

Léa Chuzel

**Application of functional
metagenomics to the field
of glycobiology**



Application of functional metagenomics to the field of glycobiology

Dissertation
zur Erlangung des akademischen Grades

**Doctor rerum naturalium
(Dr. rer. nat.)**

von Dipl. Biotech. Léa Chuzel
geboren am 15. November 1993 in Lyon (Frankreich)

genehmigt durch die Fakultät für Verfahrens- und Systemtechnik der Otto-
von-Guericke-Universität Magdeburg

Promotionskommission:

Prof. Dr. rer. nat. habil. Weiß (Vorsitz)
Prof. Dr.-Ing. Udo Reichl (Gutachter)
Dr. Erdmann Rapp (Gutachter)
Dr. Christopher H. Taron (Gutachter)

eingereicht am: 30. Juni 2021

Promotionskolloquium am: 1. November 2021

Forschungsberichte aus dem Max-Planck-Institut
für Dynamik komplexer technischer Systeme

Band 55

Léa Chuzel

**Application of functional metagenomics
to the field of glycobiology**

Shaker Verlag
Düren 2022

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <http://dnb.d-nb.de>.

Zugl.: Magdeburg, Univ., Diss., 2021

Copyright Shaker Verlag 2022

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publishers.

Printed in Germany.

ISBN 978-3-8440-8437-5

ISSN 1439-4804

Shaker Verlag GmbH • Am Langen Graben 15a • 52353 Düren

Phone: 0049/2421/99011-0 • Telefax: 0049/2421/99011-9

Internet: www.shaker.de • e-mail: info@shaker.de

Abstract

Bacteria account for ~15% of the earth's biomass and are the second largest biomass contributor after plants. Bacteria are ubiquitous and have adapted to live in all habitable ecosystems on our planet resulting in a vast diversity estimated to be 1 trillion (10^{12}) different species. The genomes of these microbes represent an extraordinary resource for novel enzyme discovery. The field of metagenomics strives to access these genomes, in particular, those from uncultivable species that have been otherwise out of reach.

Glycobiology is a field of research that addresses the structure, function, and biology of glycans and glycoconjugates. Glycans have wide-ranging importance in both basic biology and pharmaceutical science. Aberrant glycosylation has been implicated in many diseases. Additionally, glycosylation of therapeutic proteins affects their safety and efficacy, and is monitored during drug development and manufacturing. Enzymes acting on glycans are important and can represent therapeutic targets (*e.g.*, influenza A neuraminidase), therapeutic agents (*e.g.*, use of sialidases in cancer immunotherapy), or essential glycoanalytical tools (*e.g.*, PNGase F) that help deconvolute glycan structure. In this thesis, a functional metagenomic workflow was created for discovery of new enzymes that act upon glycans. This was used to answer fundamental and application-driven questions in the field of glycobiology.

The functional metagenomic workflow established in this thesis relies on large-insert metagenomic libraries created in *Escherichia coli*. A collection of almost 100,000 clones was created from diverse ecosystems including extreme environments, and is estimated to contain 3-4 million of environmental genes. Three activity-based screens were executed using libraries in this collection. The first led to identification of a novel exosialidase having a unique catalytic mechanism and new protein structure that defined a new glycoside hydrolase family (GH156). The second screen isolated two sialidases with a preference for a non-human form of sialic acid, a specificity not previously described. Finally, a third screen identified

two enzymatic activities: a sugar-specific sulfatase and a sulfate-dependent hexosaminidase that can act on sulfated glycans, an important chemical modification of *N*-glycans for which well-defined analytical tools (*e.g.*, enzymes acting specifically on sulfated sugars) have yet to be established.

To summarize, this work demonstrates the benefit of using functional metagenomics to identify precise enzyme specificities that can answer questions in the field of glycobiology and address technical challenges in glycoanalytics. It highlights that this method of discovery can yield novel protein families that act on glycans, unusual enzymatic specificities, and needed analytical tools. The workflow employed in this thesis is versatile and can easily be adapted to other enzyme discovery projects.

Zusammenfassung

Mit einer Biomasse von etwa 15% sind Bakterien, nach den Pflanzen, der zweitgrößte Produzent von Biomasse auf der Erde. Bakterien sind ubiquitär und haben sich an das Leben in allen Habitaten auf unserem Planeten angepasst, was zu einer enormen Vielfalt von geschätzten 1 Billion (10^{12}) verschiedenen Arten führt. Die Genome dieser Mikroben stellen eine außergewöhnliche Ressource für die Entdeckung (und Nutzung) neuer bakterieller Enzyme dar. Der Bereich der Metagenomik strebt nach der Erforschung dieser Genome insbesondere von Arten, die nicht kultiviert werden können und sonst unzugänglich wären.

Die Glykobiologie ist ein Forschungsgebiet, das sich mit der Struktur, Funktion und Biologie von Glykanen und Glykokonjugaten befasst. Glykane haben eine weitreichende Bedeutung sowohl in der Grundlagenbiologie als auch in der (bio)pharmazeutischen Forschung und Produktion.

Veränderungen der Glykosylierung werden mit einer Vielzahl an Erkrankungen in Verbindung gebracht. Darüber hinaus beeinflusst die Glykosylierung therapeutischer Proteine hinsichtlich deren Sicherheit und Wirksamkeit, und wird daher während der Arzneimittelentwicklung und -herstellung überwacht. Enzyme, die auf Glykane wirken, sog. Glyko-Enzyme können therapeutische Ziele (z. B. Influenza-A-Neuraminidase) oder therapeutische Wirkstoffe (z. B. Verwendung von Sialidasen in Krebsimmuntherapie) darstellen oder als essenzielle glykoanalytische Werkzeuge (z.B. PNGase F) fungieren, die zur Aufklärung der Glykanstruktur beitragen. In dieser Dissertation wurde ein funktionaler metagenomischer Workflow zur Entdeckung neuer Glyko-Enzyme entwickelt. Dieser Workflow wurde zur Beantwortung grundlegender und anwendungsorientierter Fragen im Bereich der Glykobiologie eingesetzt.

Der in dieser Arbeit etablierte funktionelle Metagenomik-Workflow basiert auf metagenomischen Bibliotheken mit großen Inserts, die in *Escherichia coli* erstellt wurden.

Eine Sammlung von fast 100.000 Klonen wurde aus verschiedenen Ökosystemen einschließlich extremer Umgebungen erstellt und enthält schätzungsweise 3-4 Millionen Umweltgene. Drei aktivitätsbasierte Screenings wurden unter Verwendung von Bibliotheken in dieser Sammlung ausgeführt. Das erste Screening führte zur Identifizierung einer neuen Exosialidase mit einem einzigartigen katalytischen Mechanismus und einer neuen Proteinstruktur, die eine neue Glykosid-Hydrolase Familie (GH156) definierte. Das zweite Screening isolierte zwei Sialidasen mit einer Präferenz für eine nicht-humane Form von Sialinsäure, eine Spezifität, die zuvor nicht beschrieben wurde. Schließlich identifizierte ein drittes Screening zwei enzymatische Aktivitäten: Eine zuckerabhängige Sulfatase und eine sulfatabhängige Hexosaminidase, die auf sulfatierte Glykane einwirken können, eine wichtige chemische Modifikation von N-Glykanen, für die noch genau definierte analytische Werkzeuge (z.B. Enzyme die spezifisch auf sulfatierte Zucker wirken) entwickelt werden müssen.

Zusammenfassend demonstriert diese Arbeit den Nutzen der funktionellen Metagenomik zur Identifizierung präziser Enzymspezifitäten, die Fragestellungen im Bereichen der Glykobiologie beantworten können und technische Herausforderungen der Glykoanalytik angehen. Es unterstreicht, dass diese Entdeckungsmethode neue, auf Glykane wirkende Proteinfamilien, ungewöhnliche enzymatische Spezifitäten und benötigte analytische Werkzeuge hervorbringen kann. Der in dieser Arbeit verwendete Workflow ist vielseitig und kann problemlos für Projekte zur Entdeckung anderer Enzyme angepasst werden.

List of abbreviations	IX
1 Introduction	1
2 Theoretical background	6
2.1 Metagenomics.....	8
2.1.1 Microorganisms: a tremendous diversity previously inaccessible	8
2.1.2 Sequence-based metagenomics	10
2.1.3 Functional metagenomics.....	13
2.1.3.1 Choosing a vector-host system	14
2.1.3.2 Screening strategies	16
2.1.3.3 Application of functional metagenomics.....	18
2.1.4 Contrasting approaches to enzyme discovery	20
2.2 Glycobiology	23
2.2.1 A few definitions	23
2.2.2 Glycan synthesis	23
2.2.3 Tremendous glycan diversity	24
2.2.4 Glycan functions	26
2.2.5 Importance of protein glycosylation in the pharmaceutical industry.....	28
2.2.6 Glycoanalytics	29
3 Development of a functional metagenomic workflow for enzyme discovery	36
3.1 Introduction	38
3.2 Material and methods	39
3.2.1 Environmental DNA extraction.....	39
3.2.1.1 eDNA isolation from terrestrial samples.....	39
3.2.1.2 eDNA isolation from aquatic environments.....	41
3.2.1.3 eDNA isolation from human feces.....	42
3.2.2 Fosmid eDNA library construction in <i>E. coli</i>	43
3.2.2.1 Library principle	43
3.2.2.2 Library assembly procedure.....	46
3.2.3 Library quality assessments	47
3.2.4 High-throughput functional screening	48
3.2.4.1 Agar plate-based enzyme screening	48
3.2.4.2 Lysate-based enzyme screening	48
3.2.5 Fosmid sequencing and bioinformatics	49
3.2.5.1 Single clone sequencing.....	49
3.2.5.2 Multiplexed-sequencing	50
3.2.5.3 Blue Pippin size-selection.....	51
3.2.5.4 RSII sequencing and <i>de novo</i> assembly.....	51
3.2.5.5 ORF prediction and ORF map drawings.....	52
3.3 Results and discussion	53
3.3.1 Metagenomics libraries and collection.....	53
3.3.1.1 Environmental DNA	53
3.3.1.2 The NEB Collection in November 2019	55

3.3.1.3 Library quality	62
3.3.2 Plate-based and lysate-based screenings	64
3.3.2.1 Plate-based screening	65
3.3.2.2 Lysate-based screening	67
3.3.2.3 Hit definition	70
3.3.3 Sequencing the hits and generation of fosmid maps	72
3.4 Chapter 3 conclusion	77
4 Screening metagenomic libraries for sialidases	79
4.1 Introduction to sialic acid biology	80
4.2 Discovery of the novel GH156 sialidase family	86
4.2.1 Material and methods	86
4.2.1.1 Screening for sialidases	86
4.2.1.2 Tn5 mutagenesis	86
4.2.1.3 <i>In vitro</i> and <i>in vivo</i> sialidase expression	87
4.2.1.4 Sialidase biochemical characterization	88
4.2.1.5 NMR spectroscopy	90
4.2.1.6 <i>Armatimonadetes</i> homolog expression	91
4.2.2 Results	91
4.2.2.1 Functional screening	91
4.2.2.2 Sialidase gene identification	93
4.2.2.3 Sialidase biochemical characterization	97
4.2.2.4 ORF12p sialidase reaction mechanism	101
4.2.2.5 ORF12p sialidase family phylogeny	104
4.2.3 Discussion	106
4.3 Biostructural characterization of GH156	108
4.3.1 Material and methods	108
4.3.1.1 Selenomethionine protein labeling	108
4.3.1.2 SEC-MALLS	110
4.3.1.3 Crystallization of EnvSia156 and EnvSia156 substrate and inhibitor complexes	110
4.3.1.4 3D structure solution	111
4.3.1.5 Site directed mutagenesis and activity assay of generated mutants	112
4.3.2 Results	113
4.3.2.1 Expression and purification of EnvSia156 SeMet mutant for MAD	113
4.3.2.2 Structure of EnvSia156 defines an unusual sialidase fold ...	116
4.3.2.3 Mechanism of EnvSia156 and definition of its active center	121
4.3.3 Discussion	128
4.4 Identification of unconventional sialidases	129
4.4.1 Material and methods	130
4.4.1.1 Screening for Neu5Gc specific sialidases	130
4.4.1.2 PacBio sequencing and enzyme identification	130
4.4.1.3 Expression and purification of C19 and C22 sialidases	130
4.4.1.4 Determination of sialidase substrate preference	131










4.4.2 Results.....	131
4.4.2.1 Functional metagenomic screening	131
4.4.2.2 Identification of C22 and C19 sialidases	134
4.4.2.3 Neu5Gc preferring sialidases activity.....	135
4.4.3 Discussion	136
4.5 Chapter 4 conclusion	140
5 Applying functional metagenomics to post-glycosylation modifications.....	141
5.1 Introduction to post glycosylation modifications with a focus on sulfation	142
5.2 Material and methods	148
5.2.1 Screening for sulfated glycan using a coupled assay	148
5.2.2 F1-ORF13 and F10-ORF19 hexosaminidase <i>in vivo</i> expression and purification.....	148
5.2.3 F1-ORF13 activity on sulfated monosaccharides	149
5.2.4 Enzyme activities on <i>N</i> -glycans	150
5.2.4.1 <i>N</i> -glycan release and APTS-labeling.....	150
5.2.4.2 Substrate preparation	150
5.2.4.3 F1-ORF13 activity on <i>N</i> -glycans released from hlgA and human urokinase.....	152
5.2.4.4 <i>N</i> -Glycan enrichment using F1-ORF13.....	153
5.2.4.5 F10-ORF19 activity on <i>N</i> -glycans released from hlgA and human urokinase.....	153
5.2.4.6 xCGE-LIF analysis.....	154
5.3 Results	154
5.3.1 Functional metagenomic screening for sulfatases	154
5.3.2 Analysis of fosmid DNA sequences.....	157
5.3.3 Identifying genes encoding active sulfatases using <i>in vitro</i> protein expression	161
5.3.4 Protein sequence analysis of active sulfatases	162
5.3.5 Characterization of F1-ORF13 sulfatase	163
5.3.5.1 Determination of F1-ORF13 sulfatase specificity using sulfated monosaccharides	163
5.3.5.2 F1-ORF13 sulfatase activity on GlcNAc-6-SO ₄ in intact <i>N</i> -glycans.....	164
5.3.5.3 F1-ORF13 binds GlcNAc-6-SO ₄ -containing <i>N</i> -glycans in absence of calcium	167
5.3.6 Identifying genes encoding active hexosaminidases using <i>in vitro</i> protein expression	171
5.3.7 Protein sequence analysis of active hexosaminidases	172
5.3.8 F10-ORF19 hexosaminidase activity upon GlcNAc-6-SO ₄ in intact <i>N</i> -glycans	174
5.4 Chapter 5 conclusion	177
6 Conclusion and outlook	180
Bibliography.....	186
List of tables and figures	211

List of abbreviations

2-AB	2-aminobenzamide
3'-SLN-2AB	2-AB labeled 3'-sialyl-N-acetylglucosamine
4-MU	4-methylumbelliferone
4-MU-Gal	4-methylumbelliferyl-D-galactopyranoside
4-MU-GlcNAc	4-methylumbelliferyl-N-acetyl- β -glucosaminide
4-MU-GlcNAc-6-SO ₄	4-methylumbelliferyl-N-acetyl- β -glucosaminide-6-sulfate
4-MU-SO ₄	4-methylumbelliferyl-sulfate
4-MU- α -Neu5Ac	4-methylumbelliferyl- α -D-N-acetylneuraminic acid
4-MU- α -Neu5Gc	4-methylumbelliferyl- α -D-N-glycolylneuraminic acid
6'-SLN-2AB	2-AB labeled 6'-sialyl-N-acetylglucosamine
ADCC	Antibody-dependent cell-mediated cytotoxicity
anSME	Anaerobic sulfatase maturing enzyme
APTS	8-aminopyrene-1,3,6-trisulfonic acid
ATCC	American Type Culture Collection
BAC	Bacterial artificial chromosome
BED	Browser Extensible Data
BHK	Baby hamster kidney
CAZy database	Carbohydrate Active Enzymes database
CCRC	Complex carbohydrate research centre
CE	Capillary electrophoresis
(x)CGE	(multiplexed) Capillary gel electrophoresis
CHO	Chinese hamster ovary
CMAH	CMP Neu5Ac hydroxylase
CMP	Cytidine monophosphate
COSMC	Core-1 β galactosyltransferase specific molecular chaperone
CryoEM	Cryo-electron microscopy
DANA	Deoxy-2,3-dehydro-N-acetylneuraminic acid
DMF	dimethylformamide
DTT	Dithiothreitol
<i>E. coli</i>	<i>Escherichia coli</i>
EDGE	Epitope-directed glycan enrichment
eDNA	Environmental DNA
EMA	European Medicines Agency
EU	Emission unit
FACS	Fluorescence activated cell sorting
FDA	US Food and Drug Administration

FGE	Formylglycine generating enzyme
FGly	Formylglycine
FLR	Fluorescence detector
GAG	Glycosaminoglycan
GBP	Glycan binding protein
GDP	Guanosine diphosphate
GFP	Green fluorescent protein
GH	Glycoside hydrolase
HA	Hemagglutinin
HPAEC-PAD	High-performance anion exchange chromatography-Pulse Amperometric detection
HPLC	High performance liquid chromatography
IPTG	Isopropyl β -D-1-thiogalactopyranoside
LC	Liquid chromatography
LIF	Laser-induced fluorescence detection
LMP (agarose)	Low-melting point (agarose)
MAD	Multiple anomalous dispersion
MALDI-TOF	Matrix-assisted laser desorption ionization-Time of flight
Man-6-P	Mannose-6-phosphate
MIR	Molecular isomorphous replacement
MPI	Max Planck Institute
MR	Molecular replacement
MS	Mass spectrometry
NEB	New England Biolabs
NGS	Next-generation sequencing
NMR	Nuclear magnetic resonance
NuIO	Nonulosonic acids
ORF	Open reading frame
PacBio	Pacific Bioscience
PAPS	3'-phosphoadenosine-5'-phosphosulfate
PBS	Phosphate buffer saline
PC	Phosphorylcholine
PE	Phosphoethanolamine
PGM	Post-glycosylation modification
PUL	Polysaccharide utilization locus
RFU	Relative fluorescence unit
SDS	Sodium dodecyl sulfate

SeMet	Selenomethionine
Sias	Sialic acids
SIGEX	Substrate-induced gene-expression screening
SIGLEC	Sialic acid-binding immunoglobulin-type lectin
SMRT	Single molecule real time
SNFG	Symbol nomenclature for glycans
UDP	Uridine diphosphate
UPLC	Ultra-performance liquid chromatography
WGS	Whole genome sequencing
WT	Wild-type
X-Gal	5-bromo-4-chloro-3-indolyl- β -D-galactopyranoside
X-Neu5Ac	5-bromo-4-chloro-3-indolyl- β -D-N-acetylneuraminic acid

Full name	Abbreviation	SNFG symbol
Fucose	Fuc	
Galactose	Gal	
Glucose	Glc	
Ketodeoxynonulosonic acid	Kdn	
Mannose	Man	
N-acetylgalactosamine	GalNAc	
N-acetylglucosamine	GlcNAc	
N-acetylneuraminic acid	Neu5Ac	
N-glycolylneuraminic acid	Neu5Gc	

1 Introduction

1. Introduction

Enzymes are biocatalysts that facilitate numerous chemical and biological reactions. Valuable to many industries, their use enables formation of compounds with commercial value (*e.g.*, proteases, lipases, amylases and cellulases for the production of detergents [1]). Enzymes also form the basis of many medical tests (*e.g.*, polymerases used in quantitative polymerase chain reaction ‘qPCR’ tests for pathogens detection [2]) and are essential to disease diagnosis. In addition, scientific research in the fields of biology and biochemistry would not exist as it does today if it was not empowered by enzymes, essential reagents in numerous experiments (*e.g.*, restriction enzymes for cloning of DNA [3]). Consequently, the desire for new enzymes is constantly growing and is motivated by both private and public sectors of industry (*e.g.*, food, detergent and pharmaceutical) and by academic research.

Microorganisms are ubiquitous on earth and their diversity is tremendous. Species have adapted to survive in any ecosystem. Consequently, microbial genomes represent a gold mine for discovery of new enzymes and biocatalysts. Cultivation of microorganisms in laboratories dates to the second half of the 19th century with the pioneering work of Louis Pasteur and Robert Koch [4]. Although many efforts have been made towards improving the culturing of microbes, today, only a small percentage of microorganisms can be cultivated in the laboratory. The concept of metagenomics emerged at the end of the 20th century with the intention of gaining access to microorganisms previously out of reach due to their inability to be cultivated. An introduction to metagenomics and its benefits in enzyme discovery programs is provided in Chapter 2, section 2.1.

While almost any live sciences field can benefit from the discovery of new enzymes, this thesis is entirely centered around glycobiology. This field encompasses the study of glycans and glycoconjugates, and their role in biological processes. The biological functions of glycans are as diverse as their structures. Glycans are involved in both health (*e.g.*, lymphocyte rolling, lysosomal protein trafficking or pituitary hormones clearance [5–7]) and disease (*e.g.*, viral infections, tumor malignancy or systemic lupus erythematosus [8–

10]). An introduction to glycobiology is provided in Chapter 2, section 2.2. The many roles of glycans in the biology of vertebrates and invertebrates is far from being fully understood. Basic research in the field is still critical. In addition, because many bioproducts are glycosylated and glycosylation affects drug efficacy, protein glycosylation is monitored as a critical quality attribute in pharmaceutical manufacturing. Several analytical workflows were developed over the past two decades to facilitate the characterization of glycans and glycoconjugates. Glycoside hydrolases, a class of enzymes that act upon glycans, are fundamental to many of these workflows. The current glycoanalytical toolbox contains many well-established glycoside hydrolases that help to ascertain glycan structure. However, several enzyme specificities are still needed to improve analysis of the tremendous structural diversity of glycans.

This thesis is the result of a scientific collaboration between New England Biolabs (NEB) and the Max Planck Institute (MPI) for Dynamics of Complex Technical Systems. NEB is a company that specializes in the production of enzyme reagents for the Life Sciences industry and is committed to supporting basic research in various areas, including glycobiology. The Bio/Process Analytics team at MPI specializes in quantitative and qualitative analysis of glycosylation.

The first goal of my thesis work was to adapt functional metagenomics methodology and develop a workflow to enable discovery of a variety of novel enzymes that act upon glycans. To this end, I first established processes to extract microbial genomic DNA from diverse environments. From this material I constructed large-insert metagenomic DNA libraries that enabled micro-expression of genes in *Escherichia coli* (*E. coli*). I established high-throughput screening assays for enzymes that acted upon glycoconjugates in various ways, and devised methods for next-generation sequencing and *de novo* data assembly of identified clones. Finally, I adopted methods for identification of gene candidates. Development of the workflow is described in more detail in Chapter 3.

1. Introduction

To make use of the devised workflow, the second goal of this thesis was to construct an extensive metagenomic DNA library collection. In total, almost 100,000 clones were created from 20 metagenomic and genomic DNA sources. Archived in 259 x 384-well plates, the collection is estimated to theoretically contain ~3-4 million microbial genes. The collection represents a rich hunting ground for enzyme screening and is presented in more detail in Chapter 3, section 3.3.1.2.

The third goal of this thesis was to use the established screening workflow and library collection to find new enzymes that address either fundamental or applied questions in the field of glycobiology. In my first screening project, I was investigating the hydrolysis of sialic acids, a family of structurally diverse monosaccharides that decorate most classes of glycans and that are involved in many biological and pathological processes. Two projects stemmed from the area of sialic acid biology and are described in Chapter 4. In a first project, I looked for the presence of sialidases in an extreme hot spring environment. This work identified a novel exosialidase whose sequence was unlike any known glycoside hydrolases. This enzyme proved to be the founding member of a new family of glycoside hydrolases termed GH156 in the Carbohydrate Active Enzymes (CAZy) database. This discovery is described in Chapter 4, section 4.2. Motivated by its unique protein sequence, the enzyme structure and mechanism of action were determined in collaboration with a team of structural biologists from the University of York. This characterization is presented in Chapter 4, section 4.3. Extending my work on sialidases, I then sought enzymes able to hydrolyze the non-human sialic acid N-glycolylneuraminic acid (Neu5Gc). I identified two sialidases with a strong preference for Neu5Gc over its human counterpart N-acetylneuraminic (Neu5Ac). These unprecedented specificities have a potential therapeutic application in preventing or treating Neu5Gc-induced colitis. Identification of these enzymes is presented in Chapter 4, section 4.4. Finally, I applied the enzyme discovery workflow to a glycoanalytics-oriented question. In this third screening project, I addressed the technical challenge of detection and characterization of sulfated *N*-glycans for

which there is a lack of analytical tools. I conducted a screen and identified a sulfatase and a sulfate-specific hexosaminidase that specifically act on sulfated *N*-glycans. I proved that both enzymes represent useful new additions to the glycoanalytical toolbox. This work is presented in Chapter 5. A schematic diagram of all enzymatic activities described in this thesis work is shown in Chapter 2, Figure 4.

2 Theoretical background

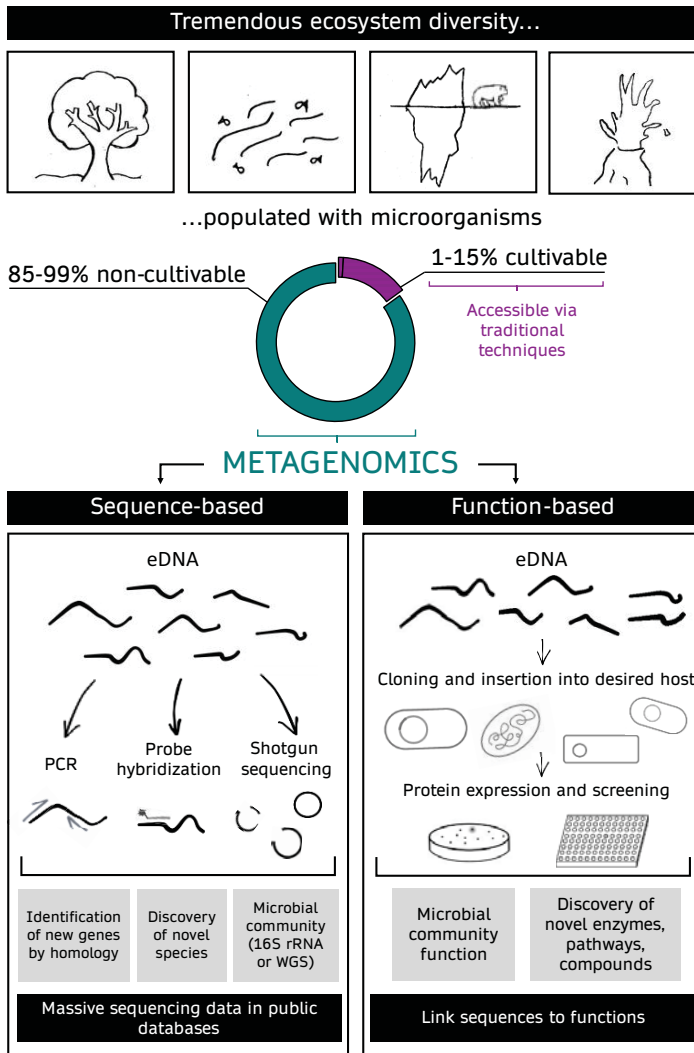


Figure 1. Sequence- and function-based metagenomics. Metagenomics provides access to genes derived from both cultivable and uncultivable species. Abbreviations: environmental DNA (eDNA), whole genome sequencing (WGS).

2.1 Metagenomics

2.1.1 Microorganisms: a tremendous diversity previously inaccessible

Microorganisms have been found in vast environments, ranging from hospitable ecosystems to habitats thought to be too extreme to support life. Some microbes live in deep sea hydrothermal vents [11] or desert soils [12]. Yet, since their appearance on earth billions of years ago, microbes have evolved to populate vast ecological niches. In fact, it is now assumed that no environment on earth is devoid of microorganisms. This long evolution and their ubiquity have resulted in the globe's bacterial species diversity being estimated in excess of 1 trillion (10^{12}) species [13].

Microorganisms have thrived in every ecosystem on earth through their ability to adapt to their environment. They encode pathways and produce biocatalysts specific for their survival in their surroundings. For example, bacteria that populate our guts encode numerous carbohydrate degrading enzymes to feed on these biomolecules naturally present in this habitat [14]. In contrast, archaea living in deep sea thermal vents encode redundant DNA repair pathways to maintain their genomes intact in physically and chemically harsh conditions [15]. Thus, the vast number of ecosystems on our planet are occupied by specifically adapted microorganisms, each encoding unique pathways and interesting enzymes. The pool of earth's microbes represents a goldmine for scientists interested in discovering novel biocatalysts or metabolic pathways.

For many years, microbiology depended upon being able to cultivate a microorganism. Microbes were grown using nutrient-containing media to facilitate their study in the laboratory. However, even in the early years of the field, evidence was emerging that not all microbes could be grown in the lab. In 1932, Razumov reported for the first time what would be later termed by Staley and Konopka "the great plate count anomaly" [16]. Razumov counted bacteria by two different methods: direct microscopy and plating techniques using Koch culture medium. He observed similar results with the two methods when counting bacteria from cultured milk but obtained noticeably different counts

from aquatic habitats. Bacteria from milk were easy to cultivate, and thus, were growing efficiently on Koch medium. However, many microbes from the environment were not able to grow resulting in a much lower count from the cultivated plate [17]. Since then, many improvements have been made to culture media to improve the number of strains that can be grown in the laboratory. Yet, it is still estimated that only 1-15% of bacteria are cultivable [18, 19] and cultivating an entire microbial community remains out of reach. It is to overcome this obstacle that the field of metagenomics was born.

The term metagenomics first appeared in a publication from Handelsman's group in 1998 [20]. Initially referring to what is now called functional metagenomics, the authors described the extraction of genomic DNA from uncultivable soil microorganisms, the cloning of this DNA into *E. coli*: a cultivable host and finally the screening of the obtained clones for production of chemical compounds. This innovative approach was pioneered by Healy *et al* in 1995 [21]. In that work, cellulases from an anaerobic digester microbial community were identified using a similar methodology. Later, the meaning of metagenomics was extended by the efforts of two groups: Tyson *et al* and Venter *et al* [22, 23]. Both used shotgun sequencing to reconstruct genomes from communities of uncultivable microorganisms by assembling sequenced genome fragments originating from natural habitats: an acid mine drainage biofilm and the Sargasso Sea. This approach is now termed sequence-based metagenomics and encompasses the direct analysis of DNA isolated from the environment.

The term metagenomics now defines both sequence- and function-based analysis of collective genomes from an uncultured microbial community (Figure 1). Bypassing growth of microorganisms in the laboratory is the essence of the field and it is where all metagenomics' interests lie, with the aim of ultimately accessing the genetic information encoded by uncultivable microorganisms.

2. Theoretical background

2.1.2 Sequence-based metagenomics

For many years, sequence-based metagenomics was technically limited to Sanger sequencing. Established by Frederick Sanger and his co-workers in 1977, Sanger sequencing enables deconvolution of a single DNA fragment per reaction tube. Extremely labor- and time-consuming in its beginnings, it was automated 10 years after its establishment [24] and is still widely used today for certain applications such as the verification of cloning constructs. Early sequence-based metagenomics studies used Sanger technology for whole genome shotgun sequencing or targeted region sequencing. In shotgun sequencing approaches, limited read length (600-1000 bp) greatly challenged data assembly as illustrated by the work from Venter *et al*, in which they could only assemble a quarter of the sequences generated [23]. In targeted region sequencing such as 16S sequencing, reads do not have to be assembled. 16S sequencing exploits hypervariable regions of the 16S ribosomal RNA (16S rRNA) gene that can serve as a phylogenetic marker. In this approach, sections of a gene are first isolated from the cDNA pool and amplified by PCR with primers designed to anneal to highly conserved regions of the gene while hypervariable regions of the amplicons are sequenced and used for taxonomic assignment [25]. Both whole genome shotgun and 16S rRNA Sanger sequencing were used to provide a picture of the microbial community living in the sampled environment [22, 23, 26, 27]. But the existence of closely related species, horizontal gene transfer and the use of non-exhaustive databases greatly challenge taxonomic assignment. In addition to phylogenetic diversity assessment, Sanger sequencing was used for sequence-based gene discovery. Genes were identified from shotgun sequencing data, using PCR amplification or probe hybridization techniques. These methods enable isolation of genes of interest based on their sequence homology to a targeted gene family. Identified genes are variants of existing gene families and always have homology to known genes. New chitinases and hydrogenases, among others, have been identified this way [28, 29]. Whether Sanger sequencing is used to assess microbial diversity or to identify new genes, its monetary cost, short read length,

and low sequencing depth have been impediments for its application in sequence-based metagenomics.

Over the past two decades, many so-called ‘next generation sequencing’ (NGS) technologies have emerged. The initiation of the international Human Genome Project in 1990, an effort to sequence the first human genome, started the race to develop better, cheaper, and faster sequencing technologies. Today, three major NGS platforms are commonly employed in sequence-based metagenomic studies: Illumina, Pacific Bioscience (PacBio) and Nanopore sequencing technologies.

Developed in 2011, Illumina sequencing belongs to the second generation of NGS. Still widely used today, it is one of the preferred sequencing technologies on the market, mainly because of the number of reads it can generate and its high sequence accuracy. Generating short reads, typically 150-300 bp, the technology has proven extremely useful in read mapping analyses and for *de novo* assembly of small genomes [30]. While it has been used in some metagenomics studies [31, 32], *de novo* sequence assembly is challenged by the technology’s short read length, in particular for complex samples.

To resolve this limitation, third generation sequencing technologies based on long-reads have emerged (*e.g.*, the PacBio platform). In comparison to Illumina sequencing, PacBio sequencing decodes DNA in real time with no stop between each nucleotide incorporation step. Additionally, no DNA amplification is required during sample preparation. The signal from a single DNA molecule is monitored, hence its name ‘Single Molecule Real Time’ (SMRT) sequencing. DNA templates loaded onto the instrument are circular. During sequencing, the polymerase replicates the template. After the circular template has been replicated (and therefore read one time, termed one ‘pass’), the polymerase will continue to migrate around the circular DNA template. The error rate of a single ‘pass’ can reach 15%, however, these errors are randomly distributed and can easily be resolved by comparing several polymerase passes [33]. Hence, after data processing, reads have an extremely high accuracy of

2. Theoretical background

over 99% for a template read 15-20 times [34, 35]. The technology generates an average read length of 10-15 kb, but reads up to 30-50 kb are commonly observed [35]. This feature facilitates *de novo* assembly of large genomes or complex samples, making PacBio sequencing particularly well-suited for metagenomic data acquisition [36]. It is the NGS approach used throughout the course of this thesis.

Nanopore sequencing is considered by some to be a fourth generation of NGS sequencing technologies. In contrast to PacBio sequencing and most of the earlier technologies, it does not determine the sequence of a DNA molecule through DNA synthesis. In nanopore sequencing, the DNA molecule is first translocated via a motor protein through a pore. The use of nanopores to sequence DNA was proposed decades ago [37] but faced multiple technical challenges such that sequencing by DNA synthesis prevailed. In 2014, Oxford nanopore released the MinION, a system comprised of nanopores embedded into a membrane with an applied electrical current based sensing method. When the motor protein passes DNA through a nanopore one base at a time, the current is disrupted. The extent of current change is different for each of the 4 bases, and can then be translated into DNA reads [38]. This method does not require the sample to be amplified and reads single molecules. Additionally, in amplification-based platforms (*e.g.*, Illumina) the shortness of read-length reflects limitations in polymerase processivity. The absence of amplification in nanopore sequencing allows the technology to generate very long reads, with read lengths longer than 2,000 kb having been reported [39]. This impressive feat highlights the potential of this emerging technology to be used to sequence small prokaryotic genomes without requiring assembly. However, nanopore sequencing still has many technical challenges with respect to high molecular weight DNA. One major flaw of nanopore sequencing remains its very high error rate (up to 40% in some cases) that renders its use for *de novo* assembly cumbersome [40]. However, its low cost and extremely small footprint (90 g for a $3.3 \times 10.5 \times 2.3$ cm instrument) confer the technology some undeniable advantages, particularly in the field of metagenomics [41]. Devices can be

transported into the field and sequencing performed at a sampling site, as demonstrated by Johnson *et al*, who conducted field sequencing in Antarctica [42]. A MinION device was even transported to the international space station where sequencing of control samples was conducted [43]. It is almost certain that the technology will improve in the coming years, and an increasing number of metagenomes will be sequenced using this approach.

Development of NGS technologies has allowed longer genomic fragments to be sequenced. In combination with significant improvements in sequence assembly software, steady improvements to our capacity to decipher genomes have been realized. Reconstruction of entire draft genomes from extracted crude eDNA has allowed discovery of novel species. For example, Nayfach *et al* assembled metagenome data from various ecological niches and obtained over 50,000 draft genomes, among which, ~25% were new species [44]. As sequencing costs have decreased and technologies improved, the amount of metagenomic sequencing data has increased dramatically. There is now an accumulation of sequences from previously inaccessible microorganisms in public databases. While this abundance of sequence data represents an extraordinary resource, its functional analysis remains challenging [45, 46]. There is a notable and growing need for widespread functional characterization of the genes encoded in this data.

2.1.3 Functional metagenomics

Functional metagenomics can help connect metagenomic sequences to gene function. In this subfield, extracted eDNA is not directly subjected to DNA sequencing but is instead cloned into an appropriate vector for introduction into a surrogate host creating a ‘metagenomic library’. These libraries are then interrogated for specific enzyme activities using a high-throughput screening approach. The goal is not to gain a comprehensive assignment of function to all open reading frames encoded within the DNA library, but rather to screen for a target enzyme activity of interest encoded within the library population.

2. Theoretical background

2.1.3.1 Choosing a vector-host system

Different types of vectors can be used to construct metagenomic libraries of eDNA: plasmids, cosmids/fosmids or bacterial artificial chromosomes (BACs). Plasmids are used to construct small eDNA insert (2-10 kb) libraries. This type of vector is adequate when one seeks to find an activity encoded by a single or a few small genes. They are easy to generate, and many tools have been specifically designed to aid in their construction (*e.g.*, miniprep kits for plasmid isolation). Certain plasmids are present in high copy number within the host cell and may contain an inducible promoter, two properties that can drastically enhance expression of cloned foreign genes. To further improve gene expression, Lämmle *et al.*, used a dual-orientation expression plasmid that contains two promoters, one on either side of the cloning site such that inserted genes can be expressed in an orientation-independent manner [47]. However, due to the small insert size with plasmids, this approach is prone to cloning partial or truncated genes, that can result in a nonfunctional product. To reduce this limitation, large insert libraries can be created using cosmid/fosmid vectors (25-40 kb inserts) or BACs (up to 200 kb inserts). Such libraries encode large gene clusters and can express entire metabolic pathways. They are less trivial to produce, rely on intrinsic promoters for gene expression and are usually less stable in the host. Interestingly, while large insert vectors theoretically limit the chances of proper gene expression, the hit rate obtained when screening these libraries is no lower than that of screening small insert plasmid libraries [48]. Additionally, as large insert clones encode many more genes than plasmid clones, far fewer clones must be screened. Hence, the use of cosmids/fosmids evolved as a good compromise in terms of size, number of genes that can be cloned and efficiency of protein expression, explaining their popularity and predominance in the field.

Once eDNA has been cloned into the chosen vector, the resulting clones are inserted into a cultivable host for gene expression and subsequent screening. Most metagenomics libraries have been created in *E. coli*. It is easy to cultivate in the laboratory and has a large array of vector systems and tools available. But

gene expression is largely host-dependent and *E. coli* transcription and translation machinery cannot express all genes present on eDNA derived from every microbe. To overcome this limitation, several groups have constructed libraries in other hosts such as *Streptomyces lividans* [49], *Ralstonia mutallidurans* [50] and *Bacillus* [51]. This strategy has proven successful as illustrated by Biver *et al*, who discovered an antibiotic when expressing a metagenomic library in *Bacillus*, but could not find it when using *E. coli* as the host [51]. However, use of non-*E. coli* hosts can be challenging. It requires the use of a vector that can replicate autonomously in the chosen host or the use of shuttle vectors. Shuttle vectors contain integration sites enabling the clone to integrate into the host genomic DNA; a strategy employed by Courtois *et al*, who expressed a soil metagenomic library in *Streptomyces*. As no host will effectively express all genes from a microbial community, the use of several hosts for expressing the same metagenomic library has been proposed. Such an approach requires more clones to be screened, which can be time-consuming and cost-prohibitive, and the use of a broad-host range vector system. Such a system was developed by Cheng *et al* and consists of a Gateway LR cloning vector that can recombinantly insert the cloned eDNA into a second destination vector [52]. By changing the destination vector, one can rapidly build metagenomic libraries for multiple hosts using the same isolated eDNA.

Numerous successful screens have been conducted using *E. coli* as an expression host. In a remarkable study, Gabor *et al*, estimated the number of genes that can be heterologously expressed in *E. coli* and that are thus functionally accessible for screening [53]. The ability to find an expressed target gene within a metagenomic library depends on three main factors: i) its abundance in the library, ii) its length and iii) its mode of expression. This study defined modes of expression based on *cis* and *trans* elements. *Trans* elements (*e.g.*, chaperones or co-factors) are hard to predict bio-informatically and were thus not addressed in their work. However, *cis* elements (*e.g.*, promoters, ribosome binding sites (rbs), etc) can easily be predicted from DNA sequence. The authors analyzed 32 prokaryotic genomes and estimated the number of

2. Theoretical background

genes that were preceded by *cis* elements recognized by *E. coli*. Expression of such genes is vector-independent and is driven by intrinsic promoters and rbs sites located within the cloned eDNA insert. When expressed in *E. coli*, these genes represented 40% of those analyzed. Important expression variation was predicted based on the taxonomic group of genomes examined. Vector-dependent genes that were predicted to require either a promoter, a rbs, or both to be present in the vector to be expressed in *E. coli* were also observed. Finally, genes lacking an *E. coli*-like rbs and promoter sites were considered inaccessible and represented 30% of examined genes. Such *in silico* estimations are encouraging and support the notion that *E. coli* constitutes an attractive host for eDNA expression. *E. coli* was therefore chosen as the exclusive host for the work presented in this thesis.

2.1.3.2 Screening strategies

When stored properly, libraries are very stable and can be screened on multiple occasions using various assays, making them a particularly valuable resource. Different screening strategies can be employed depending on the product being sought. The fastest and easiest screening method uses growth of library clones on agar plates. Clones are plated on a solid medium and screened for a growth phenotype. For example, Jeon *et al* used tricapylin-containing plates, on which clones carrying an active lipase formed a clear halo [54]. The use of selective medium that restricts growth to clones harboring an activity of interest can also be utilized. This approach was applied by Forsberg *et al* to identify genes conferring tolerance to inhibitory compounds such as ferulic acid or furfuryl alcohol [55]. Another popular method involves adding a reporter substrate directly to the plate. This strategy is principally used when screening for an activity that hydrolyzes a substrate and liberates a fluorophore or a chromophore generating a measurable signal [56]. The use of an engineered reporter strain has also been fruitful. In these gain-of-function screens, clones are used to transform a genetically modified host that cannot grow unless cloned eDNA carries a gene or pathway that rescues host viability. For example, tryptophan auxotrophic strains were used to isolate clones containing the

tryptophan biosynthesis pathway [57]. These plate-based screening methods are advantageous because hundreds of clones can be screened at one time on a single agar plate. However, it is not always feasible to design an assay that functions well in a plate for every target activity. In addition, several days are usually necessary for the phenotype to develop, and the output signal is qualitative lending some subjectivity to identification of a positive clone.

To overcome these limitations, screening can be carried out in microtiter plates (*e.g.*, 96 or 384- well plates). The bacterial clones of a library are each grown as individual liquid micro-scale monocultures in multi-well plates. These cultures can be subsequently screened with a desired assay [58]. Assays often resemble any *in vitro* assay that is ordinarily performed with material from larger flask cultures. Theoretically, assay design has unlimited possibilities provided the assay can be made to work in an *E. coli* lysate. In this thesis, carbohydrate hydrolytic activities were screened for using fluorogenic substrates. As such, a microplate reader was used to rapidly and quantitatively measure the fluorescence generated by hydrolysis of the substrate. This screening format can further benefit from the use of liquid handlers or small robotic systems to achieve good reproducibility and high-throughput.

Another approach to metagenomic library screening involves substrate-induced gene-expression screening (SIGEX) [59]. In this innovative method, screening relies on the notion that catabolic gene expression is often induced by a gene's product or a metabolite that acts on regulatory elements located adjacent to the catabolic gene. Operon-trap vectors containing a green fluorescent protein (GFP) reporter gene are used. Hits responsive to the targeted substrate, and thus expressing catabolic genes, display a green fluorescence signal and can be sorted by fluorescence activated cell sorting (FACS). The method was demonstrated by screening a groundwater metagenomic library upon induction with benzoate and naphthalene. This permitted identification of both benzoate and naphthalene catabolic genes [59].

2. Theoretical background

2.1.3.3 Application of functional metagenomics

One of the earliest functional metagenomics studies was performed by Rondon *et al* in 2000 and nicely illustrated the breadth of functions one can explore within a metagenomics library. In this work, a soil BAC library was screened for various hydrolase activities like cellulase, lipase or esterase, and also for siderophore and antibacterial compounds [60]. Applications of functional metagenomics are endless: antimicrobial agents, chemical compounds, metabolic pathways, and enzymes can be identified using this approach. Novel enzyme discovery represents the top motivation for the use of functional metagenomics. In 2019, global industrial enzyme sales represented \$9.9 billion (USD), up from \$2.3 billion in 2003, highlighting a rapidly growing market [61]. Metagenomics with its ability to discover novel enzymes is full of promise for the food, pharmaceutical and detergent industries. Enzymes can significantly improve industrial production by substituting for expensive or inefficient chemical processes. One example is the replacement of chemical hydrolysis of starch in the food industry by α -amylases. Belonging to glycoside hydrolase (GH) family 13, such enzymes are usually active at mild temperature, neutral pH, and in the presence of Ca^{2+} ; characteristics incompatible with industrial processes. Richardson *et al* conducted a functional metagenomics screen and identified several new GH13 members, among which, some were compatible with industrial starch hydrolysis [62]. This study also highlighted that it is advantageous to screen with a metagenomic library made with eDNA isolated from an environment having similar conditions to the industrial process (*e.g.*, high temperature, low temperature, high salt, etc). Often such screens identify enzymes with biochemical properties mirroring the ecological environment from which they derive. For example, screening of a deep sea sediment metagenomics library yielded a cold active lipase that retained 80% of its activity at 10°C [54]. Similarly, several extreme condition-adapted enzymes have been isolated from metagenomic libraries from various extreme environments (*e.g.*, cold and alkaline, high temperature) [63, 64].

Functional metagenomics is also of interest to the pharmaceutical industry. Many antimicrobial agents have been identified from functional screens. Iqbal *et al.* found six clones with antimicrobial activity in a soil metagenomic library expressed in *Ralstonia metallidurans* [65]. Functional metagenomics can also be used to identify genes conferring antibiotic resistance. Forsberg *et al.*, identified in various soil metagenomic libraries, 110 genes yielding resistance to different types of antibiotics [55]. More recently, a compelling functional metagenomic study identified a pair of enzymes, a deacetylase and a galactosaminidase, that when used in concert were capable of converting blood group A to universal blood type O. This mixture of enzymes was more efficient than known enzymes capable of performing the same conversion, raising hopes for large scale application in organ transplantation [66].

Functional metagenomic screening can also significantly impact our understanding of biology. For example, it was employed by Chauhan *et al* to gain insights into arsenic detoxification by marine microorganisms on ocean floors [67]. They identified several genes responsible for arsenic degradation (*e.g.*, arsenate reductase or arsenite efflux pump) by screening a sea sediment metagenomic plasmid library. Similarly, insights into the assortment of enzymes required for a given biological process can be gained. The diversity of cow rumen enzymes involved in plant-polymer breakdown was evaluated by screening a metagenomic library for diverse hydrolase activities (*e.g.*, esterases, cellulases, and amylases) [68]. Functional metagenomics can also improve our understanding of symbiotic relationships. Verma *et al* applied functional metagenomics to identify salt stress tolerance-conferring genes in microbes colonizing the human gastrointestinal tract [69]. Essential for successful colonization of their host, these genes provide microbes the capacity to resist significant fluctuations in osmolarity observed in the human intestine.

Only few examples of functional metagenomics applications were highlighted here but several reviews thoroughly summarize the enzymes and compounds identified by the field [49, 70–72].

2. Theoretical background

2.1.4 Contrasting approaches to enzyme discovery

In addition to functional metagenomics, *in silico* bioprospecting and protein engineering are also used to discover or create new enzyme activities. With *in silico* bioprospecting, new enzymes are identified by exploring sequence databases. It is the easiest, fastest, and cheapest strategy for new protein discovery. Different types of database searches can be conducted using keywords (*e.g.*, an enzyme name), a specific enzyme sequence, a generated consensus sequence or conserved motifs [73]. When using a specific enzyme sequence as a query to probe a sequence database (*e.g.*, GeneBank, SwissProt, etc), obtained candidates must be carefully chosen to ensure they constitute new candidates with the potential for new properties and are not merely highly conserved homologs with similar properties to those of the query. To do so, selected candidates are commonly chosen from sequences below 80% identity to the query sequence [74].

If a large database is prospected, narrowing the search space to smaller databases will help in reducing the number of candidates. The database can also be constrained to help identify candidate proteins with a desired trait. For example, searches limited to thermophilic organisms can be performed to identify potential thermophilic enzyme candidates. Specific niches can also be surveyed by exploiting sequencing data from their extracted metagenomes as performed by Toyama *et al*, who identified a new β -glucosidase from metagenomic data from Lake Porqu e in Brasil [75]. While being a straightforward and fast approach, *in silico* searches do not usually result in discovery of new protein families. Instead, identified candidates are typically variations of known enzymes or known protein folds.

In contrast to *in silico* sequence database mining, protein engineering uses function-based screening strategies to design new enzymes. In this approach, a known protein is subjected to random or site-directed mutagenesis, and resulting mutants are screened for new biochemical traits. For example, it is possible to narrow or broaden an enzyme's substrate specificity or to improve

its activity under certain physicochemical conditions (*e.g.*, pH, temperature, etc). Three main strategies are usually used to engineer proteins: directed evolution, rational design, and semi-rational design. In directed evolution, an iterative two-step process is applied where i) a library of mutants is randomly created (using chemicals, error-prone PCR or DNA recombination techniques) and ii) the library is screened for characteristics of interest [76]. When applied to the protein subtilisin E, this approach enabled creation of a mutant with a 256 times higher activity in presence of 60% dimethylformamide in three rounds of mutation [77]. In this strategy, no prior knowledge about the protein's structure is required as mutations are randomly generated throughout the molecule. As such, many non-productive mutants are also generated that contain mutations in essential domains resulting in improper protein folding and/or enzyme inactivation. As such, numerous mutants must be screened to identify those with proper folding, retained activity, and the desired new trait. In rational design, detailed knowledge of a protein's structure is exploited. In this approach, mutations are not randomly introduced. Instead, specific residues are mutated after *in silico* modelling to potentially change a desired characteristic of the enzyme [78–80]. Semi-rational approaches combine both directed evolution and rational methods. In this strategy, knowledge of the enzyme's structure is used to identify structural or functional domains critical for its activity. These are then protected and preserved from mutation, while random mutagenesis is applied to rest of the protein [76, 81–83]. This hybrid approach enables generation of a higher proportion of mutants with proper folding and activity.

Despite being function-driven, protein engineering uses a known enzyme as a starting point. In comparison, functional metagenomics is the only approach to gene discovery that conducts fully naïve searches for which no known sequence or gene template is used. As such, any activity, including speculative activities not yet known in biology, can theoretically be screened for, provided an assay to detect its function can be designed.

2. Theoretical background

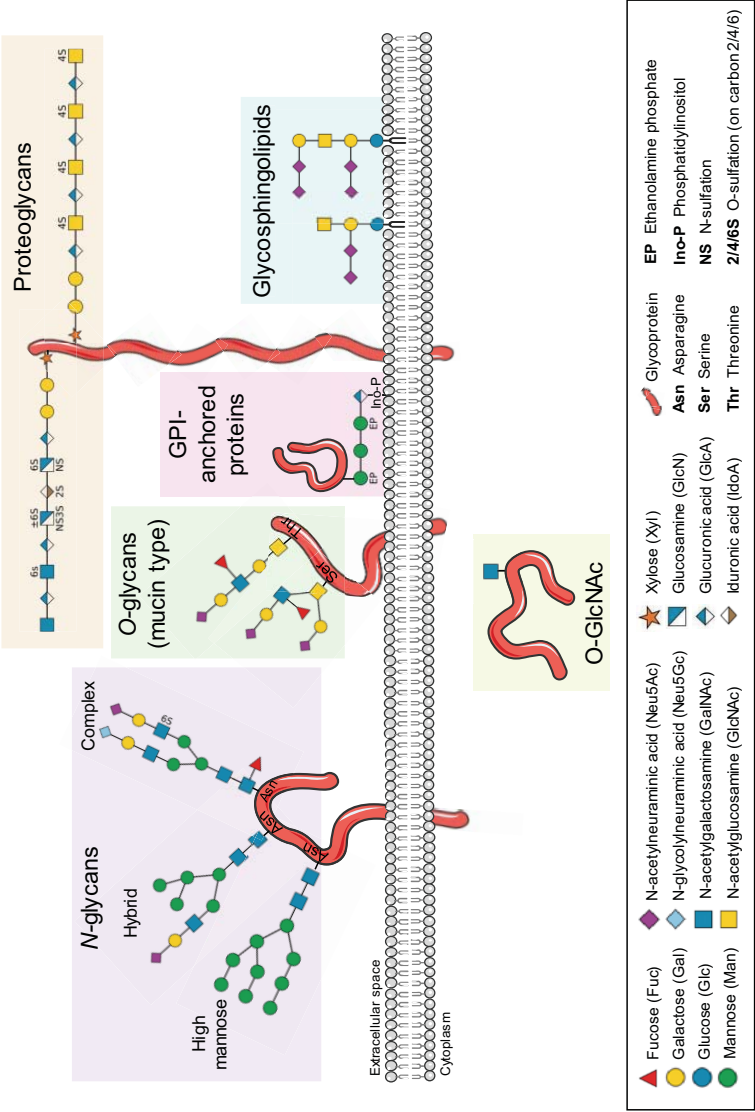


Figure 2. Diversity of glycoconjugates in mammalian cells. Inspired from [357]. Glycans are represented using the SNFG representation [235].

2.2 Glycobiology

2.2.1 A few definitions

Glycobiology is a branch of the biological sciences field that defines the role of carbohydrates ('glycans') in biological processes. The field embraces investigation of the structure, synthesis, and biological function of glycans. At the interface of chemistry and biology, this area of science is relatively new with the first appearance of the term 'glycobiology' surfacing in the 1980's [84]. Nonetheless, many crucial discoveries were made long before this time. One important example is Karl Landsteiner's discovery of human blood groups (carbohydrate ABO antigens) at the beginning of the 20th century [85]. Glycans are linear or branched chains of monosaccharides that can be appended to various types of macromolecules (also termed 'aglycones') together forming glycoconjugates. Based on the nature of the aglycon, glycoconjugates are divided into distinct classes among which: *N*- and *O*- glycoproteins, GPI-anchored glycoproteins, glycosphingolipids and proteoglycans (Figure 2) [84]. Of these, *N*- and *O*-glycoproteins are probably the most studied types of glycoconjugates. In *N*-glycoproteins, glycans are attached to proteins via the amide group of the side chain of an asparagine residue contained within the consensus sequence Asn-X-Ser/Thr (where X is any amino acid but proline). In *O*-glycoproteins, glycans are appended to proteins via the hydroxyl group of the side chain of a serine or threonine residue. Glycosylation represents the most common post-translational modification of proteins, affecting their solubility, folding, recognition and activity, therefore playing crucial biological roles. It is a remarkably dynamic process with glycan structure being subject to modification based on physiological changes such as pregnancy or aging, but also upon disease development [86–88].

2.2.2 Glycan synthesis

Biology is centered around the template-driven dogma that genetic information encoded by DNA is transcribed into RNA that will in turn be translated into protein. Glycans, whose synthesis is not template-driven, are

2. Theoretical background

excluded from this dogma, and their study has long lagged behind that of genomics and proteomics. In eukaryotic protein glycosylation, glycan synthesis takes place in the endoplasmic reticulum and the Golgi apparatus. It starts with the linking of monosaccharides that have been synthesized, recycled from degraded glycans, or scavenged from the environment [89]. Monosaccharides are first activated creating high energy sugar donors. The majority of monosaccharides are activated by coupling with uridine diphosphate (UDP) or guanosine diphosphate (GDP). Interestingly, sugar donors from the sialic acid family are activated as cytidine monophosphate (CMP)-mononucleotides [89]. Synthesis reactions encompass the activation of free sugars using kinases, or the interconversion of activated monosaccharides to another activated monosaccharide by sugar epimerization or nucleotide exchange. Synthesis pathways yielding sugar donors are highly interconnected and multiple pathways can be utilized to form the same precursor. Notably, all precursors can be derived from glucose, the principal carbon energy source for most organisms. In eukaryotes, synthesis of precursors mostly takes place in the cytosol. Sugar donors are then transported to the endoplasmic reticulum and the Golgi. Different modes of transport have been elucidated, both energy- and non-energy-dependent [89]. In the endoplasmid reticulum and the Golgi, activated monosaccharides are transferred to glycan or macromolecule acceptors by glycosyltransferases whose expression depends not only on the tissue- and cell-type, but also on the physiological state of the cell [90]. Regulation of glycosyltransferase expression and the availability of sugar donors at a certain time and place control the formation of a vast and structurally diverse array of glycan molecules.

2.2.3 Tremendous glycan diversity

The diversity of glycoconjugates is enormous and can be attributed to different factors that cause structural heterogeneity: i) the nature of the aglycone, ii) the variety of monosaccharides and degree of branching, iii) the different types of linkages that connect monosaccharides and iv) the presence of sugar modifications (Figure 2).

Glycans are formed by assembly of monosaccharides. Ten common monosaccharides are found in vertebrates: glucose (Glc), galactose (Gal), mannose (Man), fucose (Fuc), xylose (Xyl), glucuronic acid (GlcA), N-acetylglucosamine (GlcNAc), N-acetylgalactosamine (GalNAc), N-acetylneuraminic acid (Neu5Ac) and N-glycolylneuraminic acid (Neu5Gc) [91]. Additional monosaccharides exist in certain plants, bacteria, and parasites. The number of building blocks used to form glycans is greater than the four units used to make nucleic acids and is less than the 20 different amino acids that comprise proteins.

The vast breadth of glycan diversity does not lie solely in the number of building blocks used, but also in the variety of ways sugars are linked to each other. Proteins and nucleic acids are formed from amino acids and nucleotides invariably linked to one another by a single type of peptide or phosphodiester bond, respectively. In contrast, glycans building blocks are linked to each other by a range of glycosidic bonds that involve both stereochemistry of the sugar (α or β anomer) and the position of attachment to the adjacent sugar. In theory, glycosidic bonds could be formed between the anomeric carbon of a monosaccharide and any free hydroxyl group from another monosaccharide [92]. However, not all linkage possibilities have been found in nature.

Another layer of structural complexity involves modification of some glycans with chemical groups (*e.g.*, methyl, sulfate, acetyl, or phosphate moieties). Termed ‘post-glycosylation modifications’ (PGMs), these chemical moieties can be of great functional significance [93]. A notable example is mannose phosphorylation of *N*-glycans, a modification that directs trafficking of glycoproteins to lysosomes [94]. Another example is the likely roles of sulfated *N*-glycans in influenza virus replication and virulence [95, 96]. The importance of this area of glycobiology is still emerging and has been hampered by poor tools and methods to understand many PGMs. In Chapter 5 of this thesis, a screening project is presented that identified new enzymes that improves analysis of *N*-glycans bearing a sulfate PGM.

2. Theoretical background

In nature, proteins exist as a collection of glycoforms. Protein glycoforms consist of a protein backbone with different glycosylation site occupancy (termed ‘macroheterogeneity’) or glycan structures at a given site (termed ‘microheterogeneity’) [97, 98]. As an example of microheterogeneity, 6 glycoforms were identified at the single glycosite of ribonuclease B [99]. Consistent with the existence of microheterogeneities, the entire glycan repertoire of a cell (termed ‘glycome’) is vastly more complex than the genome or the proteome.

2.2.4 Glycan functions

Glycans were shaped by millions of years of evolution. As a result, they are an essential part of the cell and play many important roles within an organism [100]. Their functions are as diverse as those of proteins. As such, it is difficult to exhaustively summarize all the roles glycans play; a general summary of some important roles is provided with a focus on mammalian glycoproteins in physiology and disease. Glycans are mostly found coating the outer surface of macromolecules (*e.g.*, proteins) or cells. Thus, glycans are heavily involved in cell to cell, cell to macromolecule, and macromolecule to macromolecule interactions.

Glycans are involved in regulation of inflammatory processes by interacting with different lectins (a type of glycan binding protein (GBP)) [101]. For example, glycans expressed at the surface of neutrophils are recognized by Selectins that initiate the process of inflammation upon binding. Another class of lectins, Dectins, recognize β -linked glucose polymer present in a pathogen’s cell wall and also initiate inflammation upon binding. Finally, sialic acid-binding immunoglobulin-type lectins (SIGLECs), recognize sialic acid-containing glycans. Expressed on the surface of leukocytes, SIGLEC binding can inhibit inflammation by triggering eosinophil and neutrophil apoptosis.

Glycans are also involved in disease development and progression. The relation between glycobiology and cancer has been extensively studied with abnormal glycan epitopes being a hallmark of malignancy [102, 103]. Changes in expression of glycoproteins, glycosyltransferases, and glycosidases, or in the levels of nucleotide sugar donors available for glycan synthesis, can account for alteration of glycosylation in cancer. Truncated *O*-glycans are particularly associated with malignant transformation. Multiple mechanisms are responsible for the presence of truncated structures. Somatic mutations and hypermethylation of the Core-1 β galactosyltransferase Specific Molecular Chaperone (COSMC) were observed in pancreatic and colon cancer [104–106]. This chaperone is required for activation of T-synthase, a key enzyme in *O*-glycan synthesis. Without an active COSMC, short Tn and STn *O*-glycans accumulate while the Core-1 *O*-glycan and its complex derivatives cannot be synthesized [107]. In colon mucus, these short glycans are easily degraded by pathogens that colonize the gut generating colitis, an onset for carcinoma. Generally, truncated *O*-glycans: Tn, STn, T and ST antigens are cancer biomarkers and are responsible for metastasis and reduced immunosurveillance [103]. Simple *N*-glycan structures (without branching) are also known cancer glycan epitopes [108]. *N*-glycans with increased branching are also associated with increased cancer metastasis [8]. Certain *N*-glycan patterns, such as a sialylated galactose-galactose motif, have been found in cancers while being absent in normal tissue [109]. Chemical groups that modify glycans are also subject to change upon disease onset. The *O*-glycan T-antigen was found modified with a sulfate group in breast cancer [110].

Glycans are important cancer biomarkers, and several studies have shown their potential for diagnosis as well as disease prognosis [111–113]. Their use as a cancer biomarker is of particular interest as aberrant glycosylation typically occurs at the onset of cancer, enabling the potential for early diagnosis.

2. Theoretical background

2.2.5 Importance of protein glycosylation in the pharmaceutical industry

Protein glycosylation is also of critical importance to the pharmaceutical industry. Many recombinant proteins are used as therapeutic drugs ('biologic drugs'), among which, 60% are glycosylated [114]. Most of these biologic drugs are produced in mammalian cell lines including Chinese hamster ovary (CHO), baby hamster kidney (BHK), and murine myeloma (NS0) cells. Each can produce human-like post-translational modifications such as glycosylation.

Glycosylation of biologic drugs affects their circulating half-time and efficacy. For example, therapeutic antibodies with *N*-glycans lacking terminal sialic acid are rapidly cleared from the blood circulation due to recognition of exposed galactose residues by hepatocytes receptors [115]. In addition, the absence of a core-fucose residue on antibody *N*-glycans is associated with an increase antibody-dependent cell-mediated cytotoxicity (ADCC), enhancing the efficacy of certain biologic drugs [116]. This led to the development of FUT8 knock-out CHO cell lines to produce afucosylated *N*-glycan-bearing biologics [117].

The presence of certain glycoforms can also affect the safety of biologic drugs. Certain glycan epitopes not naturally synthesized by human cells are known antigenic structures. This is the case for the Gal- α 1-3-Gal- β 1-4-GlcNAc epitope, also commonly termed the ' α -Gal epitope'. The α -Gal epitope is not present on glycoconjugates in humans, apes or Old World monkeys as a result of an evolutionary inactivation of β -galactosyl α 1-3-galactosyl-transferase that prevents its synthesis [118]. As a consequence, these species express significant amount of anti α -Gal antibodies that represent ~1% of circulating immunoglobulins [119]. The α -Gal epitope is present in certain commercially available drugs such as Cetuximab and this raises immunogenicity concerns [120, 121]. Interestingly, anti- α -Gal antibodies also represent an opportunity to increase immunogenicity of viral vaccines. Expression of α -Gal epitopes on viral proteins such as HIV gp120 or influenza viral hemagglutinin protein generated ~100 fold higher antibody titers in mice [122, 123]. Very recently, a

similar strategy was proposed to enhance the immunogenicity of COVID-19 vaccines [124].

Neu5Gc is another human antigenic glycan epitope. Similar to the α -Gal epitope, synthesis of Neu5Gc was eliminated from the human population by mutation of the enzyme CMP Neu5Ac hydroxylase (CMAH) [125]. Circulating antibodies against this sugar moiety were found in healthy humans and were responsible for inflammation upon recognition of extrinsic Neu5Gc coming from dietary sources [126]. The presence of Neu5Gc has been reported in several biologic drugs, also raising immunogenicity concerns [127]. In this thesis, screening and identification of Neu5Gc-preferring sialidases was conducted and is reported in Chapter 4, section 4.4. These enzymes, for which an unusual substrate preference was described the first time, represent possible new preventive and/or curative treatment strategies for Neu5Gc-mediated intestinal inflammations [128].

Therefore, to ensure both efficacy and safety of recombinant therapeutic proteins, the US Food and Drug Administration (FDA) and the European Medicines Agency (EMA) regulatory administrations now require analysis and monitoring of their glycosylation [129–131]. Glycan structures are not only comprehensively characterized on a reference product, but are also routinely monitored in production lots and batches as biologic drug glycosylation depends on many culture conditions and can vary from batch-to-batch [132].

2.2.6 Glycoanalytics

Processes for determination of glycan structure has significantly improved over the past decade. Until recently, glycan analysis was highly complicated and lagged behind that of nucleic acids and proteins. However, advances in nearly every area of glycan analysis has occurred over the past two decades with better separation technologies [133, 134], development of high-throughput sample preparation workflows [135, 136], creation of reference databases [137, 138] or new glycan labeling chemistries [139]. Together, these advances have enabled not only better structural characterization of glycans in research but

2. Theoretical background

made possible routine glycosylation analyses of biologic drugs in the pharmaceutical industry [140, 141].

There is no universal method for analysis of glycoproteins, but rather a variety of techniques and workflows that enable identification of occupied glycosites and deconvolution of glycan structures [141–143]. Each method comes with a different degree of depth and throughput, and none alone enable comprehensive characterization of both glycosites and glycoforms. It is up to the experimentalist to choose a method based on available equipment, expertise and the depth of the analysis being sought. There are two general strategies that are often employed: i) characterization of glycans on glycoproteins or glycopeptides (glycoproteomics) and ii) characterization of glycans that have been released from a glycoconjugate (glycomics).

Analysis can be performed on intact glycoproteins using mass spectrometry (MS) (Figure 3A) [144, 145]. MS analysis of glycoproteins includes intact or top-down MS strategies [146–148]. In intact MS approaches, the MS¹ spectrum of the undigested glycoprotein alone is recorded while in top-down methods, if an intact protein is injected into the instrument, the protein becomes fragmented inside the mass spectrometer, and MS/MS (or MS²) spectra are generated. The key advantages of such methods are that only limited sample preparation is needed and that information on specific proteoforms can be attained. In intact glycoprotein analyses, the specific combination of different modifications of the protein can be identified and quantified. To achieve such results, a high resolution separation must precede MS analysis to enable separation of proteoforms whose mass differs by only a few Daltons [146]. Such separation is typically achieved using liquid chromatography (LC) or capillary electrophoresis (CE). In addition, lectins can be used up-front LC/CE-MS in enrichment strategies [149]. Lectins have a high specificity towards precise glycan epitopes. They can be immobilized in columns, on beads or on membranes and assist in separating protein glycoforms.

To better characterize glycan structures present on glycoproteins, bottom-up glycoproteomics methods are effective. In this approach, glycoproteins are subjected to proteolysis (with trypsin or proteinase K) (Figure 3B). Generated glycopeptides are separated and analyzed using MS/MS instruments [97, 98, 150, 151]. These workflows characterize glycans with great depth and give information on glycosites, however, they are generally low throughput. In addition, information about glycosidic linkages and isomer differentiation is not possible unless workflows are specifically adapted for such purposes [98, 152, 153].

Common approaches in glycoanalytics rely on separation of glycans that have been released from proteins (Figure 3C). *N*-glycans can be liberated from glycoproteins using amidases such as PNGase F, an enzyme capable of cleaving the bond between the innermost *N*-glycan GlcNAc residue and the asparagine to which the glycan is appended. Contrastingly, release of *O*-glycans from glycoproteins is not trivial because no broad-specificity endoglycosidase is capable of releasing complex *O*-glycans from proteins. Endo- α -N-acetylgalactosaminidase (also termed O-glycosidase) can only hydrolyze unextended core 1 and core 3 mucin-type *O*-glycans and does not cleave other cores or complex structures. For this reason, chemical release of *O*-glycans is preferred [154, 155].

2. Theoretical background

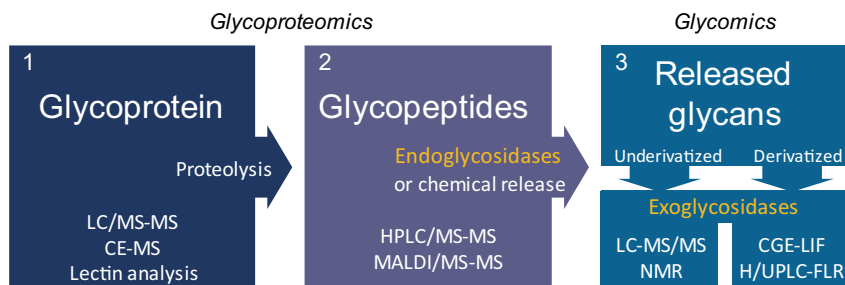


Figure 3. Main approaches employed in glycoanalytics. Adapted from [154]. The presented strategies use different starting material: intact glycoproteins (A), glycopeptides (B) or released glycans (C). Typical instrument methods for each strategy are listed (non-exhaustive). Endo- and exo-glycosidases are routinely used in analytical workflows to facilitate glycan release or determination of glycan structure respectively. Abbreviations: capillary electrophoresis (CE), capillary gel electrophoresis (CGE), high-performance liquid chromatography (HPLC), liquid chromatography (LC), matrix-assisted laser desorption ionization (MALDI), mass spectrometry (MS), nuclear magnetic resonance (NMR), ultra-performance liquid chromatography (UPLC).

Following their release, glycans can either be directly investigated or derivatized with probes for fluorescence detection or better MS ionization. Separation of labeled-glycans is commonly achieved by liquid chromatography (LC) techniques coupled with fluorescence detection (FLR) or MS [156–159] or by capillary (gel) electrophoresis (C(G)E) coupled with laser-induced fluorescence detection (LIF) [134–136, 159, 160]. C(G)E-LIF has a notable advantage in high-throughput analyses with the possibility to use 4, 16, 48 or 96 capillary-arrays in multiplexed capillary gel electrophoresis (xCGE) [134, 135, 159, 161].

Glycan derivatization can be achieved with a variety of fluorescent tags. *N*-glycans are commonly labeled by Schiff base condensation or using reactive carbamate chemistry but other labeling strategies using hydrazine or Michael addition can be performed [162]. In Schiff base condensation the transient glycosamine moiety from the released GlcNAc residue is reduced, enabling

reaction with an amine group present on the fluorescent tag. The compounds 2-aminobenzamide (2-AB) or 8-aminopyrene-1,3,6-trisulfonic acid (APTS) are common labels of this class and are used in LC-FLR and (x)C(G)E-LIF workflows, respectively. Both labels and their corresponding analytical techniques were employed in this thesis. To shorten the labeling time from a few hours to just minutes, strategies using carbamate chemistry were more recently developed [139]. In these methods, the transient glycosamine of GlcNAc is not reduced and instead directly reacts with an activated carbamate present on the label. However, this approach is limited to *N*-glycans that have been released by PNGase F which generates GlcNAc with a glycosamine moiety as part of its reaction mechanism.

Databases that contain information on common glycan structures and their corresponding retention or migration times have been developed [137, 138]. These resources facilitate glycan structure assignment through comparison of the retention/migration times of a sample with those present in the reference database. To confirm the structure assignment, highly specific exoglycosidases digestions are typically performed. Exoglycosidases are employed in a precise order to sequentially remove sugar units from the glycan's non-reducing end, effectively 'sequencing' the glycan [163–166]. The specificity of some exoglycosidases for certain linkages can permit resolution of linkage isomers. For instance, *Streptococcus pneumoniae* sialidase is specific for only α 2-3 linked sialic acid, and as such, is commonly used to differentiate α 2-3 from α 2-6 linked sialic acids. In Figure 4A, some endo- and exoglycosidases routinely employed for glycan sequencing are presented. These enzymes represent important tools in the glycoanalytical toolbox.

2. Theoretical background

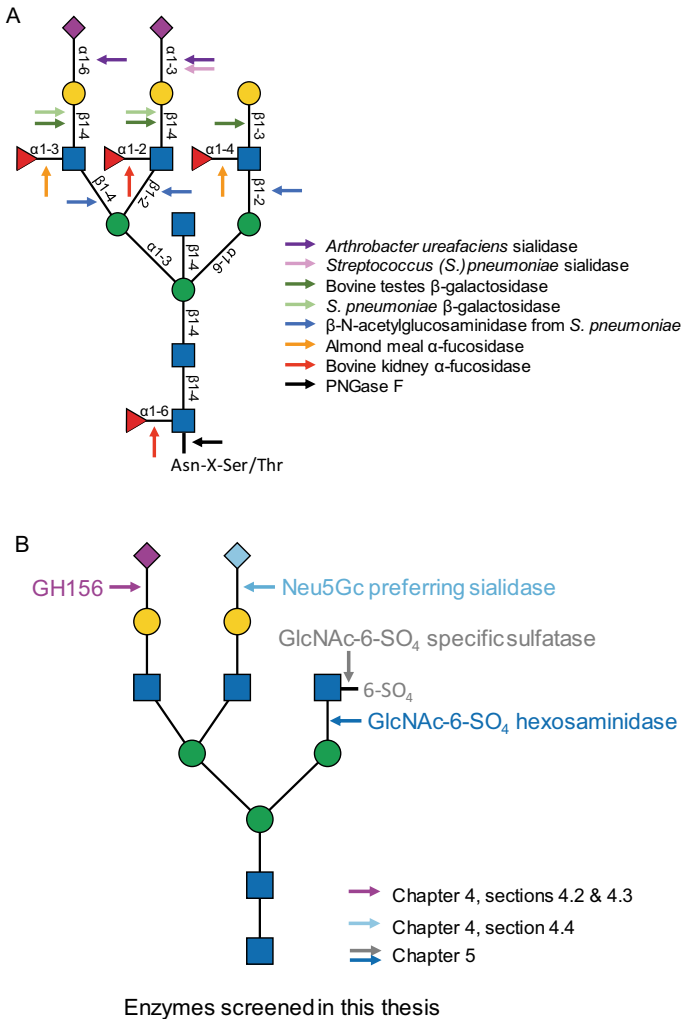


Figure 4. Enzymes in the glycoanalytical toolbox. (A) Typical exoglycosidases used in glycan sequencing. A single endoglycosidase (PNGase F) is represented but others are commonly used in glycoanalytics (e.g., Endoglycosidase H or Endoglycosidase S). Adapted from [154]. (B) Novel enzymes discovered during the course of this thesis using functional metagenomics.

While many exo- and endo-glycosidases are available for use in analytical workflows, some enzymatic activities that could help deconvolute the enormous structural complexity of the glycome are still missing. One notable example is highly specific enzymes that act on glycan PGMs which have not yet been established and validated for analytical use. Characterization of these modified glycans remains challenging, and as such, they are still rarely studied, and their biological roles poorly understood.

During the course of this thesis, three screening projects were conducted for enzymes acting on *N*-glycans resulting in identification of four activities (Figure 4B): a novel exosialidase family, termed GH156; two Neu5Gc-preferring sialidases having an unprecedented specificity; a GlcNAc-6-SO₄-specific sulfatase, and a GlcNAc-6-SO₄-specific hexosaminidase. The latter two representing attractive new glycoanalytical tools.

3 Development of a functional metagenomic workflow for enzyme discovery

Please note that Figure 5 is taken from the original publication:

Zaramela LS, Martino C, Alisson-Silva F, Rees SD, Diaz SL, **Chuzel L**, Ganatra MB, Taron CH, Secret P, Zuñiga C, Huang J, Siegel D, Chang G, Varki A, Zengler K. Gut bacteria responding to dietary change encode sialidases that exhibit preference for red meat-associated carbohydrates. *Nature Microbiology*. 2019, 4:2082–2089.

3. Development of a functional metagenomic workflow for enzyme discovery

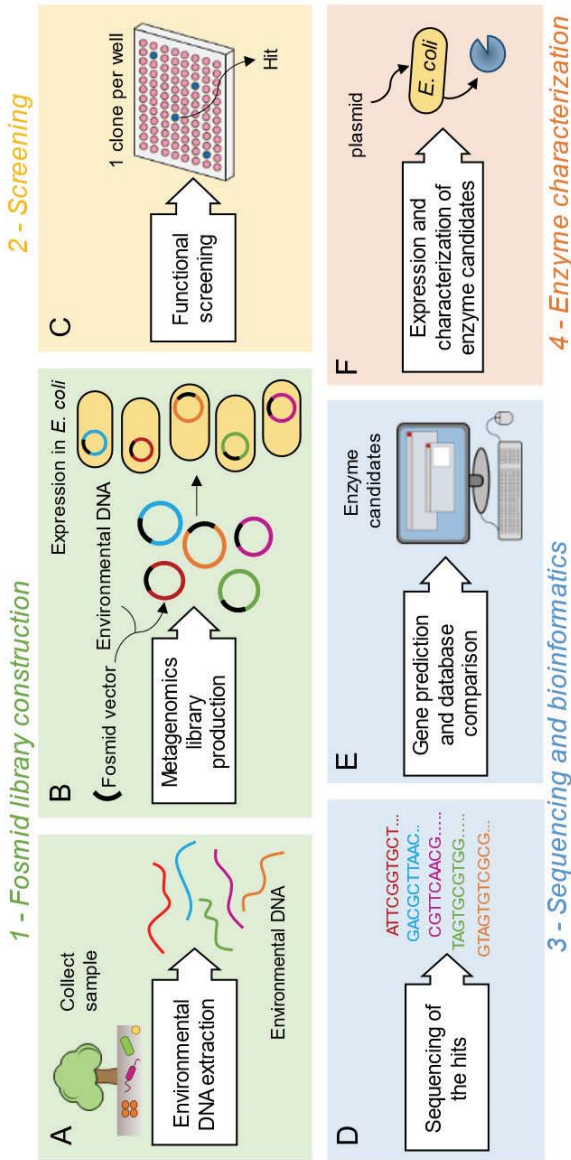


Figure 5. Functional metagenomics workflow. From [128]. (1) An ecological niche is sampled and eDNA from microorganisms living in the collected sample is isolated (A). eDNA is cloned into a fosmid backbone and inserted into *E. coli* (B), forming metagenomics libraries. (2) Metagenomic libraries are screened for enzymatic activities of interest in high-throughput assays (C). (3) Hits from the screens are fully sequenced using a next-generation sequencing technology (D). Genes encoded by the eDNA fragment are predicted *in silico* and compared to protein databases to identify enzyme candidates (E). (4) Candidates are cloned in an expression vector for protein over-expression, purification and characterization (F).

3. Development of a functional metagenomic workflow for enzyme discovery

3.1 Introduction

There is a need to identify new enzymes to help advance both basic science and applied biotechnology. From an academic perspective, discovery of new enzymes from various ecosystems (*e.g.*, extreme environments) or enzymes with biochemical properties not previously encountered adds to the body of knowledge of the world around us. Additionally, most genomes in environmental samples contain many open reading frames for which no function is known. Thus, high throughput enzyme discovery techniques can help assign function to these open reading frames (ORFs). From a biotechnology viewpoint, new enzymes with useful specificities can be developed into new analytical tools. In much of our work, analytical enzymes are applied to problems in the field of glycobiology. As introduced earlier (see Chapter 2, section 2.2.6), enzymes can improve the study and characterization of glycoconjugates. While many enzyme specificities are routinely used in glycoanalytical workflows, some specificities that can assist in comprehensive characterization of any glycostructure are still lacking. Functional metagenomics represents a powerful strategy for high-throughput enzyme discovery from any ecological niche. Therefore, a major aim of this thesis work was to establish a functional metagenomics workflows and library resources needed to enable function-based enzyme discovery, particularly for manipulation of carbohydrates. The workflow developed in this thesis is comprised of four parts (Figure 5) (1) fosmid library construction, (2) library screening, (3) sequencing and bioinformatic analysis of selected fosmid clones, and (4) characterization of identified enzyme candidates.

Construction of metagenomics libraries begins with collecting samples from various ecosystems followed by mass extraction of eDNA from its microbial community (Figure 5A). Isolated eDNA is subsequently cloned into a fosmid backbone prior to insertion into a surrogate host for micro-expression of the eDNA (Figure 5B). In this thesis, *E. coli* was exclusively used as a host. Arrayed eDNA libraries are then screened for enzymes of interest. In this thesis, unique glycoside hydrolases and sugar-specific sulfatases were sought (Figure

3. Development of a functional metagenomic workflow for enzyme discovery

5C). Clones harboring the activity of interest are then sequenced using the PacBio sequencing technology (Figure 5D). ORFs encoded by the eDNA fragment are deduced from the fosmid insert sequences and predicted genes compared to public databases to identify gene candidates responsible for the detected activity (Figure 5E). Finally, selected genes of interest are expressed *in vitro* and/or *in vivo* for biochemical characterization of the discovered enzymes (Figure 5F).

This chapter describes the development of this workflow and details the methods employed for fosmid eDNA library construction, screening and fosmid sequencing (Figure 5, steps 1, 2 &3). A substantial collection of over 100,000 arrayed clones from 20 ecosystems or individual organisms was built. This collection is a valuable resource that was central to the screens presented in this thesis and other screens currently being conducted at NEB. The entire workflow permitted discovery of the enzymes reported in both Chapters 4 and 5.

3.2 Material and methods

3.2.1 Environmental DNA extraction

Soil and water samples were stored at 4°C upon collection and DNA extraction was processed within a few days. Fecal samples were stored at -80°C prior to DNA extraction.

3.2.1.1 eDNA isolation from terrestrial samples

Microbial cells found in terrestrial samples were isolated by mixing 100 g of soil with 500 mL of extraction buffer (50 mM Tris pH 8.0, 1.5 M NaCl, 1 mM EDTA, 1 mM dithiothreitol (DTT), 0.1% Tween 20). The slurry was shaken vigorously and soil particles sedimented by letting the mixture stand for 1 h at 4°C. Supernatant was recovered in a 1 L bottle. An additional 500 mL of extraction buffer was added to the pellet and procedure repeated 3 more times. A final volume of 2 L of supernatant containing microbial cells was obtained. The supernatant was centrifuged at 10,000 x g for 3 min to pellet cells. The cell

3. Development of a functional metagenomic workflow for enzyme discovery

pellet was resuspended in 40 mL of light TE buffer (10 mM Tris-HCl pH 8.0, 0.1 mM EDTA). Four milliliters of lysozyme (100 mg/mL) (MilliporeSigma) and 10 μ L of RNase A (100 mg/mL) (Qiagen, Hilden, Germany) were added and the mixture was incubated for 30 min at 37°C. Sodium dodecyl sulfate (SDS) was then added to a final concentration of 0.5% along with proteinase K (NEB) in a ratio of 40 μ L per milliliter of sample. The sample was incubated for 1 h at 37°C. After lysis, cell debris was pelleted at 10,000 x g for 10 min at 4°C. The supernatant containing eDNA was transferred to clean tubes using wide-bore pipette tips to prevent mechanical shearing of the high molecular weight DNA. eDNA was isolated by adding an equal volume of phenol:chloroform:isoamylalcohol (IAA). The sample was carefully mixed by inversion and centrifuged at 15,000 x g for 5 min at room temperature. The top layer was transferred to a clean tube for isopropanol precipitation. Sodium acetate pH 5.2 (3 M) was added to the sample in a 1:10 buffer:sample volume ratio. Room temperature isopropanol (0.7 volume) was added, and the sample mixed gently several times. eDNA was pelleted by centrifugation at 4°C for 15 min at 15,000 x g. The eDNA pellet was washed three times with 70% ethanol. The pellet was dried and resuspended in 1.2 mL of light TE buffer. At this stage, the isolated eDNA solution still contained many impurities and had a dark brown color.

Isolated eDNA was further purified and size-selected on a 1% low melting point agarose gel. A Lambda HindIII ladder and a 40 kb control DNA ladder were run as size standards. The gel was run at 35 V in TAE buffer (in the absence of any DNA staining solution) overnight at 4°C. After electrophoresis, lanes containing size standards were cut from the gel and the gel stained with SYBR-Safe (Invitrogen, Carlsbad, CA). Gel regions containing eDNA of the desired size (30-70 kb) were cut from the gel and extracted using β -agarase I (NEB) following the manufacturer's instructions. eDNA was precipitated using isopropanol as described above. eDNA was resuspended in 200 μ L of light TE buffer and frozen until used.

3. Development of a functional metagenomic workflow for enzyme discovery

3.2.1.2 eDNA isolation from aquatic environments

eDNA from aquatic environments (pond or ocean water) was isolated using the Meta-G-Nome™ kit from Epicentre (Madison, WI). The kit is designed to isolate eDNA of ~40 kb to facilitate fosmid library construction. The manufacturer's protocol does not utilize beads nor other mechanical shearing techniques, in contrast to other commercial kits for metagenomic DNA isolation. The manufacturer's protocol was used with slight modifications. The initial filtration step performed using cheesecloth was omitted to prevent loss of microbial cells trapped in the cheesecloth. The amount of water processed was also increased from 100 mL to 250 mL.

Water was filtered through a 0.45 µm membrane using a Millipore Glass microanalysis filter holder and a water pump. The filter was then cut into half using sterile forceps and scissors, and both pieces were placed in the bottom of a 50 mL falcon tube with the upper face of the filter facing the inner part of the tube. Filter wash buffer was prepared as recommended by the manufacturer by addition of 2 µL of Tween 20 to 1 mL of the provided filter wash buffer. Freshly prepared buffer was added to the falcon tube containing the filter pieces and vortexed thoroughly for 2 min. The suspension containing recovered microbial cells was transferred to a 1.5 mL tube and centrifuged at 14,000 x g for 2 min. The obtained cell pellet was resuspended with 300 µL of TE buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA). Cells were lysed by adding 2 µL of Ready-Lyse Lysozyme solution, 1 µL of RNase A and incubating at 37°C for 30 min. Lysis was achieved by addition of an equal volume of Meta-Lysis solution and 1 µL of proteinase K (50 µg/µL). The sample was incubated at 65°C for 15 min. After incubation, the sample was brought to room temperature and cooled on ice for 5 min. A volume of 350 µL MPC Protein Precipitation Reagent was added and the sample vortexed vigorously for 10 s. Cell debris was pelleted by centrifugation at 20,000 x g for 10 min at 4°C. The supernatant containing eDNA was transferred to a clean tube and eDNA precipitated by addition of 570 µL of isopropanol. eDNA was pelleted by centrifugation for 10 min at 20,000 x g at 4°C. The supernatant was discarded and the eDNA pellet washed with

3. Development of a functional metagenomic workflow for enzyme discovery

500 μL of 70% ethanol. After centrifugation at 20,000 $\times g$ for 5 min at 4°C, ethanol was carefully removed, and the pellet dried at room temperature for 8 min. eDNA was resuspended in 50 μL of TE buffer and frozen until used.

3.2.1.3 eDNA isolation from human feces

Human microbiome DNA was isolated from a fecal sample from a healthy 29 year-old male (Donor T3806, Lee Biosolutions, Maryland Heights, MO, USA). To isolate microbial cells, 100 mg of fecal material was homogenized in 1 mL of phosphate buffer saline (PBS) pH 7.4 (Gibco™, Thermo Fischer Scientific). The sample was centrifuged at 500 $\times g$ for 1 min to remove large-sized debris. The supernatant was transferred to a clean tube. To minimize loss of microbial cells, the pellet was washed with 1 mL PBS and centrifuged at 500 $\times g$ for 1 min. Pooled supernatants were centrifuged at 3,000 $\times g$ for 10 min to pellet microbial cells. The supernatant was discarded. Pelleted cells were washed once with 1 mL PBS followed by centrifugation at 3,000 $\times g$ for 10 min. The cell pellet was washed once more with 1 mL of PBS and centrifuged at 500 $\times g$ for 1 min to remove large particles. The supernatant containing microbial cells in suspension was centrifuged at 3,000 $\times g$ for 10 min to obtain a pellet. To lyse cells, the pellet was resuspended in 500 μL of light TE buffer. To this cell suspension, 100 μL of lysozyme (10 mg/mL) and 1 μL of RNase A (100 mg/mL) were added. The sample was incubated at 37°C for 30 min, after which 20% SDS was added to a 0.5% final concentration. Proteinase K (NEB) was added in a ratio of 40 μL per milliliter of sample. The sample was incubated for 20 min at 37°C. Lysate was centrifuge at 10,000 $\times g$ for 10 min at 4°C to pellet cell debris. The supernatant was transferred to a clean tube with a wide bore pipette tip to avoid shearing of the eDNA. eDNA was isolated by addition of an equal volume of phenol:chloroform:IAA (25:24:1). To further clean the eDNA, phenol:chloroform:IAA extraction was performed twice followed by isopropanol precipitation as described previously. eDNA of 30-70 kb was size selected from a 1% low melting point agarose gel run overnight at 4°C at 35V as described for DNA extraction from soil samples.

3. Development of a functional metagenomic workflow for enzyme discovery

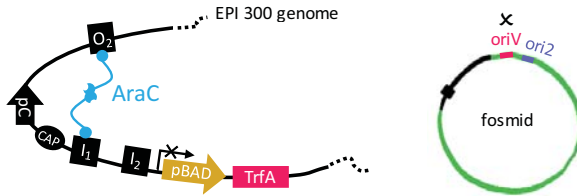
3.2.2 Fosmid eDNA library construction in *E. coli*

3.2.2.1 Library principle

Metagenomic libraries were created using one of two kits: the CopyRight® v2.0 Fosmid Cloning Kit (Lucigen Corporation, Middleton, Wi, now discontinued) for the first library produced (termed ‘Small Dixie’) or the CopyControl™ Fosmid Library Production Kit (formerly Epicentre, now Lucigen Corporation) for all others. Both kits rely on the same protocol and enable creation of large-insert DNA libraries using a fosmid vector (pSMART FOS or pCC1FOS) and λ phage particles. These kits and the fosmid cloning approach were chosen so that the collection could benefit from large cloned eDNA inserts, simple manipulation, and fosmid copy number control. Control over the fosmid copy number is an on-demand system that permits switching from 1 or 2 fosmid copies per cell to ~50-100 copies by addition of arabinose during *E. coli* growth. This feature is permitted by the presence of two distinct origins of replication on fosmid vectors: a low copy origin of replication (*ori2*, also termed *oriS*) and a high copy origin of replication (*oriV*). *Ori2* limits the number of copies to 1 or 2 per cells and is the default origin of replication [167]. *OriV* is dependent on the protein TrfA. By controlling the expression of TrfA in the host cells, *oriV* is regulated and consequently the fosmid copy number is controlled [168]. For this system to work, the expression of TrfA must be tightly regulated. Fosmids are thus inserted into modified *E. coli* strains (Replicator FOS or EPI300-T1^R), in which production of TrfA is tightly regulated by the pBAD promoter (Figure 6). In the absence of arabinose, the promoter pBAD is repressed by an AraC protein dimer. Upon binding to two DNA loci (termed *O*₂ and *I*₁) AraC forms a loop in the DNA that inhibits transcription (Figure 6A). Arabinose, by binding to AraC, changes AraC conformation and binding sites to *I*₁ and *I*₂. This relaxes the DNA loop and enables transcription of TrfA [169]. The expressed TrfA protein turns on *oriV*, resulting in fosmid replication at a high copy number (Figure 6B). During library construction and archiving, no arabinose was used to grow *E. coli*, ensuring fosmids were maintained at a low copy number for stability.

3. Development of a functional metagenomic workflow for enzyme discovery

A In absence of arabinose – single fosmid copy



B In presence of arabinose – multiple fosmid copies

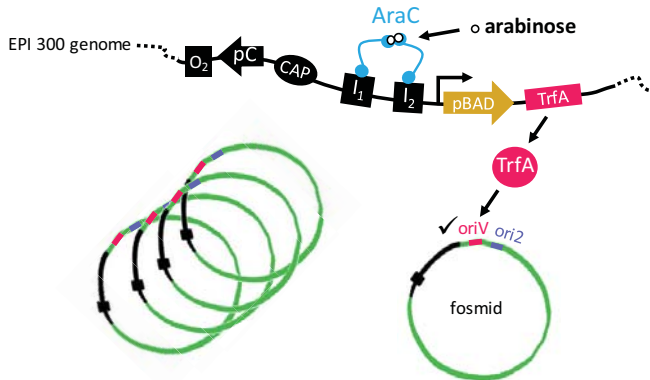


Figure 6. Fosmid copy control system. Fosmid vectors possess a tunable system via which their copy number can be switched from low to high upon addition of arabinose. (A) In the absence of arabinose, fosmid replication is performed via *ori2* a low copy origin of replication. *OriV* is then repressed as controlled by the protein TrfA which transcription is blocked. TrfA is under the tight regulation of the pBAD promoter. Transcription by pBAD is prevented by an AraC dimer which upon binding to *O2* and *I1* forms a loop structure. (B) The addition of arabinose enables transcription of TrfA by pBAD. Arabinose binds AraC changing its structure and DNA binding sites, in turns relaxing the loop. Expressed TrfA enables fosmid replication through *oriV*, a high copy origin of replication.

3. Development of a functional metagenomic workflow for enzyme discovery

Cloning of ~30-40 kb eDNA fragments into the 8 kb fosmid vector is performed by blunt-ligation. Ligation generates concatemers of ~38-48 kb units, each separated by a *cos* site on the fosmid backbone (Figure 7). eDNA is then packaged into λ particles *in vitro* by mixing ligated DNA with λ phage extracts. Packaging is initiated upon recognition of two adjacent *cos* sites separated one from another by 37-52 kb eDNA by λ phage proteins [170, 171]. *Cos* sites are then cut to form linear pieces of ~37-52 kb eDNA with cohesive ends that are packaged into phage particles. Packaged phages containing linear fosmid DNA are then mixed with the *E. coli* host. Particles bind to the *E. coli* cell surface LamB maltoporine receptor through which DNA is inserted into the host cell. Once inside the cell, the cohesive ends on the introduced DNA associate and ligate together to generate circular fosmids (Figure 7).

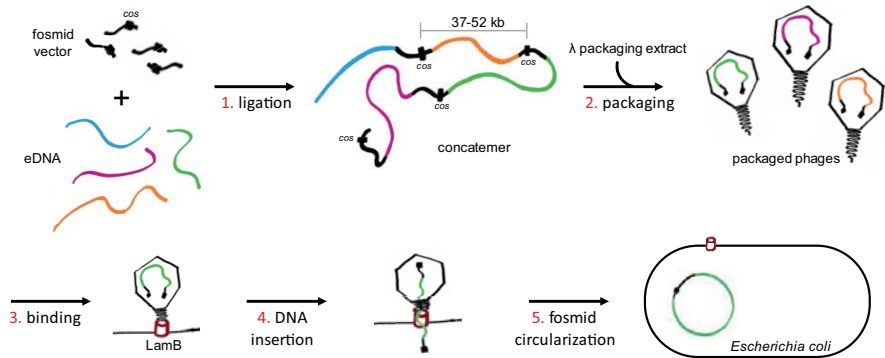


Figure 7. Construction of fosmid metagenomic libraries. eDNA fragments are ligated with the fosmid vector forming concatemer molecules (1). *Cos* sites present on the fosmid vector are recognized by λ phage proteins enabling *in vitro* packaging of individual eDNA-fosmid units (2). Phages packaged with eDNA-fosmid segments bind *E. coli* via the maltoporine LamB (3). eDNA-fosmid is injected into the *E. coli* host cell (4) where cohesive ends associate forming circular fosmid molecules (5).

3. Development of a functional metagenomic workflow for enzyme discovery

3.2.2.2 Library assembly procedure

The first metagenomic library (termed 'Small Dixie') was produced using the CopyRight® v2.0 Fosmid Cloning Kit (Lucigen Corporation, Middleton, WI). Prior to construction of the library, eDNA was concentrated using 0.45X volume of AMPure PB magnetic beads (PacBio, Menlo Park, CA). Binding of the eDNA to beads was performed at room temperature in a Thermomixer (Eppendorf, Hambourg, Germany) for 10 min at 1400 r.p.m. The supernatant was removed, and the beads were washed twice using 500 µL of 70% ethanol. eDNA was eluted from the beads with water and shaking for 2 min at 1400 r.p.m. in a Thermomixer. The library was constructed using the CopyRight® v2.0 Fosmid Cloning Kit following the manufacturer's instructions with minor adaptations. eDNA was first end-repaired and size-selected using a 1% low melting point agarose gel run overnight at 35 V. eDNA fragments from 30-70 kb were extracted from the gel using 1 U of β-agarase I (NEB) for each 100 µL of melted agarose. The end-repaired and size-selected eDNA was ligated to the pSMART FOS cloning vector respecting a 1:10 molar ratio of insert:vector. Phage packaging of the resulting clones was performed using Gigapack III XL packaging extracts (Agilent Technologies, Santa Clara, CA). Replicator FOS cells were transfected with the packaging reaction and plated on YT-CXIS agar medium (8 g Bacto-tryptone, 5 g yeast extract, 5 g NaCl, 15 g agar per liter containing 12.5 µg/mL of chloramphenicol, 40 µg/mL 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-Gal), 0.4 mM isopropyl β-D-1-thiogalactopyranoside (IPTG), and 5% (w/v) sucrose) and incubated overnight at 37°C. A total of 616 colonies were harvested and archived in two 384-well plates in 20% sterile glycerol (v/v) and were frozen at -80°C until use.

All other libraries were produced using the CopyControl™ Fosmid Library Production Kit (Lucigen Corporation, Middleton, WI). Size selection of eDNA was performed prior to end-repair of the DNA. Ideally, 1 µg of size-selected eDNA was used to start the library production procedure. However, depending upon the sample, as little as ~200 ng was used. eDNA was concentrated using 0.45X volume of AMPure PB magnetic beads as described above. Purified

3. Development of a functional metagenomic workflow for enzyme discovery

DNA was end-repaired following the kit's instructions. After the end-repair reaction, eDNA was again cleaned and concentrated with 0.45X volume of AMPure PB magnetic beads. eDNA was eluted in 7 μ L of light TE buffer and ligated to the pCC1FOS vector following the kit's instructions. Typically, 500 ng of pCC1FOS vector was mixed with 250 ng of eDNA for a 1:10 molar ratio of insert:vector. If less than 250 ng of eDNA was retrieved after end-repair and clean-up, the amount of eDNA and pCC1FOS were adjusted to maintain a 1:10 insert:vector molar ratio. Ligation was performed for 4 h at room temperature. Ligated product was then packaged into λ phage particles *in vitro* following the manufacturer's instructions. A small portion of the packaging reaction was used to transfect EPI300-T1^R cells. The transfection reaction was plated on LB-agar plates (10 g tryptone, 5 g yeast extract, 10 g NaCl, 1 g dextrose, 1 g MgCl₂-6H₂O, 20 g bacto-agar, and 2 mL 2M NaOH per liter) to determine the phage titer. The remaining packaged phage were stored as a 20% glycerol stock at -80°C. Libraries in the form of packaged phages were shipped to BioS&T (Montreal, Ontario, Canada). BioS&T performed large-scale transfection and robotic archiving of individual *E. coli* clones in 384-well barcoded plates.

The entire collection of arrayed library clones was stored in the form of micro-culture glycerol stocks in 384-well plates at -80°C. Two copies of each library plate were generated, a working copy that is regularly thawed at room temperature and replicated for screening, and an archive copy kept in a different -80°C freezer as a backup.

3.2.3 Library quality assessments

Restriction digestion was used to assess the average insert size in each fosmid eDNA library. Fosmids were individually isolated from 12 randomly selected library clones using the FosmidMAX™ DNA purification kit (Lucigen Corporation, Middleton, WI). Fosmids from the 'Small Dixie' library were digested with the SbfI endonuclease (NEB). Fosmids from other libraries were digested with FseI (NEB). Digestions were performed at 37°C for 1 h in 50 μ L reactions with 1X CutSmart® Buffer (NEB) using 10 U of SbfI or 2 U of FseI

3. Development of a functional metagenomic workflow for enzyme discovery

and 1 μg of fosmid DNA. The average insert size of cloned eDNA was estimated by analyzing digestion products on a 1% agarose gel or on a genomic DNA ScreenTape (Agilent Technologies, Santa Clara, CA).

To assess the eDNA insert diversity, the same 12 isolated fosmids were subjected to Sanger sequencing using the T7 universal primer and/or the pCC1 forward and reverse sequencing primers (Epicentre, Madison, WI) (Appendix: Supplementary table 1). Comparison of obtained nucleotide sequences was used to look for duplicate clones and assess the potential origin of the eDNA by BLASTN to probe the GeneBank sequence database.

3.2.4 High-throughput functional screening

3.2.4.1 Agar plate-based enzyme screening

Library clones were spotted onto LB agar plates supplemented with 40-60 $\mu\text{g}/\text{mL}$ of 5-bromo-4-chloro-3-indolyl- β -D-galactopyranoside (X-Gal) or 5-bromo-4-chloro-3-indolyl- α -D-N-acetylneuraminic acid (X-Neu5Ac) (MilliporeSigma) dissolved in dimethylformamide (DMF) using a 384 Slot Pin Multi-Blot™ Replicator (V&P Scientific Incorporation, San Diego, CA). Clones were cultivated overnight at 37°C. Agar plates were later stored at 4°C until blue color developed (several days).

3.2.4.2 Lysate-based enzyme screening

The library clones were grown in 50 μL LB liquid cultures (10 g tryptone, 5 g yeast extract, 10 g NaCl, 1 g dextrose, 1 g $\text{MgCl}_2 \cdot 6\text{H}_2\text{O}$, 2 mL of 2M NaOH per liter, containing 12.5 $\mu\text{g}/\text{mL}$ chloramphenicol and 1X inducing solution (Lucigen Corporation, Middleton, WI) in black 384-well plates with clear bottom (Greiner Bio-One, Kremsmünster, Austria) and incubated overnight at 37°C. Inducing solution was added prior to enzyme screening to increase the fosmid copy number and boost protein production from cloned environmental genes. Fifty microliters Y-PER™ lysis buffer (Thermo Fischer Scientific, Waltham, MA) containing 40 $\mu\text{g}/\text{mL}$ of 4-methylumbelliferyl-D-galactopyranoside (4-MU-Gal) (MilliporeSigma) were added to the liquid

3. Development of a functional metagenomic workflow for enzyme discovery

cultures. The mixtures were incubated overnight at 37°C in a static incubator. Fluorescence at $\lambda_{\text{ex}}=365$ nm and $\lambda_{\text{em}}=445$ nm was monitored over time with a SpectraMax Plus 384 Microplate Reader (Molecular Devices, Sunnyvale, CA).

3.2.5 Fosmid sequencing and bioinformatics

Fosmid DNA was extracted using the FosmidMAX™ DNA purification kit following the manufacturer's instructions. DNA was isolated from 40 mL LB cultures supplemented with 12.5 µg/mL chloramphenicol and 1X inducing solution to increase the fosmid copy number and obtain higher fosmid DNA yields. Enzymes used for preparation of the PacBio SMRTbell libraries were from NEB (Ipswich, MA).

3.2.5.1 Single clone sequencing

A 10-kb SMRTbell library was prepared using 10 µg of fosmid DNA dissolved in 150 µL light TE buffer. DNA was sheared using g-tubes (Covaris, Woburn, MA). The g-tubes were centrifuged at 4500 r.p.m. (Eppendorf 5424 centrifuge) for 30 s with the cap up, followed by 30 s with the cap down, three times each for a total of 6 spins. Sheared fosmid DNA was concentrated using 0.45X volume of AMPure PB magnetic beads as described previously (see section 3.2.2). DNA was eluted from the beads with 20 µL of light TE buffer. Two microliters of PreCR® Repair Mix and 5 µL of NEBNext® End Repair Reaction Buffer in a 50 µL total volume reaction was added to repair the damaged DNA template. The reaction was incubated at 37°C for 20 min. DNA was end-repaired by addition of 2.5 µL NEBNext® End repair mix (incubation at 25°C for 5 min) and then purified with 0.45X volume of AMPure PB magnetic beads and eluted with 20 µL of light TE buffer. DNA was blunt-ligated with 2.5 µL of 80 µM SMRT bell adaptors (Integrated DNA Technologies, San Jose, CA) using 17.5 µL of Blunt/TA Ligase Master Mix. The ligation reaction was incubated at 25°C for 15 min. Ligase was removed using the Monarch® PCR and DNA Cleanup kit. Unligated DNA fragments were removed by digestion with Exonuclease III and VII (100 U each) in the presence of 4 µL of buffer for T4 DNA ligase with 10 mM ATP. Finally, two

3. Development of a functional metagenomic workflow for enzyme discovery

AMPure PB magnetic bead purifications were performed. The SMRTbell library was ultimately eluted with 15 μ L light TE buffer.

3.2.5.2 Multiplexed-sequencing

To lower the cost and time associated with sequencing, multiplexed libraries were prepared when large number of fosmid clones were to be sequenced. Twelve fosmids were multiplexed in a single 10-kb SMRT-bell library, enabling sequencing of 12 fosmids using a single SMRT cell. Fosmid DNA was first sheared into 10 kb pieces using g-tubes (Covaris, Woburn, MA). Shearing was performed individually for each fosmid using 5 μ g of fosmid diluted in 150 μ L as described previously (3.2.5.1. Single clone sequencing). Sheared DNA was purified using 0.45X volume of AMPure PB beads with elution in 10 μ L light TE buffer. Concentration of fosmid DNA after shearing and cleaning was quantified using the Qubit[®] dsDNA broad-range assay (Thermo Fischer Scientific, Waltham, MA). The DNA was end-repaired and ligated in a single reaction. For each clone, 400 ng of DNA in 5 μ L was mixed with 1.5 μ L of NEBNext[®] End Repair Reaction Buffer, 1 μ L of NEBNext[®] End Repair mix, 1 μ L of T4 DNA ligase high concentration, 1.5 μ L of MQ water and 5 μ L of PacBio barcoded adapter. Reactions were incubated for 20 min at 37°C, followed by 15 min at 25°C, and 10 min at 65°C. All twelve reactions were pooled as each fosmid possessed a unique barcode. The barcoded fosmid pool was purified using 0.45X volume of AMPure PB beads. DNA was eluted with 43 μ L of light TE buffer. Any DNA damage was repaired by addition of 5 μ L of ThermoPol buffer, 2 μ L of PreCR[®] Repair Mix and 0.5 μ L of 100X NAD⁺ and incubation at 37°C for 20 min. DNA fragments that were not ligated to adapters, and thus did not form circular SMRT-bell templates, were degraded by addition of 1 μ L of Exonuclease III and 1 μ L of Exonuclease VII and incubation for 1h at 37°C. The SMRT-bell library was then purified twice using 0.45X volume of AMPure PB magnetic beads. Final elution was performed with 10 μ L of light TE buffer.

3. Development of a functional metagenomic workflow for enzyme discovery

3.2.5.3 Blue Pippin size-selection

Multiplexed SMRT-bell libraries were size-selected to remove SMRT-bell templates smaller than 8 kb using a BluePippin instrument (Sage Science, Beverly, MA). About 400 ng of SMRT bell library in 30 μ L light TE buffer was mixed with 10 μ L of loading buffer and pipetted into the sample well of a 0.75% agarose cassette. The instrument was set to collect fragments larger than 8 kb. To minimize sample loss, once electrophoresis was finished and sample recovered from the elution module, 50 μ L of 0.1% Tween 20 solution was added to the elution module. After incubation for 20 min at room temperature this 50 μ L aliquot was collected and pooled with the previously collected sample. Finally, size-selected libraries were cleaned by performing three 0.45X volume AMPure PB bead purifications. Final elution from the beads was performed in 10 μ L of light TE buffer.

3.2.5.4 RSII sequencing and *de novo* assembly

MagBead complexes were prepared from the sequencing libraries on the day of sequencing using enzymes from PacBio. The standard MagBead OCPW protocol provided by the Binding Calculator tool was used with default parameters for non-multiplexed samples. For multiplexed sequencing libraries, the MagBead OCPW protocol was used with standard parameters except an on-plate concentration of 0.2 nM.

MagBead complexes were sequenced on a PacBio RS II instrument using the P6 chemistry and one single-molecule real-time (SMRT) cell for a 360 min movie. For multiplexed libraries, reads were first demultiplexed into different sets of reads (one set per barcode) each containing data for a single fosmid. After demultiplexing, the HGAP.3 protocol was used to *de novo* assemble the reads. Non-multiplexed libraries were directly assembled using the HGAP.3 protocol available on the PacBio secondary analysis portal. Obtained contigs typically contained an overlap at their ends because of the circular structure of fosmids (Figure 8).

3. Development of a functional metagenomic workflow for enzyme discovery

Assembly overlaps at each end



Figure 8. De novo assembler overlap error on circular fosmids. Adapted from [172]. The HGAP.3 pipeline used to assemble Pacific Bioscience sequencing reads from fosmids generates overlaps at the contig ends. Commonly observed for circular sequences this assembly error must be corrected.

Today's assemblers automatically assume the sequence to be assembled is linear, generating errors in the circular sequence assembly. This well-known limitation of *de novo* assemblers can be fixed using the bioinformatic tool Circlator [172]. Circlator identifies the circular sequence and generates a linearized version that does not contain overlaps. Contigs obtained from HGAP.3 were thus input into the Circlator pipeline via command line along with the non-assembled PacBio reads (the second data input required by Circlator) to correct this type of assembly error.

3.2.5.5 ORF prediction and ORF map drawings

ORFs encoded by any assembled contig were predicted using MetaGeneMark [173]. The table output files from MetaGeneMark were modified to enable automatic importation of the predicted ORFs as annotations into the software Geneious (<https://www.geneious.com/>). The MetaGeneMark table was converted into a Browser Extensible Data (BED) file format by i) replacing the first column with the clone I.D., ii) placing the strand column at the right-most position, iii) replacing the gene length column by a score of 1000 (Appendix: Supplementary Figure 2). The second MetaGeneMark output file that consists of the deduced protein sequences of each predicted ORF was subjected to BLASTP analysis against the GenBank non-redundant protein

3. Development of a functional metagenomic workflow for enzyme discovery

database. ORFs that showed good identity and coverage scores to several proteins of the same function were annotated with this putative function. ORFs that showed no significant homology to any protein of known function or that matched proteins annotated as “hypothetical proteins” were assigned as such.

3.3 Results and discussion

3.3.1 Metagenomics libraries and collection

3.3.1.1 Environmental DNA

Effective metagenomic libraries must represent as much of the microbial diversity of an ecosystem as possible. Thus, the method used to lyse cells and extract eDNA must be efficient over a broad range of species. Ideally, the cell lysis technique must be capable of breaking down the cell walls of any type of microbial cell. Microbes can be divided into three groups: archaea, Gram-positive bacteria and Gram-negative bacteria, and each have a different cell wall structure. In both Gram-positive and Gram-negative bacteria, the cell wall is composed of peptidoglycan, a rigid structure formed by polysaccharide layers associated with amino acids and derivatives [174]. The peptidoglycan layer of Gram-positive bacteria is much thicker than that of Gram-negative bacteria. However, the latter possess a unique outer membrane absent in Gram-positive bacteria. Lysozyme cleaves the bond between two components of the peptidoglycan: N-acetylmuramic acid and N-acetylglucosamine. This enzymatic treatment is efficient on Gram-positive bacteria, but most Gram-negative bacteria are insensitive to lysozyme due to the outer membrane [175]. The lysis method employed to isolate eDNA combined lysozyme with SDS, a detergent efficient at interrupting a wide range of cell wall structures [176], as well as proteinase K, a protease known to facilitate cell wall degradation [177]. In addition, by degrading DNases, proteinase K also prevents deterioration of eDNA. The combination of these three components increased the likelihood of accessing the most genetic information from a wide range of sampled microorganisms.

3. Development of a functional metagenomic workflow for enzyme discovery

The presence of eukaryotic DNA in our metagenomic libraries may affect screening in *E. coli*. As both prokaryotes and eukaryotes are ubiquitous in most environments, it was anticipated that the eDNA libraries would contain eDNA from bacteria, archaea, and eukaryotes. Eukaryotic eDNA present in cloned fosmids in *E. coli* would not be expected to express due to the presence of introns in eukaryotic genes, and a different promoter structure than that of *E. coli*. Thus, if eukaryotic eDNA fosmids were to be dominant in number within a library, it would be severely detrimental to the success of any screen. During routine quality sampling of the metagenomic libraries (see section 3.2.3) no clones containing DNA from eukaryotic species were encountered. While these samplings dealt with only small numbers of randomly selected clones, they at least show that eukaryotic eDNA fosmids are not dominating the libraries. Ongoing efforts to entirely sequence all 5,376 clones from the Dixie hot spring library (S. Fossa & C. Taron, unpublished observations) should enable a first analysis of the proportion of fosmid inserts that originating from eukaryotes in one of the libraries.

Isolation of high-quality and high molecular weight eDNA is also imperative for assembling good fosmid libraries. Obtaining eDNA of an appropriate size is crucial and has a direct impact on generating fosmid clones. In the event eDNA fragments being too big or too small, the λ -phage DNA packaging process cannot be completed and no fosmid clone is formed. eDNA of ~30-40 kb must be retrieved to construct the library (the optimal size for λ -phage). If the molecular weight of the eDNA is too high, it can become fragmented into smaller pieces by mechanical shearing during isolation. Recovery of eDNA smaller than 30 kb on the contrary represents a problem. For this reason, samples were handled gently during DNA isolation and wide-bore pipette tips were used during sample manipulation to limit mechanical shearing. Following eDNA isolation, its size was always measured. For water samples, isolated eDNA was generally greater than 30 kb and no size-selection was required (Figure 9A). However, eDNA extracted from tougher more particulate samples, such as human feces or soil, generally contained a broad

3. Development of a functional metagenomic workflow for enzyme discovery

range of sizes (Figure 9B). To remove eDNA of smaller sizes, a size selection was performed by selecting eDNA of ~30-40 kb from an overnight agarose gel. As shown in Figure 9B, this efficiently removed smaller fragments from the eDNA mixture. Using these procedures, high-quality eDNA of an appropriate high molecular weight was successfully extracted from various ecological niches and enabled construction of numerous fosmid clone libraries.

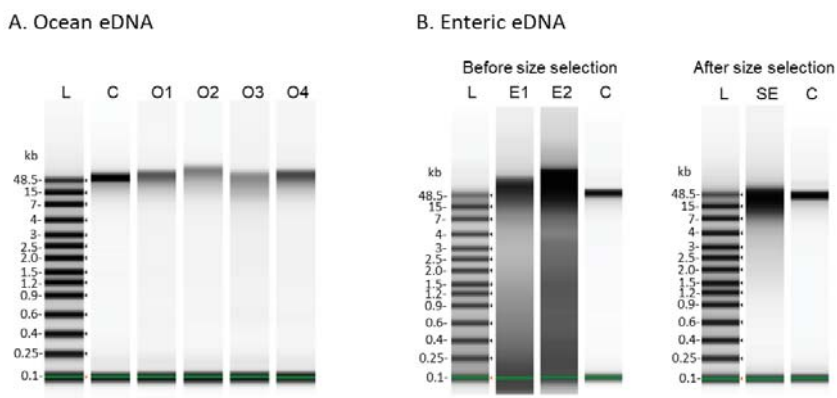


Figure 9. Environmental DNA extraction from different ecosystems. Size of the isolated eDNA was monitored on a genomic DNA tape station and compared to a ladder (L) and a 40 kb DNA control (C). eDNA isolated from ocean source (O1-4) was mostly larger than 30 kb in size (A), while DNA isolated from human feces had a broad size distribution (E1-2) with many fragments below 30 kb (B). The size distribution of enteric eDNA was narrowed by selecting fragments above 30 kb from an agarose gel. Analysis of size-selected fragments (SE) showed successful size-selection of the highest molecular weights.

3.3.1.2 The NEB Collection in November 2019

The metagenomic and genomic libraries created in this thesis work are presented in Table 1. With the exception of a first metagenomic library (named ‘Small Dixie’), the libraries I created constitute a collection herein referred to as the ‘NEB Collection’. The Small Dixie library was excluded from the NEB Collection as it was a small first test library that was created using the

3. Development of a functional metagenomic workflow for enzyme discovery

CopyRight® v2.0 Fosmid Cloning Kit, while all other libraries were produced using the CopyControl™ Fosmid Library Production Kit (see section 0). The fosmid vector backbone and the *E. coli* host strain were consequently different for this first library. A larger number of clones were obtained when using the CopyControl™ Fosmid Library Production Kit justifying my switch to this system. Despite ‘Small Dixie’ not formally belonging to the NEB Collection, the library was used in a screen that led to the discovery of a novel family of sialidases that is presented in Chapter 4, section 4.2. For the Dixie hot spring metagenomic community to be represented in the NEB Collection, a second library using the same eDNA was created using the CopyControl™ Fosmid Library Production Kit. This library termed ‘Dixie’ contains 5,376 clones and more deeply represents eDNA from the microbial community living in that ecosystem.

As of November 2019, the NEB Collection was comprised of 21 libraries with a total of 99,456 clones. Represented libraries have three origins: eDNA (constituting metagenomic libraries), genomic DNA from a single organism, and genomic DNA from mock microbial community (both constituting genomic libraries). Ten libraries are of metagenomic origin (Figure 10), representing 81,024 clones. Of these, half were constructed with eDNA generously donated by colleagues Dr. Richard D. Morgan and Dr. Andrew F. Gardner (Table 1). The other half of the metagenomic libraries were constructed from eDNA extracted within the scope of this thesis as described in section 3.2.1. For a typical metagenomic library in this collection, fourteen 384-well plates (5,376 clones) were archived. Even if a larger number of clones could have been archived for most libraries, the number of clones was intentionally limited to 5,376. It was decided that diversity of ecosystems represented in the collection was preferred over extensive numbers of clones for each library. Thus, the goal was to construct a collection representing many different ecological niches. For two metagenomics libraries from soil origin (the oak and pine tree libraries) not enough clones were obtained to archive fourteen 384-well plates. Two and five 384-well plate libraries were instead created, respectively.

3. Development of a functional metagenomic workflow for enzyme discovery

	Library name	Library type	Sample origin	Sample form	# of clones	Collection barcode	Vector
	Small Dixie	Metag.	Dixie Hot Springs, Nevada, USA	eDNA	616	n.a.	pSMART
NEB collection	Dixie	Metag.	Dixie Hot Springs, Nevada, USA	eDNA	5,376	NEB 0001-NEB 0014	pCC1
	Compost	Metag.	Brick Ends Farm, Hamilton, Massachusetts, USA	soil	5,376	NEB 0015-NEB 0028	pCC1
	NEB Pond Water	Metag.	New England Biolabs, Ipswich, Massachusetts, USA	water	5,376	NEB 0029-NEB 0042	pCC1
	Ocean	Metag.	Long beach, Gloucester, Massachusetts, USA	water	5,376	NEB 0043-NEB 0056	pCC1
	<i>Enterococcus faecalis</i>	Genomic	American Type Culture Collection	gDNA	768	NEB 0057-NEB 0058	pCC1
	<i>Akkermansia muciniphila</i>	Genomic	American Type Culture Collection	gDNA	768	NEB 0059-NEB 0060	pCC1
	<i>Elizabethkingia meningoseptica</i>	Genomic	American Type Culture Collection	gDNA	768	NEB 0061-NEB 0062	pCC1
	20 strain even mix microbial standard	Genomic mix	American Type Culture Collection	mixed gDNA	10,752	NEB 0063-NEB 0090	pCC1
	Salt Marshes	Metag.	Ipswich, Massachusetts, USA	eDNA	5,376	NEB 0091-NEB 0104	pCC1
	Oak tree	Metag.	Soil under an oak tree, Ipswich, Massachusetts, USA	eDNA	768	NEB 0105-NEB 0106	pCC1
	Pine tree	Metag.	Soil under a pine tree, Ipswich, Massachusetts, USA	eDNA	1,920	NEB 0108-NEB 0112	pCC1
	<i>Thermococcus gorgonarius</i>	Genomic	American Type Culture Collection	gDNA	768	NEB 0113-NEB 0114	pCC1

3. Development of a functional metagenomic workflow for enzyme discovery

<i>Thermus antranikianii</i>	Genomic	American Type Culture Collection	gDNA	768	NEB 0115-NEB 0116	pCC1
Blue Lagoon (Warm water, Iceland)	Metag.	Blue lagoon warm water, Iceland	eDNA	5,376	NEB 0117-NEB 0130	pCC1
Human Gut microbiome #1	Metag.	Man (29 yrs), Donor T3806, Lee Biosolutions, Maryland Heights, Missouri, USA	feces	23,040	NEB 0131-NEB 0190	pCC1
<i>Thermococcus zilligii</i>	Genomic	American Type Culture Collection	gDNA	768	NEB 0191-NEB 0192	pCC1
<i>Thermobifida fusca</i>	Genomic	American Type Culture Collection	gDNA	768	NEB 0193-NEB 0194	pCC1
<i>Archaeoglobus fulgidus</i>	Genomic	American Type Culture Collection	gDNA	768	NEB 0195-NEB 0196	pCC1
<i>Thermoplasma volcanium</i>	Genomic	American Type Culture Collection	gDNA	768	NEB 0197-NEB 0198	pCC1
Human Gut microbiome #2	Metag.	Man (29 yrs), Donor T3806, Lee Biosolutions, Maryland Heights, Missouri, USA	feces	23,040	NEB 0199-NEB 0258	pCC1
<i>Thermococcus kodakarensis</i>	Genomic	American Type Culture Collection	gDNA	768	NEB 0259-NEB 0260	pCC1

Table 1. NEB metagenomic and genomic library collection. Abbreviation: metag.: metagenomic.

3. Development of a functional metagenomic workflow for enzyme discovery

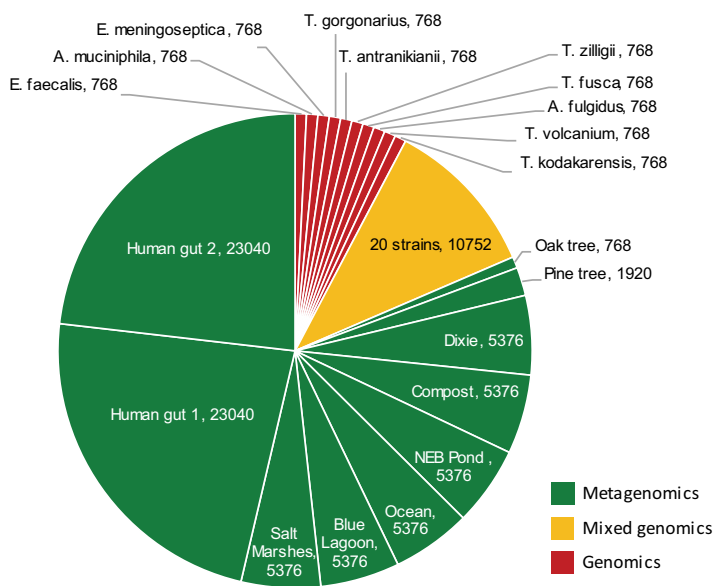


Figure 10. Distribution of libraries comprising the NEB Collection. For each library its name and the number of clones archived is indicated. Libraries are of three different types: metagenomic eDNA, mixed genomic DNA, and genomic DNA.

The largest metagenomic library in the collection consists of eDNA isolated from feces of a healthy male donor. Two libraries were created from the same eDNA extract, each comprised of 23,040 clones. The human gut of a healthy individual is estimated to carry on average ~160 different bacterial species, although, estimations of over 1000 species have also been reported [178, 179]. Members of the gut microbiota have the capacity to utilize complex carbohydrates abundant in dietary fruits, vegetables, and cereals, as well as those originating from human cells. The human gastro-intestinal tract is covered by a mucin layer comprised of heavily O-glycosylated proteins [180]. Investigation of an *in silico* modeled microbiome composed of 177 genomes revealed the presence of 15,882 different carbohydrate active enzymes (CAZymes) consisting of known glycoside hydrolases, glycoside transferases

3. Development of a functional metagenomic workflow for enzyme discovery

and polysaccharide lyases [14]. The high abundance of CAZymes in the human gut makes it an attractive hunting ground for discovery of novel enzymes in the glycobiology field. For this reason, and in an attempt to capture the diversity of this ecological niche, a larger number of clones (46,080) were archived for this library.

In addition to metagenomic libraries, ten single-genome libraries were also constructed. These libraries were primarily made to add genetic diversity to the collection from archaea and extreme hyperthermophiles. While the primary motivation of this thesis was to discover enzymes relevant to the field of glycobiology, the library collection was also intended to be useful to NEB researchers studying nucleic acid enzymes, a field where thermoresistant enzymes are often of interest. Extremophiles live in harsh environments and often possess proteins and enzymes with enhanced stability and activity in extreme conditions that can benefit industrial and biotechnological applications [64]. Single genome libraries were also created for some mesophiles that are known to produce certain carbohydrate-active enzymes. Genomic DNA from *Akkermancia muciniphila*, *Enterococcus faecalis* and *Elizabethkingia meningospectica* was used for these constructions. *Akkermancia muciniphila* populates the gastrointestinal tract of humans where it can decompose heavily O-glycosylated mucins [181]. *Elizabethkingia meningospectica* produces Peptide-N-Glycosidase F (PNGase F), an important enzyme that is capable of hydrolyzing the bond between N-glycans and glycoproteins [182, 183]. Similarly, *Enterococcus faecalis* encodes endo- α -N-acetylgalactosaminidase (O-glycosidase), an enzyme that catalyzes the removal of Core 1 and Core 3 O-glycans from glycoproteins [184]. The rationale for creation of libraries for these organisms was that they may encode additional yet-undiscovered activities in addition to those they are known to produce. To make these libraries, pure genomic DNA purchased from the American Type Culture Collection (ATCC) was used. For each single organism genomics library, two 384-well plates (768 clones) were archived. Considering the genome and

3. Development of a functional metagenomic workflow for enzyme discovery

average clone insert sizes, two plates give 6-20X genome coverage depending on the species

Finally, one library was constructed from a mixture of genomic DNA from 20 human associated bacteria. Termed the '20 strain even mix' this genomic DNA mixture represents a mock microbial community whose species were selected for their GC and Gram stain diversity. To ensure all 20 genomes were fully covered in this library, twenty-eight 384-well plates (10,752 clones) were archived.

Altogether, the collection of large insert metagenomic libraries represents about 3-4 Gb of cloned eDNA available for expression in *E. coli*. As the average prokaryotic gene size is roughly 1 kb, it is estimated that the NEB Collection contains between 3 to 4 million environmental genes. This constitutes a vast pool of genes from which new enzymes can be discovered. While these are compelling totals, it is important to note that not all cloned genes will express in *E. coli*. As discussed in the theoretical background, heterologous expression of eDNA from fosmids mostly relies on the presence of intrinsic promoters present within the cloned eDNA. The fosmid backbone used in this work contains the T7 promoter (pT7) at one of its ends which can drive expression of properly oriented genes located in close downstream proximity, but expression of genes located further away or in the opposite direction of pT7 rely on the presence of other promoters internal to the eDNA insert. The *E. coli* transcriptional machinery must be capable of recognizing these foreign promoters for gene expression to occur. Consequently, the species' phylum whose eDNA is cloned greatly affects the expression of genes by *E. coli*. Species whose promoter structure is closed to that of *E. coli* have a better chance of being expressed. In an *in silico* study, ~7% of Actinobacteria's genes were considered expressible from their own intrinsic promoters in *E. coli*, while this number reached ~73% for the more closely related phylum Firmicutes [53]. On average, however, ~40% of prokaryotic genes are predicted to be accessible for expression in *E. coli*. As such, the number of *E. coli* expressible genes in the NEB Collection may be closer to 1.2 to 1.6 million.

3. Development of a functional metagenomic workflow for enzyme discovery

In this thesis work, a solid collection of metagenomic libraries was created. Representing diverse ecological niches, from human microbiome to thermal springs, it encompasses numerous microorganisms and constitutes a useful resource for enzyme discovery programs.

3.3.1.3 Library quality

Clone archiving in 384-well plates was performed by a service provider (see section 3.2.2). Prior to shipping a library for archiving, a quality control was performed to ensure that the library eDNA inserts were of proper size and that the clones were diverse and represented species consistent with the sample's origin. To accomplish this, a small number of clones, (typically 12) were randomly selected and their fosmids isolated. Restriction analysis using a rare 8-base cutting endonuclease was performed. The enzyme SbfI was used for the 'Small Dixie' library and FseI for all other libraries. These enzymes have a single restriction site in the vector pSMART FOS and pCC1FOS, respectively. Digests were analyzed on 1% agarose gels. An example of 12 FseI digested fosmids from the human gut metagenomic library is presented in Figure 11. In this example, distinct restriction patterns were observed for each fosmid. This indicates that analyzed clones possess a unique eDNA fragment. Additionally, the size of the DNA also revealed that no empty fosmids (fosmids containing no eDNA insert) were present in the selected clones.

3. Development of a functional metagenomic workflow for enzyme discovery

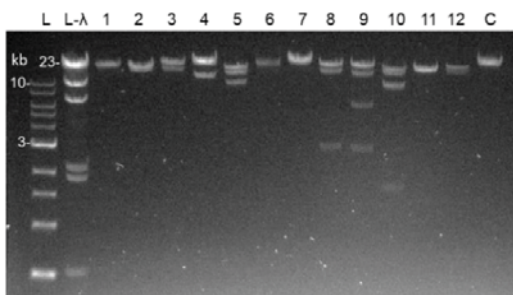


Figure 11. Restriction fragment analysis of 12 clones from the human gut metagenomic library. The 12 fosmids were digested with the enzyme FseI (1-12) and digests separated on a 1% agarose gel. Two ladders were used: 1 kb Plus DNA Ladder (L) and lambda DNA digested with Hind III (L-λ). A 40 kb fosmid control was also run to serve as a size marker (C).

In addition, isolated fosmids were sequenced using Sanger sequencing. In this analysis, the T7 universal primer that anneals to the fosmid backbone and is directed toward one end of the eDNA insert was used (Appendix: Supplementary Figure 3). Sanger sequencing generates a short (~600-800 bp) sequence of each insert. These sequence tags were used to probe the GenBank sequence repository to assist in identification of the species of origin for the cloned eDNA fragments. An example of this analysis for the Dixie library is shown in Table 2.

Associated species or genus	# of clone	Comment
<i>Thermocrinis ruber</i>	1	Hyperthermophilic bacterium, identified for the first time in Yellow stone Nation Park (USA) [185]
<i>Desulfobacula</i> genus	1	Members were found in aquatic habitat [186]
<i>Thermomicrobium roseum</i>	1	Hyperthermophilic bacterium, identified for the first time in Yellow stone Nation Park (USA) [187]
<i>Synechococcus</i>	1	Cyanobacteria widely distributed in marine environments [188]
<i>Thermus</i> genus	1	Group of thermophilic bacteria [189]
No significant match	3	Unsequenced species
Ambiguous data	1	Sequencing data matches a conserved DNA fragment present in various species

Table 2. Sanger sequencing analysis of randomly selected clones from the Dixie metagenomics library.

3. Development of a functional metagenomic workflow for enzyme discovery

Different species were observed, revealing the library was diverse and encompassed DNA from several different genomes. All organisms were consistent with the aquatic hot spring origin of the library (Table 2). Interestingly, sequence tags for 3 clones did not match any sequence in public databases, indicating the eDNA insert for these clones originates from unsequenced species. This observation also directly illustrates that functional metagenomics yields access to genomic DNA from unknown or understudied species.

Because the outcome of a screening project is directly impacted by the quality of a library, performing these simple quality controls when building a metagenomic library collection is important. These quality control checks were performed for each library constructed in this thesis work prior to arraying clones in 384-well plates. In all cases, libraries added to the NEB collection passed these quality benchmarks. Furthermore, sequence tags from unsequenced species were a recurring observation in the metagenomic libraries, emphasizing the promise of these libraries in enzyme discovery.

3.3.2 Plate-based and lysate-based screenings

Functional screening is a key step in the screening workflow. A screening assay needs to be amenable to high-throughput, sensitive, and reproducible. Because this thesis work focuses mostly on glycoside hydrolases, assays were developed to isolate this class of enzymes. To develop the assays I used several commercial reporter substrates that consisted of monosaccharides bearing a chromophore or a fluorophore molecule (Figure 12).

3. Development of a functional metagenomic workflow for enzyme discovery

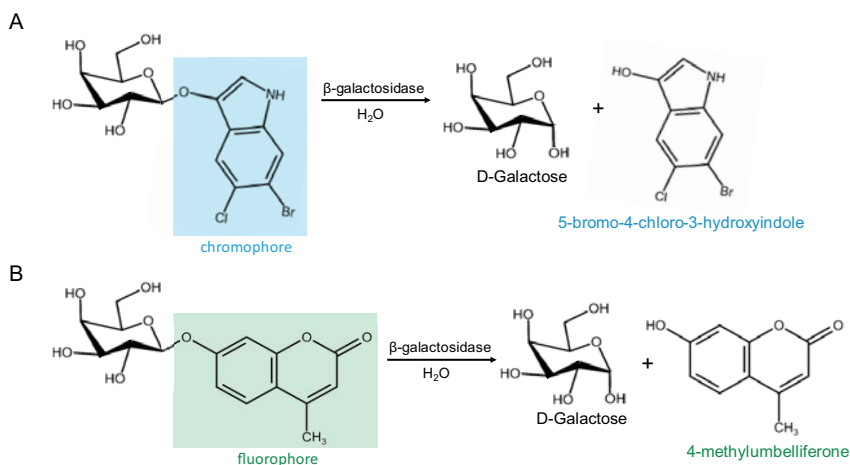


Figure 12. Chromogenic and fluorogenic substrate for glycoside hydrolase screening. (A) The X-Gal substrate can be used to assay β -galactosidases. Upon action of β -galactosidases, the sugar moiety is separated from the chromophore releasing 5-bromo-4-chloro-3-hydroxyindole which upon self-dimerization and oxidation forms a blue compound. (B) Similarly, 4-MU-Gal can be used in fluorescence-based assay to detect β -galactosidase activity. Action of the enzyme liberates the fluorophore 4-methylumbelliferone which emits fluorescence at ~ 445 nm when excited at ~ 365 nm.

When devising the screening workflow, I first developed plate-based screening assays. These were not only easy to execute but were also very high in throughput. These assays rely upon chromogenic substrates as illustrated in Figure 12A. However, because this type of assay has significant limitations (discussed below), I soon switched to lysate-based assays using soluble fluorogenic substrates (Figure 12B). Both types of assays were used in this thesis, hence their principle, features, advantages, and drawbacks will be discussed in the next two sections.

3.3.2.1 Plate-based screening

In plate-based screens, chromogenic substrates composed of a sugar residue linked to 5-bromo-4-chloro-3-indole were used. The substrate is colorless in

3. Development of a functional metagenomic workflow for enzyme discovery

solution, but upon hydrolysis by a glycoside hydrolase, 5-bromo-4-chloro-3-hydroxyindole is released. The latter dimerizes spontaneously to form 5,5'-dibromo-4,4'-dichloro-indigo, an insoluble blue compound (Figure 12A). A typical substrate of this kind is X-Gal, widely used in molecular biology for blue/white screening, a technique that distinguishes a desired cloning product from undesired ones [190]. In the plate-based assay, a chromogenic substrate was mixed with growth medium in agar plates onto which metagenomics libraries were replicated using a pin-replicator. This approach required very little material and was extremely fast to execute. An entire 384-well plate was screened on a single LB-agar plate, and many plates can be grown simultaneously. To lower the screening cost, that is mostly attributed to the chromogenic substrate, the LB-agar thickness was minimized and rectangular Petri dishes that perfectly fit a 384-well plate were used.

An example of this type of assay being used to screen the ‘Small Dixie’ library is represented in Figure 13. The library was screened for β -galactosidases and sialidases using X-Gal and X-Neu5Ac, respectively. One hit was observed for each screen as revealed by the presence of a blue colony on each plate (Figure 13A and Figure 13B- black arrows).

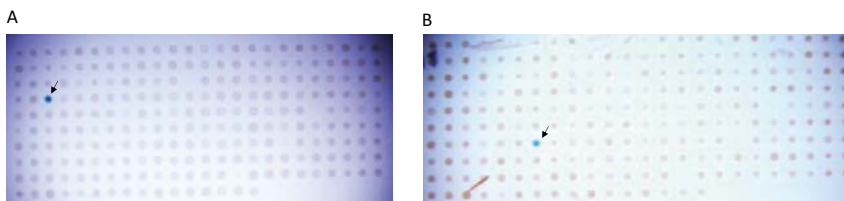


Figure 13. Plate-based screening for β -galactosidase and sialidase activities. The ‘Small Dixie’ library was screened for β -galactosidase (A) and sialidase (B) (from [191]) activity in a plate-based assay using X-Gal and X-Neu5Ac. A single hit was observed for both screens as shown by the strong blue color which developed on the clone located in position D4 (A) or G7 (B) for the β -galactosidase or sialidase screens respectively (black arrow). Clone coordinates were given with letters in alphabetic order for the rows and numbers in ascending order for the columns.

3. Development of a functional metagenomic workflow for enzyme discovery

While the limited hands-on time and the high throughput of this approach are undeniable advantages, it also suffers from several limitations. First, few monosaccharides or oligosaccharides linked to the 5-bromo-4-chloro-3-indole chromophore are commercially available, restricting the number of possible screens that can be easily executed. Second, the screening output is qualitative and relies upon the development of a color. Hits with low activity might be missed as light-blue colonies may not be perceived by the naked eye. In addition, the blue color may take a long time (many days) to develop. As a result, after initial overnight colony formation at 37°C agar plates must be stored at 4°C until enough blue pigment accumulates to be detected. In the sialidase screen of the ‘Small Dixie’ library, the blue color took more than a week to develop, and the hit illustrated in Figure 13B was almost missed. Therefore, to overcome these limitations, I transitioned from a plate-based to a lysate-based screening approach.

3.3.2.2 Lysate-based screening

Lysate-based assays were developed using substrates consisting of a sugar moiety coupled to the fluorophore 4-methylumbelliferone (4-MU). When linked to a sugar moiety, the 4-MU fluorophore is quenched. The action of a glycoside hydrolase releases 4-MU that once free in solution, emits fluorescence at ~445 nm upon excitation at ~365 nm. To execute lysate-based screens, metagenomic libraries were replicated in a black 384-well plate previously filled with liquid LB growth medium. Plate filling was performed with a robotic liquid handler to ensure a high throughput and good well-to-well reproducibility. After overnight growth of the cells at 37°C, a microculture of each individual clone was obtained. Lysis buffer containing the fluorogenic substrate was then added to each microculture.

To develop such assays, I investigated lysis solutions that upon addition to the microcultures could efficiently lyse the *E. coli* host cells without the need for additional physical breakage (*e.g.*, sonication). In this approach, hands-on time was minimized and a good screening throughput achieved. Three different

3. Development of a functional metagenomic workflow for enzyme discovery

lysis solutions were compared. Two were commercial lysis products (Y-PER™ and B-PER™ (ThermoFischer) and one termed ‘NEB’ was developed by a colleague. The B-PER™ and NEB solutions were designed to lyse prokaryotes, while Y-PER™ was developed to lyse yeast cells. To compare the three lysis solutions in a screening assay, 4 clones termed C1-C4 were assayed for β -galactosidase activity. C1 and C2 were used as negative controls as both were known to not produce β -galactosidase activity. C3 and C4 were used as example positive clones. C3 was an isolate from the ‘Small Dixie’ library that was identified as a β -galactosidase hit in the plate-based assay described above (Figure 13A, clone with the plate coordinate D4). C4 contains the empty pCC1FOS vector backbone. The absence of an insert in this vector leaves the LacZ gene from pCC1FOS intact enabling the clone to produce β -galactosidase activity. Clones C1-C4 were cultivated in 384-well plates in duplicate. Microcultures obtained after an overnight growth at 37°C were lysed with the three different lysis solutions containing the substrate 4-MU-Gal (Figure 14).

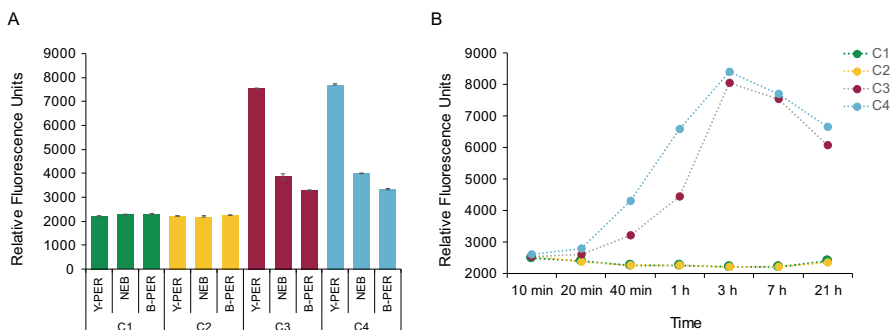


Figure 14. Development of a lysate-based screening assay. Three lysis buffers: Y-PER™, NEB and B-PER™ were tested for screening metagenomic libraries in a lysate-based format. Assay was performed using 4-MU-Gal on 4 clones: 2 clones with no β -galactosidase activity (C1 and C2) and 2 clones encoding a known active β -galactosidase (C3 and C4). Fluorescence at $\lambda_{ex}=365/\lambda_{em}=445$ nm after 7 h incubation at 37°C is reported for all three buffers tested (A). Fluorescence over time of the 4 clones is shown for the best performing buffer: Y-PER™ (B).

3. Development of a functional metagenomic workflow for enzyme discovery

The fluorescence displayed by the clones was monitored at 365/445 nm for 21 hours. Comparison of measured fluorescence for each lysis solution is shown after 7 hours (Figure 14A). A significantly higher fluorescence was detected for C3 and C4 compared to C1 and C2 regardless the lysis solution tested. However, the fluorescence signal was noticeably higher when lysis was performed with Y-PER™. This difference points to the lysis efficiency of Y-PER™ being better than that of B-PER™ or NEB buffers. However, the pH of the three lysis buffers may also influence performance of the 4-MU substrate. The released 4-MU fluorophore has a pKa ~8 and shows a stronger fluorescence in higher pH solutions [192]. It is then possible that the pH of *E. coli* lysates generated with Y-PER™ is better suited for the 4-MU fluorophore. Considering the fluorescence displayed by clones devoid of activity (C1 and C2) compared to that displayed by clones expressing an activity (C3 and C4), Y-PER™ had the best signal to noise ratio in this assay. Consequently, Y-PER™ was chosen as the lysis reagent for the lysate-based screens.

4-MU based assays can be executed in continuous or discontinuous methods. In 4-MU discontinuous protocols, the assay is stopped by addition of a strong base before measuring the fluorescence signal. Addition of a base elevates the mixture pH, increasing the fluorescence signal of 4-MU [192, 193]. While better sensitivity can be achieved using a discontinuous 4-MU assay, accumulation of fluorescence generation over-time cannot be monitored. In a continuous 4-MU assay, a clones' ability to generate fluorescence over time can be monitored. Increases in fluorescence over time can increase the confidence in defining hits. A clone whose fluorescence is measured only once over the hit threshold during the course of the assay might be a false positive. However, as shown in Figure 14B, the fluorescence of true hits (clones C3 and C4) increases over-time, and as such, crosses the hit threshold at multiple timepoints. To take advantage of collecting data overtime, I elected to perform the 4-MU screening assays using a continuous protocol using Y-PER™ to lyse cells. In a typical screen, after addition of the Y-PER™ and substrate mixture, plates were incubated at 37°C and fluorescence typically measured after 1 h, 3 h, 7 h, 24 h

3. Development of a functional metagenomic workflow for enzyme discovery

and 48 h. Fluorescence was typically not measured past 48 h due to fluorescence quenching and lysate evaporation.

A lysate-based screening assay using a continuous protocol was developed to enable high-throughput screening of metagenomic libraries for glycoside hydrolases. Conditions for efficient lysis of the *E. coli* host cells in a single step were established. The assay was validated for β -galactosidases using 4-MU-Gal but was later used for various enzyme activities in different screens with other 4-MU-coupled monosaccharides (see Chapter 4 and 5). This screening approach overcomes the two major limitations of plate-based screening as the assay signal fully develops within 48 h and is quantitative.

3.3.2.3 Hit definition

The development of a lysate-based assay was principally motivated by the need to obtain quantitative data to numerically discriminate a hit from a non-hit. Concurrent with establishing a lysate-based screening protocol, a scheme for systematic data analysis was also framed and used to evaluate hits.

Despite the lack of a universal mathematical interpretation of a hit, one common definition is any measurement that is at least 3 standard deviations greater than the mean [194]. I adopted this definition and considered primary screens hits to be clones that generated a fluorescence signal higher than the mean + 3 standard deviations (Figure 15).

3. Development of a functional metagenomic workflow for enzyme discovery

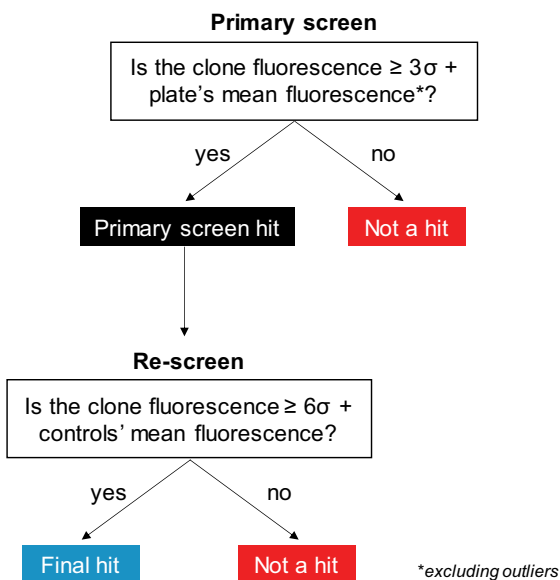


Figure 15. Definition of screening hits. Decision tree followed to discriminate hits from non-hits upon high-throughput screening of metagenomic libraries.

The mean and the standard deviation were always calculated separately for each plate to account for plate-to-plate variability. By doing so, it was ensured that the calculated standard deviation remained low, enabling more hits to be found, including those with the lowest activities. A flaw of this definition is the detrimental impact statistical outliers have on the calculated mean and standard deviation. To overcome this limitation, obvious outliers, whose fluorescence value is far above the background noise were excluded from the mathematical calculations. Because most of the clones in a 384-well plate are not hits, at most, only a handful of clones for each plate were pulled from the calculation. For large screens, clones considered hits in a first screening assay ('primary hits') were collected and archived in a new plate. They were then re-screened in the same manner with the intention of identifying possible false positives. In the re-screen, the plate contains mostly hits which prevents the mean and standard

3. Development of a functional metagenomic workflow for enzyme discovery

deviation to be calculated as described above. To circumvent this, a negative control clone containing the empty fosmid vector backbone is also cultivated in each well of an entire row of the archived hits plate. The mean and standard deviation are then calculated from the control replicates. Due to the smaller variability observed amongst the control wells, a stricter hit definition is used, with clones having a signal 6 standard deviations above the control mean being considered a hit (Figure 15). Examples using 4-MU fluorescent screening are presented in Chapters 4 and 5.

The transition from a plate-based to a lysate-based screening approach enabled to establish a mathematical definition of a hit, making screening data analysis reliable and independent from the screener's visual perception.

3.3.3 Sequencing the hits and generation of fosmid maps

Following their identification, fosmids from hits were sequenced using the PacBio next generation sequencing technology. Analysis of a hit's cloned insert DNA is a key step towards selection of gene candidates thought responsible for the enzymatic activity detected upon screening. To sequence hits, isolated fosmids that typically range from 30-50 kb were mechanically sheared into ~8-10 kb fragments using Covaris g-tubes. These fragments were ligated to DNA sequencing adaptors to obtain PacBio sequencing libraries that were run on a RSII instrument.

In the workflow, different protocols for PacBio sequencing were employed depending on the number of hits identified during screening (Appendix: Supplementary Figure 1). When the number of hits was limited to a few (*e.g.*, <8), sequencing was done individually, and a distinct sequencing library was prepared for each clone. When more than a few hits were identified, fosmid sequencing was performed using a multiplexed approach. To reduce the time and costs associated with sequencing, up to 12 clones were sequenced together. To do so, isolated fosmids from 12 clones were barcoded with different DNA sequencing adapters enabling distinction from one another. PacBio offers 96 distinct barcodes permitting 96 samples to be sequenced as part of a single

3. Development of a functional metagenomic workflow for enzyme discovery

sequencing library. However, the amount of sequencing data produced per sample decreases linearly with the multiplexing level. To ensure enough data were produced and proper assembly could be achieved, no more than 12 clones were multiplexed at once. This multiplexing level already represented a challenge and required the sequencing library preparation to be optimized. Optimization involved the introduction of a size-selection step in the multiplexed sequencing library protocol. The RSII instrument preferentially sequences the smaller fragments of a sequencing library, a phenomenon attributed to the tendency for smaller molecules to better load into the PacBio SMRT cells [195]. This bias constitutes a problem for *de novo* assembly of the data, as smaller fragments are more difficult to assemble. To overcome this issue, the sequencing library size distribution was narrowed using a BluePippin device (Sage Science, Beverly, MA) to remove DNA fragments smaller than 8 kb. This step efficiently increased the average size distribution of the PacBio sequencing library as illustrated in Figure 16.

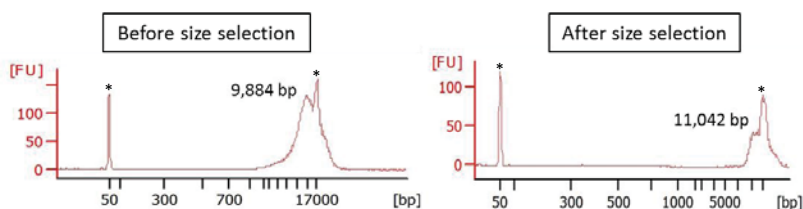


Figure 16. Size distribution of a multiplex PacBio library before and after size selection. Size distribution was analyzed on 12000 DNA Chip (Bioanalyzer, Agilent Technologies, Santa Clara, CA) of a multiplex PacBio library before and after size selection with the BluePippin device (8 kb threshold). Peaks marked with (*) correspond to internal standards used during analysis. Abbreviations: FU: fluorescence units, bp: base pair.

3. Development of a functional metagenomic workflow for enzyme discovery

Following sequencing, PacBio reads were assembled *de novo*, using no reference sequence. This was achieved using the HGAP.3 protocol, a data analysis workflow provided by PacBio. Reads from non-multiplexed PacBio sequencing libraries were directly input to the HGAP.3 pipeline without the need for data pre-processing. However, the multiplexing approach to fosmid sequencing also involved making changes to the computational workflow for read assembly. The HGAP.3 protocol provided by PacBio does not handle multiplexed data. It was then necessary to pre-process obtained reads and sort them according to the 12 barcodes prior to *de novo* assembly (Figure 17). This was achieved using a computational tool developed at NEB by my colleague Dr. Vladimir Potapov. Once reads were sorted into 12 sets, they were individually input to the HGAP.3 *de novo* assembly pipeline.

After *de novo* assembly, a few data processing steps were still required to obtain the full fosmid insert from the contigs generated by the HGAP.3 protocol (Figure 17). First, contigs were circularized to remove overlaps at their ends (for details see section 3.2.5.4. RSII sequencing and *de novo* assembly). Then, the vector backbone pCC1FOS or pSMART FOS was trimmed from the circularized contig to obtain only the eDNA insert (Figure 17).

3. Development of a functional metagenomic workflow for enzyme discovery

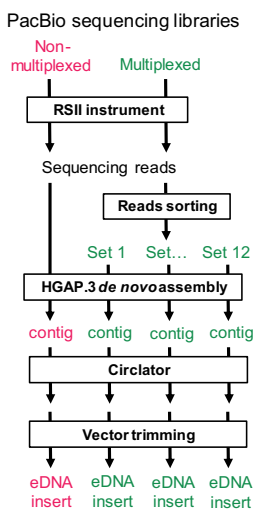


Figure 17. Fosmid sequencing data processing. Fosmids were sequenced using the PacBio sequencing technology. Data from non-multiplexed sequencing libraries were directly assembled using the HGAP.3 pipeline. Reads from multiplexed sequencing libraries were first sorted in different sets of reads, one set per barcode. Individual sets of reads were then assembled using HGAP.3. Contigs obtained after de novo assembly with HGAP.3 were corrected using the Circlator command-line software and the fosmid vector backbone sequence was trimmed. DNA sequences obtained after data processing correspond to the full eDNA fragment insert.

Each predicted ORF protein sequence was then compared to the nonredundant GenBank protein sequence database by BLASTP analysis. Based on the observed similarity of each ORF to homologous protein in GenBank, they were classified as proteins of known function or hypothetical proteins. Proteins of known function were then sub-divided into different groups to highlight activities relevant to the screening project. This classification was represented by coloring the annotations on the ORF map. As an example, the ORF map of a sulfatase screen hit (described in Chapter 5) is illustrated in Figure 18.

3. Development of a functional metagenomic workflow for enzyme discovery

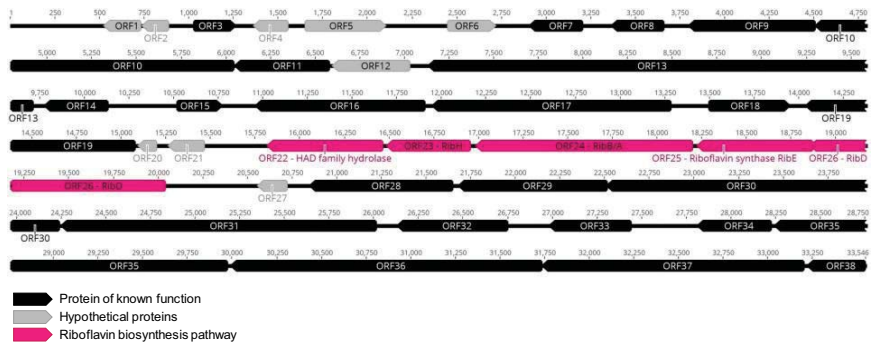


Figure 18. ORF map of a fosmid clone from the human gut metagenomic library. The fosmid was sequenced using the Pacific Bioscience technology. ORFs were predicted with MetaGeneMark and annotations imported into the software Geneious for drawing. ORFs were classified into three categories based on their homology to proteins from the nonredundant protein database (NCBI). In black, ORFs with a known annotated function. In grey, ‘hypothetical proteins’ with no annotated function. In pink, proteins from the riboflavin biosynthesis pathway.

In this example, 38 ORFs were predicted by MetaGeneMark. A majority shared homology with proteins of known function, but 9 were assigned as ‘hypothetical proteins’. These could result from mis-predicted proteins or proteins for which homologs with a known function were not detected. This again highlights the capacity of functional metagenomics to give access to unknown gene families. Interestingly, in this example, 5 ORFs, ORF22 to 26 encoded genes from the riboflavin biosynthesis pathway (represented in pink) [196]. Multiple metabolites produced by this pathway including riboflavin and lumazine are fluorescent compounds having excitation/emission spectra that partially overlap with that of 4-MU in our screen [197, 198] strongly suggesting this particular hit is a false positive.

The DNA sequence of each hit’s insert was determined using next generation sequencing technology. This enabled prediction of ORFs encoded by each fosmid insert and their representation under the form of color-coded

3. Development of a functional metagenomic workflow for enzyme discovery

maps. These maps contain essential information about the orientation, proximity, and possible function of encoded ORFs.

3.4 Chapter 3 conclusion

A functional metagenomics workflow was assembled to enable screening environmental genes for sought enzymatic activities. Each step of the workflow (Figure 5) was implemented during my work at NEB. Key sections of the screening workflow were described and discussed in this chapter and include creation of metagenomics libraries and the NEB Collection, establishing plate- and lysate-based high-throughput enzyme screening assay formats, establishing physical and data processing method to achieve next generation sequencing of fosmid using the PacBio long-read technology.

To exploit the established workflow a collection of metagenomics clones was built. In total, almost 100,000 clones were generated from diverse ecological niches: thermal springs, compost, soil, ocean, pond water or human gastrointestinal microbiome. Single organism genomic libraries were also created to enrich the collection with hyperthermophiles or known species of interest. It is estimated than ~3-4 million genes were represented in the collection as of November 2019. This resource continues to be built upon with 3,072 clones being added by NEB colleagues from November 2019 to March 2021. Stored at -80°C, the metagenomics libraries are stable and will remain available for any future screening projects.

The workflow presented in this chapter is currently in-use at NEB and ongoing efforts are improving it. From a time- and labor-perspective, fosmid sequencing is a limiting step. To shorten the time to go from hit identification to enzyme discovery, all clones from certain libraries in the collection are being fully sequenced, and a searchable sequence repository database has been built. With these new resources, within minutes of visualizing a hit by biochemical assay it will be possible to examine that clone's DNA sequence that already resides in the repository. This will dramatically increase the speed of discovery.

3. Development of a functional metagenomic workflow for enzyme discovery

The remaining chapters of this thesis illustrate projects in which I used this workflow and the fosmid collection. In each project, selected libraries from the collection were screened. The ‘Small Dixie’ library was screened for Neu5Ac sialidases (Chapter 4, section 4.2), the compost library was screened for Neu5Gc sialidases (Chapter 4, section 4.4), and the human gut microbiome library was screened for sugar-specific sulfatases (Chapter 5). The collection has also been successfully used by several other groups at NEB for different screening projects that are beyond the scope of this thesis.

4 Screening metagenomic libraries for sialidases

Please note that parts of this chapter are taken from the original publications:

Chuzel L, Ganatra MB, Rapp E, Henrissat B, Taron CH. Functional metagenomics identifies an exosialidase with an inverting catalytic mechanism that defines a new glycoside hydrolase family (GH156). *Journal of Biological Chemistry*. 2018, 293:18138–18150.

Bule P, **Chuzel L**, Blagova E, Wu L, Gray MA, Henrissat B, Rapp E, Bertozzi CR, Taron CH, Davies GJ. Inverting family GH156 sialidases define an unusual catalytic motif for glycosidase action. *Nature Communications*. 2019, 10:4816.

Zaramela LS, Martino C, Alisson-Silva F, Rees SD, Diaz SL, **Chuzel L**, Ganatra MB, Taron CH, Secrest P, Zuñiga C, Huang J, Siegel D, Chang G, Varki A, Zengler K. Gut bacteria responding to dietary change encode sialidases that exhibit preference for red meat-associated carbohydrates. *Nature Microbiology*. 2019, 4:2082–2089.

4. Screening metagenomic libraries for sialidases

4.1 Introduction to sialic acid biology

Substances referred to as ‘sialic acids’ were originally named from their discovery in salivary proteins. Independently, the same type of substance was later identified in the brain, and because its similarity to sialic acid was initially unknown, it was given the new name ‘neuraminic acid’. Ultimately referring to the same group of molecules, both terms ‘sialic acid’ and ‘neuraminic acid’ spread in the literature and are today interchangeably employed [199].

The family of sialic acids (Sias) is part of a wider family of nonulosonic acids (NulOs). All share the same 9-carbone backbone structure and have an anomeric center at position C2. The axial orientation of the carboxyl group linked to C2 distinguishes the α - from the β -anomer. The α -anomers are represented in Figure 19 and are the common forms of Sias in the bound state when these are attached to other sugar moieties and part of glycan chains. The β -anomer on the other hand, is the favored form of Sia in the unbound state, when it is free in solution [200].

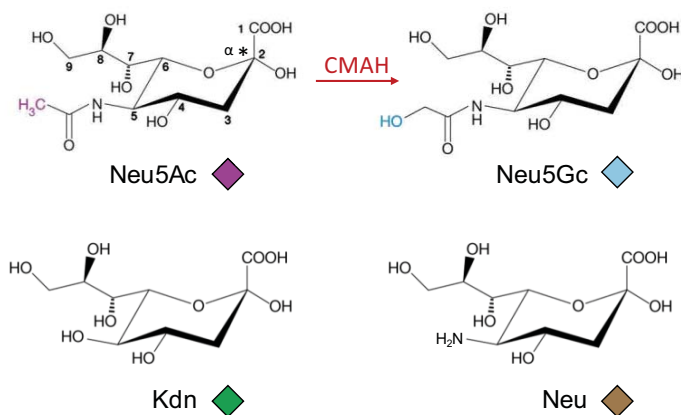


Figure 19. Sialic acid core structures. Adapted from [199]. Carbon numbering is shown for Neu5Ac. Structures are represented in their α -anomeric configuration at C2. The mammalian cytidine monophospho-N-acetylneuraminic acid hydroxylase (CMAH) enzyme is responsible for conversion of Neu5Ac into Neu5Gc.

4. Screening metagenomic libraries for sialidases

The four common forms of sialic acid are N-acetyl neuraminic acid (Neu5Ac), N-glycolylneuraminic acid (Neu5Gc), ketodeoxynonulosonic acid (Kdn), and the free amine form of neuraminic acid (Neu). Their structures are illustrated in Figure 19 along with their corresponding SNFG symbol. Each of these structures can be further modified with different substituents on their hydroxyl groups (*e.g.*, O-acetyl, O-methyl or O-sulfate) resulting in a family of diverse molecules [201]. The bigger family of NulOs from which Sias belong, have a similar organization with a 9-carbon backbone but an N-acetyl group in position C7 in place of the hydroxyl group and no hydroxyl group in C9. All NulOs, including the sub-family of Sias, share the same biosynthesis pathways. Their synthesis is initiated by condensation of a 6-carbon monosaccharide with a pyruvate forming the 9-carbon backbone [199]. Sias are all activated under the form of CMP-sugar, a reaction that occurs in the nucleus in eukaryotes. In eukaryotes, CMP-Sias are then transported into the Golgi where they are used to enable addition of Sias to glycan acceptors by sialyltransferases (see Chapter 2, section 2.2.2).

The diversity of sialoglycans does not simply arise from the different substitutions one can find on the 4 Sia core structures, but also from the different types of linkage via which they are appended to glycan chains. Sias are attached to other sugar moieties via their anomeric carbon C2. This typically occurs with Gal at position C3 or C6, GalNAc at position C6 or other Sias at position C8 or C9, forming the linkages α 2-3, α 2-6, α 2-8 and α 2-9, respectively [199].

Commonly found at the terminal position of *N*-glycans, *O*-glycans and glycosphingolipids, the roles of Sias are intrinsically linked to their recognition by other molecules. Sias are essential to several physiological processes. For instance, E- and P-selectins, localized on endothelial cells interact with sialoglycans on leukocytes. This sia-mediated communication is employed to target circulating leukocytes at inflammatory sites [5, 202]. Other Sia-recognizing molecules (SIGLECs) are found at the surface of various blood cell types and play important roles in immune response modulation [203].

4. Screening metagenomic libraries for sialidases

The exposure of Sias in the outer position on glycoconjugates also makes them a target for pathogens. Many virus-host cell interactions are mediated by sialylated glycans. Influenza virus hemagglutinin (HA) is anchored to the virus envelop and recognizes α 2-6 linked sialic acid on human upper airway cells, mediating virus entry. Interestingly, avian influenza viruses recognize α 2-3 linked sialic acid, the predominant linkage found in bird intestines [204]. Other infectious pathogens rely upon recognition of sialoconjugates such as *Helicobacter pylori* or *Plasmodium falciparum* [205, 206]. In addition to infections, Sias appear to have roles in cancer progression. Certain cancers are associated with hypersialylation. The increase in cell surface Sia-containing molecules is hypothesized to help malignant cells evade the immune system, coated with these ‘healthy self’ signature molecules [207].

In humans, the most abundant form of Sia is Neu5Ac. Neu5Gc is commonly found in mammals, including the close human-related species of bonobos and chimpanzees, but is absent in humans. A deletion in the CMP-N-acetylneuraminic acid hydroxylase (CMAH) gene that converts Neu5Ac into Neu5Gc (Figure 19) is responsible for the absence of Neu5Gc in the human sialome [208]. Sias are used by many pathogens to infect vertebrate by virtue of their terminal localization within glycans, as discussed above. The loss of a functional CMAH presumably provided an evolutionary advantage by preventing infections from certain pathogens for which host interaction depends upon binding to Neu5Gc. Such pathogens may include the parasite *Plasmodium reichenowi* that causes great ape malaria, the *E. coli* strain K99, and simian virus 40 [208]. Small traces of Neu5Gc have been found in humans tissues [209]. It is believed that these Neu5Gc residues come from dietary sources, such as meat and dairy products. Studies have shown that Neu5Gc-containing glycoconjugates can be incorporated into human tissues [210]. Because Neu5Gc is a human xenoantigen, its exposure can initiate an immune response and formation of anti-Neu5Gc antibodies which are detected at varying levels in most healthy humans [126, 211]. The use of animal-derived cell lines to produce pharmaceutical products results in the introduction of Neu5Gc into certain

4. Screening metagenomic libraries for sialidases

glycosylated bioproducts. Detectable levels of Neu5Gc have been found in Cetuximab raising biosafety concerns [127].

Enzymes that catalyze hydrolysis of the glycosidic bond that links a sialic acid to its adjacent sugar are termed sialidases or neuraminidases. Similar to ‘sialic acids’ and ‘neuraminic acids’, both terms are interchangeably used although for viral enzymes the term ‘neuraminidases’ is preferred. Widely distributed in biology, they have been found in animals, fungi, protozoa, bacteria, and many viruses. A database named the Carbohydrate-Active Enzymes Database (CAZy) classifies all enzymes that assemble and deconstruct glycans based on their amino acid sequences [212]. Sialidases have been compiled into four glycoside hydrolase (GH) CAZy families. Families GH33, GH34, and GH83 each contain exosialidases, enzymes that release a terminal sialic acid from oligosaccharides. In addition, family GH58 comprises endosialidases, enzymes that cleave within a polysialic acid chain [213]. Families GH34 and GH83 are exclusive to viral neuraminidases, whereas family GH33 contains all known bacterial and mammalian exosialidases. Despite their lack of primary amino acid sequence similarity, all four of these sialidase families share a common 6-fold-propeller tertiary structure [214–221]. Mechanistically, known exosialidases release sialic acid with overall retention of its anomeric conformation [222–225]. For example, α -Sias at the outer end of glycan chains are released in as an α -anomer. In contrast, endosialidases function via an inverting catalytic mechanism [226] (Figure 20). Upon internal cleavage by endosialidases, chains of α -Sias are reduced to smaller polymers with the Sia that was formerly engaged in the glycosidic linkage being converted from an α to β anomer.

4. Screening metagenomic libraries for sialidases

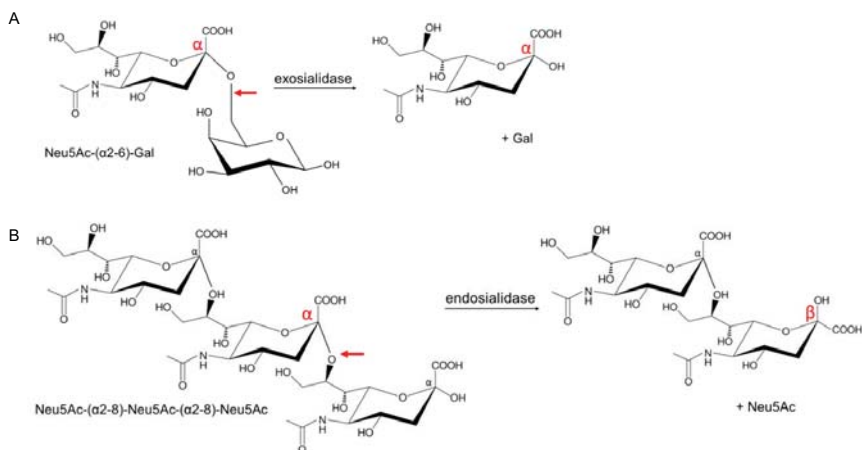


Figure 20. Retaining exosialidase and inverting endosialidases. Exosialidases (A) are retaining glycoside hydrolases, their product conformation is maintained upon catalysis and identical to that of the substrate. Endosialidases (B) are inverting glycoside hydrolases, their product conformation is changed upon hydrolysis from α to β anomer.

Sialidases are important enzymes that play many roles in biology and biotechnology. Viral neuraminidases have been particularly studied, especially that of influenza virus considering their key role in the viral life cycle. Upon the sialic acid-mediated entry into the host cell, influenza virus replicates and newly formed virions bud from the host cell surface to which they adhere via interaction of HA with the host cell sialoconjugates. Detachment of virions from the host cell surface requires the action of a viral neuraminidase that upon cleavage of Sias from the host cell surface disrupt the HA-sialoglycan interaction. Influenza neuraminidase was thus explored as a therapeutic target in regards to its essential role in the virus propagation [227]. This strategy proved effective and two inhibitors of the viral neuraminidase are today commercialized as anti-viral drugs under the names of zanamivir and oseltamivir [228].

4. Screening metagenomic libraries for sialidases

As mentioned above, hypersialylation was reported in several cancers and correlated with poor prognosis and decrease tumor immunogenicity. A proposed explanation is that recognition of sialoconjugates by SIGLECs generates inhibitory signals that enable cancer cells to evade the immune response. The use of sialidase was recently proposed as a promising new cancer immunotherapy strategy termed glycoalyx editing [229]. In this approach, sialidases are coupled to therapeutic monoclonal antibodies enabling precise targeting of the enzyme localization to the cancer cell. The sialidase, by stripping Sias from the cancer cell glycoconjugates, prevents their interaction with inhibitory receptor while enhancing antibody-dependent cell-mediated cytotoxicity (ADCC). Both contribute to increasing the cell's susceptibility to the immune system.

In view of the importance of sialidases to both biology and clinical/pharmaceutical science, I executed two projects aimed at discovery of novel sialidases. In the present chapter, I first report on functional metagenomic screening of a freshwater thermal hot spring fosmid library for sialidase activity. This screen led to the discovery of a novel exosialidase that defines a new CAZy glycoside hydrolase family (GH156) (section 4.2). I show that this new enzyme catalyzes sialic acid hydrolysis via an inverting reaction mechanism, the first example of a wild-type exosialidase with this ability. In collaboration with Prof. Gideon Davies's group, I also undertook determination of the 3D structure of this first member of GH156 and identified its catalytic center and mechanism (section 4.3). I also report functional metagenomic screening of a compost library for sialidases with enhanced activity towards non-human Neu5Gc. This project resulted in the identification of two sialidases with a significant substrate preference for non-human Neu5Gc versus Neu5Ac, a previously undescribed specificity (section 4.4).

4. Screening metagenomic libraries for sialidases

4.2 Discovery of the novel GH156 sialidase family

The sialidase screen reported in this section was one of the first functional metagenomic screens I executed in this thesis work. It was motivated by two desires: i) to look for novel sialidases in a microbial community from an extreme environment (a hot spring), and ii) to test the workflow devised in Chapter 3 and ensure each of its steps were working.

4.2.1 Material and methods

4.2.1.1 Screening for sialidases

The metagenomics library screened was termed ‘Small Dixie’. The ‘Small Dixie’ metagenomics library (see Chapter 3, section 3.3.1.2) was composed of 606 clones. It was assayed for sialidase activity using both agar plate-based and cell lysate-based assays using 5-bromo-4-chloro-3-indolyl- α -D-N-acetylneuraminic acid (X-Neu5Ac) (MilliporeSigma) and 4-methylumbelliferyl- α -D-N-acetylneuraminic acid (4-MU- α -Neu5Ac) (Toronto Research Chemicals, North York, ON), respectively, as described in Chapter 3, section 3.2.4.2.

4.2.1.2 Tn5 mutagenesis

A library of random mutants of fosmid G7 was generated using the EZ-Tn5™ <Kan-2> Insertion Kit (Lucigen Corporation, Middleton, WI) following the manufacturer’s instructions. Briefly, an equimolar amount (0.007 pmol) of the EZ-Tn5™ <KAN-2> transposon and target DNA (G7 fosmid) were used to maximize the insertion efficiency while minimizing multiple insertion events. The 10 μ L reaction was incubated at 37°C for 2 h and stopped upon addition of 1 μ L EZ-Tn5 10X stop solution and incubation at 70°C for 10 min. One microliter of the reaction was added to 50 μ L of thawed electrocompetent cells EC300110 (Lucigen Corporation) in a pre-chilled tube. Electroporation was performed with a Gene Pulser Xcell™ Electroporation System (Bio-Rad, Hercules, CA) following the manufacturer’s instruction (1 mm gap, 1800 V, 25 μ F and 200 Ω). SOC medium (NEB) was added immediately after

4. Screening metagenomic libraries for sialidases

electroporation (950 μL) and the cells were transferred to a 15 mL tube for incubation at 37°C with shaking for 1 h. Transformed cells were plated on LB + 12.5 $\mu\text{g}/\text{mL}$ chloramphenicol and 50 $\mu\text{g}/\text{mL}$ kanamycin. One hundred ninety-two random mutants were arrayed in two 96-well plates. Sialidase activity in both plates was detected using the lysate-based assay with 4-MU- α -Neu5Ac substrate as described in Chapter 3, section 3.2.4.2.

4.2.1.3 *In vitro* and *in vivo* sialidase expression

Sialidase candidates ORF9 and ORF12 were each expressed *in vitro* using the PURExpress® *in Vitro* Protein Synthesis Kit (NEB) following the manufacturer's instruction. PURExpress® DNA templates were generated by PCR using primers specific to ORF9 or ORF12 (Appendix: Supplementary table 1). PCR was performed using 250 ng of fosmid G7 as template, 0.75 μL of each primer (20 μM) and 15 μL Q5 Hot Start High-Fidelity 2X Master Mix (NEB) in a 30 μL total reaction volume. Thermocycling consisted of 25 cycles (98°C for 10 s, 72°C for 30 s, 72°C for 90 s). The amplified product was purified using the Monarch® PCR Clean-up Kit (NEB). For *in vitro* protein synthesis, 1 μL of amplified ORF9 or ORF12 DNA was mixed with 10 μL solution A, 7.5 μL solution B and 0.5 μL RNase Inhibitor Murine (NEB) and incubated for 2 h at 37°C to express the desired protein. Expression was verified by separating 2.5 μL of the reaction on a Novex 10-20% Tris-Glycine gel (Thermo Fisher Scientific, Waltham, MA). Additionally, sialidase activity of *in vitro* expressed proteins was assessed by incubating 20 μL of PURExpress® product with 4 μL of 100 $\mu\text{g}/\text{mL}$ 4-MU- α -Neu5Ac at 37°C for 1 h and reading fluorescence at $\lambda_{\text{ex}}=365$ nm and $\lambda_{\text{em}}=445$ nm in a SpectraMax microplate fluorometer (Molecular Devices, Sunnyvale, CA).

ORF12 was fused with a hexahistidine tag coding sequence at its 3' end (for a C-terminal tag) and cloned into the pJS119K vector [230]. Primers were designed using the NEBuilder assembly online tool (Appendix: Supplementary table 1). Linearization of the vector was performed by PCR with the Q5 Hot Start High-Fidelity 2X Master Mix (NEB) for 25 cycles (98°C for 10 s, 60°C

4. Screening metagenomic libraries for sialidases

for 30 s, 72°C for 2 min). An insert was also prepared by PCR using 25 cycles of 98°C for 10 s, 60.5°C for 30 s, 72°C for 2 min with the Q5 Hot Start High-Fidelity 2X Master Mix. The construct was assembled using the NEBuilder HiFi DNA Assembly Cloning Kit (NEB) and transfected into NEB 5-alpha competent *E. coli* cells following the manufacturer's instructions. The clone was verified by Sanger sequencing and used to transform NEB Express competent cells. Expression under the P_{ta}c promoter was performed at 18°C overnight upon addition of IPTG to 0.4 mM. Cells from 1L of LB culture were harvested by centrifugation at 15,000 x g for 10 min at 4°C. Cell paste was resuspended in 30 mL of 20 mM sodium phosphate, pH 7.4, 500 mM NaCl and 20 mM imidazole buffer and lysed using a TS Benchtop Series cell disruptor (Constant Systems Limited, Daventry, UK) at 32 kPsi. Expressed protein was purified on a 5 mL His-Trap™ FF column (GE Healthcare, Little Chalfont, UK). The bound ORF12-6xHis protein was eluted with 20 mM sodium phosphate, pH 7.4, 500 mM NaCl and 500 mM imidazole by gradient elution over 20 column volumes (from 0 to 100% of elution buffer). Fractions containing pure protein were pooled and dialyzed against 20 mM sodium phosphate, pH 7.4, containing 500 mM NaCl, 1 mM EDTA.

4.2.1.4 Sialidase biochemical characterization

One unit of the ORF12 sialidase was defined as the amount of enzyme required to cleave Neu5Ac from 1 nmol of 4-MU- α -Neu5Ac in 1 h at 37°C in 20 mM sodium phosphate pH 7.4. The effect of pH and temperature on enzyme activity was investigated by incubating 0.5 U of enzyme with 1.7 pmol of 3'-sialyl-N-acetylglucosamine-2-AB (Prozyme, Hayward, CA) for 3 h at 37°C. These conditions intentionally gave incomplete cleavage so that both substrate and product peaks could be monitored. A range of pH from 4.5 to 9.5 (sodium acetate 50 mM buffer for pH between 4.5 and 5.5, sodium phosphate 20 mM buffer for pH between 5.5 and 8.0 and Tris-HCl 50 mM buffer for pH between 8.0 and 9.5) at 37°C and temperature ranges from 15°C to 70°C in sodium acetate 50 mM, pH 5.0 were tested. The effect of metal ions was tested by incubating 0.5 U of enzyme with 1.7 pmol of 3'-sialyl-N-acetylglucosamine-

4. Screening metagenomic libraries for sialidases

2AB in 50 mM sodium acetate, pH 5.0 with 5 mM of NiSO₄, CaCl₂, MnSO₄, MgCl₂, FeSO₄, ZnCl₂ or CuSO₄, respectively, for 3 h at 37°C. After incubation, reactions were dried by vacuum evaporation and resuspended in 1.8 µL water and 13.2 µL acetonitrile for a 12:88 ratio. Twelve microliters were injected into a Waters Acquity BEH glycan amide column (2.1 X 150 mm, 1.7 µm) on a Waters ACQUITY UPLC H-Class instrument (Waters Corporation, Milford, MA) equipped with a quaternary solvent manager and a fluorescence detector. Fifty millimolar ammonium formate buffer pH 4.4 and 100% acetonitrile were used, respectively, as solvent A and B. The gradient used was 0-1.50 min, 12% solvent A; 1.5-35 min, 47% solvent A; 35-36 min, 70% solvent A; 36.5-42 min, 12% solvent A with a flow rate of 0.561 mL/min. Samples were kept at 5°C prior to injection and separation was performed at 30°C. The fluorescence detection wavelengths were $\lambda_{\text{ex}}=330$ nm and $\lambda_{\text{em}}=420$ nm with a data collection rate of 20 Hz. The amount of uncleaved substrate and released product were each analyzed with Empower 3 chromatography workstation software (Waters Corporation). Peak areas were calculated by integration and relative activity was calculated.

Sialidase specificity was assessed by testing its efficacy on 2-AB-labeled 3' and 6'-sialyl-N-acetylglucosamine substrates (termed 3'-SLN-2AB and 6'-SLN-2AB, respectively) (Prozyme, Hayward, CA), di-Neu5Ac-terminated biantennary complex N-glycan (termed G2S2-Ac) (Prozyme, Hayward, CA), di-Neu5Gc-terminated biantennary complex N-glycan (termed G2S2-Gc) (Tokyo Chemical Industry, Japan) and the GD3 ganglioside released glycan (Neu5Ac(α 2-8)Neu5Ac(α 2-3)Gal(β 1-4)Glc-2AB). The released glycan head group substrate from GD3 ganglioside was prepared as follows: glycans from 10 nmol of GD3 ganglioside dissolved in methanol were released with 20 mU of endoglycoceramidase I (EGCase I, NEB) in a 10 µL reaction and incubated for 24 h at 37°C. The reaction mix was passed through a Nanosep 10K Omega centrifugal device (Pall Corporation, Westborough, MA) and centrifuged at 12,000 r.p.m. for 4 min to remove the enzyme. The filtrate was dried in a vacuum evaporator prior to 2-AB labeling for 2 h at 65°C with 10 µL of 2-AB

4. Screening metagenomic libraries for sialidases

labeling mix (350 mM 2-AB, 1 M sodium cyanoborohydride in 7:3 DMSO:acetic acid). Labeled glycans were cleaned-up with a HILIC detergent removal Microspin cartridge (The Nest Group Inc., Southborough, MA).

Enzyme (1 or 10 U) was mixed with each of the substrates in 20 μ L reactions in 50 mM sodium acetate pH 5.0 and incubated at 37°C overnight. Negative and positive control experiments were performed in the same conditions with no enzyme and 1 μ L of α 2-3,6,8,9 Neuraminidase A (NeuA, NEB), respectively. Reactions were dried in a vacuum evaporator and analyzed by UPLC-HILIC-FLR as described above. ORF12p-His activity was also tested on a fetuin *O*-glycan library from Ludger (Oxfordshire, UK) and 2-AB labeled as described above.

To determine the kinetics of the reaction catalyzed by ORF12p-His, 20 U of ORF12p-His were incubated at 37°C with different concentrations of 4-MU- α -Neu5Ac (60 - 280 μ M) in 50 mM sodium acetate, pH 5.0. Each substrate concentration was tested in triplicate. Aliquots (50 μ L) of each reaction were removed at different time points over a 5 min incubation. They were immediately mixed with 50 μ L of 1 M sodium carbonate (pH 10.9) to stop the reaction. Fluorescence was read at $\lambda_{\text{ex}}=365$ nm and $\lambda_{\text{em}}=445$ nm in a SpectraMax microplate fluorometer.

4.2.1.5 NMR spectroscopy

NMR was performed by the Complex Carbohydrate Research Centre (CCRC, Georgia). A solution was prepared in a 5-mm NMR tube with 320 μ L 1.78 mg/mL 4-MU- α -Neu5Ac in D₂O (2 mM final concentration after addition of enzyme), 60 μ L of 200 mM sodium phosphate in D₂O, pH 7.4 (20 mM final concentration after addition of enzyme), and 190 μ L D₂O. The solution was mixed, and the NMR tube placed into a 600 MHz Inova NMR spectrometer (Agilent Technologies, Santa Clara, CA) at 25°C. A 1D proton NMR experiment with water presaturation (using the 2-step purge option) was recorded as the $t=0$ measurement. Then, 30 U ORF12p-His sialidase or 240 U of NeuA (NEB) was added, the NMR tube was inverted several times to ensure

good mixing, and the tube was replaced into the NMR spectrometer. 1D proton spectra with water presaturation were recorded at time intervals with 32 transients each. Chemical shifts were referenced relative to the residual HDO peak, set at 4.78 ppm.

4.2.1.6 *Armatimonadetes* homolog expression

The gene encoding the hypothetical protein OIO94155 from *Armatimonadetes* was synthesized by Genscript (Piscataway, NJ) in pUC57. The gene was sub-cloned into pJS119K using the HiFi DNA Assembly Cloning Kit (NEB) with primers designed with NEBuilder assembly online tool (Appendix: Supplementary table 1). The clone was verified by Sanger sequencing and used to transform NEB Express competent cells. Expression under the Ptac promoter was performed at 30°C for 4 h upon addition of IPTG to 0.4 mM. Crude lysate was prepared by sonication and assayed with 4-MU- α -Neu5Ac by mixing 5 μ L of crude lysate with 4 μ L of 100 μ g/mL 4-MU- α -Neu5Ac and 15 μ L of 20 mM MES buffer pH 6.5. After 1 h incubation at 37°C the fluorescence of the reaction was read at λ_{ex} =365 nm and λ_{em} =445 nm in a SpectraMax microplate fluorometer. Lysate from NEB Express cells containing the empty pJS119K vector and lysate from the pJS119K-ORF12p construct were prepared and assayed in the same conditions to serve as negative and positive controls, respectively.

4.2.2 Results

4.2.2.1 Functional screening

A small metagenomic DNA library, termed ‘Small Dixie’ (see Chapter 3, section 3.3.1.2), was constructed from environmental DNA isolated from hot spring mats collected in Dixie valley, Nevada. Introduced into *E. coli*, 616 individual clones were retrieved and arrayed into two 384-well plates to facilitate enzyme screening. Restriction fragment analysis of 12 randomly selected library clones indicated an average cloned insert size of ~30-40 kb (Figure 21).

4. Screening metagenomic libraries for sialidases

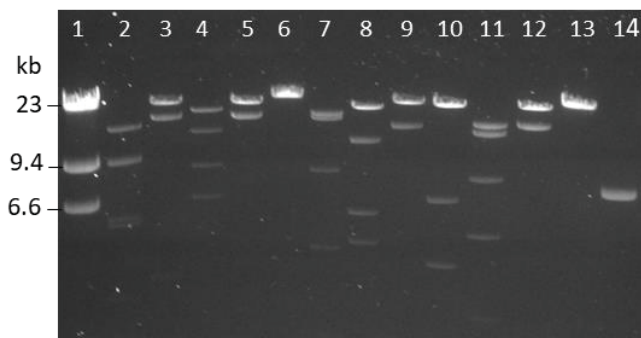


Figure 21. Restriction fragment analysis of the ‘Small Dixie’ metagenomic library. From [191]. Twelve randomly selected clones (lanes 2-13) from the hot spring metagenomic library were isolated and digested with an 8-base cutting endonuclease SbfI. Digested fosmids were separated on a 1% agarose gel along with a λ HindIII size marker (lane 1) and a linearized pSMART FOS empty vector control (lane 14; contains one SbfI site).

Additionally, Sanger sequencing of the ends of each insert revealed that the cloned DNA originated from both known and unknown bacterial species. Amongst the known species were the thermophiles *Caldilinea aerophila* and *Thermomicrobium roseum* [231, 232]. Other clones were related to *Chloroflexus sp.*, *Thermocrinis ruber* or *Acidobacteria bacterium*, and 5 clones contained DNA from unknown origins. Thus, even in a small sampling of randomly isolated clones, the library harbored genetic material from a broad range of microbial diversity.

To identify active sialidases, the arrayed clones were screened with (X-Neu5Ac) and 4-MU- α -Neu5Ac substrates in both agar plate and cell lysate assays, respectively. In the agar plate screen, a single colony (designated G7) hydrolyzed X-Neu5Ac and turned blue after an overnight incubation at 37°C and several days of incubation at 4°C, as shown in Chapter 3 (Figure 13). In the cell lysate screen, the same clone, G7, was identified by measuring an increase in fluorescence due to the hydrolysis of 4-MU- α -Neu5Ac (Figure 22).

4. Screening metagenomic libraries for sialidases

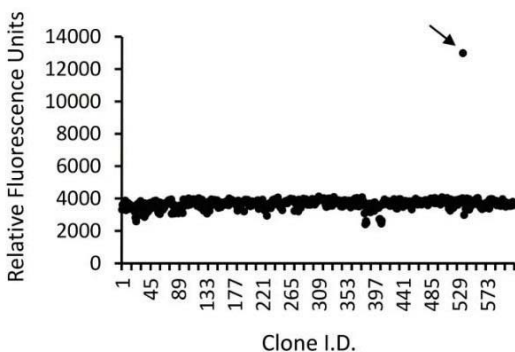


Figure 22. Screening for sialidase activity from a hot spring metagenomic library. From [191]. Lysate from microcultures of *E. coli* cells harboring individual fosmid clones were assayed for sialidase activity with 4-MU- α -Neu5Ac. A single positive clone is denoted with an arrow. (Figure from [191])

4.2.2.2 Sialidase gene identification

To identify the gene responsible for the observed sialidase activity, fosmid G7 was sequenced using the PacBio RS II platform. A total of 44,770 reads were generated with an average insert size of 7,523 bp and an average polymerase read length of 15,285 bp. The reads were assembled with the HGAP.3 algorithm [233] and a single contig was obtained. After removal of the repeat region and the sequence of pSMART, the full insert nucleotide sequence (41,198 bp) was obtained (GenBank accession number MH016668). The sequence coverage was approximately 8,000X. The fosmid insert did not match any DNA sequence in GenBank indicating it was from an unsequenced organism.

The fosmid G7 DNA sequence encoded 40 putative open reading frames (ORFs) that were predicted by MetaGeneMark [234]. BLASTP analysis of the non-redundant GenBank protein sequence database was performed using each predicted ORF protein sequence as a query. Based on the observed homology of each ORF to homologous proteins in GenBank, the 40 ORFs were classified

4. Screening metagenomic libraries for sialidases

into three categories: i) proteins of known function, ii) proteins involved in saccharide utilization, and iii) hypothetical proteins of unknown function (Figure 23). Of the 40 ORFs, four showed similarity to proteins involved in saccharide utilization, and none showed any similarity to proteins from the known CAZY sialidase families GH33, GH34, GH58 and GH83.

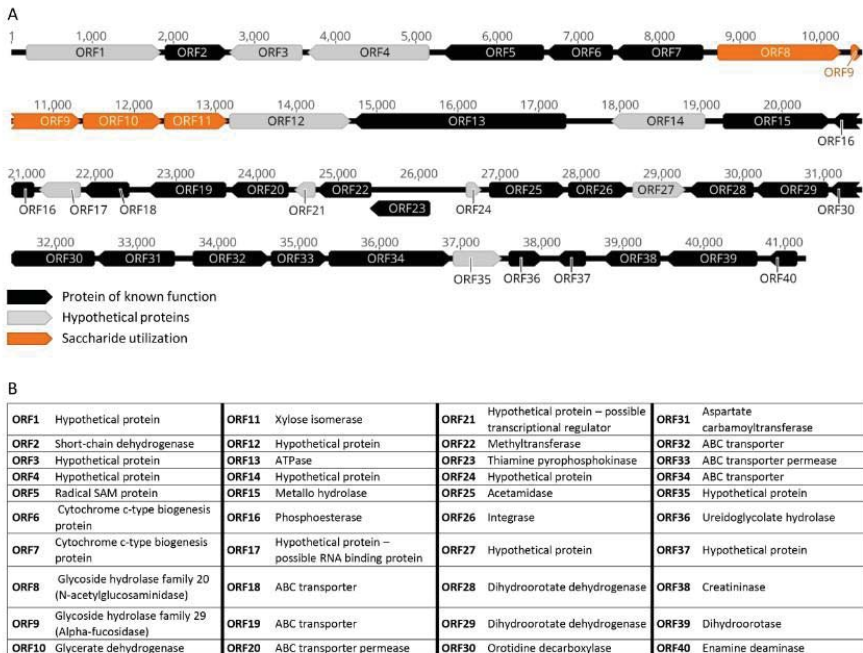


Figure 23. Predicted ORFs encoded in the fosmid G7 nucleotide sequence. From [191]. (A) Fosmid G7 was sequenced on the PacBio RSII platform. ORFs were predicted with MetaGeneMark and classified into three categories based on their homology to proteins from the nonredundant protein database (NCBI). Black, proteins with a known annotated function; gray, “hypothetical proteins” with no annotated function; orange, proteins involved in saccharide utilization. (B) Database annotations for all 40 ORFs.

4. Screening metagenomic libraries for sialidases

As sequence analysis did not identify an obvious sialidase candidate amongst the 40 sequenced ORFs, transposon mutagenesis was performed to disrupt expression of the sialidase activity. Tn5 mutagenesis was used to randomly insert a kanamycin cassette into fosmid G7 under conditions that minimized multiple insertion events. A total of 192 mutants were assayed for sialidase activity using the substrate 4-MU- α -Neu5Ac (Figure 24A). Seventeen mutants with abolished sialidase activity were identified. Each of these mutants was bidirectionally Sanger-sequenced with transposon-specific primers to identify the element's insertion site. Multiple transposon insertion events were observed for 7 mutants, and they were not further investigated. A map of the G7 insertion sites for the remaining 10 mutants is shown in Figure 24B. All loss-of-function insertions clustered to a ~5 kb region of fosmid G7 containing several ORFs with similarity to enzymes that hydrolyze or metabolize sugars (Figure 23 and Figure 24B). Most insertions occurred in ORF9 (annotated as encoding a putative α -L-fucosidase) or ORF12 (annotated as encoding a "protein of unknown function").

4. Screening metagenomic libraries for sialidases

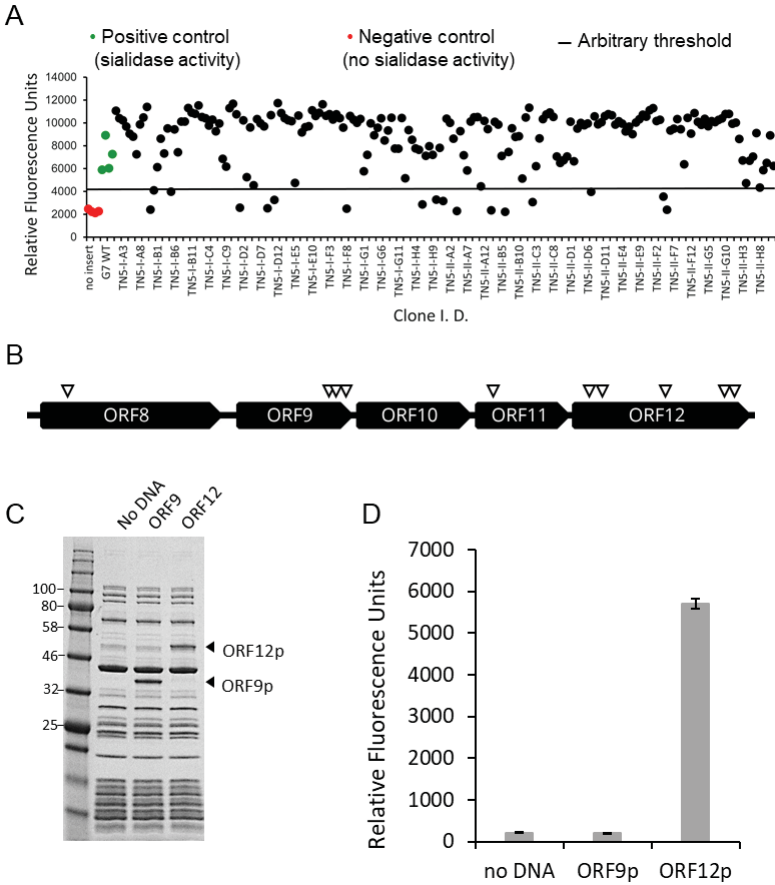


Figure 24. Identification of the sialidase-encoding ORF on fosmid G7 and its in vitro expression. From [191]. (A) Sialidase screening of a library of G7 fosmid Tns5 mutants. Four G7 wild-type clones were also assayed as a positive activity control (green dots) and four no insert fosmid clones (red dots) as a negative activity control. An arbitrary threshold was set to estimate clones with abolished sialidase activity. (B) a map of fosmid G7 transposon insertion sites (white triangles) in mutants with abolished sialidase activity. (C) SDS-PAGE of ORF9 and ORF12 proteins expressed in vitro using the PURExpress® system. D, sialidase activity produced in PURExpress® reaction mixtures was assessed using the substrate 4-MU- α -Neu5Ac.

4. Screening metagenomic libraries for sialidases

To determine if either ORF9 or ORF12 encoded a protein with sialidase activity, each was expressed using the PURExpress® *in vitro* protein synthesis system (Figure 24C) and tested for its ability to cleave 4-MU- α -Neu5Ac (Figure 24D). The expressed product from ORF12 (ORF12p) generated a fluorescence signal ~27 times higher than the no DNA control reaction (the PURExpress® mix alone). Expression of ORF9 yielded no sialidase activity. As the deduced protein sequence of ORF9 (ORF9p) had sequence homology to α -L-fucosidases, the ability of *in vitro* expressed ORF9p to cleave the chromogenic fucosidase substrate 2-chloro-4-nitrophenyl α -L-fucopyranoside (CNP-fucose) was also tested. Light absorbance at 405 nm was ~6 times higher for ORF9 than the negative control, confirming the ORF9 encoded an α -fucosidase, not an exosialidase. Together, these data support the conclusion that the gene product from ORF12 (Figure 24, GenBank accession number MH016668) is solely responsible for the hydrolysis of 4-MU- α -Neu5Ac observed during primary screening. Furthermore, as ORFs 8-12 are co-directionally oriented and are each related in some way to saccharide metabolism, it is plausible that these genes are co-transcribed in *E. coli*. This could account for the observation of reduced sialidase activity upon insertion of Tn5 into multiple positions in this region.

4.2.2.3 Sialidase biochemical characterization

To facilitate biochemical characterization of ORF12p, an expression DNA construct encoding ORF12p with a C-terminal hexahistidine tag (ORF12p-His) was assembled. ORF12p-His was expressed in *E. coli* and purified using nickel affinity chromatography (Figure 25A).

4. Screening metagenomic libraries for sialidases

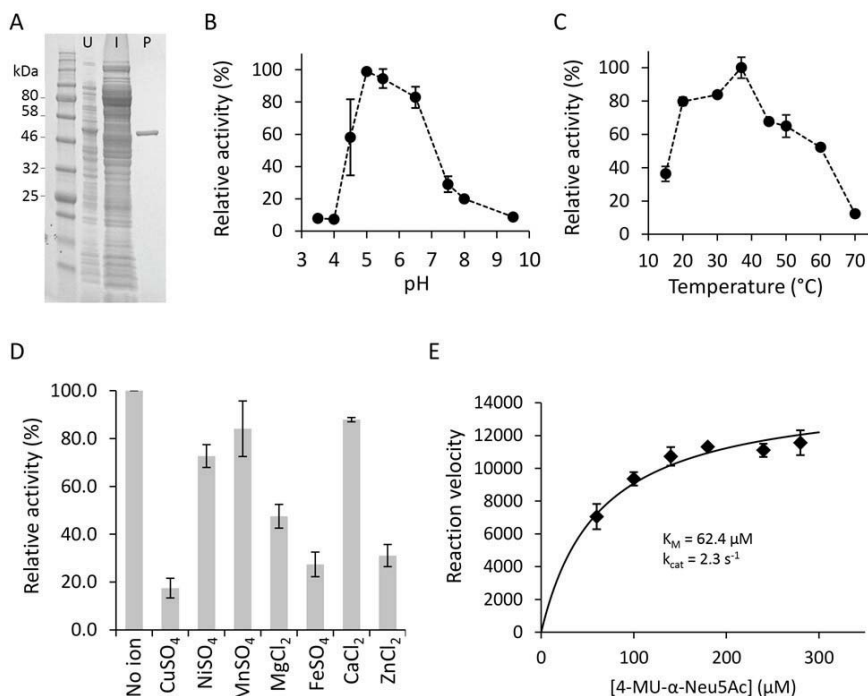


Figure 25. Purification and biochemical characterization of recombinant ORF12p. From [191]. (A) His-tagged ORF12p sialidase was expressed in *E. coli* and purified using a His-trap column as described under “Experimental procedures.” Shown is SDS-PAGE separation of lysates from uninduced cells (lane U), induced cells (lane I), and nickel-purified ORF12p-His (lane P). Purified ORF12p-His was used to determine the pH (B) and temperature (C) optima of ORF12p-His and the effect of metal ions on its catalysis (D). In these experiments, reactions were performed in triplicate using the substrate 3'-sialyl-N-acetylactosamine-2AB. Reaction products were analyzed by UPLC-HILIC-FLR and quantitated by peak integration. (E) Michaelis-Menten plot of ORF12p catalyzed hydrolysis of 4-MU- α -Neu5Ac. The initial velocity was determined in triplicate for each 4-MU- α -Neu5Ac concentration.

The substrate 3'-SLN-2AB was used to determine the optimal pH and temperature of the ORF12 sialidase as described in section 4.2.1.4. A pH

4. Screening metagenomic libraries for sialidases

optimum of 5.0-5.5 was observed (Figure 25B). The enzyme showed optimal activity at 37°C and retained at least 50% activity at 60°C (Figure 25C). No metal ions were required for activity (Figure 25D). Reaction kinetics were determined using the substrate 4-MU- α -Neu5Ac. The initial reaction rates were determined at several substrate concentrations (Figure 25D). The K_M was calculated to be 62 μ M and the k_{cat} 2.3 s^{-1} .

The specificity of purified ORF12p-His was assayed using 3'-SLN-2AB, 6'-SLN-2AB, a GD3 ganglioside head group glycan (Neu5Ac(α 2-8)Neu5Ac(α 2-3)Gal(β 1-4)Glc-2AB) and G2S2 bi-antennary *N*-glycans with two terminal Neu5Gc or Neu5Ac (G2S2-Ac or G2S2-Gc).

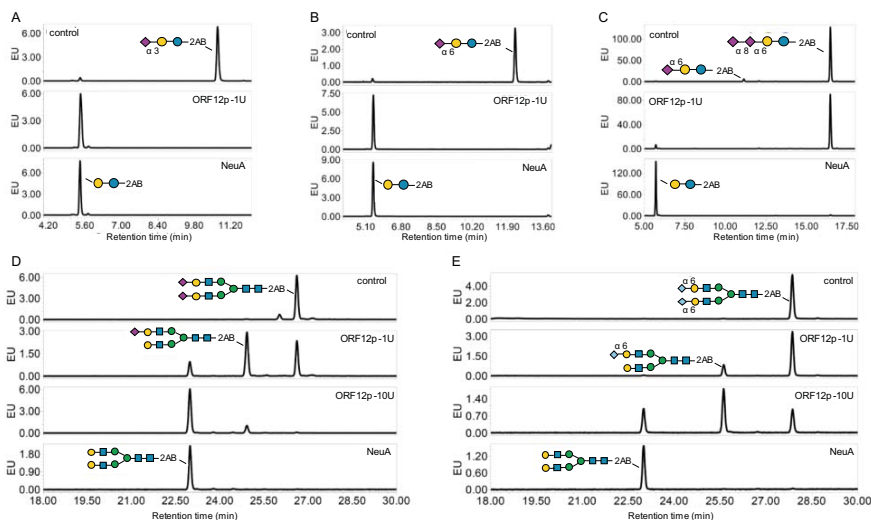


Figure 26. Specificity of ORF12p on sialic acid containing substrates using UPLC-HILIC-FLR analysis. From [191]. (A and B) The ability of ORF12p to cleave the fluorescently labeled substrates 3'- or 6'-sialyl-*N*-acetylglucosamine-2AB. Undigested substrates 3'- or 6'-sialylglucosamine-2AB run at ~10.6- or 12.3-min retention times, respectively (A and B, top panels). Control digestion with the NeuA sialidase shifted both substrate peaks to ~5.5-min retention time (A and B, bottom panels). Digestion of these substrates with 1 unit of ORF12p-His resulted in the same peak shift (A and B, middle panels). (C) ORF12p's ability to hydrolyze α 2-8 Neu5Ac was assessed using a 2AB-labeled GD3 ganglioside headgroup substrate that

4. Screening metagenomic libraries for sialidases

contains two sialic acid residues linked via an α 2–8 bond. Undigested substrate ran at \sim 16.5-min retention time with a very minor peak at \sim 11-min retention time corresponding to partially degraded substrate comprised of a single α 2–6 terminal sialic acid (C, top panel). NeuA-treated substrate shifted at \sim 5.5 min retention time (C, bottom panel). Treatment with 1 unit of ORF12p did not shift the major substrate peak (C, middle panel). (D and E) Activity of ORF12p on biantennary complex N-glycans with terminal sialic acid residues (Neu5Ac, D; or Neu5Gc, E). Undigested substrates run at \sim 26.6- and 27.9-min retention time, respectively (D and E, top panels). NeuA treatment shifted both substrate peaks at \sim 23-min retention time (D and E, bottom panels). Incubation of the substrates with 1 or 10 units of ORF12p resulted in the same peak shift, but incomplete substrate desialylation was observed resulting in another smaller peak shift at \sim 24.9- and 25.6-min retention time, respectively (D and E, middle panels). Symbolic representation of glycan structures was drawn following the SNFG [235]. Abbreviation: EU, emission units.

Digestion products were analyzed by UPLC-HILIC-FLR (Figure 26). Negative control reactions consisting of each substrate and no enzyme yielded a single major peak for all 5 substrates (Figure 26A, B, C, D & E top panel). A positive control reaction using commercial α 2-3,6,8,9 Neuraminidase A (NeuA) resulted in a shift of each substrate, illustrating complete removal of α 2-3, α 2-6 and α 2-8 terminal sialic acid (Figure 26A, B, C, D & E, bottom panel). This same peak shift was observed for 3'-SNL-2AB and 6'-SNL-2AB substrates when incubated with purified ORF12p-His, indicating that this enzyme was able to hydrolyze both terminal α 2-3 and α 2-6 Neu5Ac (Figure 26A & B, middle panel) and confirming the ability of ORF12p-His to function as an exosialidase. The peak corresponding to the released glycan from GD3 ganglioside did not shift upon incubation with ORF12p-His confirming the enzyme's inability to hydrolyze α 2-8 linked sialic acid (Figure 26C, middle panel). N-Glycan substrates G2S2-Ac and G2S2-Gc were both hydrolyzed by ORF12p-His indicating that this enzyme can hydrolyze both terminal Neu5Ac and Neu5Gc from complex N-glycan structures (Figure 26D & E, middle panels). The enzyme showed a preference for Neu5Ac. In addition, the enzyme was assayed on a 2-AB-labeled O-glycan library (Figure 27) and showed the ability to hydrolyze terminal sialic acid linked to N-acetylgalactosamine or galactose residues.

4. Screening metagenomic libraries for sialidases

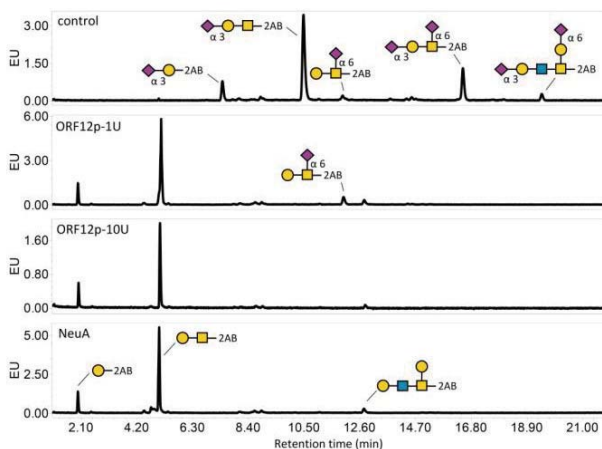


Figure 27. ORF12p activity on a fetuin 2AB-labeled O-glycan library. From [191]. The ability of ORF12p to cleave 3' or 6' terminal sialic acid linked to galactose or N-acetylgalactosamine residue was monitored by HILIC-UPLC-FLR. A library of undigested 2AB-labeled O-glycan from bovine fetuin showed 5 major structures at ~7.4, 10.5, 12, 16.5 and 19.5 min retention time (top panel). Neuraminidase A (NeuA) treatment shifted all 5 substrate peaks to 3 peaks at ~2, 5 and 12.8 min retention time (bottom panel). Incubation of the substrates with 10 U of purified ORF12 sialidase resulted in the same shifts (middle panel). Symbolic representation of glycan structures was drawn following the the SNFG [235]. Abbreviation: EU, emission units.

4.2.2.4 ORF12p sialidase reaction mechanism

Glycoside hydrolases are grouped into two mechanistic classes: inverting or retaining enzymes, depending on whether the stereochemistry of the substrate's anomeric carbon is inverted or retained upon hydrolysis (Figure 20). To date, known exosialidases function as retaining glycoside hydrolases, as they hydrolyze terminal Neu5Ac with net retention of its overall anomeric configuration [222–225].

4. Screening metagenomic libraries for sialidases

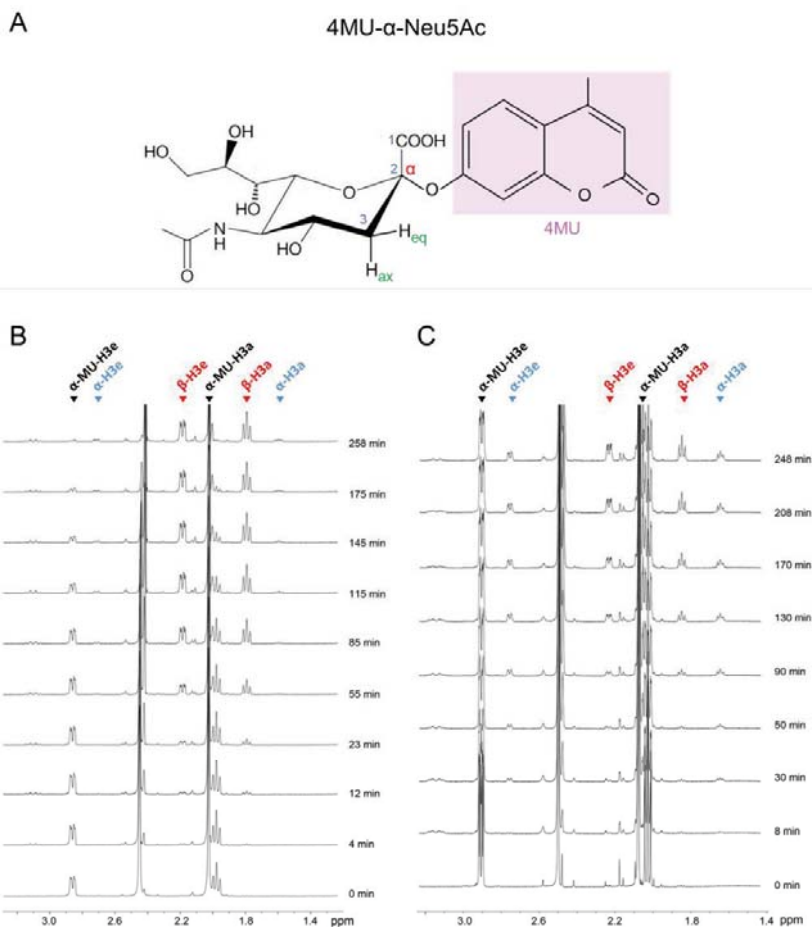


Figure 28. Stereochemical course of the hydrolysis reaction catalyzed by ORF12p (B) compared with NeuA (C). Adapted from [191]. One-dimensional proton NMR was used to monitor the reaction products formed over time upon hydrolysis of 4-MU- α -Neu5Ac by ORF12p and NeuA. (A) Structure of 4-MU- α -Neu5Ac, in green are represented the axial (ax) and equatorial (eq) protons from the position C₃ which were monitored during ¹H-NMR. (B) The up-field region of the NMR spectra showed two groups of peaks at 1.99 and 2.86 ppm corresponding to the substrate H_{3a} (axial) and H_{3e} (equatorial) of 4-MU- α -Neu5Ac (black triangles). Over time, groups of peaks appeared at 1.81 and 2.20 ppm corresponding to H_{3a}

4. Screening metagenomic libraries for sialidases

and H3e of released β -Neu5Ac (red triangles). After 1 h, two sets of signals at 1.61 and 2.72 ppm appear as a result of spontaneous mutarotation of β -Neu5Ac to α -Neu5Ac (blue triangles). (C) The spectra showed the substrate 4-MU- α -Neu5Ac groups of peaks at 2.03 and 2.89 ppm (black triangles). After 30 min, H3a and H3e peaks corresponding to the α -anomer appeared at 1.64 and 2.76 ppm (blue triangles). These were quickly converted by mutarotation to the β -anomer at 1.85 and 2.23 ppm (red triangles).

Proton nuclear magnetic resonance (^1H NMR) spectroscopy was used to determine if ORF12p functions via a retaining or inverting mechanism. To that end, purified ORF12p-His was sent to the CCRC (Georgia, USA) to conduct the ^1H NMR experiment. Enzyme was incubated with the 4-MU- α -Neu5Ac substrate. Data analysis focused on the up-field region of the NMR spectra, where signals from the axial and equatorial protons in the 3-position of Neu5Ac appear. The initial NMR spectrum (Figure 28A) showed only the axial (α -MU-H3a) and equatorial (α -MU-H3e) protons from the substrate 4-MU- α -Neu5Ac at 1.99 and 2.86 ppm, respectively. The first additional set of signals (H3a and H3e) appeared after only a few minutes at 1.81 and 2.20 ppm. These signals corresponded to β -Neu5Ac that had been released from the substrate by the enzyme. The signal set of H3a and H3e corresponding to α -Neu5Ac (at 1.61 and 2.72 ppm) appeared after about 1 h and remained very small throughout the remaining reaction. This reflects spontaneous mutarotation of the β -Neu5Ac product to its α conformer and is not a result of the action of the sialidase. These data unambiguously show that the ORF12 sialidase initially produces β -NeuAc from 4-MU- α -Neu5Ac and supports the conclusion that the enzyme functions via an inverting mechanism.

A control experiment was run with purified NeuA, a known retaining exosialidase from GH33 family [236]. The initial NMR spectrum (Figure 28B) showed the axial (H3a) and equatorial (H3e) protons corresponding to 4-MU- α -Neu5Ac at 2.03 and 2.89 ppm, respectively. Upon initiation of the reaction, the first additional signals appeared after about 30 minutes and corresponded to the α -anomer (1.64 and 2.76 ppm). These peaks were initially larger than the H3a and H3e β -Neu5Ac peaks (1.85 ppm and 2.03, respectively). After 130

4. Screening metagenomic libraries for sialidases

min, the peaks of the β -anomer became more prominent than those of the α -anomer, consistent with the spontaneous mutarotation of the α -Neu5Ac product to its β -Neu5Ac anomer, its preferred configuration in solution. This was previously reported to reach an equilibrium mixture of 92.1% of β -anomer and 7.5% of α -anomer and three open-chain species at pH 8.0 [200]. This stereochemical pattern of product release is characteristic of retaining exosialidases and is markedly different than that of ORF12p-His, further supporting our conclusion that ORF12p-His functions via an inverting mechanism of hydrolysis.

4.2.2.5 ORF12p sialidase family phylogeny

ORF12p homologs were identified in the GenBank sequence repository using the BLASTP algorithm. The 38 nearest homologs of ORF12 were retrieved and used for study of ORF12p phylogeny. Signal peptides that were present in many sequences were trimmed along with C-terminal extensions present for two homologs (OGV64360|*Lentisphaerae* and KXX35835|*Omnitrophica*). Trimmed sequences ranging from 477 to 578 amino acids were then aligned with Muscle and a phylogenetic tree was made using BLOSSUM62 substitution matrix and the Neighbor-Joining method (Figure 29A). It is noteworthy that all 38 ORF12p homologs showed no similarity to known CAZy glycoside hydrolases or other proteins of known function, and all were annotated as hypothetical proteins.

4. Screening metagenomic libraries for sialidases

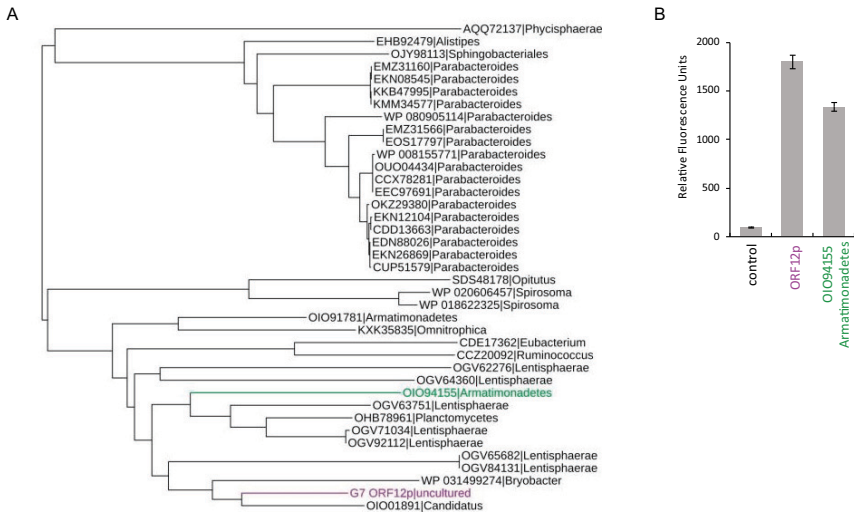


Figure 29. ORF12 protein family. Adapted from [191]. (A) ORF12p homologs were identified with the BLASTP algorithm and aligned with Muscle. A phylogenetic tree was then generated with the BLOSSUM62 matrix and the neighbor-joining method. ORF12p and Armatimonadetes OIO94155 are shown in purple and green text, respectively. (B) the Armatimonadetes OIO94155 protein homolog was expressed in *E. coli* and crude lysate was tested with 4-MU- α -Neu5Ac.

All sequences comprising the ORF12p family originated from bacteria and were distributed within 8 prokaryotic phyla. Most sequences originated from organisms in the phylum Bacteroidetes or the superphylum PVC that contains the phyla Planctomycetes, Verrucomicrobia, Omnitrophica and Lentisphaerae. Other sequences were from organisms from the phyla Firmicutes, Armatimonadetes and Acidobacteria. Based on our phylogenetic analysis, ORF12p was most closely related to sequences from organisms of the PVC superphylum. Notably, no homolog was found within the phylum Proteobacteria that contains most of the well-studied bacteria, or in the phyla Actinobacteria or Cyanobacteria that contains many species widely represented

4. Screening metagenomic libraries for sialidases

in soil and aquatic environments. Similarly, no eukaryotic or archaeal proteins were found to have homology with ORF12p.

To determine if a second member of the ORF12p sequence family also had sialidase activity, the homologous protein OIO94155 from Armatimonadetes was expressed in *E. coli* and crude lysate was assayed with 4-MU- α -Neu5Ac. Lysate from *E. coli* expressing ORF12p was also assayed. A high fluorescence signal was detected for both ORF12p and OIO94155 lysates indicating that OIO94155 is also an exosialidase (Figure 29B).

The nucleotide sequence encoding ORF12p was also unique and showed no significant match to DNA sequences from known organisms. On the cloned fosmid DNA fragment, the gene encoding ORF12p resided in a gene cluster that also encoded a GH29 family α -fucosidase and a GH20 family N-acetylglucosaminidase (Figure 23). While this locus did not conform to the currently defined structural features of a polysaccharide utilization locus (PUL) [237, 238], the proximity of these genes and broad specificity of ORF12p suggests that these proteins may function in a coordinated manner, perhaps in degradation of eukaryotic complex *N*-glycans.

4.2.3 Discussion

We utilized functional metagenomic screening to search for environmental enzymes with the ability to hydrolyze the fluorogenic sialidase substrate 4-MU- α -Neu5Ac. Our screen identified a 41 kb environmental DNA fragment from a hot spring metagenomic library that produced sialidase activity in *E. coli*. The DNA fragment originated from an unknown microorganism and encoded no ORFs with homology to known sialidase families. Tn5 mutagenesis was used to identify a 505-amino acid ORF responsible for the observed sialidase activity (ORF12). We undertook the phylogenetic analysis of the ORF12 protein (ORF12p) family and biochemical characterization of this new sialidase. Glycoside hydrolases and other enzymes that assemble or modify oligo- and polysaccharides are classified into families based on their amino acid sequence similarity [212]. These enzymes are catalogued by family in the curated CAZy

4. Screening metagenomic libraries for sialidases

database. Currently, there are three distinct CAZy families of exosialidases; GH33, GH34, and GH83. The ORF12p sialidase identified in this study showed no homology to any members of these exosialidase families or to members of GH58, a family of endosialidases. Searches of public sequence repositories revealed that ORF12p was a member of a small family of proteins that are distributed only amongst bacteria from approximately eight phyla, several of which have been found in aquatic environments. No ORF12p family members were found in the most common bacterial phyla (*i.e.*, Proteobacteria), archaea, or eukaryotes. Thus, this new sialidase family appears to be highly niche-specific in bacterial biology. The ORF12 protein sequence family defines a novel CAZy glycoside hydrolase family that has been designated GH156. Biochemical characterization of ORF12p revealed both functional similarities and differences with known sialidases. ORF12p was capable of removing $\alpha(2,3)$ - and $\alpha(2,6)$ -Neu5Ac from the terminal position of oligosaccharides, indicating that it functions as an exosialidase. Despite lacking protein sequence similarity with known exosialidases, ORF12p shares similar basic biochemical properties with them [239–241]. For example, ORF12p was broadly active from pH 4.5 to 8.0 (optimal at pH 5.0) and from 20 to 60°C (optimal at 37°C) and did not require metal ions, all features consistent with enzymes from the GH33, GH34, and GH83 families. Additionally, its reaction kinetics (K_M 62 μ M; k_{cat} 2.3 s^{-1}) were similar to reported values for other enzymes from the GH33 family of bacterial exosialidases (Figure 25) [242–244].

A key difference between ORF12p and known exosialidases was its mechanism of catalysis. Glycoside hydrolases are grouped into two main mechanistic classes: inverting or retaining enzymes, depending on whether the stereochemistry of the substrate's anomeric carbon is inverted or retained upon hydrolysis. Our proton 1H NMR data unambiguously showed that ORF12p liberated β -Neu5Ac as the primary product of hydrolysis of the substrate (4-MU- α -Neu5Ac) (Figure 28), indicating that it naturally functions via an inverting catalytic mechanism. This observation is unprecedented as all wild-type (WT) exosialidases that have been described to date (from

4. Screening metagenomic libraries for sialidases

prokaryotes, eukaryotes, and viruses) hydrolyze terminal Neu5Ac with retention of its anomeric configuration [222–225]. However, Watson *et al* [245] showed that mutations introduced into the active site of one bacterial sialidase could induce a change in catalytic mechanism from retaining to inverting. The authors presented a model whereby water can more easily access the active site and directly acts as a nucleophile.

In summary, we have identified a novel exosialidase: EnvSia156 using functional screening of a hot spring metagenomic library in *E. coli*. This enzyme is the first member of the novel CAZy family GH156 and is the first reported WT exosialidase to function via an inverting mechanism. This study also highlights the benefits of using function-based screening to identify unique new carbohydrate hydrolases within interesting ecological niches.

4.3 Biostructural characterization of GH156

To understand how ORF12p achieves inverting catalysis, a deeper exploration of its reaction mechanism was needed. To this end, I collaborated with Prof. Gideon Davies' group from the University of York to resolve the crystal structure of ORF12p and characterize the molecular determinants of its substrate recognition and catalysis. In this study, ORF12p was termed EnvSia156. This study yielded the first 3D structure for a member of GH156 using x-ray crystallography. In this collaboration, I expressed and purified the WT and selenomethionine-labeled EnvSia156, generated the site-specific EnvSia156 mutants, and assayed their activity. SEC-MALLS analysis, preparation of the protein crystals, solving of the protein structure and proposal of a catalytic mechanism was performed by Prof. Gideon Davies' group members.

4.3.1 Material and methods

4.3.1.1 Selenomethionine protein labeling

Recombinant sialidase was labeled with selenomethionine (SeMet) in the *E. coli* auxotrophic T7 Express Crystal strain background (New England

4. Screening metagenomic libraries for sialidases

Biolabs). The enzyme was expressed from a total of 12 L of culture as follows: a starter culture was grown at 37°C with shaking at 250 r.p.m. for 18 h in minimal medium (1X M9 salts [246], 0.4% glucose, 0.1 mM CaCl₂, 2 mM MgSO₄, 0.0002% ferric ammonium citrate, 50 µg/mL kanamycin). Twelve flasks of 1 L minimal medium supplemented with 50 µg/ml L-methionine were inoculated with 10 mL of starter culture each and grown at 37°C with shaking at 250 r.p.m. until the optical density reached ~0.6-0.8. Cells were pelleted at 4500 x g for 15 min at 4°C and resuspended in pre-warmed minimal medium lacking methionine or SeMet. Resuspended cultures were grown for 2.5 h at 37°C with shaking at 250 r.p.m. to deplete cells of methionine at which point SeMet was added to the cultures to a final concentration of 50 µg/mL. Cultures were grown with shaking for 30 min at 37°C, after which EnvSia156 sialidase expression was induced by addition of IPTG to 0.4 mM and continued shaking at 37°C for 4 h. SeMet-labeled cells were harvested by centrifugation at 4500 x g for 25 min at 4°C.

Combined cell paste from 12 L of culture was resuspended in 240 mL of 20 mM sodium phosphate, pH 7.4, 500 mM NaCl, 20 mM imidazole buffer, and lysed using a TS Benchtop Series cell disruptor at 32 kPsi. Purification of the SeMet labeled EnvSia156-6His was performed in two steps. First, EnvSia156-6His was bound to a 5 mL His-Trap™ FF column (GE Healthcare, Little Chalfont, UK) and eluted with a 100 mL 0 to 500 mM imidazole gradient (in 20 mM sodium phosphate, pH 7.4, 500 mM NaCl). Fractions of 3 mL containing partially pure EnvSia156 were pooled and the 63 mL resulting solution was concentrated to 2 mL using Vivaspin 20, 30,000 molecular weight cut off tubes (Sartorius, Göttingen, Germany). Second, gel filtration chromatography was performed using a HiPrep 16/60 Sephacryl S-200 High Resolution column (GE Healthcare, Chicago, IL) and 50 mM Tris, pH 8.5, 200 mM NaCl buffer. The sample was prepared by diluting 1 mL of concentrated EnvSia156-6His with 1 mL of the column buffer and injecting it onto the column with a 2 mL sample loop.

4. Screening metagenomic libraries for sialidases

4.3.1.2 SEC-MALLS

Size Exclusion Chromatography-Multi-Angle Laser Light Scattering (SEC-MALLS) experiments were run in 20 mM HEPES 7.4, 300 mM NaCl buffer. The injected sample comprised 100 μ L of EnvSia156 at 1.8 mg/mL in 20 mM HEPES pH 7.4, 100 mM NaCl, 1 mM DTT. Experiments were conducted on a system comprising a Superdex 200 10/30 GL (GE Healthcare) size exclusion chromatography column, a Wyatt HELEOS-II multi-angle light scattering detector and a Wyatt rEX refractive index detector linked to a Shimadzu HPLC system (SPD-20A UV detector, LC20-AD isocratic pump system, DGU-20A3 degasser, and SIL-20A autosampler). Work was conducted at room temperature ($20 \pm 2^\circ\text{C}$). All solvents and buffers were 0.2 μ m filtered before use, and a further 0.1 μ m filter was present in the flow path. Shimadzu LC Solutions software was used to control the HPLC and Astra V software for the HELEOS-II and rEX detectors. All data were analyzed using the Astra V software. Molecular masses were estimated using the Zimm fit method with degree 1. A value of 0.16 mL/g was used for protein refractive index increment (dn/dc), after calibration with a 2.5 mg/mL sample of BSA.

4.3.1.3 Crystallization of EnvSia156 and EnvSia156 substrate and inhibitor complexes

EnvSia156 at 14.6 mg/mL was tested against a range of commercial crystallization screens with the addition of 5 mM Tris(2-carboxyethyl)phosphine (TCEP) to every condition. The first crystals found consisted in stacks of plates and were used to create a seed stock. Optimization screens were set with and without seeds. Well-diffracting single crystals were obtained by the sitting drop diffusion method at 293 K with 0.8 M sodium formate, 15% PEG 4000, 0.1 M sodium acetate, pH 6.0 and with 0.2 M MgCl_2 15% PEG 4000, 0.1 M sodium citrate, pH 5.6 using a protein per well ratio of 1:1. SeMet-labeled EnvSia156 at 21 mg/mL was crystallized using a similar protocol. The best diffracting crystals were obtained in 0.2 M lithium sulfate, 15% PEG 4000, 0.1 M sodium acetate, pH 6.0 using a protein per well solution

4. Screening metagenomic libraries for sialidases

ratio of 1:1. Crystals were cryoprotected in well solution with 25% glycerol and flash cooled in liquid N₂ for data collection. Protein concentration was determined in an Epoch microplate spectrophotometer (BioTek, Winooski, VT), using a calculated extinction coefficient of 124,705 M⁻¹.cm⁻¹.

EnvSia156 complexes were generated by co-crystallization with 20 mM Neu5Ac, Neu5Gc, 2-deoxy-2,3-dehydro-N-acetylneuraminic acid (DANA), or Kdn. Initial screens and optimizations were performed using the same protocol described for the apocrystals. Ligands were introduced by addition of 0.2 μL of 100 mM stock solution in water to a 0.8 μL drop with a protein/well solution ratio of 1/1. Well-diffracting single crystals were obtained with 0.2 M potassium bromide, 15% PEG 4000, 0.1 M sodium acetate, pH 6.0 (DANA), 0.15 M potassium thiocyanate, 20% PEG 1500, 0.1 M sodium acetate, pH 6.0 (Kdn), 0.8 M sodium formate, 12% PEG 4000, 0.1 M sodium acetate, pH 6.0 (Neu5Ac and Neu5Gc).

4.3.1.4 3D structure solution

The EnvSia156 SeMet derivative crystal structure was solved by multiple anomalous dispersion (MAD). Data were collected at beamline I04 of Diamond Light Source at the selenium peak, inflection, and high-energy remote wavelengths (0.97946, 0.97965, and 0.97873 Å), processed with DIALS [247], reduced with Aimless [248] and phased with the Crank2 pipeline [249]. The initial model was subsequently subjected to alternating rounds of manual building and refinement with Coot [250] and REFMAC5 [251]. The SeMet model was then used to solve the native apo and ligand crystal structures by molecular replacement. Data for the apo, NeuGc, and Neu5Ac complex crystals were collected at beamline I04 of the Diamond Light Source, while data for the remaining crystals were collected at beamline I04-1 of the same synchrotron. All datasets were processed with DIALS, reduced with Aimless and phased with PhaserMR [252]. The obtained models were refined as described above. Ligand coordinates were built using JLigand [253]. wwPDB Validation Service [254]

4. Screening metagenomic libraries for sialidases

was used to validate the structures before deposition in the PDB. 3D structure figures were generated using UCSF Chimera [255].

4.3.1.5 Site directed mutagenesis and activity assay of generated mutants

To identify potential catalytic residues, D14A, D14N, H134A and H134N mutants were generated by site-directed mutagenesis on the pJS119K-EnvSia156 construct using the Q5® Site-Directed Mutagenesis Kit (NEB) and primers shown in Supplementary table 1 (Appendix).

Activity of the mutants D14A, D14N, H134A and H134N was assayed on 4-MU- α -Neu5Ac and compared to wild type EnvSia156. For each mutant protein and wild type EnvSia156, 0.3 μ g of protein in 20 μ L 50 mM sodium acetate pH 5.0 was mixed with 4 μ L of 100 μ g/mL 4-MU- α -Neu5Ac. Reactions were performed in triplicate and incubated at 37°C prior measuring fluorescence at $\lambda_{\text{ex}}=365$, $\lambda_{\text{em}}=445$ nm.

The enzymatic activity of wild type EnvSia156 and the H134A mutant were also assayed on procainamide labeled 3'sialyllactose by incubating 0.05, 0.15, 0.3, and 0.6 μ g of enzyme in 50 mM sodium acetate pH 5.5 with the substrate for 1 h at 37 °C. After incubation, 2.4 μ L of the reaction was mixed with 17.6 μ L acetonitrile for a 12:88 ratio. Sixteen microliters were injected into a Waters Acquity BEH glycan amide column (2.1 \times 150 mm, 1.7 μ m) on a Waters ACQUITY UPLC H-Class instrument (Waters Corporation, Milford, MA) equipped with a quaternary solvent manager and a fluorescence detector using the same solvents, gradient and temperatures as described in section 4.2.1.4. The fluorescence detection wavelengths were $\lambda_{\text{ex}}=308$ nm and $\lambda_{\text{em}}=359$ nm with a data collection rate of 20 Hz. Data were analyzed with Empower 3 chromatography workstation software (Waters Corporation).

4.3.2 Results

4.3.2.1 Expression and purification of EnvSia156 SeMet mutant for MAD

The three major experimental methods for solving the 3D structure of a protein are cryo-electron microscopy (cryoEM), nuclear magnetic resonance (NMR), and x-ray crystallography. NMR has the benefit of determining a protein's structure while it is in solution, a condition more closely resembling its native surroundings. In contrast, x-ray crystallography determines a protein's structure by analysis of diffraction patterns produced when x-rays are passed through a crystallized protein. While NMR is well-suited for analysis of small proteins (<35 kDa), larger proteins are better addressed with x-ray crystallography [256]. If cryoEM, a more recent technology is gaining in popularity it tends to generate lower resolution structure than x-ray crystallography [257]. In this study, as EnvSia156 is close to 60 kDa, I pursued determination of EnvSia156 structure using x-ray crystallography in a collaboration with Prof. Gideon Davies at the University of York in the UK.

X-ray crystallography relies on the formation of a protein crystal that upon exposure to an x-ray beam generates a diffraction pattern. The atom arrangement from the crystal can be deduced from two parameters: the amplitude of the diffraction pattern and the phase of the diffracted radiations. Intensities of the diffracted spots enable calculation of the amplitude from the acquired data. However, phase information is missing (an issue referred to as 'the phase problem'), which once solved, enables determination of a protein's structure [258, 259].

Three techniques can be employed to solve the phase problem: multiple isomorphous replacement (MIR), multiwavelength anomalous diffraction (MAD), and molecular replacement (MR) [259, 260]. The latter is a favored approach as it requires only a single set of diffraction data to be collected. MR uses knowledge about the structure of a protein whose sequence is similar to the protein being studied. The higher the sequence identity between the two proteins, the more likely it is to succeed. Typically a minimum of 25% identity

4. Screening metagenomic libraries for sialidases

is required to employ such technique [261]. In this study, MR could not be applied as EnvSia156 does not resemble other known sialidases and is the founding member of a new enzyme family. This dictated our need to use MIR or MAD to solve the phase problem for this protein.

MAD and MIR are experimental approaches to phasing. Both rely on the incorporation of heavy atoms into the protein crystal. The heavy atom-containing crystal is exposed to an x-ray beam and additional sets of data are collected. In MIR, heavy atoms are incorporated into the crystal by soaking native crystals into a solution of mercury, platinum or gold [258]. In contrast, MAD is based on the *in vivo* metabolic incorporation of SeMet into the protein [262, 263]. It has become the favored experimental approach for x-ray phase determination and was the method chosen for this work.

To introduce SeMet into EnvSia156, the pJS119K-EnvSia156 expression construct was introduced into *E. coli* T7 Express Crystal cells, a methionine auxotrophic strain. Cells were cultivated in growth medium supplemented with L-methionine until the exponential growth phase was reached. Cells were then harvested and resuspended in media containing SeMet. Expression of EnvSia156 was then induced, enabling metabolic incorporation of SeMet in place of methionine into EnvSia156.

A culture volume of 12 L was required to produce 14 mg of SeMet-labeled EnvSia156. In comparison, 24 mg of WT EnvSia156 required only a 2 L culture, indicating there was some SeMet toxicity observed during production of the labeled protein. WT and SeMet EnvSia156 were both purified using a two-step protocol. First an affinity chromatography was performed using a nickel column, exploiting the 6-His tag present on the C-terminus of EnvSia156. Second, size exclusion chromatography using a Sephacryl S-200 High Resolution was performed. A final purity over 95% was obtained for both WT and SeMet EnvSia156 as shown in Figure 30A and C. This level of purity was key for forming protein crystals. One guiding principal of this largely empirical

4. Screening metagenomic libraries for sialidases

process is that the chances of forming protein crystals increases proportionally to increased protein purity.

Introduction of SeMet into EnvSia156 was confirmed by intact mass spectrometry analysis of both WT and SeMet EnvSia156 (Figure 30B and D). Two major masses corresponding to the approximated size of WT EnvSia156 were detected at 58.888 kDa and 57.582 kDa respectively (Figure 30B). A single major peak was detected for SeMet EnvSia156 at 59.267 kDa (Figure 30D). The mass of 57.582 kDa was too low to be that of the full-length WT EnvSia156 when compared with the mass of SeMet EnvSia156. This mass presumably corresponds to a truncated version of WT EnvSia156. Consequently, the number of SeMet incorporated within EnvSia156 was estimated considering the mass of 58.888 kDa for WT EnvSia156:

$$\begin{aligned} & \# \text{ SeMet incorporated} \\ & = \frac{\text{Mass of SeMet EnvSia156 (Da)} - \text{Mass of WT EnvSia156 (Da)}}{\text{Mass SeMet (Da)} - \text{Mass Met (Da)}} \\ & \# \text{ SeMet incorporated} = \frac{59266.9684 - 58887.6367}{49} \approx 7.7 \end{aligned}$$

The EnvSia156 protein sequence is comprised of 505 amino acids coupled to 6 terminal histidine residues added to facilitate purification. Of these 511 amino acids, 12 are Met residues. Typically, one SeMet for every ~75-100 amino acids is necessary for successful MAD [264]. Based on our mass spectrometry data, on average 8 Met were substituted by SeMet in the expressed SeMet EnvSia156 mutant. Thus, a sufficient number of heavy atoms were becoming inserted into EnvSia156 to enable formation of SeMet-derived crystals for use in MAD.

4. Screening metagenomic libraries for sialidases

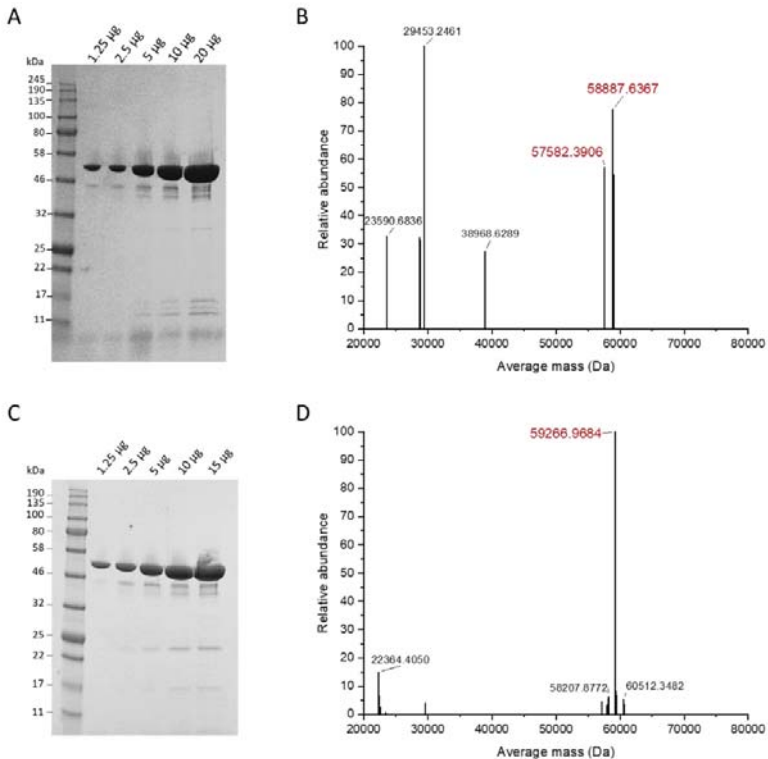


Figure 30. Wild-type and SeMet-labeled EnvSia156 expressed proteins. His-tagged WT (A) and SeMet-labeled (C) EnvSia156 were expressed in *E. coli* for growth of protein crystals. Intact mass spectrometry analysis of wT (B) and SeMet-labeled (D) EnvSia156.

4.3.2.2 Structure of EnvSia156 defines an unusual sialidase fold

We sought to determine the oligomeric status of EnvSia156. Because the chosen approach to solve EnvSia156 3D structure was x-ray crystallography, the determination of its oligomeric status in solution was necessary. X-ray crystallography solves the 3D structure of molecules in a crystal form. Consequently, interactions of protein monomers observed in x-ray

4. Screening metagenomic libraries for sialidases

crystallography may differ from true oligomeric contacts occurring in solution. To evaluate EnvSia156 oligomeric status in solution we used SEC-MALLS analysis. SEC-MALLS combines gel filtration chromatography with MALLS detection and is one of the most reliable methods for determination of proteins molecular weight and oligomeric status. If SEC alone can be used to estimate these values, it suffers two principal limitations. In SEC analyses, the protein molecular weight is correlated to its retention time and compared to standard proteins. However, for this method to be accurate, the studied protein must have the same conformation as the standards, and must not interact with the column by electrostatic or hydrophobic interactions [265]. Unlike SEC alone, SEC-MALLS enables absolute determination of a protein or protein complex molecular weight without the use of any calibration or standards [266]. In this approach, SEC is solely used to separate the different species in solution and not for assessing the protein molecular weight. The latter is determined from the measurement of the light scattered into multiple angles by the protein illuminated with a laser beam [265]. SEC-MALLS analysis of EnvSia156 indicated it is a dimer in solution with an estimated absolute molecular mass of 119.2 kDa (Figure 31).

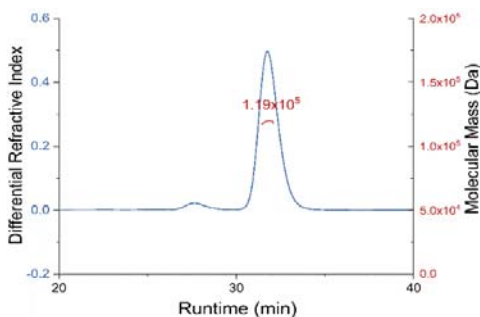


Figure 31. Analysis of the solution oligomeric state of EnvSia156 by SEC-MALLS. From [267]. SEC-MALLS trace showing a large peak for purified EnvSia156, with an estimated molecular mass of 119 kDa, consistent with a dimeric assembly. A second discrete peak can be observed which may correspond to a tetrameric assembly (dimer of a dimer).

4. Screening metagenomic libraries for sialidases

To further explore the 3D structure of EnvSia156, the protein was crystallized in the absence of a substrate, and in complex with Neu5Ac, Neu5Gc, Kdn and the sialidase inhibitor DANA. SeMet-labeled EnvSia156 was also crystallized and data collected at three different wavelengths (selenium peak, inflection, and high-energy remote). The resulting refined SeMet structure was subsequently used as the starting model for refinement of unliganded EnvSia156 and complexes with 20 mM Neu5Ac, Neu5Gc, DANA, Kdn. The best model was obtained from a crystal of EnvSia156 with bound Neu5Ac (EnvSia156-Neu5Ac), at a resolution of 2.0 Å (Figure 32). Two copies of the protein were present in the asymmetric unit, each composed of two distinct domains. The two copies appear to be related by non-crystallographic symmetry resulting in a dimer consistent with the SEC-MALLS results.

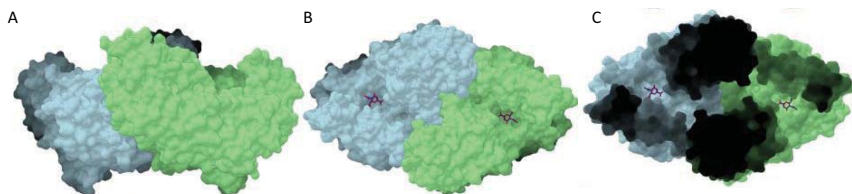


Figure 32. 3D structure of EnvSia156. From [267]. (A and B) Van der Waals' surface of a EnvSia156 homodimer with one unit colored in light blue and the other in light green. (C) "Top-down" view of the dimer with depth shading highlighting the continuous grooves connecting the two binding sites.

The catalytic domain (residues 6–376) comprises a complete $(\beta/\alpha)_8$ -barrel fold while the C-terminal domain (residues 377–502) consists of an eight-stranded β -sandwich (Figure 33A). The functional side of a barrel enzyme is normally present on the "N-terminal" face of the barrel, defined by the $\beta\alpha$ -loops (loops with a β -strand on the amino end and an α -helix on the carboxy-terminus end), which are therefore longer than the purely structural $\alpha\beta$ -loops [268]. That considered, EnvSia156 $\beta\alpha$ -loops are still remarkably elongated. A particularly extensive loop is formed by residues 50–85, connecting β -strand 2 to α -helix 2, which is itself an unusually large helix (residues 86–119) and one of the most

distinctive features of EnvSia156 structure. The C-terminal β -sandwich domain is formed by eight antiparallel β -strands organized into two β -sheets, with one sheet formed by strands 1, 3, 6, and 8 and the other by strands 2, 4, and 7, while strand 5 is shared between the two faces. The larger face of the β -sandwich folds on itself as the loop connecting β -strands 5 and 6 bends towards the catalytic domain and interacts with the long 7th $\beta\alpha$ -loop of the barrel, contributing for structural stabilization. The first β -strand of the C-terminal domain is interrupted by a long loop (residues 382–399) that likely plays a role both in overall structural stabilization and dimer assembly. With B -factors values ranging from 16 to 20 it is one of the most rigid features in the structure, which is unusual for a large loop. This loop makes several important hydrophobic interactions with both its dimeric counterpart and the catalytic domain, contributing to the formation of a continuous cleft connecting the active sites of the two units in the dimer (Figure 32C). According to PDBsum [269], 20% of the residues establishing non-bonded contacts between the two units are located in this loop, highlighting its role in the dimerization process. An assembly analysis with PISA calculated an interaction surface area between the two monomers of 1725 \AA^2 with a predicted solvation energy gain (ΔG) of -13.2 kcal/mol, suggesting that the observed homodimer formed by the two copies present in the asymmetric unit is the natural biological assembly of EnvSia156. The structure of EnvSia156 is clearly different from all the known families of both exosialidases (Figure 33B, C & D) and endosialidases (Figure 33E) which, despite some differences regarding accessory domains, substrate binding clefts or catalytic residues; all possess a six-bladed β -propeller topography [217, 218, 220, 270, 271]. A structural similarity search on DALI server [272] using the catalytic domain of EnvSia156 (residues 6–375) returns Cwp19 “functional” region as the closest structural homolog (PDB code 5OQ2, z score of 27.2 and rmsd of 3.4 \AA over 299 aligned residues [273]) with 14% identity between aligned regions. Cwp19, produced by *Clostridium difficile* during stationary phase to induce autolysis, has been recently described as a lytic transglycosylase with activity on peptidoglycan [274], but beyond the superficial fold

4. Screening metagenomic libraries for sialidases

resemblance there is no conservation of active center residues and indeed the ligand complexes of GH156 clash with Cwp19 main chain following overlap.

The C-terminal β -sandwich domain (residues 376–502) shows distant homology to a domain from a *Bacteroides uniformis* uncharacterized family 86 glycoside hydrolase possessing agarase activity (PDB code 5TA5, z score of 11.6 and rmsd of 2.1 over 101 aligned residues). It is also distantly related to a CBM4 domain appended to a porphyranase produced by *Bacteroides plebeius*, similarly classified as a GH86 (PDB code 4AW7, z score of 9.7 and rmsd of 2.0 Å over 102 aligned residues). Both are bacteria found in the human gut microbiome [275, 276] thought to have acquired the GH86 genes by horizontal gene transfer with sea dwelling organisms. Spatially equivalent surface aromatics may imply a carbohydrate binding module-like role for the β -sandwich domain of EnvSia156, but that remains to be established.

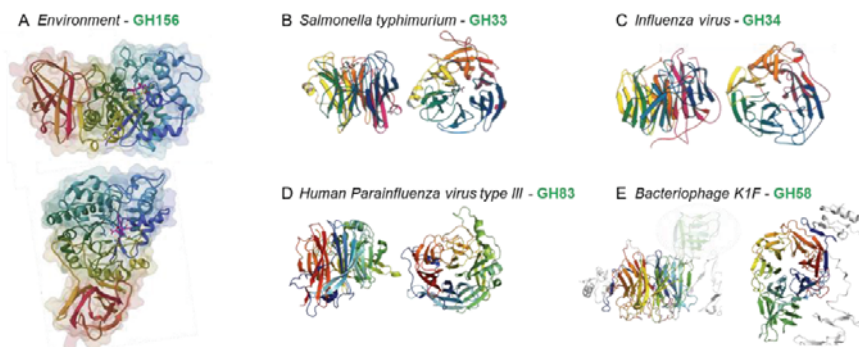


Figure 33. Ribbon representation with rainbow coloring of a representative member of each sialidase family. Families GH33 (B) [216], GH34 (C) [216], GH83 (D) (PDB 4MZA, drawn with Pymol) and GH58 (E) (PDB 1VoE, drawn with Pymol) share a similar six-bladed β -propeller fold unlike family GH156 (A) ([267]). GH156 catalytic module (blue-green) comprises a $(\beta/\alpha)_8$ -barrel fold while the C-terminal module (red-orange) consists of an eight-stranded β -sandwich (A).

4.3.2.3 Mechanism of EnvSia156 and definition of its active center

Having ascertained that EnvSia156 indeed displays an unusual sialidase fold, definition of the active center of this family was next sought. A sequence comparison with the closest primary structure homologs found in GenBank using the BLASTP tool (Appendix: Supplementary Figure 4), shows that areas with the highest degree of conservation are mostly found within the β -strands of the $(\beta/\alpha)_8$ -barrel or just after it, on the “N-terminal” face of the $(\beta/\alpha)_8$ -barrel. By coloring the structure according to homology, using the same color scheme as in the alignment, it is possible to identify a highly conserved pocket on the “N-terminal” face of the $(\beta/\alpha)_8$ -barrel that sits at both ends of the continuous groove formed by the dimer (Figure 35). Considering its position and conservation, this site is a likely candidate to represent the EnvSia156 catalytic center. Classic retaining exosialidase enzymes have an active center characterized by a triad of arginines that stabilize the carboxylate of sialic acid, three catalytic residues - a glutamic acid, a tyrosine, and an aspartic acid - and an hydrophobic pocket that accommodates the C5 moiety [216] (Figure 36A). The identified conserved pocket indeed shows the presence of three 100% conserved arginines (Arg129, Arg202, and Arg246), hinting at a sialoside binding region, although one of them, Arg246, seems to be too buried to be able to interact with the carboxylate moiety of sialic acid. EnvSia156 also displays several highly conserved hydrophobic residues forming a pocket that could potentially accommodate the aglycone moiety of the sialoside substrates. To better define, the active center residues, 3D structures were determined with the product Neu5Ac and with the inhibitor DANA (Figure 34A&B).

4. Screening metagenomic libraries for sialidases

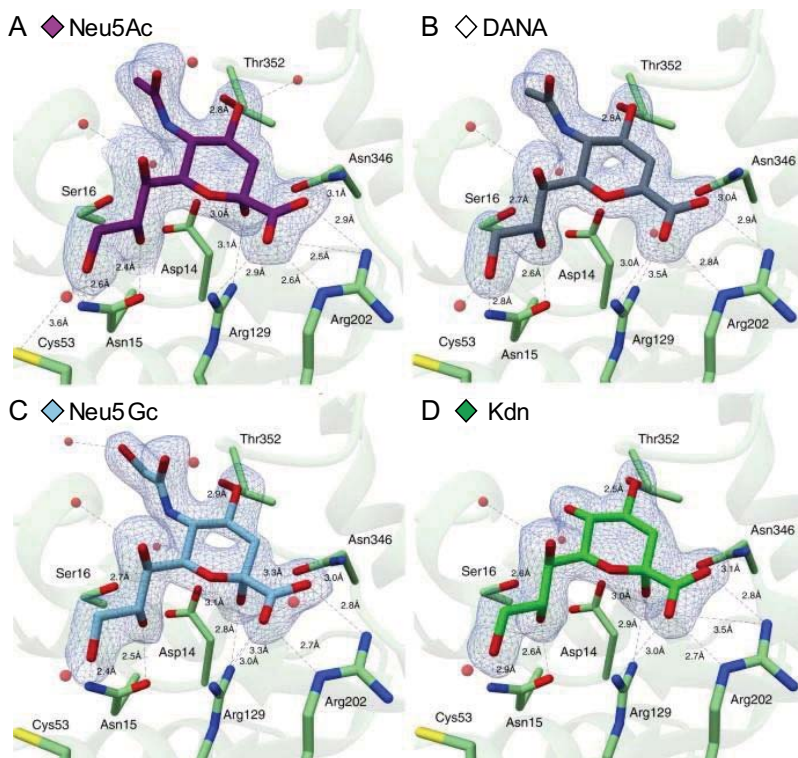


Figure 34. Crystal structures of EnvSia156-ligand complexes. From [267]. The four panels show a detailed view of the crystallized EnvSia156-ligand complexes within the active center. The enzyme residues making hydrogen-bond contacts (black dashed lines) with the ligands are displayed in stick representation. The ligands are surrounded by a mesh representation of the Refmac5 maximum-likelihood σ A-weighted $2F_o - F_c$ electron density map contoured at 1σ (0.46 electrons/ \AA^3), and are colored purple (A – Neu5Ac), gray (B – DANA), light blue (C – Neu5Gc) and green (D – Kdn). Water molecules are colored red.

The structures with Neu5Ac show the β -anomer (equating to a product complex following catalytic attack on an α -sialoside with inversion) of the ligand interacting with EnvSia156 at the predicted catalytic center based on

4. Screening metagenomic libraries for sialidases

sequence homology, through 13 hydrogen bonds (with 8 more mediated by water molecules) and several hydrophobic interactions (Figure 34A). The -1 subsite, which accommodates the sialic acid monosaccharide, shows several differences from what is commonly observed in other sialidases. It is possible to identify the characteristic triad of residues coordinating the carboxylate group, although it consists of two arginine side-chains (Arg129 and Arg202) and one asparagine (Asn346), instead of the three arginines displayed in conventional sialidases. Invariant Asp14 lies adjacent to the (anomeric) C2 carbon and makes a hydrogen bond with the hydroxyl group. Thus, Asp14 is the most likely candidate to act as a general catalytic base residue.

Another major difference between EnvSia156 and other sialidases is the absence of a hydrophobic pocket that accommodates the C5 moiety of sialic acids. Rather, the more open structure of the -1 sub-site orientates the sialic acid in such way that the functional group at the C5 position is pointing away from the enzyme. The structure of EnvSia156 with Neu5Ac shows a single discrete non-bonded contact between Gln351 and the acetamide methyl group and one water-mediated hydrogen bond between N5 and Asp14/Ser16. This suggests that EnvSia156 can tolerate different C5 moieties and explains why, while showing some preference for Neu5Ac, it can also hydrolyze terminal Neu5Gc (Figure 26E). To test this supposition, the structure of EnvSia156 in complex with Kdn was obtained, showing that this sialic acid with a simple hydroxyl group at the C5 position can also be accommodated (Figure 34D). All complex structures show the protein forming hydrogen bonds with all three hydroxyl groups in the glycerol chain, although the residues involved are not as conserved as those mentioned above. Tyr20 and Tyr135 bind with O7 via a water molecule, Ser16 and Asn15 coordinate O8, and O9 contacts Asp132 and His134 via a water molecule, while binding directly to Asn15 and Cys53. This is not a common feature in sialidases and suggests that the glycerol group is a key feature for substrate recognition.

4. Screening metagenomic libraries for sialidases

EnvSia156 acts, in solution on a variety of α 2-3, and α 2-6 linked sialic acid glycosides including complex *N*-glycans and *O*-glycan-linked sialic acids (Figure 26 & Figure 27). The diversity of structures accepted as leaving groups suggests a more open and less hydrogen-bonded environment. We indeed observed a hydrophobic platform, adjacent to the -1 sub-site (Figure 35B). Four highly conserved aromatic residues, namely Trp164, Phe203, Phe278, and Trp279, forming a hydrophobic pocket that could possibly accommodate the aglycone moiety in the $+1$ binding site (Figure 35B).

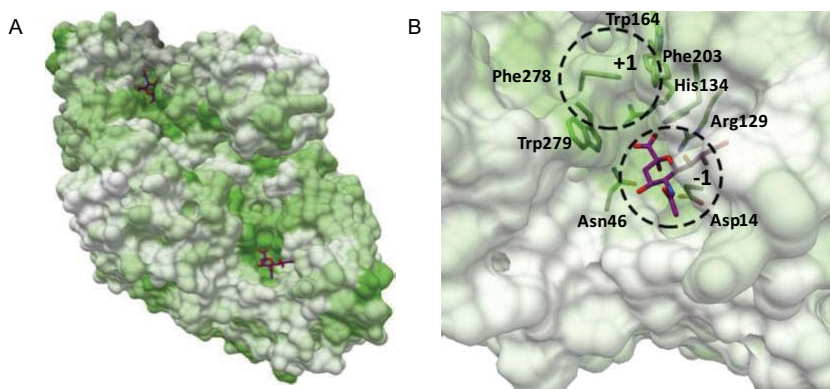


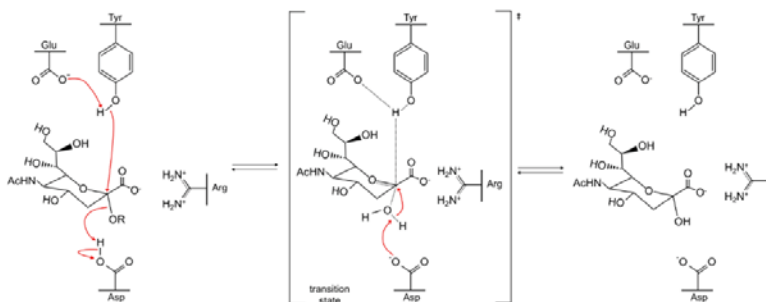
Figure 35. EnvSia156 homology and binding subsites. From [267]. (A) Birdseye view of an EnvSia156 dimer with the van der Waals' surface colored according to sequence conservation, calculated by aligning sequences (ClustalOmega) with the closest primary structure homologs found in GenBank using BLASTP tool. The color pattern matches the alignment used to render the structure (Appendix: Supplementary Figure 4), with darker shades of green corresponding to full conservation and white to no conservation. (B) Detailed view of the catalytic site of EnvSia156 with putative substrate coordinating residues in a stick representation, colored according to sequence conservation. The -1 and putative $+1$ binding sub-sites are highlighted with dashed circles.

In the classic retaining mechanism of exosialidases, the carboxylate group of a glutamate forms a hydrogen bond with the hydroxyl group of the catalytic tyrosine, increasing its nucleophilic character, which in turn attacks the

4. Screening metagenomic libraries for sialidases

anomeric carbon to form a covalent glycosyl-enzyme intermediate. An aspartate functions as the acid/base residue donating the proton on this initial step and then facilitating the nucleophilic attack of the glycosyl-enzyme intermediate [277] (Figure 36A).

A Retaining exosialidase



B Inverting glycoside hydrolase

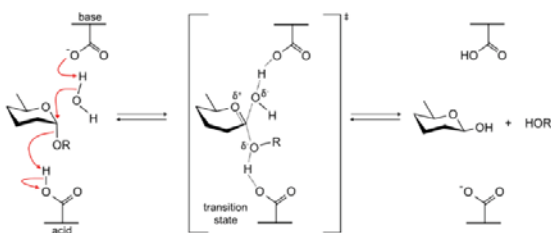


Figure 36. Catalytic mechanism of retaining exoglycosidases in contrast to a typical inverting glycoside hydrolases. (Figure adapted from Steve Withers, Spencer Williams "Glycoside Hydrolases" in CAZypedia, available at URL <http://www.cazypedia.org/>, accessed 07 April 2021) (A) Hydrolysis by retaining exosialidases involves three catalytic residues: a tyrosine (Tyr), a glutamate (Glu) and an aspartate (Asp). (B) Hydrolysis by a general inverting glycoside hydrolase typically involves a Brønsted acid and base.

Contrastingly, but as expected for an inverting enzyme, none of the signature retaining sialidase residues are identifiable in the EnvSia156 active site. EnvSia156 is an inverting enzyme (Figure 28) which demands two catalytic residues: a Brønsted acid to protonate the leaving-group and a Brønsted general

4. Screening metagenomic libraries for sialidases

base to activate the nucleophilic water molecule for nucleophilic attack [278–281] (Figure 36B). Asp14 is ideally placed to assume the role of general base, through its position on the “beta” face and its direct interaction with the product OH in the Neu5Ac complex. Furthermore, in the DANA complex, Asp14 coordinates a single water molecule below the anomeric C2 carbon, a position that mimics what would be seen in a substrate complex. On the “alpha” face side of the sugar, His134 could potentially act as the Brønsted acid catalyst that donates the hydrogen to the leaving group (Figure 37A). To test the hypothesis that Asp14 and His134 are key catalytic residues, D14A, D14N, H134A and H134N variants were created and tested on 4-MU- α -Neu5Ac. Asp14 mutants were inactive, consistent with the role of Asp14 as general base (Figure 37C). In contrast, His134 variants suffered a significant drop (~30%) in activity (Figure 37C), but mutation of this site did not fully abolish EnvSia156 activity. This is consistent with the reduced need for protonic assistance to liberate the activated sugar analog substrate, a good leaving group. If H134 functioned as a general acid, we would predict it to be less active on less-activated substrates having higher pKa leaving groups such as sialyllactose. To test this hypothesis, the activity of the H134A mutant was tested on procainamide-labeled α -2,3-sialyllactose and compared to the activity of WT enzyme. Reaction conditions leading to partial digestion of the substrates were used and both the unreacted substrate and the product were monitored on a UPLC. While in the tested conditions, 0.15 μ g of WT enzyme was able to hydrolyze $77.5 \pm 1.8\%$ of the substrate no activity was detected using the same amount of the H134A-mutant enzyme (Figure 37D). Up to four times more H134A-mutant were tested but did not permit to detect the formation of product. This observation strongly supports the notion that H134 acts as a general acid during catalysis of natural substrates. An analogous Asp-His catalytic dyad has been similarly proposed for members of the GH117 family [282, 283]. Members of this family act through a single displacement inverting mechanism, similar to the one here proposed for EnvSia156.

4. Screening metagenomic libraries for sialidases

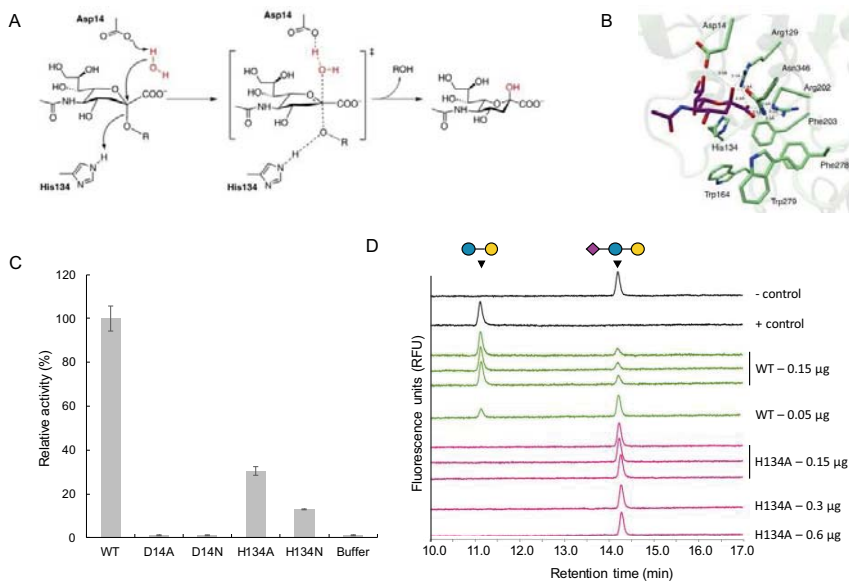


Figure 37. EnvSia156 proposed catalytic mechanism. Adapted from [267]. (A) Proposed inverting mechanism for EnvSia156 with Asp14 acting as the general catalytic base and His134 as the acid catalyst. (B) Structure of the EnvSia156-Neu5Ac complex highlighting the two putative catalytic residues (Asp14 and His134), the carboxylate-coordinating triad (Arg129, Arg202, Asn346) and the hydrophobic pocket (Trp164, Phe203, Trp279, Phe278). (C) Activity of EnvSia156 mutants (D14A, D14N, H134A and H134N) versus wild-type (WT) on 4-MU- α -Neu5Ac. (D) Activity of H134A EnvSia156 mutant compared to WT on procainamide-labeled α 2-3-sialyllactose monitored by UPLC.

A prior study by Newstead *et al* showed that a retaining sialidase could be converted into an inverting sialidase by mutation of the catalytic tyrosine and having a water molecule act as the nucleophile [284]. Thus, we speculate that the inverting catalytic mechanism of EnvSia156 likely involves Asp14-activation of an analogous nucleophilic water molecule that attacks the anomeric carbon, with His134 donating a proton to the leaving group (Figure 37A).

4. Screening metagenomic libraries for sialidases

4.3.3 Discussion

The structure of EnvSia156 was determined by x-ray crystallography using a MAD approach to phasing. A SeMet-labeled variant of EnvSia156 was produced and used to form crystals that along with native EnvSia156 crystals enabled solving the 3D protein structure. To date, EnvSia156 is the only member of the CAZy family GH156 that has been biochemically characterized. Elucidation of its structure establishes basic knowledge for this new family of glycoside hydrolases.

Consistent with EnvSia156's unique inverting mechanism for exosialidases, its structure is also unlike that of known exo- and endo-sialidases. The EnvSia156 catalytic center was explored, and a model of its hydrolysis mechanism was proposed. The resolution obtained for the structure of EnvSia156 alone as well as in complex with products and inhibitors enabled identification of two candidate catalytic residues. Site-directed mutagenesis was performed and confirmed that Asp14 and His134 are part of the enzyme's active site. Our mechanism proposes that Asp14 acts as the general base while His134 plays the role of the acid, enabling hydrolysis of the glycan via a single-displacement mechanism. Such facets have not been previously observed in the canon of literature for sialidases

At the end of 2020, a breakthrough in the field of structural biology was made with the development of the algorithm AlphaFold2 [285]. This software uses artificial intelligence to predict protein structure from an amino acid sequence. The algorithm demonstrated its performance in the 2020 Critical Assessment of protein Structure Prediction (CASP) challenge where it was capable of successfully predicting the structure of a hundred of proteins. In most cases, structures predicted by AlphaFold2 were very close to those obtained by experimental means. This big advancement in the determination of protein structures may, out of curiosity, raise the following question: could the structure of EnvSia156 have been *in silico* determined using such algorithm? If the question is worth asking, it is important to keep in mind that good diffracting

4. Screening metagenomic libraries for sialidases

crystals were successfully generated for both EnvSia156 and SeMet EnvSia156, and that traditional x-ray crystallography techniques enabled solving its structure. In fact, the best application of AlphaFold2 could be found for proteins that are difficult to crystallize. The algorithm represents an interesting alternative to experimental structure determination when obtaining experimental data is an obstacle.

In certain types of cancer, hypersialylated cells are often observed [207]. This phenotype is typically associated with poor prognosis and decreased immunogenicity. Sialylated glycans on cancer cells are involved in cancer immune escape. Previous studies have shown that by conjugating sialidases with antibodies it is possible to direct their activity towards the specific desialylation of tumor cells. Termed ‘glycocalyx editing’, this approach has great potential in cancer therapy [286, 287]. As the GH156 family operates using an inverting mechanism, its potential to improve glycocalyx editing was explored. Initial experiments performed by Prof. Carolyn Bertozzi’s group (pioneers in the glycocalyx editing field) showed that the ability of EnvSia156 to cleave sialic acid from the surface of a chronic myelogenous leukemia-derived K562 cell line was limited in comparison to the well-known *Vibrio cholerae* sialidase (data not shown). However, it is possible that protein engineering of EnvSia156 (*e.g.*, conjugation with antibodies or a lectin) to better target the enzyme to the glycocalyx might improve its utility in editing.

4.4 Identification of unconventional sialidases

Neu5Gc is a human xenoantigen that may have a detrimental effect when incorporated into normal human tissues [125, 126, 209]. Its presence in therapeutics might also impact treatment efficacy, and some transplant rejection may be induced by the presence of Neu5Gc in a donated organ [127, 132, 288, 289]. Pending specific regulation by the FDA, pharmaceutical companies monitor and control the presence of Neu5Gc in their manufactured products.

Most sialidases can cleave both Neu5Ac and Neu5Gc in a similar manner. However, all three sialidases encoded by *Streptococcus pneumoniae* (NanA,

4. Screening metagenomic libraries for sialidases

NanB and NanC) are reported to have a strong preference for Neu5Ac over Neu5Gc [290, 291]. Conversely, no sialidase has been reported with specificity bias for Neu5Gc over Neu5Ac. In respect to the increase interest in Neu5Gc, the availability of a Neu5Gc specific enzyme would be a noteworthy addition to the glycoanalytical toolbox. To this end, I executed a second screen aimed at potential identification of sialidases with a specificity bias for Neu5Gc.

4.4.1 Material and methods

4.4.1.1 Screening for Neu5Gc specific sialidases

A compost-derived metagenomic library was differentially screened with fluorogenic 4-methylumbelliferyl- α -D-N-acetylneuraminic acid (4-MU- α -Neu5Ac) (Toronto Research Chemicals, North York, ON) and 4-methylumbelliferyl- α -D-N-glycolylneuraminic acid (4-MU- α -Neu5Gc) substrates as described in Chapter 3, section 3.2.4.2. In a primary screen, library lysates were assayed using 4-MU- α -Neu5Gc. Positive clones that met our definition of a ‘primary screen hit’ definition (see Chapter 3, section 3.3.2.3) were archived in a fresh 384-well plate in sterile 20% (v/v) glycerol. Each positive clone was grown and comparatively re-screened in separate assays containing 4-MU- α -Neu5Ac and 4-MU- α -Neu5Gc substrates (reactions run in duplicate).

4.4.1.2 PacBio sequencing and enzyme identification

Fosmids C19 and C22 were sequenced from a multiplexed PacBio sequencing library prepared as described in Chapter 3, section 3.2.5.2.

4.4.1.3 Expression and purification of C19 and C22 sialidases

Enzyme cloning, expression and purification was performed by a colleague, Mehul B Ganatra. Primers used for cloning can be found in Supplementary table 1 (Appendix). Briefly, exosialidases were isolated by PCR from their respective fosmids C19 and C22. Genes were cloned in the pET21a(+) vector with 6 C-terminal histidine-tag using the NEB HiFi cloning kit. Enzymes were expressed

and purified following standardized protocols, similar to what is described in section 4.2.1.3.

4.4.1.4 Determination of sialidase substrate preference

Commercial sialidases from *Clostridium perfringens*, *Arthrobacter ureafaciens* and *Streptococcus pneumoniae* were from NEB. Enzymes from *Vibrio cholerae*, *Micromonospora viridifaciens* and *Salmonella typhimurium* were from MilliporeSigma, Bio-Techne Corporation (Minneapolis, MN) and Megazyme (Bray, Ireland), respectively.

Each enzyme was assayed in triplicate. One microliter of enzyme was incubated with 2 μL of 100 $\mu\text{g}/\text{mL}$ 4-MU- α -Neu5Ac or 4-MU- α -Neu5Gc in a 20 μL reaction with 50 mM sodium acetate pH 5.5. Enzymes from *Clostridium perfringens*, *Arthrobacter ureafaciens* and *Streptococcus pneumoniae* were diluted 1:5 for the assay so the reaction would be slow enough to allow the initial reaction velocity to be monitored. Fluorescence at $\lambda_{\text{ex}}=365$ nm $\lambda_{\text{em}}=445$ nm was monitored for 1 h at 37°C. The activity of each enzyme (RFU/min) for both substrates were calculated from the initial fluorescence values, before the reactions plateaued.

4.4.2 Results

4.4.2.1 Functional metagenomic screening

A compost metagenomic library (see Chapter 3.3.1.2) comprised of 5,376 clones was screened for sialidase activity using 4-MU- α -Neu5Gc. Clones meeting the hit definition described in Chapter 3, section 3.3.2.3, were identified, arrayed in a new plate, and comparatively re-screened using two separate assays (4-MU- α -Neu5Gc versus 4-MU- α -Neu5Ac). In total, 8 significant hits were observed (clones A12, A24, B23, C19, C22, D8, D9 and D13) (Figure 38A). Six of these (A12, A24, B23, D8, D9 and D13) displayed strong activity on 4-MU- α -Neu5Ac, consistent with the prior observation that most exosialidases have the ability to hydrolyze both Neu5Gc and Neu5Ac [292–294]. The fluorescence signal from both re-screen assays after a 7 h

4. Screening metagenomic libraries for sialidases

incubation was used to compare the clones' activity on 4-MU- α -Neu5Gc versus 4-MU- α -Neu5Ac. A Neu5Gc/Neu5Ac fluorescence ratio over 1 was observed for 3 clones (C19, C22 and D9) indicating a preference for the 4-MU- α -Neu5Gc substrate. Among these 3 clones, 2 (C19 and C22) showed limited activity on 4-MU- α -Neu5Ac. In comparison, 5 clones (A12, A24, B23, D8 and D13) showed a Neu5Gc/Neu5Ac fluorescence ratio below 1, indicating a preference for the 4-MU- α -Neu5Ac substrate (Figure 38B). Clones C19 and C22 were the most compelling hits as their activity on 4-MU- α -Neu5Ac was minor but they had more pronounced activity on 4-MU- α -Neu5Gc (Figure 38A).

4. Screening metagenomic libraries for sialidases

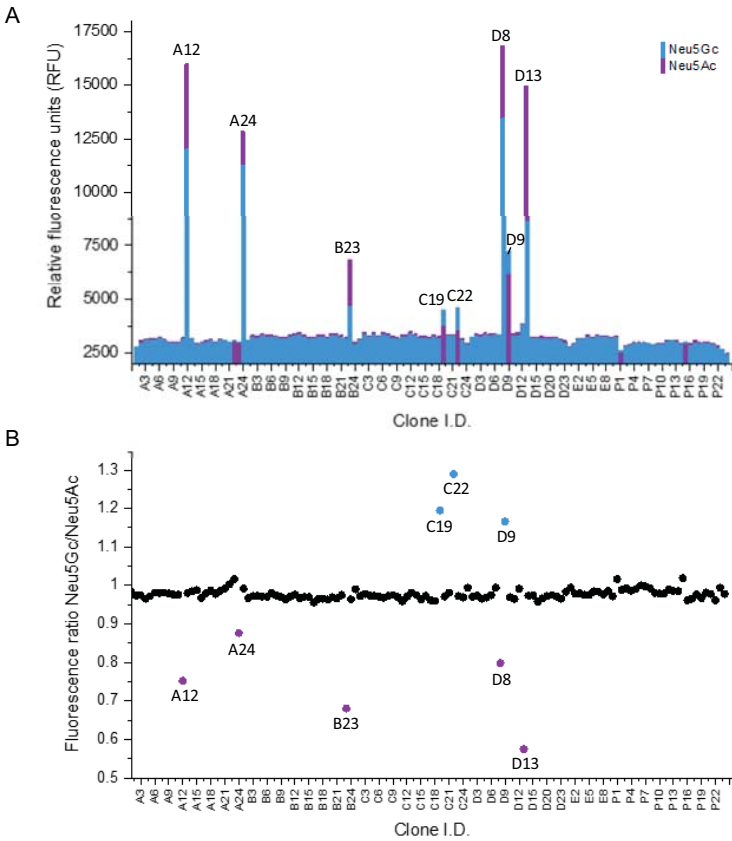


Figure 38. Neu5Gc and Neu5Ac differential screening of a compost metagenomic library. Lysate from microcultures of *E. coli* cells harboring individual fosmid clones were assayed for sialidase activity with 4-MU- α -Neu5Ac. Clones considered initial hits were collected and re-screened separately in a differential assay with 4-MU- α -Neu5Ac and 4-MU- α -Neu5Gc (A) Bars in purple represent the fluorescence signal in the 4-MU- α -Neu5Ac assay while bars in blue represent the signal in the 4-MU- α -Neu5Gc assay. The signal ratio of 4-MU- α -Neu5Gc over 4-MU- α -Neu5Ac was plotted (B). Five clones (purple circles) with a preferential activity for 4-MU- α -Neu5Ac were identified as well as 3 clones (blue circles) with a preferential activity for 4-MU- α -Neu5Gc.

4. Screening metagenomic libraries for sialidases

4.4.2.2 Identification of C22 and C19 sialidases

Fosmid clones C22 and C19 were isolated and sequenced using the PacBio technology. The DNA sequence of cloned inserts from C19 and C22 each encoded a single bacterial sialidase gene, termed ORF3 (Figure 39A) and ORF6 (Figure 39B), respectively. Analysis of the deduced amino acid sequences of C19 and C22 sialidases showed high similarity to each other (56% identity) (Appendix: Supplementary Figure 5). Both belong to the GH33 family of bacterial exosialidases and are most similar to proteins from terrestrial bacteria of the genera *Rhodopirellula* and *Verrucomicrobia*, respectively.

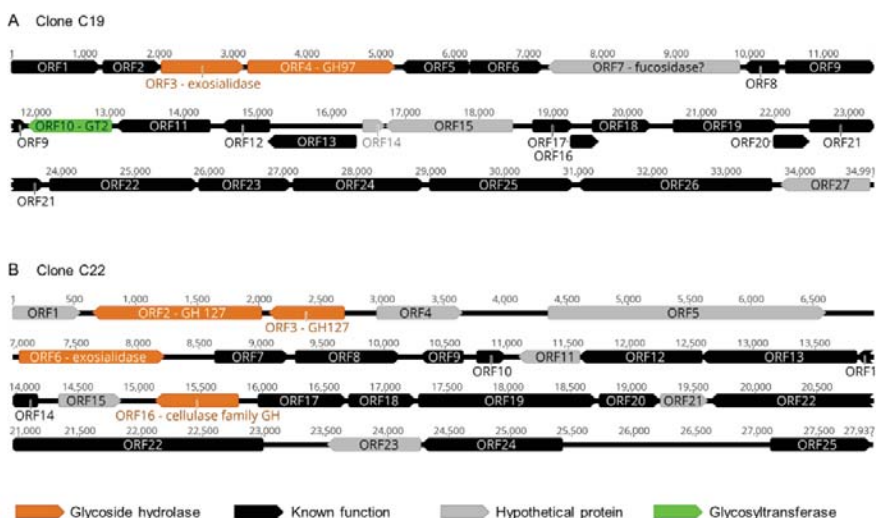


Figure 39. Clone C19 and C22 ORF map. Fosmid C19 (A) and C22 (B) were sequenced on the PacBio RSII platform. ORFs were predicted with MetaGeneMark and classified into four categories based on their homology to proteins from the nonredundant protein database (NCBI). Orange, glycoside hydrolases; black, proteins of known annotated function; gray, “hypothetical proteins” with no annotated function; green, glycosyltransferases.

4.4.2.3 Neu5Gc preferring sialidases activity

Recombinant putative sialidases were evaluated for their substrate preferences. ORF3 from C19 and ORF6 from C22 were cloned in a pET21a(+) vector with a histidine tag for over-expression and purification. Purified enzymes were then tested for their ability to hydrolyze 4-MU- α -Neu5Gc and 4-MU- α -Neu5Ac. The recombinant enzymes were compared to a panel of commercially available exosialidases from 6 other bacterial sources: *Arthrobacter ureafaciens*, *Clostridium perfringens*, *Micromonospora viridifaciens*, *Salmonella typhimurium*, *Streptococcus pneumoniae* and *Vibrio cholerae*. EnvSia156, for which discovery by functional metagenomics is described in Chapter 4.2, was also included in this panel of exosialidases.

Exosialidases C19-ORF3 and C22-ORF6 showed a clear statistical preference ($P \leq 0.001$) for 4-MU- α -Neu5Gc over 4-MU- α -Neu5Ac while most of the tested sialidases did not show statistical preference for one substrate over the other (Figure 40). As previously reported, the sialidase from *Streptococcus pneumoniae* showed selective activity for 4-MU- α -Neu5Ac [290, 291] and did not hydrolyze 4-MU- α -Neu5Gc. The exosialidase from *Clostridium perfringens* showed a slight preference for 4-MU- α -Neu5Ac, and the exosialidase from *Arthrobacter ureafaciens* showed a slight preference for 4-MU- α -Neu5Gc ($P \leq 0.05$) (Figure 40).

4. Screening metagenomic libraries for sialidases

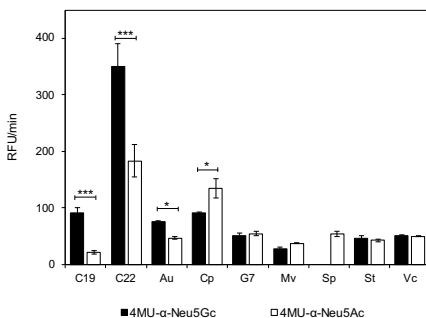


Figure 40. Neu5Gc and Neu5Ac activity of a bacterial exosialidase panel. From [128]. Statistical significance was determined by Student's two-tailed t-test. * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$. Bars represent the geometric mean \pm s.e.m. Abbreviations: RFU: relative fluorescence units, C19: C19-ORF3, C22: C22-ORF6, Au: *Arthrobacter ureafaciens*, Cp: *Clostridium perfringens*, G7: 'Small Dixie' clone G7-EnvSia156, Mv: *Micromonospora viridifaciens*, St: *Salmonella typhimurium*, Sp: *Streptococcus pneumoniae* and Vc: *Vibrio cholerae*.

Significant preference of exosialidases for Neu5Gc over Neu5Ac hydrolysis as observed for C19-ORF3 and C22-ORF6 has not been reported previously. This observation underscores that sialidases with such substrate bias exist and that the differential screening methodology that was applied could identify them. This work was included in a published study on this topic with collaborators at the University of California San Diego (See Discussion; [128]).

4.4.3 Discussion

In this project, I sought to determine if sialidases with a preference for hydrolysis of Neu5Gc exist in nature and if they could be identified using a differential functional metagenomic screen. Such a specificity would be useful in glycoanalytics, and possibly in therapeutic or probiotic development. The screen identified two sialidases termed C19-ORF3 and C22-ORF6 that harbor a strong preference for hydrolysis of a 4-MU- α -Neu5Gc substrate over its acetylated counterpart 4-MU- α -Neu5Ac. Despite their substrate preference,

4. Screening metagenomic libraries for sialidases

both enzymes still can hydrolyze Neu5Ac. However, their considerable preference for Neu5Gc suggests that it might be possible to develop reaction conditions (pH, temperature and incubation time) or engineer mutant version of these proteins that further enhances cleavage of Neu5Gc residues while excluding Neu5Ac moieties. Furthermore, characterization of C19-ORF3 and C22-ORF6 sialidases within glycoanalytical workflows was not performed in this initial discovery study but would be of interest.

While identifying the C19-ORF3 and C22-ORF6 exosialidases, we learned of ongoing work in Karsten Zengler's group at the University of California San Diego. They had a similar interest in finding exosialidases with preference for Neu5Gc hydrolysis, but they were attempting identification of such enzymes using entirely different methodology. In their study, they compared the gut microbiome of humanized mice (mice carrying a *cmah*^{-/-} mutation are unable to make Neu5Gc, a model for humans who naturally lack Neu5Gc) fed with a Neu5Gc-rich, a Neu5Ac-rich or Sia-poor diet. They looked for sialidase genes encoded by gut bacteria from the three mouse cohorts and identified several diet-dependent sialidases, amongst which was sialidase26. Sialidase26 is encoded by a bacterial species found most abundantly in the Neu5Gc-rich diet. Its appears to derive from a close relative of *B. thetaiotaomicron*. Sialidase26 belongs to family GH33 (like C19-ORF3 and C22-ORF6). In enzyme assays, Sialidase26 showed preferential hydrolysis of Neu5Gc over Neu5Ac (Figure 41A). Microbiome sequencing data was also analyzed for member of the Hazda tribe, an African group whose diet annually cycles between high and low Neu5Gc foods. In this cohort, sialidaseHz136 was identified. This enzyme showed strong sequence similarity to sialidase26 found in their mouse study. SialidaseHz136 also showed preferential activity for hydrolysis of Neu5Gc over Neu5Ac (Figure 41B).

4. Screening metagenomic libraries for sialidases

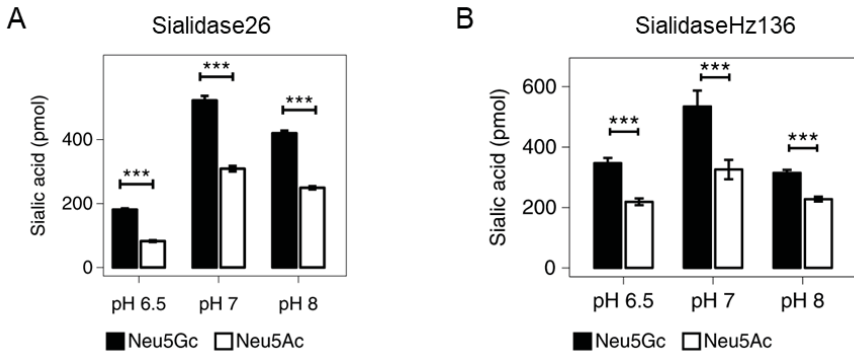


Figure 41. Sialidase26 (A) and sialidaseH136 (B) substrate preference. From [128]. Statistical significance was determined by Student's two-tailed t-test. * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$. Bars represent geometric mean \pm s.e.m.

Our findings were shared with Karsten Zengler's group and published as a collective study that describes the identification of unconventional sialidases with previously undescribed preference for Neu5Gc. Our combined data support the conclusion that exosialidases with Neu5Gc preference are widespread in nature, being found enriched in the gut microbiota of mice and humans upon consumption of a Neu5Gc-rich diet, as well as in terrestrial environments where decomposing biological material from animal sources might require such specificity preference. Interestingly, phylogenetic analysis of protein sequences from sialidase26, sialidaseH136, sialidase C19-ORF3, and C22-ORF6 shows notable divergence. The Neu5Gc-preferring sialidases obtained from mammalian gut microbiota (sialidase26 and sialidaseH136) are phylogenetically distinct from those found in compost (C19-ORF3 and C22-ORF6), suggesting that the ability of certain GH33 sialidases to efficiently hydrolyze Neu5Gc may have occurred via parallel evolutionary events within different ecological niche (Figure 42).

4. Screening metagenomic libraries for sialidases

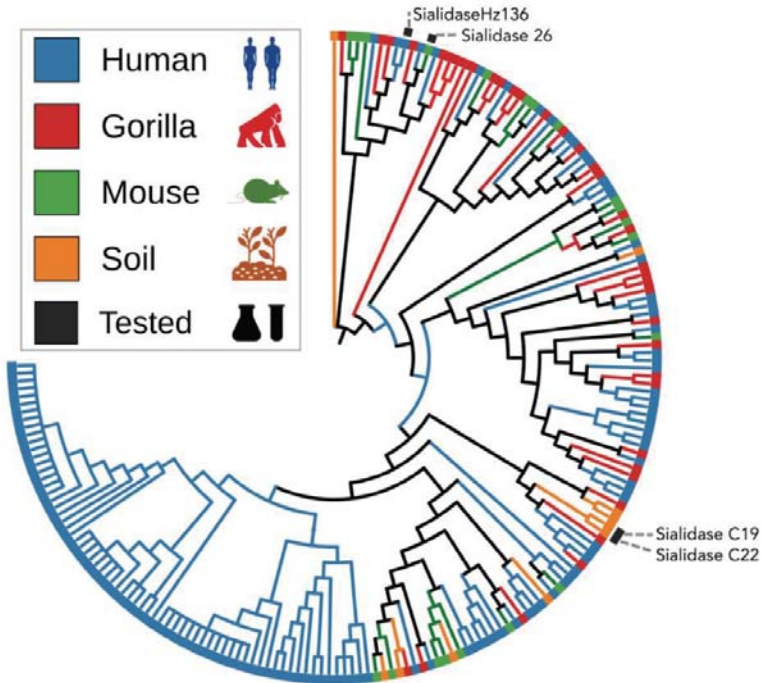


Figure 42. Phylogenetic relationship of sialidases. Adapted from [128]. Sialidase sequences retrieved from human, gorilla, mouse, and soil metagenomic data, are labeled in red, blue, green, and orange, respectively. The sialidases C19-ORF3 and C22-ORF6 identified in this work via functional metagenomic screening along with sialidase Hz136 and sialidase 26 discussed above are indicated at the end of their respective branches.

As described above, incorporation of Neu5Gc into normal human tissue can cause inflammation and increase the risk of certain cancers [295, 296]. A previous study from Dr. Karsten Zengler's group demonstrated that dietary free-Neu5Gc is not incorporated in *cmah*^{-/-} mice as compared to food-bound Neu5Gc [210]. Consequently, cleavage of Neu5Gc from food entering the gut might prevent incorporation of this non-human sugar into colon tissue. Although further *in vivo* characterization is needed, identification of the

4. Screening metagenomic libraries for sialidases

unconventional sialidases described in this chapter lays the foundation for defining a strategy for the use of pre- or probiotics (rich in these enzymes) to prevent incorporation of Neu5Gc into human tissues, thereby reducing the risk of Neu5Gc-mediated chronic inflammation.

4.5 Chapter 4 conclusion

In the present Chapter, two distinct functional screens were performed to identify new sialidases. First, a small thermophilic metagenomic library was screened for sialidases using two different assays (plate- and lysate-based), both employing a Neu5Ac substrate. This screen led to the identification of a new family (GH156) of glycoside hydrolases having an unusual inverting mechanism. The enzyme 3D structure was elucidated through collaboration with Prof. Gideon Davies' group at the University of York and enabled proposal of a catalytic mechanism and identification of active site residues.

Second, a compost metagenomic library was screened for non-conventional sialidases that act on the non-human Neu5Gc form of sialic acid. Two enzymes harboring a strong preference for Neu5Gc were identified, setting a new precedent as such specificity had not been previously reported. Through collaboration with Dr. Zengler's group at the University of California that found exosialidases with similar specificity via an orthogonal approach, we further showed that such enzymes are surprisingly present in vastly different ecosystems. This work raises the possibility of future use of such enzymes in prevention or management of chronic inflammation.

5 Applying functional metagenomics to post-glycosylation modifications

Please note that this chapter is taken from the original publication:

Chuzel, L., Fossa, S. L., Boisvert, M. L., Cajic, S., Hennig, R., Ganatra, M. B., Reichl, U., Rapp, E., Taron, C. H. Combining functional metagenomics and glycoanalytics to identify enzymes that facilitate structural characterization of sulfated *N*-glycans. *Microbial Cell Factories*. 2021, 20:162.

5. Applying functional metagenomics to post-glycosylation modifications

5.1 Introduction to post glycosylation modifications with a focus on sulfation

Glycans can be modified with diverse chemical groups that considerably increase glycan structural heterogeneity. Chemical groups include sulfate, phosphate, methyl, acetyl, acyl, pyruvate and zwitterionic (*e.g.*, phosphorylcholine (PC) and phosphoethanolamine (PE)) moieties [93, 297, 298]. Their localization on glycans is not restricted to a precise sugar position. Instead, they are found on a variety of sugar residues on a hydroxyl or amino group (*O*- or *N*-modifications, respectively). An example of *O*- and *N*-sulfation that are both found on heparin, is shown in Figure 43 [299]. It is important to note that this notation must not be confused with *N*- or *O*-glycosylation, that refers to the chemical manner in which a glycan is attached to a protein.

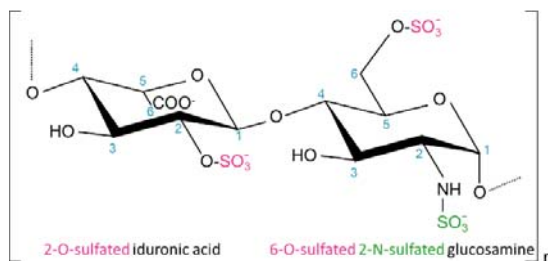


Figure 43. Structure of heparin. Heparin is a polymer of 2-*O*-sulfated iduronic acid and 6-*O*-sulfated, 2-*N*-sulfated glucosamine. Both *O*- and *N*-sulfation compose heparin repetitive unit with sulfate group located at the position C-2 of iduronic acid, C-6 of glucosamine (*O*-sulfation) and C-2 of glucosamine (*N*-sulfation).

Because these chemical groups are typically added to glycans after the formation of the glycosidic bond, they are collectively termed ‘post-glycosylation(al) modifications’ (PGMs), to be consistent with the term ‘post-translational modifications’ [297]. Their synthesis involves both transferases and chemical group donors (Table 3) [89, 300, 301]. It is believed that most PGMs derive from a single enzymatic reaction that facilitates the direct transfer

5. Applying functional metagenomics to post-glycosylation modifications

of a chemical group to a glycan moiety, with the exception of mannose-6-phosphate (Man-6-P) on *N*-glycans that results from a two-step enzymatic process [6, 94]. Man-6-P formation first involves the transfer of GlcNAc-1-phosphate from UDP-GlcNAc to the C-6 position of mannose by an N-acetylglucosamine-phosphotransferase. Following addition of GlcNAc-1-phosphate, a phosphodiester glycosidase hydrolyzes the freshly added GlcNAc residue, leaving the phosphate group exposed on mannose. Yet, the formation of PGMs has been for the most part overlooked, and some doubts remain as to the exact donors of some chemical groups and the chronology of their synthesis.

Chemical group	Donor	Enzyme
Acetyl	Acetyl CoA	O-acetyltransferase
Acyl	Acyl-CoA (?)	O-acyltransferase
Methyl	S-adenosylmethionine (SAM)	O-methyltransferase
Phosphate		
N-glycans (Man-6-P)	UDP-GlcNAc	N-acetylglucosamine-phosphotransferase + phosphodiester glycosidase
O-mannose glycans/GAGs	ATP	Kinase
Phosphorylcholine (PC)	CDP-choline or PC (?)	PC-transferase
Phosphoethanolamine (PE)	CDP-ethanolamine or PE (?)	PE-transferase
Pyruvate	Phosphoenolpyruvate	Pyruvyltransferase
Sulfate	3'-phosphate-5'-phosphosulfate (PAPS)	Sulfotransferase

(?) *hypothesized*

Table 3. Enzymes and donors involved in post glycosylation modifications (PGMs). The table is adapted from [89].

PGMs are found in all classes of glycoconjugates including protein *N*- and *O*-glycans, glycosphingolipids, GPI-anchored glycoproteins and glycosaminoglycans (GAGs) on proteoglycans. GAGs are the type of glycans for which PGMs are the most abundant. GAGs are linear polymers comprised of repeating disaccharide units. There are several different types of GAGs: heparin, heparan sulfate, keratan sulfate, chondroitin sulfate, dermatan sulfate,

5. Applying functional metagenomics to post-glycosylation modifications

and hyaluronan [302]. Most GAGs are heavily sulfated except for hyaluronan. In addition to GAGs, sulfation is found on mammalian and invertebrate *N*- and *O*-glycans. In contrast, *O*-methylation is not observed in mammals but is instead found on various bacteria and plant polysaccharides, and on mollusk and worm glycoproteins [303]. Acylation which most common form consists in the addition of an acetyl group, occurs on Sias of a wide-range of glycoconjugates and species including humans [201, 304]. Phosphorylation is commonly found on *N*-glycans of lysosomal proteins and was also reported on *O*-mannose-type glycans and GAGs [305, 306]. Pyruvate sugar substitutions are typically found on bacteria, yeast, and algae cell envelope glycans, but is believed to be absent in humans [307]. Finally PC and PE zwitterionic groups have been found on various classes of glycoconjugates of different eukaryotes and prokaryotes [298]. PC is frequently observed on polysaccharides from bacteria as well as on *N*-glycans and glycolipids from nematodes, while PE is a constituent of mammalian and fungal GPI anchored glycoproteins and has been observed on insect *N*- and *O*-glycans, fungal *N*-glycans, and cellulose from *E. coli* biofilms [300, 308–313].

PGMs influence a wide variety of biological functions. Due to the breadth of the functions attributed to PGMs, an exhaustive summary was not attempted in this section. Instead, some examples of PGM roles in the context of *O*- and *N*-glycans of eukaryotic glycoproteins is given. For example, Man-6-P present on *N*-glycans of certain secretory pathway proteins directs their transport to the lysosome [94]. *O*-acetylation of Sias plays a key role in the propagation of certain viruses. For instance, 9-*O*-acetylated-Neu5Ac (also referred to as 9-*O*-sia or Neu5,9Ac₂) constitutes the receptor determinant for influenza C virus and human coronavirus hCoV-OC43 [314, 315]. In a mechanism similar to influenza A neuraminidase-mediated virion release (see Chapter 4), influenza C virus and hCoV-OC43 express a sialate-9-*O*-acetyltransferase that directs release of newly formed virions. In addition, 9-*O*-sia is a marker of lymphoblastic leukemia [316, 317]. Sulfated *N*-glycans are prevalent on both influenza virus A neuraminidase and hemagglutinin surface proteins, and likely

5. Applying functional metagenomics to post-glycosylation modifications

play a role in virus replication and virulence [95, 96]. A GlcNAc-6-SO₄ α2-6-sialyl LacNAc *N*-glycan epitope on B lymphocytes is the preferred ligand for the CD22 receptor, indicating a role in immunoregulation [93]. Finally, sulfated *N*-glycans present on serum immunoglobulin G were recently proposed as potential biomarkers for rheumatoid arthritis [318].

The study of PGMs has been largely hampered by technical challenges, and better tools are needed to aid their analysis. Because sulfation represents a major type of PGMs, I made it the focus of the last chapter of this thesis. I utilized functional metagenomics combined with glycoanalytics to identify enzymes that facilitate structural characterization of sulfated *N*-glycans. Sulfation has been observed on all classes of glycans, including *N*- and *O*-glycans. In eukaryotes, sulfation occurs in the Golgi where a sulfate group is transferred from 3'-phosphoadenosine-5'-phosphosulfate (PAPS) to specific positions on sugars by sulfotransferases (Table 3) [319]. Sulfation of N-acetylglucosamine (GlcNAc-6-SO₄), N-acetylgalactosamine (GalNAc-6-SO₄ and GalNAc-4-SO₄), galactose (Gal-3-SO₄ and Gal-6-SO₄) and mannose (Man-6-SO₄) has been observed in *N*- and *O*-glycans [95, 110, 320–324] (Figure 44).

5. Applying functional metagenomics to post-glycosylation modifications

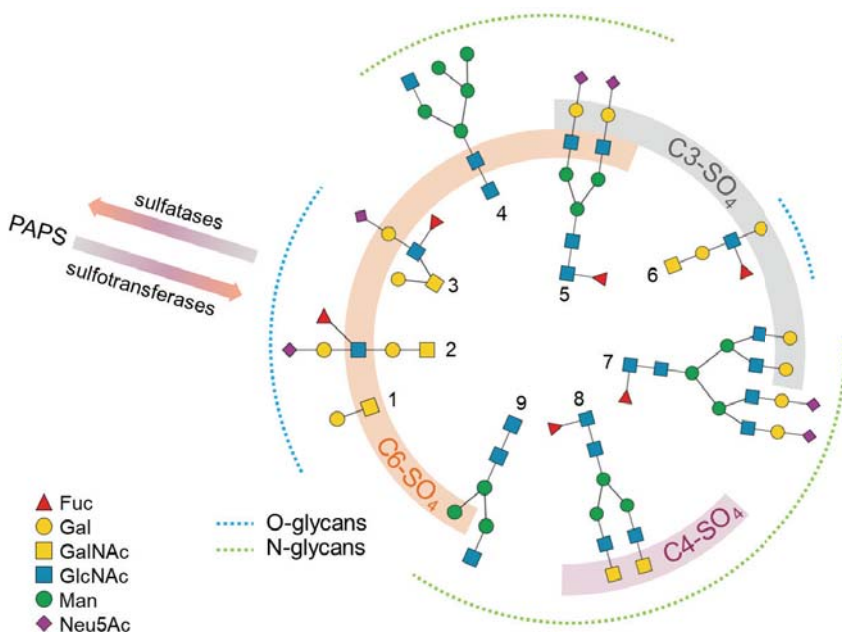


Figure 44. *N*- and *O*-glycans can be chemically modified with sulfate. Sulfation typically occurs at position C6 on GlcNAc, GalNAc, Gal or Man (orange highlight), at position C4 on GalNAc (grey highlight) or at position C3 on Gal (purple highlight). Examples of sulfated *O*-glycans from human glycoproteins (1, 2, 3 and 6) [110, 320–322], and sulfated *N*-glycan structures from Influenza virus (4), human peripheral blood cells (5), human Tamm-Horsfall glycoprotein (7), human pituitary hormones (8) [95] and *Panulirus interruptus* hemocyanin (9) [323] are shown. Glycans are represented using SNFG nomenclature [235].

Structural characterization of sulfated glycans has been accomplished using mass spectrometry (*e.g.*, LC-MS/MS or MALDI-TOF) of released glycans or glycopeptides [95, 318, 325–327]. These methods can be compromised by poor ionization of sulfated species and the natural low abundance of sulfated *N*- and *O*-glycans. Methods to enrich samples for sulfated glycans could aid their analysis. Glycan-specific binding proteins have recently been used to isolate glycans possessing certain structural features from complex samples [328–332].

5. Applying functional metagenomics to post-glycosylation modifications

However, no sugar sulfate-specific binding proteins have yet been applied to enrich sulfated glycans. Additional analytical methods involving multiplexed capillary gel electrophoresis with laser-induced fluorescence detection (xCGE-LIF) [134, 159, 161, 333, 334] and high/ultra-performance liquid chromatography coupled to fluorescence detector (H/UPLC-FLR) methods [134, 136, 138, 156, 158, 159] are also routinely used to separate glycans and match their migration- and retention-times, respectively, to those of known glycans in reference databases. These methods of glycan analysis often use highly specific exoglycosidases to enzymatically confirm the presence or absence of certain sugars [163, 164, 166]. However, enzymes that remove site-specific sugar sulfation or sulfated sugars are currently missing from the glycoanalysis toolbox.

I sought to improve the range of analytical enzymes available for analysis of sulfated glycans. To do so, the high-throughput functional metagenomic screening workflow described in Chapter 3 was applied to identify enzymes that manipulate GlcNAc-6-SO₄, a sulfated sugar found in some *O*-glycans and certain mammalian *N*-glycan outer arms. A novel coupled screening assay using a fluorescent GlcNAc-6-SO₄ analog and an exogenous hexosaminidase was devised. In this assay the coupling enzyme can act upon the substrate and release the fluorophore only if the sulfate group has first been removed by an enzyme expressed from a metagenomic clone. The screen identified a sulfatase and hexosaminidase that each act exclusively upon GlcNAc-6-SO₄. xCGE-LIF-based glycoanalysis was further employed to evaluate these enzymes activity on *N*-glycan substrates. The suitability of these enzymes for structural characterization and enrichment of *N*-glycans bearing terminal GlcNAc-6-SO₄ is also shown.

5.2 Material and methods

5.2.1 Screening for sulfated glycan using a coupled assay

Clones from the metagenomic library were screened in a lysate-based format using a coupled-assay following the protocol described in Chapter 3 (section 3.2.4.2) with adaptations. Following clone cultivation, 50 μ L of Y-PER™ lysis buffer supplemented with 40 μ g/mL of 4-Methylumbelliferyl N-acetyl- β -D-glucosaminide-6-sulfate (4-MU-GlcNAc-6-SO₄) (Dextra Laboratories Ltd, Reading, UK) and 1 U/mL of β -N-Acetylhexosaminidase_r (NEB) were added to each well. For the re-screen the following substrates: 4-methylumbelliferyl-N-acetyl- β -D-glucosaminide (4-MU-GlcNAc) (Dextra Laboratories Ltd) and 4-methylumbelliferyl-sulfate (4-MU-SO₄) (MilliporeSigma) were used in a similar protocol.

5.2.2 F1-ORF13 and F10-ORF19 hexosaminidase *in vivo* expression and purification

F1-ORF13 sulfatase and F10-ORF19 hexosaminidase were cloned for *in vivo* expression. F1-ORF13 was synthesized with codon optimization for expression in *E. coli* and cloned with a C-terminal 6xHis-tag into a pET21a(+) vector by GenScript (Piscataway, NJ, USA). F10-ORF19 hexosaminidase was cloned without its predicted signal sequence (residues 1-22) and with a C-terminal 6-His-tag using the NEBuilder® HiFi DNA Assembly Cloning kit. Primers were designed using the NEBuilder® online tool (<https://nebuilder.neb.com/>) (Appendix: Supplementary table 1). The pET28c(+) vector was modified to contain a transcription terminator sequence upstream of the promoter region for tighter regulation of protein expression.

F1-ORF13 and F10-ORF19 were expressed in NEB T7 Express *E. coli* cells. For each ORF, a 1 L culture of LB containing 100 μ g/mL ampicillin was inoculated with a single colony and induced for protein expression by addition of IPTG to 0.4 mM final concentration once the culture OD₆₀₀ reached 0.4-0.6. Induction was performed overnight at 16°C. Cells were harvested and

5. Applying functional metagenomics to post-glycosylation modifications

resuspended in loading buffer (20 mM sodium phosphate, 500 mM NaCl, 20 mM imidazole, pH 7.4). Lysis was performed using a HPL6 High Pressure Homogenizer (Maximator GmbH, Nordhausen, Germany) for 3 passes at 10 kPsi. Proteins were purified on a HisTrap Fast Flow column. Bound proteins were eluted from the column with a linear gradient of imidazole concentrations from 20 to 500 mM. Fractions containing the protein of interest were pooled and dialyzed against imidazole-free buffer (20 mM sodium phosphate, 500 mM NaCl, 1 mM EDTA, pH 7.4) using a 3.5 kDa MWCO Slide-A-Lyzer dialysis cassette (ThermoFisher Scientific).

5.2.3 F1-ORF13 activity on sulfated monosaccharides

Specificity of F1-ORF13 was tested on the following sulfated monosaccharides: N-acetyl-D-glucosaminide-6-O-sulfate (GlcNAc-6-SO₄) (MilliporeSigma), D-glucosamine-6-O-sulfate (GlcN-6-SO₄), N-acetyl-D-galactosamine-4-O-sulfate (GalNAc-4-SO₄) and D-galactose-4-O-sulfate (Gal-4-SO₄) (Dextra Laboratories Ltd, Reading, UK). Reactions were performed in 20 µL volume with 10 µL of *in vitro* PURExpressed F1-ORF13 or PURExpress® control (no DNA template) mixtures in 25 mM HEPES pH 8.3, 10 mM CaCl₂ with 160 µM sulfated monosaccharide. Incubation was at 37°C overnight. Proteins were precipitated by addition of 80 µL of cold acetone to each 20 µL reaction and incubation at -20°C for 1 h. Samples were centrifuged at 14,000 x g for 10 min at 4°C and supernatant recovered. An additional 500 µL of acetone was added to the pellet, samples were quickly vortexed and centrifuged a second time. The supernatant was recovered and pooled with the previous one. Pooled supernatants were dried in a speed vacuum concentrator (Martin Christ Gefriertrocknungsanlagen GmbH, Osterode am Harz, Germany) and resuspended in 125 µL of water. Samples were analyzed on a Dionex ISC-5000+ a high-performance anion exchange chromatography system with pulsed amperometric detection (HPAEC-PAD) with a CarboPac™ PA200 3x50 mm guard column and a CarboPac™ PA200 3x250 mm analytical column (Dionex, Sunnyvale, CA). A 25 µL volume of sample or 25 µM standard monosaccharide dissolved in water was injected using the full loop injection mode. Separation

5. Applying functional metagenomics to post-glycosylation modifications

was performed at 30°C using water as eluent A, 1 M sodium acetate in 1 mM NaOH as eluent B and 100 mM NaOH as eluent C and the following gradient: from 0-12min: 90% A - 10% C, from 12-25 min: 70% A - 20% B - 10% C, from 25-28 min: 40% B - 60% C and from 28-39: 40% B - 60% C.

5.2.4 Enzyme activities on *N*-glycans

5.2.4.1 *N*-glycan release and APTS-labeling

N-glycans were released from human Immunoglobulin A (hIgA) (Athens Research and Technology, Athens, GA) or human urokinase (Active Bioscience, Hamburg, Germany) with PNGase F. To that end, four aliquots each containing 420 µg of urokinase or 210 µg of hIgA were mixed with 2% SDS for a final SDS concentration of ~0.3%. Aliquots were incubated for 10 min at 60°C. SDS was then neutralized by addition of 8% IGEPAL® CA-630 (MilliporeSigma) in PBS to a final concentration of 1.8%. Enzymatic glycan release was achieved with 7 U of PNGase F and aliquots incubated for 3 h at 37°C. Released *N*-glycans were dried and labeled via reductive amination with APTS (MilliporeSigma) as previously described [88]. Briefly, 28 µL of MQ water, 28 µL of 20 mM APTS in 3.6 M citric acid and 28 µL of reducing agent were added to each of the four aliquots. Labeling was performed for 3 h at 37°C. A volume of 1.4 mL of 80% acetonitrile (in water) was added to each of the four aliquots prior to pooling them together. Labeled *N*-glycans were cleaned-up using HILIC-SPE as previously described [88].

5.2.4.2 Substrate preparation

Substrates used for testing F1-ORF13 activity were the following APTS-labeled *N*-glycans: FA2G2S2-SO₄, FA2G2-SO₄, FA2G0-SO₄ and an *N*-glycan composed of the core FA2 with one branched fucose and two terminal sulfated GalNAcs (termed here FA2FGalNAc-(SO₄)₂). FA2G2S2-SO₄ and FA2FGalNAc-(SO₄)₂ were isolated from hIgA and urokinase respectively by *N*-glycan fractionation. Glycan fractionation was performed on a Dionex Ultimate 3000 HPLC equipped with a Jasco FP-2020+ fluorescence detector

5. Applying functional metagenomics to post-glycosylation modifications

using a TSKgel Amide-80 HR column (5 μm particle size, 25 cm x 4.6 mm; length x internal diameter) preceded by a TSKgel Amide-80 Guard (5 μm particle size, 1 cm x 4.6 mm; length x internal diameter). MQ water with 0.2% acetic acid and 0.2% triethylamine was used as solvent A and acetonitrile containing 0.1% acetic acid was used as solvent B. The gradient used for separation was: 0 min, 80% B; 0-10 min, 70% B; 10-120 min, 55% B; 120-123 min, 0% B; 130-136 min, 80% B, 136-174, 80% B. Samples were kept at 4°C prior to injection. A volume of 90 μL of sample containing 80% acetonitrile was injected and separation was performed at 30°C. The fluorescence detection wavelengths were: $\lambda_{\text{ex}}=448$ nm and $\lambda_{\text{em}}=510$ nm. Fractions were collected from 20 – 70 min by automatic peak detection. Fractions containing *N*-glycans of interest were dried and resuspended with MQ water for use as enzyme substrates. FA2G2-SO₄ was prepared from FA2G2S2-SO₄ by digestion with sialidase A. Purified FA2G2S2-SO₄ resulting from fractionation was dried, treated with sialidase A (Agilent, Santa Clara, CA) as recommended by the supplier and digest performed overnight at 37°C. Digested *N*-glycans were cleaned-up by HILIC-SPE as described previously [88] with the exception that the 5 washes were omitted. FA2G0-SO₄ was subsequently prepared by digesting FA2G2-SO₄ with $\beta(1-4,6)$ -galactosidase (Agilent, Santa Clara, CA), following the supplier instructions. Reaction was incubated overnight at 37°C and cleaned-up by HILIC-SPE as described for the sialidase A digest.

Substrates for F10-ORF19 testing were desialylated and degalactosylated APTS-labeled total *N*-glycans released from hIgA and urokinase FA2FGalNAc-(SO₄)₂ APTS-labeled glycan obtained by fractionation. Total *N*-glycans from hIgA were desialylated and degalactosylated using sialidase A and $\beta(1-4,6)$ -galactosidase as described above.

5. Applying functional metagenomics to post-glycosylation modifications

5.2.4.3 F1-ORF13 activity on N-glycans released from hlgA and human urokinase

F1-ORF13 was assayed separately using APTS-labeled FA2G2S2-SO₄, FA2G2-SO₄, FA2G0-SO₄ from hlgA and FA2FGalNAc-(SO₄)₂ from human urokinase. Each 10 μL reaction contained 4 μL of substrate, 4 μL of 50 mM HEPES buffer, pH 8.3, 1 μL of 100 mM CaCl₂ and 6 μg of F1-ORF13. Reactions were incubated at 37°C for 3 h and cleaned-up with HILIC-SPE. Samples were dried in a speed vacuum concentrator and resuspended with 15 μL of MQ water prior to xCGE-LIF.

F1-ORF13 optimal pH was determined using the APTS-labeled FA2G0-SO₄ N-glycan in 10 μL reactions containing 4 μL of 50 mM buffer, 1 μL of 100 mM CaCl₂ and 6 μg of F1-ORF13. Buffers tested were sodium acetate, pH 4.0 and 5.0, MES buffer, pH 5.5 and 6.6, sodium phosphate buffer, pH 7.5, HEPES, pH 8.3 and Tris buffer, pH 9.0. After 3 h incubation at 37°C, reactions were cleaned-up by HILIC-SPE and product formation monitored by xCGE-LIF.

To test F1-ORF13's metal ion requirement, 10 μl reactions were performed as described above using HEPES pH 8.3 as buffer and 1 μL of 100 mM solution of CaCl₂, MgCl₂, MnSO₄ or ZnSO₄. 1 μL of MQ water was used for the control sample. Incubation was performed for 3h at 37°C prior HILIC-SPE clean-up and xCGE-LIF analysis.

To investigate the ability of F1-ORF13 to bind the APTS-labeled FA2G0-SO₄ N-glycan substrate in absence of calcium, 10 μL solutions containing 2 μL of MES buffer pH 5.5 and 0.6 or 6 μg F1-ORF13 were incubated at 37°C for 3 h. Reactions were performed in duplicate. One set of reactions was directly subjected to HILIC-SPE, the other was digested with 1 μL of 20 mg/mL proteinase K (AppliChem GmbH, Darmstadt, Germany). Proteinase K treated samples were incubated for 2 h at 42°C, cleaned with HILIC-SPE, and analyzed by xCGE-LIF.

5. Applying functional metagenomics to post-glycosylation modifications

5.2.4.4 *N*-Glycan enrichment using F1-ORF13

For EDGE-profiling, 10 µg of F1-ORF13 was mixed with 10 µL of hIgA-APTS-labeled desialylated and degalactosylated *N*-glycans in a 110 µL final volume in 25 mM MES buffer pH 5.5. Mixture was incubated in a Thermomixer for 1 h at room temperature with gentle agitation (300 r.p.m.). Mixture was loaded onto a Nasosep® 30 KDa filter (Pall Corporation, Port Washington, NY) pre-washed twice with 500 µL of water. Filter was centrifuged for 5 min at 11,000 x *g* to recover the flow-through. Bound material was eluted by addition of 130 µL of a 0.2% SDS, 100 mM DTT solution and incubation for 20 min at 60°C. Elution fraction was recovered by centrifugation for 5 min at 11,000 x *g*. A control sample was processed in parallel in absence of F1-ORF13. Samples were dried and cleaned by HILIC-SPE as described previously [88].

5.2.4.5 F10-ORF19 activity on *N*-glycans released from hIgA and human urokinase

F10-ORF19 was assayed on the desialylated and degalactosylated APTS-labeled hIgA *N*-glycan pool or APTS-labeled urokinase FA2FGalNAc-(SO₄)₂ *N*-glycan in 10 µL reactions. Reactions were performed in GlycoBuffer 1 (NEB) using 1 µg of F10-ORF19. Incubation was performed at 37°C for 1 h. Prior to xCGE-LIF measurement, samples were cleaned-up with HILIC-SPE as described previously [88].

F10-ORF19 optimal pH was determined using the desialylated and degalactosylated APTS-labeled hIgA *N*-glycan pool as the substrate. 10 µL reactions with 4 µL of 50 mM buffer and 0.1 µg of F10-ORF19 were incubated for 30 min at 37°C. Tested buffers were: sodium acetate, pH 4.0, 4.5 and 5.0, MES buffer, pH 5.5 and 6.6, sodium phosphate, pH 7.5 and HEPES buffer, pH 8.3. Prior to xCGE-LIF analysis, samples were cleaned-up with HILIC-SPE as described previously [88].

5. Applying functional metagenomics to post-glycosylation modifications

5.2.4.6 xCGE-LIF analysis

xCGE-LIF-based glycoanalysis was performed as described [88]. Samples were run on glyXbox^{CE} systems (glyXera GmbH, Magdeburg, Germany) based on modified Applied Biosystems 3130 or 3130xl genetic analyzers using a 50 cm capillary array filled with POP-7TM polymer (ThermoFisher Scientific). For injection, samples were prepared using 1-3 μL of HILIC-SPE purified APTS-labeled samples, adding 0.5 μL of GeneScanTM 500 LIZTM size standard (ThermoFisher Scientific) diluted 1:50 in Hi-DiTM formamide (ThermoFisher Scientific), 0.5 μL of xCGE-LIF migration time normalization standard 2nd NormMix (glyXera GmbH, Magdeburg, Germany) and Hi-DiTM formamide up to a total sample volume of 12 μL . Samples were electrokinetically injected and separation was at 30°C at 15 kV. Data were analyzed using the glycoanalysis software glyXtool^{CE} (glyXera).

5.3 Results

5.3.1 Functional metagenomic screening for sulfatases

A fosmid library containing large DNA inserts (~40 kb) from human gut microbiota was created in *E. coli* (see Chapter 3). The library diversity was assessed by analyzing cloned fragments from 24 randomly picked transformants. Sequencing of fosmid inserts revealed DNA originating from various enteric bacteria species such as *Bifidobacterium longum* (3 clones), *Bifidobacterium adolescentis* (4 clones), *Bacteroides* sp. (2 clones), *Phocaeicola dorei* (2 clones), *Faecalibacterium prausnitzii* (1 clone), *Adlercreutzia* sp. (2 clones) and *Collinsella aerofaciens* (1 clone). All identified species belonged to the phyla Bacteroides, Actinobacteria or Firmicutes, consistent with the composition of a healthy individual's gut microbiota [41]. Interestingly, 4 clones (>15% of those analyzed) contained insert DNA from unknown or un-sequenced species. For 5 clones, the quality of the Sanger sequencing data did not permit a conclusion about the origin of the cloned DNA.

5. Applying functional metagenomics to post-glycosylation modifications

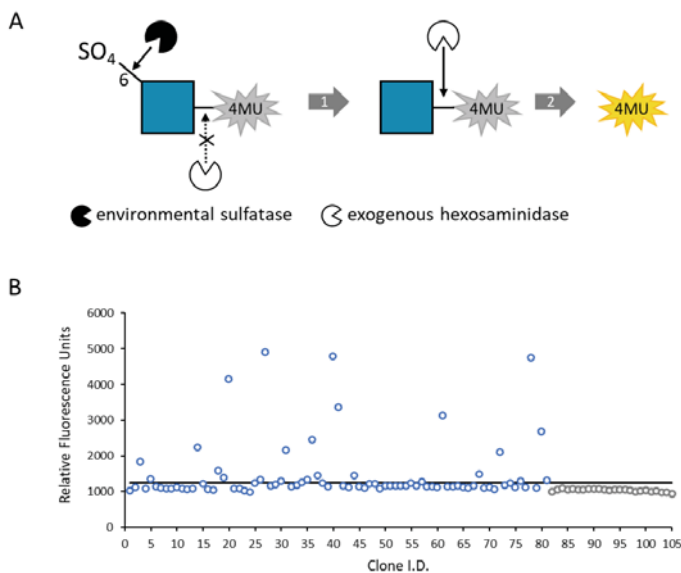


Figure 45. Screening a human gut metagenomic library for sulfatases. (A) Coupled assay for sulfatase screening. The substrate consists of a GlcNAc residue linked to the fluorophore 4-methylumbelliferone (4-MU). The GlcNAc molecule is modified at carbon 6 with a sulfate group (SNFG notation, GlcNAc, blue square). In a first reaction, a sulfatase expressed from a metagenomic clone removes the C6 sulfate. In a second reaction, an exogenous hexosaminidase (that is inactive on the sulfated substrate) in the assay mixture liberates the fluorophore from GlcNAc, generating a fluorescence signal. (B) Re-screening human gut metagenomic clones for sulfatase activity. Represented are fluorescence values for the 30h timepoint. Values above the mean + 6σ (black line) were considered ‘hits’. Control lysates from clones carrying an empty pCC1 fosmid (grey circles) were assayed along with the metagenomic clones (blue circles) from the primary screen.

In this study, identification of enzymes proficient at removing sulfate from the 6-carbon (C6) of GlcNAc was sought. A coupled assay using the substrate 4-methylumbelliferyl N-acetyl- β -D-glucosaminide-6-sulfate (4-MU-GlcNAc-6-SO₄) and an exogenous GlcNAc-6-SO₄-resistant hexosaminidase was devised (Figure 45A). In this assay, the activity of a sulfatase alone is not sufficient to

5. Applying functional metagenomics to post-glycosylation modifications

release the 4-MU fluorophore. Thus, the assay contains an exogenous hexosaminidase that is blocked by the presence of sulfate at the C6 position of GlcNAc. This enzyme enables fluorescence generation only if the C6-sulfate has been removed by a cloned environmental sulfatase. It is important to note that this screening strategy will also identify hexosaminidases that are able to directly hydrolyze C6-sulfated GlcNAc from 4-MU.

A primary screen was conducted using the 4-MU-GlcNAc-6-SO₄/hexosaminidase coupled assay with 11,520 cell lysates (half the library) comprising the human gut metagenomic DNA clone collection. A total of 81 “hits” were identified by measuring an increase in fluorescence over time. A hit was defined as a fluorescence reading 3 standard deviations over the mean background fluorescence value in at least one measured time point. This definition was intentionally liberal to capture all potential hits. The 81 hits were then re-screened using the same assay, but with a more stringent hit definition consisting of a fluorescence reading over 6 standard deviations from the mean control background fluorescence in at least two timepoints. This secondary analysis yielded 24 hits (Figure 45B) indicating an overall screen hit rate of 0.2%.

Enzyme activity in lysates from these 24 clones was then assessed using different substrates: i) 4-MU-SO₄ to detect general sulfatase activity, ii) 4-MU-GlcNAc (no sulfate) to detect hexosaminidase activity, and iii) 4-MU-GlcNAc-6-SO₄ (without exogenous hexosaminidase) to detect the activity of hexosaminidases that are not inhibited by the sulfate moiety. Twenty out of the 24 hits showed activity on 4-MU-GlcNAc-6-SO₄ in absence of exogenous hexosaminidase while 11 of these also retained activity on non-sulfated GlcNAc. These data suggest that most fosmids likely encoded a combination of both hexosaminidases and sulfatases. Only 6 hits were active on the general sulfatase substrate 4-MU-SO₄, suggesting the presence of a sulfatase that does not exclusively recognize GlcNAc-6-SO₄, and implying that the observed sulfatases from other clones may be sugar specific-sulfatases that strictly

5. Applying functional metagenomics to post-glycosylation modifications

hydrolyze sulfate located on a sugar ring. These biochemical observations were further reconciled with nucleotide sequencing of fosmid inserts.

5.3.2 Analysis of fosmid DNA sequences

Two multiplexed Pacific Bioscience (PacBio) libraries were constructed as described in Chapter 3, section 3.2.5.2 to sequence fosmids isolated from all 24 clones. Nineteen were successfully assembled in single contigs with insert sizes of 30-45 kb. These nucleotide sequences termed F1-F19 were deposited to GenBank under the accession numbers MW677166-MW677184. Five fosmids could not be sequenced properly with attempted assemblies resulting in multiple contigs or failing due to a lack of sequence coverage. Contamination of the fosmid preparation with *E. coli* genomic DNA, the presence of long repeats, or a co-culture of two distinct clones in the same well of the 384 well plate may be responsible for the failed assemblies [335]. As such, these 5 clones were not further analyzed.

Open reading frames (ORFs) encoded by each of the remaining 19 clones were predicted using MetaGeneMark. For each ORF, a MegaBLAST search against the NCBI protein repository was performed, and ORF maps for each of the 19 inserts were drawn (Appendix: Supplementary Figure 6). For two clones (F13 and F15), genes constituting the riboflavin biosynthesis pathway (ribD, ribE, ribB/A and ribH) were observed (Appendix: Supplementary Figure 6N and P) [196]. Multiple metabolites produced by this pathway, including riboflavin and luminazine, are fluorescent compounds having excitation/emission spectra that partially overlap with that of 4-MU [197, 336]. These clones were also fluorescent in our assay in the absence of the 4-MU-GlcNAc-6-SO₄ substrate (Figure 46), indicating they were false positives.

5. Applying functional metagenomics to post-glycosylation modifications

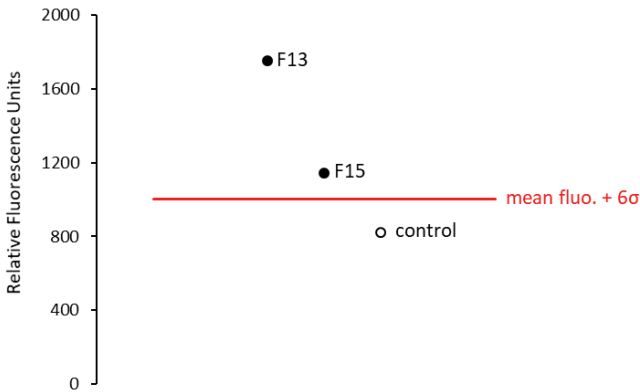


Figure 46. Background fluorescence of F13 and F15. Fluorescence at 365/445 nm was measured for F13 and F15 in absence of 4-MU-substrate. Fluorescence generated by both clones is higher than the control and higher than 6 standard deviation above the mean (red line).

It is noteworthy that the riboflavin synthesis pathway has been consistently identified in other screens conducted in our laboratory where 4-MU substrates have been used. Sequence analysis also showed that 5 clones (F14, F16, F17, F18, and F19) did not encode obvious sulfatase, hexosaminidase, or fluorescent metabolite producing pathways (Appendix: Supplementary Figure 6O, Q, R, S, T). However, these five clones had very low fluorescence signals in secondary screening. As such, 7 fosmids (F13-F19) were dismissed from this initial study and the remaining 12 fosmids (F1-F12) were further analyzed.

The sequences of fosmids F1-F12 were assessed for the presence of sulfatases and hexosaminidases from known protein families (Figure 47A). In total, 16 sulfatase genes encoded by 9 of the 12 clones were identified by homology, with 7 clones each encoding two putative sulfatases. Interestingly, of the 9 clones having putative sulfatase genes, 7 also encoded a putative hexosaminidase. Overall, 10 hexosaminidases were identified (Figure 47A). The fosmids F1-F12 were also compared to each other using Circos plots

5. Applying functional metagenomics to post-glycosylation modifications

generated by Circoletto (Figure 47B) [337]. This revealed that clones F3 and F4 both harbored the same region of the *Phocaeicola dorei* genome and were closely related to portions of F9. These three clones lacked predicted sulfatase genes but contained the same hexosaminidase gene, suggesting they might encode a protein with the ability to directly release intact GlcNAc-6-SO₄ from 4-MU-GlcNAc-6-SO₄. Clones F8 and F12 were also highly similar fosmid containing the same genome locus of a species related to *Bacteroides cellulosilyticus* with ~60% sequence identity for the analyzed fragments. Finally, a group of 4 highly related sequences was comprised of fosmids F6, F7, F11 and F2 (with F6 and F7 being nearly identical). Considered together, the Circos analysis showed that 8 sulfatase genes out of the 16 initially identified were distinct. The redundancy in fosmids F1-F12 suggests that the number of clones we screened was sufficient to explore the full diversity of the library. Thus, the 8 distinct sulfatase genes and the hexosaminidase gene (common to fosmids F3, F4 and F9) were selected for further biochemical exploration.

5. Applying functional metagenomics to post-glycosylation modifications

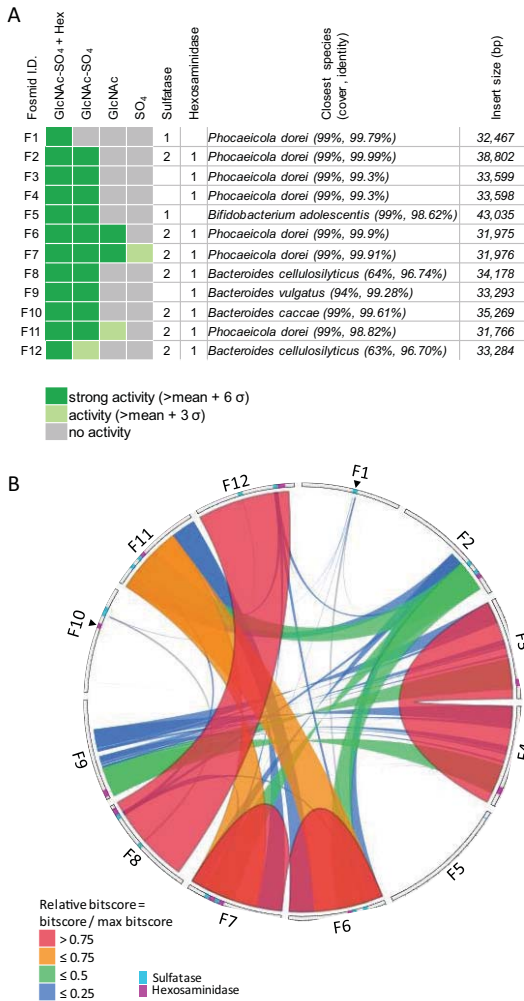


Figure 47. Function- and sequence-based analysis of fosmids F1-F12. (A) Rescreening analysis of fosmids F1-F12. Fosmids F1-F12 were tested using four distinct assays: using 4-MU-GlcNAc-6-SO₄ with exogenous hexosaminidase (GlcNAc-SO₄ + Hex), 4-MU-GlcNAc-6-SO₄ in the absence of exogenous enzyme (GlcNAc-SO₄), asulfated 4-MU-GlcNAc (GlcNAc), and 4-MU-SO₄ (SO₄). For each fosmid, shown is its activity in each assay, the predicted number of

5. Applying functional metagenomics to post-glycosylation modifications

sulfatase and hexosaminidase genes found within its DNA sequence, the closest species of origin as determined by BLAST against the entire NCBI nucleotide collection, and the size of the cloned insert. (B) Sequence alignment of fosmids F1-F12. A Circos plot made with Circoletto software illustrates relatedness in the DNA sequences of the 12 clones. Ribbons are colored using BLAST bitscores within Circoletto.

5.3.3 Identifying genes encoding active sulfatases using *in vitro* protein expression

To rapidly test candidate sulfatase genes for activity, each was expressed *in vitro* using the PURExpress® *in vitro* transcription/translation system and assayed for activity as described in Chapter 4.2.1.3 (primers are reported in Appendix: Supplementary table 1) (Figure 48).

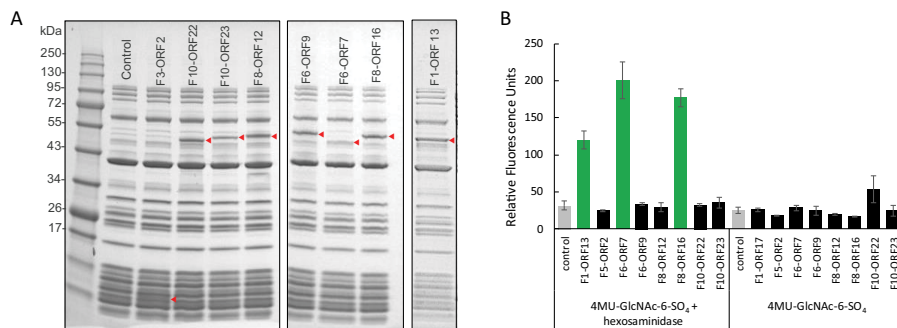


Figure 48. *In vitro* expression of sulfatase candidates. (A) SDS-PAGE of 8 sulfatases produced with the PURExpress® system. Expressed enzymes are shown with a red triangle. (B) Activity of *in vitro* expressed sulfatases. Sulfatases were assayed on 4-MU-GlcNAc-6-SO₄ in the presence and absence of exogenous hexosaminidase.

Each of the 8 sulfatase genes expressed protein of the expected size (Figure 48A). Of these, three proteins (F1-ORF13, F6-ORF7 and F8-ORF16) also showed activity on 4-MU-GlcNAc-6-SO₄ supplemented with hexosaminidase but showed no activity in control reactions lacking hexosaminidase (Figure 48B). Interestingly, F10 encoded two putative sulfatases (F10-ORF22 and F10-ORF23) and neither showed activity (Figure 48B).

5. Applying functional metagenomics to post-glycosylation modifications

5.3.4 Protein sequence analysis of active sulfatases

The deduced protein sequences of the 3 active sulfatases (F1-ORF13, F6-ORF7 and F8-ORF16) were compared to proteins present in GenBank. All three enzymes had strong homology to enzymes from *Phocaeicola* and *Bacteroides* species (Figure 49).

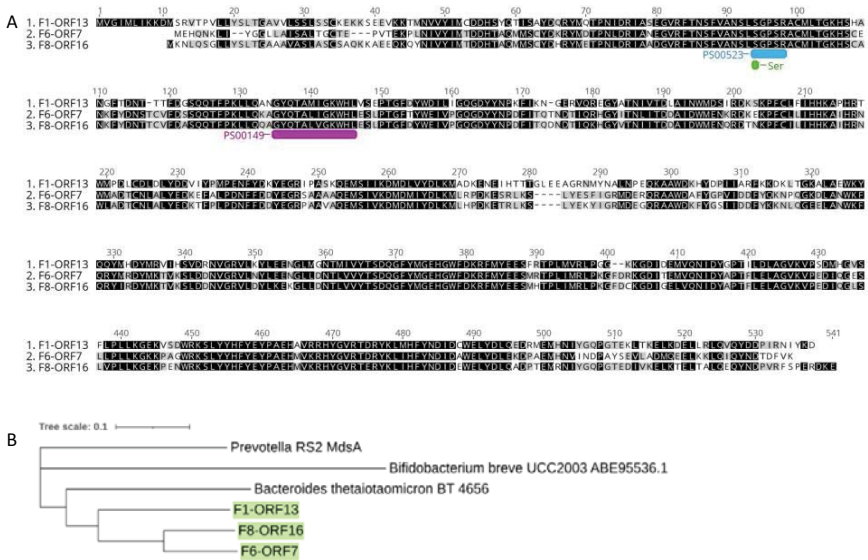


Figure 49. ORF7, 13 and 16 relationships to other GlcNAc-6-sulfatases. (A) F6-ORF7, F1-ORF13 and F8-ORF16 protein sequence alignment using the BLOSSUM62 score matrix. All three protein have a high sequence similarity. They belong to the family S1 formylglycine-dependent sulfatases. The consensus motifs from this family PS00149 and PS00523 are annotated in purple and blue respectively. The critical catalytic residue Ser is highlighted in green. (B) Phylogenetic tree of members of the family S1 sulfatase. Protein highlighted in green were identified in this work.

During secondary screening, clone F1 was the only clone that retained activity solely on 4-MU-GlcNAc-6-SO₄ in presence of an exogenous

5. Applying functional metagenomics to post-glycosylation modifications

hexosaminidase (Figure 47A). As such, F1-ORF13 was chosen for further biochemical characterization.

5.3.5 Characterization of F1-ORF13 sulfatase

5.3.5.1 Determination of F1-ORF13 sulfatase specificity using sulfated monosaccharides

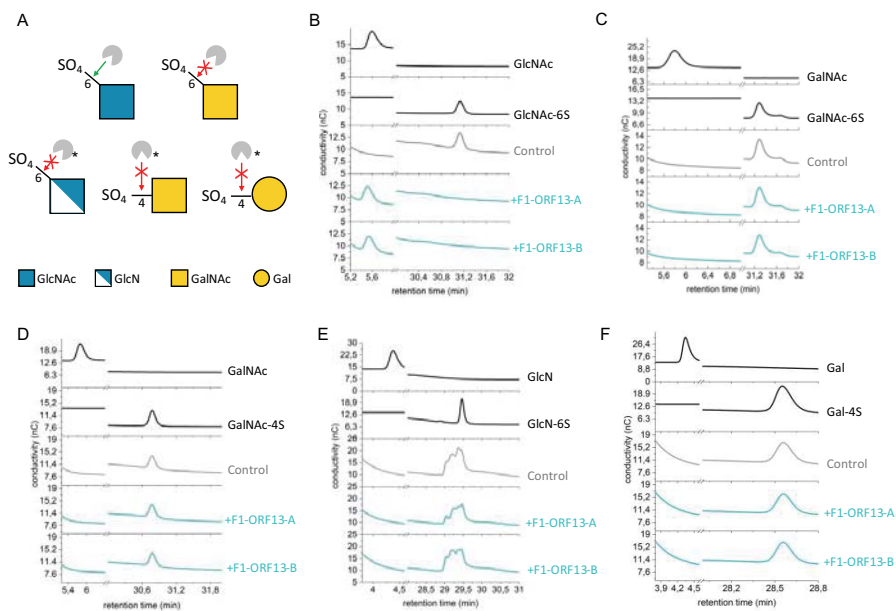


Figure 50. F1-ORF13 sulfatase specificity. (A) Summary of the specificity of F1-ORF13 on monosaccharide substrates. Monosaccharides are represented using SNFG nomenclature [235]. A red crossed arrow indicates no activity of F1-ORF13 on these substrates. (B-F) Activity of in vitro expressed F1-ORF13 sulfatase on GlcNAc-6-SO₄ (B), GalNAc-6-SO₄ (C), GalNAc-4-SO₄ (D), GlcN-6-SO₄ (E) and Gal-4-SO₄ (F). Analysis was performed by high-performance anion-exchange chromatography with pulsed amperometric detection. Reactions were performed in duplicate (F1-ORF13-A and -B – blue chromatograms). Control reactions consists of in vitro expression in the absence of F1-ORF13 (grey chromatogram). Sulfated and non-sulfated monosaccharides dissolved in water were run as retention time standards (black chromatograms).

5. Applying functional metagenomics to post-glycosylation modifications

To determine the sulfate specificity of F1-ORF13, its activity on the sulfated monosaccharides GlcNAc-6-SO₄, GlcN-6-SO₄, GalNAc-6-SO₄, GalNAc-4-SO₄ and Gal-4-SO₄ was evaluated (Figure 50).

Sulfated monosaccharides and their corresponding asulfated forms were each incubated with F1-ORF13, and reaction products were separated by high-performance anion-exchange chromatography with pulsed amperometric detection (HPAEC-PAD). Sulfate removal was only observed for GlcNAc-6-SO₄ and no other substrates, indicating that F1-ORF13 is highly selective for sulfation of carbon-6 of GlcNAc. Furthermore, the enzyme's inability to remove sulfate from similar monosaccharides GlcN-6-SO₄ or GalNAc-6-SO₄ implies the acetyl group on GlcNAc or the stereochemistry around carbon-4 are likely important determinants for substrate recognition.

5.3.5.2 F1-ORF13 sulfatase activity on GlcNAc-6-SO₄ in intact *N*-glycans

While the screen was designed to isolate enzymes that act upon a sulfated monosaccharide analog 4-MU-GlcNAc-6-SO₄, a broader aim of the study was to find enzymes that act on sulfated GlcNAc in the context of an intact *N*-glycan. As such, the activity of the F1-ORF13 sulfatase was evaluated on a sulfated *N*-glycan substrate isolated from a mammalian glycoprotein via analysis by electrophoretic separation using xCGE-LIF. To enable this experiment, the F1-ORF13 sulfatase was cloned and expressed *in vivo* in *E. coli* (Appendix: Supplementary Figure 7).

Partially purified enzyme was incubated with *N*-glycans enzymatically released from human immunoglobulin A (hIgA) and reductively labeled with the fluorophore APTS. HIgA contains 2 or 5 *N*-glycans per heavy chain for the subclasses A1 or A2, respectively [338]. The pool of glycans released from hIgA contains a form of a sialylated biantennary *N*-glycan having a single sulfate group on one outer-arm GlcNAc residue (referred to as FA2G2S2-SO₄) (Cajic S, Hennig R, Grote V, Reichl U, Rapp E; manuscript in preparation). To create a substrate with sulfated GlcNAc positioned in the terminal position of

5. Applying functional metagenomics to post-glycosylation modifications

the outer arm, FA2G2S2-SO₄ was treated with sialidase A and β (1-4,6)-galactosidase to generate FA2G2-SO₄ and FA2G0-SO₄ substrates. Enzymatically treated and untreated *N*-glycans were analyzed and compared using glyXbox^{CE}, based on xCGE-LIF. The F1-ORF13 sulfatase cleaved sulfate from terminal GlcNAc on FA2G0-SO₄ (Figure 51A bottom panel) but did not show activity on FA2G2-SO₄ (Figure 51A middle panel) nor FA2G2S2-SO₄ (Figure 51A top panel).

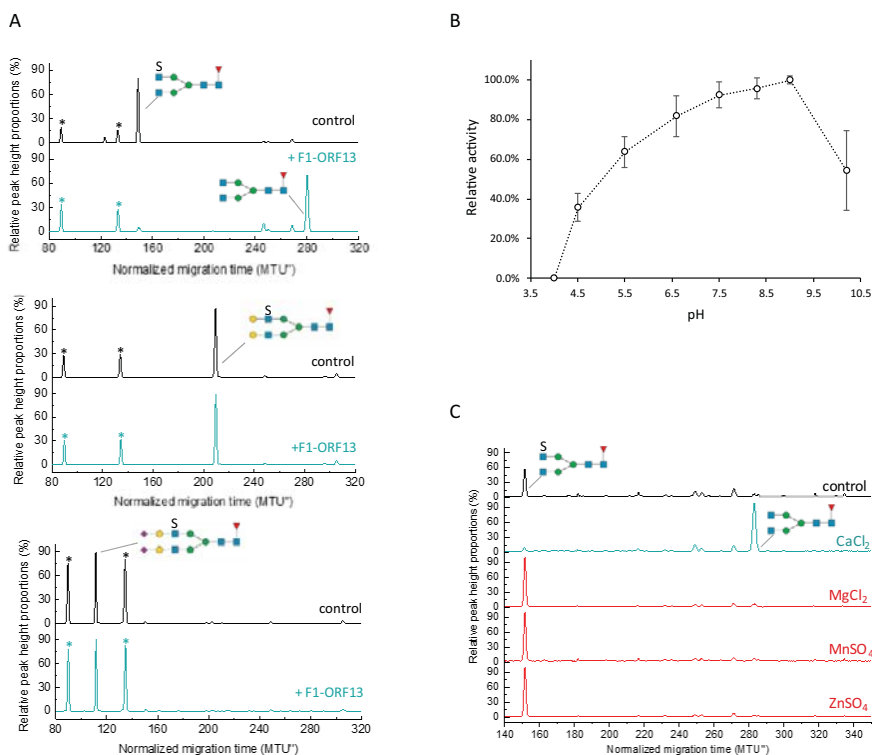


Figure 51. F1-ORF13 sulfatase activity on *N*-glycans. (A) F1-ORF13 sulfatase substrate specificity. F1-ORF13 was assayed on different APTS-labeled *N*-glycan substrates generated from human immunoglobulin. Substrate and product were monitored by xCGE-LIF. Structure assignment was enabled by comparing the normalized migration times with those of a *N*-glycan database and confirmed by exoglycosidase digests (Cajic S, Hennig R, Grote V, Reichl

5. Applying functional metagenomics to post-glycosylation modifications

U, Rapp E; manuscript in preparation). (B) F1-ORF13 optimal pH. Optimal pH for F1-ORF13 was determined using APTS-labeled FA2G0-SO₄ and following product formation by xCGE-LIF. C. F1-ORF13 ion requirement. Activity of F1-ORF13 was assayed in the presence of different metal ions. Formation of the reaction product was monitored by xCGE-LIF. The x-axis (time) of the electropherograms was normalized to two internal standards by glyXtoolCE, resulting in double normalized migration time units (MTU^{''}). Signal Intensities were normalized to the total peak height, resulting in relative peak height proportions (%). Peaks marked with (*) correspond to the internal standard used for migration time normalization in xCGE-LIF. Glycans are represented using the SNFG nomenclature [69].

F1-ORF13 was also unable to remove sulfate from an APTS-labeled *N*-glycan released from human urokinase (Figure 52).

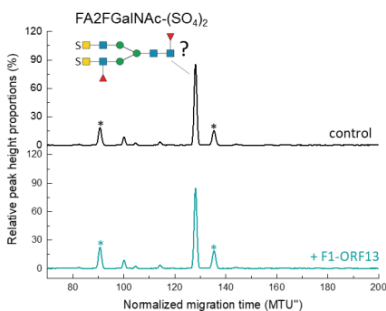


Figure 52. F1-ORF13 sulfatase activity on an APTS-labeled *N*-glycan isolated from human urokinase. F1-ORF13 sulfatase activity on an APTS-labeled *N*-glycan isolated from human urokinase was assessed by xCGE-LIF. A urokinase APTS-labeled *N*-glycan containing two terminal sulfated GalNAc residues (FA2FGalNAc-(SO₄)₂) was used as a substrate to test F1-ORF13 sulfatase (black electropherogram). The exact structure of the substrate is marked with (?) as assigned based on the literature [339, 340] and not fully experimentally confirmed. No activity could be detected as shown with the absence of migration of the peak correspond to the substrate (blue electropherogram). The x-axis (time) of the electropherograms was normalized to two internal standards by glyXtoolCE, resulting in double normalized migration time units (MTU^{''}). Signal Intensities were normalized to the total peak height, resulting in relative peak height proportions (%). Peaks marked with (*) correspond to the internal standard used for migration time normalization in xCGE-LIF. Glycans are represented using the SNFG nomenclature [235].

In this protein the dominant *N*-glycan (termed FA2FGalNAc-(SO₄)₂) contains two sulfated terminal GalNAc residues, with an additional antenna fucose at the adjacent GlcNAc [339, 340]. These experiments support the

5. Applying functional metagenomics to post-glycosylation modifications

conclusion that the F1-ORF13 sulfatase can remove sulfate from GlcNAc-6-SO₄ in the context of an intact *N*-glycan, but only if it is the terminal residue in an outer-arm branch.

The FA2G0-SO₄ substrate was also used to determine pH and ion requirements of F1-ORF13 via xCGE-LIF. The enzyme is active over pH 5-10 with a peak around pH 8-9 (Figure 51B). Consistent with other proteins of this enzyme family, F1-ORF13 is calcium-dependent (Figure 51C) [341, 342]. No hydrolysis of sulfate from FA2G0-SO₄ was detected unless Ca²⁺ ions were present (Figure 51C). Nevertheless, in the absence of calcium, the height of the peak corresponding to FA2G0-SO₄ decreased upon addition of F1-ORF13 sulfatase suggesting that F1-ORF13 was binding to the substrate.

5.3.5.3 F1-ORF13 binds GlcNAc-6-SO₄-containing *N*-glycans in absence of calcium

While testing F1-ORF13 metal ion requirement via xCGE-LIF, F1-ORF13 was incubated with APTS-labeled FA2G0-SO₄ *N*-glycans in absence of calcium to serve as a control. During this experiment I noticed that the peak corresponding to FA2G0-SO₄ was decreasing but not shifting in migrating time. I hypothesized that in the absence of Ca²⁺, F1-ORF13 bound sulfated-GlcNAc *N*-glycans. The complex formed would be too large to be eluted during the HILIC-SPE clean-up performed before each xCGE-LIF run. To test this hypothesis, different amounts of F1-ORF13 were incubated with FA2G0-SO₄ in the absence of calcium. After incubation, samples were either directly cleaned-up with HILIC-SPE or first digested with proteinase K and then, cleaned with HILIC-SPE (Figure 53). Proteinase K digestion was used to destroy F1-ORF13 and thus disrupt a possible F1-ORF13-FA2G0-SO₄ complex.

5. Applying functional metagenomics to post-glycosylation modifications

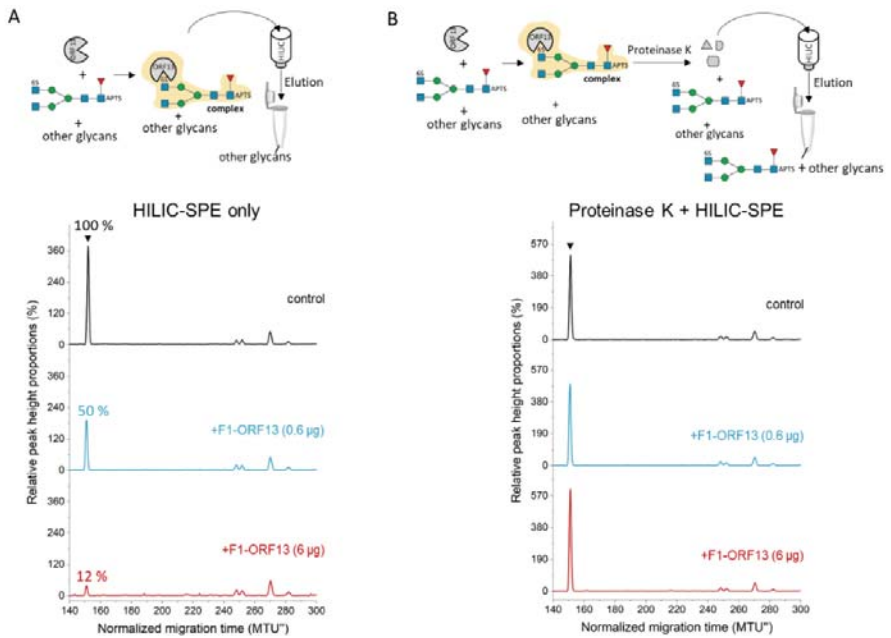


Figure 53. F1-ORF13 activity in absence of calcium. F1-ORF13 activity on an APTS-labeled N-glycan FA2Go-SO₄ (black triangle) was evaluated in absence of its cofactor. (A) HILIC-SPE cleaned reactions analyzed by xCGE-LIF. Following reaction, samples were directly cleaned-up by HILIC-SPE for subsequent analysis by xCGE-LIF. (B) Proteolyzed and HILIC-SPE cleaned reactions analyzed by xCGE-LIF. Following reaction, samples were treated with proteinase K to destroy F1-ORF13 before clean-up by HILIC-SPE and analysis by xCGE-LIF. The x-axis (time) of the electropherograms was normalized to two internal standards by glyXtoolCE, resulting in double normalized migration time units (MTU''). Signal Intensities were normalized to the height of the peaks from 240-285 MTU'', resulting in relative peak height proportions (%). Glycans are represented using the SNFG nomenclature [87]

Samples directly cleaned with HILIC-SPE, showed a clear decrease in the amount of FA2G0-SO₄ (Figure 53A). Reduction of the amount of substrate was F1-ORF13 concentration-dependent. In the samples treated with proteinase K followed by HILIC-SPE clean-up, no significant change in the relative substrate peak height was detected (Figure 53B). This confirmed the inability

5. Applying functional metagenomics to post-glycosylation modifications

of F1-ORF13 to hydrolyze SO₄ from GlcNAc in the absence of calcium as well as its capacity to recognize and bind to GlcNAc-SO₄-containing glycans. This characteristic of F1-ORF13 expands its potential as a tool to study *N*-glycan sulfation. For example, in the absence of calcium it can be used as a terminal GlcNAc-SO₄ *N*-glycan-binding protein, while in the presence of calcium, it is a GlcNAc-SO₄ sulfatase. The use of an enzyme in its apo form is an approach already employed for some nonconsuming substrate sensors [343].

To further investigate the use of F1-ORF13 as a terminal GlcNAc-SO₄-containing *N*-glycan binding protein it was tested for epitope-directed glycan enrichment (EDGE)-profiling [332]. In a proof of principle experiment, the desialylated and degalactosylated APTS-labeled hIgA *N*-glycan pool was mixed with F1-ORF13 in absence of calcium. The mixture was filtered through a Nanosep[®] 30 KDa MWCO filter. *N*-glycans unbound to F1-ORF13 were collected in the flow-through fraction after centrifugation. Bound *N*-glycans were eluted after addition of SDS and DTT. Flow-through and elution fractions were analyzed by xCGE-LIF.

5. Applying functional metagenomics to post-glycosylation modifications

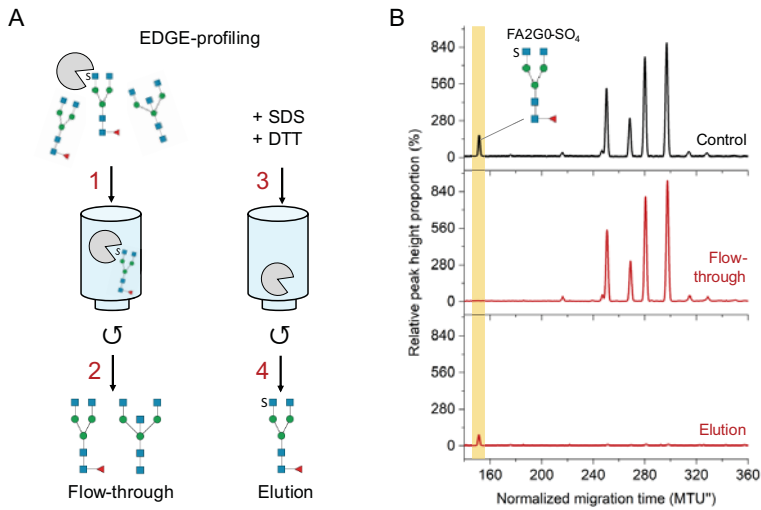


Figure 54. Use of F1-ORF13 in a N-glycan enrichment strategy. (A) Schematic of EDGE-profiling using F1-ORF13. A N-glycan pool is mixed with F1-ORF13 in absence of calcium. Mixture is applied to a 30 kDa MWCO filter. After centrifugation, N-glycans unbound to F1-ORF13 are collected in the flow-through. N-glycans containing the GlcNAc-6-SO₄ epitope at the terminal outer-arm position bind F1-ORF13 and are retained on top of the filter. Bound N-glycans are eluted with SDS and DTT. (B) xCGE-LIF analysis of the flow-through and elution fractions. The x-axis (time) of the electropherograms was normalized to two internal standards by glyXtool^{CE}, resulting in double normalized migration time units (MTU''). Signal Intensities were normalized to the height of the 2nd NormMIX standards, resulting in relative peak height proportions (%). Glycans are represented using the SNFG nomenclature [235]

The results showed that FA2G0-SO₄ was the only N-glycan structure absent in the flow-through, demonstrating selective capture of N-glycans bearing the GlcNAc-6-SO₄ epitope (Figure 54B, yellow highlight). FA2G0-SO₄ was only partially recovered in the elution fraction suggesting tight interaction with F1-ORF13.

5. Applying functional metagenomics to post-glycosylation modifications

5.3.6 Identifying genes encoding active hexosaminidases using *in vitro* protein expression

Two hexosaminidase candidates were investigated: F3-ORF26 (a gene common to F3, F4 and F9, described above) and F10-ORF19. Neither of the two putative sulfatases encoded by clone F10 showed activity on 4-MU-GlcNAc-6-SO₄ (Figure 48B). Yet, clone F10 demonstrated activity on 4-MU-GlcNAc-6-SO₄ in the presence and in the absence of exogenous hexosaminidase (Figure 47A). We hypothesized that the putative hexosaminidase gene also encoded by this clone (F10-ORF19) might be solely responsible for the fluorescent signal detected during screening. To test candidate hexosaminidase genes for activity, both genes (F3-ORF26 and F10-ORF19) were expressed *in vitro* using the PURExpress® system and assayed for activity, as described Chapter 4, section 4.2.1.3 (primers are reported in Appendix: Supplementary table 1) (Figure 55).

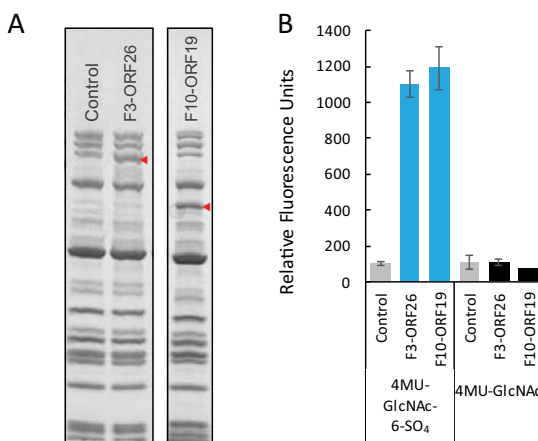


Figure 55. *In vitro* expression of hexosaminidase clones F3-ORF26 and F10-ORF19. (A) SDS-PAGE of 2 hexosaminidases produced with the PURExpress® system. Expressed enzymes are shown with a red triangle. (B) Activity of *in vitro* expressed hexosaminidases. Hexosaminidases were assayed on 4-MU-GlcNAc-6-SO₄ and its asulfated counterpart 4-MU-GlcNAc.

5. Applying functional metagenomics to post-glycosylation modifications

Both F3-ORF26 and F10-ORF19 were produced efficiently in the PURExpress® system (Figure 55A). Furthermore, they both showed the ability to hydrolyze 4-MU-GlcNAc-6-SO₄ indicating they are not inhibited by the presence of C6 sulfate on GlcNAc (Figure 55B). Interestingly, both enzymes showed no activity on asulfated 4-MU-GlcNAc, indicating they require sulfated GlcNAc for hydrolysis.

5.3.7 Protein sequence analysis of active hexosaminidases

The protein sequences of F10-ORF19 and F3-ORF26 hexosaminidases were compared and were 62.5% similar (Figure 56).

5. Applying functional metagenomics to post-glycosylation modifications

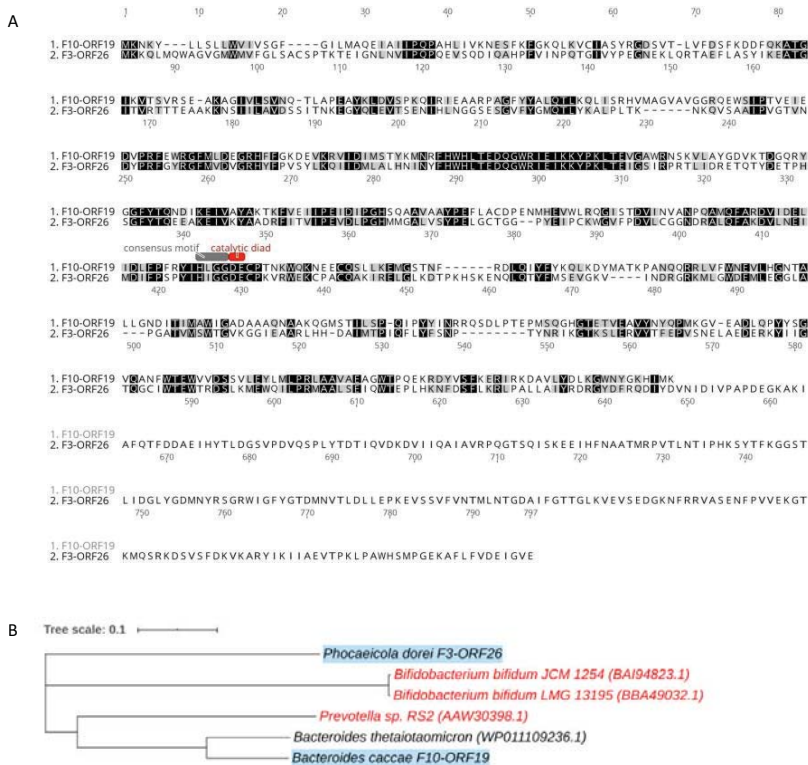


Figure 56. F10-ORF19 and F3-ORF26 are members of glycoside hydrolase family 20 (GH20). (A) F10-ORF19 and F3-ORF26 protein sequence alignment using the BLOSSUM62 score matrix. Both proteins are members of the GH20 family and contain the H-X-G-G consensus motif of this family (grey box). Catalytic diad formed with amino acid D-E is annotated in red. (B) Phylogenetic tree of GH20 members. Proteins in red were characterized in other studies and their ability to hydrolyze sulfated GlcNAc formerly demonstrated [344, 345]. Protein highlighted in blue were identified in this work.

5. Applying functional metagenomics to post-glycosylation modifications

F3-ORF26 contained a C-terminal region of ~230 amino acids absent in F10-ORF19. Both proteins belong to glycoside hydrolase family 20 (GH20) and perfectly match protein sequences of N-acetyl-hexosaminidases from *Bacteroides caccae* (F10-ORF19) and *Phocaeicola dorei* (F3-ORF26). To the best of our knowledge, neither the *Bacteroides caccae* nor *Phocaeicola dorei* hexosaminidases have been previously biochemically characterized. However, hexosaminidases from other organisms that are able to hydrolyze sulfated GlcNAc residues have been reported under the names ‘sulfoglycosidase’ or ‘mucin-desulfating glycosidase’ [346, 347]. Originating from *Prevotella* or *Bifidobacterium bifidum* these enzymes were tested on synthetic p-nitrophenyl substrates or mucin type O-glycans but not on more complex structures like sulfated N-glycans. A phylogenetic tree showing the evolutionary relationship between F10-ORF19, F3-ORF26, and other GH20 members is shown in Figure 56B. The F10-ORF19 hexosaminidase from *Bacteroides caccae* was selected for further biochemical characterization on sulfated N-glycans.

5.3.8 F10-ORF19 hexosaminidase activity upon GlcNAc-6-SO₄ in intact N-glycans

F10-ORF19 hexosaminidase was expressed *in vivo* in *E. coli* and purified to be tested in the context of N-glycans. The activity of F10-ORF19 hexosaminidase was investigated by xCGE-LIF on the pool of hIgA APTS-labeled glycans previously digested with both sialidase A and $\beta(1-4,6)$ -galactosidase to expose GlcNAc-SO₄ at the terminal end (Figure 57A black electropherogram).

5. Applying functional metagenomics to post-glycosylation modifications

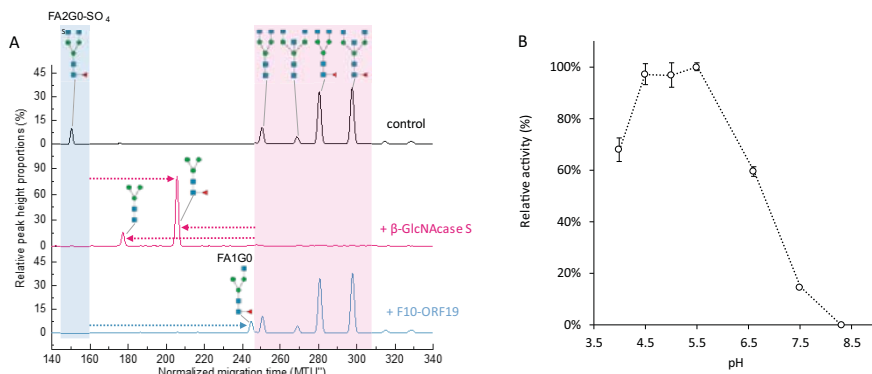


Figure 57. F10-ORF19 hexosaminidase activity on N-glycans. (A) F10-ORF19 hexosaminidase substrate specificity. F10-ORF19 was assayed on a pool of human APTS-labeled immunoglobulin A (hIgA) N-glycans pretreated with sialidase and β -galactosidase (black electropherogram). Substrate and product were analyzed by xCGE-LIF. The activity of F10-ORF19 (blue electropherogram) is indicated with a blue arrow that shows the migration time shift of the FA2G0-SO₄ peak from \sim 150 MTU'' to the FA1G0 peak at \sim 245 MTU'' due to the loss of a sulfated GlcNAc. The activity of β -N-acetylhexosaminidase S (β -GlcNAcase S; pink electropherogram) is shown with 3 pink arrows illustrating the collapse of all structures to paucimannose at (\sim 178 MTU'') and fucosylated paucimannose (\sim 207 MTU''), respectively. Structure assignment was enabled by matching the normalized migration times with those of a N-glycan database and confirmed by exoglycosidase digests (Cajic S, Hennig R, Grote V, Reichl U, Rapp E; manuscript in preparation). (B) F10-ORF19 optimal pH. Optimal pH for F10-ORF19 was determined using the hIgA-APTS labeled glycan pool as substrate and following product formation by xCGE-LIF. The x-axis (time) of the electropherograms was normalized to two internal standards by glyXtoolCE, resulting in double normalized migration time units (MTU''). Signal intensities were normalized to the total peak height, resulting in relative peak height proportions (%). Glycans are represented using SNFG nomenclature [235].

F10-ORF19 was capable of hydrolyzing GlcNAc-SO₄ while showing no activity on asulfated GlcNAc (Figure 57A, blue electropherogram). In the presence of a large excess of enzyme, weak activity on asulfated GlcNAc is observed (data not shown) but in the conditions reported here, the enzyme showed selectivity for sulfated GlcNAc residues. This specificity contrasts with

5. Applying functional metagenomics to post-glycosylation modifications

that of a well-characterized hexosaminidase from *Streptococcus pneumoniae* that is commonly used in glycoanalytics (β -N-acetylglucosaminidase S) that hydrolyzed both sulfated and non-sulfated GlcNAc as well as bisecting GlcNAc (Figure 57, shaded pink box). F10-ORF19 was active from pH 4 to 7 with an optimal activity around pH 5 (Figure 57B). F10-ORF19 also showed no activity on urokinase APTS-labeled FA2FGalNAc-(SO₄)₂ (Figure 58) indicating it does not act upon sulfated GalNAc, and is specific for GlcNAc-6-SO₄.

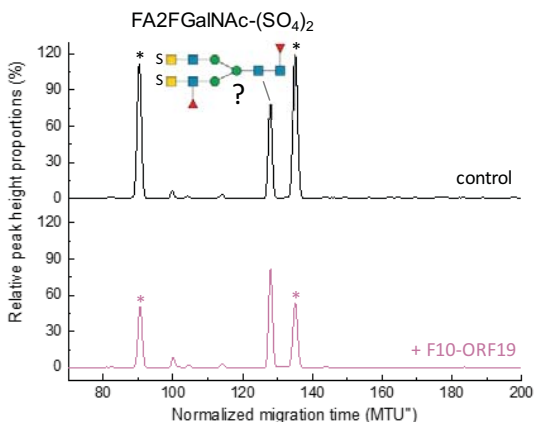


Figure 58. F10-ORF19 hexosaminidase activity on a urokinase isolated N-glycan. Activity was assessed by xCGE-LIF. A urokinase APTS-labeled N-glycan containing two terminal sulfated GalNAc residues (FA2FGalNAc-(SO₄)₂) was used as a substrate to test F10-ORF19 sulfatase (black electropherogram). The exact structure of the substrate is marked with (?) as assigned based on the literature [339, 340] and not fully experimentally confirmed. No activity could be detected as shown by the absence of migration time shift of the peak corresponding to the substrate (pink electropherogram). The x-axis (time) of the electropherograms was normalized to two internal standards by glyXtoolCE, resulting in double normalized migration time units (MTU^{*}). Signal Intensities were normalized to the total peak height, resulting in relative peak height proportions (%). Peaks marked with (*) correspond to the internal standard used for migration time normalization in xCGE-LIF. Glycans are represented using the SNFG nomenclature [235].

5.4 Chapter 5 conclusion

A human enteric metagenomic library was explored for clones encoding sugar-specific sulfatases using function-based enzyme screening with a coupled assay strategy. Enzymes that could address site-specific sulfation in *N*-glycans, and potentially be used as analytical tools were sought. The screen identified both sulfatases and hexosaminidases that act specifically upon GlcNAc-6-SO₄. Two enzymes isolated from our screen, a sulfatase (F1-ORF13) and a hexosaminidase (F10-ORF19), were cloned, purified, and further characterized. I showed that F1-ORF13 is a calcium-dependent sulfatase that exclusively removes sulfate from carbon-6 of GlcNAc. The enzyme will catalyze sulfate removal from both a GlcNAc-6-SO₄ monosaccharide or GlcNAc-6-SO₄ when it is in the terminal position of an *N*-glycan outer-arm. Furthermore, in the absence of Ca²⁺ ions, the apo-enzyme will act as a GlcNAc-6-SO₄ glycan binding protein (GBP). Lectins and GBPs have been employed in glycoanalytic workflows to enrich or deplete glycans/glycopeptides bearing specific epitopes [191, 329, 331] from complex mixtures, but none specifically address sulfated glycosylation to date. This feature is of particular interest considering the low abundance of sulfated glycans in most biological samples. Traces of GlcNAc-6-SO₄-containing *N*-glycans could be captured prior to analysis, enabling their detection otherwise concealed by non-sulfated structures. An additional study further evaluating F1-ORF13 as a GlcNAc-6-SO₄ GBP for use in glycoanalytics is ongoing. Finally, we demonstrated that the F10-ORF19 hexosaminidase is capable of hydrolyzing GlcNAc-6-SO₄ from the terminal position of *N*-glycans while showing negligible activity on asulfated GlcNAc. Considered together, the narrow specificity of both enzymes for terminal GlcNAc-6-SO₄ in *N*-glycans, and their ability to be used with other exoglycosidases, highlights their utility as novel tools to confirm the presence or absence of GlcNAc-6-SO₄ in glycoanalytics.

The F1-ORF13 sulfatase belongs to the S1 family of sulfatases in which members possess an intriguing modification of a critical serine or cysteine to formylglycine (FGly) forming two subtypes Ser-type and Cys-type S1

5. Applying functional metagenomics to post-glycosylation modifications

sulfatases [341, 348]. Discovered by von Figura's group in 1995, the FGly modification is unique to S1 sulfatases and thought to be essential for their activity [349]. To date, two main enzymatic systems that enable formation of FGly have been identified: i) formylglycine generating enzymes (FGEs) found in eukaryotes and some prokaryotes, and ii) anaerobic sulfatase maturing enzymes (anSMEs) exclusively found in prokaryotes [350, 351]. Surprisingly, in this work, expression of active F1-ORF13 was obtained both *in vitro* and *in vivo* in *E. coli* in the absence of an obvious FGE or anSME. Two hypotheses may explain these observations. First, *E. coli* might have an intrinsic ability to convert F1-ORF13 Ser94 to FGly. Similar observations have been reported for an active *Klebsiella* Ser-type sulfatase that was successfully produced in *E. coli* [352]. Sulfatase maturation systems in *E. coli* are still not completely understood, yet two putative anSME proteins (aslB and ydeM) have been described and a third, still unidentified FGly-forming system is suspected to exist [350]. The potential presence of a putative sulfatase maturation system within fosmid F1 was also investigated. While none of the 29 ORFs predicted on F1 shared homology with known FGEs or anSMEs, F1-ORF2 encoded a hypothetical protein containing a predicted 4Fe-4S cluster, a known hallmark of sulfatase maturing enzymes [353]. However, our initial attempts to co-express F1-ORF2 and F1-ORF13 did not increase the activity of F1-ORF13 (data not shown). A second hypothesis is that the F1-ORF13 sulfatase might possess some activity even in the absence of the FGly modification. This notion is supported by the activity of F1-ORF13 we detect after *in vitro* expression.

The assay system used in this study proved successful at identifying enzymes that act on or require a highly specific chemical modification of GlcNAc in high-throughput screening. While functional metagenomics identified enzyme candidates, xCGE-LIF was used to evaluate their capability to act on real *N*-glycan substrates. Extending the activity testing of enzyme candidates to natural carbohydrates was essential to prove their utility in glycoanalytics. The screening assay permitted identification of enzymes that address sulfation of the 6-carbon of GlcNAc. However, it could be easily

5. Applying functional metagenomics to post-glycosylation modifications

adapted to screen for enzymes that address different chemical modifications, a different position of modification, or modification of other sugars. For example, it would be of interest to identify sulfatases that can specifically act on GalNAc-4-SO₄ or Gal-3-SO₄, both modifications found on *N*-glycans, a class of glycoconjugates for which the impact of sulfate groups has been poorly understood [95]. In addition, the assay could be extended to identify enzymes that act upon sugars having other PGMs (*e.g.* methyl, acetyl, phosphate, etc.) that are found in a wide range of eukaryotic glycans [354]. Furthermore, no enzymes have been described that can hydrolyze zwitterionic glycan modifications (*e.g.*, phosphoethanolamine, phosphocholine) that are commonly found on glycolipid anchor glycans [312], bacterial biofilm cellulose [355] and invertebrate *N*-glycans [354]. The biological importance of PGM modifications appears vast [93, 94], and expansion of the analytical enzyme toolbox to enable their characterization will help the field gain further insight into the roles of PGMs in glycobiology.

6 Conclusion and outlook

The aims of this thesis were to i) establish a functional metagenomics workflow for enzyme discovery, ii) create a large collection of metagenomic libraries, and iii) use the devised workflow and constructed libraries to address fundamental and application-driven questions in the field of glycobiology.

In this thesis, functional metagenomics was applied to the field of glycobiology, and screening was established for enzymes that act on glycoconjugates. Screening methodologies that used either plate-based and lysate-based assays were employed. However, the quantitative data obtained using lysate-based screening was a significant advantage. Screening projects presented in this thesis, as well as those currently ongoing at NEB, now exclusively use this strategy. In my thesis work, screening assays used fluorogenic monosaccharide analog substrates. Enzymes were identified by their ability to cleave the bond between the fluorophore and the monosaccharide. However, these substrates are synthetic sugar analogs, and the extent of their hydrolysis often differs from that of natural, underivatized glycans. Thus, a key element in the success of the projects presented in my work was the characterization of enzyme function on natural substrates using glycoanalytical workflows such as H/UPLC or xCGE-LIF.

In the presented work, three metagenomic libraries were screened for enzymes that act on *N*-glycans. A thermal spring metagenomic library was screened for sialidases and led to the discovery of a novel enzyme family, termed GH156. This screen illustrated the full potential of adopting a sequence-independent approach to enzyme discovery, as the enzyme family that was discovered would not have been identified using computational bioinformatics. In a second screen, a compost metagenomic library was surveyed for sialidases able to hydrolyze an important form of animal sialic acid (Neu5Gc). Two enzymes were discovered with a novel and unusual preference for Neu5Gc. This project illustrated that this screening approach can identify enzymes with entirely new specificities. Finally, a human gut microbiome library was screened for GlcNAc-6-SO₄ sulfatases. This screen isolated sugar-specific sulfatases and sulfate-specific hexosaminidases. Both activities

6. Conclusion and outlook

represent new and important additions to the glycoanalytical toolbox and can enhance characterization of sulfated *N*-glycans. This project emphasized the ability of this screening methodology to identify enzymes with specificities that can help solve applied problems.

A major aspect of this thesis was the creation of a large collection of metagenomic and genomic fosmid DNA libraries. Large-insert metagenomic DNA libraries from diverse sources including compost, soils, ocean water, pond water, hot spring water, and human feces, were constructed in *E. coli*. Sanger sequencing of random clones from each library revealed that they all contained genomic DNA from previously unsequenced or unknown species, emphasizing the power of this resource to drive enzyme discovery. The library collection was further enhanced with large-insert genomic DNA libraries from single organisms, primarily those that populate extreme environments. The collection I assembled comprised 99,456 clones archived in 259 barcoded 384-well plates that are stored at -80°C in the form of microculture glycerol stocks.

The library collection and screening workflows presented in this thesis are currently being utilized at NEB to conduct other screens not included in this thesis. An NEB research group led by Dr. Gardner, mines thermophilic libraries for DNA-repair enzymatic activities to better understand archaeal DNA repair pathways. Most archaea are extremophiles that are capable of living in stressful environments with physicochemical conditions that subject their genomic DNA to damages. As such, archaea have adapted and encode numerous DNA repair pathways, many of which still need to be elucidated [356]. Dr. Gardner's group has identified several DNA repair activities in the *Thermococcus kodakarensis* (*Tko*) using the fosmid library constructed in *E. coli* (manuscript in preparation). In a second example, Samantha Fossa from the Taron Lab has adopted assay design inspired by the one used in my sulfatase screen (Chapter 5) to find enzymes capable of removing zwitterionic modifications of glycans. Screening the human gut microbiome library, she has identified a sugar-specific enzyme that is able to remove phosphocholine (PC) from GlcNAc in *N*-glycans, a specificity not previously described.

While there is no doubt that this method of enzyme discovery is effective and powerful, there are several adjustments that could be made to this process to make it more efficient. For application at an enzyme company like NEB, one remaining bottleneck is the speed at which screens can be executed. There are several ways screening throughput can be improved. First, the three glycobiology screens I conducted in this thesis used a single substrate in the screening assay. To increase screening throughput, multiplexed screening assays could be considered (Figure 59C). Several fluorogenic substrates could be mixed in an initial screen. Identified hits could then be re-screened to determine which individual substrate(s) was hydrolyzed. Second, additional screening strategies could be examined. For example, high-throughput mass-spectrometry-based assays (*e.g.*, MALDI) may be particularly attractive (Figure 59C). Using rapid and precise detection of signature ions from an enzymatically produced product, would permit natural glycans, glycoproteins, or nucleic acids to be used as substrates. Such an approach could both improve throughput and enable screening for many additional classes of enzymes (*e.g.*, endoglycosidases, glycosyltransferases, proteases, DNA/RNA-active enzymes, etc). In principal, if a product mass differs from the substrate mass, an assay could be devised.

Improvements to screening infrastructure could also have a dramatic impact on screening throughput. One ongoing project seeks to shorten the time it takes to go from finding a hit to discovery of an enzyme. To this end, certain libraries from the collection are currently being fully sequenced, and sequence data for each fosmid compiled in an internal database (Figure 59D). The long-term goal is to have sequence data for all fosmids in the library collection. Thus, after discovery of a hit during screening, one could retrieve its fosmid eDNA insert sequence from the database in minutes using its plate coordinates as a unique identifier. Finally, the screening workflow could benefit from automation. NEB's plans include incorporation of automated liquid handling and an integrated robotic arm to maximize the number of samples that can be quickly processed for screening. Thus, future projects will benefit from complete

6. Conclusion and outlook

automation of screening assays, rendering screening much faster, easier, and more reproducible (Figure 59C).

In summary, in this thesis, I devised functional metagenomic workflows and libraries. These were applied to discover new enzymes in the field of glycobiology. My work both advanced our basic understanding of enzymes that act upon glycans and identified interesting specificities that have application in glycoanalytical workflows.

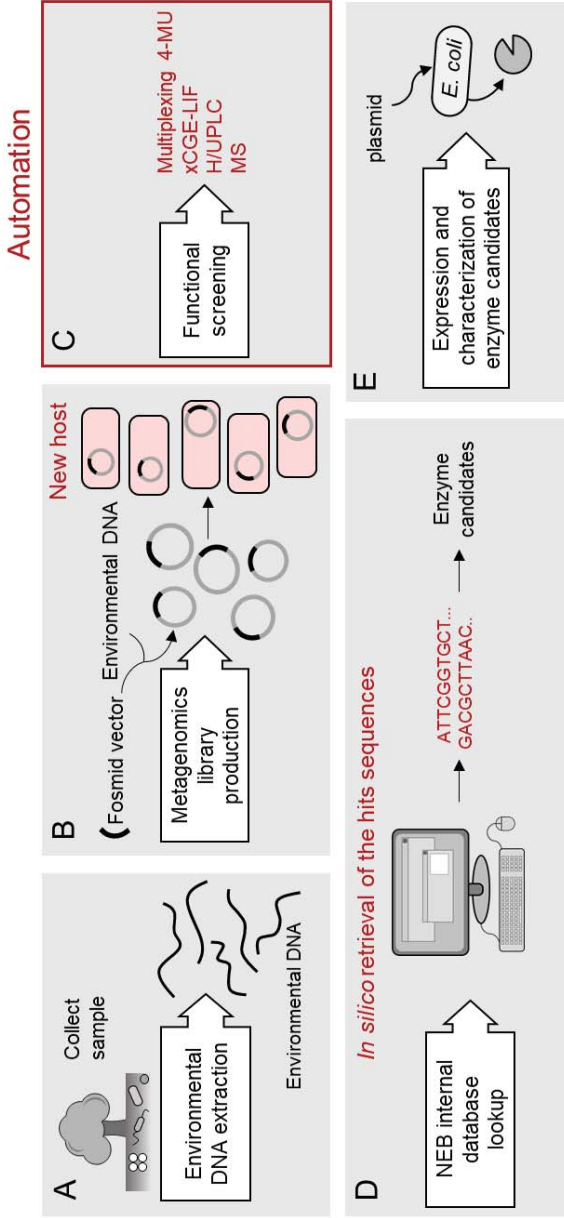


Figure 59. Improving the existing workflow. The workflow devised in this thesis would benefit from several improvements illustrated in red. Metagenomic libraries from the existing workflow were all constructed in *E. coli*. Construction of new libraries in a different host could improve expression of environmental genes (B). Devising new screening methodologies to enhance the screening assay specificity and broaden the type of enzymes one can survey would benefit future screening projects (C). Increasing the screening throughput also appears essential for faster execution of large screening projects (C). Finally, an ongoing effort is made toward sequencing existing metagenomic libraries upfront screening and compiling sequence data in a repository (D). This would enable to greatly shorten the time it takes to go from finding a hit to identify enzyme candidates. Abbreviations: 4-MU: 4-methylumbelliferone, xCGE-LIF: multiplexed capillary gel electrophoresis with laser-induced fluorescence detection, H/UPLC: high/ultra performance liquid chromatography, MS: mass spectrometry.

Bibliography

1. Olsen HS, Falholt P. The role of enzymes in modern detergency. *J Surfactants Deterg.* 1998;1:555–67. doi:10.1007/s11743-998-0058-7.
2. Kralik P, Ricchi M. A basic guide to real time PCR in microbial diagnostics: Definitions, parameters, and everything. *Frontiers in Microbiology.* 2017;8 FEB:108. doi:10.3389/fmicb.2017.00108.
3. Di Felice F, Micheli G, Camilloni G. Restriction enzymes and their use in molecular biology: An overview. *Journal of Biosciences.* 2019;44. doi:10.1007/s12038-019-9856-8.
4. Bonnet M, Lagier JC, Raoult D, Khelaifia S. Bacterial culture through selective and non-selective conditions: the evolution of culture media in clinical microbiology. *New Microbes and New Infections.* 2020;34:100622.
5. Zarbock A, Ley K, McEver RP, Hidalgo A. Leukocyte ligands for endothelial selectins: specialized glycoconjugates that mediate rolling and signaling under flow. *Blood.* 2011;118:6743–51. doi:10.1182/blood-2011-07-343566.
6. Braulke T, Bonifacino JS. Sorting of lysosomal proteins. *Biochimica et Biophysica Acta - Molecular Cell Research.* 2009;1793:605–14.
7. Thotakura NR, Blithe DL. Glycoprotein hormones: Glycobiology of gonadotrophins, thyrotrophin and free α subunit. *Glycobiology.* 1995;5:3–10. doi:10.1093/glycob/5.1.3.
8. Taniguchi N, Kizuka Y. Glycans and Cancer. In: *Advances in Cancer Research.* Academic Press Inc.; 2015. p. 11–51. doi:10.1016/bs.acr.2014.11.001.
9. Thompson AJ, de Vries RP, Paulson JC. Virus recognition of glycan receptors. *Curr Opin Virol.* 2019;34:117–29. doi:10.1016/j.coviro.2019.01.004.
10. Tomana M, Schrohlenloher RE, Reveille JD, Arnett FC, Koopman WJ. Abnormal galactosylation of serum IgG in patients with systemic lupus erythematosus and members of families with high frequency of autoimmune diseases. *Rheumatol Int.* 1992;12:191–4. doi:10.1007/BF00302151.
11. He T, Zhang X. Characterization of Bacterial Communities in Deep-Sea Hydrothermal Vents from Three Oceanic Regions. *Mar Biotechnol.* 2016;18:232–41. doi:10.1007/s10126-015-9683-3.
12. Sayed AM, Hassan MHA, Alhadrami HA, Hassan HM, Goodfellow M, Rateb ME. Extreme environments: microbiology leading to specialized metabolites. *J Appl Microbiol.* 2020;128:630–57. doi:10.1111/jam.14386.
13. Locey KJ, Lennon JT. Scaling laws predict global microbial diversity. *Proc Natl Acad Sci U S A.* 2016;113:5970–5. doi:10.1073/pnas.1521291113.
14. Kaoutari A El, Armougom F, Gordon JI, Raoult D, Henrissat B. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nat Rev | Microbiol.* 2013;11. doi:10.1038/nrmicro3050.
15. White MF, Allers T. DNA repair in the archaea—an emerging picture. *FEMS Microbiol*

Rev. 2018;42:514–26. doi:10.1093/femsre/fuy020.

16. Staley JT, Konopka A. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu Rev Microbiol.* 1985;39:321–46. doi:10.1146/annurev.mi.39.100185.001541.

17. Razumov A. S. The direct method of calculation of bacteria in water: comparison with the Koch method. *Mikrobiol.* 1932;:131–46.

18. Amann RI, Ludwig W, Schleifer KH. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev.* 1995;59:143–69. doi:10.1128/mr.59.1.143-169.1995.

19. Schloss PD, Handelsman J. Biotechnological prospects from metagenomics. *Curr Opin Biotechnol.* 2003;14:303–10. doi:10.1016/S0958-1669(03)00067-3.

20. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol.* 1998;5:R245-9. doi:10.1016/s1074-5521(98)90108-9.

21. Healy FG, Ray RM, Aldrich HC, Wilkie AC, Ingram LO, Shanmugam KT. Direct isolation of functional genes encoding cellulases from the microbial consortia in a thermophilic, anaerobic digester maintained on lignocellulose. *Appl Microbiol Biotechnol.* 1995;43:667–74. doi:10.1007/BF00164771.

22. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature.* 2004;428:37–43. doi:10.1038/nature02340.

23. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science.* 2004;304:66–74. doi:10.1126/science.1093857.

24. Hood LE, Hunkapiller MW, Smith LM. Automated DNA sequencing and analysis of the human genome. *Genomics.* 1987;1:201–12. doi:10.1016/0888-7543(87)90046-2.

25. Klindworth A, Priesse E, Schweer T, Peplies J, Quast C, Horn M, et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* 2013;41:e1–e1. doi:10.1093/nar/gks808.

26. Suzuki MT, Beja O, Taylor LT, DeLong EF. Phylogenetic analysis of ribosomal RNA operons from uncultivated coastal marine bacterioplankton. *Environ Microbiol.* 2001;3:323–31. doi:10.1046/j.1462-2920.2001.00198.x.

27. Quaiser A, Ochsenreiter T, Klenk H-P, Kletzin A, Treusch AH, Meurer G, et al. First insight into the genome of an uncultivated crenarchaeote from soil. *Environ Microbiol.* 2002;4:603–11. doi:10.1046/j.1462-2920.2002.00345.x.

28. Hjort K, Bergström M, Bergström B, Adesina MF, Jansson JK, Smalla K, et al. Chitinase genes revealed and compared in bacterial isolates, DNA extracts and a metagenomic library from a phytopathogen-suppressive soil. *FEMS Microbiol Ecol.* 2010;71:197–207. doi:10.1111/j.1574-6941.2009.00801.x.

29. Schmidt O, Drake HL, Horn MA. Hitherto Unknown [Fe-Fe]-Hydrogenase Gene

Diversity in Anaerobes and Anoxic Enrichments from a Moderately Acidic Fen †. *Appl Environ Microbiol.* 2010;76:2027–31. doi:10.1128/AEM.02895-09.

30. Ravi RK, Walton K, Khosroheidari M. MiSeq: A Next Generation Sequencing Platform for Genomic Analysis. In: DiStefano JK, editor. *Disease Gene Identification: Methods and Protocols, Methods in Molecular Biology.* Springer Science+Business Media; 2018. p. 223–32. doi:10.1007/978-1-4939-7471-9_12.

31. Lazarevic V, Whiteson K, Huse S, Hernandez D, Farinelli L, Østerås M, et al. Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *J Microbiol Methods.* 2009;79:266–71. doi:10.1016/j.mimet.2009.09.012.

32. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature.* 2010;464:59–65. doi:10.1038/nature08821.

33. Rhoads A, Au KF. PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics.* 2015;13:278–89. doi:10.1016/j.gpb.2015.08.002.

34. Goodwin S, McPherson JD, Richard McCombie W. Coming of age: ten years of next-generation sequencing technologies. *Nat Publ Gr.* 2016. doi:10.1038/nrg.2016.49.

35. Slatko BE, Gardner AF, Ausubel FM. Overview of Next-Generation Sequencing Technologies. *Curr Protoc Mol Biol.* 2018;122. doi:10.1002/cpmb.59.

36. Xie H, Yang C, Sun Y, Igarashi Y, Jin T, Luo F. PacBio Long Reads Improve Metagenomic Assemblies, Gene Catalogs, and Genome Binning. *Front Genet.* 2020;11. doi:10.3389/fgene.2020.516269.

37. Kasianowicz JJ, Brandin E, Branton D, Deamer DW. Characterization of individual polynucleotide molecules using a membrane channel. In: *Proceedings of the National Academy of Sciences of the United States of America.* National Academy of Sciences; 1996. p. 13770–3. doi:10.1073/pnas.93.24.13770.

38. Kono N, Arakawa K. Nanopore sequencing: Review of potential applications in functional genomics. *Dev Growth Differ.* 2019;61:316–26. doi:10.1111/dgd.12608.

39. Payne A, Holmes N, Rakyan V, Loose M. Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files. *bioRxiv.* 2018;;312256. doi:10.1101/312256.

40. Fu S, Wang A, Au KF. A comparative evaluation of hybrid error correction methods for error-prone long reads. *Genome Biol.* 2019;20. doi:10.1186/s13059-018-1605-z.

41. Leggett RM, Clark MD. A world of opportunities with nanopore sequencing. 2017. doi:10.1093/jxb/erx289.

42. Johnson SS, Zaikova E, Goerlitz DS, Bai Y, Tighe SW. Real-Time DNA Sequencing in the Antarctic Dry Valleys Using the Oxford Nanopore Sequencer. *J Biomol Tech.* 2017;28:2–7. doi:10.7171/jbt.17-2801-009.

43. Castro-Wallace SL, Chiu CY, John KK, Stahl SE, Rubins KH, Mcintyre ABR, et al. Nanopore DNA Sequencing and Genome Assembly on the International Space Station. *Sci Rep.* 2017;7. doi:10.1038/s41598-017-18364-0.

44. Nayfach S, Roux S, Seshadri R, Udway D, Varghese N, Schulz F, et al. A genomic catalog of Earth's microbiomes. *Nat Biotechnol.* 2021;39:499–509. doi:10.1038/s41587-020-0718-6.
45. Wooley JC, Ye Y. Metagenomics: Facts and artifacts, and computational challenges. *J Comput Sci Technol.* 2010;25:71–81. doi:10.1007/s11390-010-9306-4.
46. Howe A, Chain PSG. Challenges and opportunities in understanding microbial communities with metagenome assembly (accompanied by IPython Notebook tutorial). *Front Microbiol.* 2015;6. doi:10.3389/fmicb.2015.00678.
47. Lämmle K, Zipper H, Breuer M, Hauer B, Buta C, Brunner H, et al. Identification of novel enzymes with different hydrolytic activities by metagenome expression cloning. *J Biotechnol.* 2007;127:575–92. doi:10.1016/j.jbiotec.2006.07.036.
48. Uchiyama T, Miyazaki K. Functional metagenomics for enzyme discovery: challenges to efficient screening. *Curr Opin Biotechnol.* 2009;20:616–22. doi:10.1016/j.copbio.2009.09.010.
49. Streit WR, Schmitz RA. Metagenomics – the key to the uncultured microbes. *Curr Opin Microbiol.* 2004;7:492–8. doi:10.1016/j.mib.2004.08.002.
50. Craig JW, Chang FY, Brady SF. Natural products from environmental DNA hosted in *Ralstonia metallidurans*. *ACS Chem Biol.* 2009;4:23–8.
51. Biver S. *Bacillus subtilis* as a Tool for Screening Soil Metagenomic Libraries for Antimicrobial Activities. *J Microbiol Biotechnol.* 2013;23:850–5. doi:10.4014/jmb.1212.12008.
52. Cheng J, Pinnell L, Engel K, Neufeld JD, Charles TC. Versatile broad-host-range cosmids for construction of high quality metagenomic libraries. *J Microbiol Methods.* 2014;99:27–34. doi:10.1016/j.mimet.2014.01.015.
53. Gabor EM, Alkema WBL, Janssen DB. Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environ Microbiol.* 2004;6:879–86. doi:10.1111/j.1462-2920.2004.00640.x.
54. Jeon JH, Kim J-T, Kim YJ, Kim H-K, Lee HS, Kang SG, et al. Cloning and characterization of a new cold-active lipase from a deep-sea sediment metagenome. *Appl Microbiol Biotechnol.* 2009;81:865–74. doi:10.1007/s00253-008-1656-2.
55. Forsberg KJ, Patel S, Witt E, Wang B, Ellison TD, Dantas G. Identification of Genes Conferring Tolerance to Lignocellulose-Derived Inhibitors by Functional Selections in Soil Metagenomes. 2016. doi:10.1128/AEM.02838-15.
56. Soares F, Marcon J, Pereira e Silva M, Khakhum N, Cerdeira L, Ottoni J, et al. A Novel Multifunctional β -N-Acetylhexosaminidase Revealed through Metagenomics of an Oil-Spilled Mangrove. *Bioengineering.* 2017;4:62. doi:10.3390/bioengineering4030062.
57. Li Y, Wexler M, Richardson DJ, Bond PL, Johnston AWB. Screening a wide host-range, waste-water metagenomic library in tryptophan auxotrophs of *Rhizobium leguminosarum* and of *Escherichia coli* reveals different classes of cloned *trp* genes. *Environ Microbiol.* 2005;7:1927–36. doi:10.1111/j.1462-2920.2005.00853.x.
58. Mewis K, Armstrong Z, Song YC, Baldwin SA, Withers SG, Hallam SJ. Biomining active cellulases from a mining bioremediation system. *J Biotechnol.* 2013;167:462–71.

doi:10.1016/j.jbiotec.2013.07.015.

59. Uchiyama T, Abe T, Ikemura T, Watanabe K. Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. *Nat Biotechnol.* 2005;23:88–93. doi:10.1038/nbt1048.

60. Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR, et al. Cloning the Soil Metagenome: a Strategy for Accessing the Genetic and Functional Diversity of Uncultured Microorganisms. *Appl Environ Microbiol.* 2000;66:2541–7. doi:10.1128/AEM.66.6.2541-2547.2000.

61. Enzymes Market Size & Share | Industry Report, 2020-2027. <https://www.grandviewresearch.com/industry-analysis/enzymes-industry>. Accessed 1 Jun 2021.

62. Richardson TH, Tan X, Frey G, Callen W, Cabell M, Lam D, et al. A Novel, High Performance Enzyme for Starch Liquefaction DISCOVERY AND OPTIMIZATION OF A LOW pH, THERMOSTABLE-AMYLASE*. 2002. doi:10.1074/jbc.M203183200.

63. Vester JK, Glaring MA, Stougaard P. Discovery of novel enzymes with industrial potential from a cold and alkaline environment by a combination of functional metagenomics and culturing. *Microb Cell Fact.* 2014;13. doi:10.1186/1475-2859-13-72.

64. DeCastro M-E, Rodríguez-Belmonte E, González-Siso M-I. Metagenomics of Thermophiles with a Focus on Discovery of Novel Thermozyms. *Front Microbiol.* 2016;7:1521. doi:10.3389/fmicb.2016.01521.

65. Iqbal HA, Craig JW, Brady SF. Antibacterial enzymes from the functional screening of metagenomic libraries hosted in *Ralstonia metallidurans*. *FEMS Microbiol Lett.* 2014;354:19–26. doi:10.1111/1574-6968.12431.

66. Rahfeld P, Sim L, Moon H, Constantinescu I, Morgan-Lang C, Hallam SJ, et al. An enzymatic pathway in the human gut microbiome that converts A to universal O type blood. *Nat Microbiol.* 2019;4:1475–85. doi:10.1038/s41564-019-0469-7.

67. Chauhan NS, Nain S, Sharma R. Identification of Arsenic Resistance Genes from Marine Sediment Metagenome. *Indian J Microbiol.* 2017;57:299–306. doi:10.1007/s12088-017-0658-0.

68. Ferrer M, Golyshina O V., Chernikova TN, Khachane AN, Reyes-Duarte D, Martins Dos Santos VAP, et al. Novel hydrolase diversity retrieved from a metagenome library of bovine rumen microflora. *Environ Microbiol.* 2005;7:1996–2010. doi:10.1111/j.1462-2920.2005.00920.x.

69. Verma MK, Ahmed V, Gupta S, Kumar J, Pandey R, Mandhan V, et al. Functional metagenomics identifies novel genes ABC1TPP, TMSRP1 and TLSRP1 among human gut enterotypes. *Sci Rep.* 2018;8:1397. doi:10.1038/s41598-018-19862-5.

70. Coughlan LM, Cotter PD, Hill C, Alvarez-Ordóñez A. Biotechnological applications of functional metagenomics in the food and pharmaceutical industries. *Front Microbiol.* 2015;6:672. doi:10.3389/fmicb.2015.00672.

71. Sathya TA, Khan M. Diversity of Glycosyl Hydrolase Enzymes from Metagenome and Their Application in Food Industry. *J Food Sci.* 2014;79:R2149–56. doi:10.1111/1750-3841.12677.

72. Lorenz P, Eck J. Metagenomics and industrial applications. *Nat Rev Microbiol.* 2005;3:510–6. doi:10.1038/nrmicro1161.
73. Kamble A, Srinivasan S, Singh · Harinder. In-Silico Bioprospecting: Finding Better Enzymes. *Mol Biotechnol.* 2018;61:53–9. doi:10.1007/s12033-018-0132-1.
74. Sharma N, Thakur N, Raj T, Chand Bhalla T. Mining of Microbial Genomes for the Novel Sources of Nitrilases. 2017. doi:10.1155/2017/7039245.
75. Toyama D, de Moraes MAB, Ramos FC, Zanphorlin LM, Tonoli CCC, Balula AF, et al. A novel β -glucosidase isolated from the microbial metagenome of Lake Poraquê (Amazon, Brazil). *Biochim Biophys Acta - Proteins Proteomics.* 2018;1866:569–79. doi:10.1016/j.bbapap.2018.02.001.
76. Cobb RE, Chao R, Zhao H. Directed evolution: Past, present, and future. *AIChE J.* 2013;59:1432–40. doi:10.1002/aic.13995.
77. Chen K, Arnold FH. Tuning the activity of an enzyme for unusual environments: Sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proc Natl Acad Sci U S A.* 1993;90:5618–22. doi:10.1073/pnas.90.12.5618.
78. Bolon DN, Mayo SL. Enzyme-like proteins by computational design. *Proc Natl Acad Sci U S A.* 2001;98:14274–9. doi:10.1073/pnas.251555398.
79. Meiler J, Baker D. ROSETTALIGAND: Protein-small molecule docking with full side-chain flexibility. *Proteins Struct Funct Bioinforma.* 2006;65:538–48. doi:10.1002/prot.21086.
80. Zanghellini A, Jiang L, Wollacott AM, Cheng G, Meiler J, Althoff EA, et al. New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci.* 2006;15:2785–94. doi:10.1110/ps.062353106.
81. Voigt CA, Martinez C, Wang Z-G, Mayo SL, Arnold FH. Protein building blocks preserved by recombination. 2002. doi:10.1038/nsb805.
82. O'Maille PE, Bakhtina M, Tsai M-D. Structure-based Combinatorial Protein Engineering (SCOPE). *J Mol Biol.* 2002;321:677–91. doi:10.1016/S0022-2836(02)00675-7.
83. Dalby PA. Strategy and success for the directed evolution of enzymes. *Curr Opin Struct Biol.* 2011;21:473–80. doi:10.1016/j.sbi.2011.05.003.
84. Varki A, Kornfeld S. Historical Background and Overview. In: Varki A, Cummings RD, Esko JD, Stanley P, Hart GW, Aebi M, et al., editors. *Essentials of Glycobiology.* 3rd edition. Cold Spring Harbor Laboratory Press; 2017. p. 5–25. doi:10.1057/978-1-137-51150-8_2.
85. Daniels G, Reid ME. Blood groups: the past 50 years. *Transfusion.* 2010;50:281–9. doi:10.1111/j.1537-2995.2009.02456.x.
86. Jansen BC, Bondt A, Reiding KR, Lonardi E, De Jong CJ, Falck D, et al. Pregnancy-associated serum N-glycome changes studied by high-throughput MALDI-TOF-MS. *Nat Publ Gr.* 2016. doi:10.1038/srep23296.
87. Gudelj I, Lauc G, Pezer M. Immunoglobulin G glycosylation in aging and diseases. *Cell Immunol.* 2018;333:65–79. doi:10.1016/j.cellimm.2018.07.009.
88. Hennig R, Cajic S, Borowiak M, Hoffmann M, Kottler R, Reichl U, et al. Towards

personalized diagnostics via longitudinal study of the human plasma N-glycome. *Biochim Biophys Acta - Gen Subj*. 2016;1860:1728–38. doi:10.1016/j.bbagen.2016.03.035.

89. Freeze HH, Hart GW, Schnaar RL. Glycosylation Precursors. In: Ajit Varki, Richard D Cummings, Jeffrey D Esko, Pamela Stanley, Gerald W Hart, Markus Aebi, et al., editors. *Essentials of Glycobiology*. 3rd edition. Cold Spring Harbor Laboratory Press; 2015. <http://www.ncbi.nlm.nih.gov/pubmed/28876856>. Accessed 4 May 2021.

90. Stanley P, Taniguchi N, Aebi M. N-glycans. In: Varki A, Cummings RD, Esko JD et al., editor. *Essentials of Glycobiology*. 3rd edition. Cold Spring Harbor Laboratory Press; 2017. doi:10.1101/glycobiology.3e.009.

91. Varki A. *Essentials of glycobiology*. Cold Spring Harbor Laboratory Press; 2009.

92. Berg JM, Tymoczko JL, Stryer L. Complex Carbohydrates Are Formed by Linkage of Monosaccharides. In: *Biochemistry*. 5th edition. W H Freeman; 2002. <https://www.ncbi.nlm.nih.gov/books/NBK22396/>. Accessed 30 Mar 2021.

93. Muthana SM, Campbell CT, Gildersleeve JC. Modifications of Glycans: Biological Significance and Therapeutic Opportunities. *ACS Chem Biol*. 2012;7:31–43. doi:10.1021/cb2004466.

94. Kollmann K, Pohl S, Marschner K, Encarnação M, Sakwa I, Tiede S, et al. Mannose phosphorylation in health and disease. *Eur J Cell Biol*. 2010;89:117–23. doi:10.1016/j.ejcb.2009.10.008.

95. She Y-M, Li X, Cyr TD. Remarkable Structural Diversity of N-Glycan Sulfation on Influenza Vaccines. *Anal Chem*. 2019;91:5083–90. doi:10.1021/acs.analchem.8b05372.

96. Wang CC, Chen JR, Tseng YC, Hsu CH, Hung YF, Chen SW, et al. Glycans on influenza hemagglutinin affect receptor binding and immune response. *Proc Natl Acad Sci U S A*. 2009;106:18137–42. doi:10.1073/pnas.0909696106.

97. Hoffmann M, Marx K, Reichl U, Wuhrer M, Rapp E. Site-specific O-glycosylation analysis of human blood plasma proteins. *Mol Cell Proteomics*. 2016;15:624–41. doi:10.1074/mcp.M115.053546.

98. Hoffmann M, Pioch M, Pralow A, Hennig R, Kottler R, Reichl U, et al. The Fine Art of Destruction: A Guide to In-Depth Glycoproteomic Analyses-Exploiting the Diagnostic Potential of Fragment Ions. *Proteomics*. 2018;18:1800282. doi:10.1002/pmic.201800282.

99. Kawasaki N, Ohta M, Hyuga S, Hashimoto O, Hayakawa T. Analysis of Carbohydrate Heterogeneity in a Glycoprotein Using Liquid Chromatography/Mass Spectrometry and Liquid Chromatography with Tandem Mass Spectrometry. *Anal Biochem*. 1999;269:297–303. doi:10.1006/abio.1999.4026.

100. Varki A. Biological roles of glycans. *Glycobiology*. 2017;27:3–49. doi:10.1093/glycob/cww086.

101. Schnaar RL. Glycobiology simplified: diverse roles of glycan recognition in inflammation. *J Leukoc Biol*. 2016;99:825–38. doi:10.1189/jlb.3RI0116-021R.

102. Silsirivanit A. Glycosylation markers in cancer. In: *Advances in Clinical Chemistry*. Academic Press Inc.; 2019. p. 189–213. doi:10.1016/bs.acc.2018.12.005.

103. Pearce OMT. Cancer glycan epitopes: biosynthesis, structure and function. *Glycobiology*. 2018;28:670–96. doi:10.1093/glycob/cwy023.
104. Radhakrishnan P, Dabelsteen S, Madsen FB, Francavilla C, Kopp KL, Steentoft C, et al. Immature truncated O-glycophenotype of cancer directly induces oncogenic features. *Proc Natl Acad Sci U S A*. 2014;111:E4066–75. doi:10.1073/pnas.1406619111.
105. Ju T, Lanneau GS, Gautam T, Wang Y, Xia B, Stowell SR, et al. Human tumor antigens Tn and sialyl Tn arise from mutations in Cosmc. *Cancer Res*. 2008;68:1636–46. doi:10.1158/0008-5472.CAN-07-2345.
106. Jiang Y, Liu Z, Xu F, Dong X, Cheng Y, Hu Y, et al. Aberrant O-glycosylation contributes to tumorigenesis in human colorectal cancer. *J Cell Mol Med*. 2018;22:4875–85. doi:10.1111/jcmm.13752.
107. Ju T, Aryal RP, Kudelka MR, Wang Y, Cummings RD. The Cosmc connection to the Tn antigen in cancer. *Cancer Biomarkers*. 2014;14:63–81. doi:10.3233/CBM-130375.
108. Mellis SJ, Baenziger JU. Structures of the oligosaccharides present at the three asparagine-linked glycosylation sites of human IgD. *J Biol Chem*. 1983;258:11546–56. doi:10.1016/S0021-9258(17)44262-1.
109. Hoja-Łukowicz D, Link-Lenczowski P, Carpentieri A, Amoresano A, Pocheć E, Artemenko KA, et al. L1CAM from human melanoma carries a novel type of N-glycan with Galβ1-4Galβ1- motif. Involvement of N-linked glycans in migratory and invasive behaviour of melanoma cells. *Glycoconj J*. 2013;30:205–25. doi:10.1007/s10719-012-9374-5.
110. Seko A, Ohkura T, Ideo H, Yamashita K. Novel O-linked glycans containing 6'-sulfo-Gal/GalNAc of MUC1 secreted from human breast cancer YMB-S cells: Possible carbohydrate epitopes of KL-6(MUC1) monoclonal antibody. *Glycobiology*. 2011;22:181–95. doi:10.1093/glycob/cwr118.
111. Adamczyk B, Tharmalingam T, Rudd PM. Glycans as cancer biomarkers. *Biochim Biophys Acta - Gen Subj*. 2012;1820:1347–53. doi:10.1016/j.bbagen.2011.12.001.
112. Kim EH, Misek DE. Glycoproteomics-Based Identification of Cancer Biomarkers. *Int J Proteomics*. 2011;2011:1–10. doi:10.1155/2011/601937.
113. Ruhaak LR, Miyamoto S, Lebrilla CB. Developments in the Identification of Glycan Biomarkers for the Detection of Cancer. *Mol Cell Proteomics*. 2013;12:846–55. doi:10.1074/mcp.R112.026799.
114. Planinc A, Bones J, Dejaegher B, Van Antwerpen P, Delporte C. Glycan characterization of biopharmaceuticals: Updates and perspectives. *Analytica Chimica Acta*. 2016;921:13–27. doi:10.1016/j.aca.2016.03.049.
115. Park EI, Mi Y, Unverzagt C, Gabius H-J, Baenziger JU. The asialoglycoprotein receptor clears glycoconjugates terminating with sialic acid 2,6GalNAc. *Proc Natl Acad Sci*. 2005;102:17125–9. doi:10.1073/pnas.0508537102.
116. Iida S, Kuni-Kamochi R, Mori K, Misaka H, Inoue M, Okazaki A, et al. Two mechanisms of the enhanced antibody-dependent cellular cytotoxicity (ADCC) efficacy of non-fucosylated therapeutic antibodies in human blood. 2009. doi:10.1186/1471-2407-9-58.

117. Yamane-Ohnuki N, Kinoshita S, Inoue-Urakubo M, Kusunoki M, Iida S, Nakano R, et al. Establishment of fUT8 knockout Chinese hamster ovary cells: An ideal host cell line for producing completely defucosylated antibodies with enhanced antibody-dependent cellular cytotoxicity. *Biotechnol Bioeng.* 2004;87:614–22. doi:10.1002/bit.20151.
118. Macher BA, Galili U. The Gal α 1,3Gal β 1,4GlcNAc-R (α -Gal) epitope: A carbohydrate of unique evolution and clinical relevance. 2007. doi:10.1016/j.bbagen.2007.11.003.
119. Galili U. Anti-Gal: an abundant human natural antibody of multiple pathogenesis and clinical benefits. *Immunology.* 2013;140:1–11. doi:10.1111/imm.12110.
120. Chinuki Y, Morita E. Alpha-Gal-containing biologics and anaphylaxis. *Allergy International.* 2019;68:296–300. doi:10.1016/j.alit.2019.04.001.
121. Chung CH, Mirakhur B, Chan E, Le Q-T, Berlin J, Morse M, et al. Cetuximab-Induced Anaphylaxis and IgE Specific for Galactose- α -1,3-Galactose. *N Engl J Med.* 2008;358:1109–17. doi:10.1056/NEJMoa074943.
122. Abdel-Motal UM, Guay HM, Wigglesworth K, Welsh RM, Galili U. Immunogenicity of Influenza Virus Vaccine Is Increased by Anti-Gal-Mediated Targeting to Antigen-Presenting Cells. *J Virol.* 2007;81:9131–41. doi:10.1128/JVI.00647-07.
123. Abdel-Motal UM, Wang S, Awad A, Lu S, Wigglesworth K, Galili U. Increased immunogenicity of HIV-1 p24 and gp120 following immunization with gp120/p24 fusion protein vaccine expressing α -gal epitopes. *Vaccine.* 2010;28:1758–65. doi:10.1016/j.vaccine.2009.12.015.
124. Galili U. Amplifying immunogenicity of prospective Covid-19 vaccines by glycoengineering the coronavirus glycan-shield to present α -gal epitopes. *Vaccine.* 2020;38:6487–99. doi:10.1016/j.vaccine.2020.08.032.
125. Dhar C, Sasmal A, Varki A. From “Serum Sickness” to “Xenosialitis”: Past, Present, and Future Significance of the Non-human Sialic Acid Neu5Gc. *Front Immunol.* 2019;10:807. doi:10.3389/fimmu.2019.00807.
126. Padler-Karavani V, Yu H, Cao H, Chokhawala H, Karp F, Varki N, et al. Diversity in specificity, abundance, and composition of anti-Neu5Gc antibodies in normal humans: Potential implications for disease. *Glycobiology.* 2008;18:818–30. doi:10.1093/glycob/cwn072.
127. Ghaderi D, Taylor RE, Padler-Karavani V, Diaz S, Varki A. Implications of the presence of N-glycolylneuraminic acid in recombinant therapeutic glycoproteins. *Nat Biotechnol.* 2010;28:863–7. doi:10.1038/nbt.1651.
128. Zaramela LS, Martino C, Alisson-Silva F, Rees SD, Diaz SL, Chuzel L, et al. Gut bacteria responding to dietary change encode sialidases that exhibit preference for red meat-associated carbohydrates. *Nat Microbiol.* 2019;4:2082–9. doi:10.1038/s41564-019-0564-9.
129. O’Flaherty RM, Trbojević-Akmačić I, Greville G, Rudd PM, Lauc G. The sweet spot for biologics: recent advances in characterization of biotherapeutic glycoproteins. *Expert Review of Proteomics.* 2018;15:13–29. doi:10.1080/14789450.2018.1404907.
130. Zhang P, Woen S, Wang T, Liao B, Zhao S, Chen C, et al. Challenges of glycosylation analysis and control: an integrated approach to producing optimal and consistent therapeutic

drugs. *Drug Discov Today*. 2016;21:740–65. doi:10.1016/j.drudis.2016.01.006.

131. Duke Rebecca, Taron Christopher H. N Glycan Composition Profiling for Quality Testing of Biotherapeutics. *BioPharm Int*. 2015.

132. Sethuraman N, Stadheim TA. Challenges in therapeutic glycoprotein production. *Curr Opin Biotechnol*. 2006;17:341–6. doi:10.1016/j.copbio.2006.06.010.

133. Qing G, Yan J, He X, Li X, Liang X. Recent advances in hydrophilic interaction liquid interaction chromatography materials for glycopeptide enrichment and glycan separation. *TrAC Trends Anal Chem*. 2020;124:115570. doi:10.1016/j.trac.2019.06.020.

134. Cajic S, Hennig R, Burock R, Rapp E. Capillary (gel) electrophoresis-based methods for immunoglobulin (G) glycosylation analysis. In: Pezer M, editor. Springer. 2021.

135. Ruhaak LR, Hennig R, Huhn C, Borowiak M, Dolhain RJEM, Deelder AM, et al. Optimized Workflow for Preparation of APTS-Labeled N-Glycans Allowing High-Throughput Analysis of Human Plasma Glycomes using 48-Channel Multiplexed CGE-LIF. *J Proteome Res*. 2010;9:6655–64. doi:10.1021/pr100802f.

136. Huffman JE, Pučić-Baković M, Klarić L, Hennig R, Selman MHJ, Vučković F, et al. Comparative performance of four methods for high-throughput glycosylation analysis of immunoglobulin G in genetic and epidemiological research. *Mol Cell Proteomics*. 2014;13:1598–610. doi:10.1074/mcp.M113.037465.

137. Zhao S, Walsh I, Abrahams JL, Royle L, Nguyen-Khuong T, Spencer D, et al. GlycoStore: a database of retention properties for glycan analysis. *Bioinformatics*. 2018;34:3231–2. doi:10.1093/bioinformatics/bty319.

138. Campbell MP, Royle L, Radcliffe CM, Dwek RA, Rudd PM. GlycoBase and autoGU: tools for HPLC-based glycan analysis. *Bioinformatics*. 2008;24:1214–6. doi:10.1093/bioinformatics/btn090.

139. Lauber MA, Yu Y-Q, Brousmiche DW, Hua Z, Koza SM, Magnelli P, et al. Rapid Preparation of Released N-Glycans for HILIC Analysis Using a Labeling Reagent that Facilitates Sensitive Fluorescence and ESI-MS Detection. *Anal Chem*. 2015;87:5401–9. doi:10.1021/acs.analchem.5b00758.

140. Pralow A, Cajic S, Alagesan K, Kolarich D, Rapp E. State-of-the-Art Glycomics Technologies in Glycobiotechnology. Springer, Berlin, Heidelberg; 2020. p. 1–33. doi:10.1007/10_2020_143.

141. Rapp Erdmann, Reichl Udo, editors. *Advances in Glycobiotechnology*. Springer. 2021. <https://www.springer.com/gp/book/9783030695897#aboutBook>. Accessed 27 Jun 2021.

142. Yang Y, Franc V, Heck AJR. Glycoproteomics: A Balance between High-Throughput and In-Depth Analysis. *Trends Biotechnol*. 2017;35:598–609. doi:10.1016/j.tibtech.2017.04.010.

143. Harvey DJ. Proteomic analysis of glycosylation: structural determination of N- and O-linked glycans by mass spectrometry. *Expert Rev Proteomics*. 2005;2:87–101. doi:10.1586/14789450.2.1.87.

144. Mechref Y, Novotny M V. Structural Investigations of Glycoconjugates at High Sensitivity. *Chem Rev*. 2002;102:321–70. doi:10.1021/cr0103017.

145. Balaguer E, Neusüss C. Glycoprotein Characterization Combining Intact Protein and Glycan Analysis by Capillary Electrophoresis-Electrospray Ionization-Mass Spectrometry. *Anal Chem.* 2006;78:5384–93. doi:10.1021/ac060376g.
146. Bush DR, Zang L, Belov AM, Ivanov AR, Karger BL. High Resolution CZE-MS Quantitative Characterization of Intact Biopharmaceutical Proteins: Proteoforms of Interferon- β 1. *Anal Chem.* 2016;88:1138–46. doi:10.1021/acs.analchem.5b03218.
147. Baerenfaenger M, Meyer B. Intact Human Alpha-Acid Glycoprotein Analyzed by ESI-qTOF-MS: Simultaneous Determination of the Glycan Composition of Multiple Glycosylation Sites. *J Proteome Res.* 2018;17:3693–703. doi:10.1021/acs.jproteome.8b00309.
148. Lanucara F, Eyers CE. Top-down mass spectrometry for the analysis of combinatorial post-translational modifications. *Mass Spectrometry Reviews.* 2013;32:27–42. doi:10.1002/mas.21348.
149. Gupta G, Surolia A, Sampathkumar SG. Lectin microarrays for glycomic analysis. *OMICS A Journal of Integrative Biology.* 2010;14:419–36. doi:10.1089/omi.2009.0150.
150. Desaire H. Glycopeptide Analysis, Recent Developments and Applications*. *Mol Cell Proteomics.* 2013;12:893–901. doi:10.1074/mcp.
151. Ruhaak LR, Xu G, Li Q, Goonatileke E, Lebrilla CB. Mass Spectrometry Approaches to Glycomic and Glycoproteomic Analyses. *Chem Rev.* 2018;118:7886–930. doi:10.1021/acs.chemrev.7b00732.
152. Galermo AG, Nandita E, Barboza M, Amicucci MJ, Vo T-TT, Lebrilla CB. Liquid Chromatography–Tandem Mass Spectrometry Approach for Determining Glycosidic Linkages. *Anal Chem.* 2018;90:13073–80. doi:10.1021/acs.analchem.8b04124.
153. Schindler B, Barnes L, Renois G, Gray C, Chambert S, Fort S, et al. Anomeric memory of the glycosidic bond upon fragmentation and its consequences for carbohydrate sequencing. *Nat Commun.* 2017;8. doi:10.1038/s41467-017-01179-y.
154. Mariño K, Bones J, Kattla JJ, Rudd PM. A systematic approach to protein glycosylation analysis: a path through the maze. *Nat Chem Biol.* 2010;6:713–23. doi:10.1038/nchembio.437.
155. Jensen PH, Karlsson NG, Kolarich D, Packer NH. Structural analysis of N- and O-glycans released from glycoproteins. *Nat Protoc.* 2012;7:1299–310. doi:10.1038/nprot.2012.063.
156. Rowe L, Burkhart G. Analyzing protein glycosylation using UHPLC: a review. *Bioanalysis.* 2018;10:1691–703. doi:10.4155/bio-2018-0156.
157. Royle L, Campbell MP, Radcliffe CM, White DM, Harvey DJ, Abrahams JL, et al. HPLC-based analysis of serum N-glycans on a 96-well plate platform with dedicated database software. *Anal Biochem.* 2008;376:1–12. doi:10.1016/j.ab.2007.12.012.
158. Stöckmann H, Duke RM, Millán Martín S, Rudd PM. Ultrahigh Throughput, Ultrafiltration-Based N-Glycomics Platform for Ultrapformance Liquid Chromatography (ULTRA3). *Anal Chem.* 2015;87:8316–22. doi:10.1021/acs.analchem.5b01463.
159. Pralow A, Cajic S, Alagesan K, Kolarich D, Rapp E. State-of-the-Art Glycomics Technologies in Glycobiotechnology. Springer, Berlin, Heidelberg; 2020. p. 1–33. doi:10.1007/10_2020_143.

160. Hennig R, Rapp E, Kottler R, Cajic S, Borowiak M, Reichl U. N-Glycosylation Fingerprinting of Viral Glycoproteins by xCGE-LIF. In: *Methods in Molecular Biology*. Humana Press Inc.; 2015. p. 123–43. doi:10.1007/978-1-4939-2874-3_8.
161. Hennig R, Rapp E, Kottler R, Cajic S, Borowiak M, Reichl U. N-Glycosylation Fingerprinting of Viral Glycoproteins by xCGE-LIF. In: *Methods in Molecular Biology*. 2015. p. 123–43. doi:10.1007/978-1-4939-2874-3_8.
162. Ruhaak LR, Zauner G, Huhn C, Bruggink C, Deelder AM, Wuhrer M. Glycan labeling strategies and their use in identification and quantification. *Anal Bioanal Chem*. 2010;397:3457–81. doi:10.1007/s00216-010-3532-z.
163. Rudd PM, Dwek RA. Rapid, sensitive sequencing of oligosaccharides from glycoproteins. *Curr Opin Biotechnol*. 1997;8:488–97. doi:10.1016/s0958-1669(97)80073-0.
164. Kobata A. Exo- and endoglycosidases revisited. *Proc Japan Acad Ser B*. 2013;89:97–117. doi:10.2183/pjab.89.97.
165. Szigeti M, Guttman A. Automated N-Glycosylation Sequencing Of Biopharmaceuticals By Capillary Electrophoresis. *Sci Rep*. 2017;7. doi:10.1038/s41598-017-11493-6.
166. Thiesler CT, Cajic S, Hoffmann D, Thiel C, Van Diepen L, Hennig R, et al. Glycomic characterization of induced pluripotent stem cells derived from a patient suffering from phosphomannomutase 2 congenital disorder of glycosylation (PMM2-CDG). *Mol Cell Proteomics*. 2016;15:1435–52. doi:10.1074/mcp.M115.054122.
167. Kim U-J, Shizuya H, de Jong PJ, Birren B, Simon MI. Stable propagation of cosmid sized human DNA inserts in an F factor based vector. *Nucleic Acids Res*. 1992;20:1083–5. doi:10.1093/nar/20.5.1083.
168. Wild J, Hradecna Z, Szybalski W. Conditionally Amplifiable BACs: Switching From Single-Copy to High-Copy Vectors and Genomic Clones. *Genome Res*. 2002;12:1434–44. doi:10.1101/gr.130502.
169. Schleif R. AraC protein, regulation of the l-arabinose operon in *Escherichia coli*, and the light switch mechanism of AraC action. *FEMS Microbiol Rev*. 2010;34:779–96. doi:10.1111/j.1574-6976.2010.00226.x.
170. Feiss M, Catalano CE. Bacteriophage Lambda Terminase and the Mechanism of Viral DNA Packaging. *Landes Bioscience*; 2013. <https://www.ncbi.nlm.nih.gov/books/NBK6485/>. Accessed 20 Jan 2021.
171. Hohn B. In Vitro packaging of λ and cosmid DNA. In: *Methods in Enzymology*. Academic Press, Inc; 1979. p. 299–309. doi:10.1016/0076-6879(79)68021-7.
172. Hunt M, Silva N De, Otto TD, Parkhill J, Keane JA, Harris SR. Circulator: Automated circularization of genome assemblies using long sequencing reads. *Genome Biol*. 2015;16:1–10. doi:10.1186/s13059-015-0849-0.
173. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res*. 2010;38:e132–e132. doi:10.1093/nar/gkq275.
174. Dmitriev B, Toukach F, Ehlers S. Towards a comprehensive view of the bacterial cell wall. *Trends Microbiol*. 2005;13:569–74. doi:10.1016/j.tim.2005.10.001.

175. Salazar O, Asenjo JA. Enzymatic lysis of microbial cells. *Biotechnol Lett.* 2007;29:985–94. doi:10.1007/s10529-007-9345-2.
176. Harrison ST. Bacterial cell disruption: A key unit operation in the recovery of intracellular products. *Biotechnol Adv.* 1991;9:217–40. doi:10.1016/0734-9750(91)90005-G.
177. Shahriar M, Haque MR, Kabir S, Dewan I, Bhuyian MA. Effect of proteinase-K on Genomic DNA extraction from gram-positive strains. *Stamford J Pharm Sci.* 2011;4:53–7. doi:10.3329/sjps.v4i1.8867.
178. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. Diversity, stability and resilience of the human gut microbiota. *Nature.* 2012;489:220–30. doi:10.1038/nature11550.
179. Lloyd-Price J, Abu-Ali G, Huttenhower C. The healthy human microbiome. *Genome Med.* 2016;8:51. doi:10.1186/s13073-016-0307-y.
180. Bergstrom KSB, Xia L. Mucin-type O-glycans and their roles in intestinal homeostasis. *Glycobiology.* 2013;23:1026–37. doi:10.1093/glycob/cwt045.
181. Derrien M, Belzer C, de Vos WM. *Akkermansia muciniphila* and its role in regulating host functions. *Microb Pathog.* 2017;106:171–81. doi:10.1016/j.micpath.2016.02.005.
182. Plummer TH, Elder JH, Alexander S, Phelan AW, Tarentino AL. Demonstration of peptide:N-glycosidase F activity in endo-beta-N-acetylglucosaminidase F preparations. *J Biol Chem.* 1984;259:10700–4. <http://www.ncbi.nlm.nih.gov/pubmed/6206060>.
183. Sun G, Yu X, Bao C, Wang L, Li M, Gan J, et al. Identification and Characterization of a Novel Prokaryotic Peptide. *J Biol Chem.* 2015;290:7452–62. doi:10.1074/jbc.M114.605493.
184. Koutsoulis D, Landry D, Guthrie EP. Novel endo- α -N-acetylgalactosaminidases with broader substrate specificity. *Glycobiology.* 2008;18:799–805. doi:10.1093/glycob/cwn069.
185. Huber R, Eder W, Heldwein S, Wanner G, Huber H, Rachel R, et al. *Thermocrinis ruber* gen. nov., sp. nov., a pink-filament-forming hyperthermophilic bacterium isolated from Yellowstone National Park. *Appl Environ Microbiol.* 1998;64:3576–83. doi:10.1128/aem.64.10.3576-3583.1998.
186. Ahn Y-B, Kerkhof LJ, Hä Ggblom MM, Hä MM. *Desulfoluna spongiiphila* sp. nov., a dehalogenating bacterium in the Desulfobacteraceae from the marine sponge *Aplysina aerophoba*. doi:10.1099/ijs.0.005884-0.
187. Jackson TJ, Ramaley RF, Meinschein WG. *Thermomicrobium*, a new genus of extremely thermophilic bacteria. *Int J Syst Bacteriol.* 1973;23:28–36. doi:10.1099/00207713-23-1-28.
188. Scanlan DJ, West NJ. Molecular ecology of the marine cyanobacterial genera *Prochlorococcus* and *Synechococcus*. *FEMS Microbiol Ecol.* 2006;40:1–12. doi:10.1111/j.1574-6941.2002.tb00930.x.
189. Cava F, Hidalgo A, Berenguer J. *Thermus thermophilus* as biological model. *Extremophiles.* 2009;13:213–31. doi:10.1007/s00792-009-0226-6.
190. Green Michael R., Sambrook Joseph. *Molecular Cloning: A Laboratory Manual.* 4th edition. Cold Spring Harbor Laboratory Press; 2012. www.cshlpress.org.

191. Chuzel L, Ganatra MB, Rapp E, Henrissat B, Taron CH. Functional metagenomics identifies an exosialidase with an inverting catalytic mechanism that defines a new glycoside hydrolase family (GH156). *J Biol Chem.* 2018;293:18138–50. doi:10.1074/jbc.RA118.003302.
192. Chen HM, Armstrong Z, Hallam SJ, Withers SG. Synthesis and evaluation of a series of 6-chloro-4-methylumbelliferyl glycosides as fluorogenic reagents for screening metagenomic libraries for glycosidase activity. *Carbohydr Res.* 2016;421:33–9. doi:10.1016/j.carres.2015.12.010.
193. Profeta GS, Pereira JAS, Costa SG, Azambuja P, Garcia ES, Moraes C da S, et al. Standardization of a continuous assay for glycosidases and its use for screening insect gut samples at individual and populational levels. *Front Physiol.* 2017;8 MAY:308. doi:10.3389/fphys.2017.00308.
194. Malo N, Hanley JA, Cerquozzi S, Pelletier J, Nadon R. Statistical practice in high-throughput screening data analysis. *Nat Biotechnol.* 2006;24:167–75. doi:10.1038/nbt1186.
195. Nederbragt A J. Longing for the longest reads: PacBio and BluePippin | In between lines of code. 2013. <https://flxlexblog.wordpress.com/2013/06/19/longing-for-the-longest-reads-pacbio-and-bluepippin/>. Accessed 18 May 2021.
196. García-Angulo VA. Overlapping riboflavin supply pathways in bacteria. *Crit Rev Microbiol.* 2017;43:196–209. doi:10.1080/1040841X.2016.1192578.
197. Yang H, Xiao X, Zhao XS, Hu L, Xue XF, Ye JS. Study on Fluorescence Spectra of Thiamine and Riboflavin. *MATEC Web Conf.* 2016;63:03013. doi:10.1051/mateconf/20166303013.
198. Lee J. Lumazine protein and the excitation mechanism in bacterial bioluminescence. *Biophys Chem.* 1993;48:149–58. doi:10.1016/0301-4622(93)85006-4.
199. Varki A, Schnaar RL, Schauer R. Sialic Acids and Other Nonulosonic Acids. In: *Essentials of Glycobiology*. 3rd edition. Cold Spring Harbor Laboratory Press; 2015. doi:10.1101/glycobiology.3e.015.
200. Chan J, Sandhu G, Bennet AJ. A mechanistic study of sialic acid mutarotation: Implications for mutarotase enzymes. *Org Biomol Chem.* 2011;9:4818. doi:10.1039/c1ob05079f.
201. Angata T, Varki A. Chemical diversity in the sialic acids and related α -keto acids: An evolutionary perspective. *Chem Rev.* 2002;102:439–69. doi:10.1021/cr000407m.
202. Rosen SD, Bertozzi CR. The selectins and their ligands. *Curr Opin Cell Biol.* 1994;6:663–73. doi:10.1016/0955-0674(94)90092-2.
203. Crocker Paul R, Paulson James C, Varki Ajit. Siglecs and their roles in the immune system. *Nat Rev Immunol.* 2007;7:255–66. doi:10.1038/nri2056.
204. Stencel-Baerenwald JE, Reiss K, Reiter DM, Stehle T, Dermody TS. The sweet spot: Defining virus-sialic acid interactions. *Nat Rev Microbiol.* 2014;12:739–49. doi:10.1038/nrmicro3346.
205. Adams JH, Sim BK, Dolan SA, Fang X, Kaslow DC, Miller LH. A family of erythrocyte binding proteins of malaria parasites. *Proc Natl Acad Sci.* 1992;89:7085–9. doi:10.1073/pnas.89.15.7085.

206. Miller-Podraza H, Bergström J, Teneberg S, Milh MA, Longard M, Olsson B-M, et al. *Helicobacter pylori* and Neutrophils: Sialic Acid-Dependent Binding to Various Isolated Glycoconjugates. *Infect Immun*. 1999;67:6309–13. doi:10.1128/IAI.67.12.6309-6313.1999.
207. Rodrigues E, Macauley MS. Hypersialylation in Cancer: Modulation of Inflammation and Therapeutic Opportunities. *Cancers (Basel)*. 2018;10. doi:10.3390/cancers10060207.
208. Peri S, Kulkarni A, Feyertag F, Berninsone PM, Alvarez-Ponce D. Phylogenetic Distribution of CMP-Neu5Ac Hydroxylase (CMAH), the Enzyme Synthetizing the Proinflammatory Human Xenoantigen Neu5Gc. *Genome Biol Evol*. 2018;10:207–19. doi:10.1093/gbe/evx251.
209. Tangvoranuntakul P, Gagneux P, Diaz S, Bardor M, Varki N, Varki A, et al. Human uptake and incorporation of an immunogenic nonhuman dietary sialic acid. *Proc Natl Acad Sci*. 2003;100:12045–50. doi:10.1073/pnas.2131556100.
210. Banda K, Gregg CJ, Chow R, Varki NM, Varki A. Metabolism of Vertebrate Amino Sugars with N-Glycolyl Groups. *J Biol Chem*. 2012;287:28852–64. doi:10.1074/jbc.M112.364182.
211. Anjum C, Chia YC, Kour AK, Adalsteinsson O, Papacharalampous M, Zocchi ML, et al. Understanding the presence of xeno-derived Neu5Gc in the human body, and its significance: a review. *J Stem Cell Res Ther*. 2020;6:72–7. doi:10.15406/jsrt.2020.06.00144.
212. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res*. 2014;42:D490–5. doi:10.1093/nar/gkt1178.
213. Petter JG, Vimr ER. Complete nucleotide sequence of the bacteriophage K1F tail gene encoding endo-N-acetylneuraminidase (endo-N) and comparison to an endo-N homolog in bacteriophage PK1E. *J Bacteriol*. 1993;175:4354–63. doi:10.1128/jb.175.14.4354-4363.1993.
214. Varghese JN, Laver WG, Colman PM. Structure of the influenza virus glycoprotein antigen neuraminidase at 2.9 Å resolution. *Nature*. 1983;303:35–40. doi:10.1038/303035a0.
215. Burmeister WP, Henrissat B, Bosso C, Cusack S, Ruigrok RWH. Influenza B virus neuraminidase can synthesize its own inhibitor. *Structure*. 1993;1:19–26. doi:10.1016/0969-2126(93)90005-2.
216. Crennell SJ, Garman EF, Laver WG, Vimr ER, Taylor GL. Crystal structure of a bacterial sialidase (from *Salmonella typhimurium* LT2) shows the same fold as an influenza virus neuraminidase. *Proc Natl Acad Sci*. 1993;90:9852–6. doi:10.1073/pnas.90.21.9852.
217. Crennell S, Garman E, Laver G, Vimr E, Taylor G. Crystal structure of *Vibrio cholerae* neuraminidase reveals dual lectin-like domains in addition to the catalytic domain. *Structure*. 1994;2:535–44. doi:10.1016/S0969-2126(00)00053-8.
218. Crennell S, Takimoto T, Portner A, Taylor G. Crystal structure of the multifunctional paramyxovirus hemagglutinin-neuraminidase. *Nat Struct Biol*. 2000;7:1068–74. doi:10.1038/81002.
219. Luo Y, Li S-C, Chou M-Y, Li Y-T, Luo M. The crystal structure of an intramolecular trans-sialidase with a NeuAc α 2 \rightarrow 3Gal specificity. *Structure*. 1998;6:521–30.

doi:10.1016/S0969-2126(98)00053-7.

220. Buschiazzo A, Amaya MF, Cremona ML, Frasch AC, Alzari PM. The crystal structure and mode of action of trans-sialidase, a key enzyme in *Trypanosoma cruzi* pathogenesis. *Mol Cell*. 2002;10:757–68. doi:10.1016/S1097-2765(02)00680-9.

221. Buschiazzo A. Structural basis of sialyltransferase activity in trypanosomal sialidases. *EMBO J*. 2000;19:16–24. doi:10.1093/emboj/19.1.16.

222. Wilson JC, Angus DI, von Itzstein M. ¹H NMR Evidence That *Salmonella typhimurium* Sialidase Hydrolyzes Sialosides with Overall Retention of Configuration. *J Am Chem Soc*. 1995;117:4214–7. doi:10.1021/ja00120a002.

223. Chong AKJ, Pegg MS, Taylor NR, Itzstein M. Evidence for a sialosyl cation transition-state complex in the reaction of sialidase from influenza virus. *Eur J Biochem*. 1992;207:335–43. doi:10.1111/j.1432-1033.1992.tb17055.x.

224. Friebolin H, Baumann W, Keilich G, Ziegler D, Brossmer R, von Nicolai H. [¹H-NMR spectroscopy--a potent method for the determination of substrate specificity of sialidases (author's transl)]. *Hoppe Seylers Z Physiol Chem*. 1981;362:1455–63. <http://www.ncbi.nlm.nih.gov/pubmed/6273284>. Accessed 28 Mar 2018.

225. Kao YH, Lerner L, Warner TG. Stereoselectivity of the Chinese hamster ovary cell sialidase: Sialoside hydrolysis with overall retention of configuration. *Glycobiology*. 1997;7:559–63. doi:10.1093/glycob/7.4.559.

226. Morley TJ, Willis LM, Whitfield C, Wakarchuk WW, Withers SG. A New Sialidase Mechanism. *J Biol Chem*. 2009;284:17404–10. doi:10.1074/jbc.M109.003970.

227. von Itzstein M. The war against influenza: discovery and development of sialidase inhibitors. *Nat Rev Drug Discov*. 2007;6:967–74. doi:10.1038/nrd2400.

228. McNicholl IR, McNicholl JJ. Neuraminidase Inhibitors: Zanamivir and Oseltamivir. *Ann Pharmacother*. 2001;35:57–70. doi:10.1345/aph.10118.

229. Xiao H, Woods EC, Vukojicic P, Bertozzi CR. Precision glycocalyx editing as a strategy for cancer immunotherapy. *Proc Natl Acad Sci*. 2016;113:10304–9. doi:10.1073/pnas.1608069113.

230. Fürste JP, Pansegrau W, Frank R, Blöcker H, Scholz P, Bagdasarian M, et al. Molecular cloning of the plasmid RP4 primase region in a multi-host-range tacP expression vector. *Gene*. 1986;48:119–31. doi:10.1016/0378-1119(86)90358-6.

231. Sekiguchi Y, Yamada T, Hanada S, Ohashi A, Harada H, Kamagata Y. *Anaerolinea thermophila* gen. nov., sp. nov. and *Caldilinea aerophila* gen. nov., sp. nov., novel filamentous thermophiles that represent a previously uncultured lineage of the domain bacteria at the subphylum level. *Int J Syst Evol Microbiol*. 2003;53:1843–51. doi:10.1099/ijs.0.02699-0.

232. Wu D, Raymond J, Wu M, Chatterji S, Ren Q, Graham JE, et al. Complete genome sequence of the aerobic CO-oxidizing thermophile *Thermomicrobium roseum*. *PLoS One*. 2009;4:e4207. doi:10.1371/journal.pone.0004207.

233. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods*.

2013;10:563–9. doi:10.1038/nmeth.2474.

234. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* 2010;38:e132–e132. doi:10.1093/nar/gkq275.

235. Neelamegham S, Aoki-Kinoshita K, Bolton E, Frank M, Lisacek F, Lütteke T, et al. Updates to the Symbol Nomenclature for Glycans guidelines. *Glycobiology.* 2019;29:620–4. doi:10.1093/glycob/cwz045.

236. Friebolin H, Brossmer R, Keilich G, Ziegler D, Supp M. [1H-NMR-spectroscopic evidence for the release of N-acetyl-alpha-D-neuraminic acid as the first product of neuraminidase action (author's transl)]. *Hoppe Seylers Z Physiol Chem.* 1980;361:697–702. <http://www.ncbi.nlm.nih.gov/pubmed/6253376>. Accessed 1 Aug 2018.

237. Martens EC, Koropatkin NM, Smith TJ, Gordon JI. Complex glycan catabolism by the human gut microbiota: the Bacteroidetes Sus-like paradigm. *J Biol Chem.* 2009;284:24673–7. doi:10.1074/jbc.R109.022848.

238. Foley MH, Cockburn DW, Koropatkin NM. The Sus operon: a model system for starch uptake by the human gut Bacteroidetes. *Cell Mol Life Sci.* 2016;73:2603–17. doi:10.1007/s00018-016-2242-x.

239. Wang X, Long H, Shen D, Liu L. Cloning, expression, and characterization of a novel sialidase from *Brevibacterium casei*. *Biotechnol Appl Biochem.* 2017;64:195–200. doi:10.1002/bab.1475.

240. Guo J, Wang Y, Song B, Wang X, Yang G, Guan F. Identification and functional characterization of intracellular sialidase NeuA3 from *Streptomyces avermitilis*. *Process Biochem.* 2015;50:752–8. doi:10.1016/j.procbio.2015.02.005.

241. Park KH, Kim MG, Ahn HJ, Lee DH, Kim JH, Kim YW, et al. Structural and biochemical characterization of the broad substrate specificity of *Bacteroides thetaiotaomicron* commensal sialidase. *Biochim Biophys Acta - Proteins Proteomics.* 2013;1834:1510–9. doi:10.1016/j.bbapap.2013.04.028.

242. Minami A, Ishibashi S, Ikeda K, Ishitsubo E, Hori T, Tokiwa H, et al. Catalytic preference of *Salmonella typhimurium* LT2 sialidase for N-acetylneuraminic acid residues over N-glycolylneuraminic acid residues. *FEBS Open Bio.* 2013;3:231–6. doi:10.1016/j.fob.2013.05.002.

243. Useh NM, Ajanusi JO, Esievo KAN, Nok AJ. Characterization of a sialidase (neuraminidase) isolated from *Clostridium chauvoei* (Jakari strain). *Cell Biochem Funct.* 2006;24:347–52. doi:10.1002/cbf.1240.

244. Jers C, Guo Y, Kepp KP, Mikkelsen JD. Mutants of *Micromonospora viridifaciens* sialidase have highly variable activities on natural and non-natural substrates. *Protein Eng Des Sel.* 2015;28:37–44. doi:10.1093/protein/gzu054.

245. Watson JN, Dookhun V, Borgford TJ, Bennet AJ. Mutagenesis of the Conserved Active-Site Tyrosine Changes a Retaining Sialidase into an Inverting Sialidase. *Biochemistry.* 2003;42:12682–90. doi:10.1021/bi035396g.

246. M9 Salts. *Cold Spring Harb Protoc.* 2006;2006:pdb.rec614. doi:10.1101/pdb.rec614.

247. Clabbers MTB, Gruene T, Parkhurst JM, Abrahams JP, Waterman DG. Electron diffraction data processing with *DIALS*. *Acta Crystallogr Sect D Struct Biol*. 2018;74:506–18. doi:10.1107/S2059798318007726.
248. Evans PR, Murshudov GN. How good are my data and what is the resolution? *Acta Crystallogr Sect D Biol Crystallogr*. 2013;69:1204–14. doi:10.1107/S0907444913000061.
249. Pannu NS, Waterreus W-J, Skubák P, Sikharulidze I, Abrahams JP, de Graaff RAG. Recent advances in the *CRANK* software suite for experimental phasing. *Acta Crystallogr Sect D Biol Crystallogr*. 2011;67:331–7. doi:10.1107/S0907444910052224.
250. Emsley P, Cowtan K. *Coot*: model-building tools for molecular graphics. *Acta Crystallogr Sect D Biol Crystallogr*. 2004;60:2126–32. doi:10.1107/S0907444904019158.
251. Murshudov GN, Skubák P, Lebedev AA, Pannu NS, Steiner RA, Nicholls RA, et al. *REFMAC 5* for the refinement of macromolecular crystal structures. *Acta Crystallogr Sect D Biol Crystallogr*. 2011;67:355–67. doi:10.1107/S0907444911001314.
252. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ. *Phaser* crystallographic software. *J Appl Crystallogr*. 2007;40:658–74. doi:10.1107/S0021889807021206.
253. Lebedev AA, Young P, Isupov MN, Moroz O V., Vagin AA, Murshudov GN. *JLigand*: a graphical tool for the *CCP 4* template-restraint library. *Acta Crystallogr Sect D Biol Crystallogr*. 2012;68:431–40. doi:10.1107/S090744491200251X.
254. Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. *Nat Struct Mol Biol*. 2003;10:980–980. doi:10.1038/nsb1203-980.
255. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera?A visualization system for exploratory research and analysis. *J Comput Chem*. 2004;25:1605–12. doi:10.1002/jcc.20084.
256. Puthenveetil R, Vinogradova O. Solution NMR: A powerful tool for structural and functional studies of membrane proteins in reconstituted environments. *Journal of Biological Chemistry*. 2019;294:15914–31. doi:10.1074/jbc.REV119.009178.
257. Callaway E. Revolutionary cryo-EM is taking over structural biology. *Nature*. 2020;578:201–201. doi:10.1038/d41586-020-00341-9.
258. Smyth MS. x Ray crystallography. *Mol Pathol*. 2000;53:8–14. doi:10.1136/mp.53.1.8.
259. Ilari A, Savino C. Protein structure determination by X-ray crystallography. In: *Methods in Molecular Biology*. Humana Press; 2008. p. 63–87. doi:10.1007/978-1-60327-159-2_3.
260. Cowtan K. Phase Problem in X-ray Crystallography, and Its Solution. In: eLS. Wiley; 2003. doi:10.1038/npg.els.0002722.
261. Scapin G. Molecular replacement then and now. In: *Acta Crystallographica Section D: Biological Crystallography*. International Union of Crystallography; 2013. p. 2266–75. doi:10.1107/S0907444913011426.
262. Walden H. Selenium incorporation using recombinant techniques. *Acta Crystallogr*

Sect D Biol Crystallogr. 2010;66:352–7. doi:10.1107/S0907444909038207.

263. Barton WA, Tzvetkova-Robev D, Erdjument-Bromage H, Tempst P, Nikolov DB. Highly efficient selenomethionine labeling of recombinant proteins produced in mammalian cells. *Protein Sci.* 2006;15:2008–13. doi:10.1110/ps.062244206.

264. Hendrickson WA, Ogata CM. [28] Phase determination from multiwavelength anomalous diffraction measurements. In: *Methods in Enzymology*. Academic Press Inc.; 1997. p. 494–523. doi:10.1016/S0076-6879(97)76074-9.

265. Amartely H, Some D, Tsadok A, Lebendiker M. Characterization of Proteins by Size-Exclusion Chromatography Coupled to Multi-Angle Light Scattering (SEC-MALS). *Artic J Vis Exp.* 2019;:59615. doi:10.3791/59615.

266. Tarazona MP, Saiz E. Combination of SEC/MALS experimental procedures and theoretical analysis for studying the solution properties of macromolecules. *J Biochem Biophys Methods.* 2003;56:95–116. doi:10.1016/S0165-022X(03)00075-7.

267. Bule P, Chuzel L, Blagova E, Wu L, Gray MA, Henriessat B, et al. Inverting family GH156 sialidases define an unusual catalytic motif for glycosidase action. *Nat Commun.* 2019;10:1–11. doi:10.1038/s41467-019-12684-7.

268. Ochoa-Leyva A, Barona-Gómez F, Saab-Rincón G, Verdel-Aranda K, Sánchez F, Soberón X. Exploring the Structure–Function Loop Adaptability of a (β/α)8-Barrel Enzyme through Loop Swapping and Hinge Variability. *J Mol Biol.* 2011;411:143–57. doi:10.1016/j.jmb.2011.05.027.

269. Laskowski RA, Jabłońska J, Pravda L, Svobodová Va rekov R, Thornton JM. PDBsum: Structural summaries of PDB entries. *Protein Sci.* 2017;27:129–34. doi:10.1002/pro.3289.

270. Amaya MF, Watts AG, Damager I, Wehenkel A, Nguyen T, Buschiazio A, et al. Structural Insights into the Catalytic Mechanism of *Trypanosoma cruzi* trans-Sialidase. *Structure.* 2004;12:775–84. doi:10.1016/j.str.2004.02.036.

271. Schulz EC, Schwarzer D, Frank M, Stummeyer K, Mühlhoff M, Dickmanns A, et al. Structural Basis for the Recognition and Cleavage of Polysialic Acid by the Bacteriophage K1F Tailspike Protein EndoNF. *J Mol Biol.* 2010;397:341–51. doi:10.1016/j.jmb.2010.01.028.

272. Holm L, Sander C. Dali: a network tool for protein structure comparison. *Trends Biochem Sci.* 1995;20:478–80. doi:10.1016/S0968-0004(00)89105-7.

273. Bradshaw WJ, Kirby JM, Roberts AK, Shone CC, Acharya KR. The molecular structure of the glycoside hydrolase domain of Cwp19 from *Clostridium difficile*. *FEBS J.* 2017;284:4343–57. doi:10.1111/febs.14310.

274. Wydau-Demattis S, El Meouche I, Courtin P, Hamiot A, Lai-Kuen R, Saubaméa B, et al. Cwp19 is a novel lytic transglycosylase involved in stationary-phase autolysis resulting in toxin release in *clostridium difficile*. *MBio.* 2018;9. doi:10.1128/mBio.00648-18.

275. Hehemann J-H, Kelly AG, Pudlo NA, Martens EC, Boraston AB. Bacteria of the human gut microbiome catabolize red seaweed glycans with carbohydrate-active enzyme updates from extrinsic microbes. *Proc Natl Acad Sci.* 2012;109:19786–91. doi:10.1073/pnas.1211002109.

276. Linder M, Lindeberg G, Reinikainen T, Teeri TT, Pettersson G. The difference in

affinity between two fungal cellulose-binding domains is dominated by a single amino acid substitution. *FEBS Lett.* 1995;372:96–8. doi:10.1016/0014-5793(95)00961-8.

277. Kim S, Oh D-B, Kang HA, Kwon O. Features and applications of bacterial sialidases. *Appl Microbiol Biotechnol.* 2011;91:1–15. doi:10.1007/s00253-011-3307-2.

278. Davies G, Henrissat B. Structures and mechanisms of glycosyl hydrolases. *Structure.* 1995;3:853–9. doi:10.1016/S0969-2126(01)00220-9.

279. McCarter JD, Stephen Withers G. Mechanisms of enzymatic glycoside hydrolysis. *Curr Opin Struct Biol.* 1994;4:885–92. doi:10.1016/0959-440X(94)90271-2.

280. Rouvinen J, Bergfors T, Teeri T, Knowles JK, Jones TA. Three-Dimensional Structure of Cellobiohydrolase II from *Trichoderma reesei*. *Science* (80-). 1990;249:380–6. doi:10.1126/science.2377893.

281. Juy M, Amrt AG, Alzari PM, Poljak RJ, Claeysens M, Béguin P, et al. Three-dimensional structure of a thermostable bacterial cellulase. *Nature.* 1992;357:89–91. doi:10.1038/357089a0.

282. Hehemann J-H, Smyth L, Yadav A, Vocadlo DJ, Boraston AB. Analysis of Keystone Enzyme in Agar Hydrolysis Provides Insight into the Degradation (of a Polysaccharide from) Red Seaweeds. *J Biol Chem.* 2012;287:13985–95. doi:10.1074/jbc.M112.345645.

283. Pluvinage B, Grondin JM, Amundsen C, Klassen L, Moote PE, Xiao Y, et al. Molecular basis of an agarose metabolic pathway acquired by a human intestinal symbiont. *Nat Commun.* 2018;9:1043. doi:10.1038/s41467-018-03366-x.

284. Newstead S, Watson JN, Knoll TL, Bennet AJ, Taylor G. Structure and Mechanism of Action of an Inverting Mutant Sialidase. *Biochemistry.* 2005;44:9117–22. doi:10.1021/bi050517t.

285. Lupas AN, Pereira J, Alva V, Merino F, Coles M, Hartmann MD. The breakthrough in protein structure prediction. *Biochem J.* 2021;478:1885–90. doi:10.1042/BCJ20200963.

286. Xiao H, Woods EC, Vukojicic P, Bertozzi CR. Precision glycocalyx editing as a strategy for cancer immunotherapy. *Proc Natl Acad Sci.* 2016;113:10304–9. doi:10.1073/pnas.1608069113.

287. Gray MA, Stanczak MA, Mantuano NR, Xiao H, Pijnenborg JFA, Malaker SA, et al. Targeted glycan degradation potentiates the anticancer immune response in vivo. *Nat Chem Biol.* 2020;16:1376–84. doi:10.1038/s41589-020-0622-x.

288. Tector AJ, Mosser M, Tector M, Bach JM. The Possible Role of Anti-Neu5Gc as an Obstacle in Xenotransplantation. *Frontiers in Immunology.* 2020;11:622. doi:10.3389/fimmu.2020.00622.

289. Padler-Karavani V, Varki A. Potential impact of the non-human sialic acid N-glycolylneuraminic acid on transplant rejection risk. *Xenotransplantation.* 2011;18:1–5. doi:10.1111/j.1399-3089.2011.00622.x.

290. Parker RB, McCombs JE, Kohler JJ. Sialidase specificity determined by chemoselective modification of complex sialylated glycans. *ACS Chem Biol.* 2012;7:1509–14. doi:10.1021/cb300241v.

291. McCombs JE, Diaz JP, Luebke KJ, Kohler JJ. Glycan specificity of neuraminidases determined in microarray format. *Carbohydr Res.* 2016;428:31–40. doi:10.1016/j.carres.2016.04.003.
292. Kim S, Oh D-B, Kang HA, Kwon O. Features and applications of bacterial sialidases. *Appl Microbiol Biotechnol.* 2011;91:1–15. doi:10.1007/s00253-011-3307-2.
293. Juge N, Tailford L, Owen CD. Sialidases from gut bacteria: a mini-review. *Biochem Soc Trans.* 2016;44:166–75. doi:10.1042/BST20150226.
294. Chokhwalala HA, Yu H, Chen X. High-Throughput Substrate Specificity Studies of Sialidases by Using Chemoenzymatically Synthesized Sialoside Libraries. *ChemBioChem.* 2007;8:194–201. doi:10.1002/cbic.200600410.
295. Samraj AN, Läubli H, Varki N, Varki A. Involvement of a Non-Human Sialic Acid in Human Cancer. *Front Oncol.* 2014;4. doi:10.3389/fonc.2014.00033.
296. Samraj AN, Pearce OMT, Läubli H, Crittenden AN, Bergfeld AK, Banda K, et al. A red meat-derived glycan promotes inflammation and cancer progression. *Proc Natl Acad Sci.* 2015;112:542–7. doi:10.1073/pnas.1417508112.
297. Yu H, Chen X. Carbohydrate post-glycosylational modifications. *Org Biomol Chem.* 2007;5:865–72. doi:10.1039/b700034k.
298. Paschinger K, Wilson IBH. Anionic and zwitterionic moieties as widespread glycan modifications in non-vertebrates. *Glycoconj J.* 2020;37:27–40. doi:10.1007/s10719-019-09874-2.
299. Casu B, Naggi A, Torri G. Re-visiting the structure of heparin. *Carbohydr Res.* 2015;403:60–8. doi:10.1016/j.carres.2014.06.023.
300. Young NM, Foote SJ, Wakarchuk WW. Review of phosphocholine substituents on bacterial pathogen glycans: Synthesis, structures and interactions with host proteins. *Molecular Immunology.* 2013;56:563–73. doi:10.1016/j.molimm.2013.05.237.
301. Scott NE, Nothaft H, Edwards AVG, Labbate M, Djordjevic SP, Larsen MR, et al. Modification of the *Campylobacter jejuni* N-Linked Glycan by EptC Protein-mediated Addition of Phosphoethanolamine. *J Biol Chem.* 2012;287:29384–96. doi:10.1074/jbc.M112.380212.
302. Gandhi NS, Mancera RL. The structure of glycosaminoglycans and their interactions with proteins. *Chemical Biology and Drug Design.* 2008;72:455–82. doi:10.1111/j.1747-0285.2008.00741.x.
303. Staudacher E. Methylation – an uncommon modification of glycans. *Biol Chem.* 2012;393:675–85. doi:10.1515/hsz-2012-0132.
304. Klein A, Roussel P. O-Acetylation of sialic acids. *Biochimie.* 1998;80:49–57. doi:10.1016/S0300-9084(98)80056-4.
305. Zhu Q, Venzke D, Walimbe AS, Anderson ME, Fu Q, Kinch LN, et al. Structure of protein O-mannose kinase reveals a unique active site architecture. *Elife.* 2016;5. doi:10.7554/eLife.22238.
306. Wen J, Xiao J, Rahdar M, Choudhury BP, Cui J, Taylor GS, et al. Xylose

phosphorylation functions as a molecular switch to regulate proteoglycan biosynthesis. *Proc Natl Acad Sci U S A*. 2014;111:15723–8. doi:10.1073/pnas.1417993111.

307. Hager FF, Sützl L, Stefanović C, Blaukopf M, Schäffer C. Pyruvate substitutions on glycoconjugates. *International Journal of Molecular Sciences*. 2019;20. doi:10.3390/ijms20194929.

308. Houston KM, Harnett W. Structure and synthesis of nematode phosphorylcholine-containing glycoconjugates. *Parasitology*. 2004;129:655–61. doi:10.1017/S0031182004006171.

309. Maes E, Garénaux E, Strecker G, Leroy Y, Wieruszkeski JM, Brassart C, et al. Major O-glycans from the nest of *Vespula germanica* contain phospho-ethanolamine. *Carbohydr Res*. 2005;340:1852–8. doi:10.1016/j.carres.2005.05.008.

310. Hykollari A, Eckmair B, Voglmeir J, Jin C, Yan S, Vanbeselaere J, et al. More Than Just Oligomannose: An N-glycomic Comparison of *Penicillium* Species. *Mol Cell Proteomics*. 2016;15:73–92. doi:10.1074/mcp.M115.055061.

311. Hykollari A, Malzl D, Eckmair B, Vanbeselaere J, Scheidl P, Jin C, et al. Isomeric Separation and Recognition of Anionic and Zwitterionic N-glycans from Royal Jelly Glycoproteins. *Mol Cell Proteomics*. 2018;17:2177–96. doi:10.1074/mcp.RA117.000462.

312. Paulick MG, Bertozzi CR. The glycosylphosphatidylinositol anchor: A complex membrane-anchoring structure for proteins. *Biochemistry*. 2008;47:6991–7000. doi:10.1021/bi8006324.

313. Hollenbeck EC, Antonoplis A, Chai C, Thongsomboon W, Fuller GG, Cegelski L. Phosphoethanolamine cellulose enhances curli-mediated adhesion of uropathogenic *Escherichia coli* to bladder epithelial cells. *Proc Natl Acad Sci U S A*. 2018;115:10106–11. doi:10.1073/pnas.1801564115.

314. Hulswit RJG, Lang Y, Bakkers MJG, Li W, Li Z, Schouten A, et al. Human coronaviruses OC43 and HKU1 bind to 9-O-acetylated sialic acids via a conserved receptor-binding site in spike protein domain A. *Proc Natl Acad Sci U S A*. 2019;116:2681–90. doi:10.1073/pnas.1809667116.

315. Huang X, Dong W, Milewska A, Golda A, Qi Y, Zhu QK, et al. Human Coronavirus HKU1 Spike Protein Uses O -Acetylated Sialic Acid as an Attachment Receptor Determinant and Employs Hemagglutinin-Esterase Protein as a Receptor-Destroying Enzyme . *J Virol*. 2015;89:7202–13. doi:10.1128/jvi.00854-15.

316. Mandal C, Mandal C, Chandra S, Schauer R, Mandal C. Regulation of O-acetylation of sialic acids by sialate-O-acetyltransferase and sialate-O-acylesterase activities in childhood acute lymphoblastic leukemia. *Glycobiology*. 2012;22:70–83. doi:10.1093/glycob/cwr106.

317. Sinha D, Mandal C, Bhattacharya D. Identification of 9-O acetyl sialoglycoconjugates (9-OAcSGs) as biomarkers in childhood acute lymphoblastic leukemia using a lectin, AchatininH, as a probe. *Leukemia*. 1999;13:119–25. doi:10.1038/sj.leu.2401239.

318. Wang J-R, Gao W-N, Grimm R, Jiang S, Liang Y, Ye H, et al. A method to identify trace sulfated IgG N-glycans as biomarkers for rheumatoid arthritis. *Nat Commun*. 2017;8:631. doi:10.1038/s41467-017-00662-w.

319. Zulueta MML, Hung S-C. Synthesis of Sulfated Glycans. In: *Glycoscience: Biology and Medicine*. Tokyo: Springer Japan; 2014. p. 1–7. doi:10.1007/978-4-431-54836-2_107-1.
320. Dewald J, Colomb F, Bobowski-Gerard M, Groux-Degroote S, Delannoy P. Role of Cytokine-Induced Glycosylation Changes in Regulating Cell Interactions and Cell Signaling in Inflammatory Diseases and Cancer. *Cells*. 2016;5:43. doi:10.3390/cells5040043.
321. Ivetic A, Green HLH, Hart SJ. L-selectin: A major regulator of leukocyte adhesion, migration and signaling. *Frontiers in Immunology*. 2019;10. doi:10.3389/fimmu.2019.01068.
322. Fukuda M, Hiraoka N, Yeh JC. C-type lectins and Sialyl Lewis X oligosaccharides: Versatile roles in cell-cell interaction. *Journal of Cell Biology*. 1999;147:467–70. doi:10.1083/jcb.147.3.467.
323. van Kuik JA, Breg J, Kolsteeg CEM, Kamerling JP, Vliegenthart JFG. Primary structure of the acidic carbohydrate chain of hemocyanin from *Panulirus interruptus*. *FEBS Lett*. 1987;221:150–4. doi:10.1016/0014-5793(87)80370-8.
324. Freeze HH. Mannose 6-sulfate is present in the N-linked oligosaccharides of lysosomal enzymes of *Dictyostelium*. *Arch Biochem Biophys*. 1985;243:690–3. doi:10.1016/0003-9861(85)90547-8.
325. She Y-M, Farnsworth A, Li X, Cyr TD. Topological N-glycosylation and site-specific N-glycan sulfation of influenza proteins in the highly expressed H1N1 candidate vaccines. *Sci Rep*. 2017;7:10232. doi:10.1038/s41598-017-10714-2.
326. Chen J-Y, Huang H-H, Yu S-Y, Wu S-J, Kannagi R, Khoo K-H. Concerted mass spectrometry-based glycomic approach for precision mapping of sulfo sialylated N-glycans on human peripheral blood mononuclear cells and lymphocytes. *Glycobiology*. 2018;28:9–20. doi:10.1093/glycob/cwx091.
327. Jiang H, Irungu J, Desaire H. Enhanced detection of sulfated glycosylation sites in glycoproteins. *J Am Soc Mass Spectrom*. 2005;16:340–8. doi:10.1016/j.jasms.2004.11.015.
328. Abbott KL, Pierce JM. Lectin-Based Glycoproteomic Techniques for the Enrichment and Identification of Potential Biomarkers. In: *Methods in Enzymology*. Academic Press Inc.; 2010. p. 461–76. doi:10.1016/S0076-6879(10)80020-5.
329. Vainauskas S, Duke RM, McFarland J, McClung C, Ruse C, Taron CH. Profiling of core fucosylated N-glycans using a novel bacterial lectin that specifically recognizes α 1,6 fucosylated chitobiose. *Sci Rep*. 2016;6 July:1–12. doi:10.1038/srep34195.
330. Zhu R, Zacharias L, Wooding KM, Peng W, Mechref Y. Glycoprotein Enrichment Analytical Techniques. In: *Methods in Enzymology*. 2017. p. 397–429. doi:10.1016/bs.mie.2016.11.009.
331. Chen M, Shi X, Duke RM, Ruse CI, Dai N, Taron CH, et al. An engineered high affinity Fbs1 carbohydrate binding protein for selective capture of N-glycans and N-glycopeptides. *Nat Commun*. 2017;8:1–15. doi:10.1038/ncomms15487.
332. Ganatra MB, Potapov V, Vainauskas S, Francis AZ, McClung CM, Ruse CI, et al. A bi-specific lectin from the mushroom *Boletopsis grisea* and its application in glycoanalytical workflows. *Sci Rep*. 2021;11:160. doi:10.1038/s41598-020-80488-7.

333. Ruhaak LR, Hennig R, Huhn C, Borowiak M, Dolhain RJEM, Deelder AM, et al. Optimized Workflow for Preparation of APTS-Labeled N-Glycans Allowing High-Throughput Analysis of Human Plasma Glycomes using 48-Channel Multiplexed CGE-LIF. *J Proteome Res.* 2010;9:6655–64. doi:10.1021/pr100802f.
334. Schwarzer J, Rapp E, Reichl U. N-glycan analysis by CGE-LIF: Profiling influenza A virus hemagglutinin N-glycosylation during vaccine production. *Electrophoresis.* 2008;29:4203–14. doi:10.1002/elps.200800042.
335. Lam KN, Hall MW, Engel K, Vey G, Cheng J, Neufeld JD, et al. Evaluation of a Pooled Strategy for High-Throughput Sequencing of Cosmid Clones from Metagenomic Libraries. *PLoS One.* 2014;9:e98968. doi:10.1371/journal.pone.0098968.
336. Petushkov VN, Van Stokkum IHM, Gobets B, Van Mourik F, Lee J, Van Grondelle R, et al. Ultrafast fluorescence relaxation spectroscopy of 6,7-dimethyl-(8-ribityl)-lumazine and riboflavin, free and bound to antenna proteins from bioluminescent bacteria. *J Phys Chem B.* 2003;107:10934–9. doi:10.1021/jp034266e.
337. Darzentas N. Circoletto: Visualizing sequence similarity with Circos. *Bioinformatics.* 2010;26:2620–1. doi:10.1093/bioinformatics/btq484.
338. Arnold JN, Wormald MR, Sim RB, Rudd PM, Dwek RA. The Impact of Glycosylation on the Biological Function and Structure of Human Immunoglobulins. *Annu Rev Immunol.* 2007;25:21–50. doi:10.1146/annurev.immunol.25.022106.141702.
339. Bergweff AA, Thomas-Oates JE, van Oostrum J, Kamerling JP, Vliegthart JFG. Human urokinase contains GalNAc β (1-4)[Fuc α (1-3)]GlcNAc β (1-2) as a novel terminal element in N-linked carbohydrate chains. *FEBS Lett.* 1992;314:389–94. doi:10.1016/0014-5793(92)81512-K.
340. Vliegthart JFG. The complexity of glycoprotein-derived glycans. *Proc Japan Acad Ser B.* 2017;93:64–86. doi:10.2183/pjab.93.005.
341. Barbeyron T, Brillet-Guéguen L, Carré W, Carrière C, Caron C, Czjzek M, et al. Matching the Diversity of Sulfated Biomolecules: Creation of a Classification Database for Sulfatases Reflecting Their Substrate Specificity. *PLoS One.* 2016;11:e0164846. doi:10.1371/journal.pone.0164846.
342. Bond CS, Clements PR, Ashby SJ, Collyer CA, Harrop SJ, Hopwood JJ, et al. Structure of a human lysosomal sulfatase. *Structure.* 1997;5:277–89. doi:10.1016/S0969-2126(97)00185-8.
343. Staiano M, Pennacchio A, Varriale A, Capo A, Majoli A, Capacchione C, et al. Enzymes as Sensors. In: *Methods in Enzymology.* Academic Press Inc.; 2017. p. 115–31. doi:10.1016/bs.mie.2017.01.015.
344. Katoh T, Maeshibu T, Kikkawa K, Gotoh A, Tomabechi Y, Nakamura M, et al. Identification and characterization of a sulfoglycosidase from *Bifidobacterium bifidum* implicated in mucin glycan utilization. *Biosci Biotechnol Biochem.* 2017;81:2018–27. doi:10.1080/09168451.2017.1361810.
345. Rho J, Wright DP, Christie DL, Clinch K, Furneaux RH, Robertson AM. A Novel Mechanism for Desulfation of Mucin: Identification and Cloning of a Mucin-Desulfating

Glycosidase (Sulfoglycosidase) from *Prevotella* Strain RS2. *J Bacteriol.* 2005;187:1543–51. doi:10.1128/JB.187.5.1543-1551.2005.

346. Katoh T, Maeshibu T, Kikkawa K, Gotoh A, Tomabechi Y, Nakamura M, et al. Identification and characterization of a sulfoglycosidase from *Bifidobacterium bifidum* implicated in mucin glycan utilization. *Biosci Biotechnol Biochem.* 2017;81:2018–27. doi:10.1080/09168451.2017.1361810.

347. Rho J, Wright DP, Christie DL, Clinch K, Furneaux RH, Robertson AM. A Novel Mechanism for Desulfation of Mucin: Identification and Cloning of a Mucin-Desulfating Glycosidase (Sulfoglycosidase) from *Prevotella* Strain RS2. *J Bacteriol.* 2005;187:1543–51. doi:10.1128/JB.187.5.1543-1551.2005.

348. Parenti G, Meroni G, Ballabio A. The sulfatase gene family. *Curr Opin Genet Dev.* 1997;7:386–91. doi:10.1016/S0959-437X(97)80153-0.

349. Schmidt B, Selmer T, Ingendoh A, Figurat K von. A novel amino acid modification in sulfatases that is defective in multiple sulfatase deficiency. *Cell.* 1995;82:271–8. doi:10.1016/0092-8674(95)90314-3.

350. Benjdia A, Leprince J, Guillot A, Vaudry H, Rabot S, Berteau O. Anaerobic Sulfatase-Maturing Enzymes: Radical SAM Enzymes Able To Catalyze *In Vitro* Sulfatase Post-translational Modification. *J Am Chem Soc.* 2007;129:3462–3. doi:10.1021/ja067175e.

351. Carlson BL, Ballister ER, Skordalakes E, King DS, Breidenbach MA, Gilmore SA, et al. Function and Structure of a Prokaryotic Formylglycine-generating Enzyme. *J Biol Chem.* 2008;283:20117–25. doi:10.1074/jbc.M800217200.

352. Murooka Y, Ishibashi K, Yasumoto M, Sasaki M, Sugino H, Azakami H, et al. A sulfur- and tyramine-regulated *Klebsiella aerogenes* operon containing the arylsulfatase (*atsA*) gene and the *atsB* gene. *J Bacteriol.* 1990;172:2131–40. doi:10.1128/JB.172.4.2131-2140.1990.

353. Lanz ND, Booker SJ. Auxiliary iron-sulfur cofactors in radical SAM enzymes. *Biochimica et Biophysica Acta - Molecular Cell Research.* 2015;1853:1316–34. doi:10.1016/j.bbamcr.2015.01.002.

354. Paschinger K, Wilson IBH. Analysis of zwitterionic and anionic N-linked glycans from invertebrates and protists by mass spectrometry. *Glycoconj J.* 2016;33:273–83. doi:10.1007/s10719-016-9650-x.

355. Thongsomboon W, Serra DO, Possling A, Hadjineophytou C, Hengge R, Cegelski L. Phosphoethanolamine cellulose: A naturally produced chemically modified cellulose. *Science* (80-). 2018;359:334–8. doi:10.1126/science.aao4096.

356. Zatopek KM, Gardner AF, Kelman Z. Archaeal DNA replication and repair: new genetic, biophysical and molecular tools for discovering and characterizing enzymes, pathways and mechanisms. *FEMS microbiology reviews.* 2018;42:477–88. doi:10.1093/femsre/fuy017.

357. Mereiter S, Balmaña M, Campos D, Gomes J, Reis CA. Glycosylation in the Era of Cancer-Targeted Therapy: Where Are We Heading? *Cancer Cell.* 2019;36:6–16. doi:10.1016/j.ccell.2019.06.006.

List of tables and figures

Table 1. NEB metagenomic and genomic library collection.	58
Table 2. Sanger sequencing analysis of randomly selected clones from the Dixie metagenomics library.....	63
Table 3. Enzymes and donors involved in post glycosylation modifications (PGMs).	143
Figure 1. Sequence- and function-based metagenomics.	7
Figure 2. Diversity of glycoconjugates in mammalian cells.	22
Figure 3. Main approaches employed in glycoanalytics.	32
Figure 4. Enzymes in the glycoanalytical toolbox.	34
Figure 5. Functional metagenomics workflow.	37
Figure 6. Fosmid copy control system.....	44
Figure 7. Construction of fosmid metagenomic libraries.	45
Figure 8. De novo assembler overlap error on circular fosmids.	52
Figure 9. Environmental DNA extraction from different ecosystems.	55
Figure 10. Distribution of libraries comprising the NEB Collection.	59
Figure 11. Restriction fragment analysis of 12 clones from the human gut metagenomic library.	63
Figure 12. Chromogenic and fluorogenic substrate for glycoside hydrolase screening.	65
Figure 13. Plate-based screening for β -galactosidase and sialidase activities.	66
Figure 14. Development of a lysate-based screening assay.	68
Figure 15. Definition of screening hits.	71
Figure 16. Size distribution of a multiplex PacBio library before and after size selection.	73
Figure 17. Fosmid sequencing data processing.	75
Figure 18. ORF map of a fosmid clone from the human gut metagenomic library. .	76
Figure 19. Sialic acid core structures.	80
Figure 20. Retaining exosialidase and inverting endosialidases.	84
Figure 21. Restriction fragment analysis of the ‘Small Dixie’ metagenomic library. .	92
Figure 22. Screening for sialidase activity from a hot spring metagenomic library. ...	93
Figure 23. Predicted ORFs encoded in the fosmid G7 nucleotide sequence.	94
Figure 24. Identification of the sialidase-encoding ORF on fosmid G7 and its in vitro expression.	96
Figure 25. Purification and biochemical characterization of recombinant ORF12p. .	98
Figure 26. Specificity of ORF12p on sialic acid containing substrates using UPLC–HILIC–FLR analysis.....	99

Figure 27. ORF12p activity on a fetuin 2AB-labeled O-glycan library.....	101
Figure 28. Stereochemical course of the hydrolysis reaction catalyzed by ORF12p (B) compared with NeuA (C).	102
Figure 29. ORF12 protein family.	105
Figure 30. Wild-type and SeMet-labeled EnvSia156 expressed proteins.	116
Figure 31. Analysis of the solution oligomeric state of EnvSia156 by SEC-MALLS. ...	117
Figure 32. 3D structure of EnvSia156.	118
Figure 33. Ribbon representation with rainbow coloring of a representative member of each sialidase family.....	120
Figure 34. Crystal structures of EnvSia156-ligand complexes.	122
Figure 35. EnvSia156 homology and binding subsites.	124
Figure 36. Catalytic mechanism of retaining exoglycosidases in contrast to a typical inverting glycoside hydrolases.	125
Figure 37. EnvSia156 proposed catalytic mechanism.....	127
Figure 38. Neu5Gc and Neu5Ac differential screening of a compost metagenomic library.	133
Figure 39. Clone C19 and C22 ORF map.	134
Figure 40. Neu5Gc and Neu5Ac activity of a bacterial exosialidase panel.	136
Figure 41. Sialidase26 (A) and sialidaseHz136 (B) substrate preference.	138
Figure 42. Phylogenetic relationship of sialidases.	139
Figure 43. Structure of heparin.	142
Figure 44. N- and O-glycans can be chemically modified with sulfate.	146
Figure 45. Screening a human gut metagenomic library for sulfatases.....	155
Figure 46. Background fluorescence of F13 and F15.	158
Figure 47. Function- and sequence-based analysis of fosmids F1-F12.....	160
Figure 48. In vitro expression of sulfatase candidates.....	161
Figure 49. ORF7, 13 and 16 relationships to other GlcNAc-6-sulfatases.	162
Figure 50. F1-ORF13 sulfatase specificity.	163
Figure 51. F1-ORF13 sulfatase activity on N-glycans.	165
Figure 52. F1-ORF13 sulfatase activity on an APTS-labeled N-glycan isolated from human urokinase.	166
Figure 53. F1-ORF13 activity in absence of calcium.....	168
Figure 54. Use of F1-ORF13 in a N-glycan enrichment strategy.	170
Figure 55. <i>In vitro</i> expression of hexosaminidase clones F3-ORF26 and F10-ORF19.	171
Figure 56. F10-ORF19 and F3-ORF26 are members of glycoside hydrolase family 20 (GH20).	173
Figure 57. F10-ORF19 hexosaminidase activity on N-glycans.	175
Figure 58. F10-ORF19 hexosaminidase activity on a urokinase isolated N-glycan...	176
Figure 59. Improving the existing workflow.....	185

Supplementary table 1. Primer sequences used in this thesis. 216

Supplementary Figure 1. PacBio library preparation 217

Supplementary Figure 2. From MetaGeneMark table to BED file 218

Supplementary Figure 3. pCC1FOS vector map. 219

Supplementary Figure 4. Alignment of EnvSia156 with its closest homologs. 220

Supplementary Figure 5. C19-ORF3 and C22-ORF6 exosialidase alignment. 222

Supplementary Figure 6. Open reading frame (ORF) maps of 19 sulfatase hits. 228

Supplementary Figure 7. Partially purified F1-ORF13 and F10-ORF19. 229

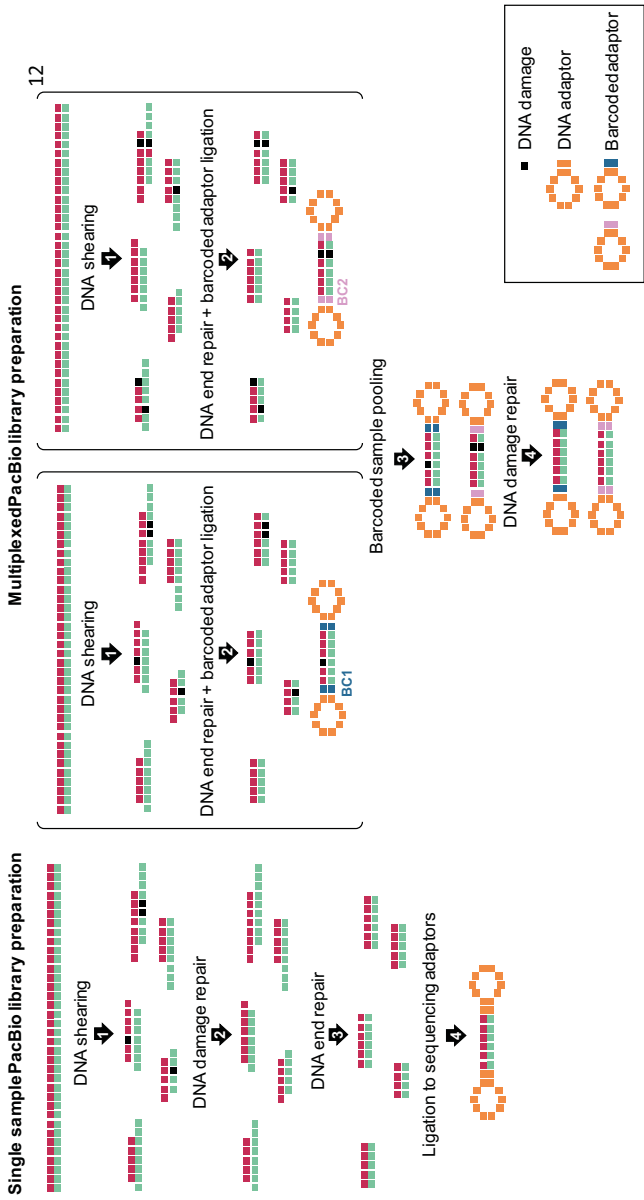
Appendix

Primer name	Primer sequence 5' → 3'	Use
T7 universal primer	TAATACGACTCACTATAGGG	Sanger sequencing primers for
pCC1 Forward primer	GGATGTGCTGCAAGGCGATTAAGTTGG	assessment of metagenomic library
pCC1 Reverse primer	CTCGTATGTTGTGTGGAATTGTGAGC	diversity
ORF12/EnvSia156 PURExpress® forward primer	GCGAATTAATACGACTCACTATAGGGCT TAAGTATAAGGAGGAAAAAATATGAGG CCGGAGACAATACCGGGGATC	PCR primers for generation of an ORF12/EnvSia156 template for <i>in vitro</i> transcription/translation
ORF12/EnvSia156 PURExpress® reverse primer	AAACCCCTCCGTTTAGAGAGGGTTATG CTAGTCAGGAGTGCCAGGGCGTATGA GAAA	
ORF12/EnvSia156-6His HiFi forward primer	TAAGCTTAGGAGGTTAACATATGAGGCC GGAGACAATACC	Primers for cloning of histidine tagged ORF12/EnvSia156 into the expression vector pJS119K
ORF12/EnvSia156-6His HiFi reverse primer	TCAGTGATGGTGATGGTATGGGAGTGC CAGGGGCGTAT	
pJS119K forward primer (for cloning ORF12/EnvSia156)	CATCACCATCACCATCACTGAGAATTCA GCTTGGCTGTTTTG	
pJS119K reverse primer (for cloning ORF12/EnvSia156)	ATGTTAACCTCCTAAGCTTAATTC	
Armatimonadetes sialidase PURExpress® forward primer	GCG AAT TAA TAC GAC TCA CTA TAG GGC TTA AGT ATA AGG AGG AAA AAA TAT GAA GGG TCC GAT CTT CAA CCA GGA CAG CAC CG	PCR primers for generation of a template for <i>in vitro</i> transcription/translation of ORF12/EnvSia156 homolog from Armatimonadetes (OIO94155)
Armatimonadetes sialidase PURExpress® reverse primer	AAA CCC CTC CGT TTA GAG AGG GGT TAT GCT AGT TAG CCG CTT TTC GCC ACC GCC AGC TC	
ORF12/EnvSia156 D14A forward primer	C TTG AAC GAG GCT AAC AGC CAT TAC	Primers for site directed mutagenesis of ORF12/EnvSia156-D14A and D14N
ORF12/EnvSia156 D14A and D14N reverse primer	GAG ATC CCC GGT ATT GTC	
ORF12/EnvSia156 D14N forward primer	C TTG AAC GAG AAT AAC AGC CAT TA C	
ORF12/EnvSia156 H134A forward primer	G AAC GAC GTG GCC TAT GTA AAT GAC G	Primers for site directed mutagenesis of ORF12/EnvSia156-H134A and H134N
ORF12/EnvSia156 H134A and H134N reverse primer	ATG CGC ATA GAG ATC CAG	
ORF12/EnvSia156 H134N forward primer	G AAC GAC GTG AAC TAT GTA AAT G	
C19-6His HiFi forward primer	GAGATATACAATGAAAAGAACCGGCTG GCTGTTGCTGCTTGTGAGCGGAC	Primers for cloning of histidine tagged C19 sialidase into the expression vector pET21a(+)
C19-6His HiFi reverse primer	GTGATGATGATGATGATGCCGGCCCGC CCTTGCGCCC	
pET21a(+) forward primer (for cloning C19)	CGCGGGCCGGCATCATCATCATCAC TGATAAG	
pET21a(+) reverse primer (for cloning C19)	TTCTTTTCATTGTATATCTCCTTCTTAAA GTTAAACAAAATTATTCTAGAGG	

C22-6His HiFi forward primer	GAGATATACAATGGTGATTGCCTGGATG GC	Primers for cloning of histidine tagged C22 sialidase into the expression vector pET21a(+)
C22-6His HiFi reverse primer	TGGTGGTGGTGGTGCCTATCAGTGATGA TGATGATGATGTGG	
pET21a(+) forward primer (for cloning C22)	TCACTGATAAGCACCACCACCACCA CT	
pET21a(+) reverse primer (for cloning C22)	CAATCACCATTGTATATCTCCTTCTTAA AGTTAAACAAAATTATTCTAGAGG	
F1-ORF13 PURExpress® forward primer	GCGAATTAATACGACTCACTATAGGGCT TAAGTATAAGGAGGAAAAAATATGGTA GGAATTATGTTAATAAAAAA	PCR primers for generation of a template for <i>in vitro</i> transcription/translation of F1-ORF13 (sulfatase)
F1-ORF13 PURExpress® reverse primer	AAACCCCTCCGTTTAGAGAGGGGTTATG CTAGTTAGTCTTTGTATATATCCGGAT AGG	
F5-ORF2 PURExpress® forward primer	GCG AAT TAA TAC GAC TCA CTA TAG GGC TTA AGT ATA AGG AGG AAA AAA TAT GAG TTA TGT TCC GGG GGT CGC CAA C	PCR primers for generation of a template for <i>in vitro</i> transcription/translation of F5-ORF2 (sulfatase)
F5-ORF2 PURExpress® reverse primer	AAA CCC CTC CGT TTA GAG AGG GGT TAT GCT AGT TAC CGG CGA ACA GCG AAT CGA AGA AA	
F6-ORF7 PURExpress® forward primer	GCG AAT TAA TAC GAC TCA CTA TAG GGC TTA AGT ATA AGG AGG AAA AAA TAT GGA ACA TCA AAA CAA ATT GAT TTA T	PCR primers for generation of a template for <i>in vitro</i> transcription/translation of F6-ORF7 (sulfatase)
F6-ORF7 PURExpress® reverse primer	AAA CCC CTC CGT TTA GAG AGG GGT TAT GCT AGT TAT TTT ACA AAG TCT GTA TCG TTA TA	
F6-ORF9 PURExpress® forward primer	GCG AAT TAA TAC GAC TCA CTA TAG GGC TTA AGT ATA AGG AGG AAA AAA TAT GGC CGG TAG CCT TGG TTT GTC CGC T	PCR primers for generation of a template for <i>in vitro</i> transcription/translation of F6-ORF9 (sulfatase)
F6-ORF9 PURExpress® reverse primer	AAA CCC CTC CGT TTA GAG AGG GGT TAT GCT AGT TAT TTT TTA TTT GTT GGA TAA TTC GG	
F8-ORF12 PURExpress® forward primer	GCG AAT TAA TAC GAC TCA CTA TAG GGC TTA AGT ATA AGG AGG AAA AAA TAT GAA ACA AAC AGT TAT AGC TTT AGG A	PCR primers for generation of a template for <i>in vitro</i> transcription/translation of F8-ORF12 (sulfatase)
F8-ORF12 PURExpress® reverse primer	AAA CCC CTC CGT TTA GAG AGG GGT TAT GCT AGT TAT AAG GTT TCC ATG TTA GAA AGT AA	
F8-ORF16 PURExpress® forward primer	GCG AAT TAA TAC GAC TCA CTA TAG GGC TTA AGT ATA AGG AGG AAA AAA TAT GAA AAA CTT ACA ATC AGG ATT ACT C	PCR primers for generation of a template for <i>in vitro</i> transcription/translation of F8-ORF16 (sulfatase)
F8-ORF16 PURExpress® reverse primer	AAA CCC CTC CGT TTA GAG AGG GGT TAT GCT AGT TAT TCT TTA TCT CTC TCT GGG GAA AA	

F10-ORF22 PURExpress® forward primer	GCG AAT TAA TAC GAC TCA CTA TAG GGC TTA AGT ATA AGG AGG AAA AAA TAT GAA TTA CAA ATC TAT ATC ATT AAT A	PCR primers for generation of a template for <i>in vitro</i> transcription/translation of F10-ORF22 (sulfatase)
F10-ORF22 PURExpress® reverse primer	AAA CCC CTC CGT TTA GAG AGG GGT TAT GCT AGT TAC TGC TTA GGC AGT GTC ACA GAA AA	
F10-ORF23 PURExpress® forward primer	GCG AAT TAA TAC GAC TCA CTA TAG GGC TTA AGT ATA AGG AGG AAA AAA TAT GAA ACA ACC TTT GCT TTT TAC CCT T	PCR primers for generation of a template for <i>in vitro</i> transcription/translation of F10-ORF23 (sulfatase)
F10-ORF23 PURExpress® reverse primer	AAA CCC CTC CGT TTA GAG AGG GGT TAT GCT AGT TAT TTG CCG ATA AGG TCT TTT CCT TC	
F3-ORF26 PURExpress® forward primer	GCG AAT TAA TAC GAC TCA CTA TAG GGC TTA AGT ATA AGG AGG AAA AAA TAT GAA AAA ACA ACT GAT GCA ATG GGC A	PCR primers for generation of a template for <i>in vitro</i> transcription/translation of F3-ORF26 (hexosaminidase)
F3-ORF26 PURExpress® reverse primer	AAA CCC CTC CGT TTA GAG AGG GGT TAT GCT AGT TAC TCC ACT CCT ATT TCG TCA ACA AA	
F10-ORF19 PURExpress® forward primer	GCG AAT TAA TAC GAC TCA CTA TAG GGC TTA AGT ATA AGG AGG AAA AAA TAT GAA AAA CAA GTA TCT TTT ATC TTT A	PCR primers for generation of a template for <i>in vitro</i> transcription/translation of F10-ORF19 (hexosaminidase)
F10-ORF19 PURExpress® reverse primer	AAA CCC CTC CGT TTA GAG AGG GGT TAT GCT AGT TAC TTC ATT ATA TGT TTG CCG TAA TT	
F10-ORF19-6His HiFi pET28c(+) forward	CTTTAAGAAGGAGATATACCATGCAAG AAATAGCTATCATTCCGC	Primers for cloning histidin tagged F10-ORF19 without its signal sequence into pET28c(+)_B1006
F10-ORF19-6His HiFi pET28c(+) reverse	CAGTGGTGGTGGTGGTGGTGCTTCATTA TATGTTTGCCGTAATTCC	
pET28c(+) forward primer (for cloning F10-ORF19)	ACGGCAAACATATAATGAAGCACCACC ACCACCACCACTGAGATC	
pET28c(+) reverse primer (for cloning F10-ORF19)	ATGATAGCTATTTCTTGCATGGTATATC TCCTTCTTAAAGTTAAACAAAATTATTT CTAGAGGGGAATTGTTATC	
pET21a(+) forward primer (for cloning F1-ORF2)	AACCTGAACTATTTATATAAGAAATAA TTTTGTTTAACTTAAAGAAG	
pET21a(+) reverse primer (for cloning F1-ORF2)	TGCCCATTTGATATCTCCTTAGAGGGGA ATTGTTATCC	Primers for cloning F1-ORF2 with an upstream RBS into pET21a(+)_F1-ORF13
RBS-F1-ORF2 forward primer	aagagatatacaATGGGCAAAAGAATAGAAA T	
RBS-F1-ORF2 reverse primer	TTATATAAATAGTTCAGGTTACGGC	

Supplementary table 1. Primer sequences used in this thesis.



Supplementary Figure 1. PacBio library preparation

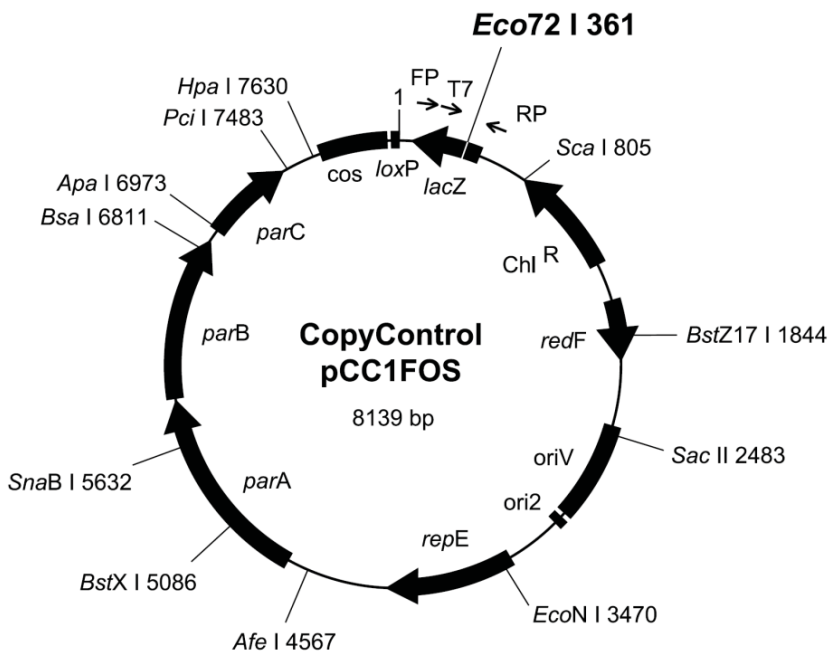
A

Gene #	MetaGeneMark output file				
	Strand	LeftEnd	RightEnd	Gene length	Class
1	-	<1	618	618	1
2	+	994	2700	1707	1
3	+	2734	4767	2034	1
4	+	4781	5350	570	1
5	+	5621	6322	702	1
6	-	6469	7158	690	1
7	-	7195	7830	636	1
8	-	8240	9586	1347	1
9	+	9980	12199	2220	1
10	+	12219	14006	1788	1
11	+	14155	15024	870	1
12	+	15085	15540	456	1
13	+	15850	16644	795	1
14	-	17006	17389	384	1
15	+	17552	18520	969	1
16	-	18686	19240	555	1
17	+	19377	19469	93	1
18	+	19454	20221	768	1
19	+	20403	21530	1128	1
20	+	21543	23012	1470	1
21	-	23083	23751	669	1
22	-	23767	25002	1236	1
23	+	25098	25904	807	1
24	+	25894	26424	531	1
25	+	26718	26963	246	1
26	-	27153	29474	2322	1
27	-	29520	30434	915	1
28	-	30671	31918	1248	1
29	+	32027	32947	921	1
30	-	33055	>33597	543	1

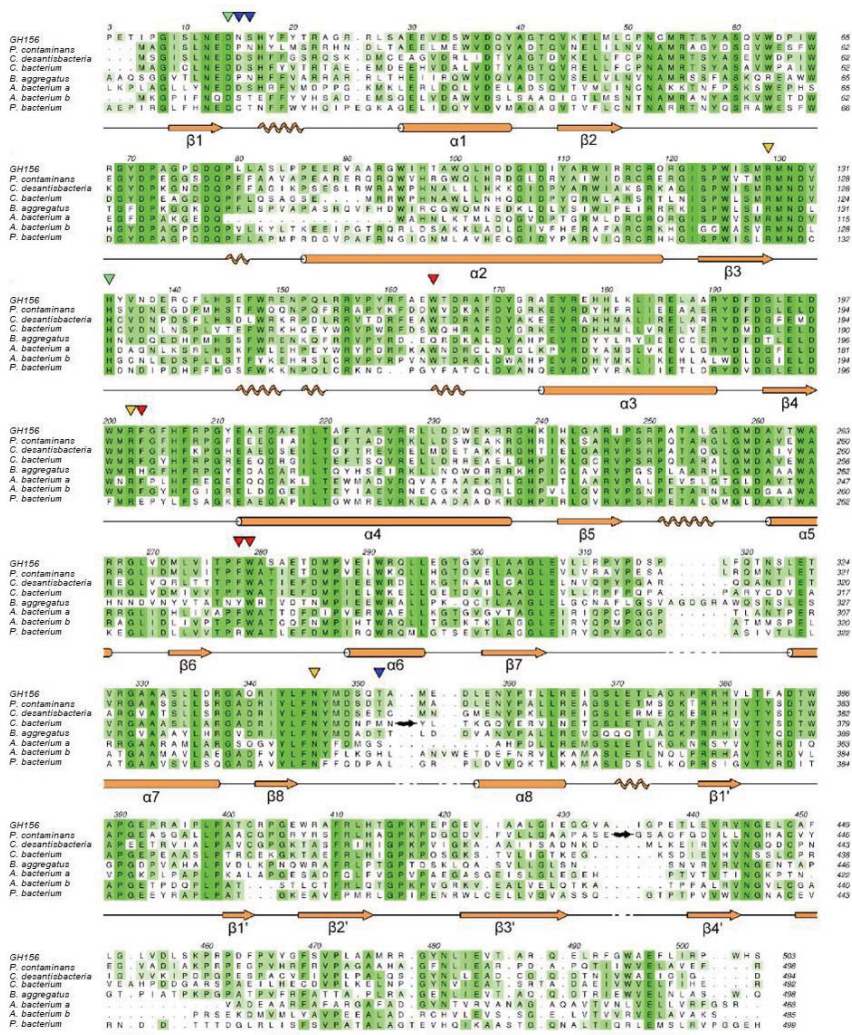
B

Clone I.D.	Bed format converted				
	LeftEnd	RightEnd	Gene name	Score	Strand
NEB136-D20	1	618	ORF1	1000	-
NEB136-D20	994	2700	ORF2	1000	+
NEB136-D20	2734	4767	ORF3	1000	+
NEB136-D20	4781	5350	ORF4	1000	+
NEB136-D20	5621	6322	ORF5	1000	+
NEB136-D20	6469	7158	ORF6	1000	-
NEB136-D20	7195	7830	ORF7	1000	-
NEB136-D20	8240	9586	ORF8	1000	-
NEB136-D20	9980	12199	ORF9	1000	+
NEB136-D20	12219	14006	ORF10	1000	+
NEB136-D20	14155	15024	ORF11	1000	+
NEB136-D20	15085	15540	ORF12	1000	+
NEB136-D20	15850	16644	ORF13	1000	+
NEB136-D20	17006	17389	ORF14	1000	-
NEB136-D20	17552	18520	ORF15	1000	+
NEB136-D20	18686	19240	ORF16	1000	-
NEB136-D20	19377	19469	ORF17	1000	+
NEB136-D20	19454	20221	ORF18	1000	+
NEB136-D20	20403	21530	ORF19	1000	+
NEB136-D20	21543	23012	ORF20	1000	+
NEB136-D20	23083	23751	ORF21	1000	-
NEB136-D20	23767	25002	ORF22	1000	-
NEB136-D20	25098	25904	ORF23	1000	+
NEB136-D20	25894	26424	ORF24	1000	-
NEB136-D20	26718	26963	ORF25	1000	+
NEB136-D20	27153	29474	ORF26	1000	-
NEB136-D20	29520	30434	ORF27	1000	-
NEB136-D20	30671	31918	ORF28	1000	-
NEB136-D20	32027	32947	ORF29	1000	-
NEB136-D20	33055	33597	ORF30	1000	-

Supplementary Figure 2. From MetaGeneMark table to BED file



Supplementary Figure 3. pCC1FOS vector map. (Figure from the CopyControl Fosmid library production kit manual, Lucigen Corporations, Middleton, WI)



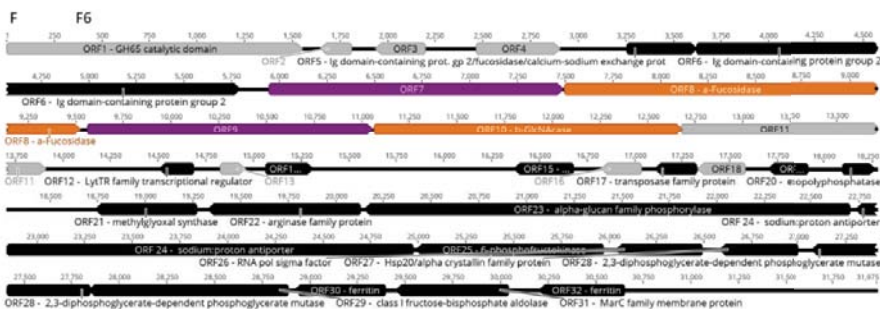
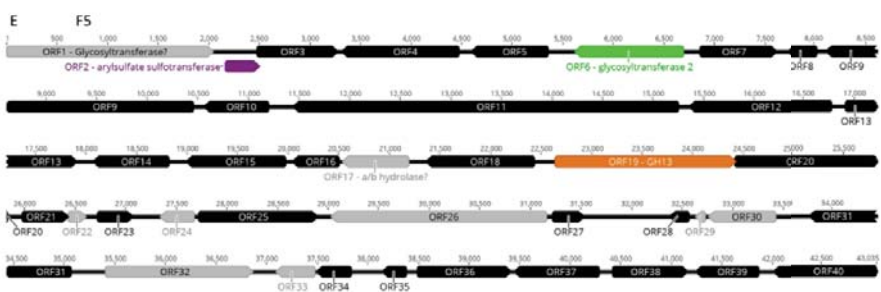
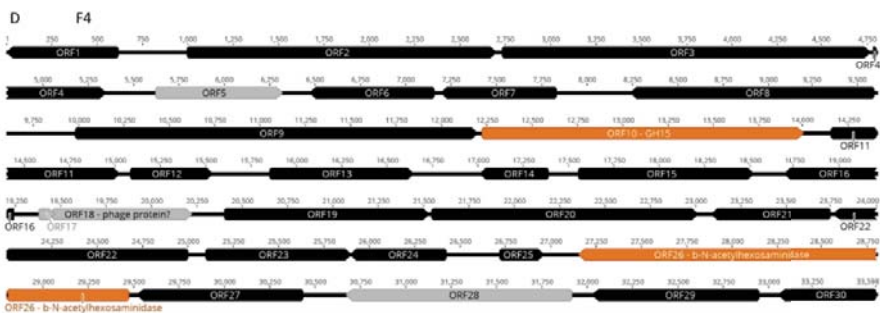
Supplementary Figure 4. Alignment of EnvSia156 with its closest homologs. Alignment of EnvSia156 with the seven closest sequence homologs found with the BLASTP tool. All proteins belong to bacteria and are from *Paenibacillus contaminans*, *Candidatus desantisbacteria*, *Chloroflexi bacterium*, *Bryobacter aggregatus*, *Armatimonadetes bacterium* (2)

sequences), and *Planctomycetes bacterium*. A cartoon representation of EnvSia156 secondary structure is displayed in orange, below the alignment. The sequences were aligned using the CLUSTAL Omega tool and further processed with ALINE. Coloring according to similarity was implemented with ALINE: dark green, identical residues; green to white, lowering color-ramped scale of conservation. Green arrows point to the EnvSia156 putative catalytic pair, yellow arrows to the carboxylate coordinating triad, red arrows to the residues defining the predicted +1 subsite and the blue arrows to the remaining residues establishing contacts with Neu5Ac as seen in the EnvSia156Neu5Ac complex structure. (From [267])

C19_exosialidase	MKRTGWLILLVSGPALCQ--PPRGYSIPLIDLAAEAWRQTVVDREPGQYLGHPTTVLLED	58
C22_exosialidase	-MVIAMW-----AGVQCTAAALVDLGVTILDISGEFERQVIVDREPGQYLGHVSTVLEED	54
	.*: . . * . . : :*: :.* **:***** :*****	
C19_exosialidase	GRTVLAVYPKGHGRGAIIVYKRSRDGGRTWSARLFPVPENWATSQETPTIHRVVDPRGRKRL	118
C22_exosialidase	DKTI LAVYPKGHGRGAI VLKRSE DGGRTWSGR LFPVPASWATS KETPTI HRVVDARGRRL	114
	.*:***** :***** :***** :***** :***** :***** :***** :*****	
C19_exosialidase	ILFSGLYPIRMSVEDDGETWTPLAPIGNFGGVVAMASVERL-RDGRYMA LFHDDGRFLR	177
C22_exosialidase	LLWSGLYPARRALSDDGRTWTELEPVGEWGGIVVMGFVEATRQPGHYVAMFHDDGRYFA	174
	:*:***** * :*:***.*** * * :*:***:*. * ** : * :*:*****:	
C19_exosialidase	GGGKPD-RFVYKTLSGDGLTWSEVPVILSHPQAHLCEPGLLRSPDGRRIAILLRENSR	236
C22_exosialidase	AQPSTNRSM TLYLTRSSDGGVSWSSPVAVWSNSAVHLCEPGGIWSPDRRLAVLLRENRR	234
	. . : : :* * * :*:***:*** ** * : * :***** : ** ** :***** *	
C19_exosialidase	KFNSFVSFS DDEGETWSEPRELPGALTGDRHTAVYARDGR LFI SFRD T-----	285
C22_exosialidase	VNNSHIMFSDDEGHTWTT PVEMPLSRAGDRHTLRYT PDGRIVCVFRAVTPVGMGRGSFGQ	294
	** : ***** :*: * * * : * : * : ***** * : * : * : * * *	
C19_exosialidase	-----HESPTRGDWVAVWGRFGDIENRGQYRVR LMKNHKLDCCYPGVLR LPDDT	337
C22_exosialidase	NEDVDTLGVGSPFEGDCVAVWGTWDDL VHNRPQGQYVRL LNNRKGWDTTYPGVEVLPDGT	354
	** . * * ***** :. * : :. * ** ***** :* . * * * * * *	
C19_exosialidase	ILTTYGHWTPEPPYIVSIRLRLELDRRAQQAGR--	373
C22_exosialidase	VVVTYGHWNAGEPPYIRSVRFRLEELDRMAAKASQKP	392
	: :***** . ***** *:***.***** * * : :	

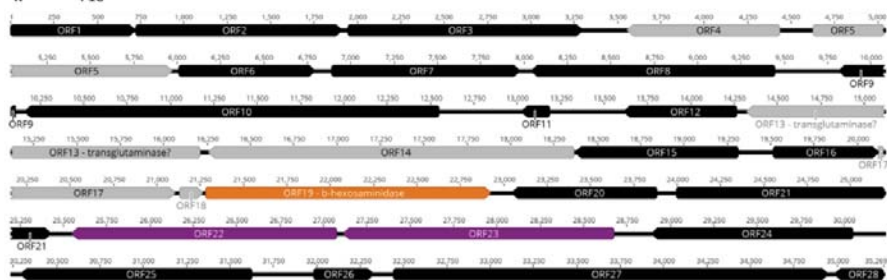
Supplementary Figure 5. C19-ORF3 and C22-ORF6 exosialidase alignment. Protein sequence alignment of C19-ORF3 and C22-ORF6 sialidases with Clustal Omega. (From [128])



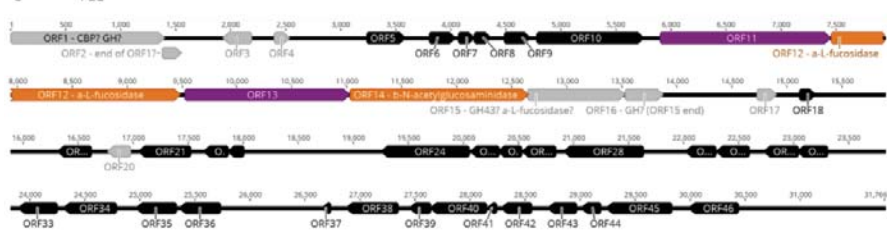




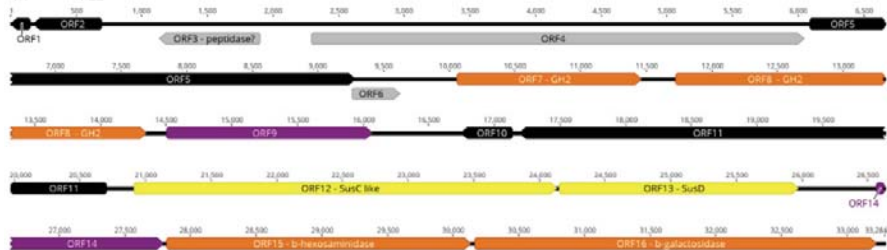
K F10

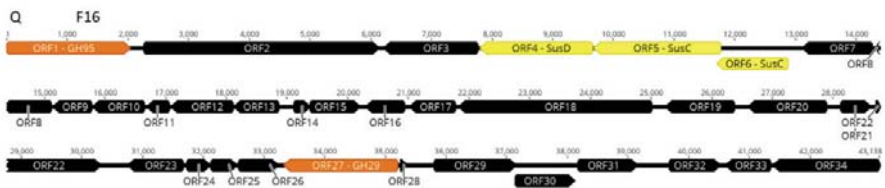
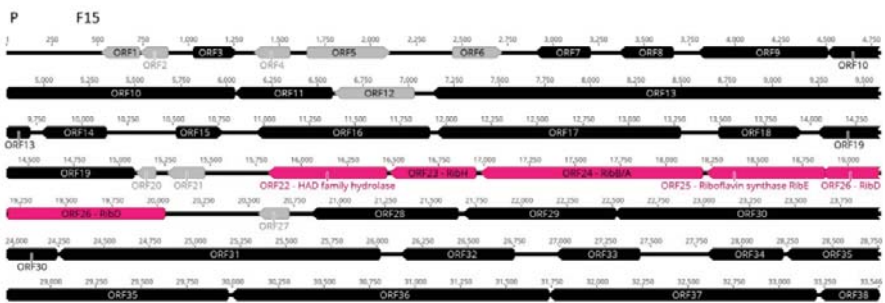
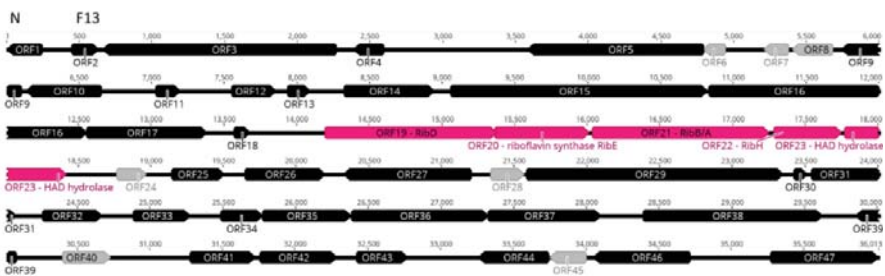


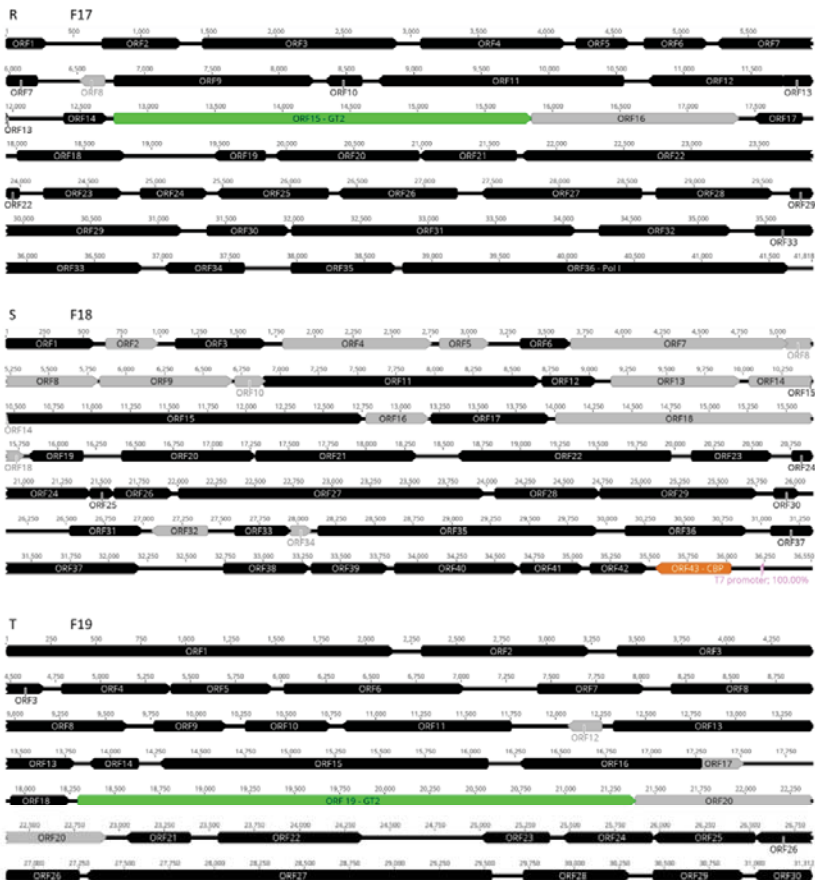
L F11



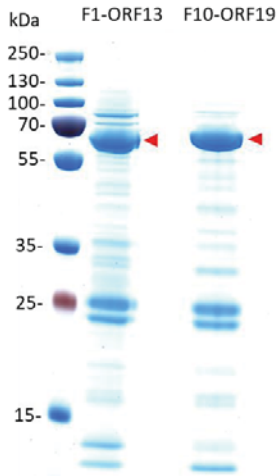
M F12







Supplementary Figure 6. Open reading frame (ORF) maps of 19 sulfatase hits. An ORF map for 19 hits was produced using Geneious (<https://www.geneious.com>). ORFs that encode for sulfatases are colored in purple, ORFs that encode for glycoside hydrolase including hexosaminidases are colored in orange, ORFs that encode for the sugar transporter SusC/D pair are colored in yellow, ORFs that encode for glycosyl transferase are colored in green, ORFs belonging to the riboflavin biosynthesis pathway are colored in pink, ORFs in black are ORFs with significant match to protein of known function while ORFs with very little to no homology to known protein are represented in grey. ORF2 from F1 (A) is colored in red as assignment of its function was attempted in this work.



Supplementary Figure 7. Partially purified F1-ORF13 and F10-ORF19. F1-ORF13 sulfatase and F10-ORF19 hexosaminidase were over-expressed in *E. coli* and partially purified on a His-Trap column.

