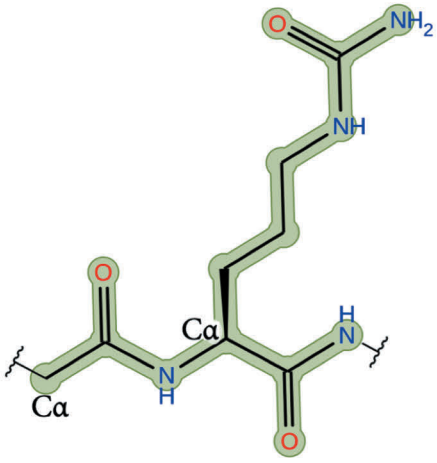
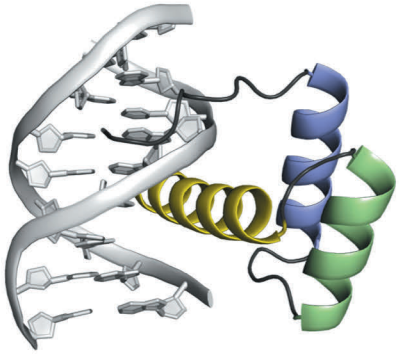
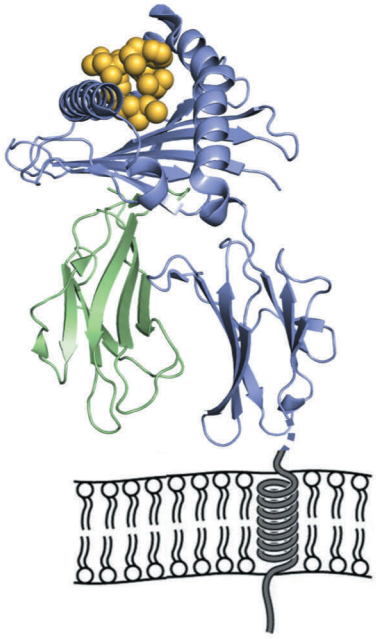
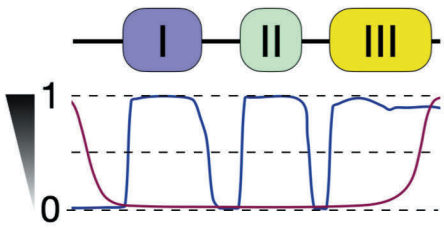
Vipin Ranga

Specificity Determining Features at the Interface of Biomolecular Complexes as Regulators of Biological Functions









Åbo Akademi University

Vipin Ranga

Born 1986, Rajasthan, India

Previous studies and degrees

Master of Technology in Computational and Systems Biology, Jawaharlal Nehru University, New Delhi, India, 2014

Bachelor of Technology in Biotechnology, Indian Institute of Technology Guwahati, India, 2010

Specificity Determining Features at the Interface of Biomolecular Complexes as Regulators of Biological Functions

Vipin Ranga

Biochemistry
Faculty of Science and Engineering
Åbo Akademi University
Åbo, Finland, 2022

From the Faculty of Science and Engineering, Åbo Akademi University; the InFLAMES Research Flagship Center; Åbo Akademi Graduate School and National Doctoral Programme in Informational and Structural Biology

Supervised by

Professor Mark S. Johnson
Structural Bioinformatics Laboratory, Biochemistry
Faculty of Science and Engineering
InFLAMES Research Flagship Center
Åbo Akademi University
Turku, Finland

Adjunct Professor Tomi T. Airenne
Structural Bioinformatics Laboratory, Biochemistry
Faculty of Science and Engineering
InFLAMES Research Flagship Center
Åbo Akademi University
Turku, Finland

Reviewed by

Professor Ramanathan Sowdhamini
National Centre for Biological Sciences
Tata Institute of Fundamental Research
Bangalore, Karnataka, India

Docent Henri Xhaard
Drug Research Program
Division of Pharmaceutical Chemistry and Technology
Faculty of Pharmacy
University of Helsinki
Helsinki, Finland

Opponent

Adjunct Professor Maija Lahtela-Kakkonen
School of Pharmacy
Faculty of Health Sciences
University of Eastern Finland
Kuopio, Finland

Cover: Left: Two-dimensional representation of the side-chain and main-chain atoms of peptidylcitrulline. Middle: Three-dimensional structural model of the human LEUTX homeodomain with double-stranded DNA motif. Plots show predicted helices and disordered regions. Right: Three-dimensional structure of the HLA-A*02 allotype (PDB ID: 5TEZ) with a predicted SARS-CoV-2 epitope. Images by Vipin Ranga.

ISBN 978-952-12-4253-3 (printed)
ISBN 978-952-12-4254-0 (digital)
Painosalama, Turku, Finland 2022

To the memory of my grandfather, Hardwari Lal Ranga

Abstract

Amino acid residues at the biomolecular interface play essential roles in many biological and cellular processes; relevant to this thesis, protein-protein interactions regulate signaling pathways and enzymatic activity, whereas protein-DNA interactions control gene expression, and protein-peptide interactions are central to the immune system. Biomolecular recognition and binding stability are largely determined by residues at the molecular interface. In this thesis, we focused on three biological datasets that are related to humans and human health: 1) dysregulated citrullination in the inflamed joints of rheumatoid arthritis patients, 2) a novel family of PRD-like transcription factors critical to the first few cell divisions in human life, and 3) epitopes that likely activate a cytotoxic T cell-mediated immune response against SARS-CoV-2 infection. For each dataset, in order to study the structural and functional consequences of molecular interactions, we applied a wide range of bioinformatics techniques to analyze sequences, structures and biological data retrieved from various databases, as well as taking into account experimental results from collaborators and from the literature.

In rheumatoid arthritis, normally cytoplasmic peptidylarginine deiminase (PAD) enzymes citrullinate arginine residues in extracellular matrix (ECM) proteins. To examine specificity determining features that regulate the citrullination activity, we analyzed the sequence and structure data of the ECM proteins that were found citrullinated in chronically inflamed human joints. For citrullination, we found that an arginine side chain needs to be exposed to solvent but can arise from β-strands, α-helices, loops and β-turns. Moreover, there is no sequence motif linked to enzymatic activity. In addition, we studied the effect of citrullination on proteins important for a normal ECM, focusing on integrin binding to fibronectin and transforming growth factor-β (TGF-β). Citrullination of these proteins was found to inhibit cell attachment and spreading since PAD-treatment of the isoDGR motif in fibronectin and the RGD motif in TGF-β significantly reduced their binding with integrin αVβ3 and αVβ6, respectively.

The expression of the human paired (PRD)-like transcription factors (TFs) are limited to the period of embryonic genome activation up to the 8-cell stage. We identified that one of these PRD-like TFs, LEUTX, binds to a TAATCC sequence motif. Sequence comparisons revealed that LEUTX protein is comprised of two domains: the DNA-binding homeodomain and a Leutx domain containing a transactivation domain. We identified specificity determining residues in the LEUTX homeodomain that are important for recognition of the TAATCC-containing 36 bp DNA motif enriched in genes involved in embryonic genome activation. We demonstrated using molecular models why a heterozygotic missense mutation A54V at the DNA-specificity determining position of LEUTX has significantly reduced overall transcriptional activity, as well as why the double mutant – I47T and A54V – form of LEUTX restores binding to the DNA motif similarly to that seen in the I47T mutation alone.

At the onset of the COVID-19 pandemic we sought to understand the molecular factors that trigger the cytotoxic T cell-mediated immune response against the SARS-CoV-2 virus, taking advantage of binding data and 3D structures for related viruses and other pathogenic organisms. We first predicted the MHC class I (MHC-I)-specific immunogenic epitopes of length 8- to 11 amino acids from the SARS-CoV-2 proteins. Next, we predicted that the 9-mer epitopes would have the highest potential to elicit a strong immune response. For experimental validation, the predicted 9-mer epitopes were matched with the SARS-CoV-derived epitopes that are known to elicit an effective T cell response *in vitro*. Furthermore, our observations provide a structural explanation for the binding of SARS-CoV-2 epitopes to MHC-I molecules, identifying conserved immunogenic epitopes essential for understanding the pathogenesis of COVID-19.

The three investigated datasets were made in concert with collaborative experimental studies and/or considering publicly available experimental data. The experimental studies generally provided the starting point for the *in silico* studies, which in turn had the objective of providing a detailed explanation of the experimental results. Furthermore, the *in silico* results could be used to devise novel and focused experiments, suggesting that bioinformatics predictions and wet-laboratory experimental investigations optimally take place with multiple advantages. Overall, this thesis demonstrates the synergy that is possible by applying this interdisciplinary approach to understanding the consequences of molecular interactions.

Sammanfattning

Aminosyror i kontaktytan mellan olika biomolekyler spelar en viktig roll i många biologiska och cellulära processer; relevanta interaktioner för den här avhandlingen är protein-protein interaktioner som reglerar signaleringsrutter och enzymatisk aktivitet, protein-DNA interaktioner som kontrollerar genexpression, samt protein-peptid interaktioner som har en central roll i immunförsvaret. Biomolekylär igenkänning och bindningsstabilitet beror till stor del på de aminosyror som finns i den molekylära kontaktytan. I den här avhandlingen fokuserade vi på tre biologiska dataset som är relaterade till människor och människors hälsa: 1) felreglerad citrullinering i inflammerade leder hos patienter med reumatoid artrit, 2) en nyupptäckt familj av PRD (human paired)-lika transkriptionsfaktorer som är nödvändiga för de första celldelningarna i människolivet, och 3) epitoper som troligen aktiverar en cytotoxisk T-cell-förmedlad immunrespons mot SARS-CoV-2 infektioner. För att studera de strukturella och funktionella konsekvenserna av de molekylära interaktionerna i varje dataset, användes en mängd olika bioinformatiska tekniker för att analysera sekvenser, strukturer och biologiska data från olika databaser och dessutom beaktades experimentella resultat från samarbetspartners och från litteraturen.

I reumatoid artrit citrullinerar vanligen PAD (cytoplasmatiska peptidyl arginin deiminas)-enzymer arginin-aminosyror i proteiner i det extracellulära matrixet (ECM). För att undersöka egenskaper som avgör specificiteten hos citrullineringsaktiviteten analyserade vi sekvens- och strukturdata för ECM-proteiner som blir citrullinerade i kroniskt inflammerade leder hos människor. Vi upptäckte att en argininsidokedja måste vara i kontakt med det omgivande lösningsmedlet för att kunna citrullineras, att de kan finnas i beta-strängar, alfa-helixar och beta-svängar, samt att det inte finns några sekvensmotiv som är kopplade till enzymatisk aktivitet. Utöver detta studerade vi effekten av citrullinering på proteiner som är viktiga för normal extracellulär matrix, med fokus på integrinbinding till fibronektin och TGF-β (transforming growth factor-β). Citrullinering av dessa proteiner upptäcktes inhibera cellvidhäftning och spridning eftersom PAD-behandling av isoDGR-motivet i fibronektin och RGD-motivet i TGF-β ordentligt reducerar deras bindning till integrin αVβ3 och αVβ6, respektive.

Expressionsnivåerna av PRD-lika transkriptionsfaktorer (TF) är begränsade till perioden av zygotens genomaktivering upp till 8-cells stadiet. Vi identifierade att en av dessa PRD-lika transkriptionsfaktorer, LEUTX, binder till ett TAATCC sekvensmotiv. Sekvensjämförelser avslöjade att LEUTX proteinet består av två domäner, det DNA-bindande homeodomänet och en leutx-domän som innehåller en transaktiveringsdomän. Vi identifierade specificitetsbestämmande aminosyror i LEUTX homeodomänen som är viktiga för igenkänning av TAATCC-innehållande 36 baspars DNA-motivet som är berikad med gener involverade i zygotens genomaktivering. Vi använde molekylära modeller för att visa varför en heterozygotisk missense-mutation, A54V, i DNA-specificitetsbestämmande

positionen i LEUTX har ordentligt minskad generell transkriptionsaktivitet, och varför dubbelmutanten I47T och A54V återställer bindning till DNA-motivet på samma sätt som observerats i enbart I47T mutationen.

När COVID-19 pandemin inleddes försökte vi förstå de molekylära faktorer som startar den cytotoxiska T-cell-förmedlade immunresponsen mot SARS-CoV-2 viruset, genom att utnyttja bindningsdata och 3D strukturer för relaterade virus och andra patogena organismer. Vi förutspådde först MHC klass I (MHC-I)-specifika immunogena epitoper av längden 8 till 11 aminosyror från SARS-CoV-2 proteiner. Därefter förutspådde vi att epitoper bestående av 9 aminosyror hade den högsta potentialen att orsaka en stark immunrespons. För experimentell validering matchades de 9 aminosyror långa epitoperna med epitoper från SARS-CoV som man vet att orsakar en effektiv T-cell respons *in vitro*. Våra observationer bidrar också med en strukturell förklaring för bindningen av SARS-CoV-2 epitoper till MHC-I molekyler, vilket identifierar konserverade immunogena epitoper som är nödvändiga för att förstår patogenesen hos COVID-19.

De tre undersökta dataseten gjordes i samarbete med experimentella studier och/eller genom att ta allmänt tillgängliga experimentella data i beaktande. De experimentella studierna gav en startpunkt för *in silico*-studierna, vilka i sin tur hade som mål att ge en detaljerad förklaring till de experimentella resultaten. *In silico*-resultaten kan också användas för att utveckla nya och fokuserade experiment, vilket indikerar att bioinformatiska förutspåelser och experimentella studier optimalt sker med många fördelar. Över lag visar denna avhandling synergin som är möjlig genom att använda detta interdisciplinära arbetssätt för att förstå konsekvenserna av molekylära interaktioner.

Table of Contents

List of original publications

This thesis is based on the following four original publications, which are referred to in Roman numerals in the text. Publications are reproduced with the permission of the publishers.

I. Sipilä KH, **Ranga V**, Rappu P, Torittu A, Pirilä L, Käpylä J, Johnson MS, Larjava H, Heino J. Extracellular citrullination inhibits the function of matrix associated TGF-β. Matrix Biology. 2016;55:77-89.

II. Sipilä KH, **Ranga V**, Rappu P, Mali M, Pirilä L, Heino I, Jokinen J, Käpylä J, Johnson MS, Heino J. Joint inflammation related citrullination of functional arginines in extracellular proteins. Scientific Reports. 2017;7:8246.

III. Katayama S, **Ranga V**, Jouhilahti E-M, Airenne TT, Johnson MS, Mukherjee K, Bürglin TR, Kere J. Phylogenetic and mutational analyses of human LEUTX, a homeobox gene implicated in embryogenesis. Scientific Reports. 2018;8:17421.

IV. **Ranga V**[*], Niemelä E[*], Tamirat MZ, Eriksson JE, Airenne TT, Johnson MS. Immunogenic SARS-CoV-2 epitopes: *In silico* study towards better understanding of COVID-19 disease – paving the way for vaccine development. MDPI Vaccines. 2020;8(3):408.

[*]Equal contribution to the work.

Additional publication

I. Vuoristo S, Bhagat S, Hydén-Granskog C, Yoshihara M, Gawriyski L, Jouhilahti E-M, **Ranga V**, Tamirat M, Huhtala M, Kirjanov I, Nykänen S, Krjutškov K, Damdimopoulos A, Weltner J, Hashimoto K, Recher G, Ezer S, Paluoja P, Paloviita P, Takegami Y, Kanemaru A, Lundin K, Airenne TT, Otonkoski T, Tapanainen JS, Kawaji H, Murakawa Y, Bürglin TR, Varjosalo M, Johnson MS, Tuuri T, Katayama S, Kere J. DUX4 is a multifunctional factor priming human embryonic genome activation. iScience. 2022;25(4):104137.

Contributions of the author

The author was involved in the conceptual design of the computational studies in publications I-IV. The author conducted computational tasks, including structural and sequence analyses, molecular modeling, sequence alignment, molecular docking, large-scale data analysis and web-based database searches. The author interpreted the computational results by integrating them with experimental observations from collaborators and from the literature. The author wrote all sections regarding his own work in publications I-IV.

Acknowledgements

The work for this thesis was carried out at the Structural Bioinformatics Laboratory (SBL), Faculty of Science and Engineering, Åbo Akademi University, during 2014-2022. During these years, I have had the amazing opportunity to explore the wonder of biological sciences and to meet many great people and excellent researchers. I would like to thank everyone who made this thesis possible.

First and foremost, I would like to express my most sincere gratitude to my supervisor *Professor Mark S. Johnson* for giving me the opportunity to work under his supervision at SBL. I am very grateful for his support, guidance, kindness, patience and encouragement throughout my PhD journey. *Mark*, you have always appreciated my scientific strengths and patiently encouraged me to overcome my professional weaknesses. Thank you for always believing in my capabilities and helping me to stay motivated until the last minute. Working with you has contributed immensely to my professional and personal growth. I am very grateful to *Professor Mark S. Johnson* and *Professor Tiina Salminen* for providing me the best research facilities that enabled me to develop research skills and to implement my project ideas.

My sincere appreciation and thanks also go to my co-supervisor *Adjunct Professor Tomi T. Airenne* for his guidance and encouragement. *Tomi*, I am grateful for your valuable insight, keen supervision and brainstorming discussions. I truly appreciate you taking the time to teach me the basics of X-ray crystallography. Thank you for always being around to listen patiently to the challenges that I faced in my projects.

Mark and *Tomi*, I will always cherish our wonderful time together, especially the beer brewing course, final tasting exams and brewery visits to Pori. Our collaborative meeting at the Karolinska Institute is one of the most memorable trips I have ever been on. Thanks *Mark* for letting me experience the peaceful silence of the beautiful Finnish forests, the fresh sea breeze and delicious Finnish food during our lab group outings and activities. *Mark* and *Tomi*, I will be forever grateful for the privilege of working with you.

I would like to acknowledge my graduate school "National Doctoral Programme in Informational and Structural Biology (ISB)" at Åbo Akademi University for accepting me as a doctoral candidate and to grant me a research scholarship to support my PhD studies. I am grateful to the director of the graduate school *Professor Mark S. Johnson* and the coordinator *Fredrik Karlsson* for providing me the opportunity to take advantage of such an excellent platform to conduct high-quality scientific studies. It has been a great pleasure to be a part of ISB and its inspiring meetings. *Fred*, thank you so much for always being supportive and helpful with all the administrative work. I would further like to convey my sincere gratitude to my thesis committee members *Professor Peter Slotte*, *Professor Jyrki Heino* and *Docent Jarmo Käpylä* for all the fruitful discussions and feedback on my thesis projects.

My sincere gratitude goes to *Professor Ramanathan Sowdhamini* and *Docent Henri Xhaard* for taking the time to review my thesis and for providing

constructive comments and suggestions in order to improve the overall quality of my thesis. I am also grateful to our collaborators *Professor Jyrki Heino*, *Professor Juha Kere*, *Professor John E. Eriksson* and their research group members. I would like to thank all the co-authors, especially *Kalle H. Sipilä*, *Shintaro Katayama* and *Erik Niemelä* for their incredible contributions to my scientific publications.

I am grateful for the research funding support I received from the Åbo Akademi University Foundation, the Magnus Ehrnrooth Foundation and the Sigrid Jusélius Foundation. Travel grants from the Åbo Akademi University Foundation and the CompLifeSci programs have supported my visits to academic conferences. I would also like to acknowledge CSC – IT Center for Science and Biocenter Finland for providing world-class high performance computing resources.

I would like to thank all the current and former members of SBL for creating a positive and fun-filled working atmosphere. It's a great pleasure working with such kind, caring and wonderful people. I am deeply indebted to *Dr. Jukka Lehtonen* for all his prompt and continuous scientific IT support. *Jukka*, you have truly made things much easier at SBL with all your technical skills and deep knowledge of computers. My gratitude goes to *Adjunct Professor Outi Salo-Ahen* for always taking the time to chat with me about my progress. Thank you *Outi* for all of your time, effort and help.

Thank you *Parthiban Marimuthu* for always sharing your invaluable research experience and scientific knowledge. Your positive attitude toward learning has always inspired and motivated me at work. Thank you *Nitin Agrawal* for being an important part of my PhD journey and doing everything to help me professionally and personally. *Nitin*, I am also grateful for your help guiding me through the technique of X-ray crystallography. *Bhanupratap Singh Chouhan*, a multi-talented person from whom I have learned so much over the years. Thank you *Parthiban, Nitin* and *Bhanu* for all the fun and laughter we have shared together at SBL and outside, and I am sure, the never-ending *lupdi* joke will always keep us united.

I would also like to thank *Mia Åstrand*, *Mahlet Tamirat*, *Mikko Huhtala*, *Ida Alanko*, *Marion Alix*, *Mikael Ilomäki*, *Serhii Vakal*, *Rajendra Bhadane*, *Christine Touma*, *Abris Bendes*, *Alexander Denesyuk, Konstantin Denessiouk, Pekka Postila*, *Polytimi Dimitriou*, *Leonor Carvalho*, *Gabriela Guédez* and *Käthe Dahlström* for creating an excellent learning environment and for all the pleasant conversations we had over the years. Thank you *Mia* for being so helpful in writing my thesis abstract in Swedish and for taking the time to clarify my doubts. Thank you *Mikko* for all of your laboratory experiments and well-organized laboratory reports. Thank you *Mahlet* and *Polytimi* for your wonderful company while traveling to conferences and meetings. Thank you *Käthe* for all your help during the initial stages of my PhD, and yes, it's great to have you back on campus. I wish to thank *Atefeh Saadabadi* for being such a wonderful colleague and friend. I will never forget our long walks with *Fetch* (the cutest dog and *Ati*'s baby) along the river Aura, adventurous hiking in forests, an amazing long-distance walk for picking berries and mushrooms, and all the wonderful memories we have made together.

I would like to thank my friends outside SBL. Thank you *Ankitha* and *Madhukar* for always being there for me. Your encouragement has always made me feel stronger in my goal of doing better. *Anki*, most of your jokes had made me laugh a lot, but few jokes were as terrible as *Maddy*'s exaggerating scientific explanations to a simple non-scientific problem, which actually made me laugh even more. I will cherish every moment we have spent together, especially our evening walks to the Halinen dam, movie nights, countless unplanned tea and dinner parties, long conversations, playing our favorite games, activities in the backyard of 8B1, clicking our last-minute pictures, seeing off, and waving good-bye until the train leaves the Kupittaa station. *Maddy*, I am so happy that you and *Uma* found love in each other, and I wish you both a wonderful and joyful life ahead. *Priyanka* (*Bittu*), thank you for making us feel special with your visits to Turku. Thank you *Kalai*, *Parthiban* and *Amaresh* for always making me feel so welcome and becoming my new family in Turku. *Poonam* and *Subhash*, thank you for always sharing your insight and wisdom. I miss all the fun we used to have together. My gratitude also goes to my friends with whom I celebrated all the Indian festivals and events in Turku. Thank you *Binu*, *Neena*, *Santosh*, *Preeti*, *Deepankar*, *Nikita*, *Shishir*, *Lav*, *Priyanka*, *Avlokita*, *Hasan*, *Arjun*, *Srikar*, *Prasanna*, *Amruta*, *Athresh*, *Kalyan*, *Swathi*, *Sachin*, *Rahul Yewale*, *Rahul Biradar*, *Shruti*, *Jismi* and *Afshan* for all the wonderful memories and fun times!

I would like to thank my old friends for all the wonderful long-distance conversations and for reminding me of the good old days. Thank you *Yogendra*, *Sushil*, *Sujata*, *Hemant, Swetapadma*, *Priyanka*, *Elluri Sheetharami*, *Barkha* and *Rahul*.

I would like to express my gratitude to my family for always standing by my side and giving me strength for everything at every moment. Respected *Mummy* and *Papa ji*, thank you so much for understanding me so well and encouraging me to follow the path that led me here today. I cannot express in words how grateful I am for all the sacrifices that you have made and for all the love that you have shown me. Dear *Sonam*, my sweet and adorable little sister, you have made us all very proud by becoming the first medical doctor in our family. Despite being young, you have always taken care of me, guided me along the right paths and supported me in every decision. Thank you for inspiring me in so many ways throughout my life. I wish you every success and happiness. Respected Mummy and Papa ji (in-laws), thank you for always inspiring me to become the best version of myself. *Jyoti*, my beloved wife, your love, care and trust over the past three years have inspired me to work hard, and believe in myself and my abilities. I cannot thank you enough for all that you have done for me, but I can certainly say that this achievement is the result of your endless support, unwavering patience and deep understanding. Thank you for coming into my life and making it full of joy and happiness.

Vipin Ranga
Turku, November 2022

Abbreviations

3D	Three-dimensional
9aaTAD	Nine-amino-acid transactivation domain
BLAST	Basic local alignment search tool
COVID-19	Coronavirus disease 2019
DPRX	Divergent-paired related homeobox
dsDNA	Double stranded deoxyribonucleic acid
DUX4	Double homeobox 4
ECM	Extracellular matrix
EGA	Embryonic genome activation
FN	Fibronectin
GRAVY	Grand average of hydropathy
HCoV	Human coronavirus
HD	Homeodomain
LAP	Latency-associated peptide
LEUTX	Leucine twenty homeobox
LTBP	Latent TGF-β binding protein
MDS	Molecular dynamics simulation
MHC	Major histocompatibility complex
MSA	Multiple sequence alignment
NCBI	National center for biotechnology information
NETs	Neutrophil extracellular traps
NLS	Nuclear localization signal
NMR	Nuclear magnetic resonance
NSP	Nonstructural protein
PAD	Peptidylarginine deiminase
PDB	Protein data bank
PRD	Paired
RA	Rheumatoid arthritis
RefSeq	Reference sequence
RMSD	Root mean square deviation
SARS	Severe acute respiratory syndrome
SF	Synovial fluid
TCR	T cell receptor
TF	Transcription factor
TGF	Transforming growth factor
TM	Transmembrane
UniProt	Universal protein resource

1. Introduction

Proteins rarely if ever act alone, and about 80% of the human proteins function in complexes (Berggård, Linse, and James 2007). Indeed, the functional properties of proteins are highly dependent on their ability to interact with specific binding partners, such as other proteins, nucleic acids, peptides, small molecules and other entities (*i.e.* ions, electrons, photons). The intermolecular interactions formed within macromolecular complexes are essential for nearly every cellular activity, some of these activities include, *e.g.*, enzyme cooperativity, molecular transport, regulation of gene expression and signal transduction. Proteins are not static entities but have different dynamic motions that allow them to perform specific biological functions (Teilum, Olsen, and Kragelund 2009) and these motions are closely linked to the molecular interactions they form.

In general, the stability of protein interfaces is maintained through both hydrophobic and charged amino acids (Brinda, Kannan, and Vishveshwara 2002; Yan et al. 2008), and the analysis of amino acid interactions is central to understanding molecular interactions and how post-translational alterations and mutations affect those interactions.

In particular, we have considered especially the role of positively charged arginine residues at interfaces. Two basic amino acids, arginine and lysine, can contribute to the stability of protein interfaces with other molecules by forming electrostatic interactions with negatively charged moieties. The arginine side chain has added potential to interact with, *e.g.*, the side chains of non-polar aromatic and aliphatic residues above and below the guanidinium plane as well as making in-plane hydrogen bonding interactions via the three nitrogen atoms of the side chain (Armstrong et al. 2016). It is highly likely that arginine side-chain modifications can readily modulate the interfacial interactions in proteins and, thus, can affect the associated physiological functions.

One such post-translational arginine modification is the hydrolytic conversion of arginine to citrulline by peptidylarginine deiminase (PAD) enzymes, citrullinating *intracellular* proteins under normal physiological conditions. However, PAD released to the extracellular space contributes to the pathogenesis of multifactorial disease such as rheumatoid arthritis (RA). In publications I and II, examination of the synovium of RA patients resulted in the identification of a wide variety of extracellular proteins that are targets of extracellular citrullination, and led to characterization of arginine recognition by PAD enzymes and the consequences of citrullination on molecular interactions. In particular, we carried out in-depth analyses on TGF-β1 and fibronectin in order to study the effect of citrullination on the integrin-mediated cellular processes such as cell adhesion and cell migration (Publications I and II).

Positively charged residues are crucial for transcription factor (TF)-DNA interactions due to their interactions with the negatively charged phosphates in the DNA backbone and electronegative atoms in the nucleotide bases (Corona and Guo 2016; Luscombe, Laskowski, and Thornton 2001), but both polar and

hydrophobic residues also play critical roles in protein-DNA interaction. Polar neutral amino acids, such as serine, threonine, asparagine, and glutamine, comprise the second largest number of hydrogen-bonding interactions with the DNA backbone and bases (Luscombe, Laskowski, and Thornton 2001). The presence of a methyl group in thymine leads to the formation of non-polar interactions with hydrophobic residues. Overall, the specificity in TF-dsDNA interactions heavily depends on the properties of individual amino acids and DNA bases. Recently, a family of paired (PRD)-like TFs has been shown to recognize specific DNA motifs to regulate the expression of target genes that are known to play an important role in early embryonic development (Madissoon et al. 2016; Töhönen et al. 2015). In publication III, we identified functionally critical amino acid residues within the human LEUTX homeodomain involved in DNA binding and showed how a loss-of-function mutation reduces its binding with the DNA motif.

Hydrophobic amino acids are crucial for the immunogenicity of cytotoxic T cell epitopes that bind to the MHC class I molecules that elicit the adaptive immune response (Chowell et al. 2015). Binding of epitopes to the antigen-binding groove of MHC-I is stabilized by both hydrophobic anchoring and hydrogen bond formation (C. Zhang, Anderson, and DeLisi 1998). The efficient presentation of an MHC-I bound epitope to a cytotoxic T cell receptor benefits from a stable and high affinity epitope-MHC-I (eMHC-I) complex (Assarsson et al. 2007; Harndahl et al. 2012).

The identification of the most immunogenic epitopes from pathogens has helped in developing sensitive diagnostic assays, effective therapeutic agents and potential vaccines for preventing various infectious diseases. In December 2019, the emergence of the SARS-CoV-2 virus threw the world into a global pandemic of COVID-19 disease. For our part, we performed computational analyses to identify the most immunogenic SARS-CoV-2-derived epitopes that are conserved in the known variants and identical to the experimentally identified epitopes of SARS-CoV (Publication IV). In addition, we give structural explanation for the binding of SARS-CoV-2 epitopes to MHC-I as well as to the T cell receptor.

In this thesis, the information derived from sequences, structures and biological data were combined with experimental observations from collaborators and from the literature in order to study the structural and functional consequences of molecular interactions. The interdisciplinary approach used in this research provides detailed structural information on the human macromolecular complexes and enabled us to probe their essential roles in human health and disease.

2. Review of the literature

In this thesis, the main focus was to analyze specificity determining features at the interface of biomolecular complexes (*i.e.* protein-protein and protein-DNA complexes) that are involved in various biochemical and cell signaling pathways associated with humans and human health. We analyzed chemical properties and three-dimensional shape of amino acids, solvent accessibility of key amino acids, intermolecular interactions, and position-specific propensities of amino acids and nucleotides, and related these features to the experimentally determined biochemical processes. In the literature section, the focus is to provide detailed information on the biomolecular complexes and their functions related to the three projects.

2.1. Peptidylarginine deiminase and citrullination

2.1.1. Peptidylarginine deiminase

Citrulline was first isolated from watermelon (*Citrullus vulgaris*) by Japanese researchers Koga and colleagues (Koga and Ohtake 1914). It was reisolated by Wada to experimentally determine the empirical formula $C_6H_{13}N_3O_3$, and later named "Citrullin" after the Latin word for watermelon (Wada 1930). Citrulline is considered as a non-essential amino acid, which means that it is synthesized naturally in mammals; for example, the free form of citrulline is produced in the urea cycle to eliminate ammonia from the blood (Krebs and Henseleit 1932), and it is also produced as a byproduct of the nitric oxide pathway (Mori 2007). Since citrulline is also a non-standard amino acid, its occurrence in a synthesized protein has to be the result of post-translational modification. The hydrolytic conversion of a peptidylarginine into a peptidylcitrulline is catalyzed by peptidylarginine deiminase (PAD) enzymes in a calcium-dependent manner (Figure 1) (Fujisaki and Sugawara 1981; Rogers, Harding, and Llewellyn-Smith 1977). The citrullination reaction, also known as hydrolytic deimination, eliminates the positive charge of the arginine residue by replacing the ketamine group (=NH) with a keto group (=O), resulting in biophysical and chemical changes in the substrate protein and therefore function (Tarcsa et al. 1996). The presence of citrulline in proteins was first detected using a color test, which was first described by Fearon in 1939 (Fearon 1939). Later, in 1958, citrulline was identified as a major constituent of the fibrous protein of the inner root sheaths of hair follicles – the first recorded occurrence of citrulline in an animal protein (Roggers and Simmonds 1958).

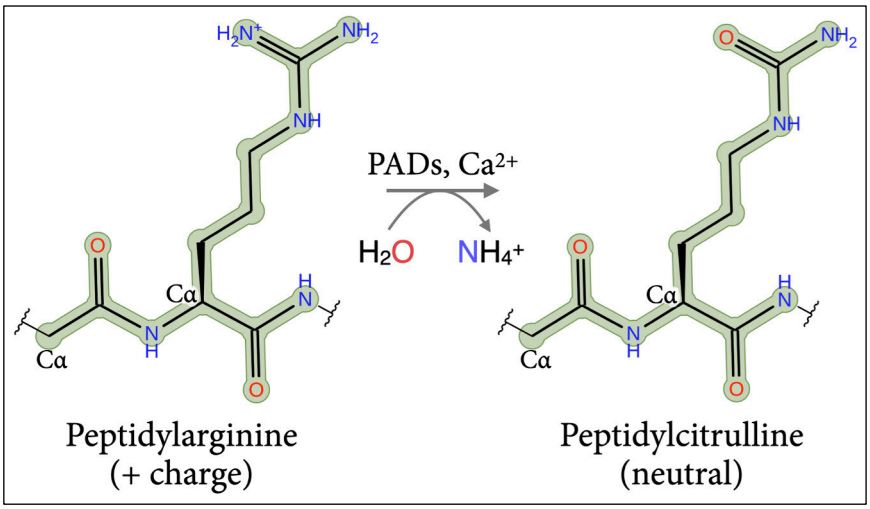


Figure 1: PAD enzymes convert a positively charged peptidylarginine residue into a neutral peptidylcitrulline.

In humans, the PAD family is composed of five isozymes (PAD1, PAD2, PAD3, PAD4 and PAD6), which are encoded by a cluster of *PADI* genes located on the short arm of chromosome 1 (Vossenaar et al. 2004). The PAD5 isozyme is missing because it was originally misannotated, later identified as the ortholog of mouse PAD4 (Mondal and Thompson 2019). The human PAD isozymes share 50% to 55% sequence identity among their protein sequences, and each PAD isozyme exhibits 70% to 95% sequence identity among mammals with a high degree of residue conservation at the catalytic site, indicating that different PAD isozymes catalyze the citrullination reaction via the same mechanism (Arita et al. 2004; Witalison, Thompson, and Hofseth 2015). However, an evolutionarily unrelated PPAD enzyme (UniProt ID: Q9RQJ2) in the prokaryotic species *Porphyromonas gingivalis*, which is known as the most pathogenic bacteria for chronic periodontitis in humans, shows calcium-independent citrullination activity (Abdullah et al. 2013; McGraw et al. 1999).

Among all the human PAD enzymes, only PAD4 possesses a nuclear localization sequence (NLS) ^{56}PPAKKKST^{63}, mediating the transport of PAD4 from the cytoplasm into the nucleus for histone citrullination (Arita et al. 2004). Although the human PAD2 lacks a putative NLS, citrullination of histones and other nuclear proteins suggests an alternative approach for PAD2 translocation into the nucleus (Falcão et al. 2019). Recently, a study conducted on carrier-mediated PAD2 transport revealed that both high levels of calcium ions in the cytoplasm and small GTPase Ran protein in the nucleus were required for PAD2 translocation across the nuclear pore complex, suggesting that the transport of larger proteins, such as the PAD2 monomer (molecular weight ~75 kDa), is highly dependent on the assistance of carrier proteins (Marfori et al. 2011). However, the diffusion of smaller proteins with molecular weight less than ~40 kDa relies on the passive forms of transport (Marfori et al. 2011).

To gain insights into the catalytic mechanism of PAD enzymes, Arita and colleagues determined the first few X-ray crystal structures of calcium-free, calcium-bound and substrate-bound human PAD4, indicating that the PAD4 isozyme contains five highly conserved calcium binding sites (Arita et al. 2004); and a fully occupied PAD4 structure with calcium ions induces conformational changes, resulting in the formation of the catalytic site for substrate binding (Arita et al. 2004). Furthermore, X-ray structures of PAD4 in complex with histone (H3 and H4) peptides showed that the molecular surface near the catalytic site recognizes a flexible peptide with an exposed arginine side chain (Arita et al. 2006). X-ray crystal structures of the human PAD2 enzyme revealed that PAD2 contains a "DXDXDG" calcium binding motif ("X" denotes any amino acid) in addition to the five conserved calcium binding sites (Slade et al. 2015). The canonical calcium binding motif is highly conserved among PAD2 orthologues (Slade et al. 2015), and interestingly, it has also been known to play key roles in a large number of unrelated calcium binding proteins; for example, the repeats of "DXDXDG" motif in the β-propeller domain of the integrin α-subunit are required for its structural integrity and function (Chouhan et al. 2011). Unlike the PAD2-4 enzymes, the structure of human PAD1 is unique because it exists in a monomeric form, and shows selectivity for diverse substrates due to its flexible structure (Saijo et al. 2016). The recently published X-ray crystal structure of human PAD3 revealed structural variability in the N-terminal domain, whereas the C-terminal catalytic domain was observed to have high sequence identity and high structural similarity to other PAD isoforms (Rechiche, Verne Lee, and Shaun Lott 2021). Human PAD6 is known to be involved in embryonic development and female fertility (Taki et al. 2011; Y. Xu et al. 2016). To date, no X-ray structures of PAD6 are available. However, *in vitro* experimental studies on wild-type PAD6 in mouse have revealed that it can form a hexameric structure but with no reported enzymatic activity (Taki et al. 2011), most likely due to the presence of different residues within the catalytic site compared to other PAD isoforms.

2.1.2. Normal functional role of PAD enzymes

PAD enzymes have a wide range of tissue distribution and play essential roles in many important physiological processes by catalyzing the citrullination of specific substrate proteins (Figure 2).

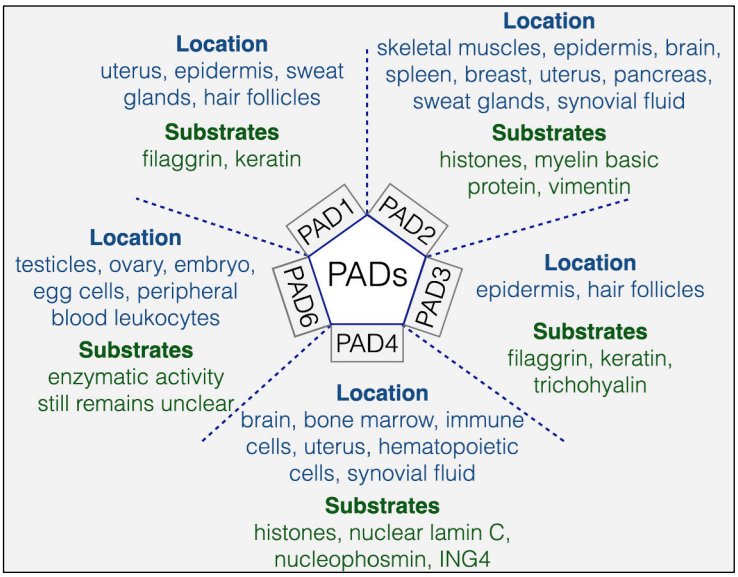


Figure 2: Target substrates and tissue-specific distribution of PADs in humans.

In humans, both PAD1 and PAD3 are mainly expressed in epidermis and hair follicles (Nachat et al. 2005; Terakawa, Takahara, and Sugawara 1991; Tsuchida et al. 1993). In the epidermis, PAD1 and PAD3 mediated citrullination of keratin and filaggrin proteins plays an important role in maintaining skin hydration and pH by stimulating filaggrin-derived natural moisturizing factors (Ishida-Yamamoto et al. 2002; Kabashima 2013). In addition, PAD3 has also been found to citrullinate trichohyalin, apoptosis-inducing factor and vimentin proteins (Witalison, Paul R Thompson, and Hofseth 2015). Citrullination of trichohyalin induces structural changes in the protein, which makes it more soluble and efficient for cross-linking with keratin filaments to provide mechanical strength to the inner root sheath of the hair follicle (Tarcsa et al. 1997). The expression of PAD1 has also been observed during the early stages of embryonic development where it catalyzes citrullination of the histone tail region, which promotes chromatin relaxation by disrupting the electrostatic interactions between histone proteins and DNA (X. Zhang et al. 2016). However, inhibition of PAD1 significantly reduced the citrullination activity in mouse embryonic cells and showed a steep drop in the transcriptional activity, causing early embryonic arrest at the 4-cell stage (X. Zhang et al. 2016).

The *PADI2* gene is widely expressed in multiple tissues, including spleen, brain, skeletal muscle, epidermis, breast, uterus, pancreas, sweat glands and synovial fluid (S. Wang and Wang 2013). PAD2-mediated citrullination of histones induces oligodendrocyte differentiation, which initiates the formation of myelin sheath around axons in the central nervous system for efficient motor function (Falcão et al. 2019). In comparison with the healthy population, patients with severe multiple sclerosis have shown a significant rise in the citrullinated myelin basic protein (MBP), which is a key component of the myelin sheath.

The expression of *PADI4* gene is detected in brain, bone marrow, uterus, breast, ovary, synovial fluid and immune cells (S. Wang and Wang 2013). When bacteria invade the human body, the innate immune system induces PAD4 mediated hypercitrullination of histones in neutrophils, initiating the process of NETosis in which the neutrophils capture and kill pathogens by releasing highly decondensed chromatin called neutrophil extracellular traps (NETs) (Yanming Wang et al. 2009). However, excessive NET formation has been observed to cause vascular endothelial injury and thrombosis (Fuchs et al. 2010). Studies on chromatin regulation have revealed that PAD4 mediated citrullination of histones promote localized chromatin decondensation in embryonic stem cells, which promotes transcriptional activation of the pluripotency inducing genes (Christophorou et al. 2014; Slade et al. 2014). Furthermore, inhibition of PAD4 in pig embryos showed embryonic arrest at the early stages of development, suggesting that the PAD4 activity plays an important role in embryonic development (Brahmajosyula and Miyake 2013).

In response to DNA damage, cells in a multicellular organism activate the DNA repair mechanism to maintain the integrity of the inherited genome, but when the mechanism fails, cells undergo the process of programmed cell death known as apoptosis. A nuclear transcription factor p53, which acts as a guardian of the genome, activates its target genes to induce apoptosis; interestingly, the *PADI4* gene was also detected as a direct transcriptional target of p53 (Tanikawa et al. 2009). The p53 mediated upregulation of PAD4 causes citrullination of various proteins, whereas, the knockdown of either p53 or PAD4 remarkably reduced the ability of proteins to citrullinate, suggesting that the citrullination activity is regulated in a p53/PAD4-dependent manner (Tanikawa et al. 2009). In addition, PAD4 mediated citrullination of nucleophosmin (NPM1), a ubiquitously expressed chaperone protein participating in histone assembly and ribosome biogenesis, initiates the apoptotic pathway by disrupting the molecular interactions between the citrullinated NPM1 and histone proteins (Tanikawa et al. 2009).

The uncitrullinated form of inhibitor of growth 4 (ING4) protein regulates the expression of genes that are involved in the p53-dependent apoptosis pathway (Q. Guo and Fast 2011). However, the PAD4 mediated citrullination of ING4 has been observed to suppress the transcription of apoptosis inducing genes by dissociating the ING4-p53 complex (Q. Guo and Fast 2011). The p53-PAD4 pathway induces the citrullination of the lamin C protein for the regulation of chromatin structure and apoptosis in response to DNA damage (Tanikawa et al. 2012). Overall, these observations support the premise that PAD4 regulates the critical steps of apoptotic cell death via the citrullination of histones, nucleophosmin and lamin C proteins.

The expression of the *PADI6* gene has been detected in the embryo, ovary and eggs (Chavanas et al. 2004; Wright et al. 2003). To date, PAD6 is the only isozyme for which no protein substrates have been detected, and, *in vitro* experiments show a lack of citrullination activity (Taki et al. 2011). In mouse, PAD6 is observed to associate with the egg cytoskeletal sheet, which is a fibrous network

of keratin-associated intermediate filaments and arises during oocyte development (Wright et al. 2003).

2.1.3. Regulation of PAD activity

PADs are involved in a wide range of vital cellular processes; therefore, it is of extreme importance that PAD enzymes balance their citrullination activity. Due to this delicate balance, it is not difficult to believe that there are numerous inflammatory and degenerative brain diseases associated with dysregulated PAD enzyme activity, promoting the abnormal citrullination of non-physiological substrates (Ishigami et al. 2005; Moscarello, Mastronardi, and Wood 2007). The exact cause of PAD dysregulation is still unclear, but experimental evidence suggests that the abnormal citrullination activity is mainly controlled by the cellular states and environmental conditions, some of which are mentioned below.

PAD activity is dysregulated at extremely high calcium concentrations displaying off-target activities, causing denaturation and loss of function in arginine-rich proteins. *In vitro* studies have demonstrated that PAD enzymes, especially PAD2 and PAD4, require millimolar concentrations of calcium for full occupancy of all the experimentally identified ion binding sites, inducing structural rearrangements and dimerization of the enzyme to achieve the maximal citrullination activity (Slade et al. 2015; Y. Zhou, Mittereder, and Sims 2018). However, under *in vivo* conditions, the cytosolic and nuclear calcium levels are maintained at low nanomolar concentrations (Slade et al. 2015; Y. Zhou, Mittereder, and Sims 2018), which raises the question – how is the functional activity of the PAD enzyme maintained inside the cell? Citrullination in humans suggests the possible involvement of another intracellular PAD-binding protein that could act as a key regulator of PAD activation even at lower calcium concentrations; for example, antibodies (anti-PAD3/4) present in the synovial fluid of rheumatoid arthritis patients have been observed to lower the calcium threshold to micromolar concentrations for PAD activity (Darrah et al. 2013; Shi et al. 2018; Zendman et al. 2007).

Another important regulatory component of PAD activity is the existence of a reducing environment, which induces the nucleophilic attack by the active site cysteine residue (C647 in PAD2 and C645 in PAD4) on the guanidinium group of the substrate arginine (Damgaard et al. 2016). In contrast to the reducing environment of the cytosol, the oxidizing nature of the extracellular environment provides protection against abnormal citrullination by PADs that may have been released from necrotic cells (Darrah and Andrade 2018).

Ubiquitously expressed transcription factors play an important role in regulating basal expression levels of *PADI* genes (Chavanas et al. 2008; S. Dong et al. 2005; S. Dong, Zhang, and Takahara 2007). For example, the sex steroid hormone 17β-estradiol (E2) has been shown to markedly increase the expression levels of the transcription factors Sp1, AP-1 and NF-Y, all of which bind to the promoter of the *PADI4* gene and contribute to regulation of gene expression in a cooperative manner (S. Dong, Zhang, and Takahara 2007). In

neutrophils, E2 mediated activation of PAD4 promotes histone citrullination and triggers the formation of NETs during pregnancy to protect both mother and fetus from pathogenic infection (Giaglis et al. 2016; Yasuda et al. 2019). However, binding of the NF-κB transcription factor to the promoter of *PADI4* resulted in reduced expression of the PAD4 encoding gene in neutrophils, suggesting that NF-κB contributes to the "resolution phase" of inflammation for restoring tissue homeostasis (Abbas et al. 2014).

Autocitrullination of PAD enzymes has been reported as a self-regulatory mechanism to control PAD activity; for example, the autocitrullinated PAD1, PAD2 and PAD3 enzymes significantly reduced the citrullination activity on filaggrin protein (Méchin et al. 2010). In addition, autocitrullination strikingly modifies the PAD4 structure to the extent that it cannot be recognized by the anti-PAD4 antibody that specifically interacts with wild-type PAD4. Indeed, the structural changes in autocitrullinated PAD4 result in a complete inactivation of citrullination activity (Andrade et al. 2010).

2.1.4. The PAD4 structure

Human PAD4 consists of 663 residues, of which the first 300 residues fold into two N-terminal immunoglobulin (IgG)-like domains: IgG1 and IgG2, and the remaining 363 residues fold into a C-terminal catalytic domain (Figure 3A). The IgG1 and IgG2 domains are composed of 19 β-strands and 4 short α-helices, whereas the C-terminal domain is composed of five circularly arranged ββαβ modules that make a pseudo 5-fold symmetric structure called an α/β propeller (Arita et al. 2004). PAD4 contains five calcium binding sites designated Ca1-Ca5 (Figure 3B). The N-terminal IgG1 domain exhibits high local flexibility compared with the C-terminal domain; the C-terminal domain contains a catalytic site (Figure 3B) and displays significant backbone similarity in PAD isotypes (Figure 3C). In the presence of high calcium concentrations (5 to 10 mM), especially in the synovial fluid of rheumatoid arthritis patients, all of the Ca1-Ca5 positions in PAD4 are occupied by calcium ions, promoting extensive structural rearrangement to form the active site and leading to the dimerization of the enzyme (G. Y. Liu et al. 2017). Interestingly, the number of calcium binding sites in PADs are not identical, varying from four (PAD1) to six (PAD2) depending on the specific isoform.

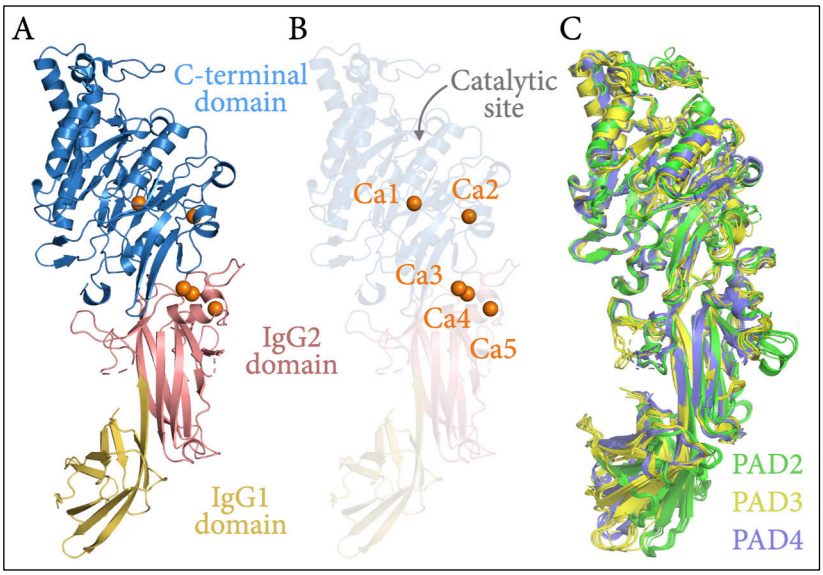


Figure 3: Structure of the human PAD enzyme. A) Cartoon representation of PAD4 (PDB ID: 2DEW). B) The C-terminal domain (blue) possesses a catalytic site and two calcium ions (Ca1 and Ca2) for PAD citrullination activity, while the other three calcium ions (Ca3-Ca5) are located in IgG2 (pink). C) Cartoon representation of the superposed structures of the human PAD2 (green; PDB IDs: 4N20, 4N22, 4N24, 4N25, 4N26, 4N28, 4N2A, 4N2B, 4N2C, 4N2D, 4N2E, 4N2F, 4N2G, 4N2H, 4N2I, 4N2K, 4N2L, 4N2M, 4N2N and 5HP5), PAD3 (yellow; PDB IDs: 6CE1, 7D4Y, 7D56, 7D5R, 7D5V, 7D8N and 7DAN) and PAD4 (indigo; PDB IDs: 1WD8, 1WD9, 1WDA, 2DEW, 2DEX, 2DEY, 2DW5, 3APM, 3APN, 3B1T, 3B1U, 4DKT and 4X8C), showing that the N-terminal IgG1 domain is more flexible than both the IgG2 and C-terminal domains.

While it is unclear whether all PAD enzymes are capable of forming a stabile homodimer, disruption of the dimer interface in PAD4 decreases the catalytic activity up to 75%, suggesting that the monomeric PAD4 has still some activity (Y. L. Liu et al. 2011). Structural studies of PADs have revealed that PAD2, PAD3 and PAD4 form a head-to-tail homodimer, which is mainly mediated by both hydrophobic and ionic interactions at the dimer interface (Arita et al. 2004; Rechiche, Verne Lee, and Shaun Lott 2021; Saijo et al. 2016; Slade et al. 2015), whereas in PAD1 the distinctive elongated N-terminal tail could make intermolecular steric clashes and most likely prevents PAD1 from forming a homodimer (Saijo et al. 2016). More specifically, studies on X-ray crystal structures of PAD4 have revealed that residues R8 and R544 respectively form ionic interactions with D547 and D273, and the mutation of arginine to glutamate at position 8 causes the dimeric enzyme to dissociate into monomers, suggesting that the ionic interactions between R8 and D547 at the interface are important for dimer pairing (Y. L. Liu et al. 2011). However, the double mutation R544A and D273A in PAD4 does not cause the enzyme to dissociate into monomers (Y. L. Liu et al. 2011).

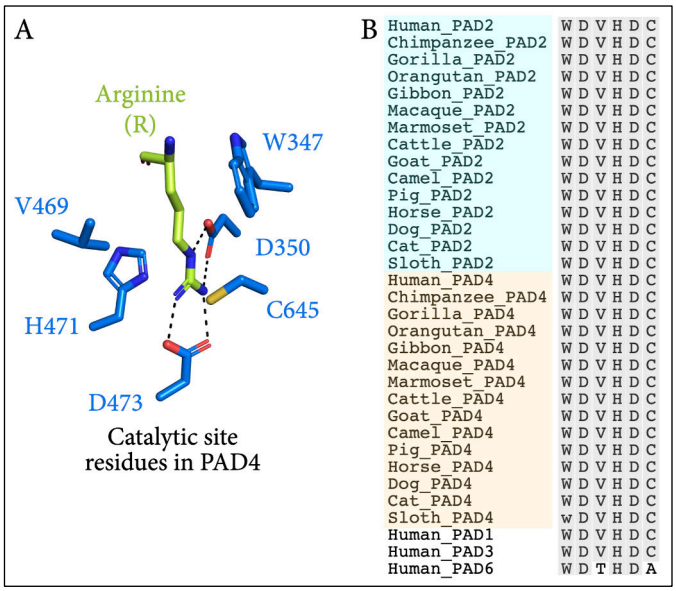


Figure 4: Residues at the catalytic site of PAD enzymes. A) Four key catalytic residues (D350, H471, D473 and C645 in PAD4) play essential roles in citrullination of the substrate arginine residue. D350 and D473 make hydrogen bonds (black dots) with the guanidinium group of arginine, while residues C645 (forms the covalent intermediate) and H471 promote catalysis via nucleophilic attack by C645 on the guanidinium carbon of arginine. Hydrophobic residues W347 and V469 provide stability to the aliphatic side chain of the substrate arginine. B) Evolutionary conservation of the catalytic site residues in placental mammalian species. Mutations in the human PAD6 provide support for a previous proposal that PAD6 is the only enzyme that lacks citrullination activity due to the missing catalytic cysteine.

Unlike other PADs, the expression of PAD2 and PAD4 has also been detected in rheumatoid synovium and synovial fluid cells (Foulquier et al. 2007; Vossenaar et al. 2004). Structurally, the catalytic site of the PAD4 enzyme is composed of four residues: H471, C645, D350 and D473 (Arita et al. 2006), forming a closely interacting catalytic tetrad for the citrullination activity. The substrate arginine residue has been observed to form “side on” hydrogen bonds with D350 and “end on” hydrogen bonds with D473 (Figure 4A), securing the position of the guanidinium group of the substrate arginine for nucleophilic attack carried out by the thiolate group of C645. Two other buried residues, W347 and V469, complete the catalytic pocket by providing hydrophobic stability to the 3-carbon aliphatic chain of the substrate arginine residue (Figure 4A). Indeed, mutation of W347 and V469 results in impaired citrullination activity of the PAD4 enzyme (C. Y. Lee et al. 2017). Although human PAD2 and PAD4 share about 50% sequence identity, the identical catalytic site residues present among distantly related placental mammals suggest the possibility of a strong connection between the enzymes´ evolution and functional constraints (Figure 4B).

Structural analysis of the PAD4 apoenzyme (PDB ID: 3APN (Horikoshi et al. 2011)) has demonstrated that the deeply buried active site is exposed to the solvent through a wide tunnel (Figure 5). However, the active site in calcium-bound PAD4 (PDB ID: 2DEW (Arita et al. 2006)) is characterized by its "U"-shaped tunnel containing two entrance doors in which the "front door" is considered as the substrate binding site accommodating the arginine side chain, whereas the "back door" provides access to the solvent through a narrow polar tunnel (Figure 5) (Fuhrmann, Clancy, and Thompson 2015; Teo et al. 2012).

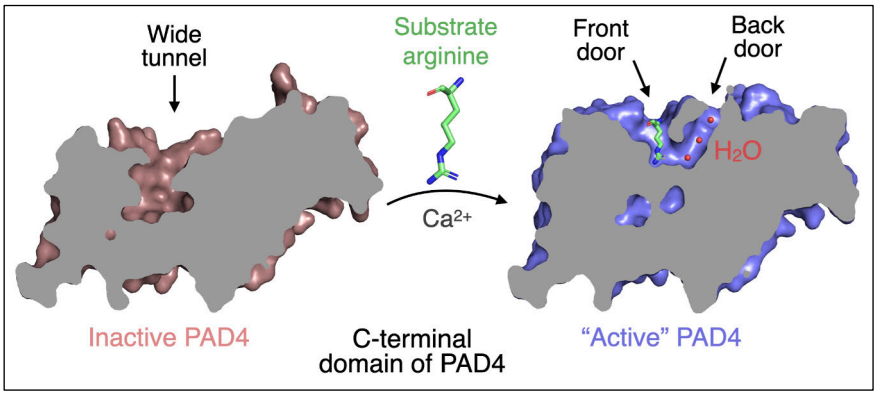


Figure 5: Sliced surface representation of the C-terminal catalytic domain in PAD4. The inactive form of PAD4 (brown; PDB ID: 3APN) possesses a single wide tunnel, whereas the arginine/calcium-bound form of PAD4 (blue; PDB ID: 2DEW) possesses a "U"-shaped tunnel labeled as "front door" and "back door". The substrate binding "front door" accommodates the side chain of arginine (green), while the "back door" tunnel provides access to water molecules (red). To enable crystallization of PAD4 in the arginine-bound form ("Active" PAD4), the catalytic cysteine of wild-type PAD4 was mutated to alanine.

2.1.5. Link between PADs and inflammation

Rheumatoid arthritis (RA) is an inflammatory disease of the joints that is thought to be intensified by PAD dysregulation. While the cause of RA is still unknown, certain factors that may increase the risk include: age, gender, genetics, smoking, obesity, physical stress and psychological disorders (Daïen and Sellam 2015; Rasch et al. 2003; Saag et al. 1997; Sleath et al. 2008; Somers et al. 2013). In RA, the immune cells misidentify the body's own proteins as harmful foreign invaders and attack them by releasing autoantibodies. Autoantibodies generated against the citrullinated proteins attack the tissue lining of joints, which in healthy conditions is known to secrete a transparent and viscous fluid for smooth joint movements (A. Hui et al. 2012). However, after onset of RA, the features of synovial fluid commonly shift to be turbid and less viscous, resulting in painful and swollen joints (A. Hui et al. 2012).

Among the five PAD enzymes, the expression levels of PAD2 and PAD4 are closely linked with the severity of disease in RA patients (Foulquier et al. 2007). Expression of *PADI2* and *PADI4* genes has been reported in a variety of immune

cells, including neutrophils, dendritic cells, mast cells, monocytes and macrophages (Foulquier et al. 2007). It has been shown that the mRNA transcripts of *PADI2* and *PADI4* were predominantly found in the peripheral blood monocytes of RA patients (Vossenaar et al. 2004). Interestingly, the PAD4 protein levels are more conserved than the mRNA levels in both monocytes and their differentiated macrophage form in which the PAD4 transcripts were hardly detected, suggesting that the PAD4 enzyme does not degrade during the maturation of monocytes into macrophages in RA patients (Vossenaar et al. 2004). However, despite the unchanged transcript levels of PAD2 in both monocytes and macrophages, high levels of the PAD2 enzyme in macrophages suggest that the transcripts are not translated into PAD2 proteins until the initiation of differentiation (Vossenaar et al. 2004).

While the exact cause of uncontrolled hypercitrullination is not known, studies have identified two independent immunological pathways: perforin and membrane attack complex (MAC), initiating the formation of pores in the plasma membrane of cells rich in PADs. A rapid entry of calcium (μM range) into the cytoplasmic compartment causes the activation of PADs (Lopez et al. 2013; Romero et al. 2013; Triantafilou et al. 2013). These activated PADs citrullinate intracellular proteins in the cytoplasm (*e.g.* vimentin, actin, aldolase, histones, α-enolase and calreticulin) as well as extracellular proteins (*e.g.* fibrinogen, fibronectin, complement C3, antithrombin-III, apolipoprotein and gelsolin) when PAD enzymes leak out of the cells into the extracellular matrix (ECM) (Van Beers et al. 2013; Darrah and Andrade 2018). Recent studies indicate that along with inflamed joints, the expression levels of *PADI2* and *PADI4* genes have also been found significantly higher in tissue samples from patients with lung carcinoma, breast carcinoma, hepatocellular carcinoma, colon adenocarcinoma and esophageal cancer (Chang et al. 2009; Cherrington et al. 2012; W. Guo et al. 2017; Yufeng Wang et al. 2021).

2.1.6. Arginine mediated interactions in extracellular proteins

The extracellular matrix is a dynamic network of tissue-specific proteins and polysaccharides that plays an essential role in regulation of fundamental cellular processes, such as cell adhesion, migration and differentiation (Frantz, Stewart, and Weaver 2010). In mammals, collagens are the most abundant extracellular proteins (~30% of the total protein mass) providing tensile strength to bone, tendons and ligaments (Kwansa, De Vita, and Freeman 2014). Interestingly, citrullination of collagen has been observed to decrease the adhesion of fibroblasts to the membrane-bound receptors integrin α10β1 and α11β1 (Sipilä et al. 2014), suggesting that citrullination of collagens dysregulates integrin-mediated cellular events, such as cell migration and proliferation.

Fibronectin (FN) is a large extracellular glycoprotein composed of two nearly identical covalently linked monomers (K. J. Johnson et al. 1999). FN is primarily involved in cell-adhesive interactions via integrins that regulate various cellular functions including cell migration, cell differentiation, phagocytosis, cytoskeletal organization, and ECM homeostasis (Blystone et al. 1994; Straus et al. 1989; J. T.

Yang et al. 1999). FN and fibrillin-1, along with other ECM proteins, provide an initial scaffold for the storage of the large latent complex of TGF-β1 in which the latent TGF-β binding protein 1 (LTBP1) is covalently linked with the latency-associated peptide (LAP) (Figure 6A). The LAP contains an integrin-binding "RGD" motif that has already been observed to form molecular interactions with several different integrins, such as αVβ3, αVβ5 and αVβ6 (Wipff et al. 2007) (Figure 6A). Since the ECM is a highly dynamic structure, mechanical stresses produced during cell contraction break the non-covalent interactions between TGF-β1 and LAP (Figure 6B), resulting in conformational changes and release of activated TGF-β1 dimer (Maeda et al. 2011). The binding of TGF-β1 with its heterodimeric transmembrane receptor, which is composed of the TGFβRI and TGFβRII subunits, initiates various signaling pathways leading to the regulation of cell proliferation, migration, differentiation and apoptosis (Vander Ark, Cao, and Li 2018). It is interesting to note that both TGF-β1 and TGF-β3 form electrostatic interactions with the TGFβRII subunit via two arginine residues, whereas the electrostatic interactions in TGF-β2 are mediated via two lysine residues (Radaev et al. 2010).

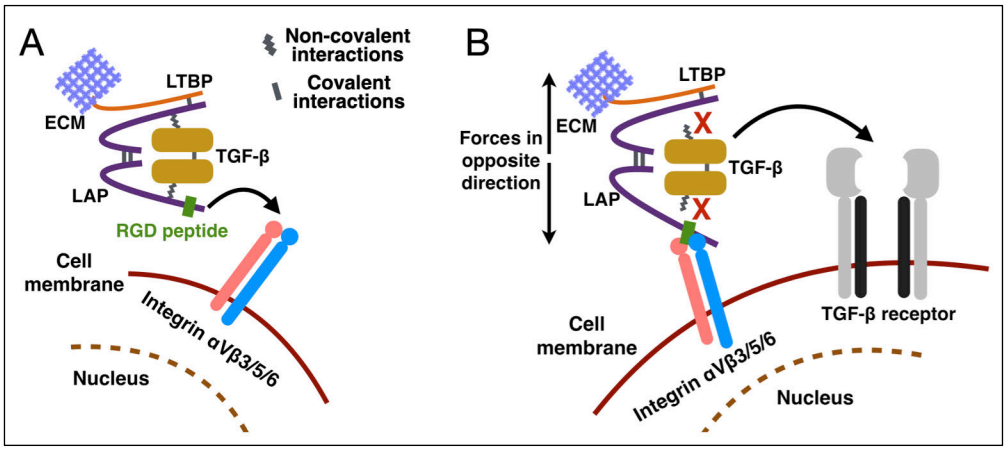


Figure 6: Integrin-mediated activation of TGF-β growth factor. A) The latency-associated peptide (LAP) of TGF-β contains a canonical integrin-binding RGD motif. Other subunit of the LAP homodimer makes covalent interactions with the latent TGF-β binding protein (LTBP), which interacts with various fibrous proteins in the extracellular matrix (ECM). B) Non-covalent interactions between LAP and homodimeric TGF-β break due to the dynamic behavior of the extracellular proteins. The released TGF-β binds to a heterodimeric TGF-β transmembrane receptor in order to initiate various intracellular signaling pathways.

FN is recognized by the α5β1, αIIbβ3, αVβ1, αVβ3, αVβ5 and αVβ6 integrins via the RGD motif located in the tenth type III module (FN-III10), while the FN-III9 module located in close proximity to the RGD motif contains an additional so-called "synergy site" PHSRN that increases the binding affinity of RGD to the

α5β1 and αIIbβ3 integrins (Bowditch et al. 1994). Another well-known integrin binding motif, NGR, is located in the FN-I5 module (Curnis et al. 2006). A spontaneous deamidation of asparagine in the NGR motif generates a new gain-of-function motif isoDGR whose binding to integrin αVβ3 has been shown to regulate endothelial cell adhesion, proliferation and tumor growth (Curnis et al. 2006). Interestingly, integrin αVβ3 showed 2-fold higher binding affinity with both the cyclic RGD (CRGDC) and isoDGR (CisoDGRC) motifs than the cyclic DGR (CDGRC) and NGR (CNGRC) motifs (Curnis et al. 2006).

Our study of synovial fluid in RA patients has identified 24 extracellular proteins, each consisting of at least one citrullinated arginine. Our 3D structural analysis of the citrullinated proteins suggests that the conversion of arginine to citrulline alters the charge of the protein and hydrogen bonding capability that most likely would lead to a weakening of the wild-type protein-protein interactions.

2.2. PRD-like family of transcription factors in human embryonic genome activation

2.2.1. Early stages of the human embryo

The development of a multicellular diploid eukaryotic embryo starts from a single-cell fertilized egg, also known as the zygote, resulting from the fusion of the haploid male (spermatozoon) and female (oocyte) gametes. Out of the hundreds of million ejaculated spermatozoa, only 1% successfully get access to the uterus. Under appropriate conditions, spermatozoa can survive up to 5 days, though only a few thousand swim to reach the uterine (fallopian) tubes (Clubb 1986). Inside the uterus, prostaglandin and oxytocin hormones induce muscular contractions to facilitate the transport of sperm towards the fallopian tube (Mastroianni 1999). In addition, a continuous supply of chemoattractants from the ovulated oocyte and its surrounding cumulus cells generate a gradient of chemicals, such as hormones, peptides and cyclic nucleotides, to attract spermatozoa for oocyte fertilization (Ralt et al. 1991). The chemical gradient is maintained for as long as the oocyte survives (~24 hours in human after ovulation) (Harper 1982). Similarly, contractions in the infundibulum region of the fallopian tube facilitate the movement of the oocyte towards the ampullary region where fertilization usually takes place (Eddy and Pauerstein 1980) (Figure 7A). During the fertilization event, the binding of spermatozoa to the zona pellucida membrane of the oocyte initiates degradation of the thick extracellular egg coat, enabling the spermatozoa to burrow into the layer (Baibakov et al. 2007; McLeskey et al. 1997). The first spermatozoon to fuse with the oocyte plasma membrane triggers the egg to complete the second meiotic cell division, resulting in the formation of two haploid male and female pronuclei (Clift and Schuh 2014). Subsequently, the fusion of pronuclei causes activation of various serine proteases, which have been observed to break molecular

interactions between the vitelline envelope and the plasma membrane of the fertilized egg for blocking polyspermy (Haley and Wessel 1999).

After fertilization, the zygote undergoes the early stages of embryonic development, including cleavage, blastula formation, implantation, gastrulation, neurulation and organogenesis. The early embryonic cleavages are rapid mitotic divisions whereby the big zygote divides into a number of smaller cells known as blastomeres (Wolpert et al. 2001). Unlike the mitotic division that takes place for cell proliferation and growth, the process of cleavage does not allow an increase in cell mass and consists of mainly three phages: 1) DNA replication, 2) mitosis, and 3) cell division (Wolpert et al. 2001). After the onset of cleavage, the following changes take place in the human embryo over time: 2-cell stage (~1 day), 4-cell stage (~2 days), 8-cell stage (~3 days), and finally 16-cell stage or morula (~4 days) (Boroviak and Nichols 2014). Furthermore, approximately on the fifth day, the morula continues to undergo mitotic divisions to form the blastocyst, in which cell differentiation and cavitation occur simultaneously to form a fluid filled cavity or blastocoel (Erb 2006). The blastocyst contains mainly two types of cells: pluripotent cells attached at one end comprise the "inner cell mass" from which the embryo develops, and the "trophectoderm" cells, which develop the placenta (Rossant 2001). As the volume of the blastocoel fluid increases, it exerts hydrostatic pressure to break the zona pellucida membrane and eventually lets the blastocyst attach to the endometrial lining near the uterine fundus (Figure 7A). This process is known as hatching and occurs about 9 days after ovulation (Sathananthan, Menezes, and Gunasheela 2003; Wilcox, Baird, and Weinberg 1999). Since the blastula in placental mammals lacks a yolk, the embryo exchanges gases and other substances with the mother's blood through the placenta, which connects the embryo to the endometrium (Gude et al. 2004).

Although the zygote receives an equal amount of genetic material from both parents, the cytoplasmic components of the zygote originate almost entirely from the oocyte (R. Li and Albertini 2013). These components include the oocyte cytoplasm, subcellular organelles, and special maternal factors, comprised of calcium ions, proteins, messenger RNAs and small non-coding RNAs (L. Li, Lu, and Dean 2013; Schier 2007). After fertilization, the proteins encoded by the maternal mRNAs regulate various processes, including maternal mRNA clearance, epigenetic modifications and the onset of embryonic genome activation (Hamm and Harrison 2018). The early stages of embryonic development are highly dependent on the process of maternal mRNA clearance. Dysregulation of this process has been observed to cause various disorders, such as early embryo arrest, defects in oocyte maturation, infertility, oocyte aging and follicle growth retardation (Sha et al. 2020). The human embryo eliminates the maternal transcripts through the action of two mRNA degradation pathways: 1) M-decay, which begins prior to the fertilization event and degrades the short-lived maternal transcripts, and 2) Z-decay, which initiates at the zygote to 2-cell stage transition and ends at the late 2-cell stage (Vastenhouw, Cao, and Lipshitz 2019) (Figure 7B). A key developmental transition takes place when the early

embryo degrades its maternal transcripts and initiates transcription of its own genome; this transition is referred to as embryonic genome activation (EGA). In human, EGA occurs gradually through two transcriptional waves: a minor wave begins at the late 2-cell stage and a major wave initiates during the 4-cell to 8-cell transition (Niakan et al. 2012) (Figure 7B) and, indeed, a dysregulated EGA is known to contribute to preimplantation pregnancy failure of an abnormal embryo (M. T. Lee, Bonneau, and Giraldez 2014). The process of EGA predominantly relies on a group of proteins known as transcription factors (TFs), allowing eukaryotes to modulate cell behavior and differentiate by switching specific genes on (expressed) or off (repressed) at a particular stage of development (Spitz and Furlong 2012).

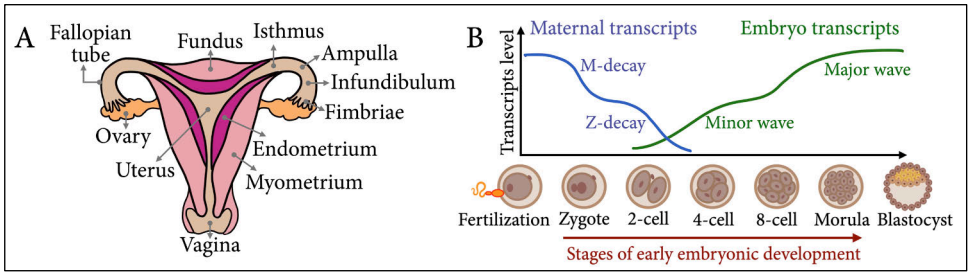


Figure 7: Embryonic development in humans. A) Schematic representation of the female reproductive tract that depicts the major sites involved in human reproduction. Fertilization of an egg by a sperm usually takes place in the fallopian tube. In implantation, the blastocyst attaches to the uterine endometrium. Figure modified from PMG Biology website. B) Degradation of maternal transcripts is essential for activation of embryonic genome activation (EGA). The short-lived maternal transcripts (M-decay) degrade prior to the fertilization event, whereas the long-lived maternal transcripts start degrading (Z-decay) at the zygote to late 2-cell stage transition phase. The minor wave of EGA starts at the 2-cell stage, while the major wave of EGA initiates at the 4-cell to 8-cell transition phase.

TFs are nuclear proteins that bind to specific DNA sites in the promoter and/or enhancer regions for the precise regulation of gene expression (Todeschini, Georges, and Veitia 2014); for example, the early embryonic TFs promote cell differentiation for the conversion of naïve cells into specific functional cell types that eventually lead to the establishment of the body plan and embryonic patterning (Rankin et al. 2000; Torlopp et al. 2014). In eukaryotes, the TFs fall into two main groups: those that are required for regulation of a large number of genes, and which are found in a broad range of cell types; and those that are required for a specific set of genes whose expression is restricted to certain types of cells and stages of development (Wolpert et al. 2001). For instance, the early embryonic development in humans is predominantly regulated by a special class of TFs known as paired (PRD)-like homeodomains (Jouhilahti et al. 2016; Töhönen et al. 2015). The PRD-like LEUTX homeodomain, which acts as a transcriptional activator, has been found to be

activated in the 4-cell stage human embryo, whereas the PRD-like DPRX homeodomain, which acts as a transcriptional repressor, has been found to be activated in the 8-cell stage human embryo (Jouhilahti et al. 2016). Interestingly, both LEUTX and DPRX recognize the same DNA motif, and therefore regulate similar target genes at different stages of embryonic development (Jouhilahti et al. 2016).

2.2.2. Homeobox genes and homeodomains

In 1983, the lab of Walter J. Gehring independently identified an important gene *Antp* (Antennapedia) in Drosophila using the chromosomal walking technique (Garber, Kuroiwa, and Gehring 1983). Soon after that the Gehring lab discovered more homeotic or Hox genes that share sequence similarities in genomic DNA from arthropods through to human (McGinnis et al. 1984). Later, the genes containing such a highly conserved sequence were named "homeobox genes". The homeobox genes have been observed to control the body plan of a developing embryo along the anteroposterior axis (McGinnis and Krumlauf 1992). Structurally, the TFs encoded by the homeobox genes contain three α-helices that fold into a DNA binding domain of ~60 amino acids (Figure 8), known as the homeodomain (HD) (Kissinger et al., 1990).

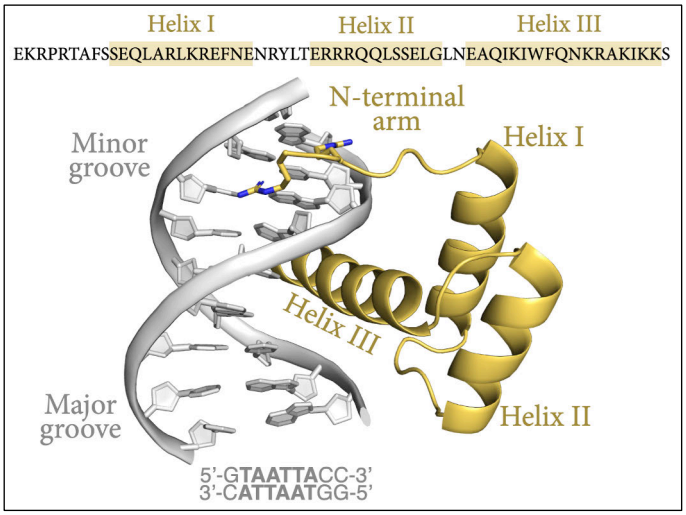


Figure 8: Three-dimensional structure of the engrailed homeodomain with double-stranded DNA motif (PDB ID: 1HDD; Kissinger et al., 1990). This is the first reported X-ray structure of a homeodomain protein. Positively charged residues of the N-terminal arm interact with the DNA minor groove, whereas residues of helix III (the recognition helix), make contact with the major groove of DNA.

Venter and colleagues analyzed the human genome using automated annotation tools and identified 160 homeobox genes (Venter et al. 2001). Since then, the development of advanced genomic tools and a more precise human

genome sequence has improved the study of homeobox gene prediction; for example, Nam and Nei identified 230 homeobox genes (Nam and Nei 2005). More recently, Holland and colleagues have analyzed the latest version of the human genome and identified 300 homeobox loci in the euchromatic region of the human genome, of which 235 loci encode functional genes while the remaining loci encode pseudogenes (Holland, Booth, and Bruford 2007). However, the total number of homeobox sequences are higher than 300 due to the fact that several genes and pseudogenes possess multiple homeobox sequences (Holland, Booth, and Bruford 2007). For example, the PRD-like human DUX4 encoding gene resides within each D4Z4 repeat located in the subtelomeric region of chromosome 4 (4q35 locus) (Schaap et al. 2013). This region consists of from 11 to more than 100 copies of the 3.3 kb D4Z4 repeat units in healthy individuals, whereas individuals suffering with facioscapulohumeral muscular dystrophy (FSHD) type 1 possess less than 10 repeats (Statland et al. 2015). Despite having so many copies of the DUX4 gene, DNA methylation of CpG islands located in the D4Z4 repeats silences the expression of the DUX4 gene in most adult tissues (Cascella et al. 2018). However, at the early pre-implantation stage in human embryos, our recent study has demonstrated that DUX4 is transiently expressed where it functions to alter chromatin accessibility, suggesting that DUX4 regulates the expression of "pioneer" transcription factors during early embryonic development (Vuoristo et al. 2022).

The HD binds DNA directly in a sequence specific manner to regulate the transcription levels of their target genes. The degree of binding affinity is enhanced when a multi-domain DNA binding protein interacts extensively with a specific DNA sequence (Pal and Levy 2020). Most of the human homeobox genes encode only one HD (Nam and Nei 2005), but some homeobox genes such as DUX4 (two HDs), ZFHX4 (four HDs) and ZHX1 (five HDs) contain more than one HD, suggesting that a multi-homeodomain homeobox protein may specifically recognize a longer DNA sequence motif (Pal and Levy 2020). Besides the conserved HD, many homeobox proteins contain additional protein domains, some of which are known to improve the DNA binding specificity either by directly interacting with a specific DNA motif or by interacting with other transcriptional regulatory proteins (Bürglin and Affolter 2016; Holland, Booth, and Bruford 2007). The human homeobox proteins are unambiguously classified into eleven distinct classes, namely ANTP, PRD, LIM, SINE, HNF, POU, ZF, CUT, PROS, CERS and TALE, based on the associated protein domains and evolutionary relationships (Holland, Booth, and Bruford 2007). In humans, the ANTP is the largest class, comprising about 39% of the homeobox genes (Holland, Booth, and Bruford 2007). For comparison, the second largest class, PRD, is comprised of about 24% of the human homeobox genes, whereas the PROS class contains the least number of homeobox genes (Holland, Booth, and Bruford 2007). Interestingly, homeobox genes belonging to the two largest classes have been observed to play pivotal roles in early patterning of the body during metazoan development (Monteiro et al. 2006). Recently, Bürglin and

Affolter presented a revised classification scheme of the animal homeobox genes in which the PRD class is considered separate from the PRD-like class (Bürglin and Affolter 2016).

2.2.3. PAIRED (PRD) and PRD-like family of transcription factors

The PRD class homeobox genes encode additional DNA binding domains other than the HD, for example the PAIRED domain, which is about twice the size of the HD and composed of two subdomains (PAI and RED), each with three alpha helices (Jun and Desplan 1996; W. Xu et al. 1995). Most of the HDs in the PRD class possess a serine residue at position 50, which is key for the DNA binding specificity of homeobox proteins (Bürglin 2011). In mammals, the PRD class consists of nine PAX1-9 members, and based on the sequence similarity and structural features, these members have been subdivided into four families: PAX1/9, PAX2/5/8, PAX3/7 and PAX4/6 (Stuart, Kioussi, and Gruss 1994).

Some of the PRD class genes encode two additional conserved motifs, the Engrailed Homology 1 (EH1) motif and the OAR motif. The EH1 motif is a short peptide of about 10 amino acids that interacts with the Groucho family proteins to mediate transcriptional repression (Fisher and Caudy 1998; Tolkunova et al. 1998), whereas the OAR motif is known to play an important role in transcriptional activation during the development of metazoans (Vorobyov and Horst 2006). Moreover, the OAR motif sequence is 15 amino acids long, which was initially identified in the C-terminal region of the PRD-like family proteins: Otp, Aristaless and Rax; hence, the motif was named OAR after the initials of the three proteins (Furukawa, Kozak, and Cepko 1997). Interestingly, the PAX2/5/8 family possesses a partial HD consisting of only the first α-helix (Czerny et al. 1997), while the PAX1/9 family lacks an HD entirely (Chi and Epstein 2002). X-ray crystal structures of the human PAX5 (PDB ID: 1K78 and 1MDM) and PAX6 (PDB ID: 6PAX) PAIRED domain-DNA complex showed that each of the subdomains, PAI and RED, directly interact with the DNA motif, and amino acids from the linker connecting the two subdomains extensively contact the DNA minor groove (Eric Xu et al. 1999; Garvie et al. 2002; Garvie, Hagman, and Wolberger 2001).

Bürglin and Affolter have separated the PRD-like class from the PRD class because the PRD-like genes do not encode the PAIRED domain, but they do encode the HD as the major conserved domain with either lysine or glutamine amino acid at position 50 (Bürglin and Affolter 2016). Despite the high sequence similarity of the PRD and PRD-like HDs, the absence of a PAIRED domain in the PRD-like class suggests that a gene encoding the PAIRED domain might have merged with a PRD-like gene during early metazoan evolution, which has probably given rise to the PRD class (Galliot, De Vargas, and Miller 1999). In animals, the PRD-like genes are almost double in number compared to the PRD genes (Bürglin and Affolter 2016). The PRD-like TFs have been classified into 28 families based on their amino acid sequences; these families include ALX, ARGFX, ARX, DMBX, DPRX, DRGX, DUX, ESX, GSC, HESX, HOPX, ISX, LEUTX, MIX, NOBOX, OTP, OTX, PHOX, PITX, PROP, PRRX, RAX, RHOX, SEBOX, SHOX, TPRX, UNCX and

VSX (Bürglin 2011). Most of the PRD-like family genes are conserved in both vertebrates and invertebrates; however, genes encoding the LEUTX, DPRX, TPRX and ARGFX family members are absent in invertebrates (Bürglin 2011). Studies from our collaborators and others have shown that many PRD-like TFs are involved in early embryonic development and reproductive tissues (Madissoon et al. 2016; Töhönen et al. 2015).

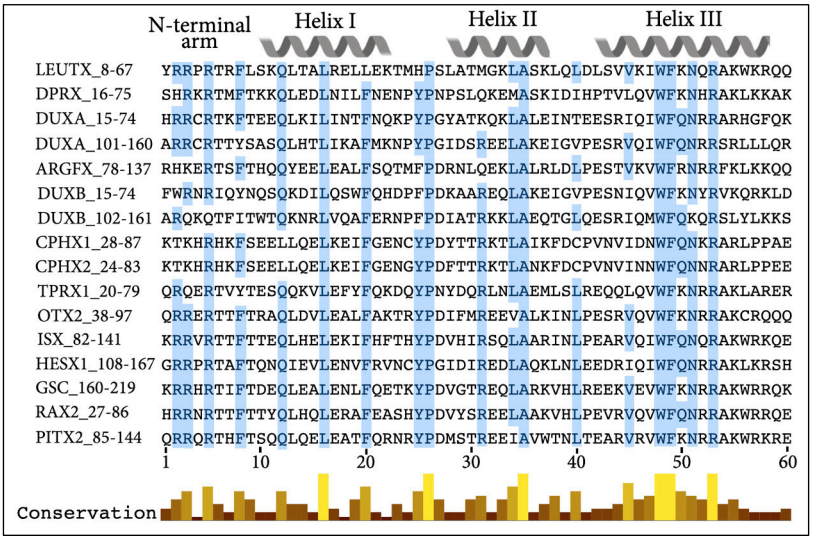


Figure 9: Amino acid sequence alignment for 14 PRD-like TFs that are highly expressed at some time point during early embryonic development and have the potential to bind a 36-bp dsDNA motif involved in EGA. Homeodomain helices are numbered above the aligned sequences and conserved residues are marked in blue; residues are numbered relative to the start of the homeodomain. A bar plot shows the level of conservation of amino acid residues.

Töhönen and colleagues identified 14 PRD-like TFs (Figure 9) that were found to be enriched for a 36-bp DNA recognition motif implicated in both early and late waves of EGA (Töhönen et al. 2015). Seven out of the 14 TFs are encoded by the maternally originated transcripts (TPRX1, OTX2, ISX, HESX1, GSC, RAX2 and PITX2), four are encoded by the embryonic genome (LEUTX, DPRX, ARGFX and DUXA), and the remaining three TFs are encoded by both the maternal and embryonic genomes (DUXB, CPHX1 and CPHX2) (Töhönen et al. 2015).

In humans, the PRD-like genes, except *OTX2*, are not expressed in the later stages of development due to the DNA methylation of their promoters, suggesting that the PRD-like genes are activated specifically during the early stages of embryonic development (Madissoon et al. 2016). Among the PRD-like TFs, DPRX has been observed to regulate the largest number of genes, the majority of which were downregulated and showed overlap with the downregulated target genes of DUXA and TPRX2 (Madissoon et al. 2016). Despite PRD-like HDs sharing high sequence identity, variability of amino acids at certain

key positions lead to significant changes in their binding affinities towards the same DNA motif. For example, in comparison to the wild type, the Q50A mutation in the engrailed HD (a member of the ANTP class) of *Drosophila melanogaster* showed a 2-fold lower binding affinity for the TAATTA motif; furthermore, the Q50K mutation induced a change in the binding site preference towards the TAATCC motif (Ades and Sauer 1994).

Similar to the PRD class, some of the PRD-like genes also consist of additional protein motifs, such as EH1 and OAR, which function to recruit activator and mediator proteins that stimulate the process of transcription (Bürglin and Affolter 2016). Although many HDs bind to a similar DNA motif, additional domains and motifs associated with HDs are likely to play a critical role in selective recognition of the DNA motif by forming additional molecular interactions with DNA and other mediator proteins. For example, HOX proteins (members of the ANTP class) interact with similar DNA motifs *in vitro* yet exhibit functional diversity under *in vivo* conditions (Svingen and Tonissen 2006).

2.2.4. LEUTX

Holland and colleagues analyzed the human genome sequence to predict the first RefSeq sequence of the *LEUTX* gene (NCBI RefSeq ID: NM_001143832.1) (Holland, Booth, and Bruford 2007). It was the only known *LEUTX* isoform in human until Jouhilahti and colleagues experimentally identified a novel full-length *LEUTX* isoform (NCBI RefSeq ID: NM_001382345.1) (Jouhilahti et al. 2016). It has been hypothesized that *LEUTX* originated by tandem duplication and asymmetric divergence from the *CRX* gene, most likely during the early radiation of placental mammalian species after their split from marsupials (Holland 2013). Molecular evolutionary analyses have suggested that the LEUTX homeodomain experienced positive selection during early placental mammalian evolution (Maeso et al. 2016; Zhong and Holland 2011). However, the HD has been reported to be absent in mice and rats, most likely due to the short gestation period in rodents (Jouhilahti et al. 2016).

Although *LEUTX* is not expressed in any human tissue under normal conditions, cell-type specific expression has been observed in the muscle biopsies of FSHD patients (Yao et al. 2014). Other than that, the expression of *LEUTX* is restricted to the 4-cell to 8-cell stage of the human embryo (Jouhilahti et al. 2016; Töhönen et al. 2015). It has been observed that LEUTX binding within the promoter region controls the expression profiles of about 25% of the upregulated genes during the 4-cell to 8-cell stage transition (Jouhilahti et al. 2016). A major gene regulatory shift has been observed at the onset of the 8-cell stage in which DPRX downregulates a large number of overlapping target genes (Jouhilahti et al. 2016). This self-regulatory mechanism, referred to as the two-stage model of human EGA, shows that DPRX counteracts the effect of LEUTX through recognition of the same DNA motif in the promoter regions of the overlapping target genes (Jouhilahti et al. 2016).

The binding of HDs to specific DNA sequences is highly dependent on amino acids located at specific key positions (Chu et al. 2012; Noyes et al. 2008). In our

studies, we modeled the HD region of LEUTX to identify specific features that match requirements for interacting with the TAATCC motif, which is present within the 36 bp DNA segment that was originally found enriched among the up-regulated genes during human EGA. Additionally, we studied the C-terminal region of LEUTX whose structure is not defined by X-ray or NMR methods, but likely plays a critical role in selective recognition of DNA by forming additional interactions with DNA and/or other proteins.

2.3. SARS-CoV-2 and epitopes recognizable by T-cells

2.3.1. SARS-CoV, MERS-CoV and SARS-CoV-2

Human coronaviruses (HCoVs) belong to the family *Coronaviridae*, the subfamily *Orthocoronavirinae*, and the order *Nidovirales* (Paules, Marston, and Fauci 2020). Members of the order *Nidovirales* generate a nested set of viral subgenomic mRNAs, which means the capped and polyadenylated subgenomic mRNAs are translated into more than one proteins from the 5' end (Pasternak, Spaan, and Snijder 2006). The *Coronaviridae* family is divided into two subfamilies, *Orthocoronavirinae* and *Torovirinae*, on the basis of the shape of nucleocapsids – a simple structure consists of genomic RNA and nucleocapsid phosphoprotein (Payne 2017). The subfamily *Orthocoronavirinae* includes four genera: *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus*, and *Deltacoronavirus* (Payne 2017). The alpha coronaviruses infecting humans include HCoV-NL63 and HCoV-229E, and the beta coronaviruses infecting humans are HCoV-OC43, HCoV-HKU1, SARS-CoV, MERS-CoV and SARS-CoV-2 (Figure 10).

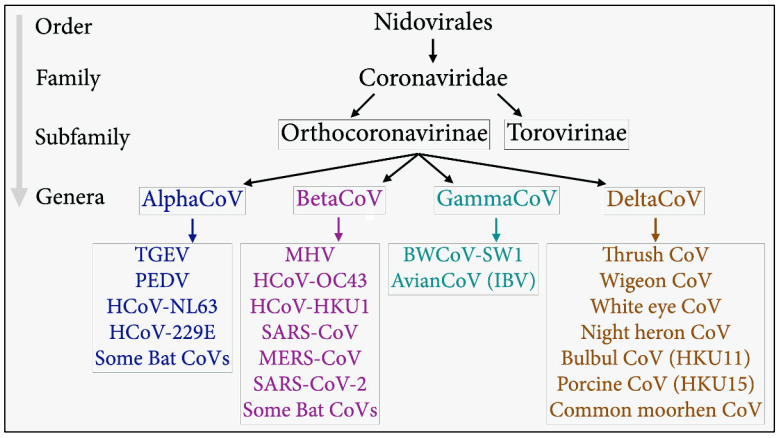


Figure 10: Classification of coronaviruses (CoVs). All CoVs belong to the order *Nidovirales*, the family *Coronaviridae* and the subfamily *Orthocoronavirinae*. The highly pathogenic human coronaviruses, namely SARS-CoV, MERS-CoV and SARS-CoV-2, belong to the genus *betacoronavirus* (BetaCoV).

At the beginning of the 21st century, three major coronavirus breakouts have been observed globally, the severe acute respiratory syndrome (SARS) in 2002, the Middle East respiratory syndrome (MERS) in 2012 and the coronavirus disease (COVID-19) in 2019. Before the first outbreak of coronavirus, it was believed that human coronaviruses (HCoV-NL63, HCoV-229E, HCoV-OC43 and HCoV-HKU1) cause mainly the upper respiratory illness in infected humans, including inflammation of the sinuses, nasal cavity, trachea (windpipe) and larynx (voice box), and these infections were not contagious from another human (D. X. Liu, Liang, and Fung 2021). However, previous studies have reported that the emergence of a new HCoV during the first outbreak in Foshan city of China has been observed to cause a severe lower respiratory tract illness, including bronchiolitis and pneumonia (D. S. Hui et al. 2005; Leung and Chiu 2004). Later in March 2003, the world health organization (WHO) announced that the causative agent of atypical pneumonia was SARS-CoV, in response to which immediate and highly effective global health measures were taken to halt the disease progression and severity (Rota et al. 2003). The SARS-CoV outbreak infected 8096 people in 29 countries/regions from November 2002 to July 2003, and resulted in 774 deaths across the globe, of which more than 83% deaths were specifically reported in China and special administrative regions of China (WHO report, published online on 24th July 2015). The overall fatality rate was around 10%, but it approached around 50% in elderly patients and those with certain underlying medical conditions (Park 2020). There was a sexual predisposition to females (53%), while the case fatality rate was higher in male patients especially in Chinese provinces (Jia et al. 2009). The median age of SARS cases in China was relatively lower (33 years) than that of other areas, for example, Taiwan (46), Hong Kong (40) and Canada (49) (Feng et al. 2009), and according to the WHO report, over 81% of cases occurred in individuals with age ≥50 years in mainland China (WHO report, produced at the global meeting on the epidemiology of SARS on 16-17 May 2003). The epidemic of SARS was halted in late June 2003 and it was thought that SARS-CoV would no longer circulate in humans due to the implementation of effective international public health measures, but a little less than a decade later, in 2012, another SARS-like illness emerged in the Middle East, which is known as MERS (Memish et al. 2013; Zaki et al. 2012).

Unlike SARS-CoV, the disease caused by MERS-CoV did not trigger a global pandemic and reflected geographical restrictions to the Arabian Peninsula and its surrounding countries (WHO report, published online on 7th April 2022). MERS-CoV infection resulted in a wide spectrum of clinical manifestations, ranging from upper respiratory tract illness to severe infection of the lungs and multiorgan failure. MERS-CoV caused 891 deaths among the 2585 infected patients (case fatality 35%) in 27 countries from September 2012 to February 2022 (WHO report, published online on 7th April 2022), indicating that MERS-CoV has continued to cause sporadic and localized outbreaks. There was a male predominance both in the number of reported cases (64%) and deaths (52% case fatality rate) (Alghamdi et al. 2014; Z. Zhu et al. 2020). While people from

all age groups were affected by the MERS-CoV outbreak similarly to that of the SARS epidemic, older persons with a median age of 50.6 years (range 2 to 109 years) were found severely affected (Salamatbakhsh et al. 2019).

The recent coronavirus breakout initiated when a cluster of acute pneumonia cases with unknown etiology were reported to the WHO by the health commission of Wuhan city, Hubei province of China on 31st December 2019 (C. Huang et al. 2020). Then on 7th January 2020, the causative agent was identified as a novel coronavirus (2019-nCoV), which was later renamed as SARS-CoV-2 (Gorbalenya et al. 2020), and the disease caused by it was termed COVID-19. The WHO officially declared the COVID-19 outbreak a worldwide pandemic on 11th March 2020 and, by then, the virus resulted in more than 118 thousand infections across 114 countries and over 4200 deaths. Compared to females, males affected with COVID-19 are higher in number (55.9% of total cases) (Nguyen et al. 2021) and showed a high degree of mortality (a male to female case fatality ratio >1) (Dehingia and Raj 2021). The median age of cases varies considerably by region and country due to the fact that the population of elderly people (age ≥65 years) who are more susceptible to disease is unevenly distributed across the globe. For example, the proportion of elderly people with respect to the total population of a country in 2020 and median age of death due to COVID-19 (in brackets) are as follows: 22.5% in Finland (84 years), 18.6% in the UK (82 years), 16.6% in the USA (77 years), 11.9% in China (70 years), and 6.5% in India (64 years) (Haravuori et al. 2020; Ioannidis 2020; Koya et al. 2021; Q. Bin Lu et al. 2021; Mohamed et al. 2020).

A study performed by Lopez-Leon and colleagues on more than 47 thousand infected patients (age ranging from 17 to 87 years) with SARS-CoV-2 reported that the most common symptoms (ranked in descending order) are fatigue, headache, attention disorder, hair loss, and dyspnea (Lopez-Leon et al. 2021). However, particularly in severe cases, the virus caused multi-organ dysfunction, which is characterized by acute lung and liver failure, acute renal injury, cardiovascular disease and neurological disorders (Gupta et al. 2020).

Current evidence suggests that SARS-CoV-2 is a highly infectious virus (R_0 value 2 to 3.5), which can survive for more than 3 hours in the air, with a half-life of more than 1 hour in aerosols at a temperature of between 21 to 23 °C (Doremalen et al. 2020). SARS-CoV-2 can directly be transmitted via airborne droplets ejected from an infected person through coughing and sneezing over an extended distance of 7 to 8 meters (Bourouiba 2020), while indirect transmission is possible via contaminated surfaces. It is worth noting that infected people can spread COVID-19 already 2 to 4 days before they have any symptoms and the mean incubation period for SARS-CoV-2 is 4 to 8 days (Tindale et al. 2020). Similarly, asymptomatic individuals (*i.e.*, showing no symptoms throughout the infection) also transmit the virus to others even when they feel fine (Oran and Topol 2020). A brief comparison among the three highly pathogenic HCoVs SARS-CoV-2, SARS-CoV and MERS-CoV is summarized in Table 1.

Table 1: A comparison among the highly pathogenic human coronaviruses: SARS-CoV-2, SARS-CoV and MERS-CoV.

	Coronavirus		
	SARS-CoV-2	**SARS-CoV**	**MERS-CoV**
Disease	Coronavirus disease 2019 (COVID-19)	Severe acute respiratory syndrome (SARS)	Middle East respiratory syndrome (MERS)
Genus	*Betacoronavirus*	*Betacoronavirus*	*Betacoronavirus*
Viral genome	+ssRNA	+ssRNA	+ssRNA
Genome length	29.90 kb	29.75 kb	30.11 kb
Genome reference sequence (NCBI)	NC_045512.2	NC_004718.3	NC_019843.3
Total number of encoded protein	26	25	25
First emergence	December 2019	November 2002	April 2012
First case location	Wuhan, China	Foshan, China	Zarqa, Jordan
Countries/regions with cases	226	29	27
Cumulative cases	Above 4.4 billion[a]	8096	2428[b]
Cumulative deaths	Above 5.9 million[a]	774	838[b]
Fatality	0.13%	9.5%	34.5%
Median incubation period	5.1 days (95% CI, 4.5 to 5.8 days)[c]	4 days (95% CI, 3.6 to 4.4 days)[d]	5.2 days (95% CI, 1.9 to 14.7 days)[e]
Time to infect first 1000 people (days)	48	130	903 (~2.5 years)
Ratio (male/female)[f]	1.27	0.88	1.78
Basic reproduction number (R_0)[g]	2 to 3.58	1.7 to 1.9	< 1
Recent status	Pandemic ongoing	Completely controlled	Sporadic continuous

According to data released by the WHO on 4th March 2022[a]; (David S. Hui et al. 2018)[b]; (Lauer et al. 2020)[c]; (Lessler et al. 2009)[d]; (Assiri et al. 2013)[e]; (Z. Zhu et al. 2020)[f]; (P. Wu et al. 2020)[g]; CI: Confidence Interval.

2.3.2. Origin of SARS-CoV-2

The origin of SARS-CoV-2 is still controversial. It is suspected that the SARS-CoV-2 virus is the product of natural evolution (Andersen et al. 2020), while some scientists proposed that it arose via human intervention (Amarilla et al. 2021; Segreto and Deigin 2021). The reference genome of SARS-CoV-2 is more closely identical to the genome of RaTG13-CoV (a SARS-like coronavirus that infects several species of horseshoe bats; GenBank ID: MN996532.2) than the genome

of SARS-CoV (96% vs 79.6% sequence identity) (P. Zhou et al. 2020). In addition, evolutionary studies on coronaviruses have reported that RaTG13-CoV is the closest known relative to SARS-CoV-2 (Cagliani et al. 2020; Gorbalenya et al. 2020; P. Zhou et al. 2020). A recent study has found that two bat coronaviruses, RaTG13 and BANAL-52, exhibit a higher level of nucleotide identity with the S1 subunit of SARS-CoV-2 spike protein, especially in the receptor binding domain (RBD) of the spike protein, suggesting that viruses similar to SARS-CoV-2 might have circulated between different species of bats living in the same geographical area (Temmam et al. 2022).

Amarilla and colleagues suspect that SARS-CoV-2 may have leaked from a laboratory (Amarilla et al. 2021). The "lab leak hypothesis" gained little traction when a research group at Wuhan Institute of Virology, who conducted the first evolutionary study on SARS-CoV-2 (P. Zhou et al. 2020), released the complete genomic sequence of RaTG13-CoV after the outbreak of SARS-CoV-2, even though the original samples of RaTG13-CoV were already collected in 2013. Subsequently, many comparative studies have been published to investigate the level of similarity between SARS-CoV-2 and other human and bat CoVs. For example, amino acid residues, ^{481}NGEVGFN^{487}, in the receptor binding motif (RBM) of SARS-CoV-2 (PDB ID: 6VW1) (Shang, Ye, et al. 2020) that directly interact with the human ACE2 receptor are more similar to the RBM residues in RaTG13-CoV (^{163}NGQTGLN^{169}, PDB ID: 7DRV) (K. Liu et al. 2021) than that of the SARS-CoV residues (^{468}TPPALN^{473}, PDB ID: 2AJF) (F. Li et al. 2005). Of the all human and bat CoVs, only SARS-CoV-2 spike protein encoding mRNA contains an insertion of 12 bases, coding for a stretch of four amino acids (^{681}PRRA^{684}) (Coutard et al. 2020). The inserted segment forms an exposed tribasic S1/S2 cleavage site (^{681}PRRAR↓SV^{687}) for the endoprotease enzyme furin, which upon peptidase activity facilitates the process of virus entry into the host cell (Coutard et al. 2020). However, the more distantly related MERS-CoV contains an analogous S1/S2 cleavage site (PRSVR↓SV), which, due to the serine residue, corresponding to R683 in SARS-CoV-2, showed the absence of furin-mediated recognition and cleavage of the MERS-CoV spike protein (Temmam et al. 2022).

It is now nearly three years since the WHO declared COVID-19 a pandemic, and multiple lines of evidence support the hypothesis of natural evolution and zoonotic transmission for the origin of SARS-CoV-2, a not uncommon mechanism of viral transmission across different species. In human, it is estimated that about 60% of emerging infectious diseases are zoonoses, *i.e.*, disease that has transmitted naturally from infected vertebrate animals to humans, and resulted in more than one billion cases of illness and millions of deaths every year (Jones et al. 2008). There are over 200 zoonotic diseases worldwide, of which over 54% of diseases are caused by bacteria, especially rickettsia, and over 25% of diseases are linked to viral or prion pathogens (Jones et al. 2008). Furthermore, most of these viral pathogens in humans originated from wild animals, for example, several major outbreaks that have been caused by RNA viruses such as Nipah, Marburg, Ebola, SARS-CoV and MERS-CoV are associated with bats (Letko et al. 2020). Recent genomic studies have reported that SARS-CoV-2 may also have

been originated from bats because the SARS-CoV-2 shares a high whole-genome identity with SARS-CoV-like coronavirus originating in bats (Gussow et al. 2020; Irving et al. 2021; H. Wang et al. 2020). The involvement of an intermediate host in the transmission of coronaviruses began to be recognized with the detection of SARS-CoV-like coronaviruses in open-air market animals, including Himalayan palm civets and raccoon dog (Guan et al. 2003). More recently, the discovery of SARS-CoV-2-like coronaviruses in Malayan pangolins has raised the possibility of intermediate host species involvement (Figure 11), where SARS-CoV-2 acquires genetic modifications, such as via mutation and recombination, for efficient transmission in humans (Lam et al. 2020). Furthermore, investigations on the transmission channel of SARS-CoV-2 revealed involvement of additional wild animals as the potential hosts, which include mink, turtle, snake and ferrets (Zhao, Cui, and Tian 2020). Besides wild animals, several domestic animals, such as dogs and cats, were also found infected with SARS-CoV-2 (Zhao, Cui, and Tian 2020).

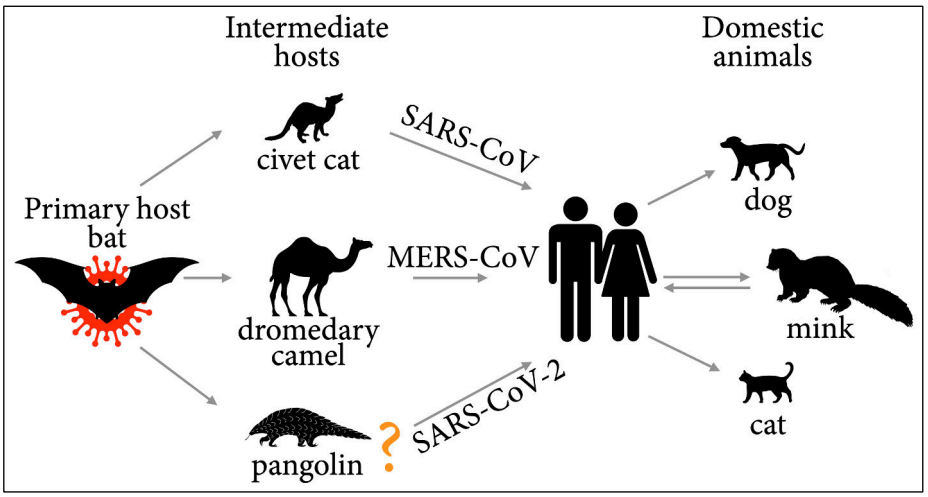


Figure 11: Schematic representation of the zoonotic spill-over of SARS-CoV, MERS-CoV and SARS-CoV-2 from bat to humans through intermediate hosts. Domestic animals were also found infected with the SARS-CoV-2 virus, most likely via infected humans.

2.3.3. SARS-CoV-2 genome, structure and evolution

Wu and colleagues at Fudan university in Shanghai have reported the first complete genome sequence of the SARS-CoV-2, a single stranded positive RNA (+ssRNA) virus of 29.9 kb in size (F. Wu et al. 2020). This sequence is considered as the reference sequence (RefSeq accession number: NC_045512.2), meaning that any newly discovered variant of SARS-CoV-2 will be compared against it to gain a better understanding of the pathogenesis by identifying similarities and differences between genomes. The SARS-CoV-2 genome shares ~79% identity with SARS-CoV and ~50% identity with MERS-CoV (R. Lu et al. 2020), and shares

more than 99% identity to SARS-CoV-2 isolates collected from different countries (Khan et al. 2020). Similar to SARS-CoV and MERS-CoV, the genome of SARS-CoV-2 encodes structural proteins, nonstructural proteins (NSPs), and accessory proteins; however, differences in the number of encoded proteins and amino acids were observed during comparative analysis of coronavirus genomes (Brant et al. 2021). At the 5' end, two-thirds of the SARS-CoV-2 genome consists of two long overlapping open reading frames (ORFs), ORF1a and ORF1b, encoding two polyprotein precursors (pp1a and pp1b) that upon proteolytic cleavage generate 16 NSPs that are involved in virus replication and transcription (V'kovski et al. 2021). The remaining one-third of the genome of SARS-CoV-2 encodes four major structural proteins that are common to all CoVs: the spike or surface (S) glycoprotein, membrane (M) glycoprotein, nucleocapsid (N) phosphoprotein, and the envelope (E) protein, and also encodes 6 ORFs encoding accessory proteins that vary in number between different CoVs (Figure 12). Both the polyproteins are processed by the viral proteases, papain-like protease (PLpro, NSP3) and 3-chymotrypsin-like protease (3CLpro, NSP5), of which PLpro is responsible for proteolytic cleavage of NSP1 to NSP3, and 3CLpro is responsible for processing of the remaining NSPs, *i.e.*, NSP5 to NSP16; however, NSP4 is cleaved by both PLpro and 3CLpro at the N-terminus and C-terminus, respectively (Moustaqil et al. 2021).

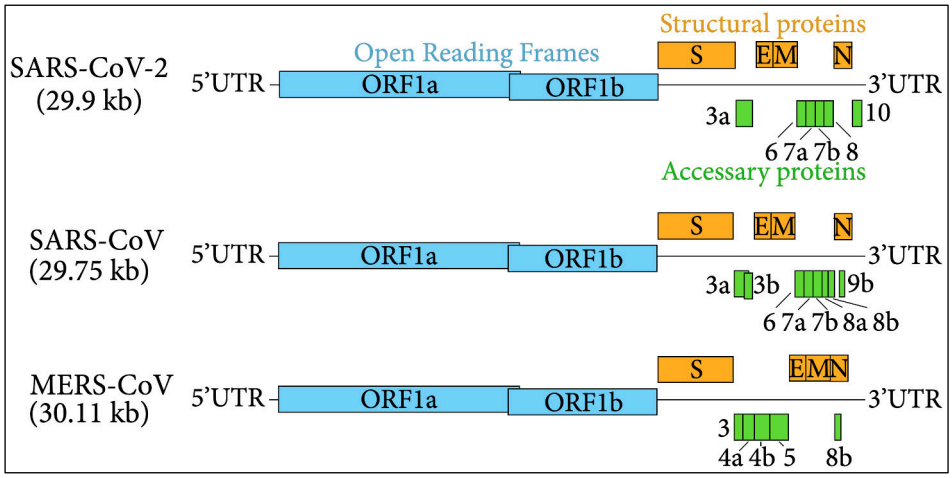


Figure 12: Comparison of SARS-CoV-2, SARS-CoV and MERS-CoV genomes (genome size for each CoV shown in brackets). Two partially overlapping ORFs, ORF1a and ORF1b, are common to all three CoVs. Unlike SARS-CoV-2, SARS-CoV and MERS-CoV contain additional overlapping genes, such as ORF3b and ORF9b in SARS-CoV, and ORF8b in MERS-CoV. Each CoV contains four structural proteins: the spike (S) glycoprotein, membrane (M) glycoprotein, envelope (E) protein, and nucleocapsid (N) phosphoprotein. Different numbers of accessary proteins are encoded by each CoV.

In order to preserve the genome integrity, SARS-CoV-2 encodes 3'-5' exoribonuclease protein (ExoN, also named NSP14), which enhances the overall

fidelity of RNA synthesis by correcting nucleotide incorporation errors catalyzed by the RNA dependent RNA polymerase (RdRp, also named NSP12) and its two accessory proteins NSP7 and NSP8 that form a heterodimer complex (Kirchdoerfer and Ward 2019; Moeller et al. 2022). However, mutations in NSP14 have shown a strong association with less efficient replication (low viral fidelity) (Moeller et al. 2022), resulting in the emergence of new variants that pose a threat to public health. As of 17th March 2022, the WHO has designated five variants of SARS-CoV-2 (alpha, beta, gamma, delta, and omicron) as variants of concern (VOC) that have changed the course of the pandemic in different countries at different points in time (Table 2).

Table 2: List of SARS-CoV-2 variants of concern (as of 17th March 2022). The data were compiled from the WHO website and the Stanford University's coronavirus antiviral and resistance database.

SARS-CoV-2 variant	**First detected**	**Declared VOC (month and year)**	**Mutations in Surface/Spike (S) protein**
Alpha (B.1.1.7)	UK September 2020	December 2020	N501Y, A570D, D614G, P681H, T716I, S982A, D1118H, Δ69-70, Δ144-145
Beta (B.1.351)	South Africa October 2020	December 2020	D80A, D215G, K417N, E484K, N501Y, D614G, A701V, Δ241-243
Gamma (P.1)	Brazil December 2020	January 2021	L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, D614G, H655Y, T1027I, V1176F
Delta (B.1.617.2)	India December 2020	May 2021	T19R, G142D, E156G, L452R, T478K, D614G, P681R, D950N, Δ157-158
Omicron (B.1.1.529)	South Africa November 2021	November 2021	A67V, T95I, G142D, N211I, G339D, S371L, S373P, S375F, K417N, N440K, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, Y505H, T547K, D614G, H655Y, N679K, P681H, N764K, D796Y, N856K, Q954H, N969K, L981F, Δ69-70, Δ143-145, Δ212, 214EPEins

The S protein is a homotrimeric type I transmembrane fusion protein that forms large protrusions from the virus surface, giving the appearance of a radiate crown, hence the name coronavirus (corona in Latin means crown) (F. Li 2016). In addition to mediating cell adhesion and virus internalization, the S protein also determines the range of cell types and host species that a coronavirus can infect (F. Li 2016). Structurally, the S protein is synthesized as a 1273 amino acid polyprotein precursor, which has a large N-terminal ectodomain, a

transmembrane domain (1213-1237 residues) and a short cytoplasmic domain (1237-1273 residues). The 180 kDa monomeric S glycoprotein is comprised of two functional subunits S1 (14-685 residues) and S2 (686-1273 residues) that are connected by a sequence of polybasic amino acids ^{680}SPRRARSV^{687}, forming a cleavage site for the furin and other proteases (Y. Huang et al. 2020; Örd, Faustova, and Loog 2020). The receptor binding domain (RBD, residues from 319 to 541) of S1 subunit infects the human cells by binding to angiotensin-converting enzyme 2 (ACE2) receptor (Hoffmann et al. 2020), while the S2 subunit mediates the process of membrane fusion and internalization of SARS-CoV-2, which is initiated after a proteolytic processing event at S2' cleavage site within the S2 subunit by the TMPRSS2 serine protease (Hoffmann et al. 2020). A recent study demonstrates that the S protein of SARS-CoV-2 binds to ACE2 receptor with higher affinity than the S protein of SARS-CoV (Shang, Wan, et al. 2020), indicating that the residues contributing to ACE2 binding are quite different due to the mutations between the S protein of SARS-CoV-2 and SARS-CoV (Shang, Ye, et al. 2020). Cryogenic electron microscopy (cryo-EM) studies have shown the different dynamic states of the RBD in different coronaviruses; for example, the RBD in SARS-CoV is mostly in a "standing-up" state (Gui et al. 2017), whereas the RBD in SARS-CoV-2 is mostly in the "lying-down" state (Wrapp et al. 2020).

As compared with the bat-CoVs, the S protein of SARS-CoV-2 binds with much greater efficiency to human ACE2 and thus promotes the rapid spread of SARS-CoV-2 throughout global populations (Damas et al. 2020). Several groups of researchers conducted comparative genomics studies to understand the molecular properties of the S protein that allow it to bind with the ACE2 receptor, and it was found that the S protein exhibits stronger purifying selection than the adaptive changes in amino acids, which were found to be highly localized to specific positions in the RBD domain of S protein (Cagliani et al. 2020; Damas et al. 2020; Tang et al. 2020). During the COVID-19 pandemic, the newly evolved variants of SARS-CoV-2 in human showed more immediate outcome of infections and more resistance to vaccines than the original Wuhan strain (R. Wang et al. 2022). In the S protein of the Alpha and Beta variants of SARS-CoV-2, a predominant D614G mutation (also observed in the Delta, Kappa and Omicron variants) results in an increased propensity of the S protein to adopt the open conformation to enhance the ACE2 recognition efficiency (Benton et al. 2021; J. Zhang et al. 2021). In addition to D614G, another K417N mutation in the Beta variant (also observed in the Omicron variant), enhances the stability of the open state of the trimeric S protein, suggesting that the virus has achieved greater transmissibility in human during its evolution (Wrobel et al. 2022).

The M protein of SARS-CoV-2 is composed of 222 amino acids and shares more than 90% sequence identity with the M protein of SARS-CoV. Unlike the S protein, structurally, the M protein belongs to the type III transmembrane protein classification, containing a small glycosylated ectodomain, a triple-span transmembrane domain, and a large C-terminal endodomain (Mariano et al. 2020). Among the structural proteins, the M protein is the most abundant

protein that defines the shape of the viral envelope (Neuman et al. 2011). During the formation of mature virions, the M protein is considered as the central organizer of CoV assembly, co-localizing almost entirely with the S and E proteins in the rough endoplasmic reticulum (RER) and the Golgi complex of the infected host cell (Duan et al. 2020). Additionally, molecular interactions between the C-terminal domains of the M and N proteins play an essential role in the SARS-CoV-2 life cycle (S. Lu et al. 2021).

In SARS-CoV-2, the 45 kDa N protein, composed of 419 amino acids, forms complex with the negatively charged genomic RNA to aid its packaging into the enveloped virion of about 100 nm in diameter (Jack et al. 2021; Ning et al. 2021). Unlike other structural proteins in SARS-CoV-2, the N protein lacks cysteine residues and predominantly contains intrinsically disordered regions (IDRs), which could be the reason of its less stability than other structural proteins (Jack et al. 2021). The IDRs include mainly three segments: the N-terminal arm (1-40 residues), a central serine and arginine rich flexible linker region (174-249 residues) and a C-terminal tail (365-419 residues) (M. Yang et al. 2021). The linker region connects the two structurally and functionally conserved regions of the N protein, including an N-terminal domain (NTD, 41-173 residues) that binds with the viral RNA genome and a C-terminal domain (CTD, 250-364 residues) that directly participates in N protein dimerization (M. Yang et al. 2021). The N protein is the most abundant protein in SARS-CoV-2, is highly immunogenic, and it shares a sequence more than 90% identical with the N protein of SARS-CoV (Dutta, Mazumdar, and Gordy 2020). Gao and colleagues had reported that monkeys vaccinated with an adenoviral-based SARS-CoV vaccine showed a strong antibody response against the S1 subunit of S protein, and also showed a consistent T-cell response against the N protein (Gao et al. 2003).

The SARS-CoV-2 E protein is a small (75 amino acids) structural protein whose sequence shares 94.7% identity with the E protein of SARS-CoV (Duart, García-murria, and Mingarro 2021). The E protein folds into three domains, namely, the N-terminal domain (1-16 residues) on the virion surface, a transmembrane domain (17-37 residues) and a big intravirion C-terminal domain (38-75 residues) (Duart, García-murria, and Mingarro 2021). The SARS-CoV-2 E protein forms a homopentameric cation channel similar to that of the SARS-CoV E protein, which is involved in critical aspects of the viral life cycle, such as assembly, budding and pathogenesis (Mandala et al. 2020). Nieto-Torres and colleagues have reported that the channel formed by the SARS-CoV E protein has a higher ion permeability than the other two channels that are formed by ORF3a and ORF8a proteins, and mutations in the E protein channel reduce viral pathogenicity to some extent (Nieto-Torres et al. 2014). However, the SARS-CoV-2 genome encodes only two ion channels: the E protein channel and the ORF3a channel (Kern et al. 2021). Indeed, calcium is a necessary component for the virus life cycle and virulence, and the E protein channel plays an essential role in maintaining calcium ion homeostasis (Torres et al. 2021).

The functions of some of the SARS-CoV-2 proteins are not known yet but their homology to other HCoVs suggests their function might be the same. As of 2nd April 2022, the protein data available from UniProt database show that most of the SARS-CoV-2 proteins are manually annotated and reviewed, and experimental data are available from the Protein Data Bank (PDB) for X-ray crystal structures of 22 SARS-CoV-2 proteins, namely, NSP1, NSP3 (PLpro), NSP5 (3CLpro), NSP7, NSP8, NSP9, NSP10, NSP12 (RNA-dependent RNA polymerase), NSP13 (helicase), NSP14 (3′-to-5′ exonuclease), NSP15 (endoribonuclease), NSP16 (2′-O-ribose methyltransferase), S, ORF3a, E, M, ORF6, ORF7a, ORF7b, ORF8, N, and ORF10.

2.3.4. T cell recognition of foreign antigens bound to major histocompatibility complex (MHC) molecules

Infection with SARS-CoV-2 induces antibody responses, especially IgG and IgM antibodies, and the levels of antibodies mainly depend on the magnitude of the initial viral load and disease severity (Cox and Brokstad 2020). Generally, in humans, immunity to coronaviruses can last from months to several years after infection (A. T. Huang et al. 2020). However, persons with mild COVID-19 showed a rapid decline in antibodies with a half-life of ~21 days for IgG, suggesting that the immunity against SARS-CoV-2 may only be maintained for few months (Ibarrondo et al. 2020; Seow et al. 2020). It is interesting to note that the induction of adaptive immunity creates memory lymphocytes (B cells and T cells) even if there are low levels of serum antibodies (Cox and Brokstad 2020). Dan and colleagues have reported that about 95% of the COVID-19 patients retained the immune memory cells 6 months after infection (Dan et al. 2021).

All three major lymphocytes of the adaptive immune system, namely, $CD4^+$ T cells (helper T cells), $CD8^+$ T cells (killer or cytotoxic T cells) and B cells (antibody generating cells) are important in immune protection against pathogens (Rydyznski Moderbacher et al. 2020). In SARS-CoV infected cells, epitopes from the two major structural proteins, S and N, have been shown to effectively enhance the MHC class I mediated immune responses in which $CD8^+$ T cells exhibit cytotoxic activity that kills virus infected cells (Tsao et al. 2006). Similarly, recent studies showed a highly correlated response of T cells with low COVID-19 severity during the acute phase, suggesting that high levels of induced $CD4^+$ and $CD8^+$ T cells may control and eliminate the viral infection (Rydyznski Moderbacher et al. 2020; P. Zhou et al. 2020). In addition, SARS-CoV-2 elicits robust $CD8^+$ T cell responses in mice and the non-human primate (*Macaca mulatta*), which likely contributes to host protection from the development of severe COVID-19 (McMahan et al. 2021; Zhuang et al. 2021). Patients severely affected by COVID-19 disease showed low $CD4^+$ T cell counts, which is usually associated with intensive care unit (ICU) admission (Chen et al. 2020), whereas low $CD8^+$ T cell counts have been found associated with high mortality (Du et al. 2020; B. Xu et al. 2020). However, generally, neutralizing antibodies produced by the B cells showed no clear correlation with decreased severity of disease (Rydyznski Moderbacher et al. 2020; Wajnberg et al. 2020).

Depending on the type of T cell, the surface glycoproteins CD4 and CD8 along with T cell receptor (TCR) are involved in recognition of major histocompatibility complex (MHC) molecules (Doyle and Strominger 1987; Garcia et al. 1996). On the surface of antigen presenting cells (APCs), such as dendritic cells, macrophages and B cells, MHC molecules display fragments of the processed antigenic peptides so that approaching T cell can engage with this molecular complex via their TCRs (Harvey et al. 2007). Based on the antigen processing and presentation pathway, the MHC molecules are mainly divided into two classes: 1) MHC class I, which is expressed on the cell surface of nearly all nucleated cells (not just APCs) and typically presents endogenous antigens, such as from viruses, to $CD8^+$ T cells, and 2) MHC class II, which is expressed mainly on APCs and primarily presents exogenous antigens, such as from bacteria, to $CD4^+$ T cells (Storni and Bachmann 2004). In humans, the MHC system is known as the human leucocyte antigen (HLA) complex, which is located on the short arm of chromosome 6 and contains the most polymorphic loci of the human genome, *i.e.*, a tightly linked cluster of genes encodes many different allotypes in different individuals inside a population (T. M. Williams 2001). The genetic diversity and evolution of HLA allotypes lead to different immune responses in different ethnic populations, which either protect people against new viral infections or enhance viral persistence (Crux and Elahi 2017).

Structurally, the HLA class I molecule is a heterodimer, which is comprised of a membrane spanning α heavy chain and a light β2-microglobulin chain. In humans, the heavy chain is encoded by the HLA-A, HLA-B and HLA-C loci. The heavy chain is comprised of an N-terminal extracellular region consisting of three domains (α1, α2 and α3), a transmembrane domain and a cytoplasmic C-terminal tail (Figure 13A). The α1 and α2 domains fold together to form a deep cleft that binds noncovalently to an 8-11 amino acids long epitope – a part of an antigen capable of eliciting an immune response (Trolle et al. 2016). The length distribution of HLA class I restricted epitopes is primarily determined by both epitope accessibility and HLA allotype specific binding preference (Trolle et al. 2016). The binding core of epitopes is 9 (typically) or 10 amino acids in length (Trolle et al. 2016).

In humans, there are three different classical HLA class II isotypes: HLA-DP, HLA-DQ and HLA-DR, all of which are encoded by genes located on the HLA-D locus (Ulvestad et al. 1994). Like the HLA class I molecule, the class II molecule is also a heterodimer, but it is comprised of two membrane-spanning α and β chains (Figure 13B). Each chain consists of two domains, α1 and α2 domains in the α chain and β1 and β2 domains in the β chain (Brown et al. 1993). Both α1 and β1 domains fold together to form an epitope binding cleft, whereas the two immunoglobulin-like domains, α2 and β2, provide structural and mechanical support to the complex (Stern et al. 1994). Unlike HLA class I, the epitope binding cleft in the class II molecule is open at both ends, which can accommodate a wide range of epitope lengths (13-25 residues), with an average epitope length of 15 amino acids (Chicz et al. 1992; Rudensky et al. 1991).

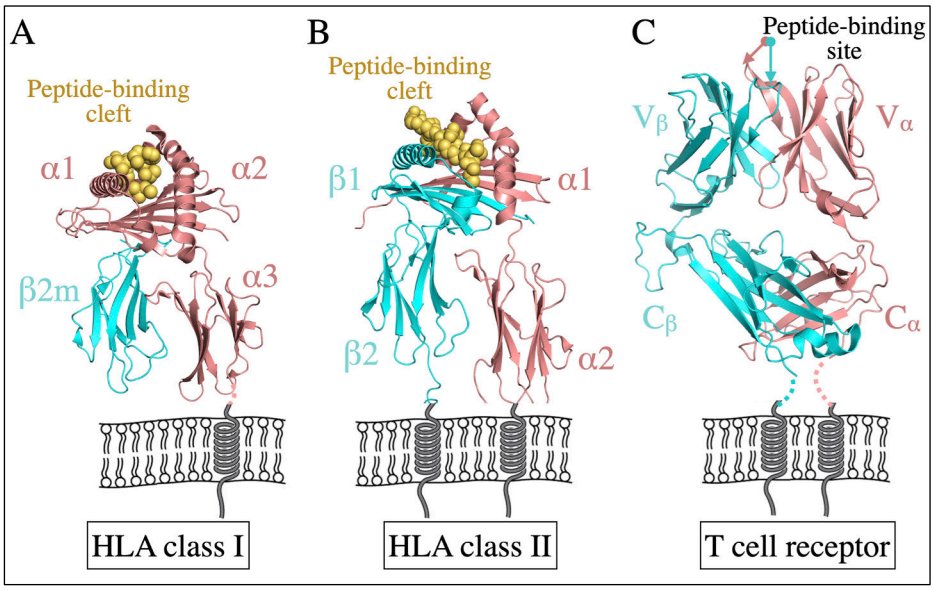


Figure 13: Structures of heterodimeric transmembrane HLA class I, HLA class II and T cell receptor (TCR) molecules. A) HLA class I molecule comprised of two subunits: the heavy α chain and β2 microglobulin (β2m). The heavy chain contains three extracellular domains (α1, α2, and α3, with α1 being at the N-terminus), a transmembrane region and a C-terminal cytoplasmic tail. The extracellular β2m domain binds non-covalently to the α chain. The immunogenic epitope (gold spheres) binds to the peptide-binding cleft located between the α1 and α2 domains. B) HLA class II molecule comprised of two subunits: an α chain and a β chain. Both the α and β chains contain two extracellular domains, a transmembrane region and a C-terminal cytoplasmic tail. The immunogenic epitope (gold spheres) binds to the peptide-binding cleft located between the α1 and β1 domains. C) T cell receptor (TCR) comprised of two subunits: an α chain and a β chain. Both the α and β chains contain two extracellular immunoglobulin-like domains: variable (V) and constant (C), a proximal connecting peptide (shown dashed), a transmembrane region and a C-terminal cytoplasmic tail. Epitopes presented by the HLA molecules bind to the peptide-binding site of TCR.

Unlike antibodies that bind to free floating antigens in blood and other body fluids, the TCRs bind to epitopes of foreign proteins that are presented on the surface of infected cell via HLA molecules. The TCR is a disulfide-linked, membrane-anchored heterodimeric complex mostly consisting of an α chain and a β chain, but a less common type of TCR is comprised of a different set of chains, one gamma (γ) and one delta (δ) (Gaulard et al. 1990). Structurally, the α and β chains are folded into an N-terminal extracellular region consisting of two immunoglobulin-like domains: variable (V) and constant (C), a membrane-proximal connecting peptide (CP), a transmembrane domain and a short cytoplasmic C-terminal tail (Figure 13C), of which the most distal extracellular domains VαVβ mediate noncovalent interactions with the epitope-bound HLA molecule (Rudolph, Stanfield, and Wilson 2006). Since the cytoplasmic regions of both chains lack signaling sequences due to their extremely short tails, other

costimulatory membrane receptors are needed to induce the cytotoxic activity in $CD8^+$ T cells, which includes the release of cytokines such as interferon gamma (IFN-γ) and tumor necrosis factor (TNF) for the killing of virus infected cell (Kaech and Wherry 2007).

It is of extreme importance to understand host immune responses during the progression from mild to potentially fatal COVID-19 disease in order to develop effective vaccines and therapeutics against SARS-CoV-2. Indeed, the HLA system is widely used to understand the cause, mechanism and effect of infectious diseases. For example, a previous study of HLA allotypes in Taiwanese population has reported that individuals with the HLA-B*46:01 allotype are most likely to fall victim to SARS (Lin et al. 2003). In addition, the X-ray crystal structure of HLA-A*24:02 allotype with SARS-CoV nucleocapsid protein-derived epitope (PDB ID: 3I6L) has led to a better understanding of the process of epitope selection and presentation strategy at the molecular level (Jun Liu et al. 2010). It is interesting to note that HLA-A*24 allotypes have been reported as the second most common allele groups in patients infected with SARS-CoV throughout the global population, the most common being the HLA-A*02 allotypes (Jun Liu et al. 2010). Likewise, HLA-A*02:01-restricted structural protein-derived epitopes in Chinese patients showed great potential for characterization and evaluation of candidate SARS vaccines (M. Zhou et al. 2006).

In our study, we performed an integrated and computational analysis to identify HLA class I restricted immunogenic SARS-CoV-2 epitopes that could elicit an effective cytotoxic T cell response. Furthermore, the predicted 9-mer epitopes were compared with the experimental data for SARS-CoV-derived cytotoxic T cell epitopes, which is available at the immune epitope database (IEDB) (Vita et al. 2019). Later, the three-dimensional molecular structures of selected ternary complexes (eHLA-TCR) were modeled to assess interactions at the atomic level. Our findings could potentially be used for designing vaccine epitopes against SARS-CoV-2 and also provide useful information for studying HLA class I restricted $CD8^+$ T cell responses.

2.3.5. COVID-19 vaccines approved for emergency use by WHO

Around the world, several vaccines are being developed against COVID-19 using different vaccine platform technologies. As of 12 April 2022, according to data provided by the WHO: COVID-19 vaccine tracker and landscape, there are 153 vaccines in clinical development and 196 vaccines in pre-clinical development. Out of the 153 clinical phase candidates, the highest number of vaccines (34%) are developed using a protein subunit (PS) platform, followed by (in decreasing order) RNA, non-replicating viral vector (VVnr), inactivated virus (IV), DNA, virus like particles (VLP) and replicating viral vector (VVr). The route of administration for most of the vaccine candidates (77%) is intramuscular. However, some vaccine candidates (3%) are designed to be administered orally. As of 12 April 2022, thirteen vaccine candidates are in phase 4 of the clinical trial, namely (vaccine platform indicated in brackets), CoronaVac (IV), BBIBP-CorV/Vero Cell (IV), mRNA-1273 (RNA), BNT162b2/Comirnaty (RNA), CV2CoV

(RNA), MIPSCo-mRNA-RBD-1 (RNA), ChAdOx1-S/AZD1222 (VVnr), Ad26.COV2.S (VVnr), Ad5-nCoV-IH (VVnr), MVC-COV1901 (PS), V-01-351/V-01D Bivalence vaccine (PS), Betuvax-CoV-2 (PS) and DoCo-Pro-RBD-1 with MF59 adjuvant (PS).

Inactivated (or killed) vaccines have been used for over a century to induce protection against pathogens that are known to cause infectious diseases, some of the infectious diseases caused by viruses include: rabies, polio, hepatitis A and influenza (Jeyanathan et al. 2020). Inactivated vaccines are developed by infecting culturing cells with a desired pathogen in a bioreactor-based cell culture system, followed by killing of the pathogen using heat, chemical or radiation to destroy the infectivity while retaining its immunogenicity (Seo 2015). Generally, with such vaccines, booster shots are given at intervals to the people who have already completed their primary vaccination course in order to maintain immunity above a protective level (Yue et al. 2022). According to the WHO report, released on 2 April 2022, three inactivated vaccine candidates, Sinovac's CoronaVac, SinoPharm's BBIBP-CorV and Bharat Biotech's COVAXIN, have been listed for COVID-19 emergency use. These inactivated vaccines have been observed to elicit humoral and T cell-mediated immune responses against the SARS-CoV-2 structural proteins (Vályi-Nagy et al. 2021; Vikkurthi et al. 2022). However, individuals vaccinated with BBIBP-CorV showed a significantly lower level of antibody response compared with those who have received the BNT162b2 mRNA vaccine (Vályi-Nagy et al. 2021).

Unlike inactivated vaccines, mRNA vaccines have a highly satisfactory safety profile without the potential risk of pathogenicity because they can be easily constructed directly from the genetic sequence of the desired pathogenic antigen without the need for pathogens (M. A. Liu 2019). In the past, various mRNA vaccines have been shown to elicit potent immunity against infectious diseases, including those caused by Zika, rabies and influenza viruses (Pardi et al. 2018). The most common delivery vector for a synthetically created mRNA vaccine is a lipid nanoparticle (LNP), which facilitates the process of endocytosis and protects the mRNA molecule from enzymatic degradation inside the host cell (Batty et al. 2021). On entering host cells, the viral mRNA is translated into protein by the host translational machinery and ribosomes, followed by proteasome-mediated degradation of the encoded protein into smaller polypeptides that, after processing, are presented by the HLA molecules to T cells (Pardi et al. 2018). As of 2 April 2022, two mRNA vaccines, Pfizer-BioNTech's BNT162b2/Comirnaty and Moderna's mRNA-1273, have received emergency validation from WHO. A group of 180 Finnish healthcare workers who had been vaccinated with BNT162b2 showed an induced antibody response against the SARS-CoV-2 spike protein, especially the S1 subunit (Jalkanen et al. 2021).

The latest report by the WHO showed that two non-replicating viral vector (VVnr) vaccines, AstraZeneca's ChAdOx1-S and Janssen's Ad26.COV2.S, have been approved for emergency use against COVID-19. The VVnr platform requires a modified virus as a deliver vehicle, such as adenovirus or poxvirus, to successfully deliver the genetic material into the host cell (Robert-Guroff 2007).

Unlike the replicating viral vector (VVr) vaccine, VVnr cannot produce additional viral antigens because the viral vector lacks the components necessary to infect a new host cell (Robert-Guroff 2007). In addition, the gene of interest does not integrate into the host cell genome after entering the nucleus, where it takes advantage of the host cell machinery to produce the viral mRNA that is subsequently exported into the cytoplasm for translation (Robert-Guroff 2007). Currently, all the VVnr based vaccines against COVID-19 are engineered with the SARS-CoV-2 spike protein, which after intracellular processing is presented in the form of polypeptide fragments on MHC class I and class II molecules (Folegatti et al. 2020; Sadoff et al. 2021; F. C. Zhu et al. 2020).

Protein subunit vaccines are composed of highly purified protein fragments of a pathogen, which have been carefully studied to identify which combinations of these fragments are capable of inducing a strong protective immune response (Foged 2011). The protein subunit vaccine is considered safer than the live attenuated and inactivated vaccines because the formulation process, known as recombinant DNA technology, does not require pathogen handling (Hansson, Nygren, and Ståhl 2000). Protein subunit vaccines are already used against infectious viruses, including hepatitis B and human papillomavirus (Leroux-Roels et al. 2000; Schiller and Lowy 2018; Zhai and Tumban 2016). Moreover, fragments of the S protein, including the S1 subunit, S2 subunit, N-terminal domain (NTD) and receptor binding domain (RBD), have already been used against SARS-CoV and MERS-CoV, suggesting that these fragments can be used as key targets for developing protein subunit vaccines against human coronaviruses (N. Wang et al. 2020). Interestingly, according to the latest WHO report, only one protein subunit vaccine, NVX-CoV2373, has been authorized for emergency use against COVID-19. The NVX-CoV2373 vaccine is comprised of a full-length SARS-CoV-2 spike protein plus Matrix-M adjuvant, and it has been observed that, especially in baboons, NVX-CoV2373 elicits robust humoral and cellular immune responses regardless of disease severity (Tian et al. 2021).

Since the emergence of the global pandemic, there has been an explosion of vaccine development. Currently, around 350 novel vaccine candidates have been developed using several different vaccine platform technologies, of which at least a dozen of vaccines have been granted an emergency use authorization by the WHO. However, the emergence of new SARS-CoV-2 variants that harbor a large number of mutations, especially in the spike protein, may cause resistance at the immunity level against current vaccines. For example, Finnish healthcare workers vaccinated with BNT162b2 showed a decreased antibody response against the Beta variant compared to the Alpha mutant (Jalkanen et al. 2021). Therefore, it is of extreme importance to update vaccines to protect better against the new variants, including Omicron as well as other variants that are expected to appear in the future.

3. Aims of the study

The general aims of this thesis were to apply a variety of sequence-based and structure-based bioinformatics techniques to characterize proteins, mutations and molecular interactions, and exploit the observations towards understanding the biological phenomena, especially as revealed by the experimental studies of our collaborators as well as observations reported in the literature.

3.1. Extracellular citrullination in rheumatoid arthritis patients (Publications I and II)

Our collaborators experimentally detected citrullinated peptides in the enriched ECM proteins from the synovial fluid samples of rheumatoid arthritis patients. In publication I and II, our aim was to map the identified peptides onto known 3D structures and obtain structural insights into how a single heavy atom change could affect the protein function via influencing key molecular interactions at the biomolecular interface. Specific emphasis was placed on the mechanism of TGF-β activation (publication I) and integrin mediated fibronectin assembly (publication II). In publication II, we also aimed to look at sequence information and structural details of the citrullination sites in the X-ray crystal structures of the corresponding proteins to provide insight on structural features controlling the substrate specificity of the PAD2 and PAD4 enzymes.

3.2. The human LEUTX regulates early embryonic development (Publication III)

In this study, the aim was to identify functionally critical amino acid residues within the LEUTX homeodomain involved in DNA binding using biomolecular sequence and structural information. We also aimed to understand how a loss-of-function mutation could affect LEUTX binding with DNA.

3.3. Prediction of potential immunogenic epitopes in SARS-CoV-2 (Publication IV)

We started our investigation in March 2020 with the aim to identify the most likely epitopes from SARS-CoV-2 that would elicit immunogenic responses, and to characterize their physicochemical properties and likely molecular interactions with human leucocyte antigens (HLA). The overall aim was to identify immunogenic epitopes with potential for use as peptide-based vaccines.

4. Materials and methods

4.1. Sequence analyses

Experimentally verified human protein sequences were retrieved from the UniProt database (Bateman et al. 2021) (publications I–IV). Protein sequences encoded by the SARS-CoV-2 genome were retrieved from the NCBI RefSeq database (Brister et al. 2015) (publication IV). Multiple sequence alignment (MSA) was performed using the L-INS-i method in the MAFFT tool (Katoh and Standley 2013) (publications I, III & IV).

In publication IV, linear SARS-CoV-2 epitopes and their binding affinities (expressed as IC_{50} values, nM) with HLA class I allotypes were predicted using the IEDB (Vita et al. 2019) and NetCTL1.2 (Larsen et al. 2007) web servers. The IEDB tool predicts the binding affinities using an integrated model that combines an artificial neural network (ANN), stabilized matrix method (SMM) and combinatorial library (CombLib). The IEDB tool is trained on high-quality experimental data and its prediction accuracy is considered to be better in comparison to other binding-affinity prediction programs. The NetCTL1.2 server integrates scores obtained from three prediction methods: proteasomal cleavage, TAP transport, and peptide-binding to HLA-I molecules.

The predicted 9-mer epitopes from IEDB with an immunogenicity score ≥0.25 (http://tools.iedb.org/immunogenicity/) were compared to those predicted using the NetCTL1.2 server and the common hits from both methods were compared with the experimentally identified eHLA complexes in SARS-CoV; the experimental data was retrieved from the IEDB database (Vita et al. 2019). The binding stability (complex half-life) of the eHLA complex was predicted using the NetMHCstabpan-1.0 web server (Rasmussen et al. 2016). NetMHCstabpan-1.0 predicts the stability of eHLA complexes based on an artificial neural network (ANN) algorithm, which is trained on experimental quantitative stability data covering 75 different HLA class I molecules. The grand average of hydropathy (GRAVY) score for the predicted epitopes was calculated using the Kyte-Doolittle hydropathy index scale (Kyte and Doolittle 1982). Potential transmembrane segments in the membrane proteins were predicted with the TMHMM2.0 tool (Krogh et al. 2001), using a hidden Markov model (HMM) approach. The Jpred4 tool was used to predict secondary structures (α-helix, β-strand and coil) in the non-transmembrane proteins (Drozdetskiy et al. 2015). The strategy and methodology for sequence and structure-based epitope prediction has been described in detail in publication IV.

4.2. Molecular modeling and structural analyses

X-ray crystal structures of biomolecular complexes were obtained from the PDB database (Berman et al. 2000) (publications I–IV). Complexes were visualized in BODIL (Lehtonen et al. 2004) and PyMOL (The PyMOL Molecular Graphics System, Version 2.5 Schrödinger, LLC.) in order to study the atomic interactions,

secondary structures and amino acid residue exposure to solvent (publications I–IV).

The basic protocol of the homology modeling approach consists of the following four processes, i) identifying one or more three-dimensional structures (template) that most closely match the sequence of the protein to be modeled (target), ii) amino acid sequence alignment of the target with the template protein, iii) model building based on the sequence alignment and structural details from the template and protein structures in general, and iv) inspection of the model for gross errors. Protein sequences of unknown 3D structures were searched against the PDB database using blastp at the NCBI database to identify the best matching templates for homology modeling (M. Johnson et al. 2008). The structure-based sequence alignments were performed using Vertaa and Malign in BODIL (Lehtonen et al. 2004). Sequence alignments were manually modified, and the positions of insertions and deletions were determined. The manually curated sequence alignments were used to construct models of the target protein using MODELLER and homodge in BODIL (Lehtonen et al. 2004; Sali and Blundell 1993). Preparation of the models of the homeodomain-DNA interactions in human LEUTX are thoroughly described in publication III.

In publication IV, the X-ray structure of the ternary complex of the T cell receptor (TCR)–HLA–influenza A epitope (PDB ID: 5TEZ (X. Yang et al. 2017)) was used as a template to model the ternary complex of HLA-A*02:01, SARS-CoV-2 spike protein epitope ^{1220}FIAGLIAIV^{1228} and TCR.

The models were visualized to check for side-chain clashes and altered where necessary using the rotamer utility in BODIL. The refined models were taken for energy minimization – the OPLS_2005 force field in the Maestro protein preparation wizard panel (Schrödinger suit) was used to optimize the bond lengths and angles. Structural features of the models were examined for standard acceptable values using the MolProbity web server (C. J. Williams et al. 2018).

4.3. Peptides modeling and molecular docking

In publication II, cyclic peptides NGR and isoDGR were sketched manually by connecting the termini with a dipeptide phenyl-N-methylvaline using the 2D sketcher tool in the Maestro suite. These structures were converted into 3D conformation using the 3D builder tool and the LigPrep wizard was used to optimize the cyclic peptides for docking. The ectodomain structure of integrin αVβ3 (PDB ID: 1L5G (Xiong et al. 2002)) was preprocessed, minimized and refined using the protein preparation wizard. Molecular docking was performed by placing the cyclic peptides within the context of the RGD motif at the integrin αVβ3 binding site using Bodil, followed by energy minimization in Maestro.

In publication IV, the predicted immunogenic 9-mer epitopes (hits) in SARS-CoV-2 were modeled using PyMOL: the side chains of the most similar peptide structures available in the PDB were mutated to match the hits and placed within the cleft between the α1 and α2 helices of HLA allotypes. The docking of epitope-HLA complexes were refined using the Rosetta FlexPepDock web server (London

et al. 2011). This server performs a Monte Carlo-Minimization-based approach to generate models in which peptides have both main-chain and side-chain conformational flexibility, whereas only the side chains are flexible with the protein that binds the peptide. The backbone conformations of the docked epitopes were analyzed using PyMOL.

4.4. Experimental studies

The original publications I, II and III also include details of a wide range of experimental studies that were conducted by our collaborators. Publication IV is solely based on computational analyses made using data from the literature and from publicly accessible databases.

5. Results

5.1. Citrullination of extracellular proteins in the synovial fluid of patients with rheumatoid arthritis (Publications I and II)

In order to identify arginine residues in proteins that have been citrullinated, our collaborators performed mass spectrometry analysis on 40 synovial fluid samples collected from the knee joints of inflammatory arthritis patients; citrullination is typically a cytoplasmic post-translational modification and not expected in the extracellular spaces. A total of 55 citrullinated sites were identified in 24 ECM proteins. We examined the sequences and the 3D structural context of arginine residues in the ECM proteins that were found citrullinated in order to identify features key to enzymatic recognition and citrullination activity. Furthermore, we considered the effect of citrullination on protein-protein interactions for complexes that contribute to the health and repair of the ECM.

Table 3: Arginine residues at the interface of ECM-associated growth factors and their receptors. Table modified from publication I.

Growth factor	Growth factor receptor	PDB ID	Arginine residues at the interface
TGF-β1	TGFBR1 and TGFBR2	3KFD	R25, R94
TGF-β3	TGFBR1 and TGFBR2	2PJY	R25, R94
FGF	FGFR1-FGF1	1EVT	R35, R37, R88
	FGFR1-FGF2	1CVS	R44, R60, R97, R109
	FGFR2-FGF1	1DJS	R35, R37, R88
	FGFR2-FGF2	1EV2	R44, R60, R97, R109, R120
	FGFR2B-FGF10	1NUN	R78, R80
	FGFR2C-FGF8B	2FDB	R48, R52, R59, R155, R177
	FGFR3C-FGF1	1RY7	R50, R52, R103
LIF	LIFR-IL6R	1PVH	R15, R123
PDGF2	PDGFR	3MJG	R27, R56, R73
TNF-α	TNFR1B	3ALQ	R31, R32
IL-2	IL2R	2B5I	R38, R81, R120
IL-7	IL7R	3DI2	R64

5.1.1. Interfaces between ECM associated growth factors and their receptors contain at least one arginine

The examination of the protein complexes formed between ECM associated growth factors and their receptors revealed that one or more arginine residues play a crucial role (Figure 1B, publication I). All the growth factor-receptor complexes where 3D structures were available were identified to have at least one arginine residue at the protein-protein interface (Table 3). Interestingly, about half of the complexes have more than two interfacial arginine residues. It is likely that disruption of electrostatic and hydrogen bonding interactions due to citrullination affects binding, altering the structure and function in comparison to the wild-type complex.

5.1.2. Solvent-exposed arginine residues are a primary target for the PAD4 enzyme

Our structural analysis of the arginine residues in X-ray crystal structures corresponding to the observed 55 citrullination sites seen in the 40 synovial fluid samples revealed that the vast majority of the arginine residues have a high tendency to be exposed to solvent (Figure 2A and J, publication II).

The early study by Arita and colleagues (Arita et al. 2006) proposed that arginine residues citrullinated by the PAD4 enzyme would be located on a flexible unstructured peptide, which changes conformation to a β-turn-like bent upon binding to the catalytic site of PAD4. This structural motif was proposed based on three X-ray crystal structures (PDB IDs: 2DEW, 2DEX and 2DEY (Arita et al. 2006)) solved by the authors. In contrast, our structural analysis of the proteins that have been identified as substrates of PAD4 revealed that although arginine residues are predominantly located on loops, nearly 50% of the citrullination sites were found on α-helices, β-strands and β-turns (Figure 2B, publication II). In particular, with fibronectin whose domains are composed of antiparallel β-strands, about 20% of the solvent-exposed arginines in fibronectin are located on loops (Table S4, publication II) but most of the citrullinated sites are located on β-strands (Figure 2K, publication II). In fibrinogen, the partially-exposed R104 is located on an α-helix, whereas R230, also partially-exposed in cathepsin G, is located on a β-strand (Figure 14). Our analysis showed that the vast majority of arginine side chains that are citrullinated are highly exposed to solvent, suggesting that in the few cases of buried or partially exposed arginines the substrate arginine must become solvent exposed before conversion into citrulline; this is also in agreement with our analysis of the PAD4 structures and the motifs that are recognized.

A

Protein	Citrullinated peptide sequence	PDB ID
Angiogenin	GNPH(Cit)ENLR	1H52
Apolipoprotein E	AIRE(Cit)LGPL	2L7B
Cathepsin G	EVFT(Cit)VSSF	1AU8
Neutrophil elastase	AQGR(Cit)LGNG	2Z7F
Fibrillin-1	RAKR(Cit)GGGG	2M74
Fibrinogen	SANN(Cit)DNTY	3GHG
Fibronectin	SEGR(Cit)DNMK	3M7P
Prothrombin	GYKG(Cit)VTGW	4NZQ

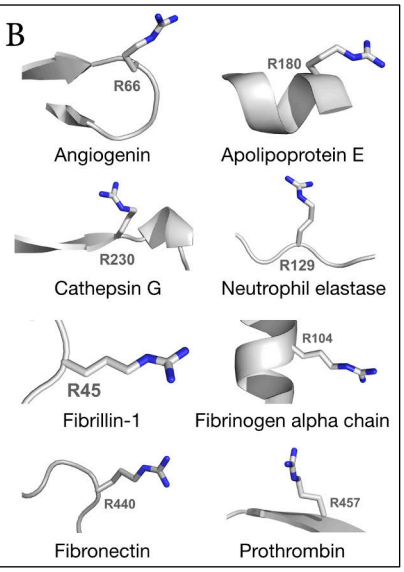


Figure 14: A few examples of extracellular proteins that were found citrullinated in the synovium of patients with RA. A) Amino acid residues surrounding the citrulline residue. B) Location of the citrullinated arginine residues on different secondary structures.

Based on the X-ray crystal structures of PAD4 in complex with histone peptides (PDB IDs: 2DEW, 2DEX and 2DEY), the authors reported that two additional amino acids on either sides of the citrullinated arginine residue are bound within the active site of the PAD4 enzyme (Arita et al. 2006). In addition to the side-chain of arginine that penetrates deeply into the PAD4 active site (Figure 15A), the main-chain atoms of residues at positions N-2, N-1 and N (numbering relative to arginine) hydrogen bond with the active site residues of PAD4 (Figure 15B-C), but the lack of side-chain interactions at N-1, N+1 and N+2 positions provide no specificity related to the type of amino acid involved. The side-chain of residue at the N-2 position, however, does hydrogen bond with the PAD4 active site (Figure 15C) – the authors suggested PAD4 specificity for residues that are small and polar (Arita et al. 2006). In contrast, our more extensive analysis of the sequences and structures show that the N-2 position can be occupied by almost any amino acid side-chain and does not show specificity for amino acids with a small side chain (Figure 2C and I, publication II). Thus, our findings indicate that there is no distinctive consensus sequence or structural motif that would lead to the citrullination of the substrate arginine, and that the arginine only needs to become exposed to solvent.

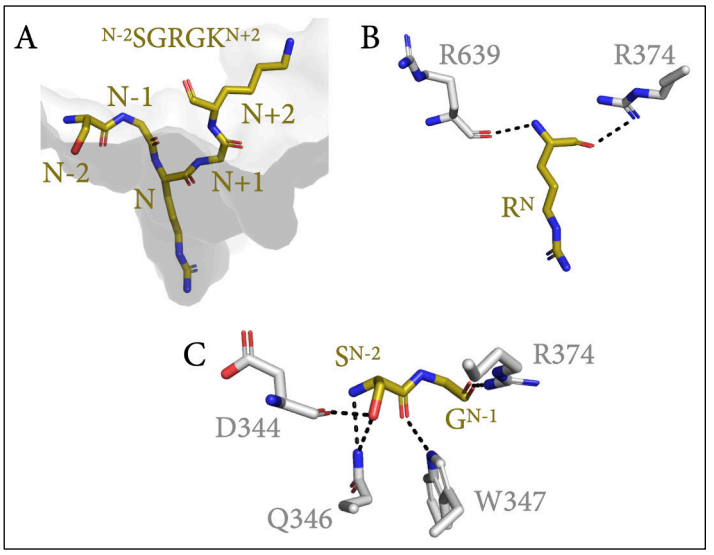


Figure 15: Hydrogen-bonding interactions (black dashed lines) between a histone peptide (amino acid sequence: SGRGK) and amino acid residues within the catalytic site cleft of PAD4 enzyme (PDB ID: 2DEW). A) Positions N-2, N-1, N, N+1 and N+2 are assigned to the five residues of the histone peptide, where "N" represents the position of the substrate arginine. B) The main-chain carbonyl oxygen and nitrogen atoms of R^N hydrogen bond with R374 and R639 of PAD4. C) The other nitrogen atom of the guanidinium group of R374 interacts with the carbonyl oxygen of the G^{N-1} residue. The most conserved residue, W347 in PAD4, hydrogen bonds with the main-chain carbonyl oxygen of S^{N-2}, while the residues D344 and Q346 in PAD4 hydrogen bond with the side chain of S^{N-2}. In addition to the side chain of R^N, the side-chain oxygen of S^{N-2} hydrogen bonds with Q346, but this is not a conserved feature of PAD4 substrate recognition.

5.1.3. Citrullination of isoDGR motif in fibronectin reduces its binding with integrin αVβ3

Among the experimentally identified citrullinated peptides, three arginine residues have specific roles in integrin-ECM interactions, namely R572 within the RGD tripeptide motif of fibrinogen, R285 within the GAOGER motif (where O is hydroxyproline) of the collagen α1(III) chain and R234 within the NGR tripeptide motif of fibronectin (FN). The FN case is especially interesting and we performed a detailed structural analysis to examine the potential effect of citrullinated FN on its ability to coordinate integrin interactions.

The intact NGR motif in FN does not bind to integrin αVβ3 (Curnis et al. 2006; Spitaleri et al. 2008). However, this motif undergoes a non-enzymatic and spontaneous post-translational modification into the isoDGR motif, in which the main-chain carbonyl group of asparagine becomes a negatively charged carboxylate side chain of isoaspartate (isoD) and the amide side chain of asparagine is then incorporated into the main chain of FN (Curnis et al. 2006, 2010). Previous studies showed that the isoDGR motif binds to integrin αVβ3 in the reverse main-chain orientation in comparison to the binding of the classical

RGD motif (Curnis et al. 2006; Spitaleri et al. 2008). The negatively charged carboxylate side chain in both RGD and isoDGR peptides makes electrostatic interactions with the cationic metal ion dependent adhesion site (MIDAS) in the β I-like domain of the integrin β3 subunit, and the arginine in either case interacts with the side chains of aspartate and tyrosine residues in the β-propeller domain of the integrin αV subunit (Figure 6C, publication II). The most probable cause for the failure of NGR binding is the absence of a negative charge, which appears to be essential for binding with the positively charged metal ion (Figure 6A, publication II). In addition, the backbone of NGR is less flexible than isoDGR because the isoD residue inputs a methylene (-CH_2-) group into the main chain that provides additional flexibility and makes isoDGR a suitable integrin binding site.

A citrullinated 30 kDa FN fragment, in which the positive charge on isoDGR is neutralized, showed markedly reduced binding to integrin αVβ3 (Figure 5A, publication II). The loss of the positive charge on arginine would disrupt the strong electrostatic interaction between isoDGR and the negatively charged residues on the integrin αV subunit (Figure 6D, publication II).

5.1.4. Citrullination interrupts the integrin-mediated TGF-β1 growth factor activation pathway

TGF-β is an important growth factor and cytokine that has various immunosuppressive properties in chronic arthritis (M. O. Li et al. 2006). Since knockdown of PAD4 has been observed to induce TGF-β signaling (Stadler et al. 2013), we investigated the effect of citrullination on the integrin-mediated TGF-β activation pathway.

β3-LAP of TGF-β3 contains a RGD motif on an exposed loop that is recognized by integrin αVβ6 (PDB ID: 4UM9 (X. Dong et al. 2014)) where the motif serves to bridge the β-propeller domain of the αV subunit and the β I-like domain of the β6 subunit. Similar to the RGD motif in FN, the negatively charged aspartate forms a strong ionic interaction with the cation at MIDAS on the β I-like domain; and the positively-charged arginine side chain of the peptide interacts with the negatively charged aspartate in the β-propeller domain (Figure 16B). Based on the multiple sequence alignment of the human proprotein of TGF-β3 (UniProt ID: P10600) with that of TGF-β1 (UniProt ID: P01137) and TGF-β2 (UniProt ID: P61812) (Figure 16A), the RGD motif of the β1-LAP of TGF-β1 has potential to bind integrin αVβ6. Interestingly, a recent structural study proved that the binding of β1-LAP to integrin αVβ6 is very similar to that of β3-LAP (Figure 16C) (Campbell et al. 2020). Since the human β2-LAP lacks an RGD motif where arginine is replaced with the amino acid serine, we speculate that the activation of TGF-β2 might not be regulated by integrins.

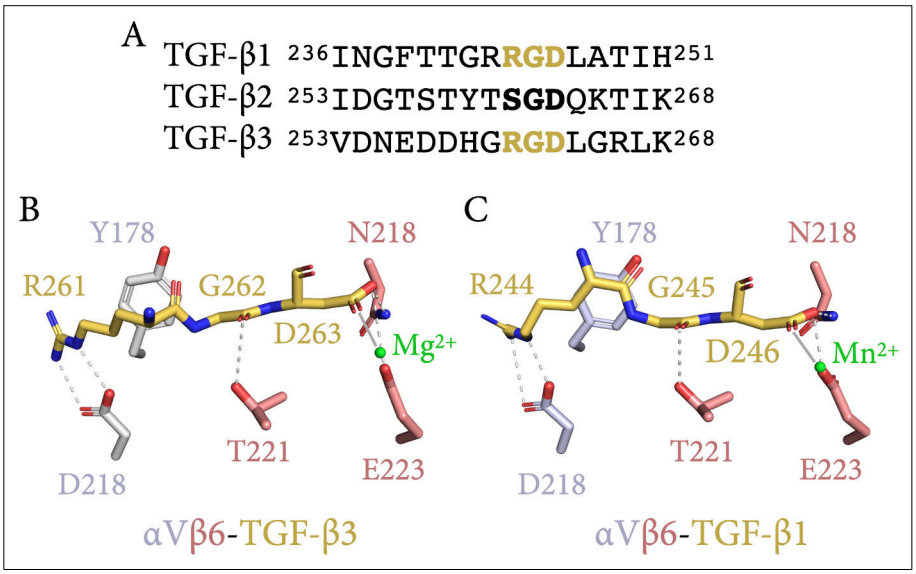


Figure 16: The RGD peptide in human TGF-β and its binding to human integrin αVβ6. A) The latency-associated peptide (LAP) of TGF-β1 and TGF-β3 contains a RGD peptide, which is absent in the LAP of TGF-β2. B) Binding of the RGD peptide from TGF-β3 to integrin αVβ6. C) Binding of the RGD peptide from TGF-β1 to integrin αVβ6. RGD peptide, gold carbon atoms; residues of integrin αV, gray carbon atoms; residues of integrin β6; pink carbon atoms; dashed gray lines are potential hydrogen bonds, a solid gray line connects the negatively charged aspartate of the RGD peptide to the cation at MIDAS (green sphere) on the integrin β6 subunit.

Citrullination of arginine within the RGD motif leads to the loss of the strong ionic interaction with the negatively charged aspartate residue in the β-propeller domain of the αV subunit (Figure 4D, publication I). Our structural analysis is in complete agreement with the experimental observations in which citrullination of the RGD site significantly reduced the binding of the β1-LAP protein with integrin αVβ6 (Figure 4B, publication I). Since the side chain of citrulline is stabilized by both hydrophobic and hydrogen bonding interactions with the αV subunit, a complete elimination of the binding of CitGD to integrin αVβ6 might not be possible. These findings indicate that citrullination would likely reduce the integrin mediated activation of both TGF-β1 and TGF-β3.

5.1.5. Citrullination interrupts active TGF-β1 binding to the TGF-βRII receptor

X-ray crystal structures are available for the human TGF-βRI/TGF-βRII complex bound to human TGF-β1 (PDB ID: 3KFD (Radaev et al. 2010)) and TGF-β3 (PDB ID: 2PJY (Groppe et al. 2008)). In each ternary complex, two arginine residues from the TGF-β1 and TGF-β3 growth factors make strong ionic interactions with amino acids that have acidic side chains present on TGF-βRII (Figure 6B in

publication I shows interactions between TGF-β1 and TGF-βRII). Experimental observations from our collaborators showed that PAD2 mediated citrullination of human TGF-β1 significantly reduced the binding of TGF-β1 with TGF-βRII (Figure 6A, publication I). Elimination of these strong ionic interactions due to citrullination may destabilize the TGF-β1/TGF-βRII complex (Figure 6C, publication I), largely contributing to the loss of binding affinity since the hydrogen bonding capacity would likely remain unchanged and a complete loss of binding would be unlikely. Structural analysis of human TGF-β2 (PDB ID: 2TGI (Daopin et al. 1992)) showed that the two arginine residues in TGF-β2 are replaced by lysine residues. Thus, our findings suggest that while TGF-β1 and TGF-β3 are likely to be affected by citrullination, TGF-β2 should not be.

5.2. Regulatory function of the PRD-like homeodomain LEUTX in early embryonic development (Publication III)

Our collaborators identified a family of transcription factors – PRD-like homeobox proteins – that function during the very earliest cell divisions in human embryos, including LEUTX and several other family members that recognize a 36 bp DNA motif; a luciferase reporter assay was used for identification (Jouhilahti et al. 2016; Töhönen et al. 2015).

In order to examine the effect of each PRD-like HD and exclude the involvement of other non-PRD-like TFs binding to this motif, our collaborators engineered a luciferase reporter construct with an 11 bp motif containing the predicted LEUTX binding site TAATCC. LEUTX, OTX2 and TPRX1 strongly activated the reporter gene, while DPRX showed downregulation, and ARGFX, CPHX1 and CPHX2 showed no significant effect (Figure 1A and C, publication III). Previous *in vitro* studies reported that HDs with a lysine residue at position 50 show a strong preference for the canonical TAATCC motif (Noyes et al. 2008; Tucker-Kellogg et al. 1997) and sequence alignment of the PRD-like HDs revealed that human LEUTX, OTX2, TPRX1 and DPRX have residue K50, while CPHX1 and CPHX2 have Q50, and ARGFX has R50 (Figure 1D, publication III). Therefore, the non-K50 type HDs are not expected to bind to the TAATCC motif with the same affinity as the K50 HDs.

5.2.1. Specificity determining residues of LEUTX recognize the major and minor grooves of the DNA motif TAATCC

To identify the important determinants that play key roles for HD-DNA interactions in LEUTX, we examined the X-ray crystal structures of HD-DNA complexes. Residues at positions 2, 3 and 5 on the N-terminal arm and at positions 47, 50, 51, 54 and 55 on the recognition helix of the HDs have been shown to contribute to the DNA binding specificity (Noyes et al. 2008). Since, no experimental structure of LEUTX existed, we prepared a 3D structural model of the human LEUTX-HD in complex with the dsDNA motif $^{5'}$TAATCC$^{3'}$ using the homology modeling approach; note that the motif includes the complementary

strand. Positively charged residues, R53, K55 and K57 in the LEUTX recognition helix, have the potential to form electrostatic interactions with the negatively charged phosphate backbone of duplex DNA, which might regulate the scanning mechanism of LEUTX-HD recognition of its target sequence. Residues K50, N51 and R58 in the LEUTX recognition helix, when inserted into the DNA major groove, form a network of direct and water mediated hydrogen bonds with the TAATCC motif (Figure 4B, D and F, publication III). In both primates and rodents, position 47 is mostly occupied by nonpolar residues (Figure 2, publication III); I47 in the human LEUTX model would stabilize the HD-DNA interface where it forms hydrophobic contacts with the methyl group of the thymine base (Figure 4B and F, publication III).

Generally, the recognition helix of an HD is responsible for interacting with DNA at the major groove, while the flexible N-terminal arm facilitates the search for and binding to the adjacent minor groove (Vuzman, Azia, and Levy 2010). In the LEUTX model, residues R2/R3 and R5 in the N-terminal arm would bind directly to nucleotide bases in the DNA minor groove (Figure 4B, publication III). Among all three arginines, R5 is the most conserved residue observed across eutherian species (Figure S1, publication III) and is involved in direct contact with the TAATCC motif; whereas, the second most conserved arginine, R2, forms a network of direct and water-mediated hydrogen bonds (Figure 4B, publication III). The details of molecular interactions are described in publication III.

5.2.2. Loss-of-function mutation A54V reduces LEUTX binding with the DNA motif, whereas I47T is compensatory

Our collaborators examined seven available human genotype resources for LEUTX variants that could potentially affect the function of LEUTX. Interestingly, an individual in the 1000 Genome data (Auton et al. 2015) was identified with a single heterozygous missense mutation A54V in the recognition helix of the LEUTX homeodomain; however, this rare allele was not passed on from father to son. In the LEUTX recognition helix, the absence of the I47-V54 combination in the inter-species comparison of the amino acid sequences from eutherian mammals (Figure 2, publication III), raises the possibility that the I47-A54V combination in LEUTX would lead to compromised DNA binding.

Our structural analysis showed that residue A54 in both the LEUTX model and the engrailed homeodomain template structure (PDB ID: 2HDD (Tucker-Kellogg et al. 1997)) faces the DNA motif: the small side chain of alanine provides more free space for water molecules to form a network of water mediated hydrogen bonds linking other specificity determining residues that are essential for DNA motif recognition. The substitution of alanine by a bulkier valine residue at position 54 would likely displace the interfacial water molecules (Figure 4F and H, publication III); disruption of the water mediated hydrogen-bonding network is the most plausible cause for the poor binding of the A54V mutant. These structure-based predictions are in good agreement with the experimental data (Figure 4L-N, publication III).

In contrast to the I47-A54V residue combination, the single I47T mutation and the double I47T-A54V mutation in the LEUTX recognition helix showed no change in the luciferase reporter assay activity compared to wild-type (Figure 4L-N, publication III). Interestingly, both residue combination, T47-A54 and T47-V54, were found in the inter-species sequence alignment of the LEUTX recognition helix (Figure 2, publication III). In the human LEUTX model, the substitution of I47 by a more compact but polar T47 residue provides extra potential to stabilize interaction with the TAATCC motif (Figure 4G and I, publication III). The interactions mediated by the side chain of residue T47 are most likely sufficient to compensate for any interruption of the hydrogen-bonding network due to the A54V mutation.

5.2.3. The C-terminal Leutx domain may have transcriptional regulatory properties

Sequence analysis revealed that the LEUTX protein is comprised of two distinct domains: the N-terminal HD, and the C-terminal Leutx domain of about 110 amino acid residues (Figure 17). Our sequence analysis revealed that the human Leutx domain consists of a long disordered region of about 75 amino acids followed by three small α-helices.

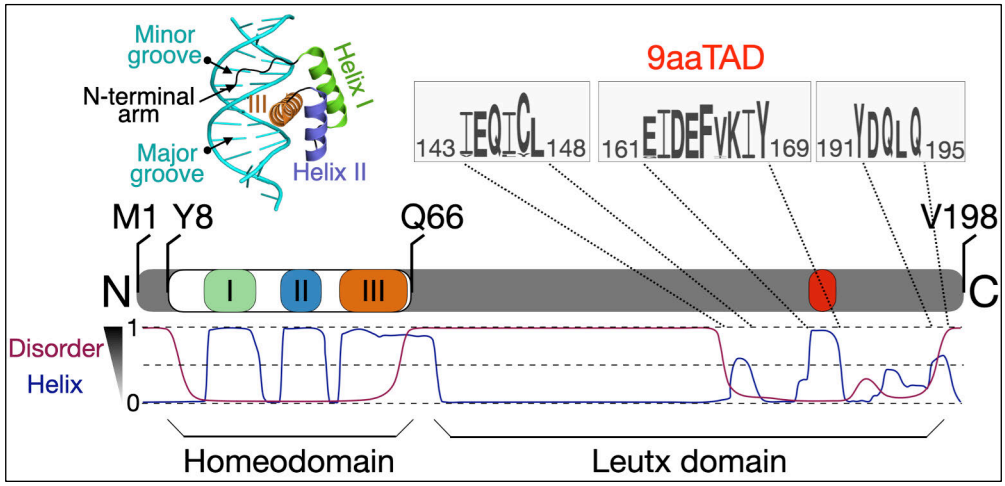


Figure 17: The *LEUTX* gene encodes the LEUTX protein that consists of the N-terminal homeodomain (HD) and the C-terminal Leutx domain. The LEUTX-HD folds into three α-helices, of which the latter two (helix II and helix III) form a helix-turn-helix structural motif. Helix III (orange), known as the recognition helix, forms key molecular interactions with DNA bases in the major groove (cyan), while the flexible N-terminal arm preceding helix I inserts itself into the DNA minor groove (cyan). In the Leutx domain, three helical regions (blue curve) are predicted within the structurally ordered region (brown curve). The second predicted helix (red box) has potential to function as a nine-amino-acid transactivation domain (9aaTAD) found in several transcription factors reported to interact with the KIX domain of the CREB-binding protein (CBP).

One of the α-helices, ^{161}EIDEFVKIY^{169} rich in hydrophobic and acidic amino acids, was predicted as a likely nine-amino-acid transactivation domain (9aaTAD), which could mediate interactions between the LEUTX transcription factor and transcriptional co-activator proteins. Interestingly, the 9aaTAD motif is absent in DPRX, which could explain why DPRX functions differently than LEUTX in the luciferase reporter assays. Recently, we identified the 9aaTAD motif in the C-terminal region of the human DUX4 transcription factor and experimentally showed its binding with the KIX domain of CREB-binding protein (Vuoristo et al. 2022).

5.3. Prediction of SARS-CoV-2 vaccine epitopes to elicit T cell-mediated immunity (Publication IV)

In order to help us understand COVID-19 disease progression, we performed a thorough literature search to understand the process of T cell-mediated destruction of virus-infected cells. In the infected cell, the cytosolically derived viral peptide fragments bind to the MHC class I (MHC-I) protein molecules assembled within the endoplasmic reticulum and are subsequently translocated to the cell surface for recognition by the T cell receptor (TCR) of the cytotoxic T cells, ultimately leading to the elimination of the infected cells. Since epitopes from SARS-CoV proteins have already been shown to stimulate a strong cytotoxic T cell immune response against the infected host cells (Kohyama et al. 2009; Tsao et al. 2006), we performed computational analyses to narrow down the MHC-I specific epitopes that would likely elicit an effective cytotoxic T cell response to SARS-CoV-2 infection.

5.3.1. MHC class I allotypes should bind specifically to 9-mer epitopes derived from structural and non-structural SARS-CoV-2 proteins

We first predicted the binding affinity (expressed as IC_{50}) of all possible linear 8- to 11-mer epitopes derived from the SARS-CoV-2 proteins (Table 1, publication IV) to HLA-A and HLA-B allotypes (Table S1A, publication IV for the list of allotypes), using the immune epitope database (IEDB) web server (Vita et al. 2019). Based on the predicted binding affinity scores, the epitope-HLA (eHLA) complexes were classified into three different groups: strong binders ($IC_{50} \leq 50$ nM), weak binders (50 nM < $IC_{50} \leq 500$ nM) and non-binders ($IC_{50} > 500$ nM). Comparison of the top one percentile of highest scoring eHLA complexes showed that the 9- and 10-mer epitopes have, on average, a higher predicted binding affinity with the HLA allotypes than do either the 8- or 11-mer epitopes (Figure 1A, publication IV). More specifically, there were about 52% more 9-mer epitopes predicted to bind to HLAs in comparison with 10-mer epitopes (Figure 1B, publication IV). Thus, the top one percentile of 9-mer epitopes were considered for further analysis.

The top one percentile 9-mer eHLA complexes ($IC_{50} \leq 50$ nM) with an immunogenicity score ≥ 0.25 (Vita et al. 2019) were compared to those predicted by the NetCTL1.2 web server using default parameters (Larsen et al. 2007). The eHLA complexes identified from this combined approach showed that the majority of epitopes (~68%) are derived from nsp3, nsp4, spike and membrane proteins and are predicted to elicit a strong immune response for the elimination of infected host cells (Figure 1C, publication IV).

5.3.2. Comparison of the *in silico* predicted SARS-CoV-2 epitopes with the experimentally validated SARS-CoV epitopes

The predicted 9-mer epitopes were then selected for comparison with the *experimentally* identified epitopes of SARS-CoV strains (data retrieved from the IEDB database). We identified 29 eHLA complexes that are common to both SARS-CoV and SARS-CoV-2 (Table S8, publication IV) and a total of six epitopes were identified that interact very strongly (experimental IC_{50} from 0.23 to 24.6 nM) with both the HLA-A*02:01 and HLA-A*02:06 allotypes (Table 3, publication IV). These immunogenic epitopes that bind strongly ($IC_{50} \leq 50$ nM) to HLAs were selected for further analysis.

Since mutations at certain positions in an immunogenic epitope interfere with peptide binding to HLAs (Falk et al. 1991), we examined the evolutionary conservation of the epitopes by performing a blastp search against the non-redundant sequence database of SARS-CoV-2. Five epitopes were found to have 100% sequence identity with the proteins of SARS-CoV-2 (Table 3, publication IV), suggesting that these fully conserved peptide sequences also bind to HLAs and activate the cytotoxic T cell response against SARS-CoV-2.

5.3.3. Sequence properties that regulate the efficiency of epitope presentation to stimulate an effective immune response

In order to gain insight into understanding the sequence properties of the predicted immunogenic epitopes that bind to MHC-I molecules, we analyzed the experimentally known complexes of HLA-A*02 allotypes bound to 9-mer epitopes from a wide range of pathogens and retrieved from the IEDB database. We first examined the enrichment of hydrophobic residues in the experimentally identified epitopes using the Kyte-Doolittle grand average of hydropathicity (GRAVY) scale (Kyte and Doolittle 1982). Our analysis showed that the immunogenic epitopes ($IC_{50} \leq 50$ nM) are more enriched in hydrophobic residues – particularly aromatic residues – in comparison to the non-immunogenic epitopes ($IC_{50} > 500$ nM) (Figure 18A) and about 67% of the immunogenic epitopes do not contain charged residues (Figure 18C). The average hydropathy scores at positions 1, 3, 5, 6 and 7 in the immunogenic epitopes are greater than 0.5, suggesting that these positions are likely occupied by hydrophobic residues (Figure 18B), and this agrees with published results (Chowell et al. 2015). Similarly, our analysis of the *in silico* identified SARS-CoV-2-derived eHLA complexes showed that about 55% of the

immunogenic epitopes have a GRAVY score >1 (*i.e.*, enriched in hydrophobic residues), whereas 19% are found in the non-immunogenic epitopes (Table S8, publication IV), which partly explains why the most potent immunogenic epitopes are derived from either hydrophobic or transmembrane regions of SARS-CoV-2 proteins.

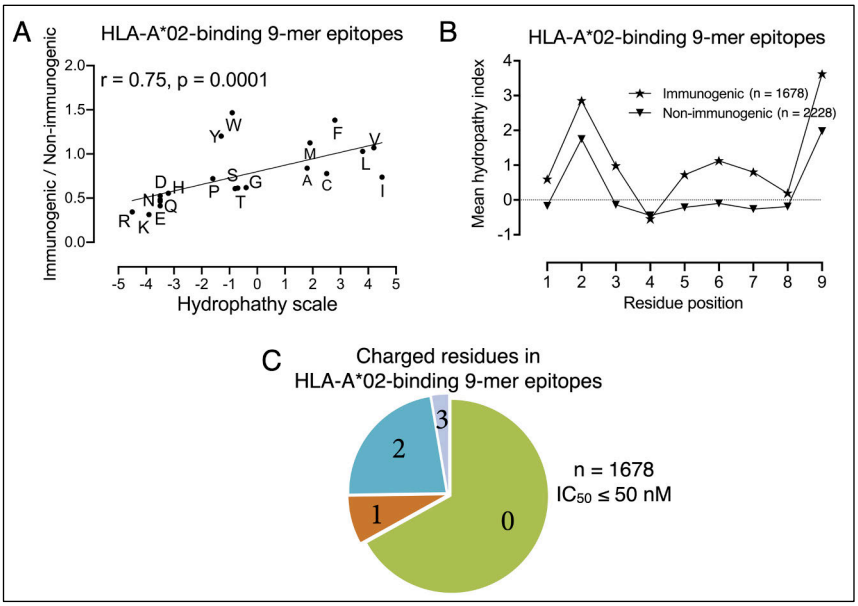


Figure 18: Comparative analysis of chemical properties and composition of amino acid residues in experimentally identified HLA-A*02-binding 9-mer epitopes. A) Propensity, immunogenic to non-immunogenic frequency ratio, of each amino acid residue as a function of its corresponding hydropathy index value. Amino acids with aromatic side chains show a high propensity for immunogenic epitopes, while charged residues show a high propensity for non-immunogenic epitopes. B) Comparison of positional hydropathy indices of the immunogenic ($IC_{50} \leq 50$ nM) and non-immunogenic ($IC_{50} > 500$ nM) HLA-A*02-binding 9-mer epitopes. The hydropathy index value was assigned to each amino acid for the calculation of the position-wise mean hydropathy index. C) Number of charged residues within the HLA-A*02-specific immunogenic epitopes. Data were plotted using GraphPad Prism (GraphPad software, version 8.0.0)

In addition to hydrophobicity, the immunogenicity of epitopes also depends on the stability (expressed as half-life) of the cell-surface eMHC-I complex (Harndahl et al. 2012). Again, we turned to examples from a wide variety of pathogens within the IEDB database, in this instance examining the half-life of each of the experimentally known epitope-HLA-A*02 complex (n = 6500) using the NetMHCstabpan1.0 web server (Rasmussen et al. 2016). This analysis revealed that the predicted half-life of the immunogenic ($IC_{50} \leq 50$ nM) eHLA-A*02:01 complexes is about four-fold higher than the non-immunogenic ($IC_{50} >$ 500 nM) complexes (Figure S1A, publication IV). Moreover, comparison within the immunogenic eHLA-A*02 complexes showed that the half-lives for eHLA-

A*02:01 complexes are about two-fold higher than the eHLA-A*02:06 complexes (Figure S1A, publication IV).

We obtained similar results from our analysis of the predicted SARS-CoV-2-derived eMHC-I complexes: both the HLA-A*02:01 and HLA-A*02:06 allotypes have longer predicted half-lives with the immunogenic epitopes in comparison to other HLAs (Table S8, publication IV). The eHLA-A*02:01 complexes were similarly observed to have a predicted half-life of more than two-fold in comparison to the eHLA-A*02:06 complexes (Table S8 and 4, publication IV).

5.3.4. Structural properties of the epitope-HLA (eHLA) complexes defining T cell receptor (TCR) recognition

To identify the important structural features that determine molecular recognition between the 9-mer epitopes from SARS-CoV-2 and MHC-I molecules, we docked the predicted top immunogenic epitopes (Table 4, publication IV) to the peptide-binding cleft that is formed between the α1 and α2 domains of the HLA-A*02:01 (PDB ID: 5TEZ (X. Yang et al. 2017)) and HLA-A*02:06 (PDB ID: 3OXR (Jingxian Liu, Chen, and Ren 2011)) allotypes. The docked backbone conformations (over Cα atoms) of the predicted epitopes matched well that observed in the 3D structure of the influenza A virus epitope ^{1}GILGFVFTL^{9} complexed with HLA-A*02:01 (PDB ID: 5TEZ) (Figure S1C-D, publication IV). More specifically, the residues at positions 1, 2, 3 and 9 of the docked epitope ^{1220}FIAGLIAIV^{1228} (S protein) are completely buried within the peptide-binding cleft of HLA-A*02:01 (Figure 2C-D, publication IV), providing structural constraints to the N- and C-termini of the epitope. However, the partially solvent-exposed residues at positions 4-8 of the docked epitope ^{1220}FIAGLIAIV^{1228} make hydrophobic interactions with the solvent-exposed residues of the α1 and α2 domains of HLA-A*02:01 (Figure 2C and E, publication IV).

The published atomic structure of the HLA-A*02:01-^{1}GILGFVFTL^{9}-TCR complex (PDB ID: 5TEZ) shows that the ^{1}GILGFVFTL^{9} epitope binds to TCR through the partially solvent-exposed residues. To understand the molecular basis of TCR binding to the SARS-CoV-2-derived eMHC-I complex, we superposed the docked HLA-A*02:01-^{1220}FIAGLIAIV^{1228} complex onto the X-ray crystal structure of HLA-A*02:01-^{1}GILGFVFTL^{9}-TCR complex and the coordinates of TCR were then extracted to generate a complete model of the HLA-A*02:01-^{1220}FIAGLIAIV^{1228}-TCR complex (Figure 3B (at t = 0 ns), publication IV). Visual analysis of the modeled ternary complex suggests that loops on the α chain and β chain of the TCR molecule play an essential role in recognition of the eMHC-I complex (Figure 2D, publication IV). More specifically, residues on the CDR3α and CDR3β loops could be important for recognizing the eHLA complex likely due to their direct contacts with the side chains of residues from the epitope ^{1220}FIAGLIAIV^{1228} and from HLA-A*02:01 (Figure 2E, publication IV). The stability of the HLA-A*02:01-^{1220}FIAGLIAIV^{1228}-TCR complex was evaluated by monitoring the root-mean-square deviation (RMSD) of the Cα atoms during

molecular dynamics (MD) simulations (Figure 3A-B, publication IV). The results from MD simulations are in agreement with the predictions that we made based on our molecular modeling approach.

6. Discussion

In this thesis, we primarily focused on *in silico* predictions, but we were also highly dependent on different levels of support from collaborative experimental data and published literature. Each of the three projects were challenging due to a lack of experimental information and therefore different computational tools were used, and the results were validated as much as possible using external data.

6.1. Citrullination of extracellular proteins in synovial fluid from patients with RA (Publications I and II)

In publications I and II, as a basis for our *in silico* study, we used the experimentally identified citrullinated peptide sequences from the ECM proteins that were found enriched in the synovial fluid samples of patients with RA. In order to understand the potential role of citrullination on protein structure and function, the citrullinated peptides were mapped onto the experimentally determined 3D structures of the ECM proteins.

In publication I, our in-depth visual inspection of the mapped structures revealed that most of the growth factors (GFs) contain at least one arginine residue at the GF-GF receptor interface. Citrullination of these GFs would most likely alter their molecular functions similar to that seen for the citrullinated TNF-α (Moelants et al. 2013) because citrullination changes the chemical characteristics of the arginine residue, including the charge and hydrogen-bonding potential. The interfacial arginine residues have been observed to play a crucial role in TGF-β1 activation in which the RGD motif of TGFβ1-LAP first binds to the ectodomain of integrin αVβ1, αVβ5, αVβ6, and αVβ8 for the release of activated TGF-β1 (X. Dong et al. 2014; Reed et al. 2015; Tatler and Jenkins 2012), and later, the released TGF-β1 binds to TGFβRII via two arginines (Groppe et al. 2008; Radaev et al. 2010). Our in-depth study demonstrated that citrullination of these arginines disrupts the biological function of TGF-β1 by inhibiting the integrin αVβ6 mediated TGF-β1 activation and by blocking the binding of TGF-β1 to its own cell surface receptor TGFβRII. Since the LAP of TGF-β2 does not contain an RGD motif, its activation is probably completely independent of an integrin receptor, unlike that of TGF-β1 and TGF-β3. Binding of the RGD peptide from TGFβ3-LAP to integrin αVβ6 (PDB ID: 4UM9 (X. Dong et al. 2014)) indicates that citrullination might affect its binding to the integrin receptor similar to that of TGFβ1-LAP. Furthermore, residues equivalent to the two arginines in the activated form of TGF-β1 and TGF-β3 are replaced by lysine residues in TGF-β2, suggesting that citrullination might not have any direct impact on the binding of TGF-β2 to TGFβRII.

This study is the first to demonstrate the effect of citrullination on ECM associated GFs, despite the fact that extracellular citrullination has been associated with inflammatory diseases (Uysal et al. 2010). The extracellular

citrullination activity of PAD2 and PAD4 enzymes is clearly making the RA disease more active and worse, since the citrullinated arginine residues in RA patients were frequently found at the interfaces of proteins critical for the ECM and its repair.

PAD enzymes lack the classical secretion signal sequence, suggesting they function to citrullinate intracellular proteins only. In RA, however, PAD enzymes apparently leak out from the cytoplasm into the ECM space. The exact cause of extracellular citrullination activity is not clear yet but various types of cellular stress stimuli can induce apoptotic or necrotic cell death, in which the lysed cells release the PAD enzymes into the calcium-rich extracellular space. Interestingly, the detection of PAD-mediated citrullination activity in other inflammatory joint diseases suggests that citrullination is not an RA-specific phenomenon (Makrygiannakis et al. 2006). However, the anti-citrullinated protein antibodies (ACPAs) in the synovial joints are highly specific for RA (Schellekens et al. 1998); and because of that, the anti-cyclic citrullinated peptide (CCP) antibody is widely used as a diagnostic biomarker of RA for the detection of ACPAs (Aggarwal et al. 2009).

In publication II, our in-depth visual inspection of the published 3D structures that were mapped with the experimentally identified citrullinated peptides revealed that about 20% of the citrullinated arginines have a potential role in protein-protein interactions. The importance of arginine residues for protein-protein interactions relevant to inflammation and tissue remodeling was shown through engineered alanine mutations of arginine residues in angiogenin (Shapiro and Vallee 1992). Mutation of R5, R33 and R66 led to significantly reduced blood vessel formation (Shapiro and Vallee 1992).

The classical integrin-binding RGD motif in fibronectin (FN) was not found citrullinated even with a high concentration of PAD2 and PAD4 enzymes, suggesting that the side chain of arginine in RGD is not accessible to the catalytic site of the PAD enzymes. Surprisingly, R234 in another integrin-binding NGR/isoDGR motif, also located in FN, was found citrullinated in inflamed synovial fluid. Experimental studies from our collaborators identified that the citrullination of R234 disrupts the binding of the isoDGR motif to integrin αVβ3. Since the isoDGR motif has been suggested to be involved in the process of ECM assembly with FN (Takahashi et al. 2007; J. Xu et al. 2010), citrullination of R234 may adversely affect the assembly of extracellular proteins in the inflamed joints of patients with RA.

Despite the fact that the isoDGR motif is the inverse sequence of the RGD motif in FN, the competitive binding assays showed that both the motifs have comparable binding affinity for integrin αVβ3 (Curnis et al. 2006). Since the isoDGR motif shares a high degree of chemical and geometric complementarity with RGD, we performed manual docking by placing the cyclic form of the isoDGR motif, in an inverted orientation, within the RGD binding pocket of integrin αVβ3. Interestingly, an additional methylene group (-CH_2-) in the isoDGR motif leads to more flexibility than is present in the canonical RGD motif, which might explain why a citrullinated form of isoDGR motif only partially reduces its binding with

integrin αVβ3 (Publication II), whereas citrullination of the less flexible RGD motif in TGFβ1-LAP drastically reduced its ability to bind to integrin αVβ6 (Publication I).

Our sequence and structural analyses do not support a previous proposal (Arita et al. 2006) that the small side chain of a polar residue at the N-2 position could determine the substrate specificity of PAD enzymes. In our analyses on available structures where citrullinated peptides were identified in RA patients, we did not identify any characteristic sequential or structural motifs, including secondary structure, which would lead to citrullination of specific arginine residues; indeed, only a solvent-exposed arginine side chain appears necessary and sufficient for binding to PADs and their citrullination activity. In a few cases, the arginine residues were not exposed to solvent in the known 3D structures; consequently, these arginine residues must become solvent-exposed by some mechanism in order to be citrullinated. To conclude, this study shows that arginine residues play a crucial function in protein-protein interactions and especially regarding the integrity of the ECM. Citrullination of arginine residues has the potential to induce inflammatory disease and tissue destruction.

6.2. Human LEUTX activates transcription of genes involved in embryonic genome activation (Publication III)

In publication III, we combined *in silico* study with experimental data from collaborators to assess the importance of the human PRD-like LEUTX protein in early embryonic development. The published sequence and structural data led us to identify functionally important DNA-binding residues in homeobox proteins. Previous studies have demonstrated that mutations of residues at positions 2, 3, and 5 on the N-terminal arm, as well as residues at positions 47, 50, 51, 54, and 55 on the recognition helix of homeodomains, are sufficient to affect the DNA-binding specificity (Ades and Sauer 1994; Noyes et al. 2008). To our knowledge, this is the first mutational study on the human PRD-like homeodomain protein suggesting that a change in the amino acids at positions 47 and 54 in LEUTX affects its binding specificity towards the TAATCC dsDNA motif, and therefore the expression level of the target gene is affected.

The HD structure of LEUTX is not known. However, sufficient structures were available to produce a high quality model structure for wild-type LEUTX as well as for several mutants indicated from our phylogenetic analysis. The mutation A54V in the recognition helix significantly reduces LEUTX activity, while the mutation I47T has no effect; and our structural models supported the results of the experimental studies. The presence of T47 would likely tune the DNA-binding affinity of LEUTX by adjusting the hydrogen-bonding interactions with other highly conserved residues at the specificity determining positions. The large hydrophobic side chain in A54V mutation has the potential to disrupt the aqueous interfacial hydrogen-bonding network in the human LEUTX-dsDNA complex; a network of water-mediated hydrogen bonds has been observed to play a crucial role in improving the fitting of homeodomain and DNA surfaces (Jayaram and Jain 2004). Surprisingly, the double mutant form of LEUTX – I47T

and A54V – restores transcriptional activity and the I47T mutation is sufficient to compensate for the loss in binding affinity due to A54V mutation. Interestingly, a previous study demonstrated that residues at positions 47 and 54 in homeodomains have a significant correlation with respect to binding a specific nucleotide (Chu et al. 2012). Similarly, our structural models and experimental data from collaborators confirm this cooperative functional behavior between the residues at these two positions.

Sequences from the C-terminal region of LEUTX of primates showed the presence of a sequence motif enriched in hydrophobic and acidic amino acids, which could most likely act as a transcription activation domain (TAD). This type of motif has been studied extensively in DNA-binding proteins where it mediates activation of transcription through interactions with general transcriptional coregulators, such as MED15, CBP and p300 (Piskacek 2009; Piskacek et al. 2016, 2021). This study does not provide any experimental support for the predicted 9aaTAD motif in the human LEUTX C-terminal region; however, our most recent study of the PRD-like human DUX4 showed that the 9aaTAD motif interacts with mediator complexes via the KIX binding domain of the coactivator CBP protein (Vuoristo et al. 2022). Thus, the HD region binding to the dsDNA motif is not sufficient for determining the function of LEUTX since the C-terminal region, including protein recognition motifs like 9aaTAD, are essential. A previous study from our collaborators showed that LEUTX is able to activate only about 25% of the upregulated genes during embryonic genome activation (EGA) (Jouhilahti et al. 2016), indicating that either PRD-like or other TFs play a crucial role in activating the full set of EGA genes. For example, our recent study showed that human DUX4 is expressed at the early pre-implantation stage in the embryo where it activates thousands of newly identified enhancer regions by altering chromatin accessibility (Vuoristo et al. 2022).

Taken together, we combined comparative sequence analysis, structural predictions, and experimental results from collaborators to address the hypothesis that LEUTX plays an essential role in early human embryonic development as suggested earlier (Jouhilahti et al. 2016). Our data show that the DNA-binding specificity is determined by a small set of residues in LEUTX and mutations in these residues are enough to affect its transcriptional activity for a TAATCC-containing promoter in genes involved in EGA, but the C-terminal region is also pinpointed for its importance.

6.3. Prediction of SARS-CoV-2-derived T cell epitopes that exactly match experimentally validated SARS-CoV epitopes (Publication IV)

In publication IV, we mainly focused on the prediction of potential SARS-CoV-2 epitopes that exactly match with the experimentally known SARS-CoV epitopes, known for eliciting cytotoxic T cell-mediated immune response. The immunogenicity of peptides bound to MHC-I molecules depends on several factors: 1) the binding affinity of peptide to the MHC-I molecule, 2) the efficiency of peptide processing by antigen presenting cells, 3) the stability of the pMHC-I complex, 4) the evolutionary conservation or “sequence stability” of the

epitopes, and 5) the hydrophobicity of the peptides. For the prediction of epitope binding affinity, a consensus approach was used in which the output from two different tools – NetCTL1.2 and IEDB – were compared with each other in order to improve the prediction results.

Our results showed that many peptides derived from nonstructural, membrane (M) and spike (S) proteins of SARS-CoV-2 are likely to be presented by MHC-I molecules. The C-terminus of the spike protein has been shown to induce a strong T cell immune response in patients infected with SARS-CoV-2 (Braun et al. 2020). One of our predicted epitopes ^{1220}FIAGLIAIV^{1228} overlaps with the C-terminal sequence of the spike protein containing the S2 subunit, which plays an essential role in SARS-CoV-2 virus entry. Therefore, we speculate that this epitope may also be important for inducing the protective immune response against the SARS-CoV-2 infection similar to that seen in patients infected with SARS-CoV (Y.-D. Wang et al. 2004). Surprisingly, the ^{1220}FIAGLIAIV^{1228} epitope is also found in four predicted MHC-II specific 15-mer epitopes, suggesting that $CD4^+$ cells could interact with MHC-I molecules. Indeed, a study that was published while our study was under review suggested that cross-reactivity may provide a "back-up" immune response and could affect disease progression in COVID-19 (Grifoni et al. 2020). Overall, our results suggest that the C-terminus of the spike protein might function as a double-edge sword, being essential for viral entry into the host cells, but also crucial for induction of protective immunity against SARS-CoV-2.

The SARS-CoV-2 genome shares about 79% sequence identity with the SARS-CoV and mutations are likely to affect the ability of MHC-I to recognize and bind SARS-derived peptides; thus the effectiveness of the cytotoxic T cell-mediated immune response is vulnerable for mutations. For example, the SARS-CoV-2-derived, predicted M protein epitope ^{148}HLR**I**AGH**H**L^{156}, which has a poor predicted binding affinity (IC_{50} = 1693 nM) towards HLA-B*15:02, shares 78% sequence identity with the experimentally identified SARS-CoV epitope ^{147}HLR**M**AGH**S**L^{155} also from a M protein and shown to stimulate a strong T cell immune response in patients with the HLA-B*15:02 allotype (O. W. Ng et al. 2016). Binding affinity prediction study showed that the SARS-CoV-derived ^{147}HLRMAGHSL^{155} epitope bind with a high affinity (IC_{50} = 232 nM) to HLA-B*15:02, which has been shown to elicit a strong protection against the severe form of SARS-CoV infection (M. H. L. Ng et al. 2010).

Previous studies have proposed that the stability of eMHC-I complex is a better predictor of immunogenicity than the binding affinity (Burg et al. 1996; Harndahl et al. 2012). Our data showed that about 86% of the predicted immunogenic epitopes ($IC_{50} \leq 50$ nM) that match with the experimental data had half-lives for the eMHC-I complexes of more than 1 hour, while only 15% of the non-immunogenic epitopes (IC_{50} > 500 nM) had half-lives in this range, suggesting that the two predictors are highly correlated and complement each other very well. Since the high mutation rate of a virus can facilitate rapid escape from adaptive immune response (Lewnard and Grad 2018), it is important to examine the sequence stability of the predicted epitopes among different

variants of concern for the development of effective vaccines against COVID-19. Our analyses showed that epitopes derived from transmembrane helices (TMHs) have not changed during the evolution of SARS-CoV-2, suggesting an important role of these epitope sequences in viral pathogenesis, and simultaneously, making these epitopes potential candidates for eliciting a strong cytotoxic T cell response against SARS-CoV-2 itself.

In order to gain better insight into how individual amino acids in an epitope contribute to MHC-I binding, we performed an in-depth structural examination of the published X-ray structures of MHC-I molecules complexed with pathogen-derived epitopes. Furthermore, this structural information was used as a basis for the protein-peptide docking of the SARS-CoV-2-derived epitopes to HLA-A*02:01 and HLA-A*02:06. The conformational flexibility of the side chains of MHC-I cannot be ignored when docking long epitopes, and therefore, a flexible docking approach was used to avoid steric clashes at the protein-epitope interface. Our data demonstrate that residues at positions 1, 2, 3 and 9 are mainly hydrophobic and are essential to provide steric constraints to the N- and C-termini of the epitopes, while the partially solvent-exposed residues at positions 4-8 interact with both MHC-I and TCR.

Taken together, this study shows that the predicted highly immunogenic epitopes of SARS-CoV-2 could provide support for the development of COVID-19 vaccines. Since the sequences of the high-ranking epitopes are 100% identical among all variants of SARS-CoV-2 and match exactly with the experimental data from SARS-CoV, a common vaccine protecting against all strains and potential future variants is possible. Moreover, the conserved epitopes with long half-lives are presented by only a few MHC-I molecules, which cover a large proportion of the world population. Therefore, developing a globally effective vaccine will probably require the involvement of multiple immunogenic epitopes restricted to specific allotypes that cover the largest human population.

6.4 Limitations of the research

Molecular modeling has been widely accepted as a valuable tool for studying macromolecular interactions and drug discovery, and the predicting methods are being developed all the time. However, a predicted model without any experimental evidence may mimic a real biomolecular complex but contain gross errors, or even has nothing to do with a real biomolecule, and may hence lead to wrong conclusions. The level of confidence one places in a model largely derives from the experimental knowledgebase that is available. What structures or complex structures are available? How similar are the sequences of the structures to your target? Mutations are frequently investigated for their effects on function: have engineered mutants been tested? Ligand binding?

In this thesis, all of the *in silico* predictions were thoroughly validated using experimental data from our collaborators and/or from the literature. Depending on the project requirements and schedule, the computational resources were utilized to the best of our knowledge and we have always sought to clearly present the likely reliability of the results and conclusions. In publication I and

II, the analysis of the effects of citrullination on protein-protein interactions was based on the static structural information from X-ray crystal structures of the protein complexes and a variety of experimental data was available from our collaborators and used to assess our prediction. In publication III, putative protein-protein interaction sites involving the 9aaTAD sequence motifs were predicted to be located in the C-terminal domain of the LEUTX protein. Experimental analyses to demonstrate the importance of these sites were not conducted during the course of this study, but the analogous site on the related transcription factor DUX4 was shown by us to be important for protein-protein interactions. In publication IV, we predicted immunogenic epitopes of SARS-CoV-2. However, the training dataset used for machine learning did not cover all of the HLA class I allotypes and was trained on SARS-CoV data, since experimental data on SARS-CoV-2-derived epitopes was not yet available when this study was conducted. Consequently, we were very conservative in our predictions and, hence, we may have missed some potential epitopes.

7. Conclusion

In order to study the structural and functional consequences of molecular interactions, we carried out an extensive *in silico* study on three biological datasets that are related to humans and human health: 1) uncontrolled citrullination of ECM proteins in the inflamed synovial joints of RA patients, 2) regulation of early embryonic development by a novel family of PRD-like transcription factors, and 3) activation of the cytotoxic T cell-mediated immune response against SARS-CoV-2 derived epitopes.

In publication I, our structure analyses of ECM-associated growth factors (GF) revealed that the GF-GF receptor complexes possess at least one arginine residue at the interaction surface of the GF. More specifically in TGF-β1, the loss of ionic interactions caused by citrullination of arginine residues significantly reduced binding to both integrin αVβ6 and TGF-βRII. In publication II, we examined the ECM proteins citrullinated in the inflamed joints of RA patients, investigating the specificity determining features of the PAD enzymes. Contrary to early reports, we found that for citrullination an arginine side chain only needs to be exposed to solvent but can arise from α-helices, β-strands, loops and β-turns. In addition, there is no sequence motif linked to the enzymatic activity of PADs. Our detailed structural analysis of one of the ECM proteins, fibronectin (FN), revealed that the loss of strong ionic interactions caused by citrullination significantly reduced FN binding to integrin αVβ3. Our observations from these two studies showed that citrullination has the potential to affect adversely the critical extracellular protein interactions that regulate immune responses and other essential cellular processes.

In publication III, we characterized the human PRD-like LEUTX transcription factor in depth and identified specificity determining residues in the DNA-binding domain that are essential for LEUTX binding to the TAATCC sequence motif of dsDNA, a motif found enriched in genes involved in embryonic genome activation. The C-terminal domain of LEUTX, the Leutx domain, contains a specific 9aaTAD sequence motif that can potentially recruit transcription coregulator proteins. A heterozygotic missense mutation A54V would likely disturb the network of water-mediated hydrogen bonds linking other specificity determining residues that are critical for TAATCC motif recognition. The double mutant form of LEUTX – I47T and A54V –restores binding to the DNA motif since the T47 residue contributes favorably to the complex stability. Our results supported a study that demonstrated the essential role of the human LEUTX during the first few cell divisions in early embryonic development.

In publication IV, we identified SARS-CoV-2 epitopes that were predicted to bind to cell surface MHC-I molecules. The predicted epitopes were matched with the experimentally known SARS-CoV-derived epitopes that are known to elicit a robust cytotoxic T cell-mediated immune response *in vitro*. The binding of the selected SARS-CoV-2 epitopes were evaluated by analyzing published X-ray crystal structures, molecular models and MD simulation trajectories. The HLA-A*02 allotypes showed the greatest potential to bind the selected epitopes,

which are mainly hydrophobic in nature and possess longer predicted half-lives than epitopes that bind to other HLAs. Our observations could assist in understanding the molecular basis for regulation of SARS-CoV-2-derived epitopes binding to MHC-I and TCR; the epitopes may enable us to provide useful information for vaccine development against SARS-CoV-2 and other related viruses that may emerge in the future.

The studies presented in this thesis demonstrate that even a tiny change in protein structure can lead to substantial changes in protein function. One atom change caused by citrullination results in a loss-of-function activity of the ECM proteins, whereas the "conservative" substitution of amino acids shows structural consequences on the PRD-like LEUTX transcription factor led to a loss of binding that could be restored by a second mutation. In addition, the structural analysis of eHLA complexes demonstrates the significance of individual amino acids in determining the physicochemical properties of SARS-CoV-2-derived immunogenic epitopes. Overall, this thesis relied on the synergy obtained by combining computational and experimental studies focused on understanding the consequences of molecular interactions on protein structure and function.

8. References

Abbas, Ali K. et al. 2014. "Negative Regulation of the Peptidylarginine Deiminase Type IV Promoter by NF-KB in Human Myeloid Cells." *Gene* 533(1): 123–31.

Abdullah, Syatirah Najmi et al. 2013. "Porphyromonas Gingivalis Peptidylarginine Deiminase Substrate Specificity." *Anaerobe* 23: 102–8.

Ades, Sarah E., and Robert T. Sauer. 1994. "Differential DNA-Binding Specificity of the Engrailed Homeodomain: The Role of Residue 50." *Biochemistry* 33(31): 9187–94.

Aggarwal, Rohit et al. 2009. "Anti-Citrullinated Peptide Antibody Assays and Their Role in the Diagnosis of Rheumatoid Arthritis." *Arthritis Care and Research* 61(11): 1472–83.

Alghamdi, Ibrahim G. et al. 2014. "The Pattern of Middle East Respiratory Syndrome Coronavirus in Saudi Arabia: A Descriptive Epidemiological Analysis of Data from the Saudi Ministry of Health." *International Journal of General Medicine* 7: 417–23.

Amarilla, Alberto A. et al. 2021. "A Versatile Reverse Genetics Platform for SARS-CoV-2 and Other Positive-Strand RNA Viruses." *Nature Communications* 12(1): 1–15.

Andersen, Kristian G. et al. 2020. "The Proximal Origin of SARS-CoV-2." *Nature Medicine* 26(4): 450–52.

Andrade, Felipe et al. 2010. "Autocitrullination of Human Peptidyl Arginine Deiminase Type 4 Regulates Protein Citrullination during Cell Activation." *Arthritis and Rheumatism* 62(6): 1630–40.

Arita, Kyouhei et al. 2004. "Structural Basis for Ca^{2+}-Induced Activation of Human PAD4." *Nature Structural and Molecular Biology* 11(8): 777–83.

Arita, Kyouhei et al. 2006. "Structural Basis for Histone N-Terminal Recognition by Human Peptidylarginine Deiminase 4." *Proceedings of the National Academy of Sciences of the United States of America* 103(14): 5291–96.

Vander Ark, Alexandra, Jingchen Cao, and Xiaohong Li. 2018. "TGF-β Receptors: In and beyond TGF-β Signaling." *Cellular Signalling* 52: 112–20.

Armstrong, Craig T., Philip E. Mason, J. L.Ross Anderson, and Christopher E. Dempsey. 2016. "Arginine Side Chain Interactions and the Role of Arginine as a Gating Charge Carrier in Voltage Sensitive Ion Channels." *Scientific Reports* 6: 1–10.

Assarsson, Erika et al. 2007. "A Quantitative Analysis of the Variables Affecting the Repertoire of T Cell Specificities Recognized after Vaccinia Virus Infection." *The Journal of Immunology* 178(12): 7890–7901.

Assiri, Abdullah et al. 2013. "Hospital Outbreak of Middle East Respiratory Syndrome Coronavirus." *New England Journal of Medicine* 369(5): 407–16.

Auton, Adam et al. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526(7571): 68–74.

Baibakov, Boris et al. 2007. "Sperm Binding to the Zona Pellucida Is Not Sufficient to Induce Acrosome Exocytosis." *Development* 134(5): 933–43.

Bateman, Alex et al. 2021. "UniProt: The Universal Protein Knowledgebase in 2021." *Nucleic Acids Research* 49(D1): D480–89.

Batty, Cole J, Mark T Heise, Eric M Bachelder, and Kristy M Ainslie. 2021. "Vaccine Formulations in Clinical Development for the Prevention Ofsevere Acute Respiratory Syndrome Coronavirus 2 Infection." *Advanced Drug Delivery Reviews* 169: 168–89.

Van Beers, Joyce J.B.C. et al. 2013. "The Rheumatoid Arthritis Synovial Fluid Citrullinome Reveals Novel Citrullinated Epitopes in Apolipoprotein E, Myeloid Nuclear Differentiation Antigen, and β-Actin." *Arthritis and Rheumatism* 65(1): 69–80.

Benton, Donald J. et al. 2021. "The Effect of the D614G Substitution on the

Structure of the Spike Glycoprotein of SARS-CoV-2." *Proceedings of the National Academy of Sciences of the United States of America* 118(9): 2–5.
Berggård, Tord, Sara Linse, and Peter James. 2007. "Methods for the Detection and Analysis of Protein-Protein Interactions." *Proteomics* 7(16): 2833–42.
Berman, H M et al. 2000. "The Protein Data Bank." *Nucleic Acids Research* 28(1): 235–42.
Blystone, S. D., I. L. Graham, F. P. Lindberg, and E. J. Brown. 1994. "Integrin α(v)B3 Differentially Regulates Adhesive and Phagocytic Functions of the Fibronectin Receptor A5β1." *Journal of Cell Biology* 127(4): 1129–37.
Boroviak, Thorsten, and Jennifer Nichols. 2014. "The Birth of Embryonic Pluripotency." *Philosophical Transactions of the Royal Society B: Biological Sciences* 369(1657): 20130541.
Bourouiba, Lydia. 2020. "Turbulent Gas Clouds and Respiratory Pathogen Emissions: Potential Implications for Reducing Transmission of COVID-19." *JAMA - Journal of the American Medical Association* 323(18): 1837–38.
Bowditch, R. D. et al. 1994. "Identification of a Novel Integrin Binding Site in Fibronectin. Differential Utilization by B3 Integrins." *Journal of Biological Chemistry* 269(14): 10856–63.
Brahmajosyula, Manjula, and Masashi Miyake. 2013. "Role of Peptidylarginine Deiminase 4 (PAD4) in Pig Parthenogenetic Preimplantation Embryonic Development." *Zygote* 21(4): 385–93.
Brant, Ayslan Castro et al. 2021. "SARS-CoV-2: From Its Discovery to Genome Structure, Transcription, and Replication." *Cell and Bioscience* 11(1): 1–17.
Braun, Julian et al. 2020. "SARS-CoV-2-Reactive T Cells in Healthy Donors and Patients with COVID-19." *Nature* 587(7833): 270–74.
Brinda, K. V., N. Kannan, and S. Vishveshwara. 2002. "Analysis of Homodimeric Protein Interfaces by Graph-Spectral Methods." *Protein Engineering* 15(4): 265–77.
Brister, J. Rodney, Danso Ako-Adjei, Yiming Bao, and Olga Blinkova. 2015. "NCBI Viral Genomes Resource." *Nucleic Acids Research* 43(D1): D571–77.
Brown, J H et al. 1993. "Three-Dimensional Structure of the Human Class II Histocompatibility Antigen HLA-DR1." *Nature* 364(6432): 33–39.
Burg, S H van der et al. 1996. "Immunogenicity of Peptides Bound to MHC Class I Molecules Depends on the MHC-Peptide Complex Stability." *The Journal of Immunology* 156(9): 3308–14.
Bürglin, Thomas R. 2011. "Homeodomain Subtypes and Functional Diversity." *Sub-Cellular Biochemistry* 52: 95–122.
Bürglin, Thomas R., and Markus Affolter. 2016. "Homeodomain Proteins: An Update." *Chromosoma* 125(3): 497–521.
Cagliani, Rachele, Diego Forni, Mario Clerici, and Manuela Sironi. 2020. "Computational Inference of Selection Underlying the Evolution of the Novel Coronavirus, Severe Acute Respiratory Syndrome Coronavirus 2." *Journal of Virology* 94(12).
Campbell, Melody G. et al. 2020. "Cryo-EM Reveals Integrin-Mediated TGF-β Activation without Release from Latent TGF-β." *Cell* 180(3): 490-501.e16.
Cascella, Raffaella et al. 2018. "Digenic Inheritance of Shortened Repeat Units of the D4Z4 Region and a Loss-of-Function Variant in SMCHD1 in a Family with FSHD." *Frontiers in Neurology* 9: 1–6.
Chang, Xiaotian et al. 2009. "Increased PADI4 Expression in Blood and Tissues of Patients with Malignant Tumors." *BMC Cancer* 9: 1–11.
Chavanas, Stéphane et al. 2004. "Comparative Analysis of the Mouse and Human Peptidylarginine Deiminase Gene Clusters Reveals Highly Conserved Non-Coding

Segments and a New Human Gene, PADI6." *Gene* 330(1–2): 19–27.
Chavanas, Stéphane et al. 2008. "Long-Range Enhancer Associated with Chromatin Looping Allows AP-1 Regulation of the Peptidylarginine Deiminase 3 Gene in Differentiated Keratinocyte." *PLoS ONE* 3(10): 1–11.
Chen, Jun et al. 2020. "Clinical Progression of Patients with COVID-19 in Shanghai, China." *Journal of Infection* 80(5): e1–6.
Cherrington, Brian D. et al. 2012. "Potential Role for PAD2 in Gene Regulation in Breast Cancer Cells." *PLoS ONE* 7(7): 1–12.
Chi, Neil, and Jonathan A. Epstein. 2002. "Getting Your Pax Straight: Pax Proteins in Development and Disease." *Trends in Genetics* 18(1): 41–47.
Chicz, R M et al. 1992. "Predominant Naturally Processed Peptides Bound to HLA-DR1 Are Derived from MHC-Related Molecules and Are Heterogeneous in Size." *Nature* 358(6389): 764–68.
Chouhan, Bhanupratap et al. 2011. "Conservation of the Human Integrin-Type Beta-Propeller Domain in Bacteria." *PLoS ONE* 6(10).
Chowell, Diego et al. 2015. "TCR Contact Residue Hydrophobicity Is a Hallmark of Immunogenic CD8+ T Cell Epitopes." *Proceedings of the National Academy of Sciences of the United States of America* 112(14): E1754–62.
Christophorou, Maria A. et al. 2014. "Citrullination Regulates Pluripotency and Histone H1 Binding to Chromatin." *Nature* 507(7490): 104–8.
Chu, Stephanie W. et al. 2012. "Exploring the DNA-Recognition Potential of Homeodomains." *Genome Research* 22(10): 1889–98.
Clift, Dean, and Melina Schuh. 2014. "Europe PMC Funders Group Re-Starting Life : Fertilization and the Transition from Meiosis to Mitosis." 14(9): 549–62.
Clubb, Elizabeth. 1986. "Natural Methods of Family Planning." *The Journal of the Royal Society for the Promotion of Health* 106(4): 121–26.
Corona, Rosario I., and Jun Tao Guo. 2016. "Statistical Analysis of Structural Determinants for Protein–DNA-Binding Specificity." *Proteins: Structure, Function and Bioinformatics* 84(8): 1147–61.
Coutard, B. et al. 2020. "The Spike Glycoprotein of the New Coronavirus 2019-NCoV Contains a Furin-like Cleavage Site Absent in CoV of the Same Clade." *Antiviral Research* 176(February): 104742.
Cox, Rebecca J., and Karl A. Brokstad. 2020. "Not Just Antibodies: B Cells and T Cells Mediate Immunity to COVID-19." *Nature Reviews Immunology* 20(10): 581–82.
Crux, Nicole B., and Shokrollah Elahi. 2017. "Human Leukocyte Antigen (HLA) and Immune Regulation: How Do Classical and Non-Classical HLA Alleles Modulate Immune Response to Human Immunodeficiency Virus and Hepatitis C Virus Infections?" *Frontiers in Immunology* 8(JUL): 1–26.
Curnis, Flavio et al. 2006. "Spontaneous Formation of L-Isoaspartate and Gain of Function in Fibronectin." *Journal of Biological Chemistry* 281(47): 36466–76.
Curnis, Flavio et al. 2010. "Critical Role of Flanking Residues in NGR-to-IsoDGR Transition and CD13/Integrin Receptor Switching." *Journal of Biological Chemistry* 285(12): 9114–23.
Czerny, Thomas, Maxime Bouchard, Zbynek Kozmi, and Meinrad Busslinger. 1997. "The Characterization of Novel Pax Genes of the Sea Urchin and Drosophila Reveal an Ancient Evolutionary Origin of the Pax2/5/8 Subfamily." *Mechanisms of Development* 67(2): 179–92.
Daïen, Claire I., and Jérémie Sellam. 2015. "Obesity and Inflammatory Arthritis: Impact on Occurrence, Disease Characteristics and Therapeutic Response." *RMD Open* 1(1): 1–9.
Damas, Joana et al. 2020. "Broad Host Range of SARS-CoV-2 Predicted by

Comparative and Structural Analysis of ACE2 in Vertebrates." *Proceedings of the National Academy of Sciences of the United States of America* 117(36): 22311–22.
Damgaard, Dres et al. 2016. "Reduced Glutathione as a Physiological Co-Activator in the Activation of Peptidylarginine Deiminase." *Arthritis Research and Therapy* 18(1): 1–7.
Dan, Jennifer M. et al. 2021. "Immunological Memory to SARS-CoV-2 Assessed for up to 8 Months after Infection." *Science* 371(6529).
Daopin, Sun, Karl A. Piez, Yasushi Ogawa, and David R. Davies. 1992. "Crystal Structure of Transforming Growth Factor-B2: An Unusual Fold for the Superfamily." *Science* 257(5068): 369–73.
Darrah, Erika et al. 2013. "Erosive Rheumatoid Arthritis Is Associated with Antibodies That Activate PAD4 by Increasing Calcium Sensitivity." *Science Translational Medicine* 5(186): 1–19.
Darrah, Erika, and Felipe Andrade. 2018. "Rheumatoid Arthritis and Citrullination." *Current Opinion in Rheumatology* 30(1): 72–78.
Dehingia, Nabamallika, and Anita Raj. 2021. "Sex Differences in COVID-19 Case Fatality: Do We Know Enough?" *The Lancet Global Health* 9(1): e14–15.
Dong, Sijun et al. 2005. "Regulation of the Expression of Peptidylarginine Deiminase Type II Gene (PADI2) in Human Keratinocytes Involves Sp1 and Sp3 Transcription Factors." *Journal of Investigative Dermatology* 124(5): 1026–33.
Dong, Sijun, Zilian Zhang, and Hidenari Takahara. 2007. "Estrogen-Enhanced Peptidylarginine Deiminase Type IV Gene (PADI4) Expression in MCF-7 Cells Is Mediated by Estrogen Receptor-α-Promoted Transfactors Activator Protein-1, Nuclear Factor-Y, and Sp1." *Molecular Endocrinology* 21(7): 1617–29.
Dong, Xianchi, Nathan E. Hudson, Chafen Lu, and Timothy A. Springer. 2014. "Structural Determinants of Integrin β-Subunit Specificity for Latent TGF-β." *Nature Structural and Molecular Biology* 21(12): 1091–96.
Doremalen, Neeltje van et al. 2020. "Aerosol and Surface Stability of SARS-CoV-2 as Compared with SARS-CoV-1." *New England Journal of Medicine* 382(16): 1564–67.
Doyle, Carolyn, and Jack L. Strominger. 1987. "Interaction between CD4 and Class II MHC Molecules Mediates Cell Adhesion." *Nature* 330(6145): 256–59.
Drozdetskiy, Alexey, Christian Cole, James Procter, and Geoffrey J. Barton. 2015. "JPred4: A Protein Secondary Structure Prediction Server." *Nucleic Acids Research* 43(W1): W389–94.
Du, Rong Hui et al. 2020. "Predictors of Mortality for Patients with COVID-19 Pneumonia Caused by SARSCoV- 2: A Prospective Cohort Study." *European Respiratory Journal* 55(5).
Duan, Liangwei et al. 2020. "The SARS-CoV-2 Spike Glycoprotein Biosynthesis, Structure, Function, and Antigenicity: Implications for the Design of Spike-Based Vaccine Immunogens." *Frontiers in Immunology* 11(October): 1–12.
Duart, Gerard, Maria J García-murria, and Ismael Mingarro. 2021. "The SARS-CoV-2 Envelope (E) Protein Has Evolved towards Membrane Topology Robustness." *BBA - Biomembranes* 1863: 183608.
Dutta, Noton K., Kaushiki Mazumdar, and James T. Gordy. 2020. "The Nucleocapsid Protein of SARS–CoV-2: A Target for Vaccine Development." *Journal of Virology* 94(13).
Eddy, C. A., and C. J. Pauerstein. 1980. "Anatomy and Physiology of the Fallopian Tube." *Clinical Obstetrics and Gynecology* 23(4): 1177–93.
Erb, Carol. 2006. The Laboratory Rat *Chapter 28: Embryology and Teratology*. Second. Elsevier Inc.
Eric Xu, H. et al. 1999. "Crystal Structure of the Human Pax6 Paired Domain-DNA Complex Reveals Specific Roles for the Linker Region and Carboxy-Terminal Subdomain in DNA

Binding." *Genes and Development* 13(10): 1263–75.
Falcão, Ana Mendanha et al. 2019. "PAD2-Mediated Citrullination Contributes to Efficient Oligodendrocyte Differentiation and Myelination." *Cell Reports* 27(4): 1090-1102.e10.
Falk, Kirsten et al. 1991. "Allele-Specific Motifs Revealed by Sequencing of Self-Peptides Eluted from MHC Molecules." *Nature* 351(6324): 290–96.
Fearon, William Robert. 1939. "The Carbamido Diacetyl Reaction: A Test for Citrulline." *Biochemical Journal* 33(6): 902–7.
Feng, Dan et al. 2009. "The SARS Epidemic in Mainland China: Bringing Together All Epidemiological Data." *Tropical Medicine and International Health* 14(SUPPL. 1): 4–13.
Fisher, Alfred L., and Michael Caudy. 1998. "Groucho Proteins: Transcriptional Corepressors for Specific Subsets of DNA-Binding Transcription Factors in Vertebrates and Invertebrates." *Genes and Development* 12(13): 1931–40.
Foged, Camilla. 2011. "Subunit Vaccines of the Future: The Need for Safe, Customized and Optimized Particulate Delivery Systems." *Therapeutic Delivery* 2(8): 1057–77.
Folegatti, Pedro M. et al. 2020. "Safety and Immunogenicity of the ChAdOx1 NCoV-19 Vaccine against SARS-CoV-2: A Preliminary Report of a Phase 1/2, Single-Blind, Randomised Controlled Trial." *The Lancet* 396(10249): 467–78.
Foulquier, Céline et al. 2007. "Peptidyl Arginine Deiminase Type 2 (PAD-2) and PAD-4 but Not PAD-1, PAD-3, and PAD-6 Are Expressed in Rheumatoid Arthritis Synovium in Close Association with Tissue Inflammation." *Arthritis and Rheumatism* 56(11): 3541–53.
Frantz, Christian, Kathleen M. Stewart, and Valerie M. Weaver. 2010. "The Extracellular Matrix at a Glance." *Journal of Cell Science* 123(24): 4195–4200.
Fuchs, Tobias A. et al. 2010. "Extracellular DNA Traps Promote Thrombosis." *Proceedings of the National Academy of Sciences of the United States of America* 107(36): 15880–85.
Fuhrmann, Jakob, Kathleen W. Clancy, and Paul R. Thompson. 2015. "Chemical Biology of Protein Arginine Modifications in Epigenetic Regulation." *Chemical Reviews* 115(11): 5413–61.
Fujisaki, Makoto, and Kiyoshi Sugawara. 1981. "Properties of Peptidylarginine Deiminase from the Epidermis of Newborn Rats." *Journal of Biochemistry* 89(1): 257–63.
Furukawa, Takahisa, Christine A. Kozak, and Constance L. Cepko. 1997. "Rax, a Novel Paired-Type Homeobox Gene, Shows Expression in the Anterior Neural Fold and Developing Retina." *Proceedings of the National Academy of Sciences of the United States of America* 94(7): 3088–93.
Galliot, Brigitte, Colomban De Vargas, and David Miller. 1999. "Evolution of Homeobox Genes: Q50 Paired-like Genes Founded the Paired Class." *Development Genes and Evolution* 209(3): 186–97.
Gao, Wentao et al. 2003. "Effects of a SARS-Associated Coronavirus Vaccine in Monkeys." *Lancet* 362(9399): 1895–96.
Garber, R L, A Kuroiwa, and W J Gehring. 1983. "Genomic and cDNA Clones of the Homeotic Locus Antennapedia in Drosophila." *The EMBO Journal* 2(11): 2027–36.
Garcia, K C et al. 1996. "An αβ T Cell Receptor Structure at 2.5 Å and Its Orientation in the TCR-MHC Complex." *Science* 274(5285): 209–19.
Garvie, Colin W., James Hagman, and Cynthia Wolberger. 2001. "Structural Studies of Ets-1/Pax5 Complex Formation on DNA." *Molecular Cell* 8(6): 1267–76.
Garvie, Colin W., Miles A. Pufall, Barbara J. Graves, and Cynthia Wolberger. 2002. "Structural Analysis of the Autoinhibition of Ets-1 and Its Role in

Protein Partnerships." *Journal of Biological Chemistry* 277(47): 45529–36.

Gaulard, P. et al. 1990. "Expression of the Alpha/Beta and Gamma/Delta T-Cell Receptors in 57 Cases of Peripheral T-Cell Lymphomas. Identification of a Subset of γ/δ T-Cell Lymphomas." *American Journal of Pathology* 137(3): 617–28.

Giaglis, Stavros et al. 2016. "Multimodal Regulation of NET Formation in Pregnancy: Progesterone Antagonizes the pro-NETotic Effect of Estrogen and G-CSF." *Frontiers in Immunology* 7: 1–14.

Gorbalenya, Alexander E. et al. 2020. "The Species Severe Acute Respiratory Syndrome-Related Coronavirus: Classifying 2019-NCoV and Naming It SARS-CoV-2." *Nature Microbiology* 5(4): 536–44.

Grifoni, Alba et al. 2020. "Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-19 Disease and Unexposed Individuals." *Cell* 181(7): 1489-1501.e15.

Groppe, Jay et al. 2008. "Cooperative Assembly of TGF-β Superfamily Signaling Complexes Is Mediated by Two Disparate Mechanisms and Distinct Modes of Receptor Binding." *Molecular Cell* 29(2): 157–68.

Guan, Y. et al. 2003. "Isolation and Characterization of Viruses Related to the SARS Coronavirus from Animals in Southern China." *Science* 302(5643): 276–78.

Gude, Neil M., Claire T. Roberts, Bill Kalionis, and Roger G. King. 2004. "Growth and Function of the Normal Human Placenta." *Thrombosis Research* 114(5-6 SPEC. ISS.): 397–407.

Gui, Miao et al. 2017. "Cryo-Electron Microscopy Structures of the SARS-CoV Spike Glycoprotein Reveal a Prerequisite Conformational State for Receptor Binding." *Cell Research* 27(1): 119–29.

Guo, Qin, and Walter Fast. 2011. "Citrullination of Inhibitor of Growth 4 (ING4) by Peptidylarginine Deminase 4 (PAD4) Disrupts the Interaction between ING4 and P53." *Journal of Biological Chemistry* 286(19): 17069–78.

Guo, Wei et al. 2017. "Investigating the Expression, Effect and Tumorigenic Pathway of PADI2 in Tumors." *OncoTargets and Therapy* 10: 1475–85.

Gupta, Aakriti et al. 2020. "Extrapulmonary Manifestations of COVID-19." *Nature Medicine* 26(7): 1017–32.

Gussow, Ayal B. et al. 2020. "Genomic Determinants of Pathogenicity in SARS-CoV-2 and Other Human Coronaviruses." *Proceedings of the National Academy of Sciences of the United States of America* 117(26): 15193–99.

Haley, Sheila A., and Gary M. Wessel. 1999. "The Cortical Granule Serine Protease CGSP1 of the Sea Urchin, Strongylocentrotus Purpuratus, Is Autocatalytic and Contains a Low-Density Lipoprotein Receptor-like Domain." *Developmental Biology* 211(1): 1–10.

Hamm, Danielle C., and Melissa M. Harrison. 2018. "Regulatory Principles Governing the Maternal-to-Zygotic Transition: Insights from Drosophila Melanogaster." *Open Biology* 8(12).

Hansson, Marianne, Per-Åke Nygren, and Stefan Ståhl. 2000. "Design and Production of Recombinant Subunit Vaccines." *Biotechnology and Applied Biochemistry* 32(2): 95.

Haravuori, Henna et al. 2020. "Personnel Well-Being in the Helsinki University Hospital during the COVID-19 Pandemic—a Prospective Cohort Study." *International Journal of Environmental Research and Public Health* 17(21): 1–9.

Harndahl, Mikkel et al. 2012. "Peptide-MHC Class I Stability Is a Better Predictor than Peptide Affinity of CTL Immunogenicity." *European Journal of Immunology* 42(6): 1405–16.

Harper, M J K. 1982. *Sperm and Egg Transport. In Austin CR and Short RV*

(Eds) Germ Cells and Fertilization. Vol 1. Cambridge University Press.
Harvey, Bohdan P. et al. 2007. "Antigen Presentation and Transfer between B Cells and Macrophages." *European Journal of Immunology* 37(7): 1739–51.
Hoffmann, Markus et al. 2020. "SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor." *Cell* 181(2): 271-280.e8.
Holland, Peter W.H. 2013. "Evolution of Homeobox Genes." *Wiley Interdisciplinary Reviews: Developmental Biology* 2(1): 31–45.
Holland, Peter W.H., H. Anne F. Booth, and Elspeth A. Bruford. 2007. "Classification and Nomenclature of All Human Homeobox Genes." *BMC Biology* 5: 1–29.
Horikoshi, Naoki et al. 2011. "Structural and Biochemical Analyses of the Human PAD4 Variant Encoded by a Functional Haplotype Gene." *Acta Crystallographica Section D: Biological Crystallography* 67(2): 112–18.
Huang, Angkana T. et al. 2020. "A Systematic Review of Antibody Mediated Immunity to Coronaviruses: Kinetics, Correlates of Protection, and Association with Severity." *Nature Communications* 11(1): 1–16.
Huang, Chaolin et al. 2020. "Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, China." *The Lancet* 395(10223): 497–506.
Huang, Yuan et al. 2020. "Structural and Functional Properties of SARS-CoV-2 Spike Protein: Potential Antivirus Drug Development for COVID-19." *Acta Pharmacologica Sinica* 41(9): 1141–49.
Hui, AlexanderY et al. 2012. "A Systems Biology Approach to Synovial Joint Lubrication in Health, Injury, and Disease." *Wiley Interdisciplinary Reviews: systems biology and medicine* 4(1): 15–37.
Hui, D. S. et al. 2005. "Impact of Severe Acute Respiratory Syndrome (SARS) on Pulmonary Function, Functional Capacity and Quality of Life in a Cohort of Survivors." *Thorax* 60(5): 401–9.
Hui, David S. et al. 2018. "Middle East Respiratory Syndrome Coronavirus: Risk Factors and Determinants of Primary, Household, and Nosocomial Transmission." *The Lancet Infectious Diseases* 18(8): e217–27.
Ibarrondo, F. Javier et al. 2020. "Rapid Decay of Anti–SARS-CoV-2 Antibodies in Persons with Mild Covid-19." *The New England Journal of Medicine* 383(11): 1085–87.
Ioannidis, John P.A. 2020. "Global Perspective of COVID-19 Epidemiology for a Full-Cycle Pandemic." *European Journal of Clinical Investigation* 50(12): 1–9.
Irving, Aaron T. et al. 2021. "Lessons from the Host Defences of Bats, a Unique Viral Reservoir." *Nature* 589(7842): 363–70.
Ishida-Yamamoto, Akemi et al. 2002. "Sequential Reorganization of Cornified Cell Keratin Filaments Involving Filaggrin-Mediated Compaction and Keratin 1 Deimination." *Journal of Investigative Dermatology* 118(2): 282–87.
Ishigami, Akihito et al. 2005. "Abnormal Accumulation of Citrullinated Proteins Catalyzed by Peptidylarginine Deiminase in Hippocampal Extracts from Patients with Alzheimer's Disease." *Journal of Neuroscience Research* 80(1): 120–28.
Jack, Amanda et al. 2021. "SARS-CoV-2 Nucleocapsid Protein Forms Condensates with Viral Genomic RNA." *PLoS Biology* 19(10): 1–30.
Jalkanen, Pinja et al. 2021. "COVID-19 MRNA Vaccine Induced Antibody Responses against Three SARS-CoV-2 Variants." *Nature Communications* 12(1): 1–11.
Jayaram, B., and Tarun Jain. 2004. "The Role of Water in Protein-DNA Recognition." *Annual Review of Biophysics and Biomolecular Structure* 33: 343–61.
Jeyanathan, Mangalakumari et al. 2020. "Immunological Considerations for

COVID-19 Vaccine Strategies." *Nature Reviews Immunology* 20(10): 615–32.
Jia, Na et al. 2009. "Case Fatality of SARS in Mainland China and Associated Risk Factors." *Tropical Medicine and International Health* 14(SUPPL. 1): 21–27.
Johnson, Kamin J., Harvey Sage, Gina Briscoe, and Harold P. Erickson. 1999. "The Compact Conformation of Fibronectin Is Determined by Intramolecular Ionic Interactions." *Journal of Biological Chemistry* 274(22): 15473–79.
Johnson, Mark et al. 2008. "NCBI BLAST: A Better Web Interface." *Nucleic acids research* 36(Web Server issue): 5–9.
Jones, Kate E. et al. 2008. "Global Trends in Emerging Infectious Diseases." *Nature* 451(7181): 990–93.
Jouhilahti, Eeva Mari et al. 2016. "The Human PRD-like Homeobox Gene LEUTX Has a Central Role in Embryo Genome Activation." *Development (Cambridge)* 143(19): 3459–69.
Jun, Susie, and Claude Desplan. 1996. "Cooperative Interactions between Paired Domain and Homeodomain." *Development* 122(9): 2639–50.
Kabashima, Kenji. 2013. "New Concept of the Pathogenesis of Atopic Dermatitis: Interplay among the Barrier, Allergy, and Pruritus as a Trinity." *Journal of Dermatological Science* 70(1): 3–11.
Kaech, Susan M., and E. John Wherry. 2007. "Heterogeneity and Cell-Fate Decisions in Effector and Memory CD8+ T Cell Differentiation during Viral Infection." *Immunity* 27(3): 393–405.
Katoh, Kazutaka, and Daron M. Standley. 2013. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution* 30(4): 772–80.
Kern, David M. et al. 2021. "Cryo-EM Structure of SARS-CoV-2 ORF3a in Lipid Nanodiscs." *Nature Structural and Molecular Biology* 28(7): 573–82.
Khan, Mohd Imran et al. 2020. "Comparative Genome Analysis of Novel Coronavirus (SARS-CoV-2) from Different Geographical Locations and the Effect of Mutations on Major Target Proteins: An in Silico Insight." *PLoS ONE* 15: 1–18.
Kirchdoerfer, Robert N., and Andrew B. Ward. 2019. "Structure of the SARS-CoV Nsp12 Polymerase Bound to Nsp7 and Nsp8 Co-Factors." *Nature Communications* 10(1): 1–9.
Kissinger, Charles R. et al. 1990. "Cryatal Structure of an Engrailed Homeodomain-DNA complex at 2.8Å Resolution: A Framework for Understanding Homeodomain-DNA interactions." *Cell* 63(3): 579-90.
Koga, Yotaro, and Ryo Ohtake. 1914. "Study Report on the Constituents of Squeezed Watermelon." *Journal of the Tokyo Chemical Society* 35: 519–28.
Kohyama, Shunsuke et al. 2009. "Efficient Induction of Cytotoxic T Lymphocytes Specific for Severe Acute Respiratory Syndrome (SARS)-Associated Coronavirus by Immunization with Surface-Linked Liposomal Peptides Derived from a Non-Structural Polyprotein 1a." *Antiviral Research* 84(2): 168–77.
Koya, Shaffi F. et al. 2021. "COVID-19 and Comorbidities: Audit of 2,000 COVID-19 Deaths in India." *Journal of Epidemiology and Global Health* 11(2): 230–32.
Krebs, H. A., and K. Henseleit. 1932. "Untersuchungen Über Die Harnstoffbildung Im Tierkörper." *Hoppe-Seyler's Zeitschrift für Physiologische Chemie* 210: 33–66.
Krogh, Anders, Björn Larsson, Gunnar Von Heijne, and Erik L.L. Sonnhammer. 2001. "Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes." *Journal of Molecular Biology* 305(3): 567–80.
Kwansa, Albert L., Raffaella De Vita, and Joseph W. Freeman. 2014. "Mechanical Recruitment of N- and C-Crosslinks in Collagen Type I." *Matrix Biology* 34: 161–69.
Kyte, Jack, and Russell F. Doolittle. 1982. "A Simple Method for Displaying the

Hydropathic Character of a Protein." *Journal of Molecular Biology* 157(1): 105–32.
Lam, Tommy Tsan Yuk et al. 2020. "Identifying SARS-CoV-2-Related Coronaviruses in Malayan Pangolins." *Nature* 583(7815): 282–85.
Larsen, Mette V. et al. 2007. "Large-Scale Validation of Methods for Cytotoxic T-Lymphocyte Epitope Prediction." *BMC Bioinformatics* 8: 1–12.
Lauer, Stephen A. et al. 2020. "The Incubation Period of Coronavirus Disease 2019 (CoVID-19) from Publicly Reported Confirmed Cases: Estimation and Application." *Annals of Internal Medicine* 172(9): 577–82.
Lee, Chien Yun et al. 2017. "Molecular Interplay between the Dimer Interface and the Substrate-Binding Site of Human Peptidylarginine Deiminase 4." *Scientific Reports* 7: 1–14.
Lee, Miler T., Ashley R. Bonneau, and Antonio J. Giraldez. 2014. "Zygotic Genome Activation during the Maternal-to-Zygotic Transition." *Annual review of cell and developmental biology* 30: 581–613.
Lehtonen, Jukka V et al. 2004. "BODIL: A Molecular Modeling Environment for Structure-Function Analysis and Drug Design." *Journal of Computer-Aided Molecular Design* 18(6): 401–19.
Leroux-Roels, Geert et al. 2000. "A Comparison of Two Commercial Recombinant Vaccines for Hepatitis B in Adolescents." *Vaccine* 19(7–8): 937–42.
Lessler, Justin et al. 2009. "Incubation Periods of Acute Respiratory Viral Infections: A Systematic Review." *The Lancet Infectious Diseases* 9(5): 291–300.
Letko, Michael et al. 2020. "Bat-Borne Virus Diversity, Spillover and Emergence." *Nature Reviews Microbiology* 18(8): 461–71.
Leung, C. W., and W. K. Chiu. 2004. "Clinical Picture, Diagnosis, Treatment and Outcome of Severe Acute Respiratory Syndrome (SARS) in Children." *Paediatric Respiratory Reviews* 5(4): 275–88.
Lewnard, Joseph A., and Yonatan H. Grad. 2018. "Vaccine Waning and Mumps Re-Emergence in the United States." *Science Translational Medicine* 10(433).
Li, Fang. 2016. "Structure, Function, and Evolution of Coronavirus Spike Proteins." *Annual review of virology* 3(1): 237–61.
Li, Fang, Wenhui Li, Michael Farzan, and Stephen C. Harrison. 2005. "Structural Biology: Structure of SARS Coronavirus Spike Receptor-Binding Domain Complexed with Receptor." *Science* 309(5742): 1864–68.
Li, Lei, Xukun Lu, and Jurrien Dean. 2013. "The Maternal to Zygotic Transition in Mammals." *Molecular Aspects of Medicine* 34(5): 919–38.
Li, Ming O. et al. 2006. "Transforming Growth Factor-β Regulation of Immune Responses." *Annual Review of Immunology* 24: 99–146.
Li, Rong, and David F. Albertini. 2013. "The Road to Maturation: Somatic Cell Interaction and Self-Organization of the Mammalian Oocyte." *Nature Reviews Molecular Cell Biology* 14(3): 141–52.
Lin, Marie et al. 2003. "Association of HLA Class I with Severe Acute Respiratory Syndrome Coronavirus Infection." *BMC Medical Genetics* 4: 1–7.
Liu, Ding X., Jia Q. Liang, and To S. Fung. 2021. "Human Coronavirus-229E, -OC43, -NL63, and -HKU1 (Coronaviridae)." *Encyclopedia of Virology* (January): 428–40.
Liu, Guang Yaw et al. 2017. "Probing the Roles of Calcium-Binding Sites during the Folding of Human Peptidylarginine Deiminase." *Scientific Reports* 7(1): 1–14.
Liu, Jingxian, Kenneth Y. Chen, and Ee C. Ren. 2011. "Structural Insights into the Binding of Hepatitis B Virus Core Peptide to HLA-A2 Alleles: Towards Designing Better Vaccines." *European Journal of Immunology* 41(7): 2097–2106.

Liu, Jun et al. 2010. "Novel Immunodominant Peptide Presentation Strategy: A Featured HLA-A*2402-Restricted Cytotoxic T-Lymphocyte Epitope Stabilized by Intrachain Hydrogen Bonds from Severe Acute Respiratory Syndrome Coronavirus Nucleocapsid Protein." *Journal of Virology* 84(22): 11849–57.

Liu, Kefang et al. 2021. "Binding and Molecular Basis of the Bat Coronavirus RaTG13 Virus to ACE2 in Humans and Other Species." *Cell* 184(13): 3438-3451.e10.

Liu, Margaret A. 2019. "A Comparison of Plasmid DNA and MRNA as Vaccine Technologies." *Vaccines* 7(2): 37.

Liu, Yi Liang, Yu Hsiu Chiang, Guang Yaw Liu, and Hui Chih Hung. 2011. "Functional Role of Dimerization of Human Peptidylarginine Deiminase 4 (PAD4)." *PLoS ONE* 6(6).

London, Nir et al. 2011. "Rosetta FlexPepDock Web Server - High Resolution Modeling of Peptide-Protein Interactions." *Nucleic Acids Research* 39(SUPPL. 2): 249–53.

Lopez-Leon, Sandra et al. 2021. "More than 50 Long-Term Effects of COVID-19: A Systematic Review and Meta-Analysis." *Scientific Reports* 11(1): 1–12.

Lopez, Jamie A. et al. 2013. "Perforin Forms Transient Pores on the Target Cell Plasma Membrane to Facilitate Rapid Access of Granzymes during Killer Cell Attack." *Blood* 121(14): 2659–68.

Lu, Qing Bin et al. 2021. "The Differential Demographic Pattern of Coronavirus Disease 2019 Fatality Outside Hubei and from Six Hospitals in Hubei, China: A Descriptive Analysis." *BMC Infectious Diseases* 21(1): 1–12.

Lu, Roujian et al. 2020. "Genomic Characterisation and Epidemiology of 2019 Novel Coronavirus: Implications for Virus Origins and Receptor Binding." *The Lancet* 395(10224): 565–74.

Lu, Shan et al. 2021. "The SARS-CoV-2 Nucleocapsid Phosphoprotein Forms Mutually Exclusive Condensates with RNA and the Membrane-Associated M Protein." *Nature Communications* 12(1).

Luscombe, Nicholas M., Roman A. Laskowski, and Janet M. Thornton. 2001. "Amino Acid-Base Interactions: A Three-Dimensional Analysis of Protein-DNA Interactions at an Atomic Level." *Nucleic Acids Research* 29(13): 2860–74.

Madissoon, Elo et al. 2016. "Characterization and Target Genes of Nine Human PRD-like Homeobox Domain Genes Expressed Exclusively in Early Embryos." *Scientific Reports* 6: 1–14.

Maeda, Toru et al. 2011. "Conversion of Mechanical Force into TGF-β-Mediated Biochemical Signals." *Current Biology* 21(11): 933–41.

Maeso, Ignacio et al. 2016. "Evolutionary Origin and Functional Divergence of Totipotent Cell Homeobox Genes in Eutherian Mammals." *BMC Biology* 14(1): 1–14.

Makrygiannakis, D. et al. 2006. "Citrullination Is an Inflammation-Dependent Process." *Annals of the Rheumatic Diseases* 65(9): 1219–22.

Mandala, Venkata S. et al. 2020. "Structure and Drug Binding of the SARS-CoV-2 Envelope Protein Transmembrane Domain in Lipid Bilayers." *Nature Structural and Molecular Biology* 27(12): 1202–8.

Marfori, Mary et al. 2011. "Molecular Basis for Specificity of Nuclear Import and Prediction of Nuclear Localization." *Biochimica et Biophysica Acta - Molecular Cell Research* 1813(9): 1562–77.

Mariano, Giuseppina, Rebecca J. Farthing, Shamar L.M. Lale-Farjat, and Julien R.C. Bergeron. 2020. "Structural Characterization of SARS-CoV-2: Where We Are, and Where We Need to Be." *Frontiers in Molecular Biosciences* 7(December).

Mastroianni, L. 1999. "The Fallopian Tube and Reproductive Health." *Journal of Pediatric and Adolescent Gynecology* 12(3): 121–26.

McGinnis, William et al. 1984. "A Homologous Protein-Coding

Sequence in Drosophila Homeotic Genes and Its Conservation in Other Metazoans." *Cell* 37(2): 403–8.
McGinnis, William, and Robb Krumlauf. 1992. "Homeobox Genes and Axial Patterning." *Cell* 68(2): 283–302.
McGraw, Walker T., Jan Potempa, David Farley, and James Travis. 1999. "Purification, Characterization, and Sequence Analysis of a Potential Virulence Factor from Porphyromonas Gingivalis, Peptidylarginine Deiminase." *Infection and Immunity* 67(7): 3248–56.
McLeskey, S. B. et al. 1997. "Molecules Involved in Mammalian Sperm-Egg Interaction." *International Review of Cytology* 177: 57–113.
McMahan, Katherine et al. 2021. "Correlates of Protection against SARS-CoV-2 in Rhesus Macaques." *Nature* 590(7847): 630–34.
Méchin, Marie Claire et al. 2010. "Deimination Is Regulated at Multiple Levels Including Auto-Deimination of Peptidylarginine Deiminases." *Cellular and Molecular Life Sciences* 67(9): 1491–1503.
Memish, Ziad A. et al. 2013. "Family Cluster of Middle East Respiratory Syndrome Coronavirus Infections." *New England Journal of Medicine* 368(26): 2487–94.
Moelants, Eva A.V. et al. 2013. "Citrullination of TNF-α by Peptidylarginine Deiminases Reduces Its Capacity to Stimulate the Production of Inflammatory Chemokines." *Cytokine* 61(1): 161–67.
Moeller, Nicholas H. et al. 2022. "Structure and Dynamics of SARS-CoV-2 Proofreading Exoribonuclease ExoN." *Proceedings of the National Academy of Sciences of the United States of America* 119(9).
Mohamed, Mohamed O. et al. 2020. "Sex Differences in Mortality Rates and Underlying Conditions for COVID-19 Deaths in England and Wales." *Mayo clinic proceedings* 95(10): 2110–24.
Mondal, Santanu, and Paul R. Thompson. 2019. "Protein Arginine Deiminases (PADs): Biochemistry and Chemical Biology of Protein Citrullination." *Accounts of Chemical Research* 52(3): 818–32.
Monteiro, A. S., B. Schierwater, S. L. Dellaporta, and P. W.H. Holland. 2006. "A Low Diversity of ANTP Class Homeobox Genes in Placozoa." *Evolution and Development* 8(2): 174–82.
Mori, Masataka. 2007. "Regulation of Nitric Oxide Synthesis and Apoptosis by Arginase and Arginine Recycling." *Journal of Nutrition* 137(6): 3–7.
Moscarello, Mario A., Fabrizio G. Mastronardi, and D. Denise Wood. 2007. "The Role of Citrullinated Proteins Suggests a Novel Mechanism in the Pathogenesis of Multiple Sclerosis." *Neurochemical Research* 32(2): 251–56.
Moustaqil, Mehdi et al. 2021. "SARS-CoV-2 Proteases PLpro and 3CLpro Cleave IRF3 and Critical Modulators of Inflammatory Pathways (NLRP12 and TAB1): Implications for Disease Presentation across Species." *Emerging Microbes and Infections* 10(1): 178–95.
Nachat, Rachida et al. 2005. "Peptidylarginine Deiminase Isoforms Are Differentially Expressed in the Anagen Hair Follicles and Other Human Skin Appendages." *Journal of Investigative Dermatology* 125(1): 34–41.
Nam, Jongmin, and Masatoshi Nei. 2005. "Evolutionary Change of the Numbers of Homeobox Genes in Bilateral Animals." *Molecular Biology and Evolution* 22(12): 2386–94.
Neuman, Benjamin W. et al. 2011. "A Structural Analysis of M Protein in Coronavirus Assembly and Morphology." *Journal of Structural Biology* 174(1): 11–22.
Ng, M. H.L. et al. 2010. "Immunogenetics in SARS: A Casecontrol Study." *Hong Kong Medical Journal* 16(5 SUPP4): 29–33.
Ng, Oi Wing et al. 2016. "Memory T Cell Responses Targeting the SARS Coronavirus Persist up to 11 Years

Post-Infection." *Vaccine* 34(17): 2008–14.
Nguyen, Ninh T. et al. 2021. "Male Gender Is a Predictor of Higher Mortality in Hospitalized Adults with COVID-19." *PLoS ONE* 16(7 July): 1–6.
Niakan, Kathy K. et al. 2012. "Human Pre-Implantation Embryo Development." *Development* 139(5): 829–41.
Nieto-Torres, Jose L. et al. 2014. "Severe Acute Respiratory Syndrome Coronavirus Envelope Protein Ion Channel Activity Promotes Virus Fitness and Pathogenesis." *PLoS Pathogens* 10(5).
Ning, Shuo, Beiming Yu, Yanfeng Wang, and Feng Wang. 2021. "SARS-CoV-2: Origin, Evolution, and Targeting Inhibition." *Frontiers in Cellular and Infection Microbiology* 11: 1–19.
Noyes, Marcus B. et al. 2008. "Analysis of Homeodomain Specificities Allows the Family-Wide Prediction of Preferred Recognition Sites." *Cell* 133(7): 1277–89.
Oran, Daniel P., and Eric J. Topol. 2020. "Prevalence of Asymptomatic SARS-CoV-2 Infection. A Narrative Review." *Annals of Internal Medicine* 173(5): 362–68.
Örd, Mihkel, Ilona Faustova, and Mart Loog. 2020. "The Sequence at Spike S1/S2 Site Enables Cleavage by Furin and Phospho-Regulation in SARS-CoV2 but Not in SARS-CoV1 or MERS-CoV." *Scientific Reports* 10(1): 1–10.
Pal, Arumay, and Yaakov Levy. 2020. "Balance between Asymmetry and Abundance in Multi-Domain DNA-Binding Proteins May Regulate the Kinetics of Their Binding to DNA." *PLoS Computational Biology* 16(5): 1–19.
Pardi, Norbert, Michael J. Hogan, Frederick W. Porter, and Drew Weissman. 2018. "MRNA Vaccines-a New Era in Vaccinology." *Nature Reviews Drug Discovery* 17(4): 261–79.
Park, Su Eun. 2020. "Epidemiology, Virology, and Clinical Features of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2; Coronavirus Disease-19)." *Pediatric Infection and Vaccine* 27(1): 1–10.
Pasternak, Alexander O., Willy J.M. Spaan, and Eric J. Snijder. 2006. "Nidovirus Transcription: How to Make Sense...?" *Journal of General Virology* 87(6): 1403–21.
Paules, Catharine I., Hilary D. Marston, and Anthony S. Fauci. 2020. "Coronavirus Infections-More Than Just the Common Cold." *JAMA - Journal of the American Medical Association* 323(8): 707–8.
Payne, Susan. 2017. "Family Coronaviridae." In *Viruses: From Understanding to Investigation*, Elsevier Science, 149–58.
Piskacek, Martin. 2009. "Common Transactivation Motif 9aaTAD Recruits Multiple General Co-Activators TAF9, MED15, CBP and P300." *Nature Precedings*.
Piskacek, Martin, Marek Havelka, Martina Rezacova, and Andrea Knight. 2016. "The 9aaTAD Transactivation Domains: From Gal4 to P53." *PLoS ONE* 11(9): 1–16.
Piskacek, Martin, Tomas Otasevic, Martin Repko, and Andrea Knight. 2021. "The 9aaTAD Activation Domains in the Yamanaka Transcription Factors Oct4, Sox2, Myc, and Klf4." *Stem Cell Reviews and Reports* 17(5): 1934–36.
Radaev, Sergei et al. 2010. "Ternary Complex of Transforming Growth Factor-B1 Reveals Isoform-Specific Ligand Recognition and Receptor Recruitment in the Superfamily." *Journal of Biological Chemistry* 285(19): 14806–14.
Ralt, Dina et al. 1991. "Sperm Attraction to a Follicular Factor(s) Correlates with Human Egg Fertilizability." *Proceedings of the National Academy of Sciences of the United States of America* 88(7): 2840–44.
Rankin, Christopher T., Tracie Bunton, Ann M. Lawler, and Se Jin Lee. 2000. "Regulation of Left-Right Patterning in Mice by Growth/Differentiation Factor-1." *Nature Genetics* 24(3): 262–65.

Rasch, Elizabeth K., Rosemarie Hirsch, Ryne Paulose-Ram, and Marc C. Hochberg. 2003. "Prevalence of Rheumatoid Arthritis in Persons 60 Years of Age and Older in the United States: Effect of Different Methods of Case Classification." *Arthritis and Rheumatism* 48(4): 917–26.

Rasmussen, Michael et al. 2016. "Pan-Specific Prediction of Peptide–MHC Class I Complex Stability, a Correlate of T Cell Immunogenicity." *The Journal of Immunology* 197(4): 1517–24.

Rechiche, Othman, T. Verne Lee, and J. Shaun Lott. 2021. "Structural Characterization of Human Peptidyl-Arginine Deiminase Type III by X-Ray Crystallography." *Acta Crystallographica Section F: Structural Biology Communications* 77: 334–40.

Reed, Nilgun I. et al. 2015. "The αVβ1 Integrin Plays a Critical in Vivo Role in Tissue Fibrosis." *Science Translational Medicine* 7(288).

Robert-Guroff, Marjorie. 2007. "Replicating and Non-Replicating Viral Vectors for Vaccine Development." *Current Opinion in Biotechnology* 18(6): 546–56.

Rogers, George E., Harry W.J. Harding, and Ida J. Llewellyn-Smith. 1977. "The Origin of Citrulline-Containing Proteins in the Hair Follicle and the Chemical Nature of Trichohyalin, an Intracellular Precursor." *BBA - Protein Structure* 495(1): 159–75.

Roggers, G E, and D H Simmonds. 1958. "Content of Citrulline and Other Amino-Acids in a Protein of Hair Follicles." *Nature* 182(4629): 186–87.

Romero, Violeta et al. 2013. "Immune-Mediated Pore-Forming Pathways Induce Cellular Hypercitrullination and Generate Citrullinated Autoantigens in Rheumatoid Arthritis." *Science Translational Medicine* 5(209).

Rossant, Janet. 2001. "Stem Cells from the Mammalian Blastocyst." *Stem Cells* 19(6): 477–82.

Rota, Paul A. et al. 2003. "Characterization of a Novel Coronavirus Associated with Severe Acute Respiratory Syndrome." *Science* 300(5624): 1394–99.

Rudensky, A et al. 1991. "Sequence Analysis of Peptides Bound to MHC II Molecules." *Nature* 353(6345): 622–27.

Rudolph, Markus G., Robyn L. Stanfield, and Ian A. Wilson. 2006. "How TCRs Bind MHCs, Peptides, and Coreceptors." *Annual Review of Immunology* 24: 419–66.

Rydyznski Moderbacher, Carolyn et al. 2020. "Antigen-Specific Adaptive Immunity to SARS-CoV-2 in Acute COVID-19 and Associations with Age and Disease Severity." *Cell* 183(4): 996-1012.e19.

Saag, Kenneth G et al. 1997. "Cigarette Smoking and Rheumatoid Arthritis." *Annals of the Rheumatic Diseases* 56: 463–69.

Sadoff, Jerald et al. 2021. "Safety and Efficacy of Single-Dose Ad26.CoV2.S Vaccine against Covid-19." *New England Journal of Medicine* 384(23): 2187–2201.

Saijo, Shinya et al. 2016. "Monomeric Form of Peptidylarginine Deiminase Type I Revealed by X-Ray Crystallography and Small-Angle X-Ray Scattering." *Journal of Molecular Biology* 428(15): 3058–73.

Salamatbakhsh, Maryam, Kazhal Mobaraki, Sara Sadeghimohammadi, and Jamal Ahmadzadeh. 2019. "The Global Burden of Premature Mortality Due to the Middle East Respiratory Syndrome (MERS) Using Standard Expected Years of Life Lost, 2012 to 2019." *BMC Public Health* 19(1): 1–7.

Sali, A, and T L Blundell. 1993. "Comparative Protein Modelling by Satisfaction of Spatial Restraints." *Journal of Molecular Biology* 234(3): 779–815.

Sathananthan, Henry, Judith Menezes, and Sulochana Gunasheela. 2003. "Mechanics of Human Blastocyst Hatching in Vitro." *Reproductive BioMedicine Online* 7(2): 228–34.

Schaap, Mireille et al. 2013. "Genome-Wide Analysis of Macrosatellite Repeat

Copy Number Variation in Worldwide Populations: Evidence for Differences and Commonalities in Size Distributions and Size Restrictions." *BMC Genomics* 14(1).
Schellekens, Gerard A et al. 1998. "Citrulline Is an Essential Constituent of Antigenic Determinants Recognized by Rheumatoid Arthritis-Specific Autoantibodies." *J. Clin. Invest* 101(1): 273–81.
Schier, Alexander F. 2007. "The Maternal-Zygote Transition: Death and Birth of RNAs." *Science* 316: 406–7.
Schiller, John, and Doug Lowy. 2018. "Explanations for the High Potency of HPV Prophylactic Vaccines." *Vaccine* 36(32): 4768–73.
Segreto, Rossana, and Yuri Deigin. 2021. "The Genetic Structure of SARS-CoV-2 Does Not Rule out a Laboratory Origin: SARS-CoV-2 Chimeric Structure and Furin Cleavage Site Might Be the Result of Genetic Manipulation." *BioEssays* 43(3): 1–9.
Seo, Ho Seong. 2015. "Application of Radiation Technology in Vaccines Development." *Clinical and experimental vaccine research* 4(2): 145–58.
Seow, Jeffrey et al. 2020. "Longitudinal Observation and Decline of Neutralizing Antibody Responses in the Three Months Following SARS-CoV-2 Infection in Humans." *Nature Microbiology* 5(12): 1598–1607.
Sha, Qian Qian et al. 2020. "Dynamics and Clinical Relevance of Maternal MRNA Clearance during the Oocyte-to-Embryo Transition in Humans." *Nature Communications* 11: 4917.
Shang, Jian, Yushun Wan, et al. 2020. "Cell Entry Mechanisms of SARS-CoV-2." *Proceedings of the National Academy of Sciences of the United States of America* 117(21): 11727-34.
Shang, Jian, Gang Ye, et al. 2020. "Structural Basis of Receptor Recognition by SARS-CoV-2." *Nature* 581(7807): 221–24.
Shapiro, Robert, and Bert L. Vallee. 1992. "Identification of Functional Arginines in Human Angiogenin By Site-Directed Mutagenesis." *Biochemistry* 31(49): 12477–85.
Shi, Jing et al. 2018. "Affinity Maturation Shapes the Function of Agonistic Antibodies to Peptidylarginine Deiminase Type 4 in Rheumatoid Arthritis." *Annals of the Rheumatic Diseases* 77(1): 141–48.
Sipilä, Kalle et al. 2014. "Citrullination of Collagen II Affects Integrin-Mediated Cell Adhesion in a Receptor-Specific Manner." *FASEB Journal* 28(8): 3758–68.
Slade, Daniel J. et al. 2015. "Protein Arginine Deiminase 2 Binds Calcium in an Ordered Fashion: Implications for Inhibitor Design." *ACS Chemical Biology* 10(4): 1043–53.
Slade, Daniel J., Sachi Horibata, Scott A. Coonrod, and Paul R. Thompson. 2014. "A Novel Role for Protein Arginine Deiminase 4 in Pluripotency: The Emerging Role of Citrullinated Histone H1 in Cellular Programming." *BioEssays* 36(8): 736–40.
Sleath, Betsy et al. 2008. "Communication about Depression during Rheumatoid Arthritis Patient Visits." *Arthritis Care and Research* 59(2): 186–91.
Somers, Emily C., Sussie Antonsen, Lars Pedersen, and Henrik Toft Sørensen. 2013. "Parental History of Lupus and Rheumatoid Arthritis and Risk in Offspring in a Nationwide Cohort Study: Does Sex Matter?" *Annals of the Rheumatic Diseases* 72(4): 528–29.
Spitaleri, Andrea et al. 2008. "Structural Basis for the Interaction of IsoDGR with the RGD-Binding Site of αVβ3 Integrin." *Journal of Biological Chemistry* 283(28): 19757–68.
Spitz, François, and Eileen E.M. Furlong. 2012. "Transcription Factors: From Enhancer Binding to Developmental Control." *Nature Reviews Genetics* 13(9): 613–26.
Stadler, Sonja C. et al. 2013. "Dysregulation of PAD4-Mediated Citrullination of Nuclear GSK3β Activates TGF-β Signaling and Induces Epithelialto-Mesenchymal Transition in Breast Cancer Cells." *Proceedings of the National Academy of Sciences of the*

United States of America 110(29): 11851–56.
Statland, Jeffrey M. et al. 2015. "Milder Phenotype in Facioscapulohumeral Dystrophy with 7-10 Residual D4Z4 Repeats." *Neurology* 85(24): 2147–50.
Stern, Lawrence J. et al. 1994. "Crystal Structure of the Human Class II MHC Protein HLA-DR1 Complexed with an Influenza Virus Peptide." *Nature* 368: 215–21.
Storni, Tazio, and Martin F. Bachmann. 2004. "Loading of MHC Class I and II Presentation Pathways by Exogenous Antigens: A Quantitative In Vivo Comparison." *The Journal of Immunology* 172(10): 6129–35.
Straus, Anita H, William G Carter, Elizabeth A Wayner, and Sen-itiroh Hakomor. 1989. "Mechanism of Fibronectin-Mediated Cell Migration : Dependence or Independence of Cell Migration Susceptibility on RGDS-Directed Receptor (Integrin)." 183(1): 126–39.
Stuart, E T, C Kioussi, and P Gruss. 1994. "Mammalian Pax Genes." *Annual review of genetics* 28: 219–36.
Svingen, T., and K. F. Tonissen. 2006. "Hox Transcription Factors and Their Elusive Mammalian Gene Targets." *Heredity* 97(2): 88–96.
Takahashi, Seiichiro et al. 2007. "The RGD Motif in Fibronectin Is Essential for Development but Dispensable for Fibril Assembly." *Journal of Cell Biology* 178(1): 167–78.
Taki, Hirofumi et al. 2011. "Purification of Enzymatically Inactive Peptidylarginine Deiminase Type 6 from Mouse Ovary That Reveals Hexameric Structure Different from Other Dimeric Isoforms." *Advances in Bioscience and Biotechnology* 02(04): 304–10.
Tang, Xiaolu et al. 2020. "On the Origin and Continuing Evolution of SARS-CoV-2." *National Science Review* 7(6): 1012–23.
Tanikawa, Chizu et al. 2009. "Regulation of Protein Citrullination through P53/PADI4 Network in DNA Damage Response." *Cancer Research* 69(22): 8761–69.
Tanikawa, Chizu et al. 2012. "Regulation of Histone Modification and Chromatin Structure by the P53-PADI4 Pathway." *Nature Communications* 3: 676.
Tarcsa, Edit et al. 1996. "Protein Unfolding by Peptidylarginine Deiminase: Substrate Specificity and Structural Relationships of the Natural Substrates Trichohyalin and Filaggrin." *Journal of Biological Chemistry* 271(48): 30709–16.
Tarcsa, Edit et al. 1997. "The Fate of Trichohyalin: Sequential Post-Translational Modifications by Peptidyl-Arginine Deiminase and Transglutaminases." *Journal of Biological Chemistry* 272(44): 27893–901.
Tatler, Amanda L., and Gisli Jenkins. 2012. "TGF-β Activation and Lung Fibrosis." *Proceedings of the American Thoracic Society* 9(3): 130–36.
Teilum, Kaare, Johan G. Olsen, and Birthe B. Kragelund. 2009. "Functional Aspects of Protein Flexibility." *Cellular and Molecular Life Sciences* 66(14): 2231–47.
Temmam, Sarah et al. 2022. "Bat Coronaviruses Related to SARS-CoV-2 and Infectious for Human Cells." *Nature* 604(7905): 330–36.
Teo, Chian Ying et al. 2012. "Discovery of a New Class of Inhibitors for the Protein Arginine Deiminase Type 4 (PAD4) by Structure-Based Virtual Screening." *BMC bioinformatics* 13 Suppl 1(Suppl 17): S4.
Terakawa, Hiroshi, Hidenari Takahara, and Kiyoshi Sugawara. 1991. "Three Types of Mouse Peptidylarginine Deiminase: Characterization and Tissue Distribution." *Journal of Biochemistry* 110(4): 661–66.
Tian, Jing Hui et al. 2021. "SARS-CoV-2 Spike Glycoprotein Vaccine Candidate NVX-CoV2373 Immunogenicity in Baboons and Protection in Mice." *Nature Communications* 12: 372.

Tindale, Lauren C. et al. 2020. "Evidence for Transmission of Covid-19 Prior to Symptom Onset." *eLife* 9: 1–34.
Todeschini, Anne Laure, Adrien Georges, and Reiner A. Veitia. 2014. "Transcription Factors: Specific DNA Binding and Specific Gene Regulation." *Trends in Genetics* 30(6): 211–19.
Töhönen, Virpi et al. 2015. "Novel PRD-like Homeodomain Transcription Factors and Retrotransposon Elements in Early Human Development." *Nature Communications* 6: 1–9.
Tolkunova, Elena N. et al. 1998. "Two Distinct Types of Repression Domain in Engrailed: One Interacts with the Groucho Corepressor and Is Preferentially Active on Integrated Target Genes." *Molecular and Cellular Biology* 18(5): 2804–14.
Torlopp, Angela et al. 2014. "The Transcription Factor Pitx2 Positions the Embryonic Axis and Regulates Twinning." *eLife* 3: e03743.
Torres, Berta et al. 2021. "Impact of Low Serum Calcium at Hospital Admission on SARS-CoV-2 Infection Outcome." *International Journal of Infectious Diseases* 104: 164–68.
Triantafilou, Kathy, Timothy R. Hughes, Martha Triantafilou, and Paul P. Morgan. 2013. "The Complement Membrane Attack Complex Triggers Intracellular $Ca2^{+}$ Fluxes Leading to NLRP3 Inflammasome Activation." *Journal of Cell Science* 126(13): 2903–13.
Trolle, Thomas et al. 2016. "The Length Distribution of Class I-Restricted T Cell Epitopes Is Determined by Both Peptide Supply and MHC Allele-Specific Binding Preference." *J Immunol* 196(4): 1480–87.
Tsao, Yeou Ping et al. 2006. "HLA-A*0201 T-Cell Epitopes in Severe Acute Respiratory Syndrome (SARS) Coronavirus Nucleocapsid and Spike Proteins." *Biochemical and Biophysical Research Communications* 344(1): 63–71.
Tsuchida, Mamoru et al. 1993. "CDNA Nucleotide Sequence and Primary Structure of Mouse Uterine Peptidylarginine Deiminase: Detection of a 3′-untranslated Nucleotide Sequence Common to the MRNA of Transiently Expressed Genes and Rapid Turnover of This Enzyme's MRNA in the Estrous Cycle." *European Journal of Biochemistry* 215(3): 677–85.
Tucker-Kellogg, Lisa et al. 1997. "Engrailed (Gln50→Lys) Homeodomain-DNA Complex at 1.9 Å Resolution: Structural Basis for Enhanced Affinity and Altered Specificity." *Structure* 5(8): 1047–54.
Ulvestad, E et al. 1994. "HLA Class II Molecules (HLA-DR, -DP, -DQ) on Cells in the Human CNS Studied in Situ and in Vitro." *Immunology* 82(4): 535–41.
Uysal, Hüseyin et al. 2010. "Antibodies to Citrullinated Proteins: Molecular Interactions and Arthritogenicity." *Immunological Reviews* 233(1): 9–33.
V'kovski, Philip et al. 2021. "Coronavirus Biology and Replication: Implications for SARS-CoV-2." *Nature Reviews Microbiology* 19(3): 155–70.
Vályi-Nagy, István et al. 2021. "Comparison of Antibody and T Cell Responses Elicited by BBIBP-CorV (Sinopharm) and BNT162b2 (Pfizer-BioNTech) Vaccines against SARS-CoV-2 in Healthy Adult Humans." *GeroScience* 43(5): 2321–31.
Vastenhouw, Nadine L., Wen Xi Cao, and Howard D. Lipshitz. 2019. "The Maternal-to-Zygotic Transition Revisited." *Development* 146(11): dev161471.
Venter, J C et al. 2001. "The Sequence of the Human Genome." *Science* 291(5507): 1304–51.
Vikkurthi, Rajesh et al. 2022. "Inactivated Whole-Virion Vaccine BBV152/Covaxin Elicits Robust Cellular Immune Memory to SARS-CoV-2 and Variants of Concern." *Nature Microbiology* 7(JulY).
Vita, Randi et al. 2019. "The Immune Epitope Database (IEDB): 2018 Update." *Nucleic Acids Research* 47(D1): D339–43.

Vorobyov, Eugene, and Jürgen Horst. 2006. "Getting the Proto-Pax by the Tail." *Journal of Molecular Evolution* 63(2): 153–64.
Vossenaar, E. R. et al. 2004. "Expression and Activity of Citrullinating Peptidylarginine Deiminase Enzymes in Monocytes and Macrophages." *Annals of the Rheumatic Diseases* 63(4): 373–81.
Vuoristo, Sanna et al. 2022. "DUX4 Is a Multifunctional Factor Priming Human Embryonic Genome Activation." *iScience* 25(4).
Vuzman, Dana, Ariel Azia, and Yaakov Levy. 2010. "Searching DNA via a 'Monkey Bar' Mechanism: The Significance of Disordered Tails." *Journal of Molecular Biology* 396(3): 674–84.
Wada, Mitsunori. 1930. "On the Occurrence of a New Amino Acid in Watermelon, Citrullus Vulgaris, Schrad." *Journal of the Agricultural Chemical Society of Japan* 6(1–5): 32–34.
Wajnberg, Ania et al. 2020. "Robust Neutralizing Antibodies to SARS-CoV-2 Infection Persist for Months." *Science* 370(6521): 1227–30.
Wang, Huihui et al. 2020. "The Genetic Sequence, Origin, and Diagnosis of SARS-CoV-2." *European Journal of Clinical Microbiology and Infectious Diseases* 39(9): 1629–35.
Wang, Ning, Jian Shang, Shibo Jiang, and Lanying Du. 2020. "Subunit Vaccines Against Emerging Pathogenic Human Coronaviruses." *Frontiers in Microbiology* 11: 298.
Wang, Rui et al. 2022. "Emerging Vaccine-Breakthrough SARS-CoV-2 Variants." *ACS Infectious Diseases* 8(3): 546–56.
Wang, Shu, and Yanming Wang. 2013. "Peptidylarginine Deiminases in Citrullination, Gene Regulation, Health and Pathogenesis." *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* 1829(10): 1126–35.
Wang, Yanming et al. 2009. "Histone Hypercitrullination Mediates Chromatin Decondensation and Neutrophil Extracellular Trap Formation." *Journal of Cell Biology* 184(2): 205–13.
Wang, Yue-Dan et al. 2004. "T-Cell Epitopes in Severe Acute Respiratory Syndrome (SARS) Coronavirus Spike Protein Elicit a Specific T-Cell Immune Response in Patients Who Recover from SARS." *Journal of Virology* 78(14): 5612–18.
Wang, Yufeng et al. 2021. "Histone Citrullination by PADI4 Is Required for HIF-Dependent Transcriptional Responses to Hypoxia and Tumor Vascularization." *Science Advances* 7: eabe3771.
Wilcox, Allen J., Donna Day Baird, and Clarice R. Weinberg. 1999. "Time of Implantation of the Conceptus and Loss of Pregnancy." *N Engl Med* 340: 1796-99.
Williams, Christopher J. et al. 2018. "MolProbity: More and Better Reference Data for Improved All-Atom Structure Validation." *Protein Science* 27(1): 293–315.
Williams, T. M. 2001. "Human Leukocyte Antigen Gene Polymorphism and the Histocompatibility Laboratory." *Journal of Molecular Diagnostics* 3(3): 98–104.
Wipff, Pierre Jean, Daniel B. Rifkin, Jean Jacques Meister, and Boris Hinz. 2007. "Myofibroblast Contraction Activates Latent TGF-B1 from the Extracellular Matrix." *Journal of Cell Biology* 179(6): 1311–23.
Witalison, Erin E., Paul R. Thompson, and Lorne J. Hofseth. 2015. "Protein Arginine Deiminases and Associated Citrullination: Physiological Functions and Diseases Associated with Dysregulation." *Current Drug Targets* 16(7): 700–710.
Wolpert, Lewis et al. 2001. *Principles of Development*. Second. Oxford University Press.
Wrapp, Daniel et al. 2020. "Cryo-EM Structure of the 2019-NCoV Spike in the Prefusion Conformation." *Science* 367(6483): 1260–63.
Wright, Paul W. et al. 2003. "EPAD, an Oocyte and Early Embryo-Abundant Peptidylarginine Deiminase-like Protein That Localizes to Egg

Cytoplasmic Sheets." *Developmental Biology* 256(1): 74–89.
Wrobel, Antoni G. et al. 2022. "Evolution of the SARS-CoV-2 Spike Protein in the Human Host." *Nature Communications* 13(1): 1–7.
Wu, Fan et al. 2020. "A New Coronavirus Associated with Human Respiratory Disease in China." *Nature* 579(7798): 265–69.
Wu, Peng et al. 2020. "Real-Time Tentative Assessment of the Epidemiological Characteristics of Novel Coronavirus Infections in Wuhan, China, as at 22 January 2020." *Eurosurveillance* 25(3): 1–6.
Xiong, Jian-Ping et al. 2002. "Crystal Structure of the Extracellular Segment of Integrin αVβ3 in Complex with an Arg-Gly-Asp Ligand." *Science* 296(5565): 151–55.
Xu, Bo et al. 2020. "Suppressed T Cell-Mediated Immunity in Patients with COVID-19: A Clinical Retrospective Study in Wuhan, China." *Journal of Infection* 81(1): e51–60.
Xu, Jielin et al. 2010. "Iso-DGR Sequences Do Not Mediate Binding of Fibronectin N-Terminal Modules to Adherent Fibronectin-Null Fibroblasts." *Journal of Biological Chemistry* 285(12): 8563–71.
Xu, Wenqing et al. 1995. "Crystal Structure of a Paired Domain-DNA Complex at 2.5 Å Resolution Reveals Structural Basis for Pax Developmental Mutations." *Cell* 80(4): 639–50.
Xu, Yao et al. 2016. "Mutations in PADI6 Cause Female Infertility Characterized by Early Embryonic Arrest." *American Journal of Human Genetics* 99(3): 744–52.
Yan, Changhui et al. 2008. "Characterization of Protein-Protein Interfaces." *Protein Journal* 27(1): 59–70.
Yang, Joy T. et al. 1999. "Overlapping and Independent Functions of Fibronectin Receptor Integrins in Early Mesodermal Development." *Developmental Biology* 215(2): 264–77.
Yang, Mei et al. 2021. "Structural Insight Into the SARS-CoV-2 Nucleocapsid Protein C-Terminal Domain Reveals a Novel Recognition Mechanism for Viral Transcriptional Regulatory Sequences." *Frontiers in Chemistry* 8: 1–12.
Yang, Xinbo, Guobing Chen, Nan ping Weng, and Roy A. Mariuzza. 2017. "Structural Basis for Clonal Diversity of the Human T-Cell Response to a Dominant Influenza Virus Epitope." *Journal of Biological Chemistry* 292(45): 18618–27.
Yao, Zizhen et al. 2014. "DUX4-Induced Gene Expression Is the Major Molecular Signature in FSHD Skeletal Muscle." *Human molecular genetics* 23(20): 5342–52.
Yasuda, Hiroyuki et al. 2019. "17-B-Estradiol Enhances Neutrophil Extracellular Trap Formation By Interaction With Estrogen Membrane Receptor." *Archives of Biochemistry and Biophysics* 663: 64–70.
Yue, Lei et al. 2022. "A Third Booster Dose May Be Necessary to Mitigate Neutralizing Antibody Fading after Inoculation with Two Doses of an Inactivated SARS-CoV-2 Vaccine." *Journal of Medical Virology* 94(1): 35–38.
Zaki, Ali M. et al. 2012. "Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia." *New England Journal of Medicine* 367(19): 1814–20.
Zendman, Albert J.W. et al. 2007. "ABAP: Antibody-Based Assay for Peptidylarginine Deiminase Activity." *Analytical Biochemistry* 369(2): 232–40.
Zhai, Lukai, and Ebenezer Tumban. 2016. "Gardasil-9: A Global Survey of Projected Efficacy." *Antiviral Research* 130: 101–9.
Zhang, Chao, Abraham Anderson, and Charles DeLisi. 1998. "Structural Principles That Govern the Peptide-Binding Motifs of Class I MHC Molecules." *Journal of Molecular Biology* 281(5): 929–47.
Zhang, Jun et al. 2021. "Structural Impact on SARS-CoV-2 Spike Protein by D614G Substitution." *Science* 372(6541): 525–30.

Zhang, Xiaoqian et al. 2016. "Peptidylarginine Deiminase 1-Catalyzed Histone Citrullination Is Essential for Early Embryo Development." *Scientific Reports* 6: 1–11.

Zhao, Jie, Wei Cui, and Bao Ping Tian. 2020. "The Potential Intermediate Hosts for SARS-CoV-2." *Frontiers in Microbiology* 11: 1–11.

Zhong, Ying Fu, and Peter W.H. Holland. 2011. "The Dynamics of Vertebrate Homeobox Gene Evolution: Gain and Loss of Genes in Mouse and Human Lineages." *BMC Evolutionary Biology* 11: 169.

Zhou, Minghai et al. 2006. "Screening and Identification of Severe Acute Respiratory Syndrome-Associated Coronavirus-Specific CTL Epitopes." *The Journal of Immunology* 177(4): 2138–45.

Zhou, Peng et al. 2020. "A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin." *Nature* 579(7798): 270–73.

Zhou, Yebin, Nanette Mittereder, and Gary P. Sims. 2018. "Perspective on Protein Arginine Deiminase Activity-Bicarbonate Is a pH-Independent Regulator of Citrullination." *Frontiers in Immunology* 9: 1–8.

Zhu, Feng Cai et al. 2020. "Safety, Tolerability, and Immunogenicity of a Recombinant Adenovirus Type-5 Vectored COVID-19 Vaccine: A Dose-Escalation, Open-Label, Non-Randomised, First-in-Human Trial." *The Lancet* 395(10240): 1845–54.

Zhu, Zhixing et al. 2020. "From SARS and MERS to COVID-19: A Brief Summary and Comparison of Severe Acute Respiratory Infections Caused by Three Highly Pathogenic Human Coronaviruses." *Respiratory Research* 21(1): 1–14.

Zhuang, Zhen et al. 2021. "Mapping and Role of T Cell Response in SARS-CoV-2-Infected Mice." *Journal of Experimental Medicine* 218(4): e20202187.

ISBN 978-952-12-4254-0