

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Studies of highly conserved amino acid residues with fixed mutations unique to the human lineage.

Permalink

<https://escholarship.org/uc/item/6968w282>

Author

Vaill, Michael

Publication Date

2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Studies of highly conserved amino acid residues with fixed mutations unique to the human lineage.

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Biomedical Sciences with a Specialization in Anthropogeny

by

Michael Ivan Vaill

Committee in charge:

Professor Ajit Varki, Chair
Professor Gen-Sheng Feng
Professor Fred H. Gage
Professor Pascal Gagneux
Professor Kamil Godula
Professor Katerina Semendeferi

2021

Copyright
Michael Ivan Vaill, 2021
All rights reserved.

The dissertation of Michael Ivan Vaill is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

To:

To my parents, who taught me to question everything.

TABLE OF CONTENTS

Dissertation Approval Page	iii
Dedication	iv
Table of Contents	v
List of Figures	vii
Acknowledgements	viii
Vita	xi
Abstract of the Dissertation	xii
Chapter 1 Anthropogeny: the study of human origin	1
1.1 What is Anthropogeny?	1
1.2 Early beliefs and theories about human origins	2
1.3 Anatomical and physiological differences between humans and “great apes”	5
1.4 Immunological and molecular comparisons of human and “great ape” proteins.	6
1.5 Discovery of candidate gene differences	10
1.6 Comparative genomics and genetics	12
1.7 Comparative studies of gene expression and networks	14
1.8 Acknowledgements	16
Chapter 2 Sialic Acids and Siglecs: A Brief Introduction	17
2.1 All mammalian cells are covered in a dense, sialic acid decorated, glycan coat.	17
2.2 Siglecs are sialic-acid binding cell surface receptors primarily found on immune cells	19
Chapter 3 Human-specific Polymorphic Pseudogenization of <i>SIGLEC12</i> Protects Against Advanced Cancer Progression	21
3.1 Introduction	22
3.2 Results	24
3.3 Discussion	31
3.4 Materials and Methods	36
3.5 Acknowledgements	40
Chapter 4 A Uniquely Human Evolutionary Change in the Polysialyltransferase ST8Sia2	48
4.1 Introduction	49
4.2 Results	51
4.3 Methods	54

	4.4 Discussion	58
	4.5 Acknowledgements	61
Chapter 5	Novel Methods to Characterize the Length and Quantity of Highly Unstable PolySialic Acids	67
	5.1 Introduction	68
	5.2 Results and Discussion	73
	5.3 Materials and Methods	79
	5.4 Conclusions and Perspectives	83
	5.5 Acknowledgements	84
Bibliography		90

LIST OF FIGURES

Figure 2.1:	Schematic figure contextualizing sialoglycan biosynthesis and cell-surface presentation of Siglec receptors	20
Figure 3.1:	Siglec-XII Induction of expression of genes associated with cancer progression.	42
Figure 3.2:	Enhanced Expression and Unexpectedly High Frequency of Siglec-XII in carcinomas.	43
Figure 3.3:	Correlation between <i>SIGLEC12</i> genomic status and frequency or progression, only of late stage cancers.	44
Figure 3.4:	Population Genetic Analysis shows signatures of selection in and around the <i>SIGLEC12</i> locus.	45
Figure 3.5:	Simple analysis of urine allows screening for Siglec-XII protein expression status.	46
Figure 3.6:	Proposed evolutionary history of human Siglec-XII.	47
Figure 4.1:	Genome-wide comparison of human and chimpanzee orthologs	62
Figure 4.2:	Human and chimpanzee <i>ST8SIA2</i> share only 1 amino acid difference at N308K	63
Figure 4.3:	N308 is a highly conserved residue throughout vertebrates	64
Figure 4.4:	Modeling of human ST8Sia2 based on the crystal structure of human ST8Sia3	65
Figure 4.5:	Surface plasmon resonance binding analysis of BDNF and FGF2 binding by polySia-NCAM-Fc polysialylated by either human or chimpanzee ST8Sia2 .	66
Figure 5.1:	Overview of polysialic acid lactonization, and Lactonization-Protection-DMB procedure	85
Figure 5.2:	DMB derivatization of colominic acid also hydrolyzes α 2-8 linkages	86
Figure 5.3:	Lactonization protects high DP polysialic acids during release from mouse brain glycopeptides	87
Figure 5.4:	Lactonization-protection-DMB HPLC long-chain polysia analysis and EndoN-sensitive Neu5Ac analysis during analysis of early postnatal mouse brains. .	88
Figure 5.5:	DMB Analysis of EndoN-sensitive Neu5Ac	89

ACKNOWLEDGEMENTS

I will forever consider myself lucky to have spent these years under the mentorship of several notable individuals. My thesis advisor, Ajit Varki, welcomed me into his lab as I found my footing in science, guided my curiosity as I tackled ambitious projects, and exemplified how to remain steadfastly curious while navigating the hurdles implicit in research. During my years as a Varkian it has become very clear why Ajit's success has continued throughout the decades. Ajit fosters creativity as well as rigor, and independence as well as collaboration. The Varki lab is a family of remarkable individuals who graciously share their experience and talents. In particular, I thank Sandra Diaz for her welcoming and patient mentorship which began on the very first day of my rotation in the Varki lab. From radiolabeled monosaccharides to glycan-arrays, her expertise in sialic acid research is unparalleled and I consider myself immensely lucky to have shared a bay with her during these years. I also want to thank Kunio Kawanishi for the many insights into science and life, preferably shared during long runs, track workouts, and plates of nachos at Rock Bottom. My key co-authors in the lab, Masaya Hane and Dillon Chen with *ST8SIA2* and Shoib Siddiqui with *SIGLEC12*, are all outstanding scientists and friends to whom I am deeply thankful. I thank all of my committee members for their time and contribution over these years. I especially thank Pascal Gagneux, who has been an enduring role model during my years of graduate school. Pascal's mentorship spans topics from anthropogeny to glycobiology to zymology, and spans classrooms from East Africa to BRF2 to his backyard.

The Center for Academic Training in Anthropogeny (CARTA), and the Glycobiology Research and Training Center (GRTC) are each unmatched academic communities thanks to an extraordinary collection of faculty, trainees, and staff. I am grateful for all of the support that I received as a student of the Biomedical Sciences Program, the Graduate Specialization in An-

thropogeny, and as a member of the UC San Diego community. Late nights trouble shooting the HPLC with Sandra, long discussions at home with my roommate Phil, and weekend adventures in the mountains and deserts of California with my friends have all contributed to this journey. Finally, my parents and family have always stood behind me no matter where my adventures take me. Their life-long love and support encourages me to pursue my dreams. Those named above contributed extensively, but the support of countless others also made this PhD possible. Thank you all.

Chapter 1, in part, is currently being prepared for submission for publication of the material. Vaill, M., Kawanishi, K., Varki, N., Gagneux, P., Varki, A. (2021) Comparative Physiological Anthropogeny: Exploring Molecular Underpinnings of Distinctly Human Phenotypes. *Physiological Reviews*. The dissertation author is the first author of this paper.

Chapter 3, in full, is a reprint of material as it appears in: Siddiqui, S. S., Vaill, M., Do, R., Khan, N., Verhagen, A. L., Zhang, W., Lenz, H.-J., Johnson-Pais, T. L., Leach, R. J., Fraser, G., Wang, C., Feng, G.-S., Varki, N., & Varki, A. (2021). Human-specific polymorphic pseudogenization of SIGLEC12 protects against advanced cancer progression. *FASEB BioAdvances*, 3(2), 69–82. <https://doi.org/10.1096/fba.2020-00092>. The dissertation author is second author of this paper.

Chapter 4, in full, is currently being prepared for submission for publication of the material. Vaill, M., Hane, M, Naito-Matsui, Y., Davies, L., Kitajima, K., Sato, C., Varki, A., & Chen, D. (2021). A Uniquely Human Evolutionary Change in the Polysialyltransferase ST8Sia2. The dissertation/thesis author is the first investigator and author of this paper.

Chapter 5, in full, has been submitted for publication of the material as it may appear: Vaill, M., Chen, D., Diaz, S., & Varki, A. (2021) Novel Methods to Characterize the Length

and Quantity of Highly Unstable PolySialic Acids. The dissertation/thesis author is the primary investigator and author of this paper.

VITA

- 2014 Bachelor of Science in Biochemistry and Molecular Biology, University of Georgia
- 2021 Doctor of Philosophy in Biomedical Sciences with a Specialization in Anthropogeny, University of California San Diego

PUBLICATIONS

Vaill, M. I., Desai, B. N., & Harris, R. B. S. (2014). Blockade of the cerebral aqueduct in rats provides evidence of antagonistic leptin responses in the forebrain and hind-brain. *American Journal of Physiology. Endocrinology and Metabolism*, 306(4), E414-423. <https://doi.org/10.1152/ajpendo.00661.2013>

Siddiqui, S. S., **Vaill, M.**, Do, R., Khan, N., Verhagen, A. L., Zhang, W., Lenz, H.-J., Johnson-Pais, T. L., Leach, R. J., Fraser, G., Wang, C., Feng, G.-S., Varki, N., & Varki, A. (2021). Human-specific polymorphic pseudogenization of SIGLEC12 protects against advanced cancer progression. *FASEB BioAdvances*, 3(2), 69–82. <https://doi.org/10.1096/fba.2020-00092>

Kawanishi, K., Saha, S., Diaz, S., **Vaill, M.**, Sasmal, A., Siddiqui, S. S., Choudhury, B., Sharma, K., Chen, X., Schoenhofen, I. C., Sato, C., Kitajima, K., Freeze, H. H., Münster-Kühnel, A., & Varki, A. (2021). Evolutionary conservation of human ketodeoxynonulosonic acid production is independent of sialoglycan biosynthesis. *The Journal of Clinical Investigation*, 131(5). <https://doi.org/10.1172/JCI137681>

Siddiqui, S. S., **Vaill, M.**, & Varki, A. (2021). Ongoing selection for a uniquely human null allele of SIGLEC12 in world-wide populations may protect against the risk of advanced carcinomas. *FASEB BioAdvances*, 3(4), 278–279. <https://doi.org/10.1096/fba.2021-00036>

ABSTRACT OF THE DISSERTATION

Studies of highly conserved amino acid residues with fixed mutations unique to the human lineage.

by

Michael Ivan Vaill

Doctor of Philosophy in Biomedical Sciences with a Specialization in Anthropogeny

University of California San Diego, 2021

Professor Ajit Varki, Chair

Anthropogeny asks the deepest questions of human curiosity: “Where did we come from? How did we get here?” This dissertation of biomedical sciences takes advantage of molecular and cellular methods to investigate these questions. The human genome is highly similar to the genome of our closest living evolutionary relative, the chimpanzee (the human and chimp genomes are more similar than the mouse and rat genomes). Parsing the molecular mechanisms responsible for distinctly-human phenotypes is a daunting challenge. Over recent decades, it has become apparent that the biology of sialic acids, monosaccharides that coat

the surface of all cells, is rapidly evolving in the human species. *SIGLEC12* and *ST8SIA2* are two specific genes involved in sialic acid biology which each contain human-specific amino acid changes. *SIGLEC12* is a cell-surface sialic acid receptor which, in addition to a fixed amino acid change in the human lineage, is experiencing negative selection throughout human populations. We discovered implications of these evolutionary changes in human cancer risk and progression. *ST8SIA2* is a sialyltransferase that is involved in neuroplasticity, neurodevelopment, and in psychiatric and neurodegenerative disease. We identify a single amino acid change in the otherwise highly conserved *ST8SIA2* that has functional consequences and may contribute to many distinctly human properties of the brain.

Chapter 1

Anthropogeny: the study of human origin

1.1 What is Anthropogeny?

Anthropogeny refers to the transdisciplinary investigation of human origins and the evolutionary processes involved. While the earliest use of this term seems to be about 2 centuries old (Hooper, 1839), it fell into disuse in the 20th Century. Comparative Anthropogeny is a systematic comparison of humans and other living non-human hominids (so-called “Great Apes”) (Varki and Gagneux, 2017). This chapter presents historical perspective, briefly describing how the field progressed from the early evolutionary insights of Darwin, Wallace and Huxley to the current emphasis on in-depth molecular and genomic investigations of “human-specific” biology. The overall conclusion of this chapter is that while many genetic differences between humans and other hominids have been revealed in the last several decades (most information is available about human-chimpanzee differences), these findings still largely fail to explain the distinctive

anatomical and physiologically divergences of humans from other living hominids. Recent advances in genome editing may also prove useful for developing organoid and animal models for validated genetic traits. Throughout this chapter the terms "distinctly human," or "human-specific," refer to human phenotypes that appear to be derived within the hominin lineage, and absent or much less prominent in other hominids.

1.2 Early beliefs and theories about human origins

Questions about human origins have long challenged the world's brightest thinkers. Early religion-centric myths assumed that our species must have a divine origin, because our physiological and behavioral phenotypes appeared so distinct (at least, when viewed from our own perspective). Prior to the 1800s, and before the understanding of evolution by natural selection, humans were seen from the western religious perspective as being at the apex of creation, situated at the peak of a *scala naturae* — the medieval conception of a natural order, in which all living organisms were arranged in a linear order from simple to complex (Linné, 1758). This misconception is also reflected in the zoological term "primates," meaning "first in rank", (Dixson, 1981, Fossey, 1982) and still unconsciously persists in the way we refer to "lower" and "higher" organisms, and with phylogenetic trees depicted with "higher" organisms at the top and humans as the topmost.

What have we learned in the 150 years since Darwin and Huxley theorized the evolutionary relationship of humans with African apes? Early evolutionary thinking depended upon comparing easily observed phenotypes, such as anatomy and behavior (Huxley, 1863, Darwin, 1871). The dramatic differences between humans and apes in such phenotypes, led these thinkers to propose a long history of distinct evolution between humans and other living ape

species.

As the concepts of evolution began to be proposed by biologists some challenged the notion of humans as another species that descends from some ancient ancestors shared with the rest of the living world, especially in complete absence of fossils indicating any intermediate forms. It was hard to deny however, that humans shared much anatomical homology with primates and, in particular, with the African apes. Evolutionary concepts at that time depended entirely upon comparing anatomy and behavior (the latter usually in captivity prior to the first long-term observations in the wild). Years of taxonomical studies based on detailed comparative studies erroneously placed the great apes (African and Asian) into a monophyletic group called “pongids”. Advances in careful post-mortem autopsy were rapidly improving understanding of human anatomy and physiology, but also contributing methods for taxonomists and evolutionary thinkers to shape their own understanding of the relationships between species. Anatomical studies seemed to suggest a close relationship between humans and the African apes. In addition to anatomy, behavior was also considered an important factor for understanding the closeness of these relationships. In fact, comparing the anatomy and behavior of different species can inform views about the sequence by which different species diverged from last common ancestors. Living non-human hominids exhibit striking contrasts in social organization, mating system and social dominance patterns, and it is still unclear which of these are ancestral or derived, possibly even independently derived in more than one lineage. Regardless of these limitations, the comparative theories of 19th century evolutionary biologists, certainly also influenced by intellectual dogma of the time, suggested that while humans shared common evolutionary ancestors with the African apes, they must also have had a long and distinct history of separate evolution. We now understand that the phenotypic differences between two

species is not strictly related to the time since those species diverged. The transition between these two scientific perspectives is better understood in the context of the last 150 years of research in evolution, from Darwin to the revolution of modern molecular biology genetics and developmental biology, much aided by numerous fossils from Africa and beyond.

The modern evolutionary synthesis, which at the beginning of the 20th century merged Darwinian evolutionary theory with modern genetics paved the way for genetic sciences (Koonin, 2009). Sewall Wright, a key architect of the modern synthesis built on his insights into gene-environment interactions in evolution to introduce the adaptive landscape model (Wright, 1932, Wright, 1980). Recalling the evolutionary concept of an adaptive landscape is invaluable tool for modern scientists and physicians alike when investigating the physiology of their species of interest, in the case of medicine the human species and its natural variation, as well as the pathology of disease. The degree to which the adaptive landscape itself, has come under the pervasive impact of human niche construction due to human culture and associated technologies cannot be underestimated. There is also renewed interest in the Baldwin effect (Baldwin, 1896) as a mechanism for shaping human biology via human culture. The Baldwin effect suggests that phenotypic changes occurring in an organism as a result of its interaction with its environment become gradually assimilated into its developmental genetic or epigenetic repertoire (Feldman et al., 1996, Crispo, 2007).

1.3 Anatomical and physiological differences between humans and “great apes”

In *Evidence as to Man's Place in Nature* (Huxley, 1863) Huxley refers to the 1598 account of Portuguese sailor Duarte Lopez, illustrated by the brothers DeBry, as the first western report of the great apes. Englishman Andrew Battell described two apes he called *Engeco* and *Pongo* in 1613, deriving the names from native names for the chimpanzee and gorilla. As the seventeenth century European exploitation of Africa unfolded, European sailors further documented the African apes, and the first captive chimpanzees were delivered to western anatomists for scientific study. In 1699, after dissecting the first chimpanzee to arrive in England, anatomist Edward Tyson published *Orang-outang, sive homo sylvestri* (Tyson, 1699) which includes detailed illustrations and meticulously describes his observations (at this time *Orang-outang* described the red Asiatic and the black African varieties). Tyson compared these animals to both humans and monkeys with lists of gross similarities (48 included) and differences (34 included) between chimpanzee and human. The differences listed by Tyson include: flatness of the nose, cranial brow ridge, curvature of the spine, roundness of the kidney, and hairiness of the body. In this very earliest comparative study of the chimpanzee, and almost two centuries before evolutionary thinking would sweep scientific thought, Tyson reported his specimen was “more resembling a Man, than any other animal.”

During the following period chimpanzees were occasionally captured and transported to Europe. After suffering short periods in captivity the animals would die and be dissected as zoological specimens. (Traill, 1821) Publications supported Tyson's belief that the larynx and respiratory anatomy of the chimpanzee is not different enough from humans to explain their lack

of speech. Traill suggested that this uniquely human trait must be derived from neurological differences. By the end of the nineteenth century the popularity of the European menagerie led to an increase in the number of apes transported to the West. Advances in husbandry for captive chimpanzees increased their lifespan and enabled reproduction. In 1930 Yale psychologist Robert Yerkes founded the National Primate Research Center, now named after him and located at Emory University. Throughout the 20th century studies performed on captive chimpanzees at Yerkes and around the world illuminated distinctly human physiological and anatomical characteristics. From the 1950s on, large numbers of chimpanzees were captured across Africa and shipped to facilities in Asia, Europe and the Americas such as the Delta primate center at the University of Louisiana (Riss and Goodall, 1976). Some biomedical research was also conducted in Africa, including on hundreds of chimpanzees in the Belgian Congo (Stanleyville now Kisangani, hepatitis studies: polio vaccine efficacy and safety studies (Osterrieth, 2001), Liberia (Monrovia, New York Blood center: HBV vaccine safety studies (van den Ende et al., 1980), and Gabon (Franceville, CIRMF, ongoing studies on virology and immunity (Georges-Courbot et al., 1996, Ollomo et al., 1997).

1.4 Immunological and molecular comparisons of human and “great ape” proteins.

With the foundations of modern understanding of evolution in place, 20th century evolutionary biologists gained access to a new toolbox: molecular biology. Nuttall published perhaps the first molecular examination of the relationship between humans and non-human primates. Nuttall also observed that serum from rabbits immunized with human serum weakly cross reacts

with non-human primates. Collecting blood samples from as many animals as possible, Nuttall tested and measured the amount of precipitated protein in each reciprocal reaction (Nuttall and Inchley, 1904).

Allan Wilson at UC Berkeley and his student Vincent Sarich used an immunological dissimilarity index in a ground-breaking papers that would begin to change the public view on how distant or closely related humans are from our closest living relatives (Sarich and Wilson, 1967, Sarich and Wilson, 1967). Because relatively conserved regions of the genome are slowly but steadily undergoing changes over time, these regions serve as good representations of how long two species independently evolved since diverging from their last common ancestor. Today this type of calculation is trivial with the ability to rapidly compare sequence conservation.

Sarich and Wilson's immunological index allowed them to read this evolutionary record without having tools to sequence genetic material. The immunological dissimilarity was determined by reading microcomplement fixation, which only required a microscope. The microcomplement fixation method measured the cross-reactivity of an antibody raised in rabbits against the serum albumin of one species against the serum albumin of another species. The closer two proteins are in primary sequence, the more reactive an antibody raised against one will be against the other. Over evolutionary time, mutations in the primary protein sequence accumulate and can be read as a molecular clock by calibrating the rate of these mutations in years using a divergence time found in the paleontological record. Applying this molecular clock approach provided an estimated time since divergence of chimpanzees and humans of only about 3-5 mya, while the estimate for divergence from gorillas was about 16 mya, and 30 mya from old world monkeys molecular time scale (Wilson and Sarich, 1969).

This molecular calculation directly falsified the contemporary theories that humans had

a long, separate evolutionary history from other living apes by suggesting that humans and chimpanzees are instead very similar. Mary-Claire King, working with Alan Wilson, later demonstrated that the DNA and amino acid sequences of most human and chimpanzee serum proteins were in fact at least 99% identical, and that the other primates also shared highly similar sequences, gorillas, followed by monkeys and other species of animals. This suggested that many of the phenotypic differences observed between humans and chimpanzees must be explained by biological mechanisms other than just accumulations of coding DNA changes. Even at this stage, with limited knowledge about mammalian genetics and molecular biology, King and Wilson's classic paper "Evolution at Two Levels in Humans and Chimpanzees" (King & Wilson, 1975) suggested that to produce the set phenotypic differences between species, small evolutionary changes must affect the regulation of gene expression in level, time and space, rather than only the protein encoded by the gene.

The earliest molecular studies on the relationship of humans and chimpanzees were limited to the study of proteins, and cytological karyotyping. In 1973 Dorothy Warburton used trypsin-Giemsa G-banding to karyotype the chimpanzee and proposed a standard nomenclature for chimpanzee chromosomes (Warburton et al., 1973). Warburton and others observed that while some chromosome banding patterns were indistinguishable between humans and chimpanzees, other chromosomes appear more prone to undergo rearrangement, particularly certain regions that he called "hot-spots". These astute observations foreshadowed the structural pliability that remains one of the great challenges in understanding mammalian genomes. Mitchell and Gosden revealed that human chromosome 2 is the result of a fusion event between two ancestral chromosomes that are still separate in all the great apes (Mitchell and Gosden, 1978). During the process of sequencing the chimpanzee genome, a proposal to modify chro-

mosome nomenclature to reflect the true homology across hominid karyotypes became widely accepted and is now standard (McConkey, 2004).

By the 1980s Charles Sibley and Jon Ahlquist developed a method which used techniques in nucleic acid hybridization to determine the relative similarity of DNA sequences and infer taxonomic relationships, providing DNA evidence to parallel prior studies of proteins. The Sibley-Ahlquist method of taxonomy was accomplished by hybridizing DNA from two species and measuring the differences in melting temperature to determine relative sequence. During these years, the phylogeny of humans and our living hominoid relatives remained a topic of debate. In 1984 Sibley and Ahlquist applied their method to resolve the phylogeny of humans and the great apes (Sibley and Ahlquist, 1984). Their methods yielded the frequently quoted >98% sequence identity figure for human and chimpanzee DNA, which is however based on the exclusion of highly repetitive DNA (which comprises about half the entire genome).

With each decade's advances in technology, scientific evidence of the relationship between humans and the great apes compounded, and the public view of human origins was beginning to shift. In 1982 a meeting of scientists from around the world was held at the Pontifical Academy of Sciences in the Vatican to direct the official position of the Catholic Church regarding the evolutionary relationship of humans, our living relatives, and our extinct paleoanthropological relatives and ancestors (Lowenstein, 1982). The published proceedings represent one of the first to suggest a much younger divergence time between humans and chimpanzees (5-7 mya) than previously assumed (20 mya).

Increased insights into mutational patterns also confirmed the theory of neutral evolution (Kimura, 1991) in as much that most DNA changes occur without causing direct phenotypic effects. This realization creates an important challenge for identifying differences with adaptive

consequences. It put a brake on pan-adaptationist interpretation but also provided a background “neutral” mutations rate which allows one to detect outcomes of natural selection when deviations from such a background rate are detected.

Work throughout the 20th century suggested that human and chimpanzee proteins bear striking similarity, and that subtle genetic changes orchestrate the differences observed between humans and great apes. The latest whole genome data have revealed that human genomes actually differ by $\sim 5\%$ from those of chimpanzees and bonobos, due to the existence of large numbers of differentially duplicated or deleted non-coding genomic elements in each lineage. Importantly, the sequence similarity of $\sim 99\%$ for most expressed proteins has been confirmed. Whole genome comparisons have also led to the appreciation of structural variation, involving complex and nested duplication and deletion events of much larger genomic segments as an important source of both, evolutionary innovation and molecular pathogenesis (Britten, 2002, Chaisson et al., 2015).

1.5 Discovery of candidate gene differences

Until 2 decades ago, there were no defined specific genetic or molecular differences between humans and other hominids with clear-cut biological, biochemical, or physiological consequences. The first specific mechanistic difference identified was a fixed pseudogenization of a *CMAH* gene encoding a sialic acid modifying enzyme, which caused the lack one of the two major types of sialic acid that typically cover the mammalian cell surface (Chou et al., 1998, Irie et al., 1998). Human *CMAH* inactivation was a distinct genetic difference between humans and chimpanzees, that has since been determined to be involved in many human-specific phenotypes.

With the availability of high-quality whole genome data, we can now make a list of genetic loci with differences and select candidates of interest to investigate. What can we learn from investigating these rare differences between human and chimpanzee genomes? As we wrote in response to the first chimpanzee genome draft, this task is much like hunting for needles in a haystack (Varki and Altheide, 2005). Besides the small fraction made up of coding genes, the vast majority of the mammalian genome consists of regulatory elements, large regions of repeat sequences and duplications, and transposable elements. Pointing to specific genetic changes that contribute to the human phenomenon is a daunting task as it must include countless regions that are not directly encoding proteins. The detailed functions of the genome, already complex in its diverse types of genetic elements and sequences, is further obscured by chromatin organization and epigenetic modification of DNA and Histone tails. Furthermore, many genomic differences could represent the result of genetic drift or neutral evolution rather than the outcome of natural selection.

For these human-specific genetic changes, a process of logical analysis must be applied to select changes that may contribute to human-specific biology and disease. After this selection, in-depth investigations can take advantage of robust biological models available today, including laboratory animals and human and non-human cell lines, to determine the consequences. Several elements can be taken into consideration in primary selection process. Chief among them are known relations to human-specific phenotypes. Humans display many traits that radically differ from phenotypes observed in the chimpanzee and other non-human hominids, such as obligate bipedality, hairlessness and large brain size. Changes in genes that are known to be involved in these human specific phenotypes are obvious prospective candidates. Many diseases also appear to be human-specific, including certain types of cancers and

many infectious diseases, suggesting human-specific biological mechanisms associated with these pathologies. In rare cases, genetic mutations identified in individuals with hereditary or congenital disorders offer insight into human phenotypes. By taking advantage of these “clues” it is possible to predict which of the many changes found in highly conserved genetic regions may contribute to human-specific biology and disease. These types of clues were involved in the famous example of *FOXP2*, a transcription factor involved with a heritable speech disorder, and later found to contain human specific evolutionary changes that appear to be involved with our unusual linguistic abilities. Gene networks (Spiteri et al., 2007, Konopka et al., 2009, Hickey et al., 2019).

1.6 Comparative genomics and genetics

With the advent of improved DNA sequencing methods it became possible to produce a draft of an entire genome, and in 2005 the chimpanzee became the fourth mammalian genome published (Mikkelsen et al., 2005). Across all genomic regions, humans and chimpanzees share ~96% sequence similarity. In accordance with King and Wilson’s 20th century discovery, we share 99% of sequences that are directly responsible for encoding proteins. Out of the more than 20,000 protein coding genes found in the human and chimpanzee genomes, approximately 3000 code for completely identical amino-acid sequences, and across the proteome there is an average of only 1 amino acid difference per protein. Minor changes in amino acid sequence can completely alter a protein’s function, including abolishing an enzyme’s activity.

Ranking of regions in the human genome manifesting significant evolutionary acceleration showed that most of these “human-accelerated” regions (HARs) do not code for proteins. The most dramatic change is seen in HAR1, which is part of a novel RNA gene (*HAR1F*) that is

expressed specifically in the developing human neocortex (Pollard et al., 2006). This extremely high similarity is a result of strong selection pressures that act on peptide-coding sequences, and only about ~6 million years of independent evolution since the divergence of the human-chimp lineages. Additionally, protein-coding DNA only represent about 1.5% of the genome and is therefore a lesser target for de novo mutation. Small changes in regulatory DNA can produce dramatic consequences for phenotypes, a key example being the massive cortical expansion during recent human evolutionary history. In addition to HARs, there are also many large segmental duplications and deletions that produced functional consequences through human specific losses and gains of paralogs, pseudogenization, and gene-conversion. A 2011 study identified 510 sites that are conserved throughout primate evolution including in chimpanzees but have undergone complete deletion in the human lineage and coined these changes hCON-DELS (McLean et al., 2011).

Recently, high-resolution assemblies of the genomes of our closest living ape relatives were constructed and annotated using long-read RNA sequencing of ape iPSCs to allow for the identification of transcripts in each species without depending on the human reference for mapping and exon identification (Kronenberg et al., 2018). Prior to this, studies of great ape genomes were dependent on the human reference genome to map sequencing reads. This analysis allowed a high-quality snapshot of the genomic differences between the great apes of humans. For example, chimpanzee and gorilla genomes are slightly larger than those of humans, due to ape-specific parallel expansions of segmental duplications. De novo genome assemblies enable unbiased interpretation of genomic differences between species, see latest bonobo genome (Mao et al., 2021). It is important to remember that a mutation which causes a phenotype or disease in humans, may be benign in another species due to differences in

genomic background.

High-resolution sequence assemblies specifically enabled the discovery of lineage-specific structural variants including segmental duplications, inversions, short tandem repeats, and changes in retrotransposons. In addition to these novel structural variants, the de novo assembly of ape genomes allows for a higher-resolution picture of human-specific protein coding features. A comparison of the human genome annotation with cDNA construction from ape iPSCs identified 57 exons uniquely gained, and 13 exons lost in the human genome. These variants are prime candidates for functional validation in experimental systems (Kronenberg et al., 2018). Massively parallel reporter assays for cis-regulatory enhancers in different human cell lines are providing treasure troves of human-specific regulatory sequences, that differ even between humans and our archaic cousins the Neanderthals (Weiss et al., 2021).

1.7 Comparative studies of gene expression and networks

Identification of lineage-specific changes affecting genes with known functions facilitates a candidate gene approach to systematic investigation of specific mechanisms involved in distinctly human phenotypes. This approach is highlighted above. However, comparative genomic studies reveal that the most rapidly evolving sequences within the genome consist of regulatory elements. While the biological consequences of these changes can be more elusive to investigate than amino acid coding changes, gene expression studies have become an important tool in understanding human specific gene regulation.

With particular interest in the distinctive aspects of human cognition, analysis of gene co-expression networks has been employed to study human-specific patterns of regional and developmental gene expression in the brain. In 2004 Khaitovich et al. used gene expression

microarrays to compare gene expression across brain regions in a collection of tissues from humans and chimpanzees (Khaitovich et al., 2004). The following year, with the completion of the chimpanzee genome draft, it became apparent that changes in protein sequences and gene expression seem to show similar patterns between tissues (Khaitovich et al., 2005). Later analysis of this expression microarray data found that a small number of changes between human and chimpanzee transcription factors can produce coordinated changes in transcriptional networks in the brain (Nowick et al., 2009). A later study by Geschwind's group used next generation sequencing and gene expression microarrays to produce a higher resolution data for weighted gene co-expression network analysis (WGCNA) comparing brain regions of humans, chimpanzees, and rhesus macaque (Konopka et al., 2012). Besides detecting elevated levels of differential expression in the human frontal lobe, this study discovered that humans seem to have more complex transcriptional programs. Networks of human brain transcriptomes contained the greatest number of modules in their systems-level analysis. Human-specific changes in transcription factors associated expression modules, particularly in glial cells (Xu et al., 2018), are directly related to the human-specific developmental neoteny, the concept involving changes (usually delays) in developmental timing that result in biological novelty (Shea, 1989, Somel et al., 2009). Most recently, single-cell RNA-seq of cerebral organoids produced from human and chimpanzee induced pluripotent stem cells was used to further discriminate the significance of human-specific expression patterns in glia (Pollen et al., 2019) and differences in the control of neurogenesis-associated retrotransposon activity (Marchetto et al., 2013).

1.8 Acknowledgements

Chapter 1, in part, is currently being prepared for submission for publication of the material. Vaill, M., Kawanishi, K., Varki, N., Gagneux, P., Varki, A. (2021) Comparative Physiological Anthropogeny: Exploring Molecular Underpinnings of Distinctly Human Phenotypes. *Physiological Reviews*. The dissertation author is the first author of this paper.

Chapter 2

Sialic Acids and Siglecs: A Brief

Introduction

2.1 All mammalian cells are covered in a dense, sialic acid decorated, glycan coat.

All mammalian cells are coated with a diverse collection of glycans affixed to membrane lipids and proteins. This glycocalyx serves critical functions throughout development and life. Constituting a large hydrated outer coat on each cell, the glycocalyx serves a physical role of protecting the cell, and serves many specific functional roles such as interactions with pathogens, immune cells, soluble factors, cell surface receptors, and countless other known and yet to be discovered processes within the cell. The complex carbohydrates of glycoproteins are categorized into two basic categories, N-glycans, asparagine linked, and O-glycans, serine or threonine linked. These carbohydrates are synthesized and affixed to proteins in the ER-Golgi

pathway. A core glycan structure is first synthesized on a lipid anchor, before it is transferred to the target protein in the ER. After being transferred to the target protein, the glycan is further elongated and modified as the secreted or membrane protein passes through the Golgi system. Oligosaccharide synthesis occurs through the action of membrane-bound glycosyltransferases that transfer individual monosaccharides from sugar nucleotide donors found in the Golgi lumen to the growing chains. This process allows for the production a highly diverse cast of glycans with specialized functions while only requiring a relatively small number of genes.

This is distinct from the template-driven synthesis of nucleic acids and protein in which each biomolecule must first be transcribed from its own gene before undergoing modifications. This diversity is orchestrated by the regulation of sugar nucleotide transporters, glycosyltransferases and competing glycosidases, as well as enzymes controlling the many chemical modifications that sugars can be subjected to such as acetylation and sulfation (essentials). Glycosyltransferases function by recognizing the substrate domain or protein, as well as the underlying glycan structure. This underlying structure has two important characteristics: the sequence in which the monosaccharides are ordered, and the type of linkage through which they are attached. Glycosyltransferases recognize these characteristics and subsequently transfer another monosaccharide upon the glycan using a specific linkage. N-Glycans can be characterized by overall structural characteristics that further distinguish them from the core glycan that is transferred to the protein from the lipid anchor. This universal core, consisting of Man₃GlcNAc₂, can be extended by the addition of branching mannose to produce the auxiliary sugar branches of glycan structures known as “triantennary” or “tetrantennary.” These core chains are decorated and further elongated and branched by the addition of a variety of sugars including mannose, N-acetylglucosamine, galactose, and fucose in different linkage configurations. An important

functional addition is the “capping” of these branches with terminal sialic acid residues. These negatively charged sialic acid caps play a critical role in glycan recognition and interactions.

2.2 Siglecs are sialic-acid binding cell surface receptors primarily found on immune cells

Sialic acid-recognizing Ig-like lectins (Siglecs) are cell surface receptors typically expressed on innate immune cells, and binding ligands bearing sialic acids (Sias), a family of glycans prominently present at the terminal end of the glycan chains on cell surface and extra-cellular glycoconjugates (Varki, Schnaar, & Schauer, 2017). Siglecs may be found on other cell types, including the notable case of Siglec-XII which is covered in chapter 4 of this dissertation.

Most CD33-related Siglecs (CD33rSiglecs) have immunoinhibitory functions mediated by immunoreceptor tyrosine-based inhibitory motifs (ITIMs) and ITIM-like motifs in the cytosolic tail (Bochner & Zimmermann, 2015; Adams et al., 2017; Varki, Schnaar, & Crocker, 2017). Upon binding sialic acid ligands, these intracellular signaling motifs recruit protein tyrosine phosphatases Shp1 and Shp2 which subsequently participate in a variety of signaling pathways in immune cells to influence cellular activation (Bochner & Zimmermann, 2015; Adams et al., 2017; Varki, Schnaar, & Crocker, 2017). Siglecs play a role in the evolutionary struggle between human innate immunity, and ongoing coevolution of bacterial pathogens to evade and exploit systems of self/non-recognition. This struggle has produced high amounts of human-specific changes in the genes encoding Siglecs.

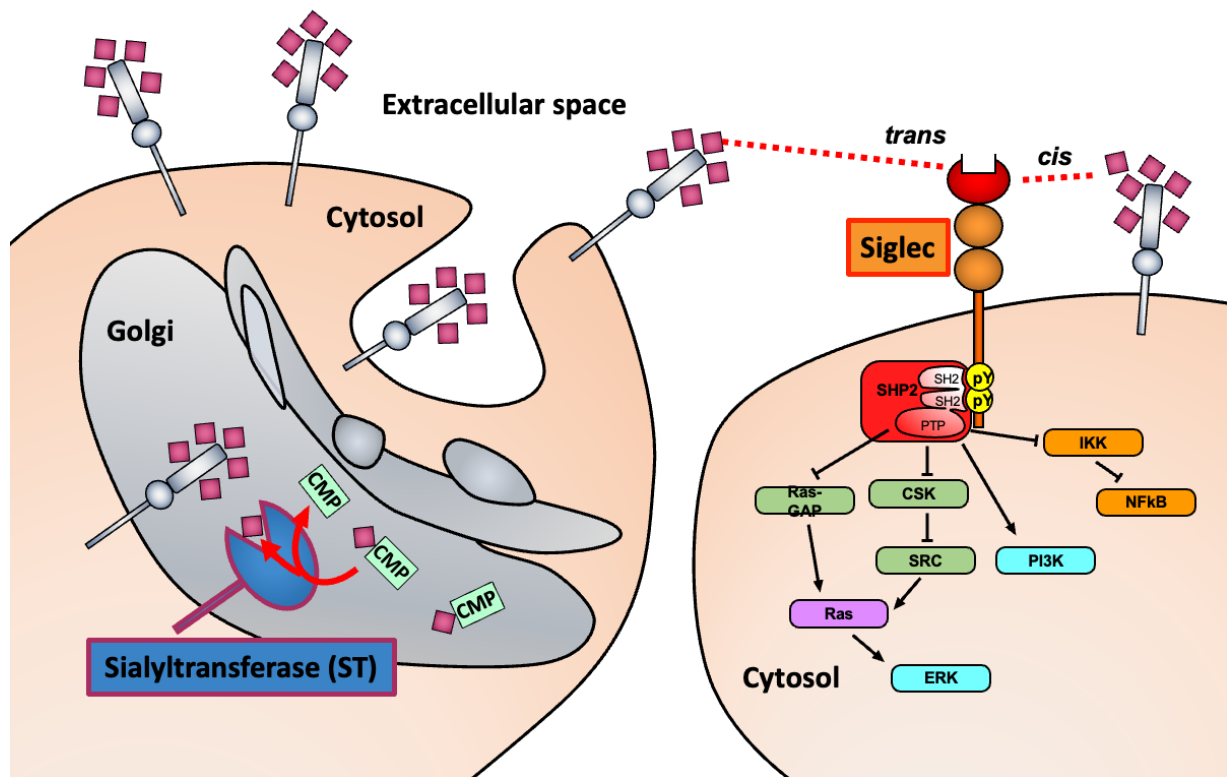


Figure 2.1: Schematic figure contextualizing sialoglycan biosynthesis (left) and cell-surface presentation of Siglec receptors (right). As glycoproteins pass through the Golgi, a series of biosynthetic enzymes process carbohydrate structures. Sialyltransferases, including ST8Sia2 which is the subject of chapter 4, transfer sialic acid from the cytidine monophosphate (CMP) sugar nucleotide donor CMP-Sia. Siglecs recognize sialic acid in both *cis* and *trans* and send signals into the cell (right), most commonly through the interaction of intracellular ITIM and ITAM motifs with cytosolic proteins including Shp1 and Shp2.

Chapter 3

Human-specific Polymorphic

Pseudogenization of *SIGLEC12*

Protects Against Advanced Cancer

Progression

Compared with our closest living evolutionary cousins, humans appear unusually prone to develop carcinomas (cancers arising from epithelia). The *SIGLEC12* gene, which encodes the Siglec-XII protein expressed on epithelial cells, has several uniquely-human features: a fixed homozygous missense mutation inactivating its natural ligand recognition property; a polymorphic frameshift mutation eliminating full-length protein expression in 60-70% of worldwide human populations; and, genomic features suggesting a negative selective sweep favoring the pseudogene state. Despite loss of canonical sialic acid binding, Siglec-XII still recruits Shp2

and accelerates tumor growth in a mouse model. We hypothesized that dysfunctional Siglec-XII facilitates human carcinoma progression, correlating with known tumorigenic signatures of Shp2-dependent cancers. Immunohistochemistry was used to detect Siglec-XII expression on tissue microarrays. PC-3 prostate cancer cells were transfected with Siglec-XII and transcription of genes enriched with Siglec-XII was determined. Genomic *SIGLEC12* status was determined for four different cancer cohorts. Finally, a dot blot analysis of human urinary epithelial cells was established to determine the Siglec-XII expressors versus non-expressors. Forced expression in a *SIGLEC12* null carcinoma cell line enriched transcription of genes associated with cancer progression. While Siglec-XII was detected as expected in 30-40% of normal epithelia, 80% of advanced carcinomas showed strong expression. Notably, >80% of late-stage colorectal cancers had a functional *SIGLEC12* allele, correlating with overall increased mortality. Thus, advanced carcinomas are much more likely to occur in individuals whose genomes have an intact *SIGLEC12* gene, likely because the encoded Siglec-XII protein recruits Shp2-related oncogenic pathways. The finding has prognostic, diagnostic and therapeutic implications.

3.1 Introduction

Humans appear unusually prone to develop carcinomas (cancers arising from epithelial cells), compared with our closest evolutionary cousins (“great apes”) (Schmidt, 1978; Puente et al., 2006; Varki & Varki, 2015). Here, we show an unexpected human-specific connection between advanced carcinomas and a member of the CD33-related family of Siglecs (Sialic acid binding Ig-like lectins) receptors (Bochner & Zimmermann, 2015; Adams, Stanczak, von Gunten, & Läubli, 2017; Varki, Schnaar, & Crocker, 2017).

Unlike other CD33-related Siglecs, the literature on Human Siglec-XII (encoded by the

gene *SIGLEC12*) is sparse (Angata, Varki, & Varki, 2001; Yu, Lai, Maoui, Banville, & Shen, 2001; Foussias et al., 2001; Mitra et al., 2011; McDonough et al., 2013). In fact, it has been largely ignored and even excluded from major reviews on Siglec biology (Varki, Schnaar, & Crocker, 2017; Macauley, Crocker, & Paulson, 2014; Bornhöfft, Goldammer, Rebl, & Galuska, 2018; Zhou, Oswald, Oliva, Kreisman, & Cobb, 2018), because the protein and the locus encoding *SIGLEC12* are atypical in several ways. First, the protein has two amino-terminal V-set domains (Yu et al., 2001), compared with only one in all other Siglecs. Second, there is a human-universal mutation of critical arginine residues in both V-set domains, rendering it unable to recognize Sias (hence the use of the Roman numeral XII for the protein, instead of Arabic numerals for functional Siglecs). Third, the Arg to Cys mutation of the V-set 1 domain is not present in orthologs of closely related “great apes” (chimpanzee, baboon, gorilla and orangutan) (Angata et al., 2001). Fourth, chimpanzee Siglec-12 preferentially recognizes a form of Sia (Neu5Gc) that was lost from the human lineage due to an independent fixed mutation of *CMAH* (Angata et al., 2001). Fifth, the *SIGLEC12* gene harbors a common polymorphic frameshift mutation causing truncation of Siglec-XII and/or alternate splicing (Flores et al., 2019) that causes loss of expression of full-length protein (Mitra et al., 2011) in the majority of humans. Finally, while the wild-type protein is expressed in some tissue macrophages, it is not found on other blood cell types, and is instead more prominent on epithelial cell surfaces (Mitra et al., 2011).

At first glance, the above features suggest a non-functional protein in the process of being eliminated from humans by pseudogenization. However, forced expression of human Siglec-XII in a genetically null human carcinoma cell line led to enhanced tumor growth in nude mice (Mitra et al., 2011). Furthermore, while human Siglec-XII does not recognize Sias, it still has ITIM and ITIM-like domains in the cytosolic tail that can be phosphorylated to recruit Shp1

and Shp2 phosphatases (Yu et al., 2001). Finally, genome-wide analysis of signals of selection in human populations identified polymorphisms that introduce nonsense-mediated-decay into human genes, including *SIGLEC12* (Yngvadottir et al., 2009). The human *SIGLEC12* locus appears to be undergoing selection favoring a null and/or truncated form, a possible example of the “less-is-more” hypothesis first proposed by Olson (Olson, 1999).

Here we focus on Siglec-XII expression in tumor and normal epithelia, identify genes upregulated upon Siglec-XII expression, address the predictive value of *SIGLEC12* status in cancer cohorts, and provide further evidence suggesting ongoing selection for the null state. Finally, we report a simple urine test to screen for the minority of individuals capable of full-length Siglec-XII expression.

3.2 Results

Expression of genes associated with cancer progression in a Siglec-XII expressing prostate cancer cell line.

Supporting the relevance of Siglec-XII expression in advanced cancer, we noted that in tissue sections where both malignant and adjacent normal tissue were present, Siglec-XII expression was higher in the malignant cells (one such example is shown in Fig 1A). To begin to explore the mechanism of action of this cell surface receptor, we took advantage of our earlier model system, Siglec-XII non-expressing PC-3 prostate carcinoma cells, which were transfected with a vector causing expression of full-length Siglec-XII. We had already observed larger tumors when this PC-3-Siglec-XII cell line was injected subcutaneously into the flanks of athymic nude mice, as compared to PC-3 cells transfected with vector alone (Mitra et al., 2011). We now

compared the RNA expression profiles between these two cell lines and found many genes to be differentially expressed (Fig 1B, 1C). Importantly, these differentially expressed genes were enriched for those known to play a role in cancer biology. A few of those up-regulated were *IDO1* (Indoleamine 2,3-Dioxygenase 1) (Zhai et al., 2015); *LCP1* (Lymphocyte Cytosolic Protein 1) (Koide et al., 2017); *BST2* (Bone Marrow Stromal Cell Antigen 2) (Mahauad-Fernandez, De-Mali, Olivier, & Okeoma, 2014); and *CEACAM6* (Carcinoembryonic Antigen Related Cell Adhesion Molecule 6) (Chiang et al., 2018), which are all involved in cancer progression. Among the down-regulated genes related to cancer progression were *CXADR* (Coxsackievirus and adenovirus receptor) (Stecker et al., 2011); *TACSTD2* (Tumor Associated Calcium Signal Transducer 2) (Wang et al., 2014); *CTSF* (Cathepsin F) (Ji et al., 2018); and, *ZNF43* (Zinc Finger Protein 43) (Jen & Wang, 2016) (Fig 1D). Taken together these data support the notion that Siglec-XII expression may facilitate late stage carcinoma progression in humans.

Enrichment analysis shows that similar gene sets are upregulated in Siglec-XII expressing cell lines and Shp2-expressing cell lines.

To query which molecular pathways are altered by Siglec-XII expression status, we performed gene set enrichment analysis (GSEA) on the expression profiles produced for each of the two PC3 cell lines. The GSEA result shows that transcriptional changes in Siglec-XII expressing cells affect the expression of many gene sets found in the Molecular Signature Database (MSigDB) (Subramanian et al., 2005). Of relevant interest are gene sets found in the Oncogenic Signatures collection, which were generated based on data produced by perturbing known cancer genes. The most dramatically enriched oncogenic signatures in Siglec-XII expressing cells include a set of genes altered in *KRAS*-addicted cancers (Singh et al., 2009), and a set of *TAZ*

associated genes found to be enriched in high-grade tumors (Cordenonsi et al., 2011).

To predict whether these gene set enrichments may be related to Siglec-XII activation of Shp2 signaling, we performed the same analysis on gene expression profiles of cancer cell lines that were found to be either dependent on Shp2 or independent of Shp2 in the development of resistance to MEK inhibition (Ahmed et al., 2019). Expression data for these samples was downloaded from the Gene Expression Omnibus (NCBI GEO accession number GSE121117). This analysis revealed that many of the same sets of genes that are enriched in Shp2-dependent cancers, are also enriched in our PC-3 cells with forced expression of Siglec-XII (Fig 1D). The full results of the enrichment analysis are available in the supplementary data.

Enhanced Expression and Unexpectedly High Frequency of Siglec-XII in carcinomas.

Immunohistochemical analyses for Siglec-XII showed low to moderate level expression in normal epithelia in a commercially available normal multi-tissue array with sample positivity of 35%. As the majority of human genomes harbor a homozygous null state abrogating full length protein expression, this low frequency is as expected. In contrast, in a multi-tissue array from the same source with multiple malignancies, we found an abundance of expression in carcinomas (malignancies arising from epithelia) (see examples in Fig 2A), with a much higher than expected frequency of Siglec-XII in cancers (80%) as compared to normal tissue (Fig 2B, C). Remarkably, 100% of the squamous cell carcinomas were positive (Fig 2D). This result was obtained from a mixed population group aged between 21-75 years. For this immunohistochemical analyses anti-Siglec-XII antibody clone 276 was used, which have been used and characterized earlier (Mitra et al., 2011). For analysis of subsets of tumor types, the samples

were divided into squamous, columnar, cuboidal, neural, and “uncategorized” (which included endothelium, mesothelium, and endocrine glands). Between 10-34 of each of these categories for both carcinoma and normal tissues were included in analysis. A paired t-test was used to determine significance between frequency of Siglec-XII+ staining in normal or carcinoma samples ($p < 0.01$). Tumors included in the “malignant” multi-tissue array are all likely to be advanced stage carcinomas. Given the prognosis of advanced carcinomas, our finding implies that minority of individuals who can express full-length Siglec-XII may be at the highest risk for dying with advanced carcinomas. A second panel of tissues, obtained from an independent source, and subjected to similar staining, confirmed this finding (Supplemental Table 1).

No correlation between *SIGLEC12* genomic status and frequency or progression of early stage cancers.

Next, we asked if Siglec-XII expression predicts early carcinoma risk or progression in a well-defined population. We had already reported that the incidence of prostate cancer was not different between men with different *SIGLEC12* genotypes (Mitra et al., 2011). From the same cohort, there is now a minimum of 5-year follow up available for many of these patients categorized into no evidence of disease (NED); Biochemical recurrence (BCR) and Metastasis (Met). There was no clear correlation of *SIGLEC12* status with the progression of these early stage carcinomas (Fig 3A). Of course, most of these cases were originally diagnosed by a measuring prostate specific antigen (PSA), which picks up many early stage cases that never progress in the lifetime of the individual (Eastham et al., 2003). Indeed, if we compare the patients with a poor outcome, versus those with no evidence of disease recurrence following prostatectomy (NED), we find that most of the patients (84 out of 122) detected by PSA screening did not have

disease progression at the time of follow up.

Seventh-day Adventists are members of a religious sect that do not smoke or consume alcohol and have a largely vegetarian diet with limited intake of red meat (Beeson, Mills, Phillips, Andress, & Fraser, 1989). As usual risk factors for cancer are limited, carcinoma incidence is much lower than in the general population. We genotyped the common *SIGLEC12* frameshift insertion mutation on genomic DNA from the peripheral blood cells of 54 Seventh-day Adventist cancer patients and 53 non-cancer patients (age and sex-matched). While we found more Siglec-XII expressers in the cancer group, this trend suggesting that Siglec-XII expressers may be more prone to develop carcinomas was not statistically significant (Fig 3B). Notably, many of these cancers were diagnosed at an early stage. Taken together, the data above suggest that the genomic status of *SIGLEC12* may not be correlated with the early cancer risk, but rather with late progression.

High frequency of *SIGLEC12* expression in advanced colorectal cancer cohort and correlation with overall survival.

Given the lack of significant correlation between *SIGLEC12* status and carcinoma risk or early stage carcinomas, we reasoned that there might instead be a correlation with late stage cancers. Indeed, in two stage IV colorectal cancer cohorts FIRE3 (592 patients from Germany and Austria) (Heinemann et al., 2014) and TRIBE (508 patients from Italy) (Loupakis et al., 2014) >80% of patients expressed Siglec-XII based on the frameshift mutation (Fig 3C, D). This recapitulates our initial immunohistochemistry-based findings. Furthermore, to see if *SIGLEC12* status has prognostic value, we checked overall survival in relation to Siglec-XII expression. Interestingly, in FIRE3 the overall survival increased from 28 months to 51 months between

Siglec-XII expressers and non-expressers (Fig 3E, F), i.e., a correlation between Siglec-XII expression and poor prognosis in late stage colorectal cancer.

Further evidence for selection at the *SIGLEC12* locus.

Earlier work suggested that this locus might be undergoing selection favoring the null state (Yngvadottir et al., 2009). To test this hypothesis, we examined the population level genetic variation and evidence of natural selection in and around the *SIGLEC12* locus and carried out tests that aimed to detect a selective sweep (Aakhus et al., 1990), deviation from neutrality (Tajima, 1989) and population differentiation (Weir & Cockerham, 1984) in three ethnic groups (YRI, CHB and CEU) (Pybus et al., 2014). Analysis of site-frequency spectrum provided a composite likelihood ratio indicative of a soft “selective sweep” acting on the gene throughout the overall human population (Fig 4A). Additionally, the common frameshift mutation (rs66949844) was present adjacent to this area. We also estimated population differentiation (measured as Wright's index of fixation; F_{ST}) and found moderately high F_{ST} values (0.3) compared to the average F_{ST} for genome-wide autosomal markers throughout the human population (Akey, Zhang, Zhang, Jin, & Shriver, 2002). The high F_{ST} value indicates stark differentiation of populations, which suggests directional selection (see Figure Legend, Fig 4B). Furthermore, we found an excess of rare alleles relative to a model of neutral evolution as indicated by negative Tajima's D values (Fig 4C); especially in CHB and YRI (African ancestry population). Individually none of these signals were very strong, but together, they suggest selection for the null state (note that ongoing selection for a null state would favor inactivating mutations, which would tend to mask the usual signatures of a selective sweep).

Dot blot analysis of bladder epithelial cells for detection of Siglec-XII status.

All the population studies above were handicapped by the fact that in addition to the common frameshift insertion mutation, we found other less common mutations that would nullify Siglec-XII expression. For example, another deleterious mutation (rs16982743) was observed at a global frequency of 18.6% that changes glutamine to a stop codon at the 29th position (McDonough et al., 2013). Thus, bi-allelic whole exome sequencing of *SIGLEC12* genomic DNA would be required for rigorous population studies. Even this approach could be confounded by selection for non-coding mutations that suppress gene expression in a given allele with an intact open reading frame. In addition, there is evidence for an alternately spliced truncated form of the protein (Flores et al., 2019). To facilitate future population and cancer cohort studies, it would be useful to have a simple assay to rapidly detect all mutations abrogating expression, without the need to do whole exome sequencing. We took advantage of the fact that among normal epithelial tissues tested by IHC, Siglec-XII was expressed in bladder epithelium, kidney tubules and salivary gland ducts, and detected the expression of Siglec-XII in cells isolated from saliva and urine (Fig 5A). It was determined via buccal swab genomic analysis that the *SIGLEC12* genomic status (*SIGLEC12* +/- or -/-) correlates with either Siglec-XII expression (+/-) or no expression (-/-). As expected, Siglec-XII expression in cells obtained from the urine of multiple healthy donors showed expression of Siglec-XII in the +/- genotypes and no expression in the Siglec-XII null genotypes. While there was a significant background in samples from saliva, results from dot blot screening of urinary cells were very clean (a typical example is shown in Fig 5B).

3.3 Discussion

We focused our initial work on a common polymorphic frameshift mutation in human populations with an allele frequency ranging from 38% in sub-Saharan Africans to 86% in the Native American population (Mitra et al., 2011). An earlier study (Yngvadottir et al., 2009) suggested selection on *SIGLEC12* based on the inactivating mutation rs16982743. However, another frameshift mutation (rs66949844) was present in the human population at an allele frequency of 59% (Consortium et al., 2015). Overall, this region of *SIGLEC12* showed reduced genetic diversity, which was supported by a sweep scan (Aakhus et al., 1990). These findings were concordant with results from a study in six different human populations (Schridder & Kern, 2016) showing a soft sweep in a region of *SIGLEC12*. The presence of excess rare alleles in and around a genomic region is also an indicator of a low level of population differentiation (Akey et al., 2002; Barreiro, Laval, Quach, Patin, & Quintana-Murci, 2008) further indicating the presence of purifying selection or balancing selection. Negative selection in *SIGLEC12* region was also evident from the result of Tajima's D (TD) especially in the YRI population (African ancestry).

Previous studies showed that while the non-Sia binding Siglec-XII can be expressed in *SIGLEC12* mutated PC-3 human prostate cancer cells, efforts to transfect the chimpanzee version of *SIGLEC12* or the arginine restored version of human *SIGLEC12* were not successful (Mitra et al., 2011). This could be either due to rapid turnover or selection against expression in vitro. Regardless, the non-Sia-binding full-length human Siglec-XII is clearly different functionally, allowing persistent surface expression in malignant cells by as yet unknown mechanisms. Chimpanzee and arginine-restored human *SIGLEC12* both display a preference for binding N-Glycolylneuraminic acid (Neu5Gc) (Fig 6A) (Angata et al., 2001) which is absent in humans due

to a homozygous fixed deletion in the gene *CMAH* (Chou et al., 1998). After losing its preferred ligand in an ancestral pre-human species, it is possible that Siglec-12 was susceptible to exploitation by a Neu5Gc-presenting pathogen (Fig 6B) or some other harmful form of activation, driving the fixation of the arginine mutation to produce the non-Sia-binding full-length Siglec-XII found in humans today (Fig 6C). It is worth pointing out that the reason the arginine mutation was fixed in humans remains unknown, and the consequences of the Sia-binding chimpanzee Siglec-12 deserve further study to contribute to our understanding of this evolutionary scenario. Finally, with the unusual derived trait of post-reproductive lifespan in modern humans, we propose that selection for survival in late life is driving the complete loss of the human *SIGLEC12* gene, as evidenced by the genomic signatures we report. Ongoing selection for null-state alleles may be acting to relieve the increased risk of advanced carcinomas produced by the archaic non-functional Siglec protein (Fig 6D).

It is also important to re-emphasize that the arginine and frameshift mutations of Siglec-XII do not occur in chimpanzees. Humans and chimpanzees are very similar in terms of genomic sequences but different phenotypically. Remarkably, while cancers are common in humans, few are reported in chimpanzees, and are usually lymphomas or soft tissue tumors, unlike those that arise in humans, who are instead prone to epithelial carcinomas (Schmidt, 1978; Puente et al., 2006; Varki & Varki, 2015). Here immunohistochemistry analyses indicate that Siglec-XII is highly expressed in advanced carcinomas, as compared to normal epithelium. Considering the multiple mutations reported (Mitra et al., 2011; McDonough et al., 2013) and others possible in the population, the overall expression in 35% in normal samples seems reasonable to represent the general mixed population. On the other hand, the high abundance of Siglec-XII in advanced carcinomas is remarkable, as is the high frequency of expression at >80%, in epithelial carcino-

mas. A second panel shows that certain types of carcinomas are even more likely to be found in Siglec-XII expressers; all squamous carcinomas show an 88% frequency, with lung showing 94% and cervix 90% (see Supplemental table 1).

To explore molecular mechanisms of Siglec-XII in cancer progression, we compared RNA expression patterns between *SIGLEC12* null PC-3 prostate cancer cells with and without transfection with a construct encoding Siglec-XII. One of the top hits among the up-regulated genes in RNA-Seq was *IDO-1* (Indoleamine 2,3-dioxygenase 1), an enzyme involved in conversion of tryptophan to kynurenine metabolites. This enzyme is highly upregulated in many types of cancers. It is known that a decrease in the levels of tryptophan and an increase in the levels of kynurenine leads to immunosuppression and enhanced tumor growth (Li, Zhang, Li, & Liu, 2017; Zhai et al., 2018; Zhai et al., 2015). The molecular mechanisms for the effects of *IDO-1* overexpression point towards maintenance of immunosuppression in tumor microenvironment, due to depletion of effector T cells and enrichment of regulatory T cells (Zhai et al., 2015). There has been a recent focus on IDO-1 targeting through small molecule inhibitors in preclinical and clinical settings (Li et al., 2017; Prendergast, Malachowski, DuHadaway, & Muller, 2017).

While Siglec-XII has lost its Sia-binding property, it still has the ability to recruit Shp1 and Shp2 (Yu et al., 2001). Shp2 (encoded by *PTPN11*) is a well-characterized oncogene that elicits cell growth, proliferation, tumorigenesis and metastasis (Bollu, Mazumdar, Savage, & Brown, 2017). Over-activation and activating mutations of Shp2 are known to be involved in breast cancer, leukemia and gliomas (Xu et al., 2005; Zhou, Coad, Ducatman, & Agazie, 2008; Bollu et al., 2017).

While not the primary objective of this study, we used RNA-sequencing data from our Siglec-XII expressing PC-3 cells to briefly investigate the hypothesis that Siglec-XII expression

enhancers tumor growth via Shp2. Using Gene Set Enrichment Analysis, we identified which pathways are altered by Siglec-XII protein expression. This analysis revealed that among the most dramatically enriched pathways are *KRAS* and *YAP/TAZ*. Notably, when we conducted a parallel analysis from a previously published transcriptomic study of Shp2-dependence the same pathways were the most highly enriched (Ahmed et al., 2019). Upregulation of these well-known oncogenic pathways in individuals with an intact *SIGLEC12* allele may explain the molecular mechanisms underpinning our discovery of increased frequency of Siglec-XII protein in advanced carcinomas.

We also performed population studies on four cancer cohorts. The first was a prostate cancer cohort we had studied earlier (Mitra et al., 2011). While a 5-year follow up for 122 patients was recorded in this cohort, we still did not see any positive correlation between *SIGLEC12* pseudogenization and outcome. One reason for this negative result may be that out of 122 patients only 10 developed metastasis (poor prognosis) and this might not be a sufficient number to find the relevance of *SIGLEC12* in prognosis. The second cohort we tested was a Seventh-day Adventist group and the lack of correlation could be due to two reasons. Firstly, most of the cancer patients in this group represented early stage cancer, where the effect of Siglec-XII is not pronounced. Secondly, many cancer risk factors such as intake of red meat, smoking, drinking alcohol etc., are minimal in this cohort, so it might be that Siglec-XII plays a role only when other obvious risk factors are involved. Overall, it appears that Siglec-XII does not play a role in early-stage carcinomas. In other populations, we discovered that the null state of the gene affects the prognosis of advanced carcinomas. Therefore, Siglec-XII expression is more likely to contribute to the advancement of benign neoplasia to deadly malignancies.

According to the well-established theoretical concept, natural selection occurs in pre-

reproductive or reproductive individuals (Ungewitter & Scrable, 2009). However, humans are a rare species that have prolonged post-reproductive lifespan (PRLS), and according to the 'grandmother hypothesis' inclusive fitness of infertile elderly caregivers can determine the fate of helpless grandchildren (Hawkes & Coxworth, 2013; Coxworth, Kim, McQueen, & Hawkes, 2015). We report selection acting on the *SIGLEC12* locus in human populations. This could be caused by deleterious fitness consequences of advanced carcinomas, which mostly occur late in middle to late life. To the best of our knowledge, our work is the first potential example of inclusive fitness effects selecting for cancer suppression, supporting a function for PRLS in humans. In contrast, an expansion in the number of *TP53* genes maybe providing late life protection against cancer risk in long-lived elephants (Abegglen et al., 2015). However, elephants do not have a PRLS, so the underlying selection mechanism must be different.

This first study of a very unusual phenomenon raises even more questions than answers. We do not know if there is still any definite ligand for Siglec-XII. It does not bind with Sias, but we cannot rule out its interaction with another unknown ligand(s). Conversely, we can also consider the hypothesis that this is a constitutively active receptor, which does not need any ligand for its activation. This aspect of Siglec biology is not extensively studied. Secondly, we did not study the signaling pathways mediated by SiglecXII-SHP2 axis. Third, we have not yet done the gene expression analysis in PC-3 cells with SHP-2 inhibitors. Moreover, a knockdown of *SIGLEC12* in a Siglec-XII expressing cell line will be useful. These are important aspects of Siglec-XII biology, which will be focused in further studies. Regardless, we have previously noted that triggering of endocytosis by antibodies against this receptor can deliver toxins into the cell (Mitra et al., 2011). In analogy to the targeting of Siglec-3/CD33 human leukemias (Lamba et al., 2017), a similar approach could be taken for treatment of late stage carcinomas. Our simple urine screen

should be of value in these and other clinical studies.

3.4 Materials and Methods

Immunohistochemistry Studies

Multi-tissue array slides were obtained from US Biomax (Rockville, Maryland), which were completely anonymized and consisted of normal human and cancer tissues. A second set of multi-tissue array slides were obtained from Novus Bio, which contained a variety of malignancies (about 476 different types) and also a set of normal multi-tissue array. The sections were de-paraffinized and blocked for endogenous biotin and peroxidase. The heat-induced epitope retrieval was performed with citrate buffer pH 6. A 5-step signal amplification method was used which includes application of mouse monoclonal anti-Siglec-XII antibody (clone 276), followed by biotinylated donkey anti-mouse, horseradish peroxidase (HRP), Streptavidin, followed by application of the enzyme biotinyl tyramide and then labeled Streptavidin. The AEC kit (Vector) was used as substrate, nuclear counterstain was with Mayer's hematoxylin, and the slides were aqueous mounted for digital photographs, taken using the Olympus BH2 microscope.

Buccal Swab

Healthy volunteers were recruited, and their buccal swab samples were used for DNA isolation with institutional review board (IRB) approval issued by the University of California, San Diego (UCSD). Before collection of the swab, the donors were asked to remove the mucous layer of their cheek by rubbing sterile gauze against it. Subsequently a sterile cotton tip was rubbed on the inner cheek cells for genomic DNA isolation. Genomic DNA was iso-

lated using the ChargeSwitch Buccal Cell gDNA isolation kit (Invitrogen, Cat No CS11021) according to the manufacturer's instructions. The PCR amplification for *SIGLEC12* gene was performed using the primers: Forward 5'-CAATGCAGAAGTCCGTGACGGTGCAGG-3' and reverse 5'-AGGATCAGGAGGGGCATCCAAGGTGC-3'. The Phusion High Fidelity Polymerase kit was used according to the manufacturer's instructions. The DNA amplicon was purified using QIAquick PCR purification kit (Qiagen, Cat no.-28106) and it was sent for sequencing at Eton Bio, San Diego using the sequencing primer: 5'-CTCTCTCTGGTGTCTCTGATGC-3' (reverse).

Dot Blot Using Urine from Healthy donors

Healthy volunteers donated 50 ml of first morning urine according to the IRB approved study. The urine sample was centrifuged at room temperature for 10 min at 500xg. The supernatant was removed, and cell pellet re-suspended in 100 μ l PBS. The sample was applied onto nitrocellulose membrane and immobilized by applying negative pressure. The membrane was blocked using 50% Licor solution (cat no-927-40000) + 50% PBST (PBS+0.01%Tween). After blocking, primary anti-Siglec-XII antibody (clone 1130) was applied at a dilution of 1:100-1:500. This clone of antibody have been used and characterized before (Mitra et al., 2011). The primary antibody dilution was performed in 90% Licor Solution + 10% PBST and incubation was carried out for 1 hour at RT. The membrane was then washed with 10 ml PBST 3 times for 5 min each. After washing, the membrane was incubated with anti-mouse-Licor-800 antibody at a dilution of 1:10000 in 90% Licor Solution + 10% PBST. The secondary antibody incubation was performed for 1 hour at RT in dark. After incubation the membrane was washed with PBST 3 times for 5 min followed by two times with PBS for 5 min. The band on the membrane was visualized by using Licor fluorescence scanning machine. Here, only PBS was used as a negative

control and Siglec-XII-Fc was used as a positive control.

***SIGLEC12* frameshift mutation in Seventh-day Adventist group**

The Seventh-day Adventist group is a diverse population group where the key carcinogenesis risk factors are less prevalent, such as consumption of red meat, alcohol and smoking. The genomic DNA was isolated from the peripheral blood cells of 53 cancer patients and 54 age-matched control subjects. The frame-shift deletion mutation of *SIGLEC12* was analyzed by first PCR amplifying the *SIGLEC12* locus using the primers 5'-ACCCCTGCTCTGTGGGAGAGT-3' (forward) and 5'AGGATCAGGAGGGGCATCCAAGGTGC-3' (reverse). The PCR was performed using Phusion High Fidelity Polymerase kit. The amplified product was purified using the QIAquick PCR purification kit (Qaigen, cat no.-28106) and sent for sequencing to EtonBio, San Diego, USA. The sequencing was performed using the primer: 5'-CTCTCTCTGGTGTCTCTGATGC-3' (reverse).

RNA-Sequence Analysis

PC-3 and PC-3-SigXII expressing cells were cultured to confluency in T25 flasks and mRNA was extracted from the cells using the Qaigen RNeasy plus mini kit extraction mini-elute kit (Cat no.- 74134). Transcriptomic analysis was performed on RNA libraries prepared from *SIGLEC12* and control PC3 cells using the TruSeq RNA Library Prep Kit v2. Each cell line was used to prepare 4 separate technical replicate libraries for sequencing. Libraries were sequenced at 1x50 bp on HiSeq 4000 (Illumina). Reads were mapped to human reference genome Hg19 using STAR v2.5.3a (Dobin et al., 2013). Mapped reads were counted at the gene level using featureCounts v1.5.2 (Liao, Smyth, & Shi, 2014) and counts were analyzed

using DESeq2 v1.14.1 (Love, Huber, & Anders, 2014). Differentially expressed genes with a p-value < 0.05 and fold change > 2 were then selected for further examination and gene set enrichment analysis using the GSEA software (Subramanian et al., 2005), the MSigDB v7.0 oncogenic signatures collection (C6) (Liberzon et al., 2011), and the Siglec-XII or Shp2 expression status as phenotype with 1,000 permutations.

Statistical Analysis

Graph Prism pad 5.0 was used. The chi-square test was performed on immunohistochemistry data, different cancer cohorts and a p value < 0.05 was considered as significant. For the RNA-Seq the two-way ANOVA was used as the statistically significant value. The p value < 0.05 and fold change of 2 was used as a cut-off for assessing the differentially expressed genes.

Population Genetics Analysis

Human genomes were accessed from the 1000 Genomes Project server (www.1000genomes.org/). Bed coordinates defining the *SIGLEC12* genomic regions were retrieved from build hg19 using the University of California, Santa Cruz (UCSC), genome browser. A region containing *SIGLEC12* gene in three different populations of West Africa, Northern European and East Asian ancestry (YRI, CHB, CEU), using the selection tools pipeline (Pybus et al., 2014). Statistical tests such as frequency-based method (Tajima's D) and population differentiation-based methods (FST) among three different populations were analyzed (Pybus et al., 2014). Each test is suited to detect selection at different timescales. Tajima's D is a commonly used summary of the site-frequency spectrum (SFS) of nucleotide polymorphism data and is based on the difference between two estimators of θ (the population mutation

rate $4N_e\mu$): nucleotide diversity that is the average number of pairwise differences between sequences, and Watterson's estimator, based on the number of segregating sites. A negative Tajima's D signifies an excess of low frequency polymorphisms, and indicates a population size expansion, selective sweep, and/or positive selection, or negative selection. A positive Tajima's D value indicates a decrease in population size and/or that balancing selection (Tajima, 1989). On the other hand, the estimator of population differentiation (F_{ST}), compares the variance of allele frequencies within and between populations (Holsinger & Weir, 2009). While large values of F_{ST} at a locus indicate complete differentiation between populations, which suggests directional selection, small values indicate the lack of differentiation in populations being compared, which might be an indicator of directional or balancing selection in both (Vitti, Grossman, & Sabeti, 2013). Human genome raw data for *SIGLEC12* (Huber, DeGiorgio, Hellmann, & Nielsen, 2016) was utilized for detecting Selective Sweep using SweepFinder2 (Aakhus, Stavem, Hovig, Pedersen, & Solum, 1990) which implements a composite likelihood ratio (CLR) test (Nielsen et al., 2005). The CLR uses the variation of the SFS of a region to compute the ratio of the likelihood of a selective sweep at a given position to the likelihood of a null model without a selective sweep. Tajima's D and sweep scans were visualized in Excel and F_{ST} were visualized in R studio platform and examined for evidence of deviation from the null expectation.

3.5 Acknowledgements

We thank Sandra Diaz for the excellent technical support in this work. We also thank Nivedita Mitra for useful discussion on the project. This study is primarily funded by

R01GM32373 (A.V.) and also by 5U01CA086402 (R.JL). MV was supported by the UCSD Genetics Training Program T32 GM008666, and Training Grant DK007202 in Gastroenterology.

Chapter 3, in full, is a reprint of material as it appears in: Siddiqui, S. S., Vaill, M., Do, R., Khan, N., Verhagen, A. L., Zhang, W., Lenz, H.-J., Johnson-Pais, T. L., Leach, R. J., Fraser, G., Wang, C., Feng, G.-S., Varki, N., & Varki, A. (2021). Human-specific polymorphic pseudogenization of *SIGLEC12* protects against advanced cancer progression. *FASEB BioAdvances*, 3(2), 69–82. <https://doi.org/10.1096/fba.2020-00092>.

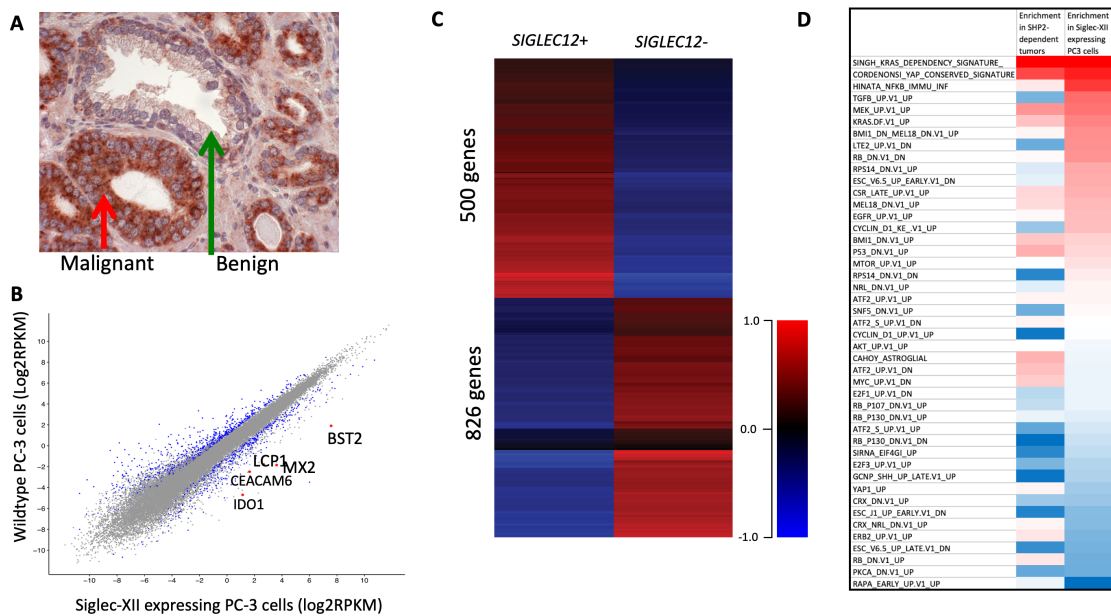


Figure 3.1: A) Example of tissue sections with adjacent normal and malignant cells from a prostate cancer patient B) Gene Expression in Siglec-XII transfected prostate cancer cells versus sham transfection (n=4). Differentially expressed genes highlighted in blue, and genes not differentially expressed are in grey color. A fold change of 2 and p value < 0.05 was used as a cut-off. C) The heatmap shows the differentially expressed genes in the Siglec-XII expressing PC-3 cell line versus parental PC-3 cells (n=4). D) Siglec-XII GSEA shows same top pathway expression as Shp2 positive tumors.

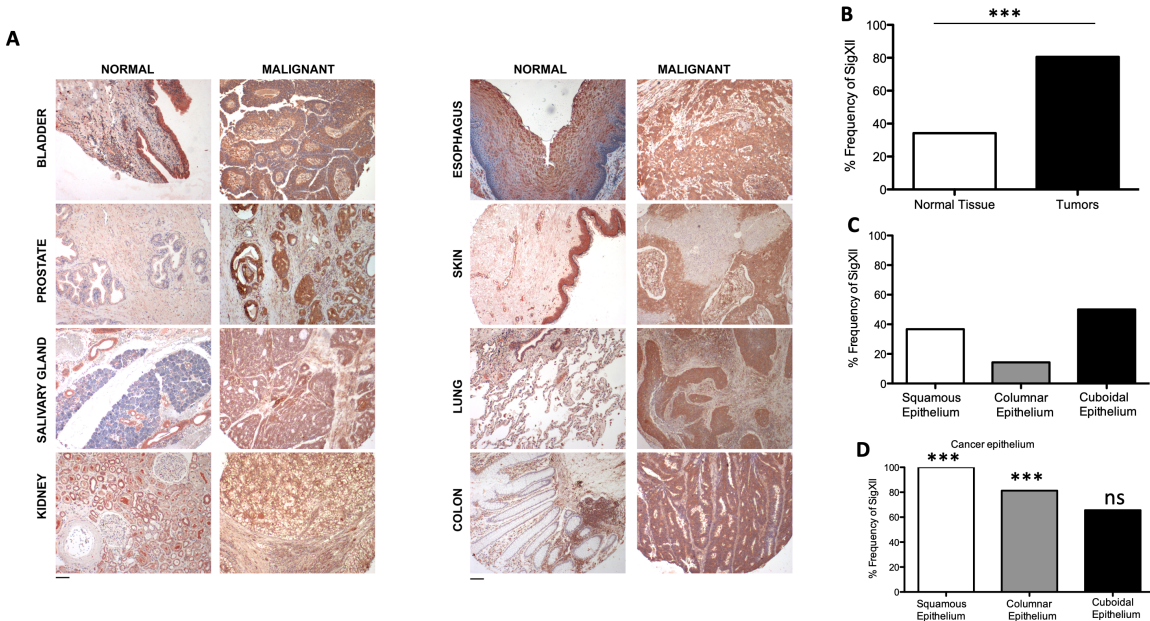


Figure 3.2: A) Expression of Siglec-XII studied in normal (benign) and cancer (malignant) human tissues using mouse monoclonal antibody clone 276 (See Materials and Methods). Representative examples of positive samples are shown. B) Frequency of Siglec-XII detection on normal and cancer tissues (n=97 for normal tissues and n=85 for tumor samples, ***p value<0.001). C) Normal epithelium divided into squamous (n=35), columnar (n=14), cuboidal (n=34). D) Carcinoma epithelium also divided into squamous (n=22), columnar (n=16) and cuboidal (n=32).

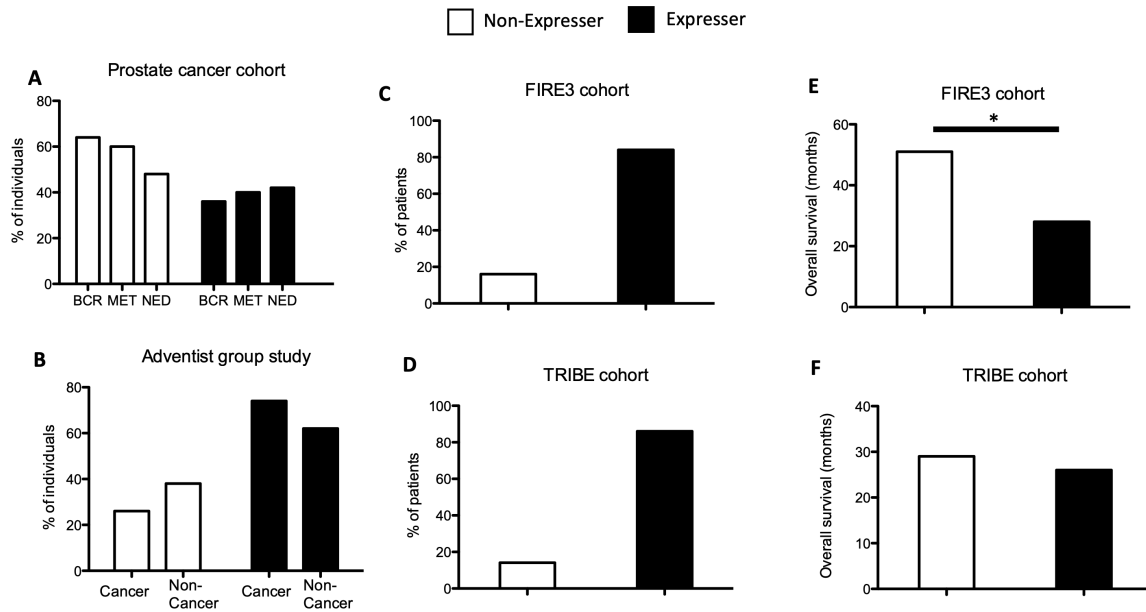


Figure 3.3: A) Prostate cancer patients diagnosed with PSA test followed up after 5 years. (NED-No evidence of disease: n=84, BCR- Biochemical Cancer Recurrence: n=28 and Met-Metastasis: n=10). B) Seventh-Day Adventist population where environmental risk factors for cancer are minimal. The percentage of patients with cancer and without cancer is shown to be either *SIGLEC12*^{-/-} (non-expresser, n for cancer=14, n for non-cancer=20) or *SIGLEC12*^{+/-} and *SIGLEC12*^{+/+} (expresser, n for cancer=40 and n for non-cancer=33). (C, D) Percentage of patients that are Siglec-XII expressers (*SIGLEC12*^{+/-} and *SIGLEC12*^{+/+}) or non-expressers (*SIGLEC12*^{-/-}) in the FIRE3 and TRIBE stage IV colorectal cancer cohorts (FIRE3 cohort: expresser n=85, non-expresser n=16 and TRIBE cohort: expresser n=177, non-expresser n=27). (E, F) Overall survival of colorectal cancer patients that are Siglec-XII expressers versus non-expressers (*p value < 0.05).

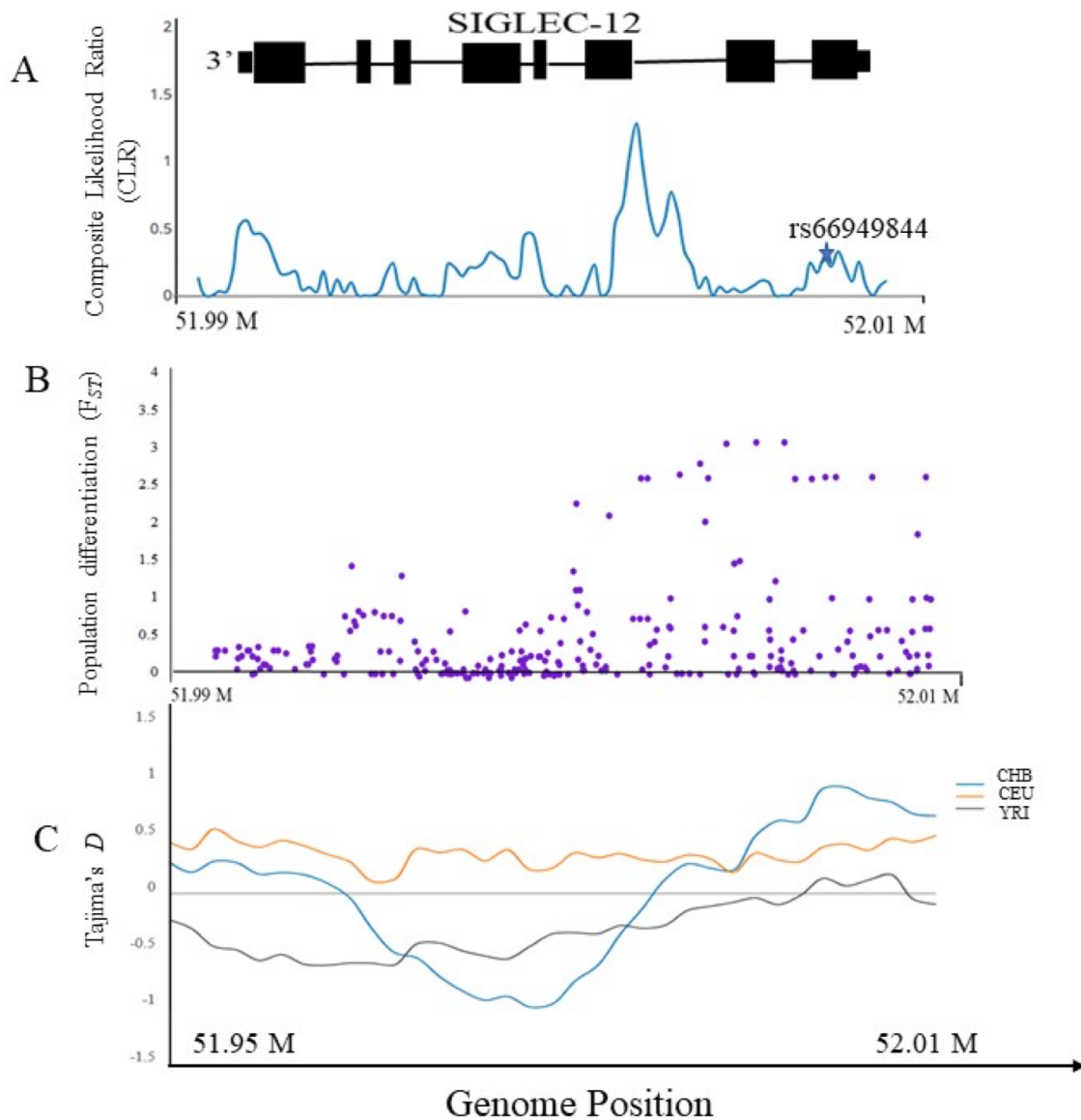


Figure 3.4: A) Signatures of "Selective sweep" in *SIGLEC12* in human population. The composite likelihood ratio (CLR) test of selective sweep based on the site frequency spectrum (SFS) is shown in blue. The star shown in the figure denotes the location of frameshift mutation. (Note. Schematic representation of *SIGLEC12* gene on top). (B) Estimation of Population differentiation " F_{ST} " (global) in three human populations (CHB, CEU, YRI). The purple dots represents F_{ST} values. (C) Estimation of Tajima's D in and around region of *SIGLEC12* in three human population are shown in different colors (Blue = CHB, Orange = CEU and Grey = YRI).

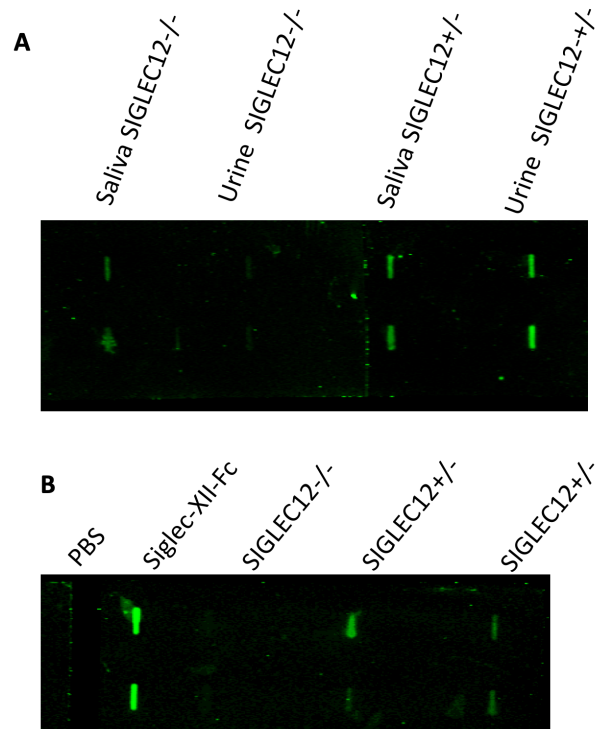


Figure 3.5: A) Urine and saliva samples were obtained from healthy individuals and used for checking protein expression of Siglec-XII by the dot blot. B) Urine samples from multiple healthy donors were used to check protein expression of Siglec-XII. One typical example is shown. (The whole blot was corrected uniformly for brightness using Photoshop, to match the visual appearance).

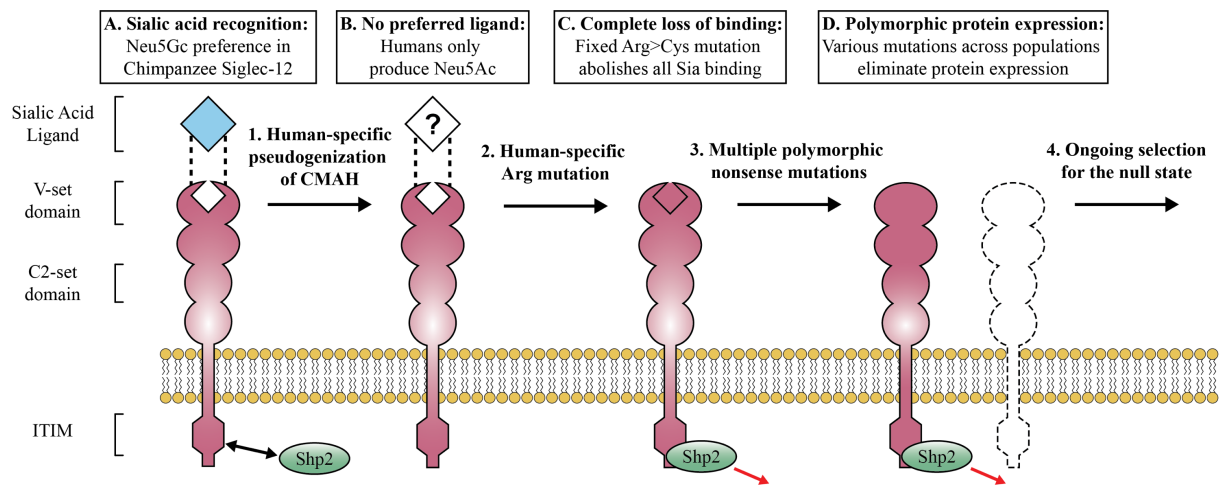


Figure 3.6: A) The last common human-chimpanzee ancestor and modern chimpanzees had a functional CMAH enzyme and an abundance of Neu5Gc-terminated cell-surface glycans. Chimpanzee Siglec-12 recognizes Neu5Gc through an arginine-dependent binding pocket in its terminal V-set domain. B) After divergence from chimpanzees CMAH was completely inactivated in human ancestor, leaving Siglec-12 with no endogenous ligand. C) Another unknown evolutionary event fixed a mutation in the critical arginine rendering human “Siglec-XII” incapable of binding any sialic acids, however, the full-length protein continues to recruit Shp2 and alter gene expression. D) Modern humans are experiencing purifying selection acting to increase the frequency of common null-state alleles across populations.

Chapter 4

A Uniquely Human Evolutionary Change in the Polysialyltransferase ST8Sia2

We have discovered a uniquely human evolutionary change in the amino acid sequence of ST8Sia2, one of the two highly conserved sialyltransferases that are responsible for the synthesis of polysialic acid (PolySia). This is the only amino acid that differs between the human ST8Sia2 enzyme and that of non-human hominids. While the ancestral N308 is conserved in all non-human mammals, and even in distantly related species through *Xenopus laevis*, K308 is fixed in all human populations and is present in the Neanderthal and Denisovan genomes, suggesting selection for this change during early human brain evolution. Despite hundreds of millions of years of conservation at this basic residue, the enzyme retains the ability to polysialylate human NCAM when K308 is mutated to an alanine, suggesting that conservation at this

site is involved in some other aspect of polysialylation. Proper spatial and temporal regulation of polySia synthesis is critical in neural development, plasticity, and regeneration, as well as implicated in psychiatric and neurodegenerative disease. Thus, the uniquely human change may be involved in some of the unique aspects of the human brain such as advanced cognitive abilities, developmental delay, and susceptibility to psychiatric and neurodegenerative diseases. We have identified a reduction in the stability of the enzyme as a result of the human mutation, implying that it may play a role in increased turnover and regulation of the transferase. We have also identified a marked difference in binding affinity of BDNF and FGF2 between polySia-NCAM molecules synthesized by the human and chimpanzee form of the enzyme. Our data indicate that the uniquely human change in ST8Sia2 has functional consequences upon which evolutionary selection could have acted.

4.1 Introduction

Despite major advances in genomics, we still largely lack an understanding of how specific genetic differences between humans and our closest ape relatives translate into molecular mechanisms representing the biological distinctions that make us human. Detailed molecular and biochemical examination is required to determine how any uniquely human genetic change affects physiology. Human brains are unique in our advanced cognition, unusual longevity, and susceptibility to psychiatric and neurodegenerative diseases. In depth investigation of genetic changes that occurred after our divergence from our closest living hominid relatives can reveal some of the specific factors contributing to the unique properties of our brains.

All mammalian cells are coated with a diverse collection of glycans affixed to membrane lipids and proteins. This glycocalyx serves critical functions throughout development and

life. Constituting a large hydrated outer coat on each cell, the glycocalyx serves a physical role of protecting the cell, and serves many specific functional roles such as interactions with pathogens, immune cells, soluble factors, cell surface receptors, and countless other known and yet to be discovered processes within the cell. Polysialic acid is an unusual homopolymer of up to 400 α 2-8-linked sialic acid residues that is added specifically to the fifth and sixth N-glycans of Neural Cell Adhesion Molecule (NCAM) i.e., polySia-NCAM. Polysialic acid is synthesized by the cooperative action of two polysialyltransferases, ST8Sia2 and ST8Sia4, coded by the human genes *ST8SIA2* and *ST8SIA4*, respectively. Polysialic acid synthesis is tightly regulated and notably restricted to specific tissues, primarily the central and peripheral nervous systems. In rodents, polysialic acid synthesis is relatively high and widespread throughout the nervous system during embryonic development; synthesis dramatically drops following birth, and after several months is restricted to specific areas of neurogenesis and plasticity. *St8sia2* knockout mice display dramatic neuroanatomical phenotypes in these regions (Angata et al., 2004).

Direct interaction of polySia with cell-surface receptors can also stimulate potentiation in Glu-N2A NMDA receptors (Kochlamazashvili et al., 2012) and inhibit potentiation in GluN2B NMDA receptors (Kochlamazashvili et al., 2010). Our lab identified a human-specific change in ST8Sia2, one of the two polysialyltransferases cooperatively responsible for synthesis of polySia. While this highly conserved enzyme shares a virtually identical sequence amongst all “great apes,” humans are a striking exception, harboring a single amino acid change (N308K) at a residue that is conserved in all mammals sequenced to date. This most unusual change is fixed (homozygous) in all human populations and is present in Neanderthal and Denisovan genomes, suggesting that it was selected for during early *Homo* brain evolution. GWAS studies identified a relationship between promoter region SNPs in the *ST8SIA2* gene and schizophre-

nia (Arai et al., 2006). polySia-NCAM expression is reduced in schizophrenic brains (Barbeau 1995). In addition, a mutated ST8Sia2 identified in a schizophrenia patient produces polySia with altered structure and function (Isomura 2011). *St8sia2*-KO mice show misguidance of infrapyramidal mossy fibers and ectopic synapse formation similar to altered hippocampal phenotypes observed in schizophrenic patients (Angata et al., 2004).

4.2 Results

Most human-chimp orthologous genes are highly similar and are undergoing purifying selection

In a genome-wide analysis of conservation between human and chimpanzee orthologs, we identify more than 3000 human and chimpanzee genes coding for completely identical amino-acid sequences. Across the proteome there is an average of only 1 amino acid difference per protein (Figure 1, black points). Genes that are undergoing rapid selection display very high (> 1) ratios of synonyms to nonsynonymous changes (Figure 1, blue points). Like many highly similar orthologous pairs, human and chimpanzee *ST8SIA2* are undergoing strong selective pressures. While a small number of genes have dramatic changes, and are undergoing rapid evolution after the human-chimp divergence, most differences lie in genes that are highly similar. We propose that selecting candidate genetic differences for investigation, based on predefined criteria, is a valuable path to understanding the molecular mechanisms behind uniquely human phenotypes. *ST8SIA2* is one example and is 0.3% (1/359 amino acids) divergent between humans and chimpanzees, with a Ka/Ks value of 0.0667.

Humans have a fixed mutation in ST8Sia2 at a highly conserved basic residue predicted to be located near the polybasic region

ST8Sia2 contains the four conserved sialyltransferase motifs found in all sialyltransferases: sialyl motif-long (SM-L), sialylmotif-short (SM-S), sialylmotif-very short (SM-VS), and sialylmotif-III (SM-III). In addition, the enzyme contains a domain found only in the ST8Sia family of α 2-8 sialyltransferases, known as the polysialyltransferase domain (PSTD), and the polybasic region (PBR).

To help predict the mechanism through which this mutation could impact the function of the protein, homology modeling was used to predict 3D structure of the enzyme and corresponding location of the residue in question. Recently the first crystal structure of an ST8 family transferase was solved – that of human ST8Sia3 (Volkers et al., 2015). ST8Sia3 is involved in synthesizing short oligoSia chains of up to 4 α 2-8-linked sialic acid residues and shares 59% sequence homology with ST8Sia2. The structure of ST8Sia2 was modeled using the Phyre2 protein fold recognition server (Kelley et al., 2015) against the structure of ST8Sia3 (structure reported in Volkers 2015, 59% sequence identity) and visualized using PyMOL software. The resulting proposed structure indicates that residue 308 sits at the C-terminus of motif-S helix 12, lying adjacent to the poly-basic region (PBR) involved in NCAM FN1 domain recognition (B). Superpositioning the Ig5 domain of NCAM, which contains the two polysialylated N-glycans, suggests that helix α 1 of the PBR, as well as the C-terminus of helix α 12 are directly oriented against NCAM Ig5 (A).

While homology modeling is rather limited in its accuracy, this exercise suggests some mechanistic hypotheses. If the mutation changes the interaction between the acceptor glycans on Ig5 it could alter the set of N-glycans targeted for polysialylation by ST8Sia2. If the enzyme

synthesizes polySia in a processed manner, as has been proposed (Nakata et al., 2006), a change in NCAM binding could change the association of the enzyme with its growing polySia product, affecting the DP of the polymers produced.

Human and African great ape ST8Sia2 are completely identical except for a single amino acid changed in humans

The mutation that we have identified in this highly conserved enzyme is very intriguing because this is the only amino acid that differs between the human ST8Sia2 enzyme and that of chimpanzees, bonobos and gorillas (Figure 2).

N308 is conserved deeply within vertebrate evolution

The chimpanzee-like N308 residue is conserved in all non-human mammals that have been sequenced as well as more distant evolutionarily related species including chicken and *Xenopus laevis*. The human K308 is fixed in all human populations and is also present in draft genome sequences of two extinct close relatives: Neanderthal and Denisovan (Figure 3).

Human ST8Sia2 produces longer polysialic acid chains

Preliminary studies have identified significantly different binding affinities of neurotrophic factors to PolySia-NCAM produced by either the human or chimpanzee form of ST8Sia-II. This indicates that there are some structural differences in the glycoprotein products of the enzyme dependent upon the single amino acid variation. Additionally, homology modeling of ST8Sia-II predicts that the altered residue lies in proximity to a region involved in substrate selectivity. The human mutation in ST8Sia-II changes the glycan structures of polySia-NCAM. The mutation

changes selectivity of the glycoprotein substrate, and/or the amount of polySia produced (degree of polymerization or number of polymers per N-glycan). The mutation could also change the glycan product by altering the substrate selectivity.

Polysialic acid synthesized by human *ST8Sia2* binds neurotrophic factors more strongly than that produced by the ancestral enzyme

Using surface plasmon resonance (SPR) we compared the binding affinity of polySia synthesized by either the human or chimpanzee enzyme. Our results reveal that polySia synthesized by the human enzyme show an increased affinity towards BDNF and FGF2 (Figure 5). This result confirms that there is some consequence upon the polySia glycan products as a direct result of the unique human coding change, plausibly encoding the mechanism that was selected for during the unique human evolution of *ST8SIA2*. Sensitivity towards BDNF requires polySia in hippocampal slice culture and cultured cortical neurons (Muller 2000, Vuskits 2001). In addition, polySia binding confers a protective effect upon the proteolytic processing and degradation of these factors (Hane 2015).

4.3 Methods

Genome wide Ka/Ks calculation and sequence alignments

Non-synonymous substitutions between human and chimpanzee were calculated from NCBI homologene data, and ensemble database (Yates et al., 2020), using biomaRt package (Durinck et al., 2009). All annotated human and chimpanzee protein coding sequences were downloaded as fasta sequences from NCBI using the Bioconductor package Biostrings. These

fasta sequences were then used for pairwise alignments (as seen in figures 2 and 3), and the total number of mismatched residues used to calculate percent identity. Lists of ENSG IDs were prepared for all genes with homologues in NCBI homogene, and attributes were retrieved from ensemble using biomaRt. Genes with an annotated chimp homologue were then sorted into bins representing 0.1% pairwise amino acid sequence identity. ggplot2 was used to plot the average Ka/Ks values for each bin against the number of homologous pairs in the bin to produce figure 1.

Surface Plasmon Resonance

The Au sensor surface is washed once with acetone and after drying, the chip is immersed in 10 μ M DBA in ethanol to form a self-assembly membrane (SAM) on the Au surface. After gently shaking for 30 min at room temperature, the sensor surface is washed with ethanol three times and allowed to dry. The chip is then placed in a solution of EDC and NHS (a 1:9 mixture of 130 μ M EDC in water and 144 μ M NHS in 1,4-dioxane) at room temperature for 30 min with gentle shaking to activate the SAM on the Au surface. After adding water, the surface is incubated for 5 min, and then washed the Au surface. The Au chip containing surface-activated SAM is placed on the sensor chip support using the sensor chip assembly unit, and is set in a Biacore 3000 instrument. After priming the system with water for 7 min, a 0.1 mg/ml protein A solution was loaded twice, each time for 7 min at a flow rate of 10 μ l/min. Immobilized streptavidin was monitored by measuring the resonance unit (RU) value, which typically reached 1300-1850 RU for protein A. To destroy excess activated groups, 1 mM ethanolamine was injected into the system for 7 min. After washing with HBS-EP (0.01M HEPES pH7.4 containing 0.15M NaCl, 3mM EDTA, and 0.0005% Surfactant P20), purified polySia-NCAM-Fc (0.1 mg/ml

in 500 mM HBS-EP) was injected into the system to allow immobilization on the Au surface. Immobilization of the polySia-NCAM-Fc was monitored based on the observed RU values, which typically reached approximately 850-1300 RU. NCAM-Fc derived from a mock transfectant was used as a negative control.

For analysis of the interactions between immobilized polySia-NCAM-Fc and the two neuroactive molecules, varying concentrations of BDNF (0-37.0 nM) or FGF2 (0-56.8 nM) in HBS-EP were injected over the polySia-NCAM-immobilized sensor chip surface at a flow rate of 20 μ l/min. After 120 s, HBS-EP was flowed over the sensor surface to monitor the dissociation phase. Following 180 s of dissociation, the sensor surface was fully regenerated by the injection of 10 μ l of 3 M NaCl. The analyses were performed three times. All values were analyzed using BIAevaluation software, and expressed as the mean \pm SD.

Plasmids

pPROTA-humanST8SIA2-V5 (pPROTA-hST8SIA2) and pPROTA-chimpST8SIA2-V5 (pPROTA-cST8SIA2) encoding soluble human and chimpST8SIA2 chimeric with protein A and V5, respectively, and pcDNA3.1-human ST8SIA2-V5/His (pcDNA-hST8SIA2) and pcDNA3.1-chimpST8SIA2-V5/His (pcDNA-cST8SIA2) encoding full-length human and chimp ST8SIA2, respectively, with V5 and 6xHis tags, were used in this study. Mutagenesis of pPROTA-hST8SIA2 and pcDNA-hST8SIA2 was performed using the QuickChange Site Directed Mutagenesis kit (Stratagene, CA, USA) and the primers listed in Table S1, resulting in the construction of pPROTA-cST8SIA2-V5 and pcDNA-cST8SIA2. A plasmid, pIG-NCAM, containing cDNA encoding NCAM-Fc was kindly gifted from Dr. Paul Crocker (University of Dundee, UK). The NCAM-Fc fragment was excised from pIG-NCAM with *Hind*III and *Not*I, purified and then inserted into the

HindIII and *NotI* sites of pcDNA4-*myc/His* to generate pcDNA4-NCAM-Fc. The sequences of all prepared constructs were confirmed by the deoxynucleotide chain termination method.

HPLC

PolySia-NCAM-Fc purified as described above was analyzed by SDS-PAGE/Western blotting using anti-polySia (12E3) (1 $\mu\text{g/ml}$) and anti-NCAM (H300) (0.2 $\mu\text{g/ml}$) antibodies at 4 °C. As the secondary antibody, either peroxidase-conjugated anti-rabbit IgG antibodies (0.1 g/ml) or anti-mouse IgG+M antibodies (0.4 $\mu\text{g/ml}$) was applied to the membranes, incubated for 60 min at 37 °C, and then color development was performed using standard reagents. The polysialylation state was also analyzed chemically by mild acid hydrolysis-anion-exchange chromatography analysis. Briefly, samples are hydrolyzed with 0.01 N trifluoroacetic acid (TFA) at 50°C for 1 h. To partially hydrolyzed samples, are added 20 μl of 0.01N TFA and 20 μl of 7 mM DMB solution in 5.0 mM TFA containing 1M 2-mercaptoethanol and 18 mM sodium hydrosulfite. These samples are incubated at 50 °C for 1h. The DMB-labeled samples are applied to an HPLC analysis. HPLC equipped with a DNA Pac PA-100 (4 x 250 mm, Dionex) anion exchange column and a fluorescence detector (FP-2025, JASCO). After equilibrating the column with 20 mM Tris-HCl (pH 8.0) at 26°C, sample is applied and eluted the DMB-labeled di/oligo/polySia at a flow rate of 0.5 ml/min with a linear gradient of NaCl (gradient from 0 to 0.6M) after 15 min wash with 20 mM Tris-HCl (pH 8.0). The fluorescence of the DMB-labeled samples are detected with a fluorescence detector at excitation 373 nm and emission 448 nm. The analyses were performed three times and all values are expressed as the mean \pm SD.

4.4 Discussion

As first discovered almost 50 years ago by King and Wilson, human and chimpanzee proteins bear striking similarity. Further insight into these differences was unlocked in 2005 when the chimpanzee became the fourth mammalian genome published. Across all genomic regions, humans and chimpanzees share 96% sequence similarity. In accordance with King and Wilson's 20th century discovery, we share 99% of sequences that are directly responsible for encoding proteins. This extremely high similarity is a result of high selection pressures that act on peptide-coding sequences, and only about 6 million years of independent evolution since divergence of the human-chimp lineages. Additionally, protein-coding DNA only represent about 1.5% of the genome and is therefore a lesser target for de novo mutation. Small changes in regulatory DNA can produce dramatic consequences for phenotype, for example the massive cortical expansion found in recent human evolutionary history. Early analysis of the human and chimpanzee genome drafts searched for regions with many human-specific changes, and found non-coding regulatory sequences that are highly conserved across mammals but show accelerated accumulation of changes in humans. These short genomic regions are known as human accelerated regions (HARs), and some were later shown to have human-specific function as neurodevelopmental enhancers. In addition to HARs, there are also many large segmental duplications and deletions that produced functional consequences through human specific paralogs, pseudogenization, and gene-conversion.

Much work has been accomplished to identify and characterize these dramatic structural changes and regions undergoing rapid evolution, however, it remains that most of the uniquely-human genetic changes are found dispersed throughout the genome. High Ka/Ks values are mostly relevant in genes that also display highly divergent amino-acid sequences. This leaves

the majority of amino acid coding differences between human and chimpanzee genomes in genes that are highly conserved ($Ka/Ks < 0.1$) and are not experiencing ongoing selection. Genomic signatures of recent and ongoing selection can be leveraged to systematically identify important sequences in the above categories, however, more ancient changes that are fixed in a lineage for more than several hundred thousand years escape such tests.

For these remaining human-specific genetic changes, a process of logical analysis must be applied to select changes that may contribute to human-specific biology and disease. After this selection, in-depth investigations can take advantage of robust biological models available today to determine the consequences. Several elements can be taken into consideration in primary selection process. First is known relations to human-specific phenotypes. Humans display many traits that are vastly departed from phenotypes observed in the chimpanzee and other non-human hominids, such as hairlessness and large brain size. Changes in genes that are known to be involved in these human specific phenotypes are prospective candidates. Many diseases appear to be human-specific, including certain types of cancers and many infectious diseases, suggesting human-specific biological mechanisms associated with these pathologies. In rare cases, genetic mutations identified in individuals with hereditary or congenital disorders offer insight into human phenotypes. By taking advantage of these “clues” it is possible to predict which of the many changes found in highly conserved genetic regions may contribute to human-specific biology and disease. Many of these types of clues were involved in the famous example of *FOXP2*, a transcription factor involved with a heritable speech disorder, and later found to contain human specific evolutionary changes that appear to be involved with our unusual linguistic abilities.

In this paper we presented a uniquely human coding change in the polysialyltransferase

ST8Sia2. This single amino acid substitution carries many of the hallmarks of a potential evolutionarily adaptive change. As we have presented in this paper, this single amino acid is the only difference between humans and chimpanzees, and additionally this residue is completely conserved throughout primates, as well as many other vertebrate lineages.

While the primary bottleneck in investigating human-specific genetics is the “list-to-lab” phase, we report functional consequences of this change. We identified a marked difference in binding affinity of BDNF between polySia-NCAM molecules synthesized by the human and chimpanzee form of the enzyme (unpublished). This change in BDNF affinity may also play a critical role in the aging process, as a reduction in BDNF signaling has been identified as a mediator of cognitive decline and neurodegeneration (Hayashi et al., 2001).

Published results from our lab describing the rapid turnover of polySia by exovesicular sialidase (Sumida et al., 2015), as well as the intrinsic instability of polySia (Manzi et al., 1994), emphasize the importance of tight control over polySia presentation. Polysialic acid functions through diverse mechanisms to alter neural plasticity e.g. modulating cell-cell contacts, altering signaling via interactions with factors including ProBDNF, BDNF, FGF2, and dopamine (Schnaar et al., 2014, Hane et al., 2015). This process is critical during neurodevelopment and throughout life for neurogenesis, synaptic plasticity, and peripheral nerve regeneration. Mutations in the polysialyltransferases and their promoter regions have been identified as risk factors for schizophrenia, autism, and bipolar disorder. Additionally, dysregulation of polysialic acid is associated with neurodegenerative disease (Murray et al., 2016).

4.5 Acknowledgements

This work was supported by R01GM32373 (to A.V.), Aviceda Therapeutics and NIH K12HL141956 (to D.C.), and Training Grant DK007202 in Gastroenterology and UCSD Genetics Training Program T32 GM008666 (to M.V.).

Chapter 4, in full, is currently being prepared for submission for publication of the material. Vaill, M., Hane, M, Naito-Matsui, Y., Davies, L., Kitajima, K., Sato, C., Varki, A., & Chen, D. (2021). A Uniquely Human Evolutionary Change in the Polysialyltransferase ST8Sia2. The dissertation/thesis author is the first investigator and author of this paper.

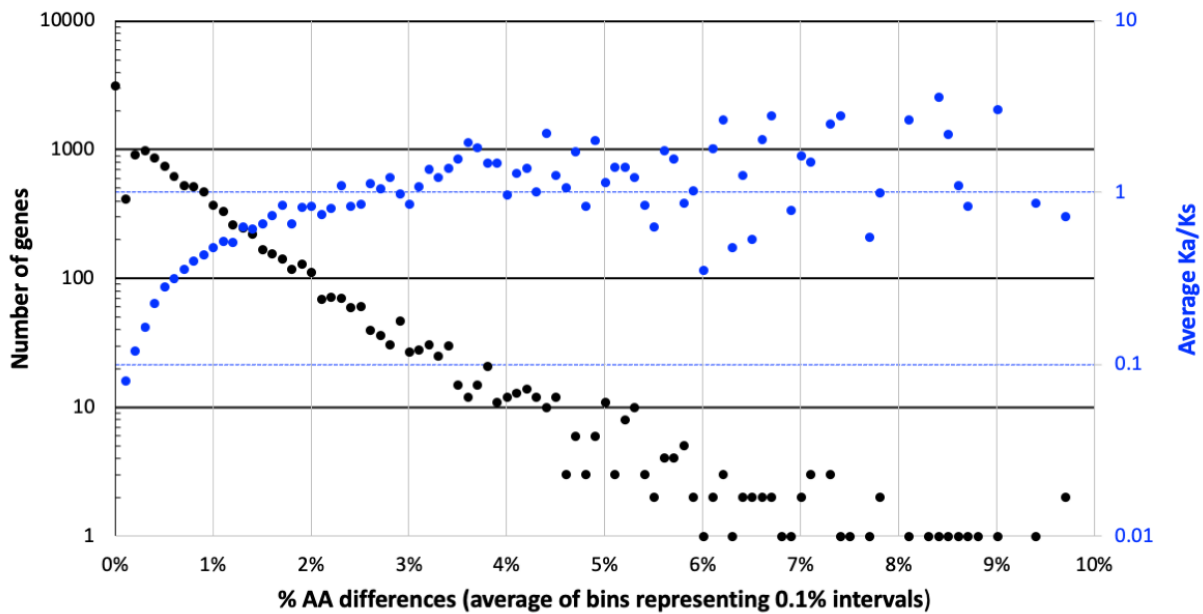


Figure 4.1: Genome-wide comparison of human and chimpanzee orthologs. Groups of orthologous pairs between human and chimpanzee are plotted on the x-axis, binned by the nearest 0.1% amino acid similarity. For each group, the total number of genes is plotted in black, on the left-hand y-axis (logarithmic), and the average Ka/Ks is plotted in blue, on the right-hand y-axis (logarithmic).


```

human      MQLQFRSWMLAALTLVFLIFADISEIEEEEIGNSGGRGTIRSAVNSLHKSNSRAEVVINGSSSPAVVDRSNESIKHNIQPASSK
chimpanzee MQLQFRSWMLAALTLVFLIFADISEIEEEEIGNSGGRGTIRSAVNSLHKSNSRAEVVINGSSSPAVVDRSNESIKHNIQPASSK
bonobo     MQLQFRSWMLAALTLVFLIFADISEIEEEEIGNSGGRGTIRSAVNSLHKSNSRAEVVINGSSSPAVVDRSNESIKHNIQPASSK
gorilla    MQLQFRSWMLAALTLVFLIFADISEIEEEEIGNSGGRGTIRSAVNSLHKSNSRAEVVINGSSSPAVVDRSNESIKHNIQPASSK
*****

human      WRHNQTLSLRIRKQILKFLDAEKDISVLKGTLPKPGDI IHYIFDRDSTMNVSQNL YELLPRTSPLKNKHFGTCAIVGNSGVLLNSG
chimpanzee WRHNQTLSLRIRKQILKFLDAEKDISVLKGTLPKPGDI IHYIFDRDSTMNVSQNL YELLPRTSPLKNKHFGTCAIVGNSGVLLNSG
bonobo     WRHNQTLSLRIRKQILKFLDAEKDISVLKGTLPKPGDI IHYIFDRDSTMNVSQNL YELLPRTSPLKNKHFGTCAIVGNSGVLLNSG
gorilla    WRHNQTLSLRIRKQILKFLDAEKDISVLKGTLPKPGDI IHYIFDRDSTMNVSQNL YELLPRTSPLKNKHFGTCAIVGNSGVLLNSG
*****

human      CGQEIDAHSFVIRCNLAPVQEYARDVGLKTDLVTMNP SVIQRAFEDLVNATWREKLLQRLHSLNGSILWIPAFMARGGKERVEWV
chimpanzee CGQEIDAHSFVIRCNLAPVQEYARDVGLKTDLVTMNP SVIQRAFEDLVNATWREKLLQRLHSLNGSILWIPAFMARGGKERVEWV
bonobo     CGQEIDAHSFVIRCNLAPVQEYARDVGLKTDLVTMNP SVIQRAFEDLVNATWREKLLQRLHSLNGSILWIPAFMARGGKERVEWV
gorilla    CGQEIDAHSFVIRCNLAPVQEYARDVGLKTDLVTMNP SVIQRAFEDLVNATWREKLLQRLHSLNGSILWIPAFMARGGKERVEWV
*****

human      NELILKHHVNVRTAYPSLRLHAVRGYWLTKNVHIKRPTTGLLMYVRTAYPSLRLHAVRGYWLTKNVHIKRPTTGLLMYTLATR
chimpanzee NELILKHHVNVRTAYPSLRLHAVRGYWLTKNVHIKRPTTGLLMYVRTAYPSLRLHAVRGYWLTKNVHIKRPTTGLLMYTLATR
bonobo     NELILKHHVNVRTAYPSLRLHAVRGYWLTKNVHIKRPTTGLLMYVRTAYPSLRLHAVRGYWLTKNVHIKRPTTGLLMYTLATR
gorilla    NELILKHHVNVRTAYPSLRLHAVRGYWLTKNVHIKRPTTGLLMYVRTAYPSLRLHAVRGYWLTKNVHIKRPTTGLLMYTLATR
*****

human      FCNKIYLYGFWPFPLDQNPVKYHYD SLKYGYTSQASPH TMPLEFKALKSLHEQGALKLTVGQCDGAT
chimpanzee FCNQIYLYGFWPFPLDQNPVKYHYD SLKYGYTSQASPH TMPLEFKALKSLHEQGALKLTVGQCDGAT
bonobo     FCNQIYLYGFWPFPLDQNPVKYHYD SLKYGYTSQASPH TMPLEFKALKSLHEQGALKLTVGQCDGAT
gorilla    FCNQIYLYGFWPFPLDQNPVKYHYD SLKYGYTSQASPH TMPLEFKALKSLHEQGALKLTVGQCDGAT
** : *****

```

Figure 4.2: Human and chimpanzee *ST8SIA2* share only 1 amino acid difference at N308K. Alignment of human *ST8SIA2* protein sequence with chimpanzee, bonobo, and gorilla. The single, human-specific, amino acid change N308K is highlighted in red.

Species	Sequence	Divergence (Mya)
Human	PTTGLLMYTLATRFCKQIYLYGF	-
Neandertal	.-----.....	0.6
Denisovan----	0.6
ChimpanzeeN.....	6.6
GorillaN.....	8.3
OrangutanNE.....	15.8
GrivetN.....	27.3
MouseN.....	90.1
RatN.....	90.1
ChickenNR.H....	320.5
XenopusI.....NR.....	355.7
ZebrafishM.....DE.H....	436.8

Figure 4.3: N308 is a highly conserved residue throughout vertebrates. Alignment of the human-specific amino acid change and surrounding motif with archaic hominin genomes, and representative vertebrate phylogeny.

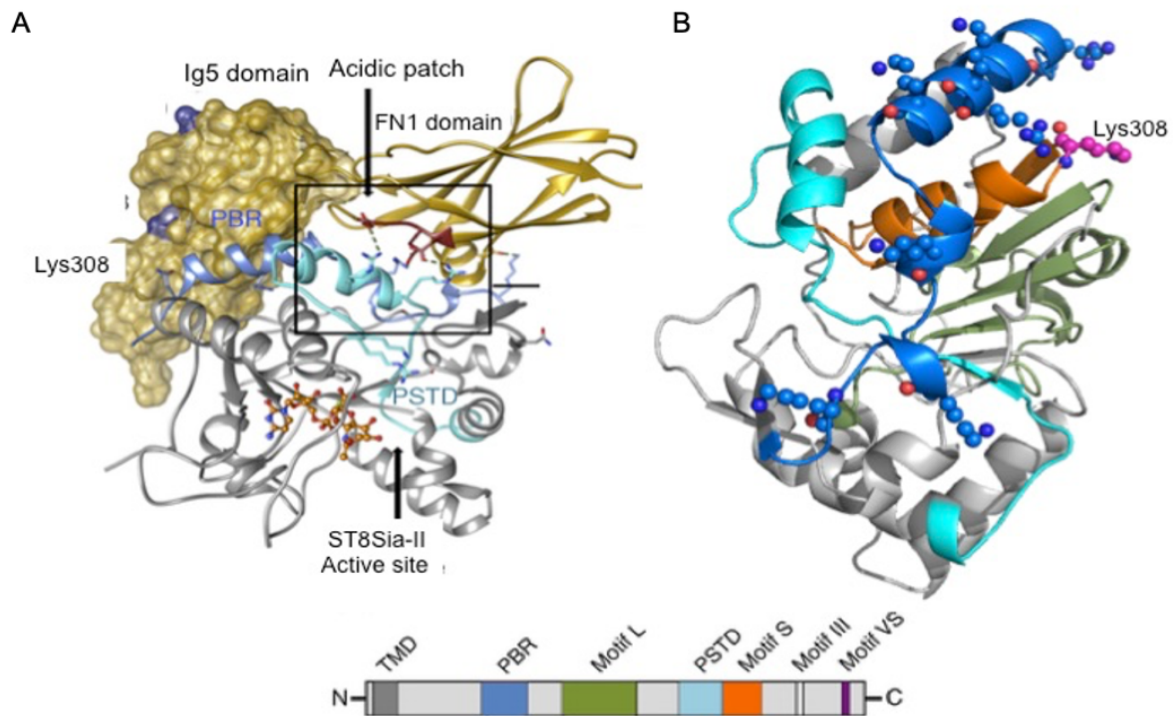


Figure 4.4: Modeling of human ST8Sia2 based on the crystal structure of human ST8Sia3. A) Superimposed structure of NCAM Ig5 and FN1 domains on show orientation of helix α C terminal Lys308 against Ig5. (B) PBR basic residues (blue, ball representation) are located in proximity to uniquely human Lys308 (Pink, ball representation).

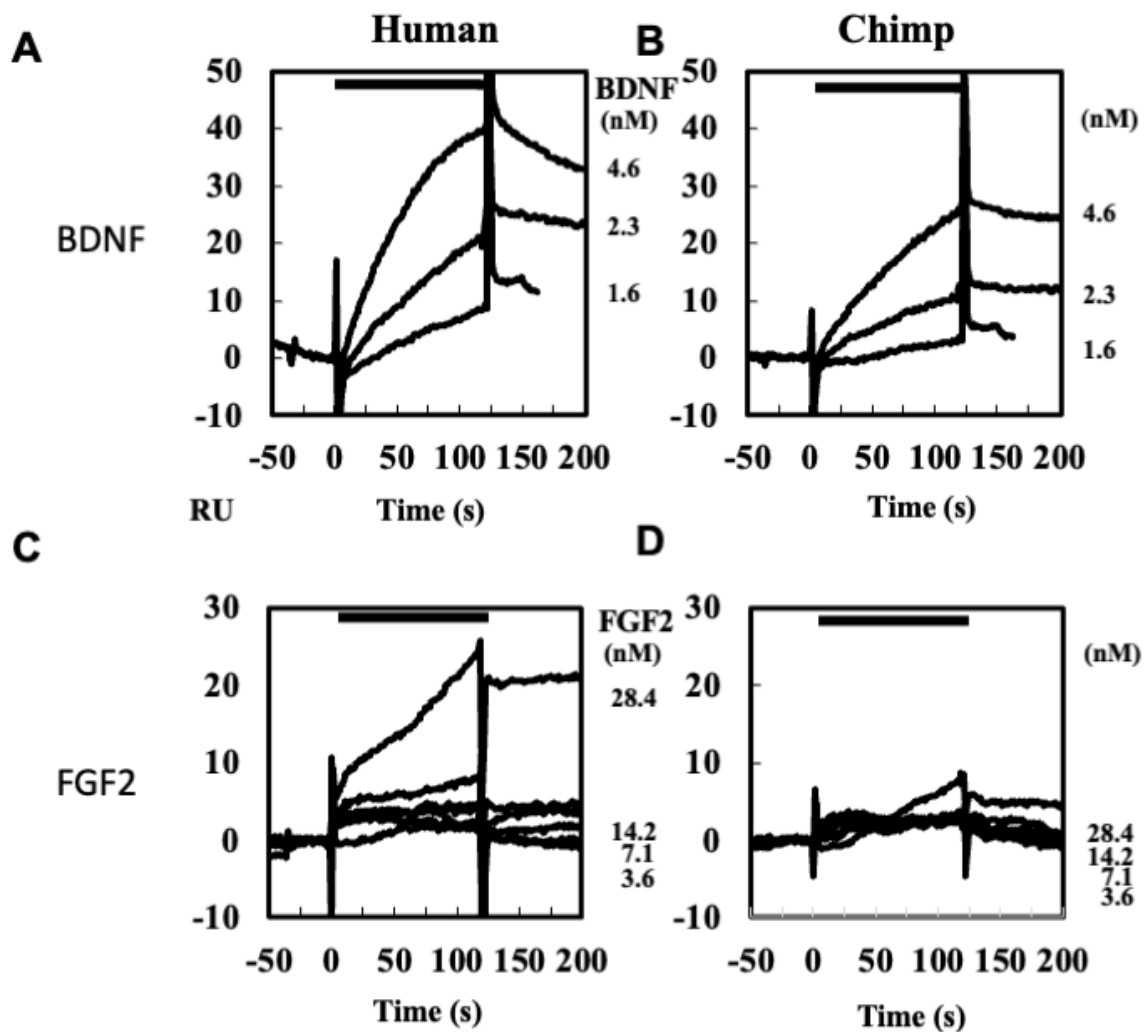


Figure 4.5: Surface plasmon resonance binding analysis of BDNF and FGF2 binding by polySia-NCAM-Fc polysialylated by either human or chimpanzee ST8Sia2. BIACORE sensorgrams showing RU (y-axis) vs time (x-axis). A) BDNF binding by immobilized human-type polySia-NCAM-Fc. B) BDNF binding by chimpanzee-type polySia-NCAM-Fc. C) FGF2 binding by immobilized human-type polySia-NCAM-Fc. D) FGF2 binding by chimpanzee-type polySia-NCAM-Fc.

Chapter 5

Novel Methods to Characterize the Length and Quantity of Highly Unstable PolySialic Acids

Polysialic acid (polySia) is a linear homopolymer of α 2-8-linked sialic acids that is highly expressed during early stages of mammalian brain development and modulates a multitude of cellular functions. While degree of polymerization (DP) can affect such functions, currently available methods do not accurately characterize this parameter, because of instability of the polymer. We have developed two novel methods to characterize the DP and total polySia content in biological samples. PolySia chains with exposed reducing termini can be derivatized with DMB for subsequent HPLC analysis. However, application to biological samples of polySia-glycoproteins requires release of polySia chains from the underlying glycan, which is difficult to achieve without concurrent partial hydrolysis of the α 2-8-linkages of the polySia chain, affecting

its accurate characterization. We report an approach to protect internal α 2-8sia linkages of long polySia chains, using previously known esterification conditions that generate stable polylactone structures. Such polylactonized molecules are more stable during acid hydrolysis release and acidic DMB derivatization. Additionally, we used the highly specific Endoneuraminidase-NF enzyme to discriminate polysialic acid and other sialic acid and developed an approach to precisely measure the total content of polySia in a biological sample. These two methods provide improved quantification and characterization of polySia.

5.1 Introduction

Polysialic acid (polySia) is a linear homopolymer of sialic acids (Sias) in α 2-8 linkages. PolySia was first discovered in the polysaccharide capsule of *E. Coli* K235L+O in 1957 (BARRY and GOEBEL, 1957). Some four years after discovering sialic acids in bacteria, Guy Barry called the capsular polysaccharide of this strain of *E. Coli* “colominic acid,” and reported that the structure of this molecule consists of a polymer of repeating residues of N-acetylneuraminic acid (Neu5Ac). Barry suggested that, given the ongoing explosion of discoveries related to sialic acids in the 1950s, such a polymeric structure of sialic acids may have some relevance in mammals (Lundblad, 2015). In fact, this prediction proved to be correct (Troy et al., 1982, Whitfield and Troy, 1984, Whitfield et al., 1984, McCoy et al., 1985, Edwards et al., 1994), and a common antigen found in vertebrate brains would later be identified as a similar polymer of sialic acids extended upon the N-glycans of the neural cell adhesion molecule (NCAM) (Cunningham et al., 1983, Finne et al., 1983, Eckhardt et al., 1995). Notably, *E. Coli* K1 expressing PolySia is a common cause of neonatal and pediatric meningitis (McCracken et al., 1974, Robbins et al., 1974, Glode et al., 1977), and can also cause infections in adults (O’Hanley et al., 1985,

Anderson et al., 2010).

In vertebrates, polySia is highly expressed during embryonic brain development and, in mice, reaches peak expression levels perinatally (Ong et al., 1998). During postnatal mouse brain development, the amounts of polySia remain high during the first week, before rapidly declining between postnatal days 9–17. This decrease continues into adulthood (Seki and Arai, 1991, Rousselot and Nottebohm, 1995). In both adult rodents and humans, polySia is selectively expressed in areas where neurogenesis persists (Rousselot et al., 1995, Bernier et al., 2000, Sanai et al., 2011). Functionally, polySia forms a highly hydrated structure on the cell surface with a steric effect (“repulsive field”) that directly affects cell-cell contacts and interactions of cell surface receptors and matrix components, while also sequestering and concentrating soluble ligands like FGF2 and BDNF (“attractive field”), enabling critical modulatory roles in many neural processes (Colley et al., 2014). PolySia-N-glycan structures depend on expression and activity of two enzymes ST8Sia2 and ST8Sia4, which act either independently or cooperatively (Close et al., 2001, Galuska et al., 2006, Galuska et al., 2008, Thompson et al., 2013).

Accurately determining DP of polySia in biological samples has remained a technical challenge. HPLC-based analysis has been used in the characterization of polySia to determine the structural characteristics, primarily the degree of polymerization (DP) (Sato et al., 1998, Galuska et al., 2008). Another approach to interpret the chain length of polySia include analysis of the relative abundance of the terminal sialic acids compared with the internal sialic acids. This can be accomplished either by comparing the relative reactivity of antibodies that recognize the polySia chain (mAb 12E3) to antibodies that bind along the polymer in an abundance relative to the length (mAb 735) (Sato et al., 1995). Alternatively, a chemical modification can be used to demarcate the non-reducing terminal sialic acid by mild periodate oxidation which cleaves a

2-carbon glycol from the side-chain of exposed sialic acids leaving a 7-carbon sugar in the non-reducing terminal position, while internal residues remain protected from reduction in a 9-carbon form (Sato et al., 1998). The C7 and C9 sugars can then be released to monomers and the relative abundance quantitated by C18-HPLC. However, even a small amount of contaminating monosialyl residues can confound the accuracy of this method.

Fluorometric analysis of sialic acids is a useful tool for quantitatively studying relative abundance of sialic acid forms, enabling highly sensitive detection of samples derived from biological samples which typically contain sialic acids on the order of picomoles-femtomoles. 1,2-diamino-4,5-methylenedioxybenzene (DMB) selectively labels alpha-keto acids like sialic acids, producing a covalent derivative that can be separated by HPLC with a quantitative fluorescence readout. Derivatization of sialic acid was first accomplished by using acidic conditions to protonate the sialic acid enabling a reaction aldehyde side chain of the fluor DMB via Schiff-base mechanism at 50°C. The acid used was 0.7M HCl (Nakamura and Sweeley, 1987), however, because sialic acid and its derivatives are labile this was later optimized to use milder conditions using 1M acetic acid (AcOH) (Sato et al., 1998). Later application of DMB derivatization to polysialic acid was accomplished by dropping the temperature of the reaction to 4°C or 10°C and using a smaller quantity of the stronger trifluoroacetic acid (TFA) at 20 mM (Sato et al., 1998).

Applications to biological samples of sialylated-glycoproteins first requires release of sialic acid from the underlying glycan structure, which in mammals is an α 2-3 or α 2-6 linkage to galactose. This release can be efficiently accomplished enzymatically by a sialidase, or chemically by a simple incubation in a mild or strong acid at high temperature to accomplish complete release (Varki and Diaz, 1984). But this step poses a significant challenge to

the analysis of polySia chains in biological samples: exosialidases can no longer work on this internal bond, and the instability of the internal α 2-8-glycosidic bonds of polySia can undergo spontaneous hydrolysis via intramolecular cleavage in a temperature and pH dependent manner (Manzi et al., 1994), resulting in degradation of polySia prior to analysis, thus affecting its detectable DP (Guo et al., 2019). Standard analysis using acid hydrolysis release and DMB derivatization under acidic conditions typically yields polySia fragments of DP \sim 30-50 from biological samples. However, DP 400 in embryonic brain has been reported following release using endo- β -galactosidase, an enzyme that selectively releases the rare polySia chains linked via a poly-N-acetyllactosamine motif (Nakata and Troy, 2005). The release of all polySia chains, rather than an enzymatically selected subset, from the underlying glycoprotein remains a challenge for the field. ELISA-capture has also been used as a method to identify polysialic carriers in serum, where a number of non-NCAM polysialylated proteins exist (Tajik et al., 2020). A recent comprehensive review of methods for polySia analysis summarized these difficulties, concluding that “a methodology for truly accurate determination of polySia chain length remains elusive” (Guo et al., 2019).

Classic studies in polymer chemistry showed that the resonant properties of poly-acids determined that the pK of carboxyl groups increases proportionally with the DP of the molecule (Katchalsky and Spitnik, 1947). This has consequences on the labile α 2-8 glycosidic linkages of polysialic acid, as they experience a self-catalyzed intramolecular cleavage (Manzi et al., 1994), dependent upon the protonation and thus the pK_a of the neighboring carboxylate groups. As the polymer is extended and carboxylate groups become protonated, the internal most bonds will become increasingly susceptible to this catalyzed hydrolysis (Manzi et al., 1994). Because of this protonation, when the very long polySia chains found in mammalian tissues are subjected

to standard sialic acid DMB derivatization conditions of pH 3-5, internal linkages experience catalyzed hydrolysis even at 10°C, or 4°C.

Previous methods involving DMB analysis of polysialylated glycoconjugates took advantage of the low-rate of unpredictable hydrolysis by directly subjecting polysialylated structures to DMB derivatization conditions. Random hydrolysis of internal bonds produces fragments of DP 20-100 with reducing sialic acid residues exposed for the DMB labeling chemistry to proceed (Inoue et al., 2001). However, this unpredictable rate of hydrolysis also presents challenges to accurately determine the DP. We achieve lactonization-protection of the α 2-8 glycosidic linkages using ice cold-acidic conditions. Under these conditions, the C9 hydroxyl group of each sialic acid undergoes esterification with the carboxylate of the adjacent sialic acid to form a stable 6-carbon lactone structure (Fig 1A) (Cheng et al., 1998, Zhang and Lee, 1999, Kakehi et al., 2001). Due to the spatial arrangement between galactose and a 2-3 linked sialic acid a similar lactone ring can be formed by using a highly reactive catalyst (Liu et al., 2008), however, it is unlikely that such a reaction could proceed under milder or biologically relevant conditions. An early study of the lactonization properties of polysialic acid suggested that lactonization may offer protection during acid hydrolysis (Zhang and Lee, 1999). This study showed protection of the highly abundant oligo/poly-Neu5Gc (DP_{leq}11) structures from salmon eggs prior to PAD (pulse amperometric detection). Lactonization was also used to enable MALDI-TOF-Mass Spectrometry of polysialic acid chains (Galuska et al., 2007). Here, we apply the lactonization-protection in mammalian glycoproteins containing large poly-Neu5Ac structures, prior to DMB labeling, and fluorescence detection and took advantage of the polysialic acid specificity of Endo-N enzyme for total polysialic acid quantification.

5.2 Results and Discussion

Overview of polySia lactonization.

To more accurately determine the DP in biological examples, we developed an approach to protect internal α 2-8sia linkages of long polySia chains during acid hydrolysis and subsequent DMB derivatization. We achieve this by first inducing the formation of stable lactone-ring structures (Fig 1A, Fig 1B – step 1) along the length of the polySia, prior to proceeding with acid hydrolysis (Fig 1B – step 2) and DMB labeling (Fig 1B – step 3). This lactonized molecule remains stable under acid hydrolysis release and during acidic DMB derivatization. Subsequent addition of base reverses lactonization (Fig 1B – step 4) before separation of the labeled polySias with anion exchange HPLC (Fig 1B – step 5) and analysis using fluorescence detection of DMB labels. Free polymers of sialic acid, such as bacterial colominic acid which is collected from *E. Coli* culture and is not covalently linked at the reducing end, can be readily labeled by DMB. However, accomplishing release from the underlying glycan structure (Fig 1B – step 2) has remained an obstacle for the analysis of mammalian polysialylated glycan structures. This lactonization-DMB allows for improved characterization of the degrees of polymerization, but it does not provide quantitation of polySia content. To quantitate polySia, we take advantage of the highly specific endosialidase Endoneuraminidase-NF (EndoN), to quantitate the total amount of polySia in a sample. EndoN is an endosialidase from the bacteriophage that evolved to specifically targets the polySia capsular polysaccharide of *E. Coli* K1. EndoN cleaves α 2-8Sia linkages within a chain of sias with degree of polymerization of 5 or greater (Hallenbeck et al., 1987, J Biol Chem, 262, 3553-3561). We applied the specificity of the EndoN enzyme as a tool to selectively release all polySia from brain samples that we also analyzed by lactonization-DMB

to derive a parallel absolute quantitation of polySia content.

Lactonization protects colominic acid from acid hydrolysis or intramolecular self-cleavage.

To investigate whether other factors such as the DMB derivatization temperature and/or the type of acid used for derivatization contribute to the ability to detect higher DP, samples of colominic acid were subjected to DMB derivatization at 50°C (Fig 2A), or 4°C (Fig 2B). Compared to DMB derivatization at 4°C, 50°C derivatization induces the hydrolysis of long polymers. We next tested whether lactonization effectively protects α 2-8 sialic acid linkages from acid hydrolysis conditions. Samples that were not subjected to the overnight lactonization step before hydrolysis were primarily detected by HPLC as peaks of Sia oligomers of DP 1-15 (Fig 2C), with the largest detectable peaks around DP 25. Lactonization produced a noticeable increase in the higher peaks and a decrease in these smaller peaks (Fig 2D). Thus, lactonization protects higher DP colominic acid during hydrolysis conditions. Furthermore, these results confirm that these oligomers are the result of the degradation of large polySia chains, which after the lactonization step were protected through acid hydrolysis and still detectable up to DP 40.

Lactonization protects mouse brain polySia structures during release by acid hydrolysis.

Our studies with colominic acid illustrate that lactonization protects highly unstable α 2-8 sialic acid linkages during acid hydrolysis. Next, we explored whether lactonization protection of α 2-8 sialic acid linkages can be used during acid hydrolysis release of polySia chains from biological polysialylated glycoproteins. PolySia is highly expressed by proliferating neural cells

during early brain development, and is found throughout neonatal mouse brains. We prepared a whole brain homogenate from postnatal day 1 (P1) mouse pups using pH 8.0 Tris-HCl buffer. We then used organic extraction to remove the abundant lipids from the whole brain homogenate, and treated the protein fraction with proteinase K to produce an aqueous P1 brain glycopeptide sample rich in polysialylated N-glycan structures (Fig 3A). We used this P1 brain glycopeptide sample to investigate the use of lactonization in the study of mammalian brains glycans.

In samples which were subjected to an initial step of lactonization prior to hydrolysis, we detected eluting polySia structures up to DP 60 in P1 mouse brains (Fig 3B). Without an initial lactonization step these large peaks were completely destroyed prior to analysis (Fig 3C). Additionally, acetic acid hydrolysis of P1 brain glycopeptides generates peaks of mono and oligo sialic acids eluting between 15 and 35 minutes. These peaks are significantly reduced in area by using lactonization. This suggests that, like colominic acid, the abundance of mono and oligo sialic acids in the non-lactonized sample are the product of hydrolysis of unprotected polySia structures, and that lactonization prior to hydrolysis preserves these structures through release and DMB labeling. This allows for higher resolution characterization of the collection of polySia structures in the brain sample. While there is likely possibly some degradation that occurs, this critical new approach permits a more accurate glimpse of the endogenous state of brain polysialylation.

Analysis of dynamic polysialic acid synthesis during early postnatal mouse brain development

Using the lactonization method, we next set out to study a hallmark paradigm of polySia expression and function in the mammalian nervous system. As previously described, polySia

expression is highest in the nervous system during embryonic and early postnatal development. As the brain matures, proliferating polysia⁺ neural precursors give rise to the terminally differentiated neural cell types of the adult brain, and eventually polysia⁺ cells are only detected in the specific brain areas of adult neurogenesis, including the dentate gyrus of the hippocampus and the subventricular zone of the lateral ventricle.

Brains were collected from mice representing infant, juvenile, and adult stages of neurodevelopment (7 days, 14 days, and 10 weeks of age, respectively). These ages were chosen to represent progressive stages of neural development and to apply our new methods to further the understanding of polySia developmental dynamics. Infant mice showed striking peaks of polySia between DP of 20 and 40, with highest detected peaks even higher (Fig 4A). Juvenile mice showed similar amounts of polySia up to DP20, but less polymers in the highest size range (Fig 4B). Adult mice had a smaller, but appreciable amount of the largest polymers (Fig 4C). Importantly, because lactonization reduces the degradation of large polymers into smaller fragments, this method improves the ability to observe the relative dynamics of polySia chains. Without lactonization, indiscriminate hydrolysis of long polySia chains during sample preparation produces a misleading increase in smaller chains.

These results not only confirm that our method is capable of capturing the dynamic regulation of polySia in mammals, but also offers some insight into the biology of polySia during mouse brain development. It appears that the abundance of large polymers decreases dramatically with age, disproportionately to the smaller and medium sized polySia chains. This may be related to the developmental regulation of the two polysialyltransferases ST8Sia2 and ST8Sia4, which are known to have slightly differing and cooperative activity *in vivo* (Galuska et al., 2006). Previous studies into polySia function have identified a diverse set of molecu-

lar mechanisms through which polysialylated glycans play a role in cellular activity. To name a few, these mechanisms include: physically affecting cell-cell contact, directly interacting with signaling molecules in cis or trans, and interacting with soluble ligands as either a co-receptor to increase ligand-receptor activity, or a molecular sink to trap ligands and reduce receptor activity (Schnaar et al., 2014). The difference in abundance of larger polySia changes in older mice may be related to medium-shorter chains having a more critical function in adult polySia⁺ cells, such as in hippocampal plasticity, relative to the long-chains roles in developing polySia⁺ cells.

The degree of polymerization (DP) is known to affect the biological functions of polysialic acid. For example, brain derived neurotrophic factor (BDNF), a neural growth factor bind to polysialic acid in a DP-dependent manner (Kanato et al., 2009), suggesting polysialic acid DP can affect the myriad of BDNF-dependent functions, including cell survival, differentiation (Huang and Reichardt, 2003), synaptic plasticity, long-term memory formation (Bramham and Messaoudi, 2005) as well being associated with neuropsychiatric disorders, such as schizophrenia, major depression and bipolar disorder (Arai et al., 2006; Cox et al., 2009; Barker et al., 2012; Shaw et al., 2014; Miranda et al., 2019). Additionally, several genetic variations of *ST8SIA2* have also been linked to such psychiatric disorders, including autism spectrum disorder. In fact, a familial schizophrenia-related *ST8Sia2* mutation appears to affect enzyme function by producing shorter polySia chains, highlighting the consequences of dysregulated polySia chain-length in humans. It is possible that, depending on context, cells depend on an abundance of chains of a specific DP for preferential between the many potential functions of polySia. Differential regulation of polySia DP may be involved in the context-specific molecular mechanisms at play for a given polysialylated protein or cell surface. With advances such as the method described this paper, these are all areas that may be approached in the future.

Complementary enzymatic method to quantitate polySia content

EndoN is a highly specific enzyme from a phage that hydrolyzes α 2-8 linkages in chains of polySia of DP \geq 7 (Rutishauser et al., 1985; Nadano et al., 1986; Hallenbeck et al., 1987). We took advantage of this specificity to quantify total polysialic acid content. We used DMB labeling following EndoN digestion to specifically identify the EndoN sensitive sialic acids. EndoN produces fragments of between 2-5 sialic acids (Pelkonen et al., 1989; Hallenbeck et al., 1987; Gross et al., 1977; Finne and Mäkelä; Galuska et al., 2006). After EndoN treatment we needed to separate the oligomeric products from the glycopeptide sample containing other, non-EndoN-sensitive, sialic acid determinants. To accomplish this, we used a centrifugal filter device with a cutoff of 3000 Da, or 9 Neu5Ac units, to collect EndoN-sensitive sialic acids before HPLC quantitation.

By digesting a sample with EndoN, labeling the filtered product with DMB, and comparing the oligomers with a control sample in which EndoN enzyme was omitted, the peaks corresponding with EndoN products can be quantified and summed together to produce a figure representing the amount of polysialic acid in the sample.

To illustrate this principle, we once again used colominic acid as a standard. We treated colominic acid with EndoN, collected the products of this reaction via filtration, and subsequently labeled with DMB. HPLC analysis of EndoN-sensitive Neu5Ac released from colominic reveals that all polymers greater than 7 sialic acid residues are enzymatically hydrolyzed into oligomers of DPs 2-5 (Fig 5A). Omitting EndoN treatment before the filtration step revealed that there were no detectable sialic acid structures less than DP 9 in our colominic acid standard (Fig 5B). Besides for the intended purpose as a control for enzyme activity, and a background sample for quantitation, this result suggests that in our previous studies, applying various DMB derivatiza-

tion methods to the same colominic acid sample (Fig 2A-E), the presence of sialic acid oligomers is purely an experimental artifact, likely secondary to hydrolysis. One of the major improvements of lactonization is a reduction in these artifacts.

Because the oligomers are separated into discrete integrable peaks, we are able to quantitate the amount of colominic acid subject to enzymatic degradation. Using colominic acid as a quantitative standard for EndoN-sensitive Neu5Ac, we determined the polySia content of the infant, juvenile, and adult brain samples previously used for lactonization length analysis. Our results confirm the known trend that polySia content decreases as the brain matures. While other methods do exist to quantitate total amount of polySia, this is a simple and efficient method, and the first quantitative method that takes advantage of the highly specific enzymatic activity of EndoN.

5.3 Materials and Methods

Lactonization and mild acid hydrolysis of colominic acid

10 nmol (as Neu5Ac) of Colominic acid in 50 mM Tris-HCl pH 8.0 was lactonized by incubation overnight in an ice water bath after the addition of ice-cold HCl to final concentration of 1M induce lactonization, or water as a non-lactonized control. After overnight lactonization, samples were frozen at -80°C and lyophilized to remove HCl. Dried samples of colominic acid were then resuspended in 50 μ l 2M AcOH and subjected to a 90 minute incubate at 80°C, to mimic the conditions known to release terminal Sias from biological glycoconjugates.

HPLC Methods

After derivatization, samples were immediately prepared for HPLC analysis. Samples were centrifuged at maximum speed for 20 minutes to prevent injecting any non-soluble particulate into the HPLC system. After centrifugation, samples were transferred into auto-sampler vials for injection. Volumes between 20-80 μ l were injected, depending on concentration of samples.

HPLC chromatography was run on a Hitachi LaChrom Elite HPLC system fitted with a CARBOPAC PA200 Analytical Column (3 x 250 mm, Thermo scientific P/N 062896), and CARBOPAC PA200 Guard Column (3 x 50 mm, Thermo scientific P/N 062895). Fluorescence was measured at 372 nm excitation and 456 nm emission (Jasco detector). Anion exchange was accomplished using a flow of 0.5 ml/min and the following gradient of Milli-Q water (E1) and 1 M sodium nitrate (E2): 0 min = 0% E2, 2 min = 2% E2, 9 min = 10% E2, 39 min = 16% E2, 99 min = 31% E2, 100 min = 2% E2, 110 min = 2% E2.

Chromatograms were produced and analyzed with OpenLab Chromatography Data System (CDS) EZChrom Edition (Agilent Technologies).

Brain tissue homogenization and delipidation

Mice were euthanized by isoflurane inhalation followed by decapitation, and the whole brain was collected immediately after dissection. Brain mass was recorded, and tissue was placed on ice in a glass centrifuge tube with ice cold water at a 4:1 w/w ratio of water:brain tissue. Keeping tube on ice, tissue was thoroughly homogenized, for about 1 minute, on a polytron at high speed. 1:1 chloroform:methanol mixture was added at 20X volume of water/brain homogenate and mixed vigorously by polytron, followed by centrifugation at 1200 x g for 15

minutes at 4°C. Supernatant containing lipids and low molecular weight molecules was carefully removed and discarded, leaving a pellet containing brain glycoproteins and other macromolecules. To wash pellet, fresh 10:10:1 chloroform:methanol:water mixture was added at the previous volume before polytron mixing and centrifugation as before.

After final wash, samples were left on ice. After excess organic solvent was completely evaporated 50mM Tris-HCl pH 8.0 was added at 2x original brain mass. Pellet was suspended completely with repeated pipetting.

Proteinase digestion

Proteinase K (Invitrogen P/N 25530031) was used to digest the protein component of brain glycoprotein samples, producing glyopeptides ready for analysis. Glycoproteins suspended in 50 mM Tris-HCl pH 8.0 were treated with 200 ng/ μ l proteinase K (0.01 volumes 20 mg/ml stock enzyme). Samples were then incubated overnight at 37°C with rapid mixing. 1 mM phenylmethylsulfonyl flouride (PMSF) was then added to inhibit proteinase (0.01 volumes 100 mM stock PMSF in isopropanol). Samples were then washed using a 3K-cutoff centrifugal filter unit (Sigma-Millipore UFC9003), with 3 serial dilution-concentration steps: dilution in 15 ml 50 mM Tris-HCl pH 8.0 followed by concentration.

Lactonization and mild acid hydrolysis of mouse brain glycopeptides

100 μ l of brain glycopeptides (derived from ~50 mg brain tissue) was placed in an ice water bath, and 100 μ l of ice cold 2M HCl (or H₂O for the non-lactonized control) was quickly added and mixed (final concentration 1M HCl). The samples were incubated overnight in ice water to induce lactonization. The samples were then centrifuged at maximum speed (17000g)

at 4°C for 30 mins, frozen at -80°C, and lyophilized until dry. The dried lactonized glycopeptides were then suspended in 100 μ l 2M AcOH and heated at 80°C to hydrolyze any non-lactonized sialic acid linkages, including the internal 2-3 and 2-6 linkages connecting the polysialic acid chains to their N-glycan antennae.

Enzymatic digestion of polysialic acid and quantitation of endoN-sensitive Neu5Ac

50 μ l of brain glycopeptides prepared (originating from 25 mg brain tissue) was treated with 1 μ l endoneuraminidase-NF on ice for 90 minutes (2 mg/ml enzyme stock, enzyme was a kind gift from Rita Gerardy-Schan). Samples were then filtered through 3k centricon filter, and flow-through fractions were collected for DMB labeling. After DMB analysis, oligosialic acid products of endoN were identified and peak areas calculated using OpenLab Chromatography Data System (CDS) EZChrom Edition (Agilent Technologies). For each sample total relative endoN-sensitive Neu5Ac was determined by calculating the sum of product oligomers, with each oligomer peak area multiplied by its respective DP. 1.25 μ g of brain tissue were used for each injection, and 25 pmol of colominic acid was analyzed in parallel and used as a standard.

DMB labeling

DMB labeling was accomplished in a reaction containing 1.35 M DMB, 1 M acetic acid, 9 mM sodium hydrosulfite, and 0.5 M β -mercaptoethanol. Samples containing sialic acid and 2 M acetic acid were placed on ice, and 50 μ l of a 2X mix containing the remaining reaction ingredients (DMB, NaSO₂, and β -ME) was added. This reaction was incubated away from light, at 4°C, with end-over-end mixing or at 50°C for 2 hours. The 4°C samples were labeled for 40

hours. After DMB labeling, 14 μ l 10N NaOH was added, and samples were mixed thoroughly with vortexing. NaOH neutralizes acetic acid, reverses esterification of the polylactone structure, and frees carboxylate groups to restore negative charge for anion exchange separation.

5.4 Conclusions and Perspectives

Here we presented two distinct analytical approaches to characterize both the total amount of polySia in given biological sample containing polysialylated glycan species using EndoN-sensitive Neu5Ac analysis, as well as the relative quantity of polySia chains of a given length within each sample. Although we did not do so, it should also be possible to combine these approaches to specifically calculate the quantitative amount of polySia of a given DP. While our studies are most likely still affected by the intrinsic instability of polySia's α 2-8linkages, we have made progress in more closely revealing the true nature of polySia structures in a biological sample.

In this study, we tested this method of polySa analysis and found that compared to our new lactonization approach, subjecting polysialylated structures to DMB derivatization without lactonization, in both colominic acid and in biological samples, resulted in more hydrolysis. While directly treating samples with DMB reagents may elucidate certain differences between two samples, and qualitatively distinguish them, it is difficult to translate the result into a meaningful representation of polySia structures present before the introduction of random hydrolysis. We applied this new lactonization methods and identified a polySia-developmental profile of mouse brain and found that smaller DP polySia remains abundant as the mouse brain matures, whereas the larger DP polySia decreases. Lastly, we took advantage of the highly specific phage enzyme EndoN, which hydrolyzes α 2-8linkages, and developed a method that allowed

us to quantitate amounts of sialic acid in biological samples. While these methods allowed us improved quantitation of polySia, several limitations remain: 1) Lactonization may not prevent all hydrolysis, 2) it is unclear whether DMB-labeling efficiency is the same for DP chains of different length, and 3) the Endo-N method does not account for sialic acids that remain attached to the underlying glycan, as it is not DMB labeled. Studying such an unstable structure will likely remain a technical challenge, but ongoing improvements will continue to refine methodologies to accurately and precisely characterize and quantitate these important molecules, and yield further insight into the regulation of biological processes via polysialylation.

5.5 Acknowledgements

This work was supported by R01GM32373 (to A.V.), Aviceda Therapeutics and NIH K12HL141956 (to D.C.), and Training Grant DK007202 in Gastroenterology and UCSD Genetics Training Program T32 GM008666 (to M.V.).

Chapter 5, in full, has been submitted for publication of the material as it may appear: Vaill, M., Chen, D., Diaz, S., & Varki, A. (2021) Novel Methods to Characterize the Length and Quantity of Highly Unstable PolySialic Acids. The dissertation/thesis author is the primary investigator and author of this paper.

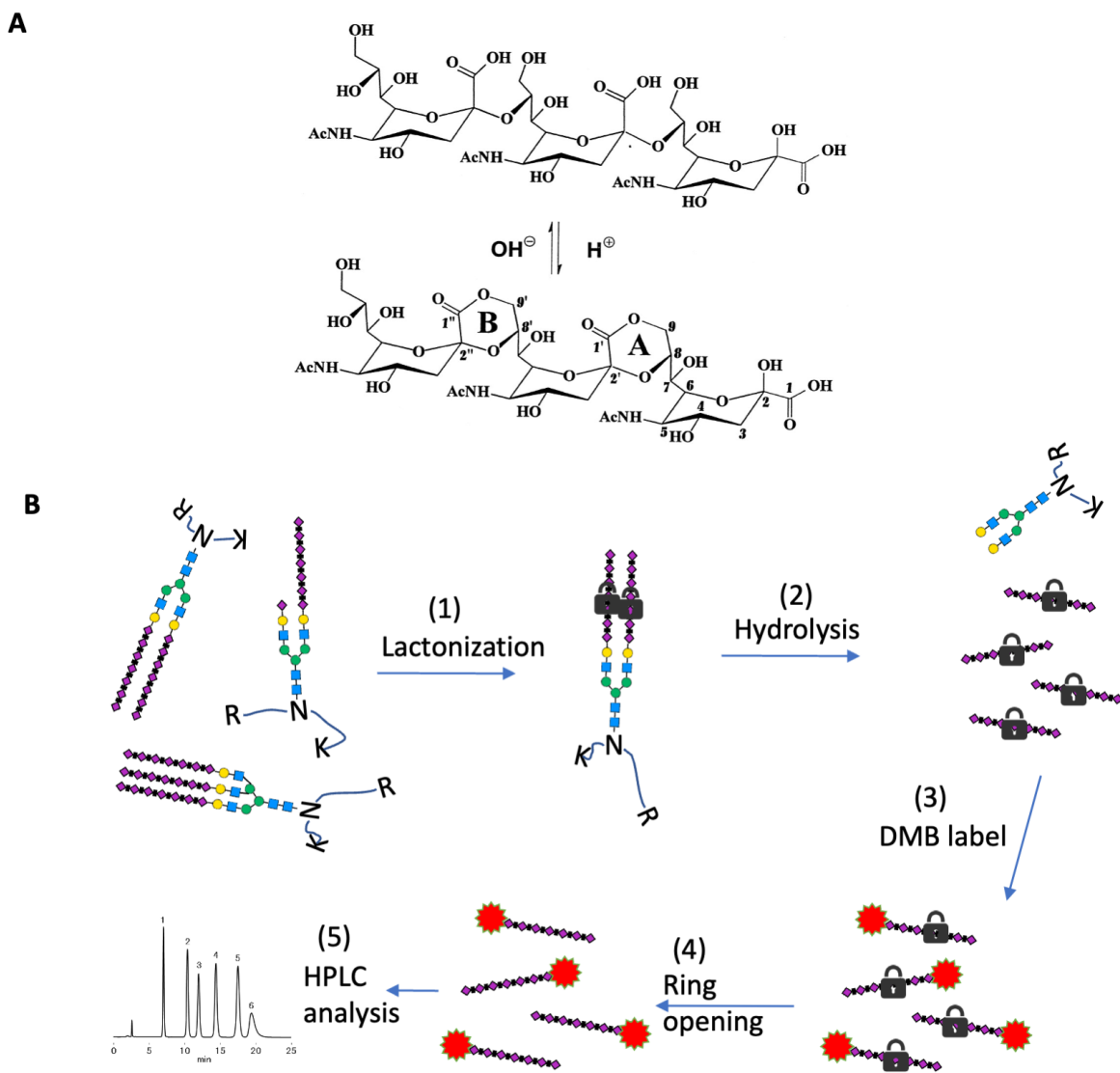


Figure 5.1: A) Structure of α 2-8 linked Neu5Ac (above), and the same following acidic lactonization. B) Overview of Lactonization-Protection-DMB procedure for DP analysis of glycoprotein polysialic acid moieties: (1) polysialic acid moieties are lactonized, (2) mild acid hydrolysis releases sialic acid structures from underlying glycans, (3) sialic acid structures are labeled with DMB fluorophore, (4) lactonization is reversed, (5) labeled structures are quantitated via anion exchange HPLC.

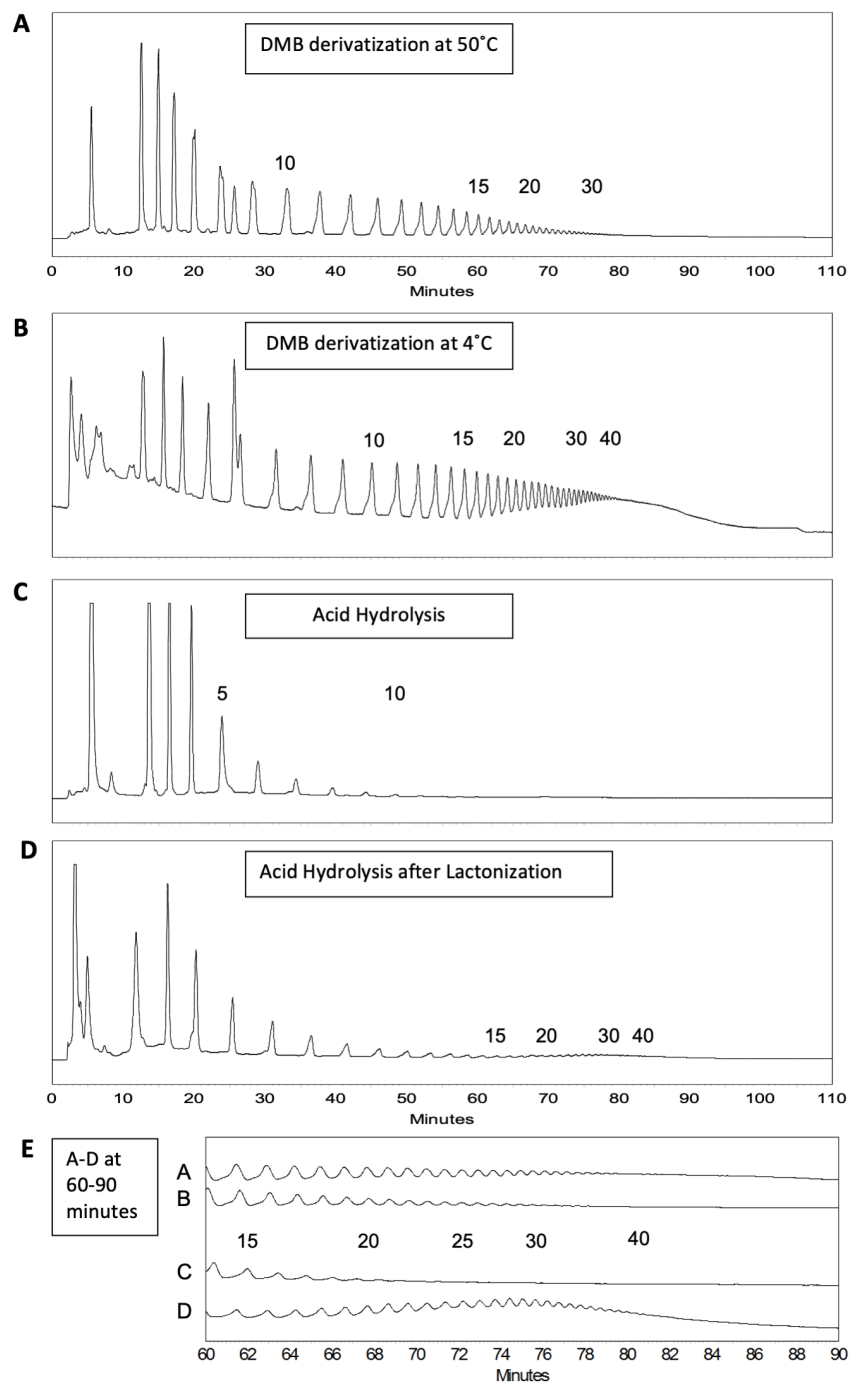


Figure 5.2: A) DMB Derivatization of Colominic acid at 50°C. B) DMB Derivatization of Colominic acid at 4°C. C) Same as B, following mild acid hydrolysis. D) Same as C, following lactonization. E) Close-up view of all chromatograms from 60-90 mins.

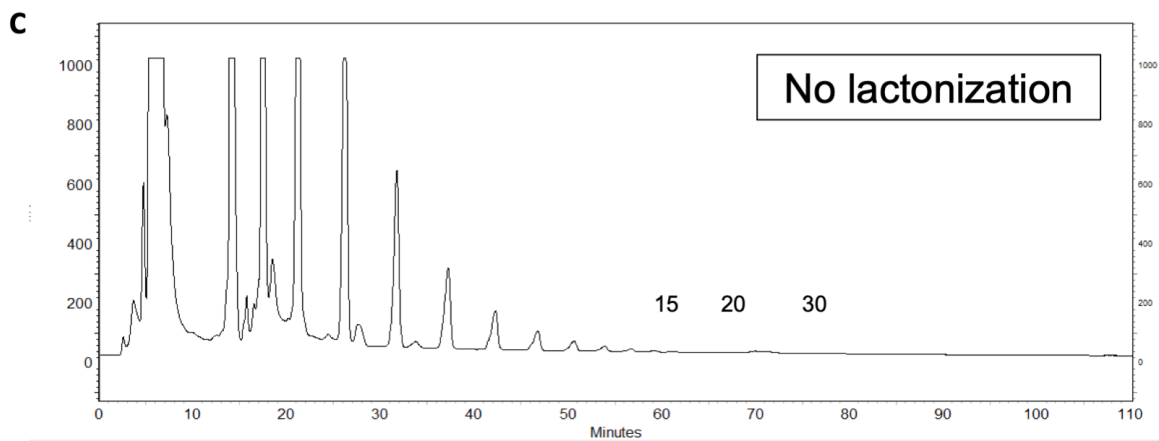
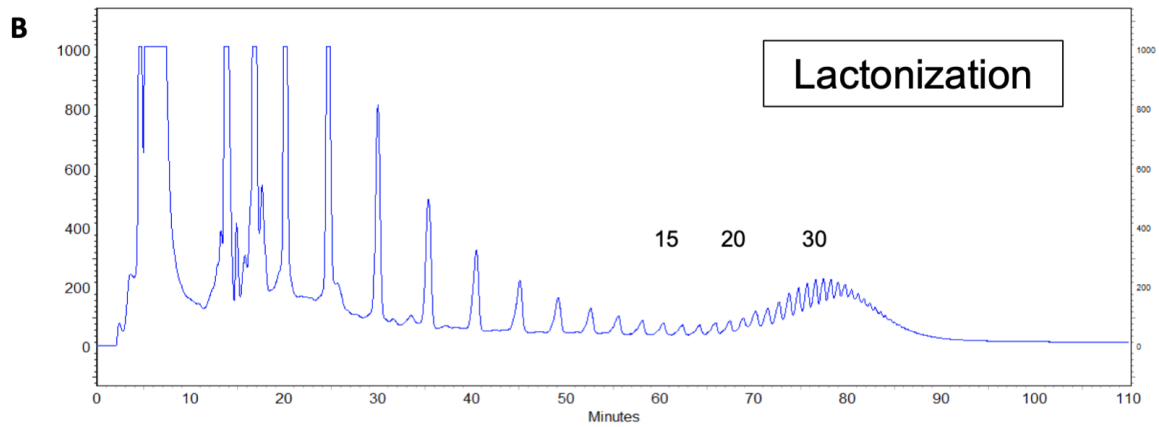
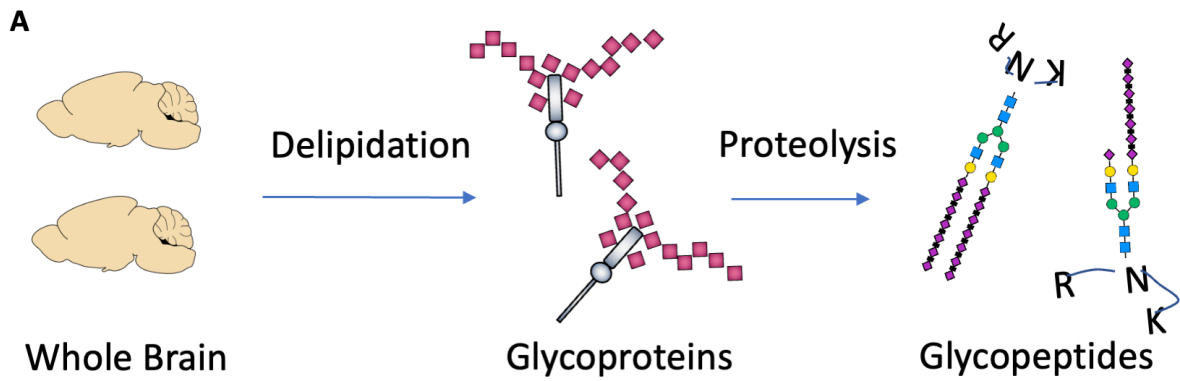


Figure 5.3: A) Overview of brain glycopeptide preparation: (1) brains are homogenized and delipidated, (2) using proteinase digestion, brain proteins are digested into glycopeptides. B) DMB-HPLC analysis of neonate mouse brain glycopeptides after lactonization step. C) DMB-HPLC analysis of neonate mouse brain glycopeptides with no lactonization.

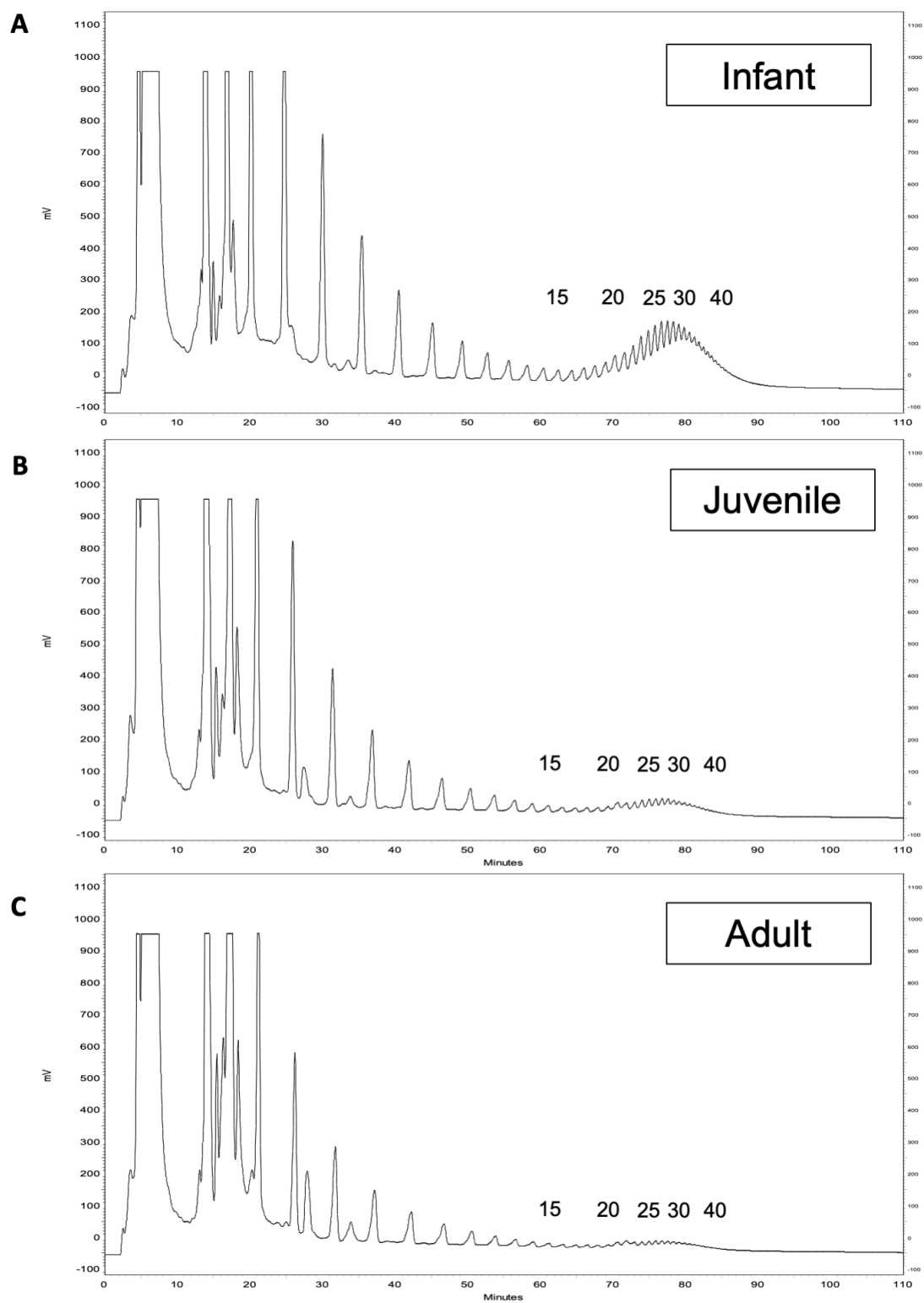


Figure 5.4: Brains subjected to Lactonization-protection-DMB HPLC analysis: A) Infant mouse (P1). B) Juvenile mouse (P14). C) Adult mouse (10 weeks).

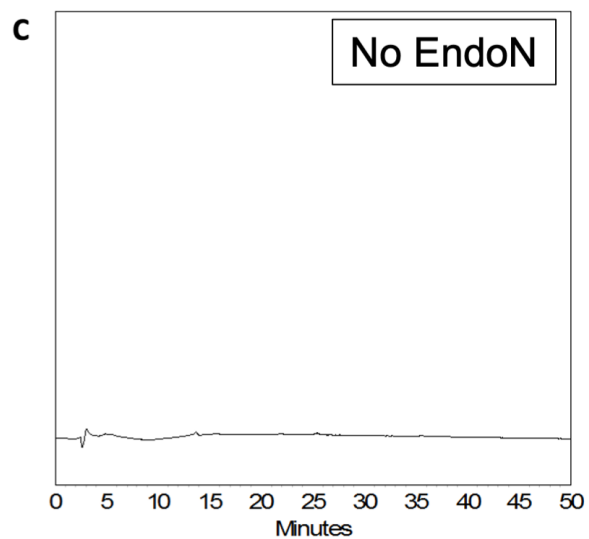
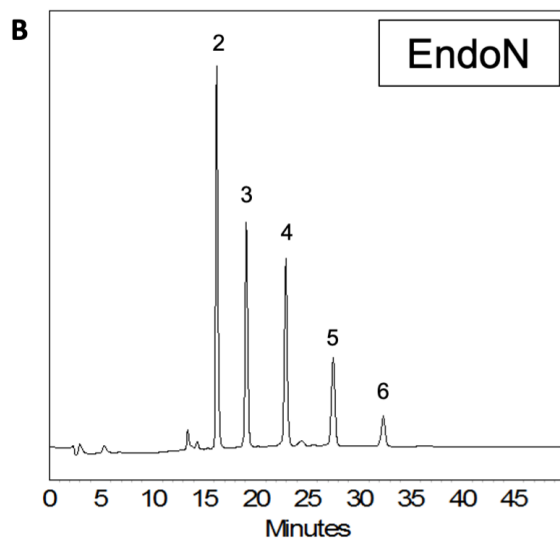
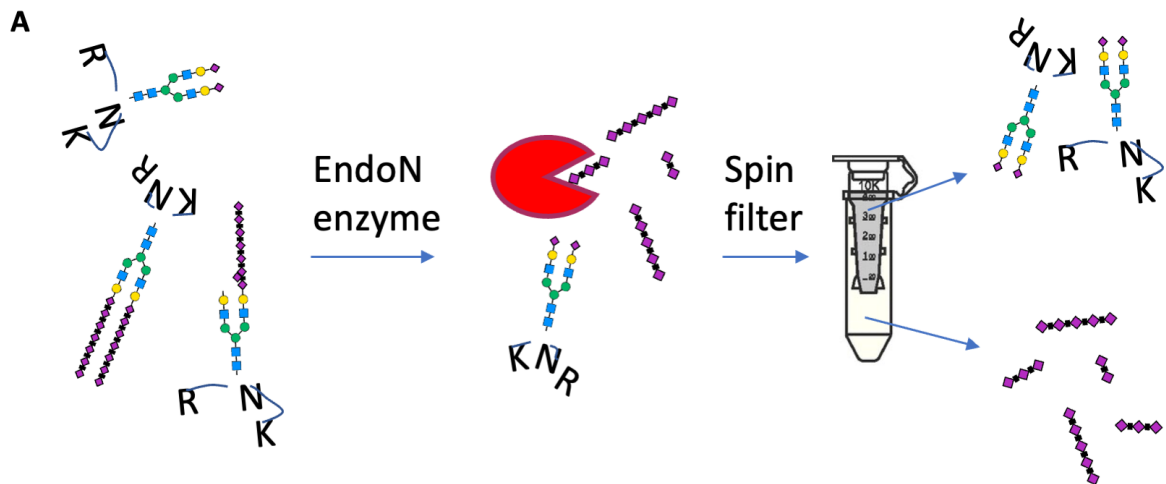


Figure 5.5: A) Overview of EndoN-sensitive Neu5Ac purification: (1) polysialic acid moieties are digested into oligosialic fragments, which are then (2) separated from glycopeptides and terminal sialic acid structures using a centrifugal filtration device. B) HPLC analysis of spin filter flow-through following EndoN digestion of colominic acid. C) HPLC analysis of spin filter flow-through of untreated colominic acid.

Bibliography

- [Aakhus et al., 1990] Aakhus, A., Stavem, P., Hovig, T., Pedersen, T., and Solum, N. (1990). Studies on a patient with thrombocytopenia, giant platelets and a platelet membrane glycoprotein with reduced amount of sialic acid. *Br J Haematol*, 74:320–329.
- [Abdullah et al., 1991] Abdullah, K., Lo, R., and Mellors, A. (1991). Cloning, nucleotide sequence, and expression of the *pasteurella haemolytica* a1 glycoprotease gene. *J Bacteriol*, 173(18):5597–5603.
- [Abegglen et al., 2015] Abegglen, L., Caulin, A., Chan, A., Lee, K., Robinson, R., Campbell, M., Kiso, W., Schmitt, D., Waddell, P., Bhaskara, S., Jensen, S., Maley, C., and Schiffman, J. (2015). Potential mechanisms for cancer resistance in elephants and comparative cellular response to dna damage in humans. *JAMA*, 314(17):1850–1860.
- [Adams et al., 2017] Adams, O., Stanczak, M., von Gunten, S., and Läubli, H. (2017). Targeting sialic acid-siglec interactions to reverse immune suppression in cancer. *Glycobiology*.
- [Ahmed et al., 2019] Ahmed, T., Adamopoulos, C., Karoulia, Z., Wu, X., Sachidanandam, R., Aaronson, S., and Poulikakos, P. (2019). Shp2 drives adaptive resistance to erk signaling inhibition in molecularly defined subsets of erk-dependent tumors. *Cell Rep*, 26(1):65–78.e5.
- [Akey et al., 2002] Akey, J., Zhang, G., Zhang, K., Jin, L., and Shriver, M. (2002). Interrogating a high-density snp map for signatures of natural selection. *Genome Res*, 12:1805–1814.
- [Akita et al., 2012] Akita, K., Yoshida, S., Ikehara, Y., Shirakawa, S., Toda, M., Inoue, M., Kitawaki, J., Nakanishi, H., Narimatsu, H., and Nakada, H. (2012). Different levels of sialyl-tn antigen expressed on muc16 in patients with endometriosis and ovarian cancer. *Int J Gynecol Cancer*, 22(4):531–538.
- [Al-Dehaimi et al., 1999] Al-Dehaimi, A., Blumsohn, A., and Eastell, R. (1999). Serum galactosyl hydroxylysine as a biochemical marker of bone resorption. *Clin Chem*, 45:676–681.
- [Anderson et al., 2010] Anderson, G., Goller, C., Justice, S., Hultgren, S., and Seed, P. (2010). Polysaccharide capsule and sialic acid-mediated regulation promote biofilm-like intracellular bacterial communities during cystitis. *Infect Immun*, 78(3):963–975.
- [Angata et al., 2001] Angata, T., Varki, N., and Varki, A. (2001). A second uniquely human mutation affecting sialic acid biology. *J Biol Chem*, 276(43):40282–40287.

- [Arai et al., 2006] Arai, M., Yamada, K., Toyota, T., Obata, N., Haga, S., Yoshida, Y., Nakamura, K., Minabe, Y., Ujike, H., Sora, I., Ikeda, K., Mori, N., Yoshikawa, T., and Itokawa, M. (2006). Association between polymorphisms in the promoter region of the sialyltransferase 8b (*siat8b*) gene and schizophrenia. *Biol Psychiatry*, 59(7):652–659.
- [Augur et al., 1995] Augur, C., Stiefel, V., Darvill, A., Albersheim, P., and Puigdomenech, P. (1995). Molecular cloning and pattern of expression of an l-fucosidase gene from pea seedlings. *J Biol Chem*, 270:24839–24843.
- [Baldwin, 1896] Baldwin, J. M. (1896). A new factor in evolution. *The American Naturalist*, 30(355):536–553.
- [Barker et al., 2012] Barker, J., Torregrossa, M., and Taylor, J. (2012). Low prefrontal *psa-ncam* confers risk for alcoholism-related behavior. *Nat Neurosci*, 15(10):1356–1358.
- [Barreiro et al., 2008] Barreiro, L., Laval, G., Quach, H., Patin, E., and Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. *Nat Genet*, 40(3):340–345.
- [BARRY and GOEBEL, 1957] BARRY, G. and GOEBEL, W. (1957). Colominic acid, a substance of bacterial origin related to sialic acid. *Nature*, 179(4552):206.
- [Beeson et al., 1989] Beeson, W., Mills, P., Phillips, R., Andress, M., and Fraser, G. (1989). Chronic disease among seventh-day adventists, a low-risk group. rationale, methodology, and description of the population. *Cancer*, 64(3):570–581.
- [Bernier et al., 2000] Bernier, P., Vinet, J., Cossette, M., and Parent, A. (2000). Characterization of the subventricular zone of the adult human brain: evidence for the involvement of *bcl-2*. *Neurosci Res*, 37(1):67–78.
- [Bochner and Zimmermann, 2015] Bochner, B. and Zimmermann, N. (2015). Role of siglecs and related glycan-binding proteins in immune responses and immunoregulation. *J Allergy Clin Immunol*, 135(3):598–608.
- [Bollu et al., 2017] Bollu, L., Mazumdar, A., Savage, M., and Brown, P. (2017). Molecular pathways: Targeting protein tyrosine phosphatases in cancer. *Clin Cancer Res*, 23(9):2136–2142.
- [Bornhöfft et al., 2018] Bornhöfft, K., Goldammer, T., Rebl, A., and Galuska, S. (2018). Siglecs: A journey through the evolution of sialic acid-binding immunoglobulin-type lectins. *Dev Comp Immunol*, 86:219–231.
- [Bramham and Messaoudi, 2005] Bramham, C. and Messaoudi, E. (2005). *Bdnf* function in adult synaptic plasticity: the synaptic consolidation hypothesis. *Prog Neurobiol*, 76(2):99–125.
- [Britten, 2002] Britten, R. (2002). Divergence between samples of chimpanzee and human dna sequences is 5 *Proc Natl Acad Sci U S A*, 99(21):13633–13635.
- [Chaisson et al., 2015] Chaisson, M., Wilson, R., and Eichler, E. (2015). Genetic variation and the de novo assembly of human genomes. *Nat Rev Genet*, 16(11):627–640.

- [Cheng et al., 1998] Cheng, M., Lin, S., Wu, S., Inoue, S., and Inoue, Y. (1998). High-performance capillary electrophoretic characterization of different types of oligo- and polysialic acid chains. *Anal Biochem*, 260(2):154–159.
- [Chiang et al., 2018] Chiang, W., Cheng, T., Chang, C., Pan, S., Changou, C., Chang, T., Lee, K., Wu, S., Chen, Y., Chuang, K., Shieh, D., Chen, Y., Tu, C., Tsui, W., and Wu, M. (2018). Carcinoembryonic antigen-related cell adhesion molecule 6 (ceacam6) promotes egf receptor signaling of oral squamous cell carcinoma metastasis via the complex n-glycosylation. *Oncogene*, 37(1):116–127.
- [Chou et al., 1998] Chou, H., Takematsu, H., Diaz, S., Iber, J., Nickerson, E., Wright, K., Muchmore, E., Nelson, D., Warren, S., and Varki, A. (1998). A mutation in human cmp-sialic acid hydroxylase occurred after the homo-pan divergence. *Proc Natl Acad Sci U S A*, 95(20):11751–11756.
- [Close et al., 2001] Close, B., Wilkinson, J., Bohrer, T., Goodwin, C., Broom, L., and Colley, K. (2001). The polysialyltransferase st8sia ii/stx: posttranslational processing and role of autopolysialylation in the polysialylation of neural cell adhesion molecule. *Glycobiology*, 11:997–1008.
- [Colley et al., 2014] Colley, K., Kitajima, K., and Sato, C. (2014). Polysialic acid: Biosynthesis, novel functions and applications. *Crit Rev Biochem Mol Biol*, 49(6):498–532.
- [Consortium et al., 2015] Consortium, . G. P., Auton, A., Brooks, L., Durbin, R., Garrison, E., Kang, H., Korbel, J., Marchini, J., McCarthy, S., McVean, G., and Abecasis, G. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.
- [Cordenonsi et al., 2011] Cordenonsi, M., Zanconato, F., Azzolin, L., Forcato, M., Rosato, A., Frasson, C., Inui, M., Montagner, M., Parenti, A., Poletti, A., Daidone, M., Dupont, S., Basso, G., Bicciato, S., and Piccolo, S. (2011). The hippo transducer taz confers cancer stem cell-related traits on breast cancer cells. *Cell*, 147(4):759–772.
- [Cox et al., 2009] Cox, E., Brennaman, L., Gable, K., Hamer, R., Glantz, L., Lamantia, A., Lieberman, J., Gilmore, J., Maness, P., and Jarskog, L. (2009). Developmental regulation of neural cell adhesion molecule in human prefrontal cortex. *Neuroscience*, 162(1):96–105.
- [Coxworth et al., 2015] Coxworth, J., Kim, P., McQueen, J., and Hawkes, K. (2015). Grandmothering life histories and human pair bonding. *Proc Natl Acad Sci U S A*, 112(38):11806–11811.
- [Crispo, 2007] Crispo, E. (2007). The baldwin effect and genetic assimilation: revisiting two mechanisms of evolutionary change mediated by phenotypic plasticity. *Evolution*, 61(11):2469–2479.
- [Cunningham et al., 1983] Cunningham, B., Hoffman, S., Rutishauser, U., Hemperly, J., and Edelman, G. (1983). Molecular topography of the neural cell adhesion molecule n-cam: surface orientation and location of sialic acid-rich and binding regions. *Proc Natl Acad Sci USA*, 80:3116–3120.
- [Darwin, 1871] Darwin, C. (1871). *The Descent of Man, and Selection in Relation to Sex*.

- [Dixon, 1981] Dixon, A. F. (1981). *The Natural History of the Gorilla*.
- [Dobin et al., 2013] Dobin, A., Davis, C., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. (2013). Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21.
- [Eastham et al., 2003] Eastham, J., Riedel, E., Scardino, P., Shike, M., Fleisher, M., Schatzkin, A., Lanza, E., Latkany, L., Begg, C., and Polyp, P. T. S. G. (2003). Variation of serum prostate-specific antigen levels: an evaluation of year-to-year fluctuations. *JAMA*, 289(20):2695–2700.
- [Eckhardt et al., 1995] Eckhardt, M., Mühlenhoff, M., Bethe, A., Koopman, J., Frosch, M., and Gerardy-Schahn, R. (1995). Molecular characterization of eukaryotic polysialyltransferase-1. *Nature*, 373:715–718.
- [Edwards et al., 1994] Edwards, U., Müller, A., Hammerschmidt, S., Gerardy-Schahn, R., and Frosch, M. (1994). Molecular analysis of the biosynthesis pathway of the alpha2,8 polysialic acid capsule by neisseria meningitidis serogroup b. *Mol Microbiol*, 14:141–149.
- [Feldman et al., 1996] Feldman, M. W., Aoki, K., and Kumm, J. (1996). Individual versus social learning: Evolutionary analysis in a fluctuating environment. *Anthropological Science*, 104(3):209–231.
- [Finne et al., 1983] Finne, J., Finne, U., Deagostini-Bazin, H., and Goridis, C. (1983). Occurrence of alpha 2-8 linked polysialosyl units in a neural cell adhesion molecule. *Biochem Biophys Res Commun*, 112(2):482–487.
- [Finne and Mäkelä, 1985] Finne, J. and Mäkelä, P. (1985). Cleavage of the polysialosyl units of brain glycoproteins by a bacteriophage endosialidase. involvement of a long oligosaccharide segment in molecular interactions of polysialic acid. *J Biol Chem*, 260(2):1265–1270.
- [Flores et al., 2019] Flores, R., Zhang, P., Wu, W., Wang, X., Ye, P., Zheng, P., and Liu, Y. (2019). Siglec genes confer resistance to systemic lupus erythematosus in humans and mice. *Cell Mol Immunol*, 16(2):154–164.
- [Fossey, 1982] Fossey, D. (1982). *The natural history of the gorilla*. by a. f. dixson. new york: Columbia university press. 1981. xviii + 202 pp., plates, figures, tables, bibliography, index. \$ 19.25 (cloth). *American Journal of Physical Anthropology*, 58(4):464–465.
- [Foussias et al., 2001] Foussias, G., Taylor, S., Yousef, G., Tropak, M., Ordon, M., and Diamandis, E. (2001). Cloning and molecular characterization of two splice variants of a new putative member of the siglec-3-like subgroup of siglecs. *Biochem Biophys Res Commun*, 284:887–899.
- [Fukasawa et al., 1999] Fukasawa, M., Nishijima, M., and Hanada, K. (1999). Genetic evidence for atp-dependent endoplasmic reticulum-to-golgi apparatus trafficking of ceramide for sphingomyelin synthesis in chinese hamster ovary cells. *J Cell Biol*, 144:673–685.
- [Galuska et al., 2008] Galuska, S., Geyer, R., Gerardy-Schahn, R., Muhlenhoff, M., and Geyer, H. (2008). Enzyme-dependent variations in the polysialylation of the neural cell adhesion molecule (ncam) in vivo. *J Biol Chem*, 283(1):17–28.

- [Galuska et al., 2007] Galuska, S., Geyer, R., Muhlenhoff, M., and Geyer, H. (2007). Characterization of oligo- and polysialic acids by maldi-tof-ms. *Anal Chem*, 79(18):7161–7169.
- [Galuska et al., 2006] Galuska, S., Oltmann-Norden, I., Geyer, H., Weinhold, B., Kuchelmeister, K., Hildebrandt, H., Gerardy-Schahn, R., Geyer, R., and Mühlenhoff, M. (2006). Polysialic acid profiles of mice expressing variant allelic combinations of the polysialyltransferases st8siaii and st8siaiv. *J Biol Chem*, 281(42):31605–31615.
- [Georges-Courbot et al., 1996] Georges-Courbot, M., Moisson, P., Leroy, E., Pingard, A., Nerrienet, E., Dubreuil, G., Wickings, E., Debels, F., Bedjabaga, I., Poaty-Mavoungou, V., Hahn, N., and Georges, A. (1996). Occurrence and frequency of transmission of naturally occurring simian retroviral infections (siv, stlv, and srv) at the cirmf primate center, gabon. *J Med Primatol*, 25(5):313–326.
- [Glode et al., 1977] Glode, M., Sutton, A., Robbins, J., McCracken, G., Gotschlich, E., Kaijser, B., and Hanson, L. (1977). Neonatal meningitis due to escherichia coli k1. *J Infect Dis*, 136 Suppl:93–S97.
- [Gross et al., 1977] Gross, R., Cheasty, T., and Rowe, B. (1977). Isolation of bacteriophages specific for the k1 polysaccharide antigen of escherichia coli. *J Clin Microbiol*, 6:548–550.
- [Guo et al., 2019] Guo, X., Elkashef, S., Loadman, P., Patterson, L., and Falconer, R. (2019). Recent advances in the analysis of polysialic acid from complex biological systems. *Carbohydr Polym*, 224:115145.
- [Hallenbeck et al., 1987] Hallenbeck, P., Vimr, E., Yu, F., Bassler, B., and Troy, F. (1987). Purification and properties of a bacteriophage-induced endo-n-acetylneuraminidase specific for poly-alpha-2,8-sialosyl carbohydrate units. *J Biol Chem*, 262:3553–3561.
- [Hane et al., 2015] Hane, M., Matsuoka, S., Ono, S., Miyata, S., Kitajima, K., and Sato, C. (2015). Protective effects of polysialic acid on proteolytic cleavage of fgf2 and probdnf/bdnf. *Glycobiology*, 25(10):1112–1124.
- [Hawkes and Coxworth, 2013] Hawkes, K. and Coxworth, J. (2013). Grandmothers and the evolution of human longevity: a review of findings and future directions. *Evol Anthropol*, 22(6):294–302.
- [Heinemann et al., 2014] Heinemann, V., von Weikersthal, L., Decker, T., Kiani, A., Vehling-Kaiser, U., Al-Batran, S., Heintges, T., Lerchenmüller, C., Kahl, C., Seipelt, G., Kullmann, F., Stauch, M., Scheithauer, W., Hielscher, J., Scholz, M., Müller, S., Link, H., Niederle, N., Rost, A., Höffkes, H., Moehler, M., Lindig, R., Modest, D., Rossius, L., Kirchner, T., Jung, A., and Stintzing, S. (2014). Folfiri plus cetuximab versus folfiri plus bevacizumab as first-line treatment for patients with metastatic colorectal cancer (fire-3): a randomised, open-label, phase 3 trial. *Lancet Oncol*, 15(10):1065–1075.
- [Hickey et al., 2019] Hickey, S., Berto, S., and Konopka, G. (2019). Chromatin decondensation by foxp2 promotes human neuron maturation and expression of neurodevelopmental disease genes. *Cell Rep*, 27(6):1699–1711.e9.

- [Holsinger and Weir, 2009] Holsinger, K. and Weir, B. (2009). Genetics in geographically structured populations: defining, estimating and interpreting f_{st} . *Nat Rev Genet*, 10(9):639–650.
- [Hooper, 1839] Hooper, R. (1839). *Lexicon Medicum; or Medical Dictionary*. Longman, London, England.
- [Huang and Reichardt, 2003] Huang, E. and Reichardt, L. (2003). Trk receptors: roles in neuronal signal transduction. *Annu Rev Biochem*, 72:609–642.
- [Huber et al., 2016] Huber, C., DeGiorgio, M., Hellmann, I., and Nielsen, R. (2016). Detecting recent selective sweeps while controlling for mutation rate and background selection. *Mol Ecol*, 25(1):142–156.
- [Huxley, 1863] Huxley, T. H. (1863). *Evidence as to man's place in nature*. D. Appelton and company, New York.
- [Inoue et al., 2001] Inoue, S., Lin, S., Lee, Y., and Inoue, Y. (2001). An ultrasensitive chemical method for polysialic acid analysis. *Glycobiology*, 11(9):759–767.
- [Irie et al., 1998] Irie, A., Koyama, S., Kozutsumi, Y., Kawasaki, T., and Suzuki, A. (1998). The molecular basis for the absence of n-glycolylneuraminic acid in humans. *J Biol Chem*, 273(25):15866–15871.
- [Jen and Wang, 2016] Jen, J. and Wang, Y. (2016). Zinc finger proteins in cancer progression. *J Biomed Sci*, 23(1):53.
- [Ji et al., 2018] Ji, C., Zhao, Y., Kou, Y., Shao, H., Guo, L., Bao, C., Jiang, B., Chen, X., Dai, J., Tong, Y., Yang, R., Sun, W., and Wang, Q. (2018). Cathepsin f knockdown induces proliferation and inhibits apoptosis in gastric cancer cells. *Oncol Res*, 26(1):83–93.
- [Takehi et al., 2001] Takehi, K., Kinoshita, M., Kitano, K., Morita, M., and Oda, Y. (2001). Lactone formation of n-acetylneuraminic acid oligomers and polymers as examined by capillary electrophoresis. *Electrophoresis*, 22:3466–3470.
- [Kanato et al., 2009] Kanato, Y., Ono, S., Kitajima, K., and Sato, C. (2009). Complex formation of a brain-derived neurotrophic factor and glycosaminoglycans. *Biosci Biotechnol Biochem*, 73(12):2735–2741.
- [Katchalsky and Spitnik, 1947] Katchalsky, A. and Spitnik, P. (1947). Potentiometric titrations of polymethacrylic acid. *J Polymer Sci*, 2:432–446.
- [Khaitovich et al., 2005] Khaitovich, P., Hellmann, I., Enard, W., Nowick, K., Leinweber, M., Franz, H., Weiss, G., Lachmann, M., and Pääbo, S. (2005). Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science*, 309(5742):1850–1854.
- [Khaitovich et al., 2004] Khaitovich, P., Muetzel, B., She, X., Lachmann, M., Hellmann, I., Dietzsch, J., Steigele, S., Do, H., Weiss, G., Enard, W., Heissig, F., Arendt, T., Nieselt-Struwe, K., Eichler, E., and Pääbo, S. (2004). Regional patterns of gene expression in human and chimpanzee brains. *Genome Res*, 14(8):1462–1473.

- [Kimura, 1991] Kimura, M. (1991). The neutral theory of molecular evolution: a review of recent evidence. *Jpn J Genet*, 66(4):367–386.
- [Koide et al., 2017] Koide, N., Kasamatsu, A., Endo-Sakamoto, Y., Ishida, S., Shimizu, T., Kimura, Y., Miyamoto, I., Yoshimura, S., Shiiba, M., Tanzawa, H., and Uzawa, K. (2017). Evidence for critical role of lymphocyte cytosolic protein 1 in oral cancer. *Sci Rep*, 7:43379.
- [Konopka et al., 2009] Konopka, G., Bomar, J., Winden, K., Coppola, G., Jonsson, Z., Gao, F., Peng, S., Preuss, T., Wohlschlegel, J., and Geschwind, D. (2009). Human-specific transcriptional regulation of cns development genes by foxp2. *Nature*, 462(7270):213–217.
- [Konopka et al., 2012] Konopka, G., Friedrich, T., Davis-Turak, J., Winden, K., Oldham, M., Gao, F., Chen, L., Wang, G., Luo, R., Preuss, T., and Geschwind, D. (2012). Human-specific transcriptional networks in the brain. *Neuron*, 75(4):601–617.
- [Koonin, 2009] Koonin, E. (2009). The origin at 150: is a new evolutionary synthesis in sight. *Trends Genet*, 25(11):473–475.
- [Kronenberg et al., 2018] Kronenberg, Z., Fiddes, I., Gordon, D., and Murali. . . , S. (2018). High-resolution comparative analysis of great ape genomes.
- [Lamba et al., 2017] Lamba, J., Chauhan, L., Shin, M., Loken, M., Pollard, J., Wang, Y., Ries, R., Aplenc, R., Hirsch, B., Raimondi, S., Walter, R., Bernstein, I., Gamis, A., Alonzo, T., and Meshinchi, S. (2017). Cd33 splicing polymorphism determines gemtuzumab ozogamicin response in de novo acute myeloid leukemia: Report from randomized phase iii children’s oncology group trial aaml0531. *J Clin Oncol*, page JCO2016712513.
- [Li et al., 2017] Li, F., Zhang, R., Li, S., and Liu, J. (2017). Ido1: An important immunotherapy target in cancer treatment. *Int Immunopharmacol*, 47:70–77.
- [Liao et al., 2014] Liao, Y., Smyth, G., and Shi, W. (2014). featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930.
- [Liberzon et al., 2011] Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. (2011). Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740.
- [Linné, 1758] Linné, C. v. (1758). *Systema Naturae*.
- [Loupakis et al., 2014] Loupakis, F., Cremolini, C., Masi, G., Lonardi, S., Zagonel, V., Salvatore, L., Cortesi, E., Tomasello, G., Ronzoni, M., Spadi, R., Zaniboni, A., Tonini, G., Buonadonna, A., Amoroso, D., Chiara, S., Carlomagno, C., Boni, C., Allegrini, G., Boni, L., and Falcone, A. (2014). Initial therapy with folfoxiri and bevacizumab for metastatic colorectal cancer. *N Engl J Med*, 371(17):1609–1618.
- [Love et al., 2014] Love, M., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol*, 15(12):550.
- [Lowenstein, 1982] Lowenstein, J. (1982). Twelve wise men at the vatican. *Nature*.

- [Lundblad, 2015] Lundblad, A. (2015). Gunnar blix and his discovery of sialic acids. fascinating molecules in glycobiology. *Ups J Med Sci*, 120(2):104–112.
- [Macauley et al., 2014] Macauley, M., Crocker, P., and Paulson, J. (2014). Siglec-mediated regulation of immune cell function in disease. *Nat Rev Immunol*, 14(10):653–666.
- [Mahauad-Fernandez et al., 2014] Mahauad-Fernandez, W., DeMali, K., Olivier, A., and Okeoma, C. (2014). Bone marrow stromal antigen 2 expressed in cancer cells promotes mammary tumor growth and metastasis. *Breast Cancer Res*, 16(6):493.
- [Manzi et al., 1994] Manzi, A., Higa, H., Diaz, S., and Varki, A. (1994). Intramolecular self-cleavage of polysialic acid. *J Biol Chem*, 269(38):23617–23624.
- [Mao et al., 2021] Mao, Y., Catacchio, C., Hillier, L., Porubsky, D., Li, R., Sulovari, A., Fernandes, J., Montinaro, F., Gordon, D., Storer, J., Haukness, M., Fiddes, I., Murali, S., Dishuck, P., Hsieh, P., Harvey, W., Audano, P., Mercuri, L., Piccolo, I., Antonacci, F., Munson, K., Lewis, A., Baker, C., Underwood, J., Hoekzema, K., Huang, T., Sorensen, M., Walker, J., Hoffman, J., Thibaud-Nissen, F., Salama, S., Pang, A., Lee, J., Hastie, A., Paten, B., Batzer, M., Diekhans, M., Ventura, M., and Eichler, E. (2021). A high-quality bonobo genome refines the analysis of hominid evolution. *Nature*, 594(7861):77–81.
- [Marchetto et al., 2013] Marchetto, M., Narvaiza, I., Denli, A., Benner, C., Lazzarini, T., Nathanson, J., Paquola, A., Desai, K., Herai, R., Weitzman, M., Yeo, G., Muotri, A., and Gage, F. (2013). Differential I1 regulation in pluripotent stem cells of humans and apes. *Nature*, 503(7477):525–529.
- [McConkey, 2004] McConkey, E. (2004). Orthologous numbering of great ape and human chromosomes is essential for comparative genomics. *Cytogenet Genome Res*, 105(1):157–158.
- [McCoy et al., 1985] McCoy, R., Vimr, E., and Troy, F. (1985). Cmp-neunac:poly-alpha-2,8-sialosyl sialyltransferase and the biosynthesis of polysialosyl units in neural cell adhesion molecules. *J Biol Chem*, 260:12695–12699.
- [McCracken et al., 1974] McCracken, G. J., Sarff, L., Glode, M., Mize, S., Schiffer, M., Robbins, J., Gotschlich, E., Orskov, I., and Orskov, F. (1974). Relation between escherichia coli k1 capsular polysaccharide antigen and clinical outcome in neonatal meningitis. *Lancet*, 2:246–250.
- [McDonough et al., 2013] McDonough, C., Gong, Y., Padmanabhan, S., Burkley, B., Langae, T., Melander, O., Pepine, C., Dominiczak, A., Cooper-Dehoff, R., and Johnson, J. (2013). Pharmacogenomic association of nonsynonymous snps in siglec12, a1bg, and the selectin region and cardiovascular outcomes. *Hypertension*, 62(1):48–54.
- [McLean et al., 2011] McLean, C., Reno, P., Pollen, A., Bassan, A., Capellini, T., Guenther, C., Indjeian, V., Lim, X., Menke, D., Schaar, B., Wenger, A., Bejerano, G., and Kingsley, D. (2011). Human-specific loss of regulatory dna and the evolution of human-specific traits. *Nature*, 471(7337):216–219.
- [Mikkelsen et al., 2005] Mikkelsen, T., Hillier, L., Eichler, E., Zody, M., and Jaffe. . . , D. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*.

- [Miranda et al., 2019] Miranda, M., Morici, J., Zanoni, M., and Bekinschtein, P. (2019). Brain-derived neurotrophic factor: A key molecule for memory in the healthy and the pathological brain. *Front Cell Neurosci*, 13:363.
- [Mitchell and Gosden, 1978] Mitchell, A. and Gosden, J. (1978). Evolutionary relationships between man and the great apes. *Sci Prog*, 65(259):273–293.
- [Mitra et al., 2011] Mitra, N., Banda, K., Altheide, T., Schaffer, L., Johnson-Pais, T., Beuten, J., Leach, R., Angata, T., Varki, N., and Varki, A. (2011). Siglec12, a human-specific segregating (pseudo)gene, encodes a signaling molecule expressed in prostate carcinomas. *J Biol Chem*, 286(26):23003–23011.
- [Nadano et al., 1986] Nadano, D., Iwasaki, M., Endo, S., Kitajima, K., Inoue, S., and Inoue, Y. (1986). A naturally occurring deaminated neuraminic acid, 3-deoxy-d-glycero-d-galactono-ulosonic acid (kdn). its unique occurrence at the nonreducing ends of oligosialyl chains in polysialoglycoprotein of rainbow trout eggs. *J Biol Chem*, 261(25):11550–11557.
- [Nakamura and Sweeley, 1987] Nakamura, M. and Sweeley, C. (1987). Glycosphingolipid glycosyltransferase assay using sephadex g-50 chromatography in aqueous phase. *Anal Biochem*, 166:230–234.
- [Nakata and Troy, 2005] Nakata, D. and Troy, F. n. (2005). Degree of polymerization (dp) of polysialic acid (polysia) on neural cell adhesion molecules (n-cams): Development and application of a new strategy to accurately determine the dp of polysia chains on n-cams. *J Biol Chem*, 280(46):38305–38316.
- [Nielsen et al., 2005] Nielsen, R., Williamson, S., Kim, Y., Hubisz, M., Clark, A., and Bustamante, C. (2005). Genomic scans for selective sweeps using snp data. *Genome Res*, 15(11):1566–1575.
- [Nowick et al., 2009] Nowick, K., Gernat, T., Almaas, E., and Stubbs, L. (2009). Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. *Proc Natl Acad Sci U S A*, 106(52):22358–22363.
- [Nuttall and Inchley, 1904] Nuttall, G. and Inchley, O. (1904). An improved method of measuring the amount of precipitum in connection with tests with precipitating antisera. *J Hyg (Lond)*, 4(2):201–206.
- [Ollomo et al., 1997] Ollomo, B., Karch, S., Bureau, P., Elissa, N., Georges, A., and Millet, P. (1997). Lack of malaria parasite transmission between apes and humans in gabon. *Am J Trop Med Hyg*, 56(4):440–445.
- [Olson, 1999] Olson, M. (1999). When less is more: gene loss as an engine of evolutionary change. *Am J Hum Genet*, 64:18–23.
- [Ong et al., 1998] Ong, E., Nakayama, J., Angata, K., Reyes, L., Katsuyama, T., Arai, Y., and Fukuda, M. (1998). Developmental regulation of polysialic acid synthesis in mouse directed by two polysialyltransferases, pst and stx. *Glycobiology*, 8:415–424.
- [Osterrieth, 2001] Osterrieth, P. (2001). Vaccine could not have been prepared in stanleyville. *Philos Trans R Soc Lond B Biol Sci*, 356(1410):839.

- [O'Hanley et al., 1985] O'Hanley, P., Lark, D., Falkow, S., and Schoolnik, G. (1985). Molecular basis of escherichia coli colonization of the upper urinary tract in balb/c mice. gal-gal pili immunization prevents escherichia coli pyelonephritis in the balb/c mouse model of human pyelonephritis. *J Clin Invest*, 75:347–360.
- [Pelkonen et al., 1989] Pelkonen, S., Pelkonen, J., and Finne, J. (1989). Common cleavage pattern of polysialic acid by bacteriophage endosialidases of different properties and origins. *J Virol*, 63(10):4409–4416.
- [Pollard et al., 2006] Pollard, K., Salama, S., Lambert, N., Lambot, M., Coppens, S., Pedersen, J., Katzman, S., King, B., Onodera, C., Siepel, A., Kern, A., Dehay, C., Igel, H., Ares, M., Van-derhaeghen, P., and Haussler, D. (2006). An rna gene expressed during cortical development evolved rapidly in humans. *Nature*, 443(7108):167–172.
- [Pollen et al., 2019] Pollen, A., Bhaduri, A., Andrews, M., Nowakowski, T., Meyerson, O., Mostajo-Radji, M., Di Lullo, E., Alvarado, B., Bedolli, M., Dougherty, M., Fiddes, I., Kronenberg, Z., Shuga, J., Leyrat, A., West, J., Bershteyn, M., Lowe, C., Pavlovic, B., Salama, S., Haussler, D., Eichler, E., and Kriegstein, A. (2019). Establishing cerebral organoids as models of human-specific brain evolution. *Cell*, 176(4):743–756.e17.
- [Prendergast et al., 2017] Prendergast, G., Malachowski, W., DuHadaway, J., and Muller, A. (2017). Discovery of ido1 inhibitors: From bench to bedside. *Cancer Res*, 77(24):6795–6811.
- [Puente et al., 2006] Puente, X., Velasco, G., Gutierrez-Fernandez, A., Bertranpetit, J., King, M., and Lopez-Otin, C. (2006). Comparative analysis of cancer genes in the human and chimpanzee genomes. *BMC Genomics*, 7:15.
- [Pybus et al., 2014] Pybus, M., Dall'Olio, G., Luisi, P., Uzkudun, M., Carreño-Torres, A., Pavlidis, P., Laayouni, H., Bertranpetit, J., and Engelken, J. (2014). 1000 genomes selection browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Res*, 42(Database issue):D903–9.
- [Riss and Goodall, 1976] Riss, D. and Goodall, J. (1976). Sleeping behavior and associations in a group of captive chimpanzees. *Folia Primatol (Basel)*, 25(1):1–11.
- [Robbins et al., 1974] Robbins, J., McCracken, G. J., Gotschlich, E., Orskov, F., Orskov, I., and Hanson, L. (1974). Escherichia coli k1 capsular polysaccharide associated with neonatal meningitis. *N Engl J Med*, 290:1216–1220.
- [Rousselot et al., 1995] Rousselot, P., Lois, C., and Alvarez-Buylla, A. (1995). Embryonic (psa) n-cam reveals chains of migrating neuroblasts between the lateral ventricle and the olfactory bulb of adult mice. *J Comp Neurol*, 351(1):51–61.
- [Rousselot and Nottebohm, 1995] Rousselot, P. and Nottebohm, F. (1995). Expression of polysialylated n-cam in the central nervous system of adult canaries and its possible relation to function. *J Comp Neurol*, 356:629–640.
- [Rutishauser et al., 1985] Rutishauser, U., Watanabe, M., Silver, J., Troy, F., and Vimr, E. (1985). Specific alteration of ncam-mediated cell adhesion by an endoneuraminidase. *J Cell Biol*, 101(5 Pt 1):1842–1849.

- [Sanai et al., 2011] Sanai, N., Nguyen, T., Ihrie, R., Mirzadeh, Z., Tsai, H., Wong, M., Gupta, N., Berger, M., Huang, E., Garcia-Verdugo, J., Rowitch, D., and Alvarez-Buylla, A. (2011). Corridors of migrating neurons in the human brain and their decline during infancy. *Nature*, 478(7369):382–386.
- [Sarich and Wilson, 1967a] Sarich, V. and Wilson, A. (1967a). Immunological time scale for hominid evolution. *Science*, 158(3805):1200–1203.
- [Sarich and Wilson, 1967b] Sarich, V. and Wilson, A. (1967b). Rates of albumin evolution in primates. *Proc Natl Acad Sci U S A*, 58(1):142–148.
- [Sato et al., 1998] Sato, C., Inoue, S., Matsuda, T., and Kitajima, K. (1998). Development of a highly sensitive chemical method for detecting alpha2-8-linked oligo/polysialic acid residues in glycoproteins blotted on the membrane. *Anal Biochem*, 261:191–197.
- [Sato et al., 1995] Sato, C., Kitajima, K., Inoue, S., Seki, T., Troy, F. n., and Inoue, Y. (1995). Characterization of the antigenic specificity of four different anti-(alpha 2-8-linked polysialic acid) antibodies using lipid-conjugated oligo/polysialic acids. *J Biol Chem*, 270(32):18923–18928.
- [Schmidt, 1978] Schmidt, R. (1978). Systemic pathology of chimpanzees. *J Med Primatol*, 7:274–318.
- [Schnaar et al., 2014] Schnaar, R., Gerardy-Schahn, R., and Hildebrandt, H. (2014). Sialic acids in the brain: gangliosides and polysialic acid in nervous system development, stability, disease, and regeneration. *Physiol Rev*, 94(2):461–518.
- [Schridder and Kern, 2016] Schridder, D. and Kern, A. (2016). S/hic: Robust identification of soft and hard sweeps using machine learning. *PLoS Genet*, 12(3):e1005928.
- [Seki and Arai, 1991] Seki, T. and Arai, Y. (1991). The persistent expression of a highly polysialylated ncam in the dentate gyrus of the adult rat. *Neurosci Res*, 12:503–513.
- [Shaw et al., 2014] Shaw, A., Tiwari, Y., Kaplan, W., Heath, A., Mitchell, P., Schofield, P., and Fullerton, J. (2014). Characterisation of genetic variation in st8sia2 and its interaction region in ncam1 in patients with bipolar disorder. *PLoS One*, 9(3):e92556.
- [Shea, 1989] Shea, B. (1989). Heterochrony in human evolution: The case for neoteny reconsidered. *American Journal of Physical Anthropology*.
- [Sibley and Ahlquist, 1984] Sibley, C. and Ahlquist, J. (1984). The phylogeny of the hominoid primates, as indicated by dna-dna hybridization. *J Mol Evol*, 20(1):2–15.
- [Singh et al., 2009] Singh, A., Greninger, P., Rhodes, D., Koopman, L., Violette, S., Bardeesy, N., and Settleman, J. (2009). A gene expression signature associated with “k-ras addiction” reveals regulators of emt and tumor cell survival. *Cancer Cell*, 15(6):489–500.
- [Somel et al., 2009] Somel, M., Franz, H., Yan, Z., Lorenc, A., Guo, S., Giger, T., Kelso, J., Nickel, B., Dannemann, M., Bahn, S., Webster, M., Weickert, C., Lachmann, M., Pääbo, S., and Khaitovich, P. (2009). Transcriptional neoteny in the human brain. *Proc Natl Acad Sci U S A*, 106(14):5743–5748.

- [Spiteri et al., 2007] Spiteri, E., Konopka, G., Coppola, G., Bomar, J., Oldham, M., Ou, J., Vernes, S., Fisher, S., Ren, B., and Geschwind, D. (2007). Identification of the transcriptional targets of *foxp2*, a gene linked to speech and language, in developing human brain. *Am J Hum Genet*, 81(6):1144–1157.
- [Stecker et al., 2011] Stecker, K., Vieth, M., Koschel, A., Wiedenmann, B., Röcken, C., and Anders, M. (2011). Impact of the coxsackievirus and adenovirus receptor on the adenoma-carcinoma sequence of colon cancer. *Br J Cancer*, 104(9):1426–1433.
- [Subramanian et al., 2005] Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E., and Mesirov, J. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550.
- [Tajik et al., 2020] Tajik, A., Phillips, K., Nitz, M., and Willis, L. (2020). A new elisa assay demonstrates sex differences in the concentration of serum polysialic acid. *Anal Biochem*, 600:113743.
- [Tajima, 1989] Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, 123(3):585–595.
- [Thompson et al., 2013] Thompson, M., Foley, D., and Colley, K. (2013). The polysialyltransferases interact with sequences in two domains of the neural cell adhesion molecule to allow its polysialylation. *J Biol Chem*, 288(10):7282–7293.
- [Traill, 1821] Traill, T. (1821). Observations on the anatomy of the orang outang. *Mem Wiener Nat Hist Soc Edinburgh.*, 1821(3):1–49.
- [Troy et al., 1982] Troy, F., Vijay, I., McCloskey, M., and Rohr, T. (1982). Synthesis of capsular polymers containing polysialic acid in escherichia coli 07-k1. *Methods Enzymol*, 83:540–548.
- [Tyson, 1699] Tyson, E. (1699). *Orang-Outang, Sive Homo Sylvestris*.
- [Ungewitter and Scrable, 2009] Ungewitter, E. and Scrable, H. (2009). Antagonistic pleiotropy and p53. *Mech Ageing Dev*, 130(1-2):10–17.
- [van den Ende et al., 1980] van den Ende, M., Brotman, B., and Prince, A. (1980). An open air holding system for chimpanzees in medical experiments. *Dev Biol Stand*, 45:95–98.
- [Varki and Altheide, 2005] Varki, A. and Altheide, T. (2005). Comparing the human and chimpanzee genomes: searching for needles in a haystack. *Genome Res*, 15(12):1746–1758.
- [Varki and Diaz, 1984] Varki, A. and Diaz, S. (1984). The release and purification of sialic acids from glycoconjugates: methods to minimize the loss and migration of o-acetyl groups. *Anal Biochem*, 137:236–247.
- [Varki and Gagneux, 2017] Varki, A. and Gagneux, P. (2017). How different are humans and “great apes”? a matrix of comparative anthropogeny on human nature. pages 151–160. Elsevier.

- [Varki et al., 2017a] Varki, A., Schnaar, R., and Crocker, P. (2017a). I-type lectins essentials of glycobiology. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY).
- [Varki et al., 2017b] Varki, A., Schnaar, R., and Schauer, R. (2017b). Sialic acids and other nonulosonic acids essentials of glycobiology. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY).
- [Varki and Varki, 2015] Varki, N. and Varki, A. (2015). On the apparent rarity of epithelial cancers in captive chimpanzees. *Philos Trans R Soc Lond B Biol Sci*, 370(1673).
- [Vitti et al., 2013] Vitti, J., Grossman, S., and Sabeti, P. (2013). Detecting natural selection in genomic data. *Annu Rev Genet*, 47:97–120.
- [Wang et al., 2014] Wang, F., Liu, X., Yang, P., Guo, L., Liu, C., Li, H., Long, S., Shen, Y., and Wan, H. (2014). Loss of *tacstd2* contributed to squamous cell carcinoma progression through attenuating *tap63*-dependent apoptosis. *Cell Death Dis*, 5:e1133.
- [Warburton et al., 1973] Warburton, D., Firschein, I., Miller, D., and Warburton, F. (1973). Karyotype of the chimpanzee, pan troglodytes, based on measurements and banding pattern: comparison to the human karyotype. *Cytogenet Cell Genet*, 12(6):453–461.
- [Weir and Cockerham, 1984] Weir, B. and Cockerham, C. (1984). Estimating f-statistics for the analysis of population structure. *Evolution*, 38(6):1358–1370.
- [Weiss et al., 2021] Weiss, C., Harshman, L., Inoue, F., Fraser, H., Petrov, D., Ahituv, N., and Gokhman, D. (2021). The cis-regulatory effects of modern human-specific variants. *Elife*, 10:e63713.
- [Whitfield and Troy, 1984] Whitfield, C. and Troy, F. (1984). Biosynthesis and assembly of the polysialic acid capsule in escherichia coli k1. activation of sialyl polymer synthesis in inactivate sialyltransferase complexes requires protein synthesis. *J Biol Chem*, 259:12776–12780.
- [Wilson and Sarich, 1969] Wilson, A. and Sarich, V. (1969). A molecular time scale for human evolution. *Proc Natl Acad Sci U S A*, 63(4):1088–1093.
- [Wright, 1932] Wright, S. (1932). *The roles of mutation, inbreeding, crossbreeding and selection in evolution*.
- [Wright, 1980] Wright, S. (1980). Genic and organismic selection. *Evolution*.
- [Xu et al., 2018] Xu, C., Li, Q., Efimova, O., He, L., Tatsumoto, S., Stepanova, V., Oishi, T., Uono, T., Yamaguchi, K., Shigenobu, S., Kakita, A., Nawa, H., Khaitovich, P., and Go, Y. (2018). Human-specific features of spatial gene expression and regulation in eight brain regions. *Genome Res*, 28(8):1097–1110.
- [Xu et al., 2005] Xu, R., Yu, Y., Zheng, S., Zhao, X., Dong, Q., He, Z., Liang, Y., Lu, Q., Fang, Y., Gan, X., Xu, X., Zhang, S., Dong, Q., Zhang, X., and Feng, G. (2005). Overexpression of *shp2* tyrosine phosphatase is implicated in leukemogenesis in adult human leukemia. *Blood*, 106(9):3142–3149.

- [Yngvadottir et al., 2009] Yngvadottir, B., Xue, Y., Searle, S., Hunt, S., Delgado, M., Morrison, J., Whittaker, P., Deloukas, P., and Tyler-Smith, C. (2009). A genome-wide survey of the prevalence and evolutionary forces acting on human nonsense snps. *Am J Hum Genet*, 84(2):224–234.
- [Yu et al., 2001] Yu, Z., Lai, C., Maoui, M., Banville, D., and Shen, S. (2001). Identification and characterization of s2v, a novel putative siglec that contains two v set ig-like domains and recruits protein-tyrosine phosphatases shps. *J Biol Chem*, 276(26):23816–23824.
- [Zhai et al., 2018] Zhai, L., Ladomersky, E., Lenzen, A., Nguyen, B., Patel, R., Lauing, K., Wu, M., and Wainwright, D. (2018). Ido1 in cancer: a gemini of immune checkpoints. *Cell Mol Immunol*.
- [Zhai et al., 2015] Zhai, L., Spranger, S., Binder, D., Gritsina, G., Lauing, K., Giles, F., and Wainwright, D. (2015). Molecular pathways: Targeting ido1 and other tryptophan dioxygenases for cancer immunotherapy. *Clin Cancer Res*, 21(24):5427–5433.
- [Zhang and Lee, 1999] Zhang, Y. and Lee, Y. (1999). Acid-catalyzed lactonization of alpha2,8-linked oligo/polysialic acids studied by high performance anion-exchange chromatography. *J Biol Chem*, 274(10):6183–6189.
- [Zhou et al., 2018] Zhou, J., Oswald, D., Oliva, K., Kreisman, L., and Cobb, B. (2018). The glycoscience of immunity. *Trends Immunol*, 39(7):523–535.
- [Zhou et al., 2008] Zhou, X., Coad, J., Ducatman, B., and Agazie, Y. (2008). Shp2 is up-regulated in breast cancer cells and in infiltrating ductal carcinoma of the breast, implying its involvement in breast oncogenesis. *Histopathology*, 53(4):389–402.