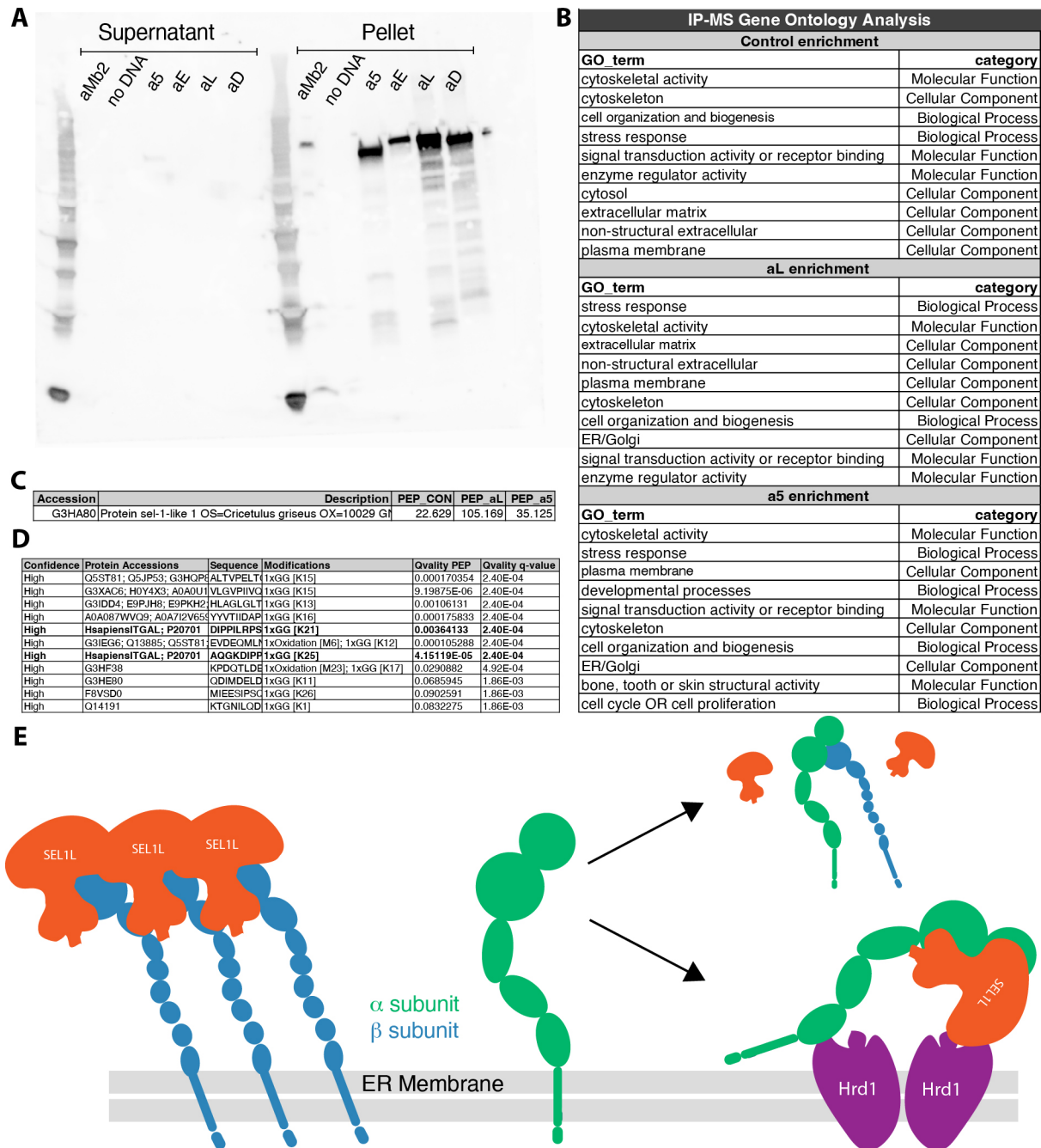
I.3 Methods

For figure I.1a, DNA encoding the ectodomain of integrins α 5, α E, α L and α D with a C-terminal Strep-Tag II in a pcDNA3.1 backbone was transfected in expiCHO cells in a 6 well plate format according to manufacturer instructions. The complete α M β 2 ectodomain was used as a positive control, and a no DNA well served as a negative control. Four days post-transfection, 100 μ L of cells were harvested, pelleted at 500xg for 5 minutes, and resuspended in 100 μ L of 25mM Tris-HCl pH 7.4. 20 μ L of either the supernatant or cell pellet suspension was boiled in 4X SDS sample buffer and processed on SDS-PAGE, transferred to a blotting membrane, probed with anti-Strep-Tag antibody and visualized via chemiluminescence.

For pulldown experiments, DNA constructs encoding the ectodomains of α 5 and α L with C-terminal 3XFLAG tags were transfected in expiCHO cells according to max titer manufacturer instructions. On day 8 post-transfection, 100mL of cells from each transfection and a no DNA control were centrifuged at 1000xg for 15 minutes at 4°C and resuspended in 2mL of lysis buffer (25mM Tris-HCl pH 7.5, 1mM EDTA, 150mM NaCl, 1% v/v NP-40, 0.5mM PMSF, 1x protease inhibitor cocktail tablet). Following 45 minutes of gentle inversion at 4°C, lysates were cleared at 30000xg for 20 minutes. Lysates were diluted to 1mg/mL in ice cold lysis buffer and incubated with 60 μ L of equilibrated anti-FLAG M2 beads for 3 hours with gentle inversion. Beads were collected on a magnetic separation rack and washed once with lysis buffer, followed by 5 washes with 1mL wash buffer (25mM Tris-HCl pH 7.5, 150mM NaCl) to remove detergent. Beads were resuspended in 30 μ L wash buffer and submitted for mass spectrometry analysis to the Fred Hutch Cancer Center proteomics core facility.

Samples were processed, desalted and run on the Orbitrap Fusion at the Fred Hutch proteomics facility. The data were searched using Proteome Discoverer 2.5 against CHO, Human and Strep-Tag II. The results were filtered for high confidence, 1% FDR at the peptide level. The data were also searched for GG on K to identify ubiquitinated targets.

Figure I.1. SEL1L may contribute to integrin heterodimer formation. (A) Western blot analysis of supernatants (left) or pellets (right) of expiCHO cells expressing various integrin α subunits. Samples were probed for anti-Strep-



TagII. Alpha subunits stay unprocessed in the cell pellet. (B) Gene ontology (GO) analysis of enriched peptides in mass spectrometry data from α subunit pulldowns. Pulldowns of integrins subunits were enriched for ER/Golgi components compared to the control. (C) Peptide representation of SEL1L from integrin pulldown data, suggesting SEL1L is enriched compared to control and may be a component of integrin quality control. (D) GG on K proteomic data from the α L subunit pulldown showing α L is ubiquitinated at multiple sites when expressed on its own. (E) A model for SEL1L-mediated quality control of the integrin heterodimer. β subunits are sequestered and protected by SEL1L while awaiting α subunit expression. When α subunits are folded they are either paired to a corresponding β , releasing SEL1L, or are targeted for proteasomal degradation via the Hrd1 ERAD pathway.

Appendix II The structural basis of orthomyxovirus restriction by MxA

**This work was done in collaboration with Dr. Rechel Geiger*

II.1 Abstract

Innate immune proteins such as the interferon-stimulated orthomyxovirus restriction factor MxA are the body's first line of defense against pathogens. MxA restricts orthomyxovirus replication via its interaction with the viral nucleoprotein (NP), but the structural basis of this restriction remains unknown. We determined negative-stain EM class averages of the MxA multimer, possibly containing internal Thogotovirus NP, suggesting a mechanism by which MxA sequesters viral components to prevent replication in a GTP-dependent manner. Future structural work on this protein complex may yield fruitful insights into the arms race interface in this stereotypical host-pathogen molecular conflict.

II.2 Main text

In order to respond immediately to pathogenic invaders, organisms encode innate immune factors that restrict virulence through a cornucopia of unique molecular pathways(536, 537). The orthomyxovirus family of viruses, including Thogotovirus and Influenza, is restricted in mammalian populations by the interferon-stimulated dynamic GTPase family member MxA(538). Extensive evolutionary, cellular, and biochemical work has determined the key molecular interactions for Orthomyxovirus restriction by MxA: restriction depends on 1) direct binding between MxA and the viral nucleoprotein (NP), 2) the GTPase activity of MxA, 3) the rapidly evolving L4 loop of MxA, and 4) MxA oligomerization(539–543). This restriction can be further amplified by targeting MxA to the nucleus(544). Still, the structural basis of this interaction has remained elusive.

Prior structural work on MxA alone suggests it can form filamentous structures via stacking of the stalk and G domains (545, 546). MxA's multimerization state is dependent on its GTPase activity and stalk domain contacts, as it can also break apart into smaller multimers such as dimers and tetramers; biochemical work with NP suggests that the MxA dimer is the most stable restriction unit(543). Beyond the direct protein-protein interaction, it is possible that MxA further depends on RNA binding to fully sequester NP; this model would be consistent with other restriction factors(547) and would prove harder to evade by viral evolution as NP must also bind RNA for viral replication. Structural information would clarify the mechanism underlying these findings and prove clinically valuable for viral epidemiology.

We cloned, expressed, and purified wildtype, dimeric and "super-restrictor" variants of MxA(548) via affinity and size exclusion chromatography. Initially we excluded nucleases from the preparation, although we found that MxA ran at a much higher molecular weight than would be expected via SEC (Fig. II.1A). Adding RNase to the preparation shifted MxA to a peak more consistent with its size (Fig. II.1A), suggesting MxA can bind to nucleotides; an A260/A280 ratio of 1.1 in the initial preparation is consistent with this hypothesis. These MxA samples were clean in nsEM (Fig. II.1B). Although we did not further test nucleotide binding, the fact that positively charged mutations in MxA L4 increase restriction(548, 549) may also indicate that nucleotide binding is important for the restriction mechanism.

To determine the structural basis of restriction, we purified monomeric Thogotovirus NP and bound it to wild-type MxA with various combinations of GTP and short RNA oligos. We then performed nsEM of these complexes and generated 2D averages of MxA oligomers. In all conditions, we identified ring-shaped structures consistent with helical MxA multimers, although adding the non-hydrolyzable GTP analog GTP γ S anecdotally reduced the presence of ring structures (Fig. II.1C), much like the paralog MxB (550). In some 2D classes from samples with NP we could see globular density inside the ring shape consistent with the size of NP while other rings had no internal density (Fig. II.1D). These data collectively suggest a model in which a MxA disassembles into dimers upon GTP hydrolysis, binds NP, and re-oligomerizes to sequester NP from the viral genome (Fig. II.1E); this model is consistent with tomographic work on MxA with the La Crosse virus nucleocapsid protein(551). Isolating these multimeric rings would likely lead to a sample clean enough for high-resolution cryoEM studies of this potential complex, revealing the molecular basis of orthomyxovirus restriction.

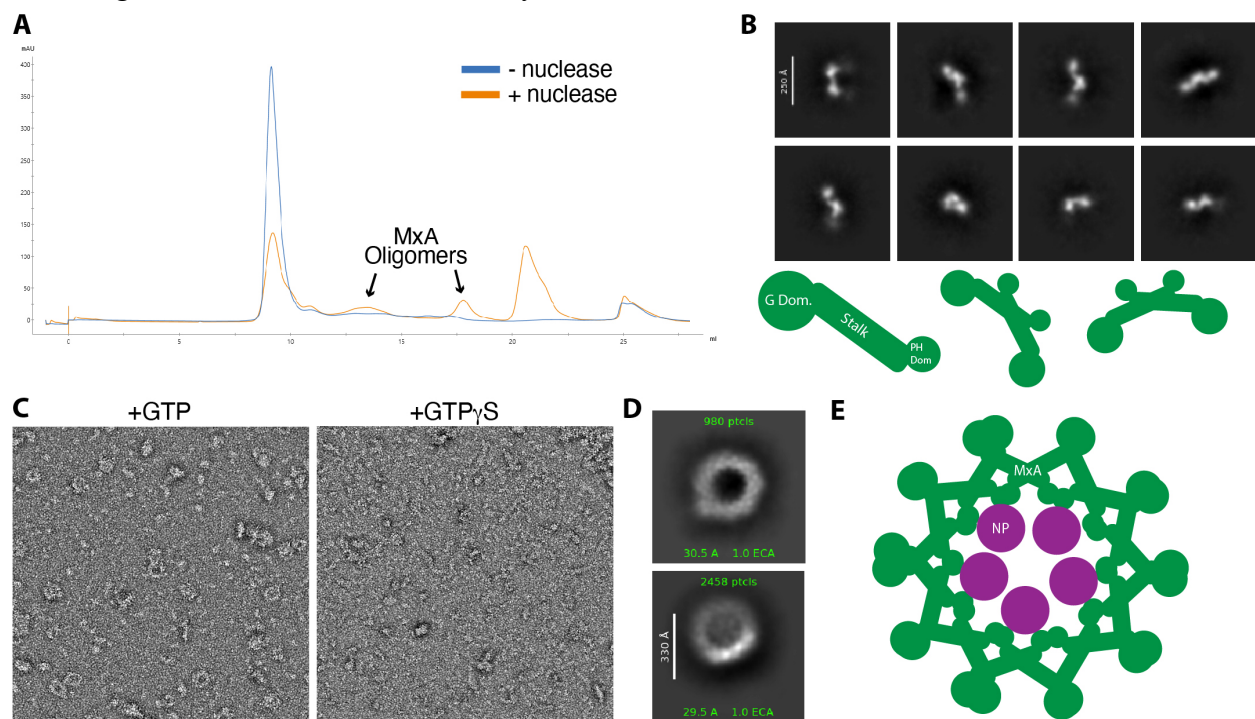


Figure II.1. nsEM of Thogotovirus NP and MxA. (A) Size exclusion chromatography of MxA in the absence (blue line) or presence (orange line) of nucleases during purification. Adding nucleases shifted peaks from near-void to sizes consistent with MxA oligomeric states (B) Negative stain of MxA in high salt shows shapes consistent with dimers; cartoons of a monomer and dimer orientations below averages. (C) Negative stain micrographs from MxA and Thogotovirus NP in either GTP or GTP γ S show ring shapes are present only with GTP, while smaller oligomeric states likely exist when GTP is not hydrolyzed. (D) 2D averages of MxA rings with no density (top) or density (bottom) in the middle, consistent with the absence or presence of NP. (E) A cartoon model of how MxA sequesters viral NP in a ring oligomer to restrict viral replication.

II.3 Methods

II.3.1 Protein purification.

Orthomyxovirus NP proteins were purified as previously reported(552). pET28 vectors containing 6XHIS-tagged NP variant coding sequences were transformed into BL21 DE3

LOBSTR cells. Following growth at 37°C to OD 06, cultures were induced to express NP with IPTG at 0.5mM final concentration and shifted to 18°C for overnight expression. Cells were then pelleted and resuspended in ice-cold lysis buffer (20mM NaPO₄ pH 7.4, 1.5M NaCl, 1mM PMSF) and lysed via sonication. The lysate was cleared at 30000xg for 20 minutes at 4°C, loaded onto a pre-equilibrated HisTRAP column and eluted with elution buffer (20mM NaPO₄ pH 7.4, 1.5M NaCl, 1mM PMSF, 500mM imidazole). NP was further purified using preparative size exclusion chromatography in 20mM NaPO₄ pH 7.4, 1.5M NaCl, aliquoted, flash-frozen in 10% v/v glycerol, and stored at -80°C for downstream use.

MxA variants were purified as described(545). Like NP, pET28 vectors containing 6XHIS-tagged MxA variants were expressed in LOBSTR cells following similar growth conditions. Cells were lysed in MxA lysis buffer (50mM HEPES pH 7.5, 400mM NaCl, 30mM imidazole, 6mM MgCl₂, 2.5mM β-mercaptoethanol, 1mM PMSF, 1x protease inhibitor cocktail tablet), lysed via sonication, and cleared at 30000xg for 20 minutes at 4°C. Lysates were loaded onto a pre-equilibrated HisTRAP column, washed with lysis buffer containing 800mM NaCl and 1mM ATP followed by a wash with lysis buffer containing 80mM imidazole, and eluted with elution buffer (50mM HEPES pH 7.5, 400mM NaCl, 300mM imidazole, 6mM MgCl₂, 2.5mM β-mercaptoethanol). Samples with nuclease treatment were incubated with lysis buffer containing 10 units RNase A and 50mM imidazole for 1 hour prior to washing and eluting. Samples were further purified using preparative SEC in 20mM HEPES pH 7.5, 400mM NaCl, 2mM MgCl₂, 2mM DTT, aliquoted, flash-frozen in 10% v/v glycerol, and stored at -80°C for downstream use.

II.3.2 Negative-stain EM sample preparation, data acquisition and processing

Protein samples were diluted to a final concentration of between 2 and 20μg/mL in buffer containing 20mM HEPES pH 7.4, 150mM NaCl (400mM for “high salt”), and the presence or absence of a 3:1 molar excess of short RNA(553) and/or 1mM GTP/GTPγS. The grids were stained with a solution containing 2% (w/v) uranyl formate as previously described(315). Data were acquired using a Thermo Fisher Scientific Talos L120C transmission electron microscope operating at 120 kV and recorded on a 4k x 4k Thermo Scientific Ceta camera at a nominal magnification of 92000x with a pixel size of 0.158 nm. Leginon(316) was used to collect 100-200 micrographs at a nominal range of 1.5-2.5 μm under focus and a dose of approximately 50 e⁻/Å² per condition.

Similar processing pipelines were used for all negative stain datasets. Micrographs were processed using cryoSPARC(317). CTF estimation was done using Patch CTF. Initially, 10000-100000 particles were picked in a reference-free manner using blob picker in cryoSPARC. The particle picks were subjected to 1 round of particle curation and 1 round of 2D classification followed by 2D class selection contributing to the final 2D classes.

Appendix III Rapid evolution of the sperm-egg fertilization complex

**An abbreviated version of this analysis is published in Deneke, Suwita et al. 2026(554).*

I.1 Abstract

Fertilization, the fusion of gametes to form a zygote, is an essential process in sexual reproduction. Despite its necessity, the molecular components implicated in the fertilization machinery evolve exceptionally rapidly. The cause of this rapid evolution, however, remains unknown. Here I perform positive selection analysis on a newly identified vertebrate fertilization complex, SPARK, to identify which regions are under selective pressure and how those regions relate to function. I find that most of the sites under positive selection are concentrated at the distal end of the fertilization complex and at sites of protein-protein interactions. I also find hotspots of adaptive pressures at exposed sites implying currently unidentified sperm-egg interactions. Finally, this analysis supports a model in which the sperm-egg fusogen is shielded prior to gamete contact. These results reinforce the power of positive selection analysis in generating and supporting biochemical hypotheses, even for essential machinery.

I.2 Main text

The life of an individual organism in a polyploid species often begins with fertilization, the process by which gametes fuse to form a one-cell zygote. In vertebrate species the parental sperm must identify, bind to, and fuse with the maternal egg cell via a complex series of chemical attraction, biophysical motility, and biochemical protein-protein interactions(555, 556). Although this process is both essential and conserved across many eukaryotes, the molecular machinery implicated in sperm-egg interactions evolves at an exceptionally rapid pace(500, 557, 558). Many hypotheses have been proposed for what causes this rapid evolution including sperm competition, sperm-egg conflict, host-pathogen interactions, and sexual selection(559). Our understanding of this extraordinary selective pressure is hampered by the fact that our exploration of the fertilization process is still in relative infancy(556). Many individual players have been proposed as essential for sperm-egg interactions, yet how they are incorporated into the larger superstructure facilitating this dynamic fusion process remains unknown.

Towards the beginning of my thesis work I began following the work of Drs. Andi Pauli and Luca Jovine, two pioneers of fertilization molecular biology, with great interest (so much so that I interviewed for a postdoctoral position with Dr. Pauli). Although the international move proved impractical at the time, I maintained a collaboration with Dr. Victoria Deneke in the Pauli lab, who had recently discovered via AlphaFold screening that a large sperm protein complex, termed SPARK, structurally synthesized all sperm components previously implicated in fertilization(554). This complex further revealed compelling interactions with egg receptors CD9 and JUNO, thus providing an opportunity to test where selective pressures are concentrated at the fertilization synapse. Conversely, positive selection analysis would help support their proposed AlphaFold model: I would expect SPARK complex proteins to exhibit strong signatures of adaptive evolution, or positive selection, measured by the relative ratio of nonsynonymous to synonymous codon substitutions (dN/dS) in closely related species. To test this hypothesis, I performed positive selection analysis using PAML(308) on components of the SPARK complex, as well as their hypothesized receptors JUNO and CD9, in *Muridae* (Fig.

III.1A, Table III.1). I found an overall average dN/dS value of 0.52 among complex members, appreciably higher than values from across *Muridae* genomes overall(455). I identified site-specific positive selection (dN/dS > 1) in 7 out of 12 tested complex members. Overall, these results implicate adaptive evolution driving the diversification of the SPARK complex.

If sperm-egg binding is driving the adaptive evolution of SPARK proteins, I would expect the amino acids under selection would be concentrated at sperm-egg interfaces(560, 561). I mapped these sites of positive selection to the predicted SPARK complex, identifying sites spread irregularly across the entirety of the complex (Fig. III.1B). Specifically, there are two main areas of exceptionally localized positive selection: the distal interface of SPACA6 and IZUMO1, where SPARK is predicted to interact with egg receptors JUNO and CD9 (Fig. III.1C), as well as the distal tip of the interface encompassing DCST1/2 and TMDD1 (Fig. III.1D). These data are consistent with the expectation that these regions are physically involved in forming the sperm-egg synapse. Further, I found reciprocal residues predicted to interact between sperm and egg proteins under positive selection between IZUMO1 and JUNO (Fig. III.1E) as well as IZUMO1-CD9-SPACA6 (Fig. III.1F). Phylogenetic analysis suggests an evolutionary pressure to diversify more pronounced in sperm genes (Fig. III.1A, III.1G), suggesting sperm diversification precedes egg gene adaptation. Identifying adaptive selection in DCST1/2 and TMDD1 implicates this region as important for physical contact with the egg, and supports the emerging model that DCST1/2 is the yet unfound fusogen bringing sperm and egg cells together.

Evolutionarily, identifying this SPARK complex provides a foundational resource for testing longstanding hypotheses on genetic conflicts in mating(562, 563). We find intense selective pressure shaping the sperm-egg interface, consistent with a tit-for-tat arms race that provides opportunities for sperm to “win” depending on their SPARK genotype(564). This is most pronounced at the interface encompassing the SPARK members IZUMO1 and SPACA6 and the hypothesized egg receptors CD9 and JUNO. Intriguingly, I found a large patch of concentrated rapidly evolving residues in SPACA6 for which there is no proposed egg receptor; this implies that further egg receptors have yet to be identified, or that this interface is dynamic and may split CD9 binding fluidly across IZUMO1 and SPACA6. I also find the distal region encompassing DCST1/2 and TMDD1 contains a high concentration of residues under positive selection, yet in the bent SPARK conformation this region is sterically blocked from contacting the egg surface. This further supports a licensing model by which DCST1/2 and TMDD1 are liberated upon sperm-egg binding, providing another interface by which sperm may gain the evolutionary upper hand.

I.3 Methods

Positive selection analyses were performed as previously published(379). Coding sequences from all SPARK components as well as CD9, CD81, and JUNO in all *Murinae* species with annotated genomes in NCBI(455) were collected and codon-aligned via webPRANK(482), and manually trimmed of gaps that contribute noise to analyses. Alignments were used to generate phylogenetic trees via phyML(565, 566). Gene-wide and site-specific positive selection was assessed via a streamlined wrapper (<https://github.com/jayoung/pamlWrapper>) for the codeml package in PAML(308). Whole gene dN/dS was calculated via model 0. For gene-wide analysis, the likelihoods of dN/dS fitting a model where either positive selection was (model 2) or was not (model 1) allowed were

assessed; the same approach was subsequently used for site-specific positive selection by comparing model 8 to model 8a; we restricted classifying residues under positive selection to those with a Bayes Empirical Bayes posterior probability greater than 0.9 under model 8. Codon alignments can be found in Supplemental Data 1, and PAML analysis results in Table S2. Residues under positive selection were mapped onto the predicted SPARK complex structure and visualized using UCSF ChimeraX.

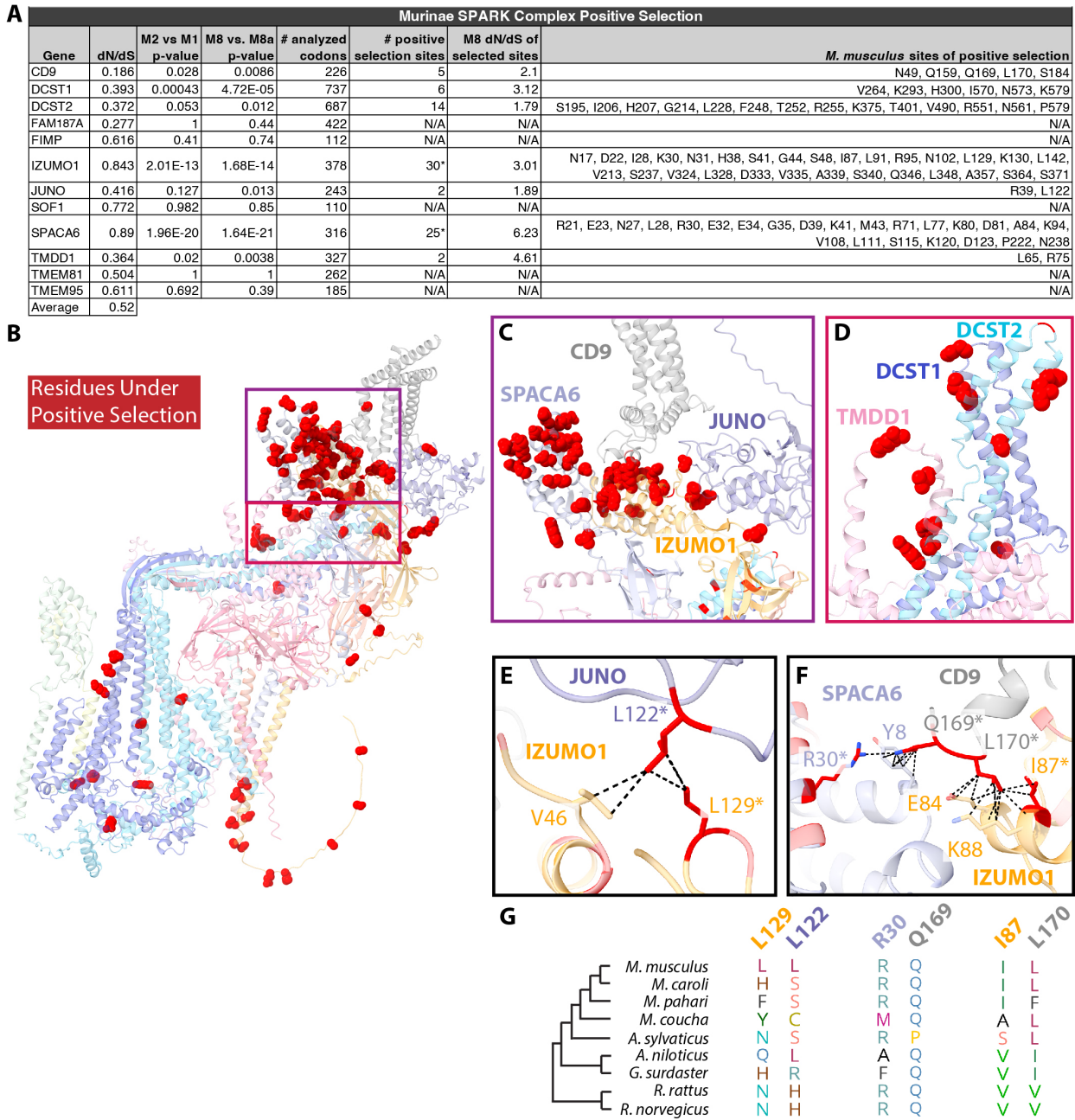


Figure III.1. Positive selection across the fertilization complex. (A) Components of the SPARK complex evolve under positive (diversifying) selection in *Murinae*. All ten components of SPARK as well as the egg receptors CD9 and JUNO were tested for the presence of both gene-level and site-specific positive selection. High SPARK gene dN/dS values suggest the presence of adaptive pressure. (B) The predicted *M. musculus* SPARK complex ribbon structure bound to CD9 and JUNO with residues under positive selection identified as red atoms. (C-D) Zoom views

from the whole SPARK complex show concentrated regions of positive selection at the IZUMO1/SPACA6 distal tip (C) as well as the DCST1/2 and TMDD1 tip (D). (E-F) Zoom views of physically interacting residues under positive selection between murine JUNO and IZUMO1 (E) and SPACA6, CD9, and IZUMO1 (F) suggest concerted evolution between egg and sperm proteins shapes the SPARK evolutionary trajectory. (G) Phylogenetic substitutions in *Murinae* at interacting sites between sperm and egg fertilization factors.

Gene_name	pam1Version	num_seqs	codon model	initial omega	cleandata	numNT	numAA	MO overall dM InL	MO fix InL	MovsMOfixed	MO total tree	MO total dN t	MO total dS t	M1 InL
CD9	4.10.6, Noverl	11	2	0.4	0	678	226	0.18572	-2028.8983	-2094.3231	2.671E-30	1.005968	0.165	0.8883
DCST1	4.10.6, Noverl	9	2	0.4	0	2211	737	0.39296	-7258.5061	-7335.4818	2.3734E-35	1.207533	0.292	0.743
DCST2	4.10.6, Noverl	9	2	0.4	0	2061	687	0.37205	-6898.8225	-6984.2637	4.7472E-39	1.246186	0.2848	0.7656
FAM187A	4.10.6, Noverl	8	2	0.4	0	1266	422	0.27661	-3707.9257	-3780.3229	2.3818E-33	1.024555	0.2014	0.7281
FIMP	4.10.6, Noverl	9	2	0.4	0	336	112	0.61646	-1114.1901	-1117.3979	0.011313	1.323096	0.3705	0.6009
IZUMO1	4.10.6, Noverl	9	2	0.4	0	1134	378	0.84342	-4551.9847	-4553.554	0.076461	1.801577	0.572	0.6781
JUNO	4.10.6, Noverl	11	2	0.4	0	729	243	0.41615	-2993.9371	-3024.9128	3.5201E-15	1.880739	0.4788	1.1504
SOX1	4.10.6, Noverl	9	2	0.4	0	330	110	0.77218	-1030.9655	-1031.6957	0.22686	1.167849	0.3574	0.4628
SPACA6	4.10.6, Noverl	9	2	0.4	0	948	316	0.89024	-3333.8632	-3334.348	0.32483	1.340916	0.4323	0.4856
TMDD1	4.10.6, Noverl	9	2	0.4	0	981	327	0.3644	-3361.699	-3405.2787	1.0011E-20	1.351037	0.3219	0.8833
TMEM81	4.10.6, Noverl	9	2	0.4	0	786	262	0.50387	-2605.5426	-2621.1844	2.2296E-08	1.275492	0.3333	0.6615
TMEM95	4.10.6, Noverl	9	2	0.4	0	555	185	0.61109	-1584.7877	-1588.5537	0.0060612	0.916784	0.2611	0.4272
Gene_name	M2 InL	M7 InL	M8 InL	M8a InL	M8a InL	M8a InL	M8a InL	M8a InL	M8a InL	M8a InL	M8a InL	M8a InL	M8a InL	M8a InL
CD9	-1971.7501	-1977.2855	-1971.8845	-1975.335	6.90112	0.0086142	10.802076	0.0045119	7.117884	0.028469	8.177	2.09517	5	49_N_0.989,159_Q_0.945;
DCST1	-7153.171	-7162.9418	-7152.6594	-7160.9372	16.555618	4.7244E-05	20.564916	3.4228E-05	15.501894	0.00043033	4.537	3.12217	6	269_V_0.950,298_K_0.976;
DCST2	-6796.1024	-6800.4389	-6795.9238	-6799.0578	6.268024	0.012294	9.030192	0.010943	5.878548	0.052804	11.904	1.79351	14	195_S_0.957,206_L_0.948;
FAM187A	-3667.927	-3667.6904	-3667.406	-3667.7033	0.594542	0.44067	0.568788	0.75247	0	1				
FIMP	-1106.6731	-1107.622	-1107.5082	-1107.5647	0.113078	0.73667	0.227742	0.89237	1.783272	0.40998				
IZUMO1	-4463.8455	-4468.1552	-4463.7028	-4493.1391	58.872584	1.6822E-14	68.904766	1.0902E-15	58.470092	2.0109E-13	23.319	3.00706	30	38_N_0.920,43_D_0.971,51
JUNO	-2948.5709	-2953.0396	-2947.4909	-2950.5456	6.109462	0.013446	11.09731	0.0038927	4.125614	0.1271	13.343	1.88852	2	58_R_0.919,141_L_0.957
SOX1	-1029.1003	-1029.2223	-1029.1003	-1029.1183	0.03614	0.84923	0.243962	0.88517	0.036136	0.98209				
SPACA6	-3232.9789	-3278.5045	-3233.0078	-3278.3755	90.735566	1.6421E-21	90.993422	1.7419E-20	90.76128	1.9563E-20	10.858	6.23139	25	10_L_0.939,45_R_0.998,47;
TMDD1	-3320.6038	-3326.2591	-3320.3643	-3324.5549	8.381178	0.0037913	11.789512	0.0027538	7.87252	0.019521	3.215	4.61387	2	65_L_0.959,75_R_0.974
TMEM81	-2591.2969	-2591.2469	-2591.2483	-2591.2483	0	1	-0.002688							
TMEM95	-1581.4772	-1582.1353	-1581.4822	-1581.8515	0.738684	0.39008	1.30619	0.52043	0.73575	0.6922				

Table III.1. Positive selection analysis of the *Murinae* fertilization complex. An expanded table of the *Murinae* positive selection analysis presented in Fig. S5D. Log likelihood values were used for chi-square statistical comparisons between models allowing or not allowing for positive selection.

References

1. C. A. Buck, A. F. Horwitz, Cell surface receptors for extracellular matrix molecules. *Annu. Rev. Cell Biol.* **3**, 179–205 (1987).
2. M. Barczyk, S. Carracedo, D. Gullberg, Integrins. *Cell Tissue Res.* **339**, 269–280 (2010).
3. R. O. Hynes, Integrins: bidirectional, allosteric signaling machines. *Cell* **110**, 673–687 (2002).
4. Y. Takada, X. Ye, S. Simon, The integrins. *Genome Biol.* **8**, 215 (2007).
5. M. Shimaoka, J. Takagi, T. A. Springer, Conformational regulation of integrin structure and function. *Annu. Rev. Biophys. Biomol. Struct.* **31**, 485–516 (2002).
6. I. D. Campbell, M. J. Humphries, Integrin structure, activation, and interactions. *Cold Spring Harb. Perspect. Biol.* **3**, a004994 (2011).
7. D. Gullberg, *I Domain Integrins* (Not Avail, ed. 2014, 2014; <http://dx.doi.org/10.1007/978-94-017-9153-3>) *Advances in Experimental Medicine and Biology*.
8. F. Hogervorst, I. Kuikman, A. E. von dem Borne, A. Sonnenberg, Cloning and sequence analysis of beta-4 cDNA: an integrin subunit that contains a unique 118 kd cytoplasmic domain. *EMBO J.* **9**, 765–770 (1990).
9. S. Suzuki, Y. Naitoh, Amino acid sequence of a novel integrin beta 4 subunit and primary expression of the mRNA in epithelial cells. *EMBO J.* **9**, 757–763 (1990).
10. J. C. Loftus, C. E. Halloran, M. H. Ginsberg, L. P. Feigen, J. A. Zablocki, J. W. Smith, The amino-terminal one-third of alpha IIb defines the ligand recognition specificity of integrin alpha IIb beta 3. *J. Biol. Chem.* **271**, 2033–2039 (1996).
11. J. D. Humphries, A. Byron, M. J. Humphries, Integrin ligands at a glance. *J. Cell Sci.* **119**, 3901–3903 (2006).
12. J. Takagi, D. P. DeBottis, H. P. Erickson, T. A. Springer, The role of the specificity-determining loop of the integrin beta subunit I-like domain in autonomous expression, association with the alpha subunit, and ligand binding. *Biochemistry* **41**, 4339–4347 (2002).
13. X. Dong, N. E. Hudson, C. Lu, T. A. Springer, Structural determinants of integrin β -subunit specificity for latent TGF- β . *Nat. Struct. Mol. Biol.* **21**, 1091–1096 (2014).
14. S. Schumacher, D. Dedden, R. V. Nunez, K. Matoba, J. Takagi, C. Biertümpfel, N. Mizuno, Structural insights into integrin $\alpha_5\beta_1$ opening by fibronectin ligand. *Science Advances* **7**, eabe9716 (2021).

15. J. P. Xiong, T. Stehle, B. Diefenbach, R. Zhang, R. Dunker, D. L. Scott, A. Joachimiak, S. L. Goodman, M. A. Arnaout, Crystal structure of the extracellular segment of integrin α V β 3. *Science* **294**, 339–345 (2001).
16. N. Miyazaki, K. Iwasaki, J. Takagi, A systematic survey of conformational states in β 1 and β 4 integrins using negative-stain electron microscopy. *J. Cell Sci.* **131** (2018).
17. J. Li, Y. Su, W. Xia, Y. Qin, M. J. Humphries, D. Vestweber, C. Cabañas, C. Lu, T. A. Springer, Conformational equilibria and intrinsic affinities define integrin activation. *EMBO J.* **36**, 629–645 (2017).
18. J. Takagi, B. M. Petre, T. Walz, T. A. Springer, Global conformational rearrangements in integrin extracellular domains in outside-in and inside-out signaling. *Cell* **110**, 599–511 (2002).
19. M. V. Nermut, N. M. Green, P. Eason, S. S. Yamada, K. M. Yamada, Electron microscopy and structural model of human fibronectin receptor. *EMBO J.* **7**, 4093–4099 (1988).
20. X. Du, M. Gu, J. W. Weisel, C. Nagaswami, J. S. Bennett, R. Bowditch, M. H. Ginsberg, Long range propagation of conformational changes in integrin α IIb β 3. *J. Biol. Chem.* **268**, 23087–23092 (1993).
21. B.-H. Luo, C. V. Carman, T. A. Springer, Structural basis of integrin regulation and signaling. *Annu. Rev. Immunol.* **25**, 619–647 (2007).
22. C. A. Lowell, T. N. Mayadas, Overview: studying integrins in vivo. *Methods Mol. Biol.* **757**, 369–397 (2012).
23. A. Cormier, M. G. Campbell, S. Ito, S. Wu, J. Lou, J. Marks, J. L. Baron, S. L. Nishimura, Y. Cheng, Cryo-EM structure of the α v β 8 integrin reveals a mechanism for stabilizing integrin extension. *Nat. Struct. Mol. Biol.* **25**, 698–704 (2018).
24. M. G. Campbell, A. Cormier, S. Ito, R. I. Seed, A. J. Bondesson, J. Lou, J. D. Marks, J. L. Baron, Y. Cheng, S. L. Nishimura, Cryo-EM reveals integrin-mediated TGF- β activation without release from latent TGF- β . *Cell* **180**, 490-501.e16 (2020).
25. J.-P. Xiong, T. Stehle, S. L. Goodman, M. A. Arnaout, Integrins, cations and ligands: making the connection. *J. Thromb. Haemost.* **1**, 1642–1654 (2003).
26. K. Zhang, J. Chen, The regulation of integrin function by divalent cations. *Cell Adh. Migr.* **6**, 20–29 (2012).
27. G. Bazzoni, L. Ma, M. L. Blue, M. E. Hemler, Divalent cations and ligands induce conformational changes that are highly divergent among β 1 integrins. *J. Biol. Chem.* **273**, 6670–6678 (1998).

28. D. A. Wilcox, J. L. Wautier, D. Pidard, P. J. Newman, A single amino acid substitution flanking the fourth calcium binding domain of alpha IIb prevents maturation of the alpha IIb beta 3 integrin complex. *J. Biol. Chem.* **269**, 4450–4457 (1994).
29. A. Masumoto, M. E. Hemler, Mutation of putative divalent cation sites in the alpha 4 subunit of the integrin VLA-4: distinct effects on adhesion to CS1/fibronectin, VCAM-1, and invasin. *J. Cell Biol.* **123**, 245–253 (1993).
30. W. B. Mitchell, J. H. Li, F. Singh, A. D. Michelson, J. Bussel, B. S. Collier, D. L. French, Two novel mutations in the alpha IIb calcium-binding domains identify hydrophobic regions essential for alpha IIbbeta 3 biogenesis. *Blood* **101**, 2268–2276 (2003).
31. T. Xiao, J. Takagi, B. S. Collier, J.-H. Wang, T. A. Springer, Structural basis for allostery in integrins and binding to fibrinogen-mimetic therapeutics. *Nature* **432**, 59–67 (2004).
32. J. Chen, A. Salas, T. A. Springer, Bistable regulation of integrin adhesiveness by a bipolar metal ion cluster. *Nat. Struct. Biol.* **10**, 995–1001 (2003).
33. J. Raborn, W. Wang, B.-H. Luo, Regulation of integrin $\alpha\text{IIb}\beta 3$ ligand binding and signaling by the metal ion binding sites in the β I domain. *Biochemistry* **50**, 2084–2091 (2011).
34. J. Zhu, B.-H. Luo, T. Xiao, C. Zhang, N. Nishida, T. A. Springer, Structure of a complete integrin ectodomain in a physiologic resting state and activation and deactivation by applied forces. *Mol. Cell* **32**, 849–861 (2008).
35. D. Valdramidou, M. J. Humphries, A. P. Mould, Distinct roles of beta1 metal ion-dependent adhesion site (MIDAS), adjacent to MIDAS (ADMIDAS), and ligand-associated metal-binding site (LIMBS) cation-binding sites in ligand recognition by integrin alpha2beta1. *J. Biol. Chem.* **283**, 32704–32714 (2008).
36. J. Chen, J. Takagi, C. Xie, T. Xiao, B.-H. Luo, T. A. Springer, The relative influence of metal ion binding sites in the I-like domain and the interface with the hybrid domain on rolling and firm adhesion by integrin alpha4beta7. *J. Biol. Chem.* **279**, 55556–55561 (2004).
37. E. C. Tozer, R. C. Liddington, M. J. Sutcliffe, A. H. Smeeton, J. C. Loftus, Ligand binding to integrin alphaIIbbeta3 is dependent on a MIDAS-like domain in the beta3 subunit. *J. Biol. Chem.* **271**, 21978–21984 (1996).
38. Y. Pan, K. Zhang, J. Qi, J. Yue, T. A. Springer, J. Chen, Cation-pi interaction regulates ligand-binding affinity and signaling of integrin alpha4beta7. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 21388–21393 (2010).
39. E. J. Park, J. R. Mora, C. V. Carman, J. Chen, Y. Sasaki, G. Cheng, U. H. von Andrian, M. Shimaoka, Aberrant activation of integrin alpha4beta7 suppresses lymphocyte migration to the gut. *J. Clin. Invest.* **117**, 2526–2538 (2007).
40. J. Zhu, J. Zhu, T. A. Springer, Complete integrin headpiece opening in eight steps. *J. Cell Biol.* **201**, 1053–1068 (2013).

41. J. Li, M. H. Jo, J. Yan, T. Hall, J. Lee, U. López-Sánchez, S. Yan, T. Ha, T. A. Springer, Ligand binding initiates single-molecule integrin conformational activation. *Cell* **187**, 2990–3005.e17 (2024).
42. T. N. Durrant, M. T. van den Bosch, I. Hers, Integrin α IIb β 3 outside-in signaling. *Blood* **130**, 1607–1619 (2017).
43. O. Vinogradova, A. Velyvis, A. Velyviene, B. Hu, T. Haas, E. Plow, J. Qin, A structural mechanism of integrin α IIb β 3 “inside-out” activation as regulated by its cytoplasmic face. *Cell* **110**, 587–597 (2002).
44. T. A. Springer, M. L. Dustin, Integrin inside-out signaling and the immunological synapse. *Curr. Opin. Cell Biol.* **24**, 107–115 (2012).
45. C. L. Abram, C. A. Lowell, The ins and outs of leukocyte integrin signaling. *Annu. Rev. Immunol.* **27**, 339–362 (2009).
46. B. D. Adair, J.-P. Xiong, M. Yeager, M. A. Arnaout, Cryo-EM structures of full-length integrin α IIb β 3 in native lipids. *Nat. Commun.* **14**, 1–7 (2023).
47. T. Huo, H. Wu, Z. Moussa, M. Sen, V. Dalton, Z. Wang, Full-length α IIb β 3 cryo-EM structure reveals intact integrin initiate-activation intrinsic architecture. *Structure* **32**, 899–906.e3 (2024).
48. J. Takagi, Structural basis for ligand recognition by integrins. *Curr. Opin. Cell Biol.* **19**, 557–564 (2007).
49. A. Sonnenberg, Integrins and their ligands. *Curr. Top. Microbiol. Immunol.* **184**, 7–35 (1993).
50. K. Burridge, Focal adhesions: a personal perspective on a half century of progress. *FEBS J.* **284**, 3355–3361 (2017).
51. K. Legerstee, A. B. Houtsmuller, A layered view on focal Adhesions. *Biology (Basel)* **10**, 1189 (2021).
52. J. L. Sechler, S. A. Corbett, J. E. Schwarzbauer, Modulatory roles for integrin activation and the synergy site of fibronectin during matrix assembly. *Mol. Biol. Cell* **8**, 2563–2573 (1997).
53. M. A. Horton, The α v β 3 integrin “vitronectin receptor.” *Int. J. Biochem. Cell Biol.* **29**, 721–725 (1997).
54. A. M. Belkin, M. A. Stepp, Integrins as receptors for laminins. *Microsc. Res. Tech.* **51**, 280–301 (2000).
55. K. R. Gehlsen, L. Dillner, E. Engvall, E. Ruoslahti, The human laminin receptor is a member of the integrin family of cell adhesion receptors. *Science* **241**, 1228–1229 (1988).

56. T. Kokubo, H. Uchida, E. T. Choi, Integrin $\alpha(v)\beta(3)$ as a target in the prevention of neointimal hyperplasia. *J. Vasc. Surg.* **45 Suppl A**, A33-8 (2007).
57. K. A. Kwakwa, J. A. Sterling, Integrin $\alpha\beta3$ signaling in tumor-induced bone disease. *Cancers (Basel)* **9**, E84 (2017).
58. Y. Gu, B. Dong, X. He, Z. Qiu, J. Zhang, M. Zhang, H. Liu, X. Pang, Y. Cui, The challenges and opportunities of $\alpha\beta3$ -based therapeutics in cancer: From bench to clinical trials. *Pharmacol. Res.* **189**, 106694 (2023).
59. L. M. Memmo, P. McKeown-Longo, The $\alpha v\beta 5$ integrin functions as an endocytic receptor for vitronectin. *J. Cell Sci.* **111 (Pt 4)**, 425–433 (1998).
60. Y. Oishi, K. Hashimoto, A. Abe, M. Kuroda, A. Fujii, Y. Miyamoto, Vitronectin regulates the axon specification of mouse cerebellar granule cell precursors via $\alpha\beta5$ integrin in the differentiation stage. *Neurosci. Lett.* **746**, 135648 (2021).
61. A. L. Tatler, A. E. John, L. Jolly, A. Habgood, J. Porte, C. Brightling, A. J. Knox, L. Pang, D. Sheppard, X. Huang, G. Jenkins, Integrin $\alpha\beta5$ -mediated TGF- β activation by airway smooth muscle cells in asthma. *J. Immunol.* **187**, 6094–6107 (2011).
62. Y. Asano, H. Ihn, M. Jinnin, Y. Mimura, K. Tamaki, Involvement of $\alpha v\beta 5$ integrin in the establishment of autocrine TGF- β signaling in dermal fibroblasts derived from localized scleroderma. *J. Invest. Dermatol.* **126**, 1761–1769 (2006).
63. A. K. Kumawat, C. Yu, E. A. Mann, A. Schridde, S. C. Finnemann, A. M. Mowat, Expression and characterization of $\alpha\beta5$ integrin on intestinal macrophages. *Eur. J. Immunol.* **48**, 1181–1187 (2018).
64. D. Mu, S. Cambier, L. Fjellbirkeland, J. L. Baron, J. S. Munger, H. Kawakatsu, D. Sheppard, V. C. Broaddus, S. L. Nishimura, The integrin $\alpha(v)\beta8$ mediates epithelial homeostasis through MT1-MMP-dependent activation of TGF- β 1. *J. Cell Biol.* **157**, 493–507 (2002).
65. J. S. Munger, X. Huang, H. Kawakatsu, M. J. Griffiths, S. L. Dalton, J. Wu, J. F. Pittet, N. Kaminski, C. Garat, M. A. Matthay, D. B. Rifkin, D. Sheppard, The integrin $\alpha v\beta 6$ binds and activates latent TGF β 1: a mechanism for regulating pulmonary inflammation and fibrosis. *Cell* **96**, 319–328 (1999).
66. H. Hendesi, M. F. Barbe, F. F. Safadi, M. A. Monroy, S. N. Popoff, Integrin mediated adhesion of osteoblasts to connective tissue growth factor (CTGF/CCN2) induces cytoskeleton reorganization and cell differentiation. *PLoS One* **10**, e0115325 (2015).
67. N. I. Reed, H. Jo, C. Chen, K. Tsujino, T. D. Arnold, W. F. DeGrado, D. Sheppard, The $\alpha\beta1$ integrin plays a critical in vivo role in tissue fibrosis. *Sci. Transl. Med.* **7**, 288ra79 (2015).

68. Z. Han, Y. Ma, G. Cao, Z. Ma, R. Chen, M. E. Cvijic, D. Cheng, Integrin $\alpha V\beta 1$ regulates procollagen I production through a non-canonical transforming growth factor β signaling pathway in human hepatic stellate cells. *Biochem. J.* **478**, 1689–1703 (2021).
69. I. Ezzat, M. Zallocchi, Integrin Alpha8 Beta1 (81): An in-depth review of an overlooked RGD-binding receptor. *Biocell* **49**, 789–811 (2025).
70. J. S. Bennett, Structure and function of the platelet integrin $\alpha IIb\beta 3$. *J. Clin. Invest.* **115**, 3363–3369 (2005).
71. K. Bledzka, S. S. Smyth, E. F. Plow, Integrin $\alpha IIb\beta 3$: from discovery to efficacious therapeutic target: From discovery to efficacious therapeutic target. *Circ. Res.* **112**, 1189–1200 (2013).
72. R. Nishiuchi, J. Takagi, M. Hayashi, H. Ido, Y. Yagi, N. Sanzen, T. Tsuji, M. Yamada, K. Sekiguchi, Ligand-binding specificities of laminin-binding integrins: a comprehensive survey of laminin-integrin interactions using recombinant $\alpha 3\beta 1$, $\alpha 6\beta 1$, $\alpha 7\beta 1$ and $\alpha 6\beta 4$ integrins. *Matrix Biol.* **25**, 189–197 (2006).
73. M. Yamada, K. Sekiguchi, Molecular basis of laminin–integrin interactions. *Current topics in membranes* **76**, 197–229 (2015).
74. J. A. Kreidberg, M. J. Donovan, S. L. Goldstein, H. Rennke, K. Shepherd, R. C. Jones, R. Jaenisch, Alpha 3 beta 1 integrin has a crucial role in kidney and lung organogenesis. *Development* **122**, 3537–3547 (1996).
75. J. A. Kreidberg, Functions of $\alpha 3\beta 1$ integrin. *Curr. Opin. Cell Biol.* **12**, 548–553 (2000).
76. N. L. Flintoff-Dye, J. Welser, J. Rooney, P. Scowen, S. Tamowski, W. Hatton, D. J. Burkin, Role for the $\alpha 7\beta 1$ integrin in vascular development and integrity. *Dev. Dyn.* **234**, 11–21 (2005).
77. U. Mayer, G. Saher, R. Fässler, A. Bornemann, F. Echtermeyer, H. von der Mark, N. Miosge, E. Pöschl, K. von der Mark, Absence of integrin $\alpha 7$ causes a novel form of muscular dystrophy. *Nat. Genet.* **17**, 318–323 (1997).
78. L. E. Reynolds, G. D’Amico, T. Lechertier, A. Papachristodoulou, J. M. Muñoz-Félix, A. De Arcangelis, M. Baker, B. Serrels, K. M. Hodivala-Dilke, Dual role of pericyte $\alpha 6\beta 1$ -integrin in tumour blood vessels. *J. Cell Sci.* **130**, 1583–1595 (2017).
79. G. O. Delwel, A. A. de Melker, F. Hogervorst, L. H. Jaspars, D. L. Fles, I. Kuikman, A. Lindblom, M. Paulsson, R. Timpl, A. Sonnenberg, Distinct and overlapping ligand specificities of the $\alpha 3A\beta 1$ and $\alpha 6A\beta 1$ integrins: recognition of laminin isoforms. *Mol. Biol. Cell* **5**, 203–215 (1994).

80. N. Ahmed, C. Riley, K. Oliva, G. Rice, M. Quinn, Ascites induces modulation of $\alpha 6 \beta 1$ integrin and urokinase plasminogen activator receptor expression and associated functions in ovarian carcinoma. *Br. J. Cancer* **92**, 1475–1485 (2005).
81. A. M. Mercurio, I. Rabinovitz, L. M. Shaw, The $\alpha 6 \beta 4$ integrin and epithelial cell migration. *Curr. Opin. Cell Biol.* **13**, 541–545 (2001).
82. R. L. Stewart, K. L. O'Connor, Clinical significance of the integrin $\alpha 6 \beta 4$ in human malignancies. *Lab. Invest.* **95**, 976–986 (2015).
83. J. Heino, Cellular signaling by collagen-binding integrins. *Adv. Exp. Med. Biol.* **819**, 143–155 (2014).
84. J. Elango, C. Hou, B. Bao, S. Wang, J. E. Maté Sánchez de Val, W. Wenhui, The molecular interaction of collagen with cell receptors for biological function. *Polymers (Basel)* **14**, 876 (2022).
85. C. Zeltz, D. Gullberg, The integrin-collagen connection--a glue for tissue repair? *J. Cell Sci.* **129**, 653–664 (2016).
86. A. Madamanchi, S. A. Santoro, M. M. Zutter, $\alpha 2 \beta 1$ integrin. *Adv. Exp. Med. Biol.* **819**, 41–60 (2014).
87. S. W. Hamaia, N. Pugh, N. Raynal, B. Némoy, R. Stone, D. Gullberg, D. Bihan, R. W. Farndale, Mapping of potent and specific binding motifs, GLOGEN and GVOGEA, for integrin $\alpha 1 \beta 1$ using collagen toolkits II and III. *J. Biol. Chem.* **287**, 26019–26028 (2012).
88. H. Gardner, Integrin $\alpha 1 \beta 1$. *Adv. Exp. Med. Biol.* **819**, 21–39 (2014).
89. M. M. Barczyk, L.-H. B. Olsen, P. da Franca, B. G. Loos, K. Mustafa, D. Gullberg, A. I. Bolstad, A role for $\alpha 1 \beta 1$ integrin in the human periodontal ligament. *J. Dent. Res.* **88**, 621–626 (2009).
90. J.-N. Schulz, C. Zeltz, I. W. Sørensen, M. Barczyk, S. Carracedo, R. Hallinger, A. Niehoff, B. Eckes, D. Gullberg, Reduced granulation tissue and wound strength in the absence of $\alpha 1 \beta 1$ integrin. *J. Invest. Dermatol.* **135**, 1435–1444 (2015).
91. E. Lundgren-Åkerlund, A. Aszodi, Integrin $\alpha 10 \beta 1$: a collagen receptor critical in skeletal development. *Adv. Exp. Med. Biol.* **819**, 61–71 (2014).
92. T. Bengtsson, A. Aszodi, C. Nicolae, E. B. Hunziker, E. Lundgren-Akerlund, R. Fässler, Loss of $\alpha 10 \beta 1$ integrin expression leads to moderate dysfunction of growth plate chondrocytes. *J. Cell Sci.* **118**, 929–936 (2005).
93. C. G. Knight, L. F. Morton, A. R. Peachey, D. S. Tuckwell, R. W. Farndale, M. J. Barnes, The collagen-binding A-domains of integrins $\alpha (1) \beta (1)$ and $\alpha (2) \beta (1)$ recognize the same specific amino acid sequence, GFOGER, in native (triple-helical) collagens. *J. Biol. Chem.* **275**, 35–40 (2000).

94. I. Mitroulis, V. I. Alexaki, I. Kourtzelis, A. Ziogas, G. Hajishengallis, T. Chavakis, Leukocyte integrins: role in leukocyte recruitment and as therapeutic targets in inflammatory disease. *Pharmacol. Ther.* **147**, 123–135 (2015).
95. G. Gorfu, J. Rivera-Nieves, K. Ley, Role of beta7 integrins in intestinal lymphocyte homing and retention. *Curr. Mol. Med.* **9**, 836–850 (2009).
96. C. Berlin, E. L. Berg, M. J. Briskin, D. P. Andrew, P. J. Kilshaw, B. Holzmann, I. L. Weissman, A. Hamann, E. C. Butcher, Alpha 4 beta 7 integrin mediates lymphocyte binding to the mucosal vascular addressin MAdCAM-1. *Cell* **74**, 185–195 (1993).
97. K. L. Cepek, S. K. Shaw, C. M. Parker, G. J. Russell, J. S. Morrow, D. L. Rimm, M. B. Brenner, Adhesion between epithelial cells and T lymphocytes mediated by E-cadherin and the alpha E beta 7 integrin. *Nature* **372**, 190–193 (1994).
98. C. Wu, A. J. Fields, B. A. Kapteijn, J. A. McDonald, The role of alpha 4 beta 1 integrin in cell motility and fibronectin matrix assembly. *J. Cell Sci.* **108 (Pt 2)**, 821–829 (1995).
99. T. D. Schreiber, C. Steinl, M. Essl, H. Abele, K. Geiger, C. A. Müller, W. K. Aicher, G. Klein, The integrin alpha9beta1 on hematopoietic stem and progenitor cells: involvement in cell adhesion, proliferation and differentiation. *Haematologica* **94**, 1493–1501 (2009).
100. S. Xu, T. Zhang, Z. Cao, W. Zhong, C. Zhang, H. Li, J. Song, Integrin- $\alpha 9\beta 1$ as a novel therapeutic target for refractory diseases: Recent progress and insights. *Front. Immunol.* **12**, 638400 (2021).
101. S. C. Fagerholm, C. Guenther, M. Lloret Asens, T. Savinko, L. M. Uotila, Beta2-integrins and interacting proteins in leukocyte trafficking, immune suppression, and immunodeficiency disease. *Front. Immunol.* **10**, 254 (2019).
102. M. Bednarczyk, H. Stege, S. Grabbe, M. Bros, B2 integrins-multi-functional leukocyte receptors in health and disease. *Int. J. Mol. Sci.* **21**, E1402 (2020).
103. C. Lamers, C. J. Plüss, D. Ricklin, The promiscuous profile of complement receptor 3 in ligand binding, immune modulation, and pathophysiology. *Front. Immunol.* **12**, 662164 (2021).
104. T. Vorup-Jensen, R. K. Jensen, Structural immunology of complement receptors 3 and 4. *Front. Immunol.* **9**, 2716 (2018).
105. H.-W. Wang, J.-W. Wang, How cryo-electron microscopy and X-ray crystallography complement each other: Cryo-EM and X-Ray Crystallography Complement Each Other. *Protein Sci.* **26**, 32–39 (2017).
106. B. S. Ludwig, H. Kessler, S. Kossatz, U. Reuning, RGD-binding integrins revisited: How recently discovered functions and novel synthetic ligands (re-)shape an ever-evolving field. *Cancers (Basel)* **13**, 1711 (2021).

107. M. D. Pierschbacher, E. Ruoslahti, Cell attachment activity of fibronectin can be duplicated by small synthetic fragments of the molecule. *Nature* **309**, 30–33 (1984).
108. R. O. Hynes, Integrins: a family of cell surface receptors. *Cell* **48**, 549–554 (1987).
109. R. Pytela, M. D. Pierschbacher, E. Ruoslahti, Identification and isolation of a 140 kd cell surface glycoprotein with properties expected of a fibronectin receptor. *Cell* **40**, 191–198 (1985).
110. E. Ruoslahti, M. D. Pierschbacher, New perspectives in cell adhesion: RGD and integrins. *Science* **238**, 491–497 (1987).
111. J.-P. Xiong, T. Stehle, R. Zhang, A. Joachimiak, M. Frech, S. L. Goodman, M. A. Arnaout, Crystal structure of the extracellular segment of integrin α V β 3 in complex with an Arg-Gly-Asp ligand. *Science* **296**, 151–155 (2002).
112. M. Nagae, S. Re, E. Mihara, T. Nogi, Y. Sugita, J. Takagi, Crystal structure of α 5 β 1 integrin ectodomain: atomic details of the fibronectin receptor. *J. Cell Biol.* **197**, 131–140 (2012).
113. T. G. Kapp, F. Rechenmacher, S. Neubauer, O. V. Maltsev, E. A. Cavalcanti-Adam, R. Zarka, U. Reuning, J. Notni, H.-J. Wester, C. Mas-Moruno, J. Spatz, B. Geiger, H. Kessler, A comprehensive evaluation of the activity and selectivity profile of ligands for RGD-binding integrins. *Sci. Rep.* **7**, 39805 (2017).
114. X. Dong, B. Zhao, R. E. Iacob, J. Zhu, A. C. Koksall, C. Lu, J. R. Engen, T. A. Springer, Force interacts with macromolecular structure in activation of TGF- β . *Nature* **542**, 55–59 (2017).
115. M. Jin, R. I. Seed, G. Cai, T. Shing, L. Wang, S. Ito, A. Cormier, S. A. Wankowicz, J. M. Jespersen, J. L. Baron, N. D. Carey, M. G. Campbell, Z. Yu, P. K. Tang, P. Cossio, W. Wen, J. Lou, J. Marks, S. L. Nishimura, Y. Cheng, Dynamic allostery drives autocrine and paracrine TGF- β signaling. *Cell* **0** (2024).
116. M. Takizawa, T. Arimori, Y. Taniguchi, Y. Kitago, E. Yamashita, J. Takagi, K. Sekiguchi, Mechanistic basis for the recognition of laminin-511 by α 6 β 1 integrin. *Sci. Adv.* **3**, e1701497 (2017).
117. T. Arimori, N. Miyazaki, E. Mihara, M. Takizawa, Y. Taniguchi, C. Cabañas, K. Sekiguchi, J. Takagi, Structural mechanism of laminin recognition by integrin. *Nat. Commun.* **12**, 4012 (2021).
118. J. A. Hollis, M. C. Chan, H. S. Malik, M. G. Campbell, Molecular exaptation by the integrin α I domain. *Sci. Adv.* **11**, eadx9567 (2025).
119. T. Springer, U. López-Sánchez, M. Monroy, Q. Qiao, J. Lee, S. Kesarwani, Z. Zhou, A. Brown, F.-Y. Lin, Conformational landscape of integrin α 4 β 1 and its recognition of VCAM

and conformation-stabilizing small molecules, *Research Square* (2026).
<https://doi.org/10.21203/rs.3.rs-8928859/v1>.

120. F. G. Giancotti, E. Ruoslahti, Integrin signaling. *Science* **285**, 1028–1032 (1999).
121. D. S. Harburger, D. A. Calderwood, Integrin signalling at a glance. *J. Cell Sci.* **122**, 159–163 (2009).
122. S. K. Sastry, A. F. Horwitz, Integrin cytoplasmic domains: mediators of cytoskeletal linkages and extra- and intracellular initiated transmembrane signaling. *Curr. Opin. Cell Biol.* **5**, 819–831 (1993).
123. S. Liu, D. A. Calderwood, M. H. Ginsberg, Integrin cytoplasmic domain-binding proteins. *J. Cell Sci.* **113** (Pt 20), 3563–3571 (2000).
124. D. A. Calderwood, Y. Fujioka, J. M. de Pereda, B. García-Alvarez, T. Nakamoto, B. Margolis, C. J. McGlade, R. C. Liddington, M. H. Ginsberg, Integrin beta cytoplasmic domain interactions with phosphotyrosine-binding domains: a structural prototype for diversity in integrin signaling. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 2272–2277 (2003).
125. B. Klapholz, N. H. Brown, Talin - the master of integrin adhesions. *J. Cell Sci.* **130**, 2435–2446 (2017).
126. K. L. Wegener, A. W. Partridge, J. Han, A. R. Pickford, R. C. Liddington, M. H. Ginsberg, I. D. Campbell, Structural basis of integrin activation by talin. *Cell* **128**, 171–182 (2007).
127. P. Zhang, L. Azizi, S. Kukkurainen, T. Gao, M. Baikoghli, M.-C. Jacquier, Y. Sun, J. A. E. Määttä, R. H. Cheng, B. Wehrle-Haller, V. P. Hytönen, J. Wu, Crystal structure of the FERM-folded talin head reveals the determinants for integrin binding. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 32402–32412 (2020).
128. N. J. Anthis, K. L. Wegener, F. Ye, C. Kim, B. T. Goult, E. D. Lowe, I. Vakonakis, N. Bate, D. R. Critchley, M. H. Ginsberg, I. D. Campbell, The structure of an integrin/talin complex reveals the basis of inside-out signal transduction. *EMBO J.* **28**, 3623–3632 (2009).
129. E. Rognoni, R. Ruppert, R. Fässler, The kindlin family: functions, signaling properties and implications for human disease. *J. Cell Sci.* **129**, 17–27 (2016).
130. Z. Sun, M. Costell, R. Fässler, Integrin activation by talin, kindlin and mechanical forces. *Nat. Cell Biol.* **21**, 25–31 (2019).
131. H. Li, Y. Deng, K. Sun, H. Yang, J. Liu, M. Wang, Z. Zhang, J. Lin, C. Wu, Z. Wei, C. Yu, Structural basis of kindlin-mediated integrin recognition and activation. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 9349–9354 (2017).

132. J. Aretz, M. Aziz, N. Strohmeyer, M. Sattler, R. Fässler, Talin and kindlin use integrin tail allostery and direct binding to activate integrins. *Nat. Struct. Mol. Biol.* **30**, 1913–1924 (2023).
133. C. G. Gahmberg, M. Grönholm, L. M. Uotila, Regulation of integrin activity by phosphorylation. *Adv. Exp. Med. Biol.* **819**, 85–96 (2014).
134. C. G. Gahmberg, M. Grönholm, How integrin phosphorylations regulate cell adhesion and signaling. *Trends Biochem. Sci.* **47**, 265–278 (2022).
135. E. G. Arias-Salgado, S. Lizano, S. Sarkar, J. S. Brugge, M. H. Ginsberg, S. J. Shattil, Src kinase activation by direct interaction with the integrin beta cytoplasmic domain. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 13298–13302 (2003).
136. M. Illario, A. L. Cavallo, K. U. Bayer, T. Di Matola, G. Fenzi, G. Rossi, M. Vitale, Calcium/calmodulin-dependent protein kinase II binds to Raf-1 and modulates integrin-stimulated ERK activation. *J. Biol. Chem.* **278**, 45101–45108 (2003).
137. E. M. Morse, N. N. Brahme, D. A. Calderwood, Integrin cytoplasmic tail interactions. *Biochemistry* **53**, 810–820 (2014).
138. J. Liu, M. Das, J. Yang, S. S. Ithychanda, V. P. Yakubenko, E. F. Plow, J. Qin, Structural mechanism of integrin inactivation by filamin. *Nat. Struct. Mol. Biol.* **22**, 383–389 (2015).
139. C. L. Oxley, N. J. Anthis, E. D. Lowe, I. Vakonakis, I. D. Campbell, K. L. Wegener, An integrin phosphorylation switch: the effect of beta3 integrin tail phosphorylation on Dok1 and talin binding: The effect of $\beta 3$ integrin tail phosphorylation on dok1 and Talin binding. *J. Biol. Chem.* **283**, 5420–5426 (2008).
140. S. Gupta, J. C.-Y. Chit, C. Feng, A. Bhunia, S.-M. Tan, S. Bhattacharjya, An alternative phosphorylation switch in integrin $\beta 2$ (CD18) tail for Dok1 binding. *Sci. Rep.* **5**, 11630 (2015).
141. A. Dai, F. Ye, D. W. Taylor, G. Hu, M. H. Ginsberg, K. A. Taylor, The structure of a full-length membrane-embedded integrin bound to a physiological ligand. *J. Biol. Chem.* **290**, 27168–27175 (2015).
142. Z. Sun, S. S. Guo, R. Fässler, Integrin-mediated mechanotransduction. *J. Cell Biol.* **215**, 445–456 (2016).
143. K. Katoh, Integrin and its associated proteins as a mediator for mechano-signal transduction. *Biomolecules* **15**, 166 (2025).
144. D. Tuckwell, Evolution of von Willebrand factor A (VWA) domains. *Biochem. Soc. Trans.* **27**, 835–840 (1999).

145. C. A. Whittaker, R. O. Hynes, Distribution and evolution of von Willebrand/integrin A domains: widely dispersed domains with roles in cell adhesion and elsewhere. *Mol. Biol. Cell* **13**, 3369–3387 (2002).
146. M. S. Johnson, B. S. Chouhan, Evolution of integrin I domains. *Adv. Exp. Med. Biol.* **819**, 1–19 (2014).
147. M. Michishita, V. Videm, M. A. Arnaout, A novel divalent cation-binding site in the A domain of the beta 2 integrin CR3 (CD11b/CD18) is essential for ligand binding. *Cell* **72**, 857–867 (1993).
148. M. S. Diamond, J. Garcia-Aguilar, J. K. Bickford, A. L. Corbi, T. A. Springer, The I domain is a major recognition site on the leukocyte integrin Mac-1 (CD11b/CD18) for four distinct adhesion ligands. *J. Cell Biol.* **120**, 1031–1043 (1993).
149. A. M. Randi, N. Hogg, I domain of beta 2 integrin lymphocyte function-associated antigen-1 contains a binding site for ligand intercellular adhesion molecule-1. *J. Biol. Chem.* **269**, 12395–12398 (1994).
150. T. Kamata, W. Puzon, Y. Takada, Identification of putative ligand binding sites within I domain of integrin alpha 2 beta 1 (VLA-2, CD49b/CD29). *J. Biol. Chem.* **269**, 9659–9663 (1994).
151. A. Kern, R. Briesewitz, I. Bank, E. E. Marcantonio, The role of the I domain in ligand binding of the human integrin alpha 1 beta 1. *J. Biol. Chem.* **269**, 22811–22816 (1994).
152. J. Emsley, S. L. King, J. M. Bergelson, R. C. Liddington, Crystal structure of the I domain from integrin alpha2beta1. *J. Biol. Chem.* **272**, 28512–28517 (1997).
153. J. O. Lee, L. A. Bankston, M. A. Arnaout, R. C. Liddington, Two conformations of the integrin A-domain (I-domain): a pathway for activation? *Structure* **3**, 1333–1340 (1995).
154. M. Nolte, R. B. Pepinsky, Venyaminov SYu, V. Koteliansky, P. J. Gotwals, M. Karpusas, Crystal structure of the alpha1beta1 integrin I-domain: insights into integrin I-domain function. *FEBS Lett.* **452**, 379–385 (1999).
155. M. Semmrich, A. Smith, C. Feterowski, S. Beer, B. Engelhardt, D. H. Busch, B. Bartsch, M. Laschinger, N. Hogg, K. Pfeffer, B. Holzmann, Importance of integrin LFA-1 deactivation for the generation of immune responses. *J. Exp. Med.* **201**, 1987–1998 (2005).
156. Y. Wang, D. Li, R. Nurieva, J. Yang, M. Sen, R. Carreño, S. Lu, B. W. McIntyre, J. J. Molldrem, G. B. Legge, Q. Ma, LFA-1 affinity regulation is necessary for the activation and proliferation of naive T cells. *J. Biol. Chem.* **284**, 12645–12653 (2009).
157. T. Schürpf, T. A. Springer, Regulation of integrin affinity on cell surfaces: Regulation of integrin affinity on cell surfaces. *EMBO J.* **30**, 4712–4727 (2011).

158. J. O. Lee, P. Rieu, M. A. Arnaout, R. Liddington, Crystal structure of the A domain from the alpha subunit of integrin CR3 (CD11b/CD18). *Cell* **80**, 631–638 (1995).
159. J. Emsley, C. G. Knight, R. W. Farndale, M. J. Barnes, R. C. Liddington, Structural basis of collagen recognition by integrin alpha2beta1. *Cell* **101**, 47–56 (2000).
160. C. Xie, J. Zhu, X. Chen, L. Mi, N. Nishida, T. A. Springer, Structure of an integrin with an alphaI domain, complement receptor type 4. *EMBO J.* **29**, 666–679 (2010).
161. M. Shi, V. Pedchenko, B. H. Greer, W. D. Van Horn, S. A. Santoro, C. R. Sanders, B. G. Hudson, B. F. Eichman, R. Zent, A. Pozzi, Enhancing integrin α 1 inserted (I) domain affinity to ligand potentiates integrin α 1 β 1-mediated down-regulation of collagen synthesis. *J. Biol. Chem.* **287**, 35139–35152 (2012).
162. M. Shimaoka, J. M. Shifman, H. Jing, J. Takagi, S. L. Mayo, T. A. Springer, Computational design of an integrin I domain stabilized in the open high affinity conformation. *Nat. Struct. Biol.* **7**, 674–678 (2000).
163. J. L. Alonso, M. Essafi, J. P. Xiong, T. Stehle, M. A. Arnaout, Does the integrin alphaA domain act as a ligand for its betaA domain? *Curr. Biol.* **12**, R340-2 (2002).
164. J. Lorentzen, M. U. Fruergaard, S. Lukácsi, M. H. Jørgensen, T. L. G. van Veghel, R. K. Jensen, K. J. Pietrzak-Lichwa, Z. Bajtay, V. Hořejší, R. K. Flygaard, D. Vorselen, S. A. Mortensen, G. R. Andersen, Three cryo-EM structures of complement C3d-bound α M β 2 reveal an unexpected layer of dynamics for α I-containing integrin receptors. *Sci. Adv.* **12**, eaea7241 (2026).
165. D. J. White, S. Puranen, M. S. Johnson, J. Heino, The collagen receptor subfamily of the integrins. *Int. J. Biochem. Cell Biol.* **36**, 1405–1410 (2004).
166. W.-M. Zhang, J. Kapyla, J. S. Puranen, C. G. Knight, C.-F. Tiger, O. T. Pentikainen, M. S. Johnson, R. W. Farndale, J. Heino, D. Gullberg, alpha 11beta 1 integrin recognizes the GFOGER sequence in interstitial collagens. *J. Biol. Chem.* **278**, 7270–7277 (2003).
167. P. R.-M. Siljander, S. Hamaia, A. R. Peachey, D. A. Slatter, P. A. Smethurst, W. H. Ouwehand, C. G. Knight, R. W. Farndale, Integrin activation state determines selectivity for novel recognition sites in fibrillar collagens. *J. Biol. Chem.* **279**, 47763–47772 (2004).
168. Y. K.-Y. Chin, S. J. Headey, B. Mohanty, R. Patil, P. A. McEwan, J. D. Swarbrick, T. D. Mulhern, J. Emsley, J. S. Simpson, M. J. Scanlon, The structure of integrin α II domain in complex with a collagen-mimetic peptide. *J. Biol. Chem.* **288**, 36796–36809 (2013).
169. M. Lahti, E. Bligt, H. Niskanen, V. Parkash, A.-M. Brandt, J. Jokinen, P. Patrikainen, J. Kämpylä, J. Heino, T. A. Salminen, Structure of collagen receptor integrin α (1)I domain carrying the activating mutation E317A. *J. Biol. Chem.* **286**, 43343–43351 (2011).

170. S. Shibata, H. Oki, K. Kawahara, R. Masuda, K. Fujii, T. Koide, Triple-helical ligands selectively targeting the closed α I domain of integrin α 2 β 1, *bioRxiv* (2025).
<https://doi.org/10.1101/2025.07.28.667101>.
171. R. W. Farndale, Collagen-binding proteins: insights from the Collagen Toolkits. *Essays Biochem.* **63**, 337–348 (2019).
172. M. Sen, K. Yuki, T. A. Springer, An internal ligand-bound, metastable state of a leukocyte integrin, α X β 2. *J. Cell Biol.* **203**, 629–642 (2013).
173. M. Shimaoka, T. Xiao, J.-H. Liu, Y. Yang, Y. Dong, C.-D. Jun, A. McCormack, R. Zhang, A. Joachimiak, J. Takagi, J.-H. Wang, T. A. Springer, Structures of the alpha L I domain and its complex with ICAM-1 reveal a shape-shifting pathway for integrin regulation. *Cell* **112**, 99–111 (2003).
174. M. Tulla, M. Huhtala, J. Jäälinoja, J. Kämpylä, R. W. Farndale, L. Ala-Kokko, M. S. Johnson, J. Heino, Analysis of an ascidian integrin provides new insight into early evolution of collagen recognition. *FEBS Lett.* **581**, 2434–2440 (2007).
175. D. Bouvard, C. Brakebusch, E. Gustafsson, A. Aszódi, T. Bengtsson, A. Berna, R. Fässler, Functional consequences of integrin gene mutations in mice. *Circ. Res.* **89**, 211–223 (2001).
176. W. B. M. de Lau, J. J. A. P. M. Wijnakker, G. J. F. van Son, D. Krueger, D. Wang, M. S. Abendroth, R. Schreurs, C. Y. Janda, F. L. H. van Rijt, B. J. P. Chalopin, E. Bokobza, J. H. van Es, T. A. Springer, A. Sonnenberg, H. Clevers, A single-chain derivative of an integrin-activating antibody potentiates organoid growth in Matrigel and collagen hydrogels. *Nat. Biotechnol.*, 1–11 (2025).
177. M. Sen, T. A. Springer, Leukocyte integrin α L β 2 headpiece structures: The α I domain, the pocket for the internal ligand, and concerted movements of its loops. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 2940–2945 (2016).
178. B. LaFoya, J. A. Munroe, A. Miyamoto, M. A. Detweiler, J. J. Crow, T. Gazdik, A. R. Albig, Beyond the Matrix: The Many Non-ECM Ligands for Integrins. *Int. J. Mol. Sci.* **19** (2018).
179. G. Ostermann, K. S. C. Weber, A. Zerneck, A. Schröder, C. Weber, JAM-1 is a ligand of the beta(2) integrin LFA-1 involved in transendothelial migration of leukocytes. *Nat. Immunol.* **3**, 151–158 (2002).
180. S. D. Marlin, T. A. Springer, Purified intercellular adhesion molecule-1 (ICAM-1) is a ligand for lymphocyte function-associated antigen 1 (LFA-1). *Cell* **51**, 813–819 (1987).
181. M. A. Arnaout, Biology and structure of leukocyte β 2 integrins and their role in inflammation. *F1000Res.* **5**, F1000 Faculty Rev-2433 (2016).

182. C. J. Sommer-Plüss, C. Leiggenger, E. Nikci, R. V. Mancuso, S. Rabbani, C. Lamers, D. Ricklin, Determining ligand binding and specificity within the $\beta 2$ -integrin family with a novel assay platform. *Biomolecules* **15**, 238 (2025).
183. S. L. Park, T. Gebhardt, L. K. Mackay, Tissue-resident memory T cells in cancer immunosurveillance. *Trends Immunol.* **40**, 735–747 (2019).
184. L. K. Mackay, A. Rahimpour, J. Z. Ma, N. Collins, A. T. Stock, M.-L. Hafon, J. Vega-Ramos, P. Lauzurica, S. N. Mueller, T. Stefanovic, D. C. Tschärke, W. R. Heath, M. Inouye, F. R. Carbone, T. Gebhardt, The developmental pathway for CD103(+)CD8+ tissue-resident memory T cells of skin. *Nat. Immunol.* **14**, 1294–1301 (2013).
185. H. Y. Fung, M. Teryek, A. D. Lemenze, T. Bergsbaken, CD103 fate mapping reveals that intestinal CD103- tissue-resident memory T cells are the primary responders to secondary infection. *Sci. Immunol.* **7**, eabl9925 (2022).
186. S. Corgnac, M. Boutet, M. Kfoury, C. Naltet, F. Mami-Chouaib, The emerging role of CD8+ tissue resident memory T (TRM) cells in antitumor immunity: A unique functional contribution of the CD103 integrin. *Front. Immunol.* **9**, 1904 (2018).
187. W. Xu, T. Bergsbaken, K. L. Edelblum, The multifunctional nature of CD103 ($\alpha E\beta 7$ integrin) signaling in tissue-resident lymphocytes. *Am. J. Physiol. Cell Physiol.* **323**, C1161–C1167 (2022).
188. H. Akasaka, D. Sato, W. Shihoya, O. Nureki, Y. Kise, Cryo-EM structure of I domain-containing integrin $\alpha E\beta 7$. *Biochem. Biophys. Res. Commun.* **721**, 150121 (2024).
189. M. Jin, G. Song, C. V. Carman, Y.-S. Kim, N. S. Astrof, M. Shimaoka, D. K. Wittrup, T. A. Springer, Directed evolution to probe protein allostery and integrin I domains of 200,000-fold higher affinity. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 5758–5763 (2006).
190. B. L. Walling, M. Kim, LFA-1 in T cell migration and differentiation. *Front. Immunol.* **9**, 952 (2018).
191. G. Bajic, L. Yatime, R. B. Sim, T. Vorup-Jensen, G. R. Andersen, Structural insight on the recognition of surface-bound opsonins by the integrin I domain of complement receptor 3. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 16426–16431 (2013).
192. H. Nguyen, N. P. Podolnikova, T. P. Ugarova, X. Wang, αMI -domain of integrin Mac-1 binds the cytokine pleiotrophin using multiple mechanisms. *Structure* **32**, 1184–1196.e4 (2024).
193. M. R. Jensen, G. Bajic, X. Zhang, A. K. Laustsen, H. Koldsø, K. K. Skeby, B. Schiøtt, G. R. Andersen, T. Vorup-Jensen, Structural basis for simvastatin competitive antagonism of complement receptor 3. *J. Biol. Chem.* **291**, 16963–16976 (2016).
194. J. Morgan, M. Saleem, R. Ng, C. Armstrong, S. S. Wong, S. G. Caulton, A. Fickling, H. E. L. Williams, A. D. Munday, J. A. López, M. S. Searle, J. Emsley, Structural basis of the

- leukocyte integrin Mac-1 I-domain interactions with the platelet glycoprotein Ib. *Blood Adv.* **3**, 1450–1459 (2019).
195. Y. Wang, H. Gao, C. Shi, P. W. Erhardt, A. Pavlovsky, D. A. Soloviev, K. Bledzka, V. Ustinov, L. Zhu, J. Qin, A. D. Munday, J. Lopez, E. Plow, D. I. Simon, Leukocyte integrin Mac-1 regulates thrombosis via interaction with platelet GPIIb α . *Nat. Commun.* **8**, 15559 (2017).
 196. M. S. Diamond, D. E. Staunton, S. D. Marlin, T. A. Springer, Binding of the integrin Mac-1 (CD11b/CD18) to the third immunoglobulin-like domain of ICAM-1 (CD54) and its regulation by glycosylation. *Cell* **65**, 961–971 (1991).
 197. M. S. Diamond, D. E. Staunton, A. R. de Fougères, S. A. Stacker, J. Garcia-Aguilar, M. L. Hibbs, T. A. Springer, ICAM-1 (CD54): a counter-receptor for Mac-1 (CD11b/CD18). *J. Cell Biol.* **111**, 3129–3139 (1990).
 198. S. Xu, J. Wang, J.-H. Wang, T. A. Springer, Distinct recognition of complement iC3b by integrins α X β 2 and α M β 2. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 3403–3408 (2017).
 199. V. P. Yakubenko, S. P. Yadav, T. P. Ugarova, Integrin α D β 2, an adhesion receptor up-regulated on macrophage foam cells, exhibits multiligand-binding properties. *Blood* **107**, 1643–1650 (2006).
 200. P. L. Stewart, G. R. Nemerow, Cell integrins: commonly used receptors for diverse viral pathogens. *Trends Microbiol.* **15**, 500–507 (2007).
 201. A. Scibelli, S. Roperto, L. Manna, L. M. Pavone, S. Tafuri, R. Della Morte, N. Staiano, Engagement of integrins as a cellular route of invasion by bacterial pathogens. *Vet. J.* **173**, 482–491 (2007).
 202. D. Yu, Z. Lu, Y. Chong, Integrins as a bridge between bacteria and cells: key targets for therapeutic wound healing. *Burns Trauma* **12**, tkae022 (2024).
 203. H. A. M. Hussein, L. R. Walker, U. M. Abdel-Raouf, S. A. Desouky, A. K. M. Montasser, S. M. Akula, Beyond RGD: virus interactions with integrins. *Arch. Virol.* **160**, 2669–2681 (2015).
 204. N. Trstenjak, D. Milić, M. A. Graewert, H. Rouha, D. Svergun, K. Djinović-Carugo, E. Nagy, A. Badarau, Molecular mechanism of leukocidin GH-integrin CD11b/CD18 recognition and species specificity. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 317–327 (2020).
 205. L. K. Shewell, C. J. Day, F. E.-C. Jen, T. Haselhorst, J. M. Atack, J. F. Reijneveld, A. Everest-Dass, D. B. A. James, K. M. Boguslawski, S. Brouwer, C. M. Gillen, Z. Luo, B. Kobe, V. Nizet, M. von Itzstein, M. J. Walker, A. W. Paton, J. C. Paton, V. J. Torres, M. P. Jennings, All major cholesterol-dependent cytolysins use glycans as cellular receptors. *Sci. Adv.* **6**, eaaz4926 (2020).

206. P. J. Muchowski, L. Zhang, E. R. Chang, H. R. Soule, E. F. Plow, M. Moyle, Functional interaction between the integrin antagonist neutrophil inhibitory factor and the I domain of CD11b/CD18. *J. Biol. Chem.* **269**, 26419–26423 (1994).
207. M. Moyle, D. L. Foster, D. E. McGrath, S. M. Brown, Y. Laroche, J. De Meutter, P. Stanssens, C. A. Bogowitz, V. A. Fried, J. A. Ely, A hookworm glycoprotein that inhibits neutrophil function is a ligand of the integrin CD11b/CD18. *J. Biol. Chem.* **269**, 10008–10015 (1994).
208. S. D. Wright, S. M. Levin, M. T. Jong, Z. Chad, L. G. Kabbash, CR3 (CD11b/CD18) expresses one binding site for Arg-Gly-Asp-containing peptides and a second site for bacterial lipopolysaccharide. *J. Exp. Med.* **169**, 175–183 (1989).
209. G. D. Ross, J. A. Cain, P. J. Lachmann, Membrane complement receptor type three (CR3) has lectin-like properties analogous to bovine conglutinin as functions as a receptor for zymosan and rabbit erythrocytes as well as a receptor for iC3b. *J. Immunol.* **134**, 3307–3315 (1985).
210. B. P. Thornton, V. Větvicka, M. Pitman, R. C. Goldman, G. D. Ross, Analysis of the sugar specificity and molecular location of the beta-glucan-binding lectin site of complement receptor type 3 (CD11b/CD18). *J. Immunol.* **156**, 1235–1246 (1996).
211. K. M. Boguslawski, A. N. McKeown, C. J. Day, K. A. Lacey, K. Tam, N. Vozhilla, S. Y. Kim, M. P. Jennings, S. B. Koralov, N. C. Elde, V. J. Torres, Exploiting species specificity to understand the tropism of a human-specific toxin. *Sci. Adv.* **6**, eaax7515 (2020).
212. O. Mukhtar, R. Dutt, A. Panda, P. Kumari, S. S. Singh, G. Paul, N. Prakash, M. Abbas, M. M. Islam, P. Arora, A. Tammour, A. Mohammed, D. Kumar, P. Malhotra, LFA-1 interaction with GBP-130 on Plasmodium falciparum-infected red blood cells mediates NK cell activation and parasite control. *Elife* **15**, RP110942 (2026).
213. N.-S. Seo, C. Q.-Y. Zeng, J. M. Hyser, B. Utama, S. E. Crawford, K. J. Kim, M. Höök, M. K. Estes, Integrins alpha1beta1 and alpha2beta1 are receptors for the rotavirus enterotoxin. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 8811–8818 (2008).
214. J. M. Bergelson, M. P. Shepley, B. M. Chan, M. E. Hemler, R. W. Finberg, Identification of the integrin VLA-2 as a receptor for echovirus 1. *Science* **255**, 1718–1720 (1992).
215. L. Xing, M. Huhtala, V. Pietiäinen, J. Käpylä, K. Vuorinen, V. Marjomäki, J. Heino, M. S. Johnson, T. Hyypiä, R. H. Cheng, Structural and functional analysis of integrin alpha2I domain interaction with echovirus 1. *J. Biol. Chem.* **279**, 11632–11638 (2004).
216. M. La Linn, J. A. Eble, C. Lübken, R. W. Slade, J. Heino, J. Davies, A. Suhrbier, An arthritogenic alphavirus uses the alpha1beta1 integrin collagen receptor. *Virology* **336**, 229–239 (2005).

217. J. A. Goldsmith, A. M. DiVenere, J. A. Maynard, J. S. McLellan, Structural basis for non-canonical integrin engagement by Bordetella adenylate cyclase toxin. *Cell Rep.* **40**, 111196 (2022).
218. F. van Kuppeveld, C. Owino, G. Dellosa, R. Luteijn, M. Zwaagstra, X. Yu, J. V. Eecke, F. Buitenwerf, W. Meijer, M. Liu, C. de Haan, D. Hurdiss, C. Porta, H. Duyvesteyn, J. Kelly, T. Tuthill, Y. Narimatsu, H. Clausen, P. Shah, E. Fry, K. Reiding, E. de Vries, D. Stuart, Identification and structural characterisation of a picornavirus integrin receptor with an essential role of two distinct glycans for virus infection, *Research Square* (2026). <https://doi.org/10.21203/rs.3.rs-9053284/v1>.
219. I. R. Kieba, K. P. Fong, H.-Y. Tang, K. E. Hoffman, D. W. Speicher, L. B. Klickstein, E. T. Lally, Aggregatibacter actinomycetemcomitans leukotoxin requires beta-sheets 1 and 2 of the human CD11a beta-propeller for cytotoxicity. *Cell. Microbiol.* **9**, 2689–2699 (2007).
220. J. Li, K. D. Clinkenbeard, J. W. Ritchey, Bovine CD18 identified as a species specific receptor for Pasteurella haemolytica leukotoxin. *Vet. Microbiol.* **67**, 91–97 (1999).
221. L. C. Ristow, V. Tran, K. J. Schwartz, L. Pankratz, A. Mehle, J.-D. Sauer, R. A. Welch, The extracellular domain of the $\beta 2$ integrin β subunit (CD18) is sufficient for Escherichia coli hemolysin and Aggregatibacter actinomycetemcomitans leukotoxin cytotoxic activity. *MBio* **10**, e01459-19 (2019).
222. F. J. Fernández, J. Santos-López, R. Martínez-Barricarte, J. Querol-García, H. Martín-Merino, S. Navas-Yuste, M. Savko, W. E. Shepard, S. Rodríguez de Córdoba, M. C. Vega, The crystal structure of iC3b-CR3 αI reveals a modular recognition of the main opsonin iC3b by the CR3 integrin receptor. *Nat. Commun.* **13**, 1–16 (2022).
223. R. K. Jensen, G. Bajic, M. Sen, T. A. Springer, T. Vorup-Jensen, G. R. Andersen, Complement receptor 3 forms a compact high-affinity complex with iC3b. *J. Immunol.* **206**, 3032–3042 (2021).
224. X. M. O'Brien, K. E. Heflin, L. M. Lavigne, K. Yu, M. Kim, A. R. Salomon, J. S. Reichner, Lectin site ligation of CR3 induces conformational changes and signaling. *J. Biol. Chem.* **287**, 3337–3348 (2012).
225. C. Cywes, N. L. Godenir, H. C. Hoppe, R. R. Scholle, L. M. Steyn, R. E. Kirsch, M. R. Ehlers, Nonopsonic binding of Mycobacterium tuberculosis to human complement receptor type 3 expressed in Chinese hamster ovary cells. *Infect. Immun.* **64**, 5373–5383 (1996).
226. K. Zen, Y. Liu, D. Cairo, C. A. Parkos, CD11b/CD18-dependent interactions of neutrophils with intestinal epithelium are mediated by fucosylated proteoglycans. *J. Immunol.* **169**, 5270–5278 (2002).
227. G. Saggu, K. Okubo, Y. Chen, R. Vattepu, N. Tsuboi, F. Rosetti, X. Cullere, N. Washburn, S. Tahir, A. M. Rosado, S. M. Holland, R. M. Anthony, M. Sen, C. Zhu, T. N. Mayadas, Cis interaction between sialylated Fc γ RIIA and the αI -domain of Mac-1 limits antibody-mediated neutrophil recruitment. *Nat. Commun.* **9**, 5058 (2018).

228. N. P. Podolnikova, S. Key, X. Wang, T. P. Ugarova, The CIS association of CD47 with integrin Mac-1 regulates macrophage responses by stabilizing the extended integrin conformation. *J. Biol. Chem.* **299**, 103024 (2023).
229. Z. Fan, S. McArdle, A. Marki, Z. Mikulski, E. Gutierrez, B. Engelhardt, U. Deutsch, M. Ginsberg, A. Groisman, K. Ley, Neutrophil recruitment limited by high-affinity bent $\beta 2$ integrin binding ligand in cis. *Nat. Commun.* **7**, 12658 (2016).
230. J. Chen, M.-C. Zhong, H. Guo, D. Davidson, S. Mishel, Y. Lu, I. Rhee, L.-A. Pérez-Quintero, S. Zhang, M.-E. Cruz-Munoz, N. Wu, D. C. Vinh, M. Sinha, V. Calderon, C. A. Lowell, J. S. Danska, A. Veillette, SLAMF7 is critical for phagocytosis of haematopoietic tumour cells via Mac-1 integrin. *Nature* **544**, 493–497 (2017).
231. C. A. Orengo, D. T. Jones, J. M. Thornton, Protein superfamilies and domain superfolds. *Nature* **372**, 631–634 (1994).
232. M. Levitt, Nature of the protein universe. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 11079–11084 (2009).
233. J. Skolnick, A. K. Arakaki, S. Y. Lee, M. Brylinski, The continuity of protein structure space is an intrinsic property of proteins. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 15690–15695 (2009).
234. I. Barrio-Hernandez, J. Yeo, J. Jänes, M. Mirdita, C. L. M. Gilchrist, T. Wein, M. Varadi, S. Velankar, P. Beltrao, M. Steinegger, Clustering predicted structures at the scale of the known protein universe. *Nature* **622**, 637–645 (2023).
235. A. M. Lau, N. Bordin, S. M. Kandathil, I. Sillitoe, V. P. Waman, J. Wells, C. A. Orengo, D. T. Jones, Exploring structural diversity across the protein universe with The Encyclopedia of Domains. *Science* **386**, eadq4946 (2024).
236. J. Durairaj, A. M. Waterhouse, T. Mets, T. Brodiazhenko, M. Abdullah, G. Studer, G. Tauriello, M. Akdel, A. Andreeva, A. Bateman, T. Tenson, V. Hauryliuk, T. Schwede, J. Pereira, Uncovering new families and folds in the natural protein universe. *Nature* **622**, 646–653 (2023).
237. J. Nomburg, E. E. Doherty, N. Price, D. Bellieny-Rabelo, Y. K. Zhu, J. A. Doudna, Birth of protein folds and functions in the virome. *Nature* **633**, 710–717 (2024).
238. Y. Loewenstein, D. Raimondo, O. C. Redfern, J. Watson, D. Frishman, M. Linial, C. Orengo, J. Thornton, A. Tramontano, Protein function annotation by homology-based inference. *Genome Biol.* **10**, 207 (2009).
239. M. Punta, P. C. Coghill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman, R. D. Finn, The Pfam protein families database. *Nucleic Acids Res.* **40**, D290–301 (2012).

240. M. Blum, A. Andreeva, L. C. Florentino, S. R. Chuguransky, T. Grego, E. Hobbs, B. L. Pinto, A. Orr, T. Paysan-Lafosse, I. Ponamareva, G. A. Salazar, N. Bordin, P. Bork, A. Bridge, L. Colwell, J. Gough, D. H. Haft, I. Letunic, F. Llinares-López, A. Marchler-Bauer, L. Meng-Papaxanthos, H. Mi, D. A. Natale, C. A. Orengo, A. P. Pandurangan, D. Piovesan, C. Rivoire, C. J. A. Sigrist, N. Thanki, F. Thibaud-Nissen, P. D. Thomas, S. C. E. Tosatto, C. H. Wu, A. Bateman, InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Res.* **53**, D444–D456 (2025).
241. A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
242. C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, J. M. Thornton, CATH--a hierarchic classification of protein domain structures. *Structure* **5**, 1093–1108 (1997).
243. M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Židek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, S. Velankar, AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
244. Y. Zhang, J. Skolnick, TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
245. T. Hamamsy, J. T. Morton, R. Blackwell, D. Berenberg, N. Carriero, V. Gligoričević, C. E. M. Strauss, J. K. Leman, K. Cho, R. Bonneau, Protein remote homology detection and structural alignment using deep learning. *Nat. Biotechnol.* **42**, 975–985 (2024).
246. V. Gligoričević, P. D. Renfrew, T. Kosciółek, J. K. Leman, D. Berenberg, T. Vatanen, C. Chandler, B. C. Taylor, I. M. Fisk, H. Vlamakis, R. J. Xavier, R. Knight, K. Cho, R. Bonneau, Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* **12**, 3168 (2021).
247. J. T. J. Lecomte, D. A. Vuletich, A. M. Lesk, Structural divergence and distant relationships in proteins: evolution of the globins. *Curr. Opin. Struct. Biol.* **15**, 290–301 (2005).
248. D. M. Fowler, S. Fields, Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).
249. H. Wei, X. Li, Deep mutational scanning: A versatile tool in systematically mapping genotypes to phenotypes. *Front. Genet.* **14**, 1087267 (2023).
250. M. Bashton, C. Chothia, The generation of new protein functions by the combination of domains. *Structure* **15**, 85–99 (2007).

251. J.-H. Han, S. Batey, A. A. Nickson, S. A. Teichmann, J. Clarke, The folding and evolution of multidomain proteins. *Nat. Rev. Mol. Cell Biol.* **8**, 319–330 (2007).
252. M. Wang, G. Caetano-Anollés, The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world. *Structure* **17**, 66–78 (2009).
253. G. Apic, J. Gough, S. A. Teichmann, Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.* **310**, 311–325 (2001).
254. C. Vogel, M. Bashton, N. D. Kerrison, C. Chothia, S. A. Teichmann, Structure, function and evolution of multidomain proteins. *Curr. Opin. Struct. Biol.* **14**, 208–216 (2004).
255. M. Bashton, C. Chothia, The geometry of domain combination in proteins. *J. Mol. Biol.* **315**, 927–939 (2002).
256. G. Apic, W. Huber, S. A. Teichmann, Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination. *J. Struct. Funct. Genomics* **4**, 67–78 (2003).
257. J.-Y. Xue, C.-A. Yang, S. Hu, H.-Y. Fan, Y.-M. Zhang, Z.-Q. Shao, L. Li, S. Wang, T. Wei, S. Dong, Y. Liu, Z. Li, Y. Van de Peer, Evolution of protein domains and protein domain combinations provides insights into the origin and diversification of land plants. *Cell Genom.* **6**, 101107 (2026).
258. C.-H. Hsu, A. W. T. Chiang, M.-J. Hwang, B.-Y. Liao, Proteins with highly evolvable domain architectures are nonessential but highly retained. *Mol. Biol. Evol.* **33**, 1219–1230 (2016).
259. I. Budimir, E. Giampieri, E. Saccenti, M. Suarez-Diez, M. Tarozzi, D. Dall’Olio, A. Merlotti, N. Curti, D. Remondini, G. Castellani, C. Sala, Intraspecies characterization of bacteria via evolutionary modeling of protein domains. *Sci. Rep.* **12**, 16595 (2022).
260. C. P. Ponting, R. R. Russell, The natural history of protein domains. *Annu. Rev. Biophys. Biomol. Struct.* **31**, 45–71 (2002).
261. N. Furnham, N. L. Dawson, S. A. Rahman, J. M. Thornton, C. A. Orengo, Large-scale analysis exploring evolution of catalytic machineries and mechanisms in enzyme superfamilies. *J. Mol. Biol.* **428**, 253–267 (2016).
262. T. Zhang, X. Yang, Y. Fu, W. Xue, Y. Zhang, S. Duan, Y. Yin, Y. Guo, C. Gao, Y. Liu, G. Li, C. Xu, H. Lu, Evolutionary dynamics of insect odorant receptors reveal ecological tuning shaping olfactory perception, *eLife* (2026). <https://doi.org/10.7554/elife.110877.1>.
263. A. F. Philip, M. Kumauchi, W. D. Hoff, Robustness and evolvability in the functional anatomy of a PER-ARNT-SIM (PAS) domain. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 17986–17991 (2010).

264. A. S. Pillai, S. A. Chandler, Y. Liu, A. V. Signore, C. R. Cortez-Romero, J. L. P. Benesch, A. Laganowsky, J. F. Storz, G. K. A. Hochberg, J. W. Thornton, Origin of complexity in haemoglobin evolution. *Nature* **581**, 480–485 (2020).
265. D. P. Anderson, D. S. Whitney, V. Hanson-Smith, A. Woznica, W. Campodonico-Burnett, B. F. Volkman, N. King, J. W. Thornton, K. E. Prehoda, Evolution of an ancient protein function involved in organized multicellularity in animals. *Elife* **5**, e10147 (2016).
266. T. N. Starr, L. K. Picton, J. W. Thornton, Alternative evolutionary histories in the sequence space of an ancient protein. *Nature* **549**, 409–413 (2017).
267. A. Papkou, L. Garcia-Pastor, J. A. Escudero, A. Wagner, A rugged yet easily navigable fitness landscape. *Science* **382**, eadh3860 (2023).
268. D. Zhang, L. M. Iyer, A. M. Burroughs, L. Aravind, Resilience of biochemical activity in protein domains in the face of structural divergence. *Curr. Opin. Struct. Biol.* **26**, 92–103 (2014).
269. L. Viljakainen, J. D. Evans, M. Hasselmann, O. Rueppell, S. Tingek, P. Pamilo, Rapid evolution of immune proteins in social insects. *Mol. Biol. Evol.* **26**, 1791–1801 (2009).
270. R. Nielsen, C. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton, M. J. Hubisz, A. Fledel-Alon, D. M. Tanenbaum, D. Civello, T. J. White, J. J. Sninsky, M. D. Adams, M. Cargill, A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3**, e170 (2005).
271. W. J. Swanson, V. D. Vacquier, The rapid evolution of reproductive proteins. *Nat. Rev. Genet.* **3**, 137–144 (2002).
272. R. Camahort, B. Li, L. Florens, S. K. Swanson, M. P. Washburn, J. L. Gerton, Scm3 is essential to recruit the histone h3 variant cse4 to centromeres and to maintain a functional kinetochore. *Mol. Cell* **26**, 853–865 (2007).
273. Z. Zhou, H. Feng, B.-R. Zhou, R. Ghirlando, K. Hu, A. Zwolak, L. M. Miller Jenkins, H. Xiao, N. Tjandra, C. Wu, Y. Bai, Structural basis for recognition of centromere histone variant CenH3 by the chaperone Scm3. *Nature* **472**, 234–237 (2011).
274. S. Stoler, K. Rogers, S. Weitze, L. Morey, M. Fitzgerald-Hayes, R. E. Baker, Scm3, an essential *Saccharomyces cerevisiae* centromere protein required for G2/M progression and Cse4 localization. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 10571–10576 (2007).
275. C.-C. Chen, M. L. Dechassa, E. Bettini, M. B. Ledoux, C. Belisario, P. Heun, K. Luger, B. G. Mellone, CAL1 is the *Drosophila* CENP-A assembly factor. *J. Cell Biol.* **204**, 313–329 (2014).
276. E. M. Dunleavy, D. Roche, H. Tagami, N. Lacoste, D. Ray-Gallet, Y. Nakamura, Y. Daigo, Y. Nakatani, G. Almouzni-Pettinotti, HJURP is a cell-cycle-dependent maintenance and deposition factor of CENP-A at centromeres. *Cell* **137**, 485–497 (2009).

277. D. R. Foltz, L. E. T. Jansen, A. O. Bailey, J. R. Yates 3rd, E. A. Bassett, S. Wood, B. E. Black, D. W. Cleveland, Centromere-specific assembly of CENP-a nucleosomes is mediated by HJURP. *Cell* **137**, 472–484 (2009).
278. L. Sanchez-Pulido, A. L. Pidoux, C. P. Ponting, R. C. Allshire, Common ancestry of the CENP-A chaperones Scm3 and HJURP. *Cell* **137**, 1173–1174 (2009).
279. M. S. Johnson, N. Lu, K. Denessiouk, J. Heino, D. Gullberg, Integrins during evolution: evolutionary trees and model organisms. *Biochim. Biophys. Acta* **1788**, 779–789 (2009).
280. C. Huang, T. A. Springer, Folding of the beta-propeller domain of the integrin alphaL subunit is independent of the I domain and dependent on the beta2 subunit. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 3162–3167 (1997).
281. E. Ruoslahti, RGD and other recognition sequences for integrins. *Annu. Rev. Cell Dev. Biol.* **12**, 697–715 (1996).
282. S. K. Shaw, M. B. Brenner, The $\beta 7$ integrins in mucosal homing and retention. *Semin. Immunol.* (1995).
283. P. J. Rutgeerts, R. N. Fedorak, D. W. Hommes, A. Sturm, D. C. Baumgart, B. Bressler, S. Schreiber, J. C. Mansfield, M. Williams, M. Tang, J. Visich, X. Wei, M. Keir, D. Luca, D. Danilenko, J. Egen, S. O’Byrne, A randomised phase I study of etrolizumab (rhuMAb $\beta 7$) in moderate to severe ulcerative colitis. *Gut* **62**, 1122–1130 (2013).
284. B. Dai, J. A. Hackney, R. Ichikawa, A. Nguyen, J. Elstrott, L. D. Orozco, K.-H. Sun, Z. Modrusan, A. Gogineni, A. Scherl, J. Gubatan, A. Habtezion, M. Deswal, M. Somsouk, W. A. Faubion, A. Chai, Z. Sharafali, A. Hassanali, Y. S. Oh, S. Tole, J. McBride, M. E. Keir, T. Yi, Dual targeting of lymphocyte homing and retention through $\alpha 4\beta 7$ and $\alpha E\beta 7$ inhibition in inflammatory bowel disease. *Cell Rep Med* **2**, 100381 (2021).
285. P. Eder, M. Kłopočka, H. Cichoż-Lach, R. Talar-Wojnarowska, M. Kopertowska-Majchrzak, A. Michalak, R. Filip, K. Waszak, K. Stawczyk-Eder, M. Janiak, K. Skrobot, A. Liebert, H. Zatorski, A. Solarska-Pólchłopek, M. Krogulecki, A. Pękała, E. Poniewierka, I. Smoła, A. Kaczka, K. Wojciechowski, S. Drygała, E. Zagórowicz, Real-world outcomes of 54-week vedolizumab therapy and response durability after treatment discontinuation in ulcerative colitis: results from a multicenter prospective POLONEZ study. *Therap. Adv. Gastroenterol.* **16**, 17562848231151296 (2023).
286. N. Green, J. Rosebrook, N. Cochran, K. Tan, J. H. Wang, T. A. Springer, M. J. Briskin, Mutational analysis of MAdCAM-1/ $\alpha 4\beta 7$ interactions reveals significant binding determinants in both the first and second immunoglobulin domains. *Cell Adhes. Commun.* **7**, 167–181 (1999).
287. J. Takagi, T. A. Springer, Integrin activation and structural rearrangement: Takagi & Springer . Integrin activation and structural rearrangement. *Immunol. Rev.* **186**, 141–163 (2002).

288. Y. Yu, J. Zhu, L.-Z. Mi, T. Walz, H. Sun, J. Chen, T. A. Springer, Structural specializations of $\alpha(4)\beta(7)$, an integrin that mediates rolling adhesion. *J. Cell Biol.* **196**, 131–146 (2012).
289. J. Takagi, K. Strokovich, T. A. Springer, T. Walz, Structure of integrin $\alpha 5\beta 1$ in complex with fibronectin. *EMBO J.* **22**, 4607–4615 (2003).
290. J. Dando, K. W. Wilkinson, S. Ortlepp, D. J. King, R. L. Brady, A reassessment of the MAdCAM-1 structure and its role in integrin recognition. *Acta Crystallogr. D Biol. Crystallogr.* **58**, 233–241 (2002).
291. K. Tan, J. M. Casasnovas, J.-H. Liu, M. J. Briskin, T. A. Springer, J.-H. Wang, The structure of immunoglobulin superfamily domains 1 and 2 of MAdCAM-1 reveals novel features important for integrin recognition. *Structure* **6**, 793–801 (1998).
292. J. L. Viney, S. Jones, H. H. Chiu, B. Lagrimas, M. E. Renz, L. G. Presta, D. Jackson, K. J. Hillan, S. Lew, S. Fong, Mucosal addressin cell adhesion molecule-1: a structural and functional analysis demarcates the integrin binding motif. *J. Immunol.* **157**, 2488–2497 (1996).
293. C. Berlin, R. F. Bargatze, J. J. Campbell, U. H. von Andrian, M. C. Szabo, S. R. Hasslen, R. D. Nelson, E. L. Berg, S. L. Erlandsen, E. C. Butcher, Alpha 4 integrins mediate lymphocyte attachment and rolling under physiologic flow. *Cell* **80**, 413–422 (1995).
294. G. J. Russell, C. M. Parker, K. L. Cepek, D. A. Mandelbrot, A. Sood, E. Mizoguchi, E. C. Ebert, M. B. Brenner, A. K. Bhan, Distinct structural and functional epitopes of the alpha E beta 7 integrin. *Eur. J. Immunol.* **24**, 2832–2841 (1994).
295. P. A. Szabo, M. Miron, D. L. Farber, Location, location, location: Tissue resident memory T cells in mice and humans. *Sci Immunol* **4** (2019).
296. J. M. Schenkel, D. Masopust, Tissue-resident memory T cells. *Immunity* **41**, 886–897 (2014).
297. R. A. Clark, Resident memory T cells in human health and disease. *Sci. Transl. Med.* **7**, 269rv1 (2015).
298. R. K. Jensen, H. Pedersen, J. Lorentzen, N. S. Laursen, T. Vorup-Jensen, G. R. Andersen, Structural insights into the function-modulating effects of nanobody binding to the integrin receptor $\alpha M\beta 2$. *J. Biol. Chem.* **298** (2022).
299. A. Punjani, D. J. Fleet, 3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. *J. Struct. Biol.* **213**, 107702 (2021).
300. P. I. Karecla, S. J. Green, S. J. Bowden, J. Coadwell, P. J. Kilshaw, Identification of a Binding Site for Integrin $\alpha E\beta 7$ in the N-terminal Domain of E-cadherin*. *J. Biol. Chem.* **271**, 30909–30915 (1996).

301. J. M. Higgins, D. A. Mandlebrot, S. K. Shaw, G. J. Russell, E. A. Murphy, Y. T. Chen, W. J. Nelson, C. M. Parker, M. B. Brenner, Direct and regulated interaction of integrin α 5 β 1 with E-cadherin. *J. Cell Biol.* **140**, 197–210 (1998).
302. S. Troyanovsky, Cadherin dimers in cell-cell adhesion. *Eur. J. Cell Biol.* **84**, 225–233 (2005).
303. J. M. Higgins, M. Cernadas, K. Tan, A. Irie, J. Wang, Y. Takada, M. B. Brenner, The role of α and β chains in ligand recognition by β 7 integrins. *J. Biol. Chem.* **275**, 25652–25664 (2000).
304. J. Vendome, S. Posy, X. Jin, F. Bahna, G. Ahlsen, L. Shapiro, B. Honig, Molecular design principles underlying β -strand swapping in the adhesive dimerization of cadherins. *Nat. Struct. Mol. Biol.* **18**, 693–700 (2011).
305. C. P. Chen, S. Posy, A. Ben-Shaul, L. Shapiro, B. H. Honig, Specificity of cell-cell adhesion by classical cadherins: Critical role for low-affinity dimerization through β -strand swapping. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 8531–8536 (2005).
306. O. J. Harrison, F. Bahna, P. S. Katsamba, X. Jin, J. Brasch, J. Vendome, G. Ahlsen, K. J. Carroll, S. R. Price, B. Honig, L. Shapiro, Two-step adhesive binding by classical cadherins. *Nat. Struct. Mol. Biol.* **17**, 348–357 (2010).
307. H. Solomon-Degefa, J. M. Gebauer, C. M. Jeffries, C. D. Freiburg, P. Meckelburg, L. E. Bird, U. Baumann, D. I. Svergun, R. J. Owens, J. M. Werner, E. Behrmann, M. Paulsson, R. Wagener, Structure of a collagen VI α 3 chain VWA domain array: adaptability and functional implications of myopathy causing mutations. *J. Biol. Chem.* **295**, 12755–12771 (2020).
308. Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
309. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Židek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
310. M. Tulla, O. T. Pentikäinen, T. Viitasalo, J. Kämpylä, U. Impola, P. Nykvist, L. Nissinen, M. S. Johnson, J. Heino, Selective binding of collagen subtypes by integrin α 11, α 21, and α 101 domains. *J. Biol. Chem.* **276**, 48206–48212 (2001).
311. Y. Zou, D. Zwolanek, Y. Izu, S. Gandhi, G. Schreiber, K. Brockmann, M. Devoto, Z. Tian, Y. Hu, G. Veit, M. Meier, J. Stetefeld, D. Hicks, V. Straub, N. C. Voermans, D. E. Birk, E. R. Barton, M. Koch, C. G. Bönnemann, Recessive and dominant mutations in

- COL12A1 cause a novel EDS/myopathy overlap syndrome in humans and mice. *Hum. Mol. Genet.* **23**, 2339–2352 (2014).
312. R.-Z. Zhang, Y. Zou, T.-C. Pan, D. Markova, A. Fertala, Y. Hu, S. Squarzoni, U. C. Reed, S. K. N. Marie, C. G. Bönnemann, M.-L. Chu, Recessive COL6A2 C-globular Missense Mutations in Ullrich Congenital Muscular Dystrophy: ROLE OF THE C2a SPLICE VARIANT*. *J. Biol. Chem.* **285**, 10005–10015 (2010).
 313. A. Irie, T. Kamata, W. Puzon-McLaughlin, Y. Takada, Critical amino acid residues for ligand binding are clustered in a predicted beta-turn of the third N-terminal repeat in the integrin alpha 4 and alpha 5 subunits. *EMBO J.* **14**, 5550–5556 (1995).
 314. B. S. Chouhan, J. Kämpylä, K. Denessiouk, A. Denesyuk, J. Heino, M. S. Johnson, Early chordate origin of the vertebrate integrin α I domains. *PLoS One* **9**, e112064 (2014).
 315. M. Ohi, Y. Li, Y. Cheng, T. Walz, Negative Staining and Image Classification - Powerful Tools in Modern Electron Microscopy. *Biol. Proced. Online* **6**, 23–34 (2004).
 316. C. Suloway, J. Pulokas, D. Fellmann, A. Cheng, F. Guerra, J. Quispe, S. Stagg, C. S. Potter, B. Carragher, Automated molecular microscopy: the new Legimon system. *J. Struct. Biol.* **151**, 41–60 (2005).
 317. A. Punjani, J. L. Rubinstein, D. J. Fleet, M. A. Brubaker, cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
 318. M. J. Humphries, Integrin activation: the link between ligand binding and signal transduction. *Curr. Opin. Cell Biol.* **8**, 632–640 (1996).
 319. I. Dransfield, C. Cabañas, A. Craig, N. Hogg, Divalent cation regulation of the function of the leukocyte integrin LFA-1. *J. Cell Biol.* **116**, 219–226 (1992).
 320. D. N. Mastronarde, Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).
 321. S. Q. Zheng, E. Palovcak, J.-P. Armache, K. A. Verba, Y. Cheng, D. A. Agard, MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
 322. J. Zivanov, T. Nakane, B. O. Forsberg, D. Kimanius, W. J. H. Hagen, E. Lindahl, S. H. W. Scheres, New tools for automated high-resolution cryo-EM structure determination in RELION-3. *Elife* **7**, e42166 (2018).
 323. E. F. Pettersen, T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. I. Croll, J. H. Morris, T. E. Ferrin, UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).
 324. M. A. Cianfrocco, M. Wong-Barnum, C. Youn, R. Wagner, A. Leschziner, “COSMIC2: A Science Gateway for Cryo-Electron Microscopy Structure Determination” in *Proceedings*

of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact (Association for Computing Machinery, New York, NY, USA, 2017) *PEARC17*, pp. 1–5.

325. R. Sanchez-Garcia, J. Gomez-Blanco, A. Cuervo, J. M. Carazo, C. O. S. Sorzano, J. Vargas, DeepEMhancer: a deep learning solution for cryo-EM volume post-processing. *Communications Biology* **4**, 1–8 (2021).
326. Y. Z. Tan, P. R. Baldwin, J. H. Davis, J. R. Williamson, C. S. Potter, B. Carragher, D. Lyumkis, Addressing preferred specimen orientation in single-particle cryo-EM through tilting. *Nat. Methods* **14**, 793–796 (2017).
327. A. Punjani, D. J. Fleet, 3DFlex: determining structure and motion of flexible proteins from cryo-EM. *Nat. Methods* **20**, 860–870 (2023).
328. K. Jamali, L. Käll, R. Zhang, A. Brown, D. Kimanius, S. H. W. Scheres, Automated model building and protein identification in cryo-EM maps. *Nature* **628**, 450–457 (2024).
329. P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L.-W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, P. H. Zwart, PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
330. A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T. A. P. de Beer, C. Rempfer, L. Bordoli, R. Lepore, T. Schwede, SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).
331. V. Nardone, A. P. Lucarelli, A. Dalle Vedove, R. Fanelli, A. Tomassetti, L. Belvisi, M. Civera, E. Parisini, Crystal Structure of Human E-Cadherin-EC1EC2 in Complex with a Peptidomimetic Competitive Inhibitor of Cadherin Homophilic Interaction. *J. Med. Chem.* **59**, 5089–5094 (2016).
332. R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstern, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, D. Hassabis, Protein complex prediction with AlphaFold-Multimer, *bioRxiv* (2022)p. 2021.10.04.463034.
333. P. Emsley, K. Cowtan, Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).
334. T. I. Croll, ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr D Struct Biol* **74**, 519–530 (2018).
335. A. Leaver-Fay, M. Tyka, S. M. Lewis, O. F. Lange, J. Thompson, R. Jacak, K. Kaufman, P. D. Renfrew, C. A. Smith, W. Sheffler, I. W. Davis, S. Cooper, A. Treuille, D. J. Mandell,

- F. Richter, Y.-E. A. Ban, S. J. Fleishman, J. E. Corn, D. E. Kim, S. Lyskov, M. Berrondo, S. Mentzer, Z. Popović, J. J. Havranek, J. Karanicolas, R. Das, J. Meiler, T. Kortemme, J. J. Gray, B. Kuhlman, D. Baker, P. Bradley, ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487**, 545–574 (2011).
336. E. Krissinel, K. Henrick, Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).
337. H. Zhang, N. S. Astrof, J.-H. Liu, J.-H. Wang, M. Shimaoka, Crystal structure of isoflurane bound to integrin LFA-1 supports a unified mechanism of volatile anesthetic action in the immune and central nervous systems. *FASEB J.* **23**, 2735–2740 (2009).
338. K. L. Brown, S. Banerjee, A. Feigley, H. Abe, T. S. Blackwell, A. Pozzi, B. G. Hudson, R. Zent, Salt-bridge modulates differential calcium-mediated ligand binding to integrin $\alpha 1$ - and $\alpha 2$ -I domains. *Sci. Rep.* **8**, 1–14 (2018).
339. T. Keren-Kaplan, L. Zeev Peters, O. Levin-Kravets, I. Attali, O. Kleifeld, N. Shohat, S. Artzi, O. Zucker, I. Pilzer, N. Reis, M. H. Glickman, S. Ben-Aroya, G. Prag, Structure of ubiquitylated-Rpn10 provides insight into its autoregulation mechanism. *Nat. Commun.* **7**, 1–12 (2016).
340. J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Židek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, J. M. Jumper, Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
341. M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, M. Steinegger, ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
342. P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, V. S. Pande, OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **13**, e1005659 (2017).
343. P. Eastman, R. Galvelis, R. P. Peláez, C. R. A. Abreu, S. E. Farr, E. Gallicchio, A. Gorenko, M. M. Henry, F. Hu, J. Huang, A. Krämer, J. Michel, J. A. Mitchell, V. S. Pande, J. P. Rodrigues, J. Rodriguez-Guerra, A. C. Simmonett, S. Singh, J. Swails, P. Turner, Y. Wang, I. Zhang, J. D. Chodera, G. De Fabritiis, T. E. Markland, OpenMM 8: Molecular dynamics simulation with machine learning potentials. *J. Phys. Chem. B* **128**, 109–116 (2024).
344. C. Tian, K. Kasavajhala, K. A. A. Belfon, L. Raguet, H. Huang, A. N. Migués, J. Bickel, Y. Wang, J. Pincay, Q. Wu, C. Simmerling, ff19SB: Amino-Acid-Specific Protein

- Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput.* **16**, 528–552 (2020).
345. A. Sengupta, Z. Li, L. F. Song, P. Li, K. M. Merz Jr, Parameterization of Monovalent Ions for the OPC3, OPC, TIP3P-FB, and TIP4P-FB Water Models. *J. Chem. Inf. Model.* **61**, 869–880 (2021).
346. C. W. Hopkins, S. Le Grand, R. C. Walker, A. E. Roitberg, Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *J. Chem. Theory Comput.* **11**, 1864–1874 (2015).
347. D. R. Roe, T. E. Cheatham 3rd, PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **9**, 3084–3095 (2013).
348. R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, V. S. Pande, MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **109**, 1528–1532 (2015).
349. L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, *arXiv [stat.ML]* (2018).
<http://arxiv.org/abs/1802.03426>.
350. R. D. Finn, J. Clements, S. R. Eddy, HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–37 (2011).
351. R. C. Edgar, MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
352. B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, R. Lanfear, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
353. A. Rambaut, *Figtree: Automatically Exported from Code.google.com/p/figtree* (Github; <https://github.com/rambaut/figtree>).
354. E. T. Boder, K. D. Wittrup, Yeast surface display for screening combinatorial polypeptide libraries. *Nat. Biotechnol.* **15**, 553–557 (1997).
355. R. M. Adams, T. Mora, A. M. Walczak, J. B. Kinney, Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. *Elife* **5**, e23156 (2016).
356. T. N. Starr, A. J. Greaney, S. K. Hilton, D. Ellis, K. H. D. Crawford, A. S. Dingens, M. J. Navarro, J. E. Bowen, M. A. Tortorici, A. C. Walls, N. P. King, D. Veasler, J. D. Bloom, Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* **182**, 1295–1310.e20 (2020).

357. J. Li, J. Yan, T. A. Springer, Low-affinity integrin states have faster ligand-binding kinetics than the high-affinity state. *Elife* **10**, e73359 (2021).
358. H. Nguyen, T. Jing, X. Wang, The Q163C/Q309C mutant of α MI-domain is an active variant suitable for NMR characterization. *PLoS One* **18**, e0280778 (2023).
359. M. Shimaoka, C. Lu, R. T. Palframan, U. H. von Andrian, A. McCormack, J. Takagi, T. A. Springer, Reversibly locking a protein fold in an active conformation with a disulfide bond: integrin α L I domains with high affinity and antagonist activity in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 6009–6014 (2001).
360. J. Cheng, G. Novati, J. Pan, C. Bycroft, A. Žemgulytė, T. Applebaum, A. Pritzel, L. H. Wong, M. Zielinski, T. Sargeant, R. G. Schneider, A. W. Senior, J. Jumper, D. Hassabis, P. Kohli, Ž. Avsec, Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
361. D. I. Beller, T. A. Springer, R. D. Schreiber, Anti-Mac-1 selectively inhibits the mouse and human type three complement receptor. *J. Exp. Med.* **156**, 1000–1009 (1982).
362. B. E. Eaton, L. Gold, D. A. Zichi, Let's get specific: the relationship between specificity and affinity. *Chem. Biol.* **2**, 633–638 (1995).
363. A. L. DuMont, P. Yoong, C. J. Day, F. Alonzo 3rd, W. H. McDonald, M. P. Jennings, V. J. Torres, Staphylococcus aureus LukAB cytotoxin kills human neutrophils by targeting the CD11b subunit of the integrin Mac-1. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 10794–10799 (2013).
364. D. Klug, S. Goellner, J. Kehrer, J. Sattler, L. Strauss, M. Singer, C. Lu, T. A. Springer, F. Frischknecht, Evolutionarily distant I domains can functionally replace the essential ligand-binding domain of Plasmodium TRAP. *Elife* **9**, e57572 (2020).
365. P. Rieu, T. Ueda, I. Haruta, C. P. Sharma, M. A. Arnaout, The A-domain of beta 2 integrin CR3 (CD11b/CD18) is a receptor for the hookworm-derived neutrophil adhesion inhibitor NIF. *J. Cell Biol.* **127**, 2081–2091 (1994).
366. S. K. Dickeson, S. A. Santoro, Ligand recognition by the I domain-containing integrins. *Cell. Mol. Life Sci.* **54**, 556–566 (1998).
367. Z. Wang, A. M. M. Thinn, J. Zhu, A pivotal role for a conserved bulky residue at the α 1-helix of the α I integrin domain in ligand binding. *J. Biol. Chem.* **292**, 20756–20768 (2017).
368. G. D. Ross, Regulation of the adhesion versus cytotoxic functions of the mac-1/CR3/ α M β 2 - Integrin glycoprotein. *Crit. Rev. Immunol.* **20**, 30 (2000).
369. J. R. Dunkelberger, W.-C. Song, Complement and its role in innate and adaptive immune responses. *Cell Res.* **20**, 34–50 (2010).

370. Y. Li, L. Zhang, The fourth blade within the beta-propeller is involved specifically in C3bi recognition by integrin alpha M beta 2. *J. Biol. Chem.* **278**, 34395–34402 (2003).
371. A. Taniguchi-Sidle, D. E. Isenman, Interactions of human complement component C3 with factor B and with complement receptors type 1 (CR1, CD35) and type 3 (CR3, CD11b/CD18) involve an acidic sequence at the N-terminus of C3 alpha'-chain. *J. Immunol.* **153**, 5285–5302 (1994).
372. K. Ley, J. Rivera-Nieves, W. J. Sandborn, S. Shattil, Integrin-based therapeutics: biological basis, clinical use and new drugs. *Nat. Rev. Drug Discov.* **15**, 173–183 (2016).
373. R. J. Slack, S. J. F. Macdonald, J. A. Roper, R. G. Jenkins, R. J. D. Hatley, Emerging therapeutic opportunities for integrin inhibitors. *Nat. Rev. Drug Discov.* **21**, 60–78 (2022).
374. M. A. Siddiq, G. K. Hochberg, J. W. Thornton, Evolution of protein specificity: insights from ancestral protein reconstruction. *Curr. Opin. Struct. Biol.* **47**, 113–122 (2017).
375. L. Martinez, X. Li, G. Ramos-Echazabal, H. Faridi, Z. M. Zigmond, N. Santos Falcon, D. R. Hernandez, S. A. Shehadeh, O. C. Velazquez, V. Gupta, R. I. Vazquez-Padron, A genetic model of constitutively active integrin CD11b/CD18. *J. Immunol.* **205**, 2545–2553 (2020).
376. L. Wen, A. Marki, Z. Wang, M. Orecchioni, J. Makings, M. Billitti, E. Wang, S. S. A. Suthahar, K. Kim, W. B. Kiosses, Z. Mikulski, K. Ley, A humanized $\beta 2$ integrin knockin mouse reveals localized intra- and extravascular neutrophil integrin activation in vivo. *Cell Rep.* **39**, 110876 (2022).
377. J. L. Steenwyk, J. T. Loucks, T. J. Buida, ClipKIT in the browser: fast online trimming of multiple sequence alignments for phylogenetics. *Nucleic Acids Res.* **53**, W169–W171 (2025).
378. J. Trifinopoulos, L.-T. Nguyen, A. von Haeseler, B. Q. Minh, W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* **44**, W232–5 (2016).
379. P. Raman, H. Khan, J. M. Young, T. Tsukiyama, H. S. Malik, Dynamic evolution of EZHIP, an inhibitor of the Polycomb Repressive Complex 2 in mammals, *bioRxiv*org (2025)p. 2025.12.12.693809.
380. C. Lu, M. Shimaoka, M. Ferzly, C. Oxvig, J. Takagi, T. A. Springer, An isolated, surface-expressed I domain of the integrin alphaLbeta2 is sufficient for strong adhesive function when locked in the open conformation with a disulfide bond. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 2387–2392 (2001).
381. A. E. Wentz, E. V. Shusta, A novel high-throughput screen reveals yeast genes that increase secretion of heterologous proteins. *Appl. Environ. Microbiol.* **73**, 1189–1198 (2007).
382. R. D. Gietz, R. H. Schiestl, High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2**, 31–34 (2007).

383. K. H. D. Crawford, J. D. Bloom, alignparse: A Python package for parsing complex features from high-throughput long-read sequencing. *J. Open Source Softw.* **4**, 1915 (2019).
384. J. A. Hollis, J. A. Stonick, I. Topalidou, J. M. Young, C. B. Moens, N. J. Lehrbach, M. G. Campbell, H. S. Malik, Remote homology and functional genetics unmask deeply preserved Scm3/HJURP orthologs in metazoans. *Sci. Adv.* **12**, eaeg5356 (2026).
385. K. L. McKinley, I. M. Cheeseman, The molecular basis for centromere identity and function. *Nat. Rev. Mol. Cell Biol.* **17**, 16–29 (2016).
386. L. Li, Q. Yuan, Y.-M. Chu, H.-Y. Jiang, J.-H. Zhao, Q. Su, D.-Q. Huo, X.-F. Zhang, Advances in holliday junction recognition protein (HJURP): Structure, molecular functions, and roles in cancer. *Front. Cell Dev. Biol.* **11**, 1106638 (2023).
387. G. Mizuguchi, H. Xiao, J. Wisniewski, M. M. Smith, C. Wu, Nonhistone Scm3 and histones CenH3-H4 assemble the core of centromere-specific nucleosomes. *Cell* **129**, 1153–1164 (2007).
388. M. C. Barnhart, P. H. J. L. Kuich, M. E. Stellfox, J. A. Ward, E. A. Bassett, B. E. Black, D. R. Foltz, HJURP is a CENP-A chromatin assembly factor sufficient to form a functional de novo kinetochore. *J. Cell Biol.* **194**, 229–243 (2011).
389. U.-S. Cho, S. C. Harrison, Recognition of the centromere-specific histone Cse4 by the chaperone Scm3. *Proceedings of the National Academy of Sciences* **108**, 9367–9371 (2011).
390. H. Hu, Y. Liu, M. Wang, J. Fang, H. Huang, N. Yang, Y. Li, J. Wang, X. Yao, Y. Shi, G. Li, R.-M. Xu, Structure of a CENP-A-histone H4 heterodimer in complex with chaperone HJURP. *Genes Dev.* **25**, 901–906 (2011).
391. M. Shuaib, K. Ouarrhni, S. Dimitrov, A. Hamiche, HJURP binds CENP-A via a highly conserved N-terminal domain and mediates its deposition at centromeres. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 1349–1354 (2010).
392. A. R. Popchock, S. Hedouin, Y. Mao, C. L. Asbury, A. B. Stergachis, S. Biggins, Stable centromere association of the yeast histone variant Cse4 requires its essential N-terminal domain. *EMBO J.* **44**, 1488–1511 (2025).
393. P. Agarwal, A. Alekar, S. Mallick, J. Shah, M. Basrai, S. K. Ghosh, Scm3 interacts with the N-terminal tail of Cse4 to regulate kinetochore assembly in budding yeast, *bioRxiv* (2025). <https://doi.org/10.1101/2025.10.20.683541>.
394. S. Le Goff, B. N. Keçeli, H. Jeřábková, S. Heckmann, T. Rutten, S. Cotterell, V. Schubert, E. Roitinger, K. Mechtler, F. C. H. Franklin, C. Tatout, A. Houben, D. Geelen, A. V. Probst, I. Lermontova, The H3 histone chaperone NASPSIM3 escorts CenH3 in Arabidopsis. *Plant J.* **101**, 71–86 (2020).

395. H. Takeuchi, S. Nagahara, T. Higashiyama, F. Berger, The chaperone NASP contributes to de Novo deposition of the centromeric histone variant CENH3 in Arabidopsis early embryogenesis. *Plant Cell Physiol.* **65**, 1135–1148 (2024).
396. B. E. Black, D. R. Foltz, S. Chakravarthy, K. Luger, V. L. Woods Jr, D. W. Cleveland, Structural determinants for generating centromeric chromatin. *Nature* **430**, 578–582 (2004).
397. B. E. Black, L. E. T. Jansen, P. S. Maddox, D. R. Foltz, A. B. Desai, J. V. Shah, D. W. Cleveland, Centromere identity maintained by nucleosomes assembled with histone H3 containing the CENP-A targeting domain. *Mol. Cell* **25**, 309–322 (2007).
398. K. L. McKinley, I. M. Cheeseman, Polo-like kinase 1 licenses CENP-A deposition at centromeres. *Cell* **158**, 397–411 (2014).
399. Y. Fujita, T. Hayashi, T. Kiyomitsu, Y. Toyoda, A. Kokubu, C. Obuse, M. Yanagida, Priming of centromere for CENP-A recruitment by human hMis18alpha, hMis18beta, and M18BP1. *Dev. Cell* **12**, 17–30 (2007).
400. D. Pan, K. Walstein, A. Take, D. Bier, N. Kaiser, A. Musacchio, Mechanism of centromere recruitment of the CENP-A chaperone HJURP and its implications for centromere licensing. *Nat. Commun.* **10**, 1–18 (2019).
401. T. Hori, Y. Mahana, M. Ariyoshi, T. Fukagawa, Dual pathways via CENP-C and Mis18C recruit HJURP for CENP-A deposition into vertebrate centromeres. *EMBO J.*, doi: 10.1038/s44318-025-00674-z (2026).
402. H. Tachiwana, S. Müller, J. Blümer, K. Klare, A. Musacchio, G. Almouzni, HJURP involvement in de novo CenH3(CENP-A) and CENP-C recruitment. *Cell Rep.* **11**, 22–32 (2015).
403. C. de Groot, J. Houston, B. Davis, A. Gerson-Gurwitz, J. Monen, P. Lara-Gonzalez, K. Oegema, A. K. Shiau, A. Desai, The N-terminal tail of *C. elegans* CENP-A interacts with KNL-2 and is essential for centromeric chromatin assembly. *MBoC* **32**, 1193–1201 (2021).
404. P. S. Maddox, F. Hyndman, J. Monen, K. Oegema, A. Desai, Functional genomics identifies a Myb domain-containing protein family required for assembly of CENP-A chromatin. *J. Cell Biol.* **176**, 757–763 (2007).
405. R. F. Prosée, J. M. Wenda, I. Özdemir, C. Gabus, K. Delaney, F. Schwager, M. Gotta, F. A. Steiner, Transgenerational inheritance of centromere identity requires the CENP-A N-terminal tail in the *C. elegans* maternal germ line. *PLoS Biol.* **19**, e3000968 (2021).
406. R. S. Keegan, E. M. Dunleavy, Distinct pools of CENP-A and CENP-C support unique phases of centromere assembly in spermatogenesis, *bioRxiv* (2026)p. 2026.01.07.698108.
407. L. F. Rosin, B. G. Mellone, Centromeres Drive a Hard Bargain. *Trends Genet.* **33**, 101–117 (2017).

408. H. S. Malik, S. Henikoff, Phylogenomics of the nucleosome. *Nat. Struct. Biol.* **10**, 882–891 (2003).
409. R. Phansalkar, P. Lapierre, B. G. Mellone, Evolutionary insights into the role of the essential centromere protein CAL1 in *Drosophila*. *Chromosome Res.* **20**, 493–504 (2012).
410. S. Erhardt, B. G. Mellone, C. M. Betts, W. Zhang, G. H. Karpen, A. F. Straight, Genome-wide analysis reveals a cell cycle-dependent mechanism controlling centromere propagation. *J. Cell Biol.* **183**, 805–818 (2008).
411. B. G. Mellone, K. J. Grive, V. Shteyn, S. R. Bowers, I. Oderberg, G. H. Karpen, Assembly of *Drosophila* centromeric chromatin proteins during mitosis. *PLoS Genet.* **7**, e1002068 (2011).
412. L. E. Kursel, F. C. Welsh, H. S. Malik, Ancient coretenation of paralogs of Cid centromeric histones and Cal1 chaperones in mosquito species. *Mol. Biol. Evol.* **37**, 1949–1963 (2020).
413. B. Medina-Pritchard, V. Lazou, J. Zou, O. Byron, M. A. Abad, J. Rappsilber, P. Heun, A. A. Jeyaprakash, Structural basis for centromere maintenance by *Drosophila* CENP-A chaperone CAL1. *EMBO J.* **39**, e103234 (2020).
414. Z. Lin, K. W. Y. Yuen, RbAp46/48LIN-53 and HAT-1 are required for initial CENP-AHCP-3 deposition and de novo holocentromere formation on artificial chromosomes in *Caenorhabditis elegans* embryos. *Nucleic Acids Res.* **49**, 9154–9173 (2021).
415. J. Chen, M. Guo, X. Wang, B. Liu, A comprehensive review and comparison of different computational methods for protein remote homology detection. *Brief. Bioinform.* **19**, 231–244 (2018).
416. P. Radivojac, Advancing remote homology detection: A step toward understanding and accurately predicting protein function. *Cell Syst.* **13**, 435–437 (2022).
417. S. Hotaling, J. L. Kelley, P. B. Frandsen, Toward a genome sequence for every animal: Where are we now? *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2109019118 (2021).
418. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
419. C. T. Amemiya, J. Alföldi, A. P. Lee, S. Fan, H. Philippe, I. MacCallum, I. Braasch, T. Manousaki, I. Schneider, N. Rohner, C. Organ, D. Chalopin, J. J. Smith, M. Robinson, R. A. Dorrington, M. Gerdol, B. Aken, M. A. Biscotti, M. Barucca, D. Baurain, A. M. Berlin, G. L. Blatch, F. Buonocore, T. Burmester, M. S. Campbell, A. Canapa, J. P. Cannon, A. Christoffels, G. De Moro, A. L. Ekins, L. Fan, A. M. Fausto, N. Feiner, M. Forconi, J. Gamielien, S. Gnerre, A. Gnirke, J. V. Goldstone, W. Haerty, M. E. Hahn, U. Hesse, S. Hoffmann, J. Johnson, S. I. Karchner, S. Kuraku, M. Lara, J. Z. Levin, G. W. Litman, E. Mauceli, T. Miyake, M. G. Mueller, D. R. Nelson, A. Nitsche, E. Olmo, T. Ota, A.

- Pallavicini, S. Panji, B. Picone, C. P. Ponting, S. J. Prohaska, D. Przybylski, N. R. Saha, V. Ravi, F. J. Ribeiro, T. Sauka-Spengler, G. Scapigliati, S. M. J. Searle, T. Sharpe, O. Simakov, P. F. Stadler, J. J. Stegeman, K. Sumiyama, D. Tabbaa, H. Tafer, J. Turner-Maier, P. van Heusden, S. White, L. Williams, M. Yandell, H. Brinkmann, J.-N. Volff, C. J. Tabin, N. Shubin, M. Schartl, D. B. Jaffe, B. Venkatesh, F. Di Palma, E. S. Lander, A. Meyer, K. Lindblad-Toh, The African coelacanth genome provides insights into tetrapod evolution. *Nature* **496**, 311–316 (2013).
420. B. J. Swalla, Deuterostome ancestors and chordate origins. *Integr. Comp. Biol.* **64**, 1175–1181 (2024).
421. H. Luo, F. Gao, Y. Lin, Evolutionary conservation analysis between the essential and nonessential genes in bacterial genomes. *Sci. Rep.* **5**, 13210 (2015).
422. J. Söding, A. Biegert, A. N. Lupas, The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33**, W244-8 (2005).
423. R. Chikhi, T. Lemane, R. Loll-Krippelber, M. Montoliu-Nerin, B. Raffestin, A. P. Camargo, C. J. Miller, M. B. Fiamenghi, D. P. Agostinho, S. Majidian, G. Autric, M. Hugues, J. Lee, R. Faure, K. D. Curry, J. A. Moura de Sousa, E. P. C. Rocha, D. Koslicki, P. Medvedev, P. Gupta, J. Shen, A. Morales-Tapia, K. Sihuta, P. J. Roy, G. W. Brown, R. C. Edgar, A. Korobeynikov, M. Steinegger, C. A. Lareau, P. Peterlongo, A. Babaian, Logan: Planetary-scale genome assembly surveys life's diversity, *bioRxiv* (2025)p. 2024.07.30.605881.
424. A. Pallavicini, A. Canapa, M. Barucca, J. Alföldi, M. A. Biscotti, F. Buonocore, G. De Moro, F. Di Palma, A. M. Fausto, M. Forconi, M. Gerdol, D. M. Makapedua, J. Turner-Meier, E. Olmo, G. Scapigliati, Analysis of the transcriptome of the Indonesian coelacanth *Latimeria menadoensis*. *BMC Genomics* **14**, 538 (2013).
425. J. Sprague, D. Clements, T. Conlin, P. Edwards, K. Frazer, K. Schaper, E. Segerdell, P. Song, B. Sprunger, M. Westerfield, The Zebrafish Information Network (ZFIN): the zebrafish model organism database. *Nucleic Acids Res.* **31**, 241–243 (2003).
426. E. A. Bassett, J. DeNizio, M. C. Barnhart-Dailey, T. Panchenko, N. Sekulic, D. J. Rogers, D. R. Foltz, B. E. Black, HJURP uses distinct CENP-A surfaces to recognize and to stabilize CENP-A/histone H4 for centromere assembly. *Dev. Cell* **22**, 749–762 (2012).
427. T. Hori, J. Cao, K. Nishimura, M. Ariyoshi, Y. Arimura, H. Kurumizaka, T. Fukagawa, Essentiality of CENP-A Depends on Its Binding Mode to HJURP. *Cell Rep.* **33**, 108388 (2020).
428. R. Bernad, P. Sánchez, T. Rivera, M. Rodríguez-Corsino, E. Boyarchuk, I. Vassias, D. Ray-Gallet, A. Arnaoutov, M. Dasso, G. Almouzni, A. Losada, Xenopus HJURP and condensin II are required for CENP-A assembly. *J. Cell Biol.* **192**, 569–582 (2011).

429. M. Bui, E. K. Dimitriadis, C. Hoischen, E. An, D. Quénet, S. Giebe, A. Nita-Lazar, S. Diekmann, Y. Dalal, Cell-cycle-dependent structural transitions in the human CENP-A nucleosome in vivo. *Cell* **150**, 317–326 (2012).
430. A. Sur, Y. Wang, P. Capar, G. Margolin, M. K. Prochaska, J. A. Farrell, Single-cell analysis of shared signatures and transcriptional diversity during zebrafish development. *Dev. Cell* **58**, 3028–3047.e12 (2023).
431. D. R. Farnsworth, L. M. Saunders, A. C. Miller, A single-cell transcriptome atlas for zebrafish development. *Dev. Biol.* **459**, 100–108 (2020).
432. Thisse, B., Pflumio, S., Fürthauer, M., Loppin, B., Heyer, V., Degraeve, A., Woehl, R., Lux, A., Steffan, T., Charbonnier, X.Q. and Thisse, C., “Expression of the zebrafish genome during embryogenesis” (R01RR15402, NIH).
433. R. B. Schittenhelm, F. Althoff, S. Heidmann, C. F. Lehner, Detrimental incorporation of excess Cenp-A/Cid and Cenp-C into Drosophila centromeres is prevented by limiting amounts of the bridging factor Cal1. *J. Cell Sci.* **123**, 3768–3779 (2010).
434. D. L. Bodor, L. P. Valente, J. F. Mata, B. E. Black, L. E. T. Jansen, Assembly in G1 phase and long-term stability are unique intrinsic features of CENP-A nucleosomes. *Mol. Biol. Cell* **24**, 923–932 (2013).
435. K. Karplus, C. Barrett, R. Hughey, Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**, 846–856 (1998).
436. S. R. Eddy, “HMMER user’s guide” (2023); <http://eddylab.org/software/hmmer/Userguide.pdf>.
437. S. C. Potter, A. Luciani, S. R. Eddy, Y. Park, R. Lopez, R. D. Finn, HMMER web server: 2018 update. *Nucleic Acids Res.* **46**, W200–W204 (2018).
438. L. Rosin, B. G. Mellone, Co-evolving CENP-A and CAL1 domains mediate centromeric CENP-A deposition across Drosophila species. *Dev. Cell* **37**, 136–147 (2016).
439. B. L. Carty, A. A. Dattoli, E. M. Dunleavy, CENP-C functions in centromere assembly, the maintenance of CENP-A asymmetry and epigenetic age in Drosophila germline stem cells. *PLoS Genet.* **17**, e1009247 (2021).
440. Y. Unhavaithaya, T. L. Orr-Weaver, Centromere proteins CENP-C and CAL1 functionally interact in meiosis for centromere clustering, pairing, and chromosome segregation. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 19878–19883 (2013).
441. R. Kitagawa, Key players in chromosome segregation in *Caenorhabditis elegans*. *Front. Biosci. (Landmark Ed.)* **14**, 1529–1557 (2009).
442. B. C. H. Lee, Z. Lin, K. W. Y. Yuen, RbAp46/48(LIN-53) Is Required for Holocentromere Assembly in *Caenorhabditis elegans*. *Cell Rep.* **14**, 1819–1828 (2016).

443. S. Tandonnet, M. Haq, A. Turner, T. Grana, P. Paganopoulou, S. Adams, S. Dhawan, N. Kanzaki, I. Nuez, M.-A. Félix, A. Pires-daSilva, De Novo genome assembly of *Auanema melissensis*, a trioecious free-living nematode. *J. Nematol.* **54**, 20220059 (2022).
444. P. Bhan, M. Muthaiyan Shanmugam, D. Wang, O. Bayansan, C.-W. Chen, O. I. Wagner, Characterization of TAG-63 and its role on axonal transport in *C. elegans*. *Traffic* **21**, 231–249 (2020).
445. X. Qing, Y. M. Zhang, S. Sun, M. Ahmed, W.-S. Lo, W. Bert, O. Holovachov, H. Li, Phylogenomic insights into the evolution and origin of Nematoda. *Syst. Biol.* **74**, 349–358 (2025).
446. A. Ghaddar, E. Armingol, C. Huynh, L. Gevirtzman, N. E. Lewis, R. Waterston, E. J. O'Rourke, Whole-body gene expression atlas of an adult metazoan. *Sci. Adv.* **9**, eadg0506 (2023).
447. J. S. Packer, Q. Zhu, C. Huynh, P. Sivaramakrishnan, E. Preston, H. Dueck, D. Stefanik, K. Tan, C. Trapnell, J. Kim, R. H. Waterston, J. I. Murray, A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. *Science* **365**, eaax1971 (2019).
448. T. Gong, F. J. McNally, *Caenorhabditis elegans* spermatocytes can segregate achiasmate homologous chromosomes apart at higher than random frequency during meiosis I. *Genetics* **223** (2023).
449. K. Oegema, A. Desai, S. Rybina, M. Kirkham, A. A. Hyman, Functional analysis of kinetochore assembly in *Caenorhabditis elegans*. *J. Cell Biol.* **153**, 1209–1226 (2001).
450. B. J. Buchwitz, K. Ahmad, L. L. Moore, M. B. Roth, S. Henikoff, A histone-H3-like protein in *C. elegans*: Cell division. *Nature* **401**, 547–548 (1999).
451. G. E. Ashley, T. Duong, M. T. Levenson, M. A. Q. Martinez, L. C. Johnson, J. D. Hibshman, H. N. Saeger, N. J. Palmisano, R. Doonan, R. Martinez-Mendez, B. R. Davidson, W. Zhang, J. M. Ragle, T. N. Medwig-Kinney, S. S. Sirota, B. Goldstein, D. Q. Matus, D. J. Dickinson, D. J. Reiner, J. D. Ward, An expanded auxin-inducible degron toolkit for *Caenorhabditis elegans*. *Genetics* **217** (2021).
452. L. Zhang, J. D. Ward, Z. Cheng, A. F. Dernburg, The auxin-inducible degradation (AID) system enables versatile conditional protein depletion in *C. elegans*. *Development* **142**, 4374–4384 (2015).
453. R. Gassmann, A. Rechtsteiner, K. W. Yuen, A. Muroyama, T. Egelhofer, L. Gaydos, F. Barron, P. Maddox, A. Essex, J. Monen, S. Ercan, J. D. Lieb, K. Oegema, S. Strome, A. Desai, An inverse relationship to germline transcription defines centromeric chromatin in *C. elegans*. *Nature* **484**, 534–537 (2012).
454. E. M. Dunleavy, N. L. Beier, W. Gorgescu, J. Tang, S. V. Costes, G. H. Karpen, The cell cycle timing of centromeric chromatin assembly in *Drosophila* meiosis is distinct from mitosis yet requires CAL1 and CENP-C. *PLoS Biol.* **10**, e1001460 (2012).

455. T. Kumon, J. Ma, R. B. Akins, D. Stefanik, C. E. Nordgren, J. Kim, M. T. Levine, M. A. Lampson, Parallel pathways for recruiting effector proteins determine centromere drive and suppression. *Cell* **184**, 4904–4918.e11 (2021).
456. S. J. Mondo, I. V. Grigoriev, A genomic perspective on fungal diversity and evolution. *Nat. Rev. Microbiol.* **23**, 718–733 (2025).
457. M. Carr, B. S. C. Leadbeater, R. Hassan, M. Nelson, S. L. Baldauf, Molecular phylogeny of choanoflagellates, the sister group to Metazoa. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 16641–16646 (2008).
458. C. W. Dunn, G. Giribet, G. D. Edgecombe, A. Hejnol, Animal phylogeny and its evolutionary implications. *Annu. Rev. Ecol. Evol. Syst.* **45**, 371–395 (2014).
459. S. Henikoff, K. Ahmad, H. S. Malik, The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**, 1098–1102 (2001).
460. H. S. Malik, S. Henikoff, Adaptive evolution of Cid, a centromere-specific histone in *Drosophila*. *Genetics* **157**, 1293–1298 (2001).
461. L. Piovani, D. Gavriouchkina, E. Parey, L. A. Sarre, K. T. C. A. Peijnenburg, J. M. Martín-Durán, D. S. Rokhsar, N. Satoh, A. de Mendoza, T. Goto, F. Marlétaz, The genomic origin of the unique chaetognath body plan. *Nature* **645**, 964–973 (2025).
462. I. A. Drinnenberg, D. deYoung, S. Henikoff, H. S. Malik, Recurrent loss of CenH3 is associated with independent transitions to holocentricity in insects. *Elife* **3**, e03676 (2014).
463. M. I. Navarro-Mendoza, C. Pérez-Arques, S. Panchal, F. E. Nicolás, S. J. Mondo, P. Ganguly, J. Pangilinan, I. V. Grigoriev, J. Heitman, K. Sanyal, V. Garre, Early Diverging Fungus *Mucor circinelloides* Lacks Centromeric Histone CENP-A and Displays a Mosaic of Point and Regional Centromeres. *Curr. Biol.* **29**, 3791–3802.e6 (2019).
464. C. M. Hammond, C. B. Strømme, H. Huang, D. J. Patel, A. Groth, Histone chaperone networks shaping chromatin function. *Nat. Rev. Mol. Cell Biol.* **18**, 141–158 (2017).
465. L. Sanchez-Pulido, C. P. Ponting, Extending the horizon of homology detection with coevolution-based structure prediction. *J. Mol. Biol.* **433**, 167106 (2021).
466. C. M. Weisman, A. W. Murray, S. R. Eddy, Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLoS Biol.* **18**, e3000862 (2020).
467. C. Yu, S. R. Sankaranarayanan, G. Cornilleau, A. C. Howes, C. Azumaya, I. Zilberleyb, B. Brillantes, T. K. Cheung, L. Dec, D. Loew, P. Tran, C. M. Rose, I. A. Drinnenberg, C. Ciferri, S. Yatskevich, Architecture and Function of Holocentric CENP-A-Independent Kinetochores, *bioRxiv* (2025)p. 2025.07.14.664805.
468. S. Lidgard, A. C. Love, Rethinking living fossils. *Bioscience* **68**, 760–770 (2018).

469. M. Perpelescu, T. Hori, A. Toyoda, S. Misu, N. Monma, K. Ikeo, C. Obuse, A. Fujiyama, T. Fukagawa, HJURP is involved in the expansion of centromeric chromatin. *Mol. Biol. Cell* **26**, 2742–2754 (2015).
470. S. Müller, R. Montes de Oca, N. Lacoste, F. Dingli, D. Loew, G. Almouzni, Phosphorylation and DNA binding of HJURP determine its centromeric recruitment and function in CenH3(CENP-A) loading. *Cell Rep.* **8**, 190–203 (2014).
471. H. Xiao, G. Mizuguchi, J. Wisniewski, Y. Huang, D. Wei, C. Wu, Nonhistone Scm3 binds to AT-rich DNA to organize atypical centromeric nucleosome of budding yeast. *Mol. Cell* **43**, 369–380 (2011).
472. P. Agarwal, A. Alekar, S. Mallick, K. Harini, S. K. Ghosh, Evidence of centromeric histone 3 chaperone involved in DNA damage repair pathway in budding yeast. *Elife* **14**, e104431 (2025).
473. H. D. Folco, H. Xiao, D. Wheeler, H. Feng, Y. Bai, S. I. S. Grewal, The cysteine-rich domain in CENP-A chaperone Scm3HJURP ensures centromere targeting and kinetochore integrity. *Nucleic Acids Res.* **52**, 1688–1701 (2024).
474. B. T. French, F. G. Westhorpe, C. Limouse, A. F. Straight, *Xenopus laevis* M18BP1 Directly Binds Existing CENP-A Nucleosomes to Promote Centromeric Chromatin Assembly. *Dev. Cell* **42**, 190-199.e10 (2017).
475. M. A. Lampson, B. E. Black, Cellular and Molecular Mechanisms of Centromere Drive. *Cold Spring Harb. Symp. Quant. Biol.* **82**, 249–257 (2017).
476. T. Kato, N. Sato, S. Hayama, T. Yamabuki, T. Ito, M. Miyamoto, S. Kondo, Y. Nakamura, Y. Daigo, Activation of Holliday junction recognizing protein involved in the chromosomal stability and immortality of cancer cells. *Cancer Res.* **67**, 8544–8553 (2007).
477. J. Helsen, K. Ramachandran, G. Sherlock, G. Dey, Progressive coevolution of the yeast centromere and kinetochore. *Nature*, 1–8 (2025).
478. L. E. Kursel, H. S. Malik, The cellular mechanisms and consequences of centromere drive. *Curr. Opin. Cell Biol.* **52**, 58–65 (2018).
479. N. Chen, T. W. Harris, I. Antoshechkin, C. Bastiani, T. Bieri, D. Blasiar, K. Bradnam, P. Canaran, J. Chan, C.-K. Chen, W. J. Chen, F. Cunningham, P. Davis, E. Kenny, R. Kishore, D. Lawson, R. Lee, H.-M. Muller, C. Nakamura, S. Pai, P. Ozersky, A. Petcherski, A. Rogers, A. Sabo, E. M. Schwarz, K. Van Auken, Q. Wang, R. Durbin, J. Spieth, P. W. Sternberg, L. D. Stein, WormBase: a comprehensive data resource for Caenorhabditis biology and genomics. *Nucleic Acids Res.* **33**, D383-9 (2005).
480. Z. Wang, H. Chen, S. Jin, Z. Li, Y. Mei, Y. Xiao, C. Zhao, T. Zhou, F. Li, Y. Liu, K. He, InsectBase 3.0: a comprehensive multi-omics resource for insects. *Nucleic Acids Res.*, gkaf1248 (2025).

481. T. D. Schneider, R. M. Stephens, Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100 (1990).
482. A. Löytynoja, N. Goldman, webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics* **11**, 579 (2010).
483. J. C. F. Servin, A. F. Straight, Repression of CENP-A assembly in metaphase requires HJURP phosphorylation and inhibition by M18BP1, *bioRxiv* (2021).
<https://doi.org/10.1101/2021.10.28.466278>.
484. M. Westerfield, *The Zebrafish Book. A Guide for the Laboratory Use of Zebrafish (Danio Rerio)* (University of Oregon Press, Eugene, OR, 2000).
485. K. R. Siegfried, In search of determinants: gene expression during gonadal sex differentiation. *J. Fish Biol.* **76**, 1879–1902 (2010).
486. R. S. Wu, I. I. Lam, H. Clay, D. N. Duong, R. C. Deo, S. R. Coughlin, A rapid method for directed gene knockout for screening in G0 zebrafish. *Dev. Cell* **46**, 112-125.e4 (2018).
487. J. C. Talbot, S. L. Amacher, A streamlined CRISPR pipeline to reliably generate zebrafish frameshifting alleles. *Zebrafish* **11**, 583–585 (2014).
488. J. N. Rosen, M. F. Sweeney, J. D. Mably, Microinjection of zebrafish embryos to analyze gene function. *J. Vis. Exp.*, doi: 10.3791/1115 (2009).
489. J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, A. Cardona, Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
490. A. Paix, A. Folkmann, G. Seydoux, Precision genome editing using CRISPR-Cas9 and linear repair templates in *C. elegans*. *Methods* **121–122**, 86–93 (2017).
491. K. S. Ghanta, C. C. Mello, Melting dsDNA donor molecules greatly improves precision genome editing in *Caenorhabditis elegans*. *Genetics* **216**, 643–650 (2020).
492. A. E. Russo, S. Giacomazzi, A. Deshong, M. Menon, V. Ortiz, K. M. Ego, K. D. Corbett, N. Bhalla, The conserved AAA ATPase PCH-2 distributes its regulation of meiotic prophase events through multiple meiotic HORMADs in *C. elegans*. *PLoS Genet.* **19**, e1010708 (2023).
493. I. D. Wolff, N. S. Divekar, S. M. Wignall, Methods for investigating cell division mechanisms in *C. elegans*. *Methods Mol. Biol.* **2415**, 19–35 (2022).
494. L. Li, A. Y. Zinovyeva, Protein Extract Preparation and Co-immunoprecipitation from *Caenorhabditis elegans*. *J. Vis. Exp.*, doi: 10.3791/61243 (2020).

495. E. Zanin, J. Dumont, R. Gassmann, I. Cheeseman, P. Maddox, S. Bahmanyar, A. Carvalho, S. Niessen, J. R. Yates 3rd, K. Oegema, A. Desai, Affinity purification of protein complexes in *C. elegans*. *Methods Cell Biol.* **106**, 289–322 (2011).
496. M. A. Hibbs, D. C. Hess, C. L. Myers, C. Huttenhower, K. Li, O. G. Troyanskaya, Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics* **23**, 2692–2699 (2007).
497. O. Belli, K. Karava, R. Farouni, R. J. Platt, Multimodal scanning of genetic variants with base and prime editing. *Nat. Biotechnol.* **43**, 1458–1470 (2025).
498. C. B. Macdonald, D. Nedrud, P. R. Grimes, D. Trinidad, J. S. Fraser, W. Coyote-Maestas, DIMPLE: deep insertion, deletion, and missense mutation libraries for exploring protein variation in evolution, disease, and biology. *Genome Biol.* **24**, 36 (2023).
499. A. R. J. Lawson, F. Abascal, P. A. Nicola, S. V. Lensing, A. L. Roberts, G. Kalantzis, A. Baez-Ortega, N. Brzozowska, J. S. El-Sayed Moustafa, D. Vaitkute, B. Jakupovic, A. Nessa, S. Wadge, M. F. Österdahl, A. L. Paterson, D. M. Rassl, R. E. Alcantara, L. O'Neill, S. Widaa, S. Austin-Guest, M. D. C. Neville, M. J. Przybilla, W. Cheng, M. Morra, L. Sykes, M. Mayho, N. Müller-Sienerth, N. Williams, D. Alexander, L. M. R. Harvey, T. Clarke, A. Byrne, J. R. Blundell, M. D. Young, K. T. A. Mahbubani, K. Saeb-Parsy, H. C. Martin, M. R. Stratton, P. J. Campbell, R. Rahbari, K. S. Small, I. Martincorena, Somatic mutation and selection at population scale. *Nature* **647**, 411–420 (2025).
500. M. D. C. Neville, A. R. J. Lawson, R. Sanghvi, F. Abascal, M. H. Pham, A. Cagan, P. A. Nicola, T. Bayzetinova, A. Baez-Ortega, K. Roberts, S. V. Lensing, S. Widaa, R. E. Alcantara, M. P. García, S. Wadge, M. R. Stratton, P. J. Campbell, K. Small, I. Martincorena, M. E. Hurles, R. Rahbari, Sperm sequencing reveals extensive positive selection in the male germline. *Nature* **647**, 421–428 (2025).
501. X. Liao, B. Lehner, Allosteric is a widespread cause of loss-of-function variant pathogenicity, *bioRxiv* (2025)p. 2025.06. 20.660737.
502. A. Escobedo, G. Voigt, A. J. Faure, B. Lehner, Genetics, energetics, and allostery in proteins with randomized cores and surfaces. *Science* **389**, eadq3948 (2025).
503. R. Odai, M. Leemann, T. Al-Murad, M. Abdullah, L. Shyrokov, T. Tenson, V. Hauryliuk, J. Durairaj, J. Pereira, G. C. Atkinson, The Viral AlphaFold Database of monomers and homodimers reveals conserved protein folds in viruses of bacteria, archaea, and eukaryotes. *Sci. Adv.* **11**, eadz8560 (2025).
504. W. Yang, S. Wang, G. R. Lee, J. Z. Zhang, A. Courbet, D. Juergens, X. Wang, T. Schlichthaerle, M. Abedi, R. Ragotte, L. An, I. Kalvet, S. Pellock, L. Mihaljevic, C. Glasscock, A. Pillai, A. Broerman, N. Ennist, E. Haefner, N. McNamara-Bordewick, I. Haydon, L. Stewart, G. Bhardwaj, D. Baker, The past, present and future of de novo protein design. *Nature* **652**, 1139–1152 (2026).

505. A. Courbet, J. Hansen, Y. Hsia, N. Bethel, Y.-J. Park, C. Xu, A. Moyer, S. E. Boyken, G. Ueda, U. Nattermann, D. Nagarajan, D.-A. Silva, W. Sheffler, J. Quispe, A. Nord, N. King, P. Bradley, D. Veisler, J. Kollman, D. Baker, Computational design of mechanically coupled axle-rotor protein assemblies. *Science* **376**, 383–390 (2022).
506. L. N. Young, E. Villa, Bringing structure to cell biology with cryo-electron tomography. *Annu. Rev. Biophys.* **52**, 573–595 (2023).
507. M. Turk, W. Baumeister, The promise and the challenges of cryo-electron tomography. *FEBS Lett.* **594**, 3243–3261 (2020).
508. E. Nogales, J. Mahamid, Bridging structural and cell biology with cryo-electron microscopy. *Nature* **628**, 47–56 (2024).
509. E. G. Armbruster, P. Rani, J. Lee, N. Klusch, J. Hutchings, L. Y. Hoffman, H. Buschkaemper, E. Enustun, B. A. Adler, K. Inlow, A. R. VanderWal, M. Y. Hoffman, D. Daksh, A. Aindow, A. Deep, Z. K. Rodriguez, C. J. Morgan, M. Ghassemian, T. G. Laughlin, E. Charles, B. F. Cress, D. F. Savage, J. A. Doudna, K. Pogliano, K. D. Corbett, E. Villa, J. Pogliano, Sequential membrane- and protein-bound organelles compartmentalize genomes during phage infection. *Cell Host Microbe* **33**, 484-497.e6 (2025).
510. S. Chen, T. Basiashvili, J. Hutchings, M. Sanz Murillo, A. Villagran Suarez, E. Xiong, J. Alegrio Louro, A. E. Leschziner, E. Villa, Cryo-electron tomography reveals the microtubule-bound form of inactive LRRK2. *Elife* **13**, RP100799 (2025).
511. M. Bergqvist, K.-S. Park, N. Karimi, L. Yu, C. Lässer, J. Lötvall, Extracellular vesicle surface engineering with integrins (ITGAL & ITGB2) to specifically target ICAM-1-expressing endothelial cells. *J. Nanobiotechnology* **23**, 64 (2025).
512. L. N. Young, A. Sherrard, H. Zhou, F. Shaikh, J. Hutchings, M. Riggi, M. Narasimhan, W. A. Flaherty, E. J. Bennett, M. K. Rosen, A. J. Giraldez, E. Villa, ExoSloNano: multimodal nanogold labels for identification of macromolecules in live cells and cryo-electron tomograms. *Nat. Methods* **23**, 131–142 (2026).
513. J. Xu, Recent advances in integrative cryoEM/ET toward visualizing the molecular sociology of cellular organelles. *Curr. Opin. Struct. Biol.* **98**, 103287 (2026).
514. K. Kixmoeller, E. V. Tarasovets, E. Mer, Y.-W. Chang, B. E. Black, Centromeric chromatin clearings demarcate the site of kinetochore formation. *Cell* **188**, 1280-1296.e19 (2025).
515. H. Zhang, M. Yang, S. H. Kim, I. T. S. Li, Integrin force loading rate in mechanobiology: From model to molecular measurement. *QRB Discov.* **6**, e9 (2025).
516. S. J. Shattil, C. Kim, M. H. Ginsberg, The final steps of integrin activation: the end game. *Nat. Rev. Mol. Cell Biol.* **11**, 288–300 (2010).

517. M. J. Flick, X. Du, D. P. Witte, M. Jirousková, D. A. Soloviev, S. J. Busuttill, E. F. Plow, J. L. Degen, Leukocyte engagement of fibrin(ogen) via the integrin receptor α 5 β 1 is critical for host inflammatory response in vivo. *J. Clin. Invest.* **113**, 1596–1606 (2004).
518. Z. J. Deng, M. Liang, M. Monteiro, I. Toth, R. F. Minchin, Nanoparticle-induced unfolding of fibrinogen promotes Mac-1 receptor activation and inflammation. *Nat. Nanotechnol.* **6**, 39–44 (2011).
519. J. Pillay, V. M. Kamp, M. Pennings, E.-J. Oudijk, L. P. Leenen, L. H. Ulfman, L. Koenderman, Acute-phase concentrations of soluble fibrinogen inhibit neutrophil adhesion under flow conditions in vitro through interactions with ICAM-1 and MAC-1 (CD11b/CD18). *J. Thromb. Haemost.* **11**, 1172–1182 (2013).
520. Y. Chen, Z. Li, F. Kong, L. A. Ju, C. Zhu, Force-regulated spontaneous conformational changes of integrins α 5 β 1 and α V β 3. *ACS Nano* **18**, 299–313 (2024).
521. R. Alon, M. L. Dustin, Force as a facilitator of integrin conformational changes during leukocyte arrest on blood vessels and antigen-presenting cells. *Immunity* **26**, 17–27 (2007).
522. M. Lerche, M. Mathieu, H. Hamidi, M. Chastney, G. Jacquemet, B. M. H. Bruininks, S. Kaptan, L. Malerød, N. M. Pedersen, A. Brech, N. Matoba, Y. Sato, I. Vattulainen, P. Roca-Cusachs, F. Perez, G. Boncompain, S. Miserey, J. Ivaska, Regulation of cell dynamics by rapid integrin transport through the biosynthetic pathway. *J. Cell Biol.* **225**, e202508155 (2026).
523. C. Huang, C. Lu, T. A. Springer, Folding of the conserved domain but not of flanking regions in the integrin β 2 subunit requires association with the α subunit. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 3156–3161 (1997).
524. J. Heino, R. A. Ignatz, M. E. Hemler, C. Crouse, J. Massagué, Regulation of cell adhesion receptors by transforming growth factor- β . Concomitant regulation of integrins that share a common β 1 subunit. *J. Biol. Chem.* **264**, 380–388 (1989).
525. B. De Strooper, F. Van Leuven, G. Carmeliet, H. Van Den Berghe, J. J. Cassiman, Cultured human fibroblasts contain a large pool of precursor β 1-integrin but lack an intracellular pool of mature subunit. *Eur. J. Biochem.* **199**, 25–33 (1991).
526. G. Sun, E. Guillon, S. A. Holley, Integrin intra-heterodimer affinity inversely correlates with integrin activatability. *Cell Rep.* **35**, 109230 (2021).
527. M. Lenter, D. Vestweber, The integrin chains β 1 and α 6 associate with the chaperone calnexin prior to integrin assembly. *J. Biol. Chem.* **269**, 12263–12268 (1994).
528. V. Martel, L. Vignoud, S. Dupé, P. Frachet, M. R. Block, C. Albigès-Rizo, Talin controls the exit of the integrin α 5 β 1 from an early compartment of the secretory pathway. *J. Cell Sci.* **113** (Pt 11), 1951–1961 (2000).

529. S. Tiwari, J. A. Askari, M. J. Humphries, N. J. Bulleid, Divalent cations regulate the folding and activation status of integrins during their intracellular trafficking. *J. Cell Sci.* **124**, 1672–1680 (2011).
530. S. Sun, G. Shi, X. Han, A. B. Francisco, Y. Ji, N. Mendonça, X. Liu, J. W. Locasale, K. W. Simpson, G. E. Duhamel, S. Kersten, J. R. Yates 3rd, Q. Long, L. Qi, Sel1L is indispensable for mammalian endoplasmic reticulum-associated degradation, endoplasmic reticulum homeostasis, and survival. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E582-91 (2014).
531. L. L. Lin, H. H. Wang, B. Pederson, X. Wei, M. Torres, Y. Lu, Z. J. Li, X. Liu, H. Mao, H. Wang, L. E. Zhou, Z. Zhao, S. Sun, L. Qi, SEL1L-HRD1 interaction is required to form a functional HRD1 ERAD complex. *Nat. Commun.* **15**, 1440 (2024).
532. L. Guo, G. Liu, J. He, X. Jia, Y. He, Z. Wang, H. Qian, Structural insights into the human HRD1 ubiquitin ligase complex. *Nat. Commun.* **16**, 6007 (2025).
533. G. R. Diaferia, V. Cirulli, I. Biunno, SEL1L regulates adhesion, proliferation and secretion of insulin by affecting integrin signaling. *PLoS One* **8**, e79458 (2013).
534. X. Wei, Y. Lu, L. Qi, Regulation of podocyte adhesion by SEL1L-HRD1 ERAD via integrin: Th-po771. *J. Am. Soc. Nephrol.* **34**, 305–305 (2023).
535. X. Wei, Y. Lu, L. L. Lin, C. Zhang, X. Chen, S. Wang, S. A. Wu, Z. J. Li, Y. Quan, S. Sun, L. Qi, Proteomic screens of SEL1L-HRD1 ER-associated degradation substrates reveal its role in glycosylphosphatidylinositol-anchored protein biogenesis. *Nat. Commun.* **15**, 659 (2024).
536. Y. Lin, Y. Zhu, L. Jing, X. Lei, Z. Xie, Regulation of viral replication by host restriction factors. *Front. Immunol.* **16**, 1484119 (2025).
537. M. Colomer-Lluch, A. Ruiz, A. Moris, J. G. Prado, Restriction factors: From intrinsic viral restriction to shaping cellular immunity against HIV-1. *Front. Immunol.* **9**, 2876 (2018).
538. J. Verhelst, P. Hulpiau, X. Saelens, Mx proteins: antiviral gatekeepers that restrain the uninvited. *Microbiol. Mol. Biol. Rev.* **77**, 551–566 (2013).
539. J. Verhelst, E. Parthoens, B. Schepens, W. Fiers, X. Saelens, Interferon-inducible protein Mx1 inhibits influenza virus by interfering with functional viral ribonucleoprotein complex assembly. *J. Virol.* **86**, 13445–13455 (2012).
540. C. Patzina, O. Haller, G. Kochs, Structural requirements for the antiviral activity of the human MxA protein against Thogoto and influenza A virus. *J. Biol. Chem.* **289**, 6020–6027 (2014).
541. G. Kochs, O. Haller, GTP-bound human MxA protein interacts with the nucleocapsids of Thogoto virus (Orthomyxoviridae). *J. Biol. Chem.* **274**, 4370–4376 (1999).

542. P. S. Mitchell, C. Patzina, M. Emerman, O. Haller, H. S. Malik, G. Kochs, Evolution-guided identification of antiviral specificity determinants in the broadly acting interferon-induced innate immunity factor MxA. *Cell Host Microbe* **12**, 598–604 (2012).
543. P. E. Nigg, J. Pavlovic, Oligomerization and GTP-binding requirements of MxA for viral target recognition and antiviral activity against influenza A virus. *J. Biol. Chem.* **290**, 29893–29906 (2015).
544. K. Turan, M. Mibayashi, K. Sugiyama, S. Saito, A. Numajiri, K. Nagata, Nuclear MxA proteins form a complex with influenza virus NP and inhibit the transcription of the engineered influenza virus genome. *Nucleic Acids Res.* **32**, 643–652 (2004).
545. S. Gao, A. von der Malsburg, S. Paeschke, J. Behlke, O. Haller, G. Kochs, O. Daumke, Structural basis of oligomerization in the stalk region of dynamin-like MxA. *Nature* **465**, 502–506 (2010).
546. S. Gao, A. von der Malsburg, A. Dick, K. Faelber, G. F. Schröder, O. Haller, G. Kochs, O. Daumke, Structure of myxovirus resistance protein a reveals intra- and intermolecular domain interactions required for the antiviral function. *Immunity* **35**, 514–525 (2011).
547. Y.-L. Li, C. A. Langley, C. M. Azumaya, I. Echeverria, N. M. Chesarino, M. Emerman, Y. Cheng, J. D. Gross, The structural basis for HIV-1 Vif antagonism of human APOBEC3G. *Nature* **615**, 728–733 (2023).
548. R. A. Geiger, D. Khera, J. L. Tenthorey, G. Kochs, L. Graf, M. Emerman, H. S. Malik, Heterozygous and generalist MxA super-restrictors overcome breadth-specificity trade-offs in antiviral restriction. *Sci. Adv.* **11**, eadu0062 (2025).
549. R. Colón-Thillet, E. Hsieh, L. Graf, R. N. McLaughlin Jr, J. M. Young, G. Kochs, M. Emerman, H. S. Malik, Combinatorial mutagenesis of rapidly evolving residues yields super-restrictor antiviral proteins. *PLoS Biol.* **17**, e3000181 (2019).
550. F. J. D. Alvarez, S. He, J. R. Perilla, S. Jang, K. Schulten, A. N. Engelman, S. H. W. Scheres, P. Zhang, CryoEM structure of MxB reveals a novel oligomerization interface critical for HIV restriction. *Sci. Adv.* **3**, e1701264 (2017).
551. G. Kochs, C. Janzen, H. Hohenberg, O. Haller, Antivirally active MxA protein sequesters La Crosse virus nucleocapsid protein into perinuclear complexes. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 3153–3158 (2002).
552. Y.-S. Tang, S. Xu, Y.-W. Chen, J.-H. Wang, P.-C. Shaw, Crystal structures of influenza nucleoprotein complexed with nucleic acid provide insights into the mechanism of RNA interaction. *Nucleic Acids Res.* **49**, 4144–4154 (2021).
553. A. Labaronne, C. Swale, A. Monod, G. Schoehn, T. Crépin, R. W. H. Ruigrok, Binding of RNA by the nucleoproteins of influenza viruses A and B. *Viruses* **8**, E247 (2016).

554. V. E. Deneke, J. P. Suwita, H. Wang, S. Tonai, Y. Lu, K. Panser, A. Schleiffer, J. A. Hollis, M. Novatchkova, G. Dürnberger, K. Stejskal, G. Krssakova, A. Blaha, A. A. R. Andresan, M. Mirus, H. Marvanova, H.-Y. Chang, T. Noda, A. Burga, E. Roitinger, M. Ikawa, A. Pauli, The SPARK complex forms the molecular basis of vertebrate fertilization, *bioRxiv* (2026)p. 2026.05.13.724128.
555. K. Georgadaki, N. Khoury, D. A. Spandidos, V. Zoumpourlis, The molecular basis of fertilization (Review). *Int. J. Mol. Med.* **38**, 979–986 (2016).
556. K. K. Siu, V. H. B. Serrão, A. Ziyat, J. E. Lee, The cell biology of fertilization: Gamete attachment and fusion. *J. Cell Biol.* **220**, e202102146 (2021).
557. W. J. Swanson, Z. Yang, M. F. Wolfner, C. F. Aquadro, Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 2509–2514 (2001).
558. A. L. Dapper, M. J. Wade, Relaxed selection and the rapid evolution of reproductive genes. *Trends Genet.* **36**, 640–649 (2020).
559. J. A. Carlisle, W. J. Swanson, Molecular mechanisms and evolution of fertilization proteins. *J. Exp. Zool. B Mol. Dev. Evol.* **336**, 652–665 (2021).
560. N. L. Clark, J. Gasper, M. Sekino, S. A. Springer, C. F. Aquadro, W. J. Swanson, Coevolution of interacting fertilization proteins. *PLoS Genet.* **5**, e1000570 (2009).
561. V. D. Vacquier, W. J. Swanson, Selection in the rapid evolution of gamete recognition proteins in marine invertebrates. *Cold Spring Harb. Perspect. Biol.* **3**, a002931 (2011).
562. V. D. Vacquier, W. J. Swanson, Y. H. Lee, Positive Darwinian selection on two homologous fertilization proteins: what is the selective pressure driving their divergence? *J. Mol. Evol.* **44 Suppl 1**, S15-22 (1997).
563. V. D. Vacquier, Evolution of gamete recognition proteins. *Science* **281**, 1995–1998 (1998).
564. D. B. Wilburn, W. J. Swanson, From molecules to mating: Rapid evolution and biochemical studies of reproductive proteins. *J. Proteomics* **135**, 12–25 (2016).
565. S. Guindon, F. Lethiec, P. Duroux, O. Gascuel, PHYML Online—a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res.* **33**, W557–W559 (2005).
566. S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, O. Gascuel, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).