The new families of β-(1→3)-glucan phosphorylases: identification, mechanisms and applications in carbohydrate synthesis

Sakonwan Kuhaudomlarp

Student number: 100084376

A thesis submitted to the University of East Anglia for the degree of Doctor of Philosophy

John Innes Centre

Norwich

August 2018

Thesis words: 50,077

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that use of any information derived there-from must be in accordance with current UK Copyright Law. In addition, any quotation or extract must include full attribution.

Abstract

β-(1→3)-glucans have applications in human food and health supplements, owning to their reported immunomodulatory properties. Enzymatic synthesis of β-(1→3)-glucans are attractive due to stereo- and regio-specificity of the biocatalytic enzymes. Glycoside phosphorylases (GPs) catalyse phosphorolysis of glycan into sugar 1-phosphate and shortened glycan chain, the reaction of which can be reversed for glycan synthesis. In order to apply GPs as biocatalyst of β-(1→3)-gluco-oligosaccharide production, this work concentrated on identification of sequences, biochemical characterisation and structural elucidation of GPs acting on β-(1→3)-gluco-oligosaccharides, in order to understand the mechanism underlying the enzyme substrate specificity.

A GP from *Paenibacillus* sp. YM-1 (*Ps*LBP), belonging to glycoside hydrolase (GH) family 94, has been reported for its specificity towards a disaccharide, β-D-glucopyranosyl-(1→3)-D-glucopyranose**.** In this study, the activity of *Ps*LBP on sugar donors, α-D-mannose 1-phosphate and α-D-glucose 1-phosphate, was investigated. X-ray crystallography and saturation transfer difference NMR were used to investigate the interaction between *Ps*LBP and its donor substrates, and a mechanism underlying the substrate chain length specificity of the enzyme.

Two new families, GH149 and GHyyy, were discovered. Two GH149 candidates, a *Euglena gracilis* phosphorylase 1 and Pro_7066, were characterised to confirm their function as β-(1→3)-glucan phosphorylase acting on disaccharide and longer substrates. A GHyyy from *Paenibacillus polymyxa* was characterised to confirm its β-(1→3)-glucan phosphorylase activity, albeit with a distinct substrate length specificity from the GH149 enzymes. Multiple alignment of GH94, GH149 and GHyyy proteins identified conserved amino acids located in the enzyme active site, implying a conserved catalytic mechanism. Structural studies of Pro_7066 revealed two additional domains and an oligosaccharide surface binding site, which are unique to GH149 and may contribute to the enzyme preference for long oligosaccharide substrates. Bacterial *GH149* or *GHyyy* genes map to gene clusters containing genes encoding other GHs and membrane transporters, highlighting the involvement of the enzymes in polysaccharide degradation.

Acknowledgements

I would like to express my sincere gratitude toward all individuals who have made my journey through PhD memorable and enjoyable.

First and foremost: I would like to acknowledge Prof. Rob Field for his supervision, valuable trust and confidence in my ability to perform and develop as a scientist. I would like to also thank my supervisory committee, Dr Nicola Patron and Dr Stephen Bornemann for their guidance and advice.

I would like to thank all members of the Rob Field group: Dr Martin Rejzek and Dr Giulia Pergolizzi for their advice on carbohydrate analysis, Dr Ellis O'Neill, the past member who initiated the interests in Euglena glycobiology research, Dr Michael Rugen for his expertise on molecular microbiology and protein expression, Dr Sergey Nepogodiev for his expertise on NMR, and others who provided friendly working environment during my time in the laboratory.

My work would be impossible without scientific supports from JIC platforms and UEA collaboration. I would like to acknowledge them for their valuable contribution in my work: Dr Clare Stevenson and Dr David Lawson for their help on X-ray crystallographic works; Dr Gerhard Saalbach for proteomics and mass spectrometry analysis, and Dr Jesus Angulo and his group for the STD NMR analysis.

I would like to thank the funding bodies: Thai Government Scholarship and John Innes Foundation. Thanks also goes to Dr Stephen Bornemann and Prof. Nick Brewin for their support during the rotation year of the PhD programme. Special thanks go to the rotation project supervisors, Dr Diane Saunders and Dr Jacob Malone, for having me in their groups during the rotation year.

I would like to thank my family for their unwavering support for the past 10 years of my education abroad. I would also like to acknowledge people I have befriended with during my time in Norwich, especially all my squash friends, who kept me physically and mentally healthy. Special thanks to David Youngs and Karan Cox for their valuable friendship and encouragement in and out of the sport, and everyone at East Anglia Tennis and Squash club for all the fun time I had outside science.

Table of contents

Abstract **i**
Acknowledgements **ii**
Table of contents **iii**
List of Figures **ix**
List of Tables **xiv**
Abbreviations **xvi**

Chapter 1 – Introduction **1**
1.1. Significance and diversity of carbohydrates 2
1.2. β-(1→3)-Glucan structures and applications 4
1.3. Identification and classification of carbohydrate-active enzymes 9
1.4. Biosynthesis of β-(1→3)-glucans 11
1.4.1. Bacterial curdlan 11
1.4.2. Fungal β-(1→3)-glucan 11
1.4.3. Callose 12
1.4.4. Chrysolaminarin from diatoms and laminarin from brown algae 12
1.4.5. Paramylon from *Euglena gracilis* 14
1.5. *In vitro* enzymatic syntheses of β-(1→3)-glucans 16
1.5.1. Transglycosylases 16
1.5.2. Glycosynthases 16
1.5.3. Glycosyltransferases (GTs) 17
1.5.4. Glycoside phosphorylases (GPs) 18
1.6. Understanding chain length control mechanisms in carbohydrate-active enzymes 20
1.6.1. Chain length control mechanism relying on the presence of sugar binding subsites 20
1.6.2. Chain length control by terminating enzymes 21
1.6.3. Chain length control by regulating the oligosaccharide binding pocket size 24
1.6.4. Auxiliary sites and their contributions to CAZymes activity on oligosaccharide and polysaccharide substrates 26
1.6.4.1. CBMs contribution to CAZymes activity and processivity on polysaccharides 26

1.6.4.2. Surface binding sites (SBSs) for oligosaccharides and their contributions to enzyme activity on large polysaccharide substrates 28
1.7. Aims of this study 31
1.8. References 32

Chapter 2 – Unravelling the specificity of a GH94 laminaribiose phosphorylase from *Paenibacillus* sp. YM-1 towards donor substrates glucose/mannose 1-phosphate using X-ray crystallography and STD NMR spectroscopy 45
2.1. Introduction 46
2.2. Results 50
2.2.1. Recombinant protein expression and *Ps*LBP activity on the cognate donor and acceptor 50
2.2.2. Activity towards noncognate donors 52
2.2.3. Overall crystal structures of *Ps*LBP 58
2.2.4. Phosphate recognition by *Ps*LBP 61
2.2.5. Glc1P and Man1P recognition by *Ps*LBP 64
2.2.6. Catalytic loop in GH94 enzymes 67
2.2.7. STD NMR studies on the interactions between *Ps*LBP and its ligands 69
2.3. Discussion 71
2.4. Conclusion and future perspectives 73
2.5. References 74

Chapter 3 – Identification of *Euglena gracilis* β-(1→3)-glucan phosphorylase and establishment of a new glycoside hydrolase (GH) family GH149 78
3.1. Introduction 79
3.2. Results 82
3.2.1. Identification of Euglena phosphorylase candidates through proteomic analysis 82
3.2.2. Recombinant protein production and biochemical characterisation of EgP1 87
3.2.3. Identification, recombinant protein production and characterisation of a bacterial orthologue of EgP1 93
3.2.4. Multiple amino acid sequence alignment and phylogeny of EgP1 and its orthologues 96
3.2.5. Structural studies of Pro_7066 99

3.2.5.1. Crystallisation of Pro_7066 99
3.2.5.2. Overall structure of Pro_7066 99
3.2.5.3. Pro_7066 active site 101
3.2.5.4. Structural elements surrounding Pro_7066 active site 102
3.2.5.5. Structure of Pro_7066 in complex with G6 revealed an oligosaccharide SBS on the surface of the catalytic domain 105
3.2.5.6. Conformation of G6 ligand at the Pro_7066 SBS 109
3.2.5.7. Effect of BCN on Pro_7066 activity 111
3.2.6. Homology modelling of EgP1 115
3.3. Discussion 118
3.4. Conclusion and future perspectives 128
3.5. References 129

Chapter 4 – A new glycoside hydrolase (GH) family, GHyyy, comprising β-(1→3)-glucan phosphorylases with different acceptor preferences to GH94 and GH149 134
4.1. Introduction 135
4.2. Results 138
4.2.1. Analysis of *Ochromonas* transcriptome revealed candidate β-(1→3)-D-glucan phosphorylases 138
4.2.2. Cloning, heterologous expression and biochemical characterisation of a GHyyy from *Paenibacillus polymyxa* ATCC 842 141
4.2.3. Gene clusters containing *GHyyy* genes in bacterial genomes suggesting a role of GHyyy in oligosaccharide degradation 149
4.2.4. Subcellular localisation of the heterokont GHyyy candidates 156
4.3. Discussion 158
4.4. Conclusion and future perspectives 162
4.5. References 163

Chapter 5 – Summary and translational opportunities 167
5.1. Introduction 168
5.2. Application of β-(1→3)-D-glucan phosphorylases in oligosaccharide synthesis 171
5.3. Activity of GH149 phosphorylases on a noncognate sugar donor Man1P 174
5.4. Chain length control strategies 176
5.5. Protein engineering and directed evolution strategy for GP improvement 177

5.6. Strategies for high-throughput functional screening of new GP activities 179
5.7. Conclusion 181
5.8. References 182

Chapter 6 – Materials and methods 184
6.1. Euglena culture 185
6.2. Anion exchange chromatography (AIEX) 185
6.3. DNJ affinity column preparation 185
6.4. DNJ affinity chromatography 185
6.5. Proteomic analysis 186
6.6. Generation of *E. coli* competent cells for transformation by heat shock 187
6.7. General cloning protocol 187
6.8. Recombinant protein production in *E. coli* 187
6.8.1. EgP1 187
6.8.2. Pro_7066 188
6.8.3. *Ps*LBP 188
6.8.4. *Pap*P 188
6.9. Recombinant protein purification 189
6.10. Determination of protein molecular masses by gel filtration analysis 189
6.11. SDS-PAGE analysis 190
6.12. Enzyme assays 190
6.12.1. The glycan synthetic assays 190
6.12.1.1. Assay on GH149 enzymes 190
6.12.1.2. *Ps*LBP enzymatic assays 191
6.12.2. The phosphorolysis assays 191
6.13. Oligosaccharide analyses 192
6.13.1. Thin layer chromatography (TLC) 192
6.13.2. HPAEC-PAD analyses 192
6.14. Specific experiments for disaccharide **1** 192
6.14.1. Preparation of disaccharide **1** 192
6.14.2. Purification of disaccharide **1** 193
6.14.3. Optical rotation analysis 193
6.14.4. NMR analysis 193
6.14.5. TLC/MS analysis 193
6.14.6. Physical data for disaccharide **1** 194
6.15. Crystallographic methods 194
6.15.1. *Ps*LBP 194

6.15.2. Pro_7066 196
6.15.3. EgP1 198
6.15.4. *Pap*P 198
6.16. Bioinformatics analyses 199
6.16.1. GH149 199
6.16.2. GHyyy 199
6.17. References 201

Appendices **202**
Appendix 1 – Chemical shifts of β-mannopyranoside in 13C NMR spectra of methyl β-D-mannopyranoside and disaccharide **1** 203
Appendix 2 – Simulation of 1H and 13C NMR spectra of disaccharide **1** using CASPER programme 204
Appendix 3 – List of Euglena proteins identified from affinity proteomics and their predicted functions 205
Appendix 4 – Sequence IDs for GH149 members 213
App. 4.1. GH149 sequences from NCBI database 213
App. 4.2. GH149 sequences from other databases 224
Appendix 5 – BLAST analyses of GH149 amino acid sequences against EgP1 226
Appendix 6 – GC content analysis of bacterial *GH149* DNA sequences 235
Appendix 7 – Sequence ID of GHyyy members 237
App. 7.1. Bacterial GHyyy sequences from NCBI 237
App. 7.2. Heterokont GHyyy sequences from MMET 250
Appendix 8 – BLAST analysis of GHyyy members against OcP1 252
App. 8.1. Sequences from MMET 252
App. 8.2. Sequences from NCBI 254
Appendix 9 – GC content analysis of bacterial *GHyyy* genes and their associated genomes 263
Appendix 10 – Synthetic genes 265
App. 10.1. EgP1 265
App. 10.2. EgP2 266
App. 10.3. EgP3 268
App. 10.4. EgP4 270
App. 10.5. Pro_7066 272
App. 10.6. *Pap*P 274
App. 10.7. *Ps*LBP 276
App. 10.8. *Prym*P 277

App. 10.9. OcP1 279
Appendix 11 – Primers for PCR amplification and cloning into pOPINF vector for gene expression in *E. coli* 282
Appendix 12 – Michaelis-Menten kinetic plots 283
App. 12.1. EgP1 283
App. 12.2. Pro_7066 285
App. 12.3. PapP 286
References 287

Lists of Figures

Chapter 1 – Introduction

Figure 1.1. Diversity of glucan chemical structures and their morphologies. 3

Figure 1.2. Structures and morphologies of linear β-(1→3)-glucans. 5

Figure 1.3. Structure of mixed-linkage β-(1→3;1→4)-glucans. 6

Figure 1.4. Chemical structures and morphologies of branched β-(1→3)-glucan. 7

Figure 1.5. Summary of proposed chrysolaminarin metabolic pathways in diatoms based on bioinformatics analysis of the genomes. 13

Figure 1.6. Laminarin metabolism in *E. siliculosus.* 14

Figure 1.7. Paramylon metabolism in *Euglena gracilis* based on bioinformatics prediction from the transcriptomic data. 15

Figure 1.8. Transglycosylation mechanism. 16

Figure 1.9. Glycosynthase mechanism and its application in β-(1→3)-glucan synthesis. 17

Figure 1.10. GP mechanisms. 18

Figure 1.11. Nomenclature for sugar binding subsite. 20

Figure 1.12. Proposed mechanism for chain length control in cyclic β-(1→2)-glucan production. 22

Figure 1.13. Length control mechanism for core lipid A structure of O9a in *E. coli*. 23

Figure 1.14. Comparison between a GH130_1 β-(1→4)-mannobiose phosphorylase and a GH130_2 β-(1→4)-mannopyranosyl-[*N*-glycan] phosphorylase. 24

Figure 1.15. Possible mechanism distinguishing substrate chain length preferences in *Rt*CDP and *Cg*CBP. 25

Figure 1.16. A hypothesis for the mode of action of processive GHs on crystalline cellulose. 27

Figure 1.17. Intimate association between CBM and the appended catalytic module determining enzyme substrate/product chain length. 28

Figure 1.18. Proposed model describing the effect of an additional loop in *Hv*Pho1 on glucan binding. 29

Chapter 2 – Unravelling the specificity of a GH94 laminaribiose phosphorylase from *Paenibacillus* sp. YM-1 towards donor substrates glucose/mannose 1-phosphate using X-ray crystallography and STD NMR spectroscopy

Figure 2.1. Structures of sugar 1-phosphates discussed in this chapter. 46
Figure 2.2. Reactions carried out by *Ps*LBP. 49
Figure 2.3. Expression and characterisation of the recombinant *Ps*LBP**.** 51
Figure 2.4. *Ps*LBP activity on Man1P and production of disaccharide **1**. 52
Figure 2.5. Synthesis and purification of disaccharide **1.** 53
Figure 2.6. ^{13}C NMR (100 MHz, D_2O, 25 °C) of disaccharide **1**. 54
Figure 2.7. ^{1}H NMR (800 MHz, D_2O, 25 °C) of disaccharide **1**. 55
Figure 2.8. COSY spectrum of disaccharide **1**. 56
Figure 2.9. HSQC data for disaccharide **1**. 56
Figure 2.10. 2D 1H coupled HSQC spectrum of disaccharide **1**. 57
Figure 2.11. Structural comparison between *Cg*CBP, *Rt*CDP and *Ps*LBP. 59
Figure 2.12. Comparison between the active sites of *Ps*LBP and *Cg*CBP. 62
Figure 2.13. Conservation of amino acids constituted -1 subsite in GH94 enzymes. 62
Figure 2.14. Structures of -1 subsites in *Ps*LBP and *Lp*SOGP. 65
Figure 2.15. Stereo views of the active sites of *Ps*LBP. 66
Figure 2.16. Comparison between loops in *Cg*CBP in complex with β-D-Glc, 1-deoxynojirimycin and phosphate (PDB code 3QG0) and Glc1P complex of *Ps*LBP. 68
Figure 2.17. STD NMR binding epitopes of ligands in the presence of *Ps*LBP. 70

Chapter 3 – Identification of *Euglena gracilis* β-(1→3)-glucan phosphorylase and establishment of a new glycoside hydrolase (GH) family GH149

Figure 3.1. Reaction carried out by β-(1→3)-D-glucan phosphorylase. 80
Figure 3.2. Partial purification of Euglena phosphorylase from the Euglena cell-free extract. 84
Figure 3.3. Glycomimetic probe for identification of DNJ-interacting phosphorylase candidates in Euglena protein lysate. 85
Figure 3.4. Expression and characterisation of insoluble recombinant EgP1 (IS-EgP1). 87
Figure 3.5. Expression and purification of soluble EgP1 recombinant protein (EgP1). 88
Figure 3.6. Characterisation of the soluble recombinant EgP1. 90

Figure 3.7. TLC analysis of the glycan synthesis carried out by EgP1. 91
Figure 3.8. Expression of the recombinant Pro_7066 in *E. coli* and purification of the soluble protein. 93
Figure 3.9. Characterisation of Pro_7066. 94
Figure 3.10. TLC analysis of the glycan synthesis carried out by Pro_7066. 95
Figure 3.11. Phylogenetic analysis of GH149. 97
Figure 3.12. Conservation of key amino acids in GH94 and GH149 sequences. 98
Figure 3.13. Structural comparison between *Rt*CDP, *Ps*LBP and Pro_7066. 100
Figure 3.14. Additional domains in Pro_7066. 101
Figure 3.15. The active site of Pro_7066 in comparison to that of *Ps*LBP. 102
Figure 3.16. Structural elements surrounding the active sites of *Rt*CDP, *Ps*LBP and Pro_7066. 104
Figure 3.17. The SBS site on the surface of Pro_7066 catalytic domain. 106
Figure 3.18. Structures of the residues in GH94 enzymes that superposed with E990, Y993 and E994 in Pro_7066 structure. 107
Figure 3.19. Multiple amino acid sequence alignment of Pro_7066 with Euglenophyceae GH149 and bacterial GH149 sequences. 108
Figure 3.20. G6 conformations in solution and at the Pro_7066 SBS. 110
Figure 3.21. TLC analysis of Pro_7066-catalysed glycan synthetic reactions in the presence of BCN. 112
Figure 3.22. MALDI-ToF analysis of the reaction carried out by Pro_7066 in the presence of G2 and either the presence or absence of inorganic phosphate. 113
Figure 3.23. EgP1 homology model based on Pro_7066 structure**.** 116
Figure 3.24. Potential SBS sites on EgP1 and other EgP candidates. 117
Figure 3.25. Genetic organisation of putative gene clusters containing the bacterial *GH149* genes. 125
Figure 3.26. Proposed mechanisms of β-(1→3)-glucan utilisation in bacteria containing GH149 gene clusters. 127

Chapter 4 – A new glycoside hydrolase (GH) family, GHyyy, comprising β-(1→3)-glucan phosphorylases with different acceptor preferences to GH94 and GH149

Figure 4.1. Phylogenetic analysis of OcP1 and its orthologues. 139
Figure 4.2. Relationship between GH94, GH149 and GHyyy**.** 140
Figure 4.3. Production of recombinant *Pap*P protein in *E. coli* and purification of the soluble protein. 142
Figure 4.4. TLC analysis of *Pap*P activity on Glc and disaccharides as acceptors. 143

Figure 4.5. Comparison between substrate chain length preferences of *Pap*P and GH149 enzyme. 145
Figure 4.6. Glycan synthetic reactions carried out by *Pap*P in the presence of β-linked disaccharide acceptors and Glc1P**.** 146
Figure 4.7. MALDI-ToF analyses of the glycan synthetic reactions catalysed by *Pap*P in the presence of β-linked disaccharide acceptors and Glc1P**.** 147
Figure 4.8. Glycan synthetic reactions carried out by Pro_7066 from GH149 family in the presence of β-linked disaccharide acceptors (10 mM) and Glc1P (10mM). 148
Figure 4.9. MALDI-ToF analyses of the glycan synthetic reactions catalysed by Pro_7066 in the presence of β-linked disaccharide acceptors and Glc1P**.** 148
Figure 4.10. Gene clusters containing *GHyyy* genes from *Paenibacillus polymyxa* and *Bacillus weihanensis*. 150
Figure 4.11. Genetic clusters containing genes encoding *GHyyy* and other glucosidases**.** 151
Figure 4.12. Genetic loci containing *GH94-GHyyy* genes in tandem and possible roles of the enzymes encoded by the genetic loci. 153
Figure 4.13. Multiple sequence alignment of GHyyy amino acid sequences against *Pt*TGS1 to identify the C-terminal sorting signals. 157
Figure 4.14. Summary of characterised GH94, GH149 and GHyyy activities. 158
Figure 4.15. A Possible role of *Ta*CDP from *Thermosipho africanus* TCF52B based on gene cluster analysis. 160

Chapter 5 – Summary and translational opportunities
Figure 5.1. The use of GPs as biocatalysts in one-pot enzymatic reactions for oligosaccharide production. 172
Figure 5.2. Schematic representation of a one-pot enzymatic reaction using sucrose phosphorylase (SP) and Pro_7066 from GH149 as biocatalysts and sucrose and inorganic phosphate as starting materials. 173
Figure 5.3. Analysis of oligosaccharide products obtained from the one-pot enzymatic reaction. 173
Figure 5.4. Activity of EgP1 and Pro_7066 on a noncognate donor Man1P. 175
Figure 5.5. Schematic representation of ultrahigh-throughput droplet system for metagenomic libraries screening. 180
Figure 5.6. Assays for GP activity that can be cooperated into the droplet screening platform. 180

Chapter 6 – Materials and methods

Figure 6.1. Crystallisation of *Ps*LBP. 195

Figure 6.2. Crystal trials and optimisation for Pro_7066. 196

Figure 6.3. Stereo view of the Pro_7066 active site. 198

Figure 6.4. *Pap*P crystal after optimisation. 199

List of Tables

Chapter 1 – Introduction

Table 1.1. CAZymes currently covered by CAZy database. 9

Table 1.2. Summary of CAZymes involved in β-(1→3)-glucan biosynthesis. 11

Table 1.3. Summary of reported CAZy families containing GPs with regio- and stereo-chemistry specific for β-(1→3)-linkage. 18

Chapter 2 – Unravelling the specificity of a GH94 laminaribiose phosphorylase from *Paenibacillus* sp. YM-1 towards donor substrates glucose/mannose 1-phosphate using X-ray crystallography and STD NMR spectroscopy

Table 2.1. Kinetic data of *Ps*LBP acting on Glc1P and Man1P in the presence of 10 mM Glc as an acceptor; for Glc acceptor, experiment was conducted in the presence of 10 mM Glc1P. 52

Table 2.2. X-ray data collection and processing *of Ps*LBP structures. 60

Table 2.3. Summary of conserved amino acid residues that are involved in GH94 catalysis, phosphate recognition or located in -1 subsite. 63

Chapter 3 – Identification of *Euglena gracilis* β-(1→3)-glucan phosphorylase and establishment of a new glycoside hydrolase (GH) family GH149

Table 3.1. Characterised GH94 sequences that were used as queries for BLAST analysis. 82

Table 3.2. Example of protein hits from Mascot search on the DNJ-eluted protein fraction. 86

Table 3.3. Summary of the observed $[M+Na]^+$ of glucans in MALDI-ToF analysis. 91

Table 3.4. Kinetic parameters of EgP1 for the glycan synthesis using 10 mM Glc1P as a donor. 92

Table 3.5. Kinetic parameters of Pro_7066 for the glycan synthesis using 10 mM Glc1P as a donor. 95

Table 3.6. Summary of dihedral angles from published PDB structures in complex with β-(1→3)-gluco-oligosaccharides. 111

Table 3.7. X-ray data collection and processing of Pro_7066 structures. 114

Table 3.8. Summary of k_{cat}/K_M values for the synthetic reaction carried out by GPs from GH94. 120

Chapter 4 – A new glycoside hydrolase (GH) family, GHyyy, comprising β-(1→3)- glucan phosphorylases with different acceptor preferences to GH94 and GH149

Table 4.1. BLASTP analyses of *Ochromonas* sp. transcriptomic sequences using 21 sequences of GH94 family and EgP1 of GH149 as queries**.** 138

Table 4.2. Summary of the estimated molecular masses of *Ps*LBP, Pro_7066, EgP1 and *Pap*P based on their elution volumes from gel filtration column 143

Table 4.3. Kinetic parameters of the glycan synthesis reaction catalysed by *Pap*P. 145

Table 4.4. Intron analysis of the *GHyyy* genes from heterokonts of which their genomes have been reported on GenBank. 156

Table 4.5. Reported catalytic efficiency [k_{cat}/K_m ($s^{-1}mM^{-1}$)] of *Ta*CDP. 161

Abbreviations

ABC	ATP-binding cassette
AIEX	Anion Exchange Chromatography
ATP	Adenosine triphosphate
BCN	Bicine
BLAST	Basic Local Alignment Search Tool
CAZymes	Carbohydrate-active enzymes
CBAP	Cellobionic acid phosphorylase
CBM	Carbohydrate binding module
CBP	Cellobiose phosphorylase
CDP	Cellodextrin phosphorylase
CM	Cell membrane
DNA	Deoxyribonucleic acid
DNJ	1-Deoxynojirimycin
DP	Degree of polymerisation
EG	Ethylene glycol
EgP1	*Euglena gracilis* phosphorylase 1
Gal1P	α-D-Galactose 1-phosphate
GDP	Guanosine diphosphate
GH	Glycoside hydrolase
Glc	Glucose
Glc1P	α-D-Glucose 1-phosphate
GlcA1P	α-D-Glucuronic acid 1-phosphate
GlcN1P	α-D-Glucosamine 1-phosphate
GlcNAc	*N*-Acetyl-glucosamine
GlcNAc1P	α-D-GlcNAc 1-phosphate
GLNBP	Galacto-*N*-biose/Lacto-*N*-biose phosphorylase
GNB	Galacto-*N*-biose
GP	Glycoside phosphorylase
GPC	Gel permeation chromatography
GSL	Glucan synthase-like
GT	Glycosyltransferase
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HPAEC-PAD	High-performance anion-exchange chromatography with pulsed amperometric detection
HRMS	Hight resolution mass spectrometry
IMAC	Immobilised metal affinity chromatography
IN	Inner membrane
IPTG	Isopropyl β-D-1-thiogalactopyranoside
IS-EgP1	Insoluble EgP1 protein
LB	Lysogeny broth
LBP	Laminaibiose phosphorylase
LC	Liquid chromatography
LDP	Laminaridextrin phosphorylase
LNB	Lacto-*N*-biose
LNBP	Lacto-*N*-biose phosphorylase
***Lp*SOGP**	β-(1→2)-oligoglucan phosphorylase from *Lachnoclostridium phytofermentans*
MALDI-ToF	matrix-assisted laser desorption/ionization-time of flight
Man1P	α-D-Mannose1-phosphate
MES	2-(*N*-morpholino)ethanesulfonic acid
MFS	Major facilitator superfamily
MS	Mass spectrometry

MW	Molecular weight
NB	*N*-butyldeoxynojirimycin
NCBI	National Centre for Biotechnology Information
NE	No enzyme control
NMR	Nuclear magnetic resonance
OcP1	Ochromonas phosphorylase 1
OM	Outer membrane
ORF	Open reading frame
P	Periplasmic space
***Pap*P**	*Paenibacillus polymyxa* phosphorylase
PCR	Polymerase chain reaction
PDB	Protein Data Bank
PEG	Polyethylene glycol
***Prym*P**	*Prymnesium parvum* phosphorylase
PUL	Polysaccharide utilisation loci
SBP	Substrate-binding protein
SBS	Surface binding site
SDS-PAGE	sodium dodecyl sulfate–polyacrylamide gel electrophoresis
STD	Saturation transfer difference
TBDT	Ton-B dependent transporter
TLC	Thin layer chromatography

Chapter 1

Introduction

1.1. Significance and diversity of carbohydrates

Carbohydrates are the most abundant macromolecules on earth and can be found across all kingdoms of life. Given the vast number of different glycan structures present in nature, they assume diverse biological roles that can be broadly categorised as structure- and recognition-related functions.[1] Polymeric glycans found in plant cell walls (cellulose and hemicellulose) are good examples of glycans that perform a structural role.[2] Recognition-related roles can be divided into extrinsic and intrinsic recognitions. Interaction between host and pathogenic microbes, such as recognition of sialylated glycophorins on human erythrocytes by *Plasmodium falciparum,* is an example of extrinsic recognition,[3] whilst intrinsic recognition is exemplified by the role of *N*-glycosylation in regulating the process of glycoprotein quality control.[4,5] Such broad classification of glycan roles only provides a simplistic view of the subject because, in some circumstance, given glycans can assume more than one function depending on their biological context.[6] For instance, heparan sulfate proteoglycans adopt a myriad of functions, such as forming organised extracellular matrix or as a co-receptor for growth factors.[7]

The structural basis of glycan diversity derives from the presence of variety of monosaccharides which can be linked in various different ways. The glycosidic linkage can be formed between monosaccharides in two possible stereoisomers (α- or β-anomer) depending on the configuration of the hydroxyl group at the anomeric carbon. Moreover, several possible regioisomers can be formed due to the presence of many hydroxy groups at different positions on monosaccharides. Modification, such as branching, can be added to the polysaccharide backbone to create more sophisticated primary structures. Polysaccharides can further form a more complex quaternary structure, utilising hydrogen bonding between hydroxy groups and ring stacking interaction between sugars. The difference in the stereo- and regio-chemistry of glycosidic linkages and in branching pattern among different glycan structures can have a profound effect on the formation of the glycan quaternary structures. For example, both amylose and amylopectin in starch contain an α-(1$\rightarrow$4)-glucan backbone, but amylopectin is periodically branched with α-(1$\rightarrow$4)-glucan chains interlinked with α-(1$\rightarrow$6)-glycosidic bonds (Figure 1.1A). If the correct degree and position of branching is achieved, amylose and amylopectin can further form crystalline structures that subsequently assemble into granules, (Figure 1.1B).[8] On the other hand, glucan chains containing β-(1$\rightarrow$4)-glycosydic bonds have a plane of rotation of 180° between adjacent

glucose (Glc) residues, which straightens the glucan chains (Figure 1.1C). The β-(1→4)-glucan chains subsequently form strong side-by-side interactions, making cellulose fibrils with great tensile strength that is suitable for its structural role in plant cell walls (Figure 1.1D). Structural complexity of natural glycans complicates the elucidation of their structures, the process of which is important for defining structure-function relationships.

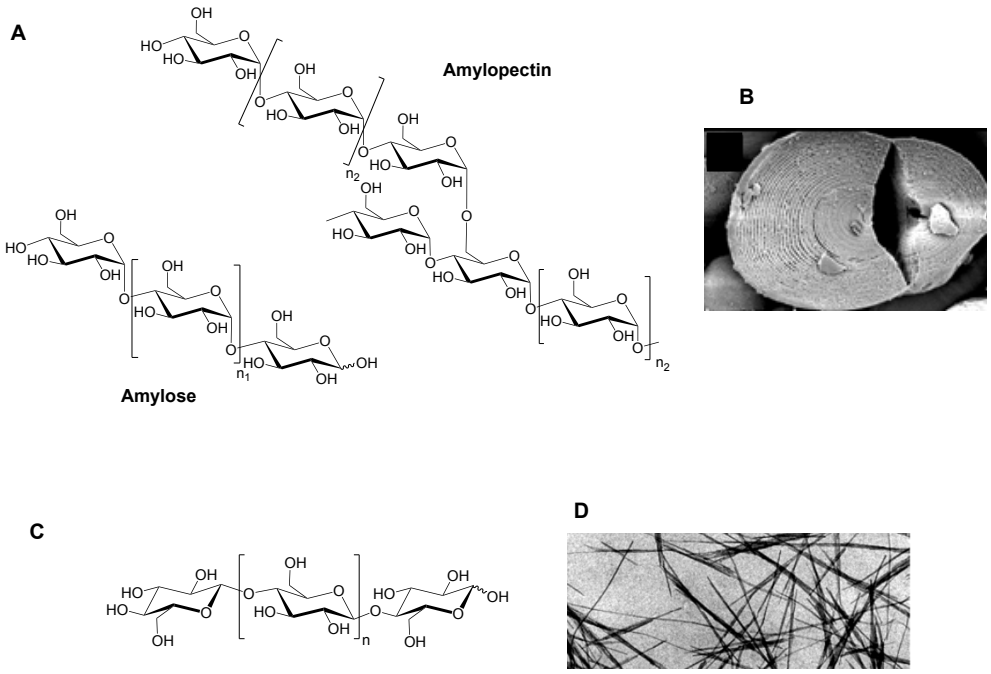


Figure 1.1. Diversity of glucan chemical structures and their morphologies. (**A**) Chemical structures of amylose and amylopectin. n_1 ~ 300-1000 and n_2 ~ 100[9]. (**B**) Morphology of starch granule. Figure reproduced from Pilling *et al.*[10] the material is copyright by the American Society of Plant Biologists and appeared with permission. (**C**) Chemical structure of cellulose. n ~1000-5000.[11] (**D**) Cellulose microfibrils. Figure reproduced from Nishiyama *et al*.[12]

1.2. β-(1→3)-Glucan structures and applications

Diverse β-(1→3)-glucan structures can be found in bacteria, fungi, plants and algae, from which β-(1→3)-glucans can be extracted for a wide range of biotechnological applications in agronomy, the food, cosmetic and therapeutics industries.[13] β-(1→3)-glucans have been used as food stabiliser and there is much interest in the application of β-(1→3)-glucans as functional dietary fibre,[14] presumably due to its reported immunomodulatory properties. Other biological impacts, such as anti-tumour properties, have been reported, although these are often described in correlation with the immunomodulatory effect of β-(1→3)-glucans.[15–18]

Linear β-(1→3)-glucan structures can be found in bacterial curdlan, plant callose and paramylon from microalga *Euglena* (Figure 1.2A). Curdlan is a linear β-(1→3)-glucan found deposited as capsular polysaccharide in *Agrobacterium*,[19] *Rhizobium* [20] and *Cellulomonas* spp.[21] The molecule contains as many as 12,000 Glc units and is insoluble in water (Figure 1.2B).[22] Callose is a linear polymer of β-(1→3)-linked glucan,[23] found deposited in the wall adjacent to the plasma membrane in response to wounding or other physiological and pathological stresses (Figure 1.2C).[24] *Euglena gracilis* produces a crystalline storage molecule named paramylon, a linear glucan chain consisting of only β-(1→3)-glycosidic linkages (Figure 1.2D).[25] Linear β-(1→3)-glucan are capable of forming triple helical structures that are stabilised by inter- and intra-strands hydrogen bonding (Figure 1.2E),[26,27] which can subsequently form high-order aggregates of crystalline structures, such that observed in curdlan and paramylon.[14]

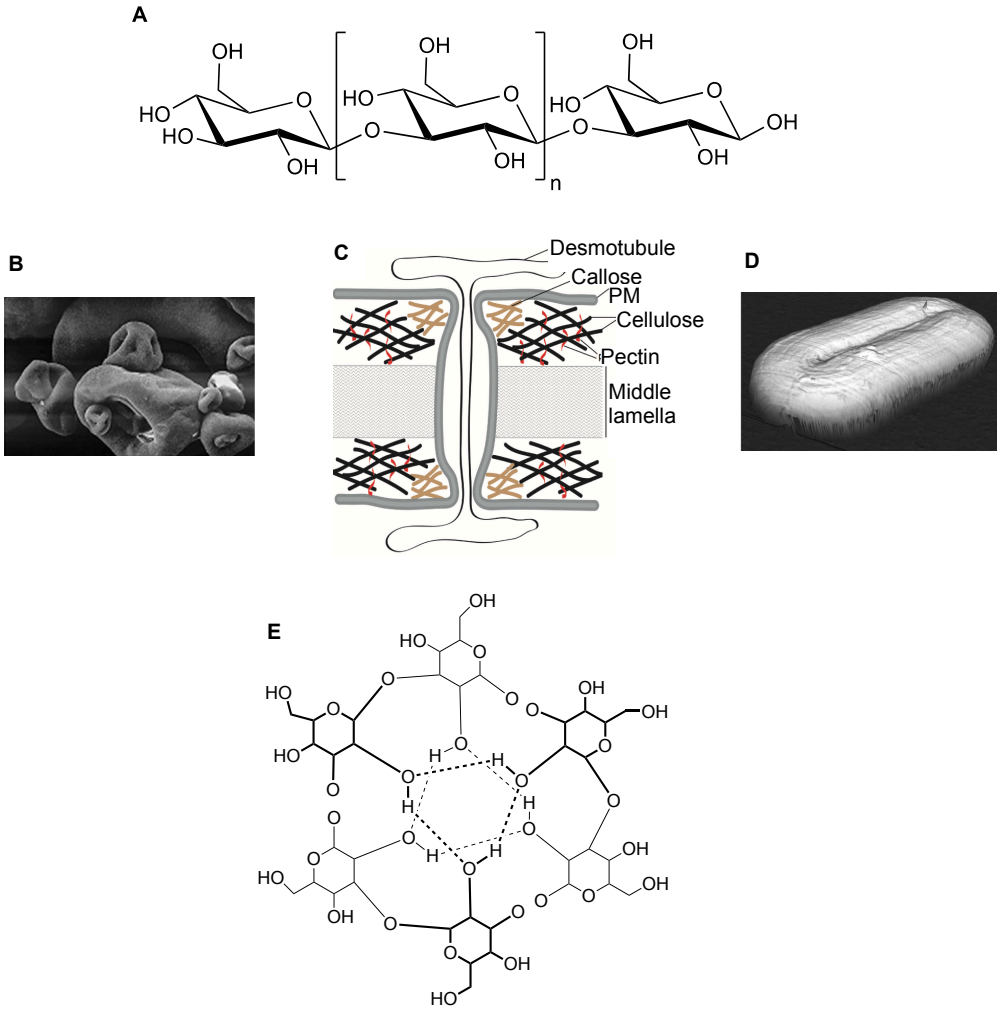


Figure 1.2. Structures and morphologies of linear β-(1→3)-glucans. (**A**) Chemical structure of linear β-(1→3)-glucan. (**B**) Scanning electron micrograph of curdlan. Figure reproduced from Mangolim *et al*.[28] (**C**) Cartoon representation of callose located near plant plasmodesmata. Figure reproduced from Knox *et al*.[29] (**D**) Scanning electron micrograph of paramylon. Figure reproduced from Barsanti *et al*.[14] (**E**) Intermolecular hydrogen bonds in triple helical structure of linear β-(1→3)-glucan chains.[30]

Albeit not containing solely β-(1→3)-linkage in its glucan backbone, it is worth mentioning linear mixed-linkage β-(1→3;1→4)-glucans found characteristically in the cell walls of grasses and cereals, which have roles as dietary fibre, immunomodulating and cholesterol lowering agents.[14] The majority of mixed-linkage β-(1→3;1→4)-glucans contain Glc residues that are joined by a single (1→3) linkage separated by two or three (1→4)-linked Glc residues (Figure 1.3A). However, the ratio of (1→3) and (1→4) linkages may vary among different sources. In grasses, mixed-linkage β-(1→3;1→4)-glucans are co-polymers of cellotriosyl and cellotetraosyl residues linked by single (1→3) linkages. The difference in the ratio of celltriosyl and cellotetraosyl units are found in different grass species and is reflected in their water solubility;[31] barley and oat glucans containing lower ratio of cellotriosyl to cellotetraosyl units are more

soluble in comparison to wheat glucans which has higher cellotriosyl:cellotetraosyl ratio.[32] The formation of stable, sheet-like complexes stabilised by hydrogen bonds between the hydroxy groups in cellotriosyl units probably accounts for the increase in water insolubility observed in β-(1→3;1→4)-glucans with higher ratios of cellotriosyl:cellotetraosyl units (Figure 1.3B).[33]

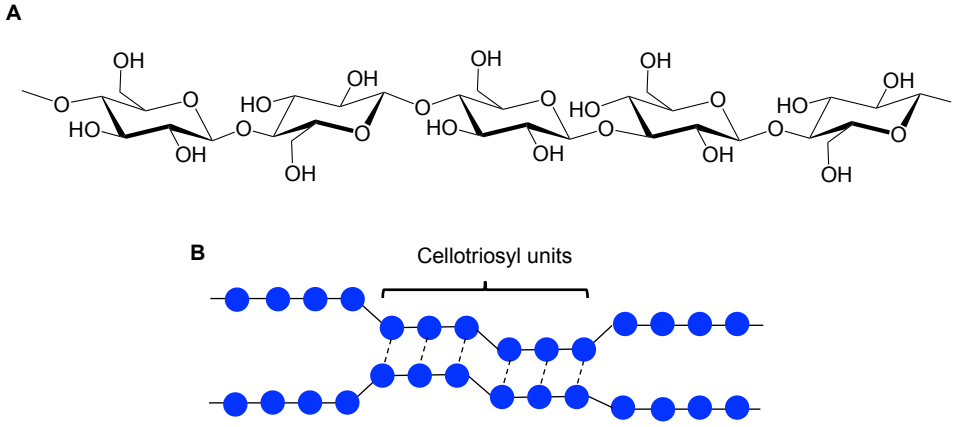


Figure 1.3. Structure of mixed-linkage β-(1→3;1→4)-glucans. (**A**) Chemical structure of general mixed-linkage β-(1→3;1→4)-glucan. (**B**) Proposed intermolecular interactions between sequential cellotriosyl units in mixed-linkage β-(1→3;1→4)-glucan chains.[33]

Branched β-(1→3)-glucan structures are found primarily in diatoms, algae and fungi. Chrysolaminarin is a soluble storage carbohydrate from photosynthetic diatoms, consisting of β-(1→3)-glucan with β-(1→6)-linked-glucosyl or short oligoglucosyl branches.[34–40] Laminarin structure in brown algae is rather similar to chrysolaminarin, with an average mass of 5,000 Da (degree of polymerisation (DP) range 31-42) and plays a role in energy storage.[41] However, two different types of laminarin have been described; 1) the M-series in which the reducing end is terminated with mannitol residue and 2) the G-series in which the reducing end is terminated with Glc (Figure 1.4A).[42] β-(1→3)-glucans in fungal cell walls has a wide variation in its degree of branching, depending on the species. In *Saccharomyces cerevisiae*, β-(1→3)-glucan has a branch-on-branch structure (Figure 1.4B) with β-(1→6)-interchain links, and an approximate molecular weight of 190-200 kDa.[43] β-(1→3)-glucan forms a core of the fungal cell wall and its non-reducing ends are crosslinked to the reducing end of chitin chains through β-(1→4)-linkages (Figure 1.4C).[14]

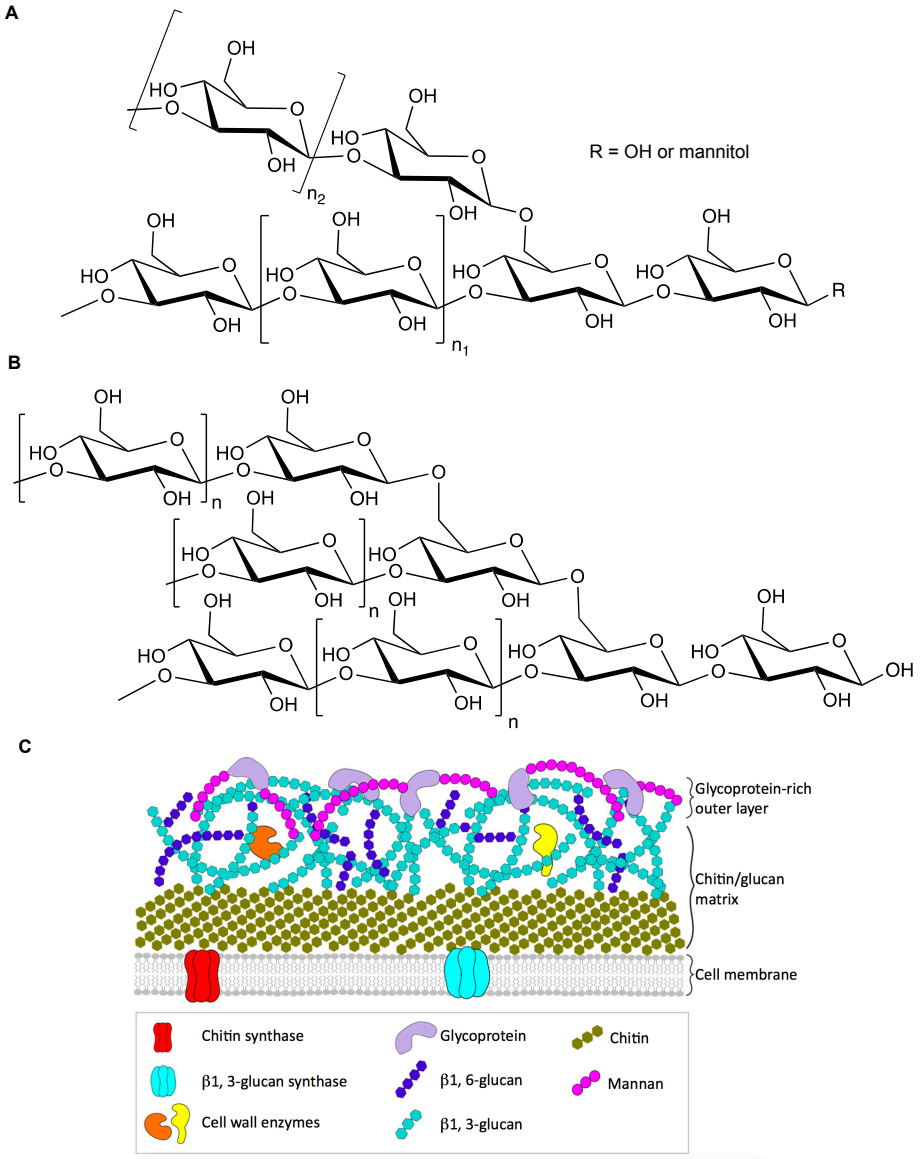


Figure 1.4. Chemical structures and morphologies of branched β-(1→3)-glucan. **(A)** linear β-(1→3)-glucan backbone with β-(1→6)-glucosyl branch found in diatoms (chrysolaminarin), and brown algae (laminarin). The non-reducing end is modified with mannitol (R = mannitol) in laminarin M-series or with Glc (R=OH) in laminarin G-series. n_1 and n_2 vary between different algal species. (**B**) Yeast branch-on-branch β-(1→3)-glucan structure, with n can reach up to 1500 units.[44] (**C**) General model of yeast cell wall structure. Each β-(1→3)-glucan chain is interspersed with either β-(1→6)-linked-β-(1→3)-branches or β-(1→6)-branches. Both β-(1→3)- and β-(1→6)-glucan branches can also be cross-linked to chitin. Figure **C** was reproduced from Geoghegan *et al.*[45] with permission.

Immunomodulatory effects of fungal β-(1→3)-glucan have been demonstrated *in vitro* and in animal and human experiments.[46] Laminarin from brown algae also possesses immunomodulatory properties.[16,17] Several reports indicate the immunomodulatory effect of paramylon from *Euglena gracilis* in humans[47] and fish.[48,49] Such immunomodulatory effects of β-(1→3)-glucan were mediated by its interaction with immune cell surface receptors, such as Dectin-1.[50–52] The β-(1→3)-glucan-receptor interaction activates signalling cascades that subsequently results in

stimulation of immune cells such as macrophage to release inflammatory factor that likely play a role in host defence against fungal invasion.[18,53] Due to the variety of β-(1→3)-glucan chain lengths and degree of β-(1→6)-branching in natural β-(1→3)-glucans, there are controversies surrounding the molecular principles underlying β-(1→3)-glucan-receptor interactions. For instance, some reports indicated the requirement of a high molecular weight, triple helical structure,[54,55] whereas others suggest that single helices are sufficient for the biological activity of β-(1→3)-glucan.[56]

Although several biological activities of β-(1→3)-glucan have been reported, there are remaining challenges that must be overcome to enable the marketing of β-(1→3)-glucan by therapeutics and food supplement industries. These challenges were summarised by Chan *et al.*:[57] 1) β-(1→3)-glucans used in published research came from extracts which contain mixture of different types of carbohydrates; 2) diverse structures of β-(1→3)-glucans from different sources and the lack of thorough quantitative and qualitative characterisation of the structures, resulted in unclear structure-function relationships; 3) the link between signalling pathways induced by β-(1→3)-glucan and subsequent immunomodulatory effects is unclear. To overcome these challenges, there is a need for defined β-(1→3)-glucan structures that can be obtained through chemical or enzymatic routes, which would subsequently enable an establishment of a clear structure-function relationship between β-(1→3)-glucan and its immunomodulatory effect. Chemical preparation of β-(1→3)-gluco-oligosaccharides can be performed through acetolysis of laminarin,[58] although it is difficult to control the extent of acetolysis to obtain products with a desired chain length. Chemical syntheses have also been used to prepare β-(1→3)-gluco-oligosaccharides,[59,60] although these approaches required further steps of protection/deprotection of the hydroxy groups on glycan to avoid anomeric mixtures and regioisomeric products. Production of β-(1→3)-glucan via enzymatic routes are considered a better alternative for β-(1→3)-glucan production due to their superior stereo and regio-specificity. The utilisation of β-(1→3)-glucan enzymatic synthesis is still challenging due to a limited availability of biocatalysts with desired activities and robust efficiency, an issue which is further complicated by a rather narrow range of operational pH and temperatures that have to be applied to ensure optimal conditions for the biocatalysts.

1.3. Identification and classification of carbohydrate-active enzymes

The diversity of glycan structures observed in nature implies that organisms possessing them must evolve a large number of carbohydrate-active enzymes (CAZymes) for synthesis, degradation and modification of the glycan structures. Indeed, a vast number of different classes of CAZymes with >150 different activities have been discovered, together with non-catalytic proteins that aid the binding of different types of carbohydrate structures (Table 1.1).

Table 1.1. CAZymes currently covered by CAZy database.[61]

Classes	Description	Numbers of reported families
Glycoside hydrolases (GHs)	Hydrolysis or rearrangement of glycosidic bonds	153
Glycosyltransferases (GTs)	Formation of glycosidic bonds by transferring sugar moieties from activated sugar molecules to specific acceptor molecules.	105
Polysaccharide lyases (PLs)	Non-hydrolytic cleavage of glycosidic bonds in uronic acid-containing polysaccharides	28
Carbohydrate esterases (CEs)	Hydrolysis of *O*- or *N*-acylation of substituted saccharides	16
Auxiliary Activities (AAs)	Redox enzymes acting in conjunction with other CAZymes	15
Carbohydrate-binding modules (CBMs)	Non-catalytic folds appended to CAZymes	83

The advent of the CAZy database[61,62] (http://www.cazy.org/) tremendously improved the discovery and classification of new CAZymes, auxiliary enzymes and non-catalytic glycan binding proteins. The basic principle of the classification involves grouping protein candidates with similar primary sequences and structural folds into the same family, regardless of their substrate specificity and activity.[63,64] More than 10,000 complete genomes from across four kingdoms (Bacteria, Viruses, Archaea and Eukaryota) have been analysed and annotated for CAZymes. Glycobiological capability of organisms can be assessed by analysing its genome and listing CAZyme candidates based on the sequence homology to known CAZy families. Using this approach, genomes from related organisms can be compared to determine how the number of CAZYmes have changed during the course of evolution. Families with related catalytic

residues and structural folds, but are distantly related by sequence similarity, are further grouped into clans.[63,65–67]

The methodology developed by CAZy has also been applied to metagenomes, which has enabled the study of genes from organisms that are difficult to culture in laboratory conditions, such as those derived from animal [68–70] and human guts, which increases the understanding of glycome beyond that of model organisms.[71,72] For instance, analysis of human gut metagenome revealed that porphyranases and agarases, enzymes involved in degradation of carbohydrate in seaweeds, may have been transferred from marine bacteria to digestive tracts of Japanese individuals, possibly due to the high consumption of seaweed in Japanese population.[73] Moreover, comparison of CAZyme numbers and expression levels in transcriptomes derived from a single species grown under different conditions enables correlation of CAZymes to biological processes such as host infection,[74–77] light/dark cycle[78] or anaerobic conditions.[79] In the context of novel enzyme discovery, CAZy annotations provide a massive repertoire of new enzyme candidates for biotechnological applications.

Despite several advantages of CAZy classification for their annotation of new CAZymes, this approach cannot be used for functional prediction and family classification of sequence candidates with very low or no sequence homology to existing proteins with precedented functions. Also, in some cases where the sequence similarity criteria suggest classification of a sequence into a particular family, the enzyme activity might be very distinct from the other family members. For example, novel GP activities were discovered in GH3, the family which previously known exclusively for GH activities.[80,81] Moreover, the existing categories in CAZy require regular expansion to accommodate new activities that are distinct from the existing classes. This has been demonstrated for lytic polysaccharide monooxygenases, originally classified into GH families, but the unique requirement for metal co-factors validate their reclassification into a new system of Auxiliary Activity (AA) families.[82] For example, GH61 has been reclassified as AA9 based on its copper-dependent lytic polysaccharide monooxygenase activity, which perform oxidative cleavage of the glycosidic bond rather than hydrolysis.[83,84] These examples emphasise the importance of experimental validation of the enzyme activity in addition to the primary sequence classification.

1.4. Biosynthesis of β-(1→3)-glucans

Understanding the enzymes that are involved in β-(1→3)-glucan biochemistry in nature could provide potential enzymatic tools for synthesis and manipulation of β-(1→3)-glucan structures. Biosynthetic enzymes involved in the synthesis of natural β-(1→3)-glucan structures are summarised in Table 1.2 and the evidence for their functions are discussed herein.

Table 1.2. Summary of CAZymes involved in β-(1→3)-glucan biosynthesis.

β-(1→3)-glucans	GT involved in β-(1→3)-glucan biosynthesis	GH involved in β-(1→6)-branching
Curdlan	GT2[19,85]	-
Paramylon	GT2,[78]* GT48[86]	-
Callose	GT48[87]	-
Chrysolaminarin	GT48[88,89]	GH16[90]
Laminarin	GT48[91]*	GH16[91]*
Fungal cell wall	GT48[92,93]	GH16[94] and/or GH17[95]

*Gene candidates identified from genomes without experimental evidence for functions.

1.4.1. Bacterial curdlan

A genetic locus encoding curdlan biosynthetic enzymes was identified in *Agrobacterium* sp. ATCC 31749, which consists of *CrdA*, *CrdS* and *CrdC*.[19] *CrdS* was proposed to be a gene encoding the glucosyltransferase (GT) named curdlan synthase based on its shared translated amino acid sequence homology to enzymes in GT2 family.[31] Mutagenesis of *CrdS* resulted in non-staining colonies on aniline blue agar, which is indicative of abolished curdlan production.[19,85]

1.4.2 Fungal β-(1→3)-glucan

The biosynthesis of fungal β-(1→3)-glucans have been comprehensively reviewed.[96] In brief, the fungal β-(1→3)-glucan backbone was proposed to be synthesised by a glucan synthase using UDP-Glc as a substrate, the activity of which was first purified from *Saccharomyces cerevisiae*.[92] Two homologous genes from *S. cerevisiae*, FK506-sensitivity 1 and 2 (*FSK1* and *FSK2)*, were proposed to encode the glucan synthase activity belonging to GT48 family. Disruption of the *FSK1* gene resulted in a decrease in level of glucan in the cells and reduced cell growth,[92,93] whilst deletion of *FSK2* had negligible effect on the glycan synthase activity or the glucan level in the

cell.[92,97] Nevertheless, double deletion of *FSK1* and *FSK2* was lethal, suggesting their importance to yeast. The activity of enzymes that generate β-(1→6)-branches in *S. cerevisiae* was predicted to be from the GH16 family,[94] but a more recent study has identified a transglycosylase responsible for the branching activity belonging to the GH17 family in *Aspergilus fumigatus*.[95]

1.4.3 Callose

Callose is synthesised by a callose synthase, the activity of which was first detected in an enriched fraction from cotton fibres and displayed sequence similarity to that of the catalytic subunit found in the fungal FKS proteins.[87] The callose synthase and its homologues in other plants have been denoted Glucan Synthase-Like (GSL) genes and were classified into GT48. The callose synthase is located in the plasma membrane and utilised UDP-Glc as a substrate for the synthesis. Multiple *GSL* genes found in a single plant species are less closely related, whilst the *GSL* genes from different species are more similar to each other, suggesting that the *GSL* genes from different species may be functional orthologues encoding callose synthases that produce callose in the same tissue or developmental stage.[98]

1.4.4. Chrysolaminarin from diatoms and laminarin from brown algae

Chrysolaminarin metabolic pathways have been postulated based on the presence of prospective gene candidates encoding enzymes involved in the pathways. A chrysolaminarin pathway has been proposed for a diatom model *Phaeodactylum tricornutum* based on its genome (Figure 1.5A).[99] Knockdown experiments were used to confirm the function of chrysolaminarin synthase (GT48) in chrysolaminarin production.[88] In addition, the expression level of GT48 was shown to be correlated with the light-dark cycle, in which maximal expression was reached at the beginning of the light period and minimum reached at the end, supporting a role of GT48 in chrysolaminarin synthesis during photosynthesis.[100] The role of β-(1→6)-branching enzymes was demonstrated by genetic complementation in yeast, where the expression of *P. tricornutum* genes encoding β-(1→6)-branching enzyme was able to rescue growth deficiency in yeast that lack the respective functional homologues.[90]

In another diatom model, *Thalassiosira pseudonana*, gene candidates involved in chrysolaminarin metabolism were also identified in the diatom genome (Figure 1.5B).[101] The expression levels of candidates responsible for chrysolaminarin

synthesis were highly upregulated during chrysolaminarin accumulation, whilst the transcript level of β-(1→3)-glucanases were upregulated during chrysolaminarin breakdown.[89] Moreover, a GT48 chrysolaminarin synthase candidate was shown to be essential for chrysolaminarin accumulation by knockdown experiments.[89]

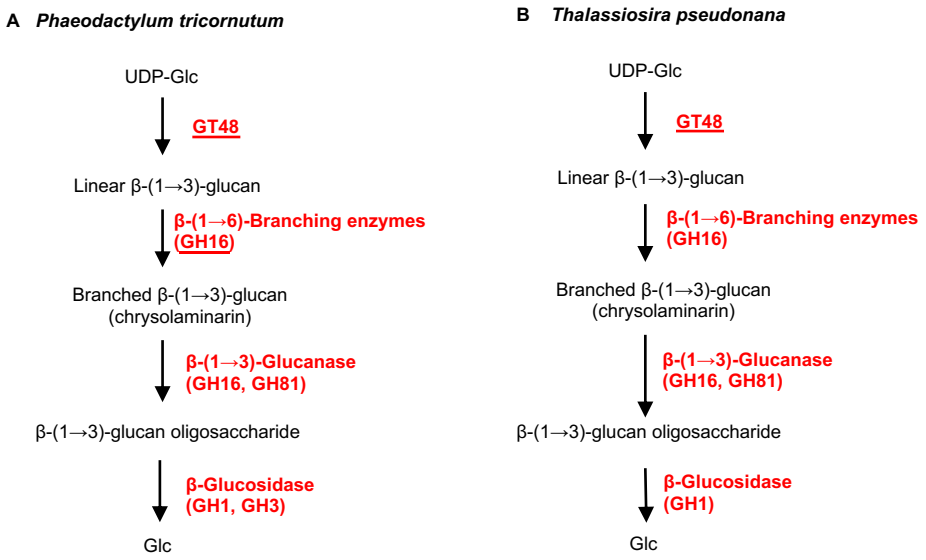


Figure 1.5. Summary of proposed chrysolaminarin metabolic pathways in diatoms based on bioinformatics analysis of the genomes. (**A**) Chrysolaminarin metabolism in *P. tricornutum*.[99] (**B**) Chrysolaminarin metabolism in *T. peudonana*.[89,101] Enzyme candidates are coloured in red. The underlined enzyme activities have been demonstrated by either genetic complementation in yeast (GH16)[90] or gene knockdown experiments (GT48 [88,89]).

Similarly, genes encoding candidate enzymes involving in laminarin metabolism have been identified in the genome of a brown alga *Ectocarpus siliculosus* (Figure 1.6),[91,102] but no experimental validation has been performed to confirm the roles of these candidates thus far. Moreover, the enzyme responsible for the addition of mannitol to the structure of laminarin M-series remains unknown.

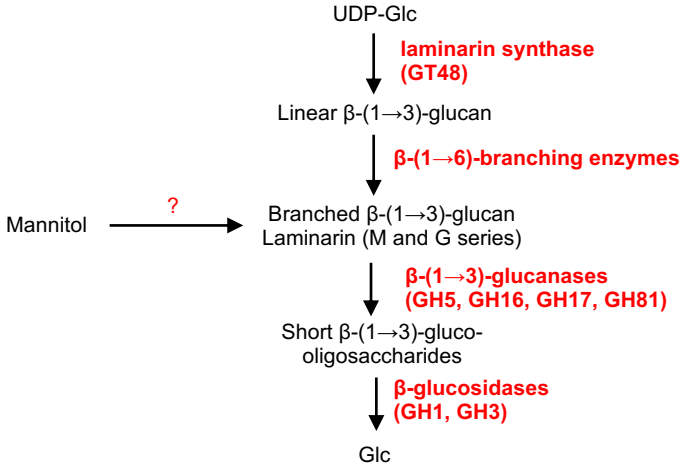


Figure 1.6. Laminarin metabolism in *E. siliculosus*. Laminarin metabolic pathway was predicted from the bioinformatics analysis of the algal genome.[91,102] Enzyme candidates are coloured in red.

1.4.5. Paramylon from *Euglena gracilis*

Paramylon synthesis is carried out by paramylon synthase, the activity of which has been detected in *E. gracilis* and required UDP-Glc as a substrate.[103,104] Paramylon synthase activity has been detected as part of a large complex with approximate MW of 670 kDa.[105] Transcriptome analyses of *E. gracilis* identified several CAZyme candidates that may be involved in the synthesis and degradation of paramylon (Figure 1.7).[78,79] Analysis of the *E. gracilis* transcriptomes identified glucan synthase-like 1 and 2 (*Eg*GSL1 and 2) belonging GT48 family.[79] Knockdown of *Eg*GSL2 inhibited accumulation of paramylon in *E. gracilis*, indicating a role of *Eg*GSL2 in paramylon synthesis.[86] A GH17 sequence candidate from *E. gracilis* was cloned and expressed in yeast to obtain recombinant protein that displayed endo-β-(1→3)-glucanase activity that is likely involved in paramylon degradation.[106] The activity of β-(1→3)-glucan phosphorylases, so called laminaribiose[107,108] or laminaridextrin phosphorylase,[109] have also been detected in *E. gracilis.* and their roles have been postulated for the degradation of linear β-(1→3)-linked gluco-oligosaccharides. However, the candidate nucleotide sequences encoding the phosphorylase activities were not identified in transcriptome analyses.[78,79]

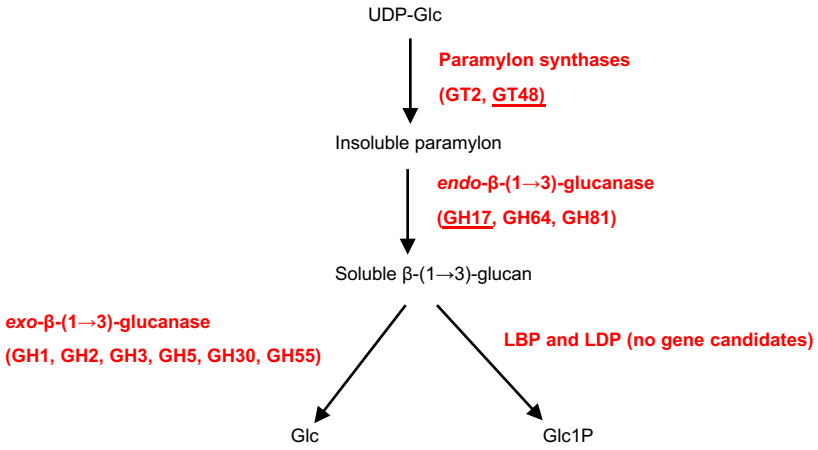


Figure 1.7. Paramylon metabolism in *Euglena gracilis* based on bioinformatics prediction from the transcriptomic data.[78,79] Enzyme candidates are coloured in red. The underlined enzyme activities have been demonstrated by *in vitro* characterisation of the recombinant protein (GH17)[106] or by gene knockdown (GT48).[86] LBP = laminaribiose phosphorylase. LDP = laminaridextrin phosphorylase.

1.5. *In vitro* enzymatic syntheses of β-(1→3)-glucans

Enzymatic synthesis is an attractive approach for oligosaccharide production due to its stereospecificity, efficiency and its reduced environmental impact, making it a 'green' choice compared to conventional chemical synthesis. Several enzymatic tools are currently available as biocatalysts for the production of linear β-(1→3)-glucan.

1.5.1. Transglycosylases

Transglycosylases are GHs with an inherent ability to catalyse the transfer of glucosyl moiety onto acceptor sugars, creating oligosaccharide products.[110] Transglycosylases utilise the same mechanism as retaining GHs,[111] which proceeds through the creation of glucosyl-enzyme intermediate, followed by transglycosylation step when appropriate sugar acceptors are present (Figure 1.8A). Several transglycosylases acting on β-(1→3)-glucan have been identified such as a GH16 β-transglycosylase from *Paecilomyces thermophila*[112] and a GH17 transglycosylase from *Euglena gracilis.*[106] These transglycosylases are promising biocatalysts because they could produce β-(1→3)-gluco-oligosaccharide with DP range 7-16.

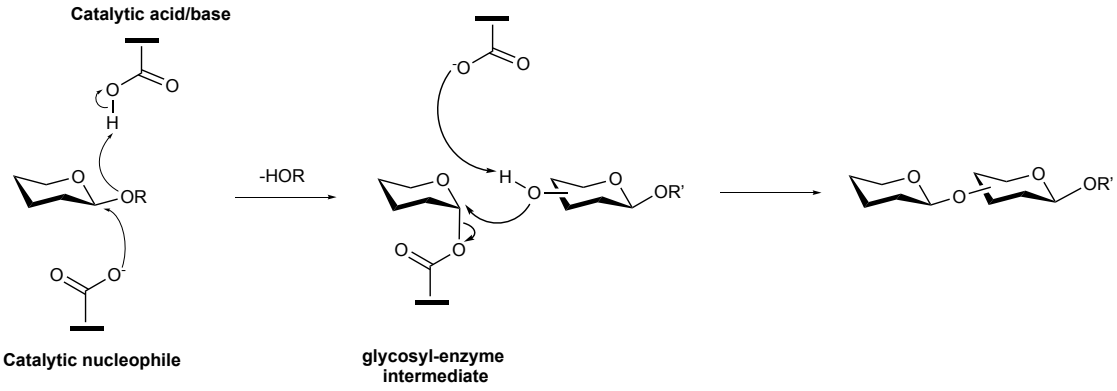


Figure 1.8. Transglycosylation mechanism.

1.5.2. Glycosynthases

Glycosynthases are enzymes that have their catalytic nucleophile substituted by a non-nucleophilic residue, enabling the modified enzymes to catalyse the transglycosylation of activated glucosyl fluoride donor with opposite anomeric configuration compared to the substrate of the wild type enzyme (Figure 1.9A).[113,114] A GH17 β-(1→3)-endoglucanase from barley was converted to a glycosynthase by mutation of E231 to glycine, and used as a biocatalyst for the *in vitro* synthesis of β-

(1→3)-glucan with DP range of 28-44 according to MALDI-ToF analysis.[115,116] The product formed crystalline material and the X-ray diffraction pattern indicated the presence of a triple-helical crystalline structure resembling those seen in anhydrous paramylon (Figure 1.9B). These helices spanned 8 nm of the crystal, with an estimated DP of 28, in agreement with the average DP of the synthetic product (Figure 1.9C).

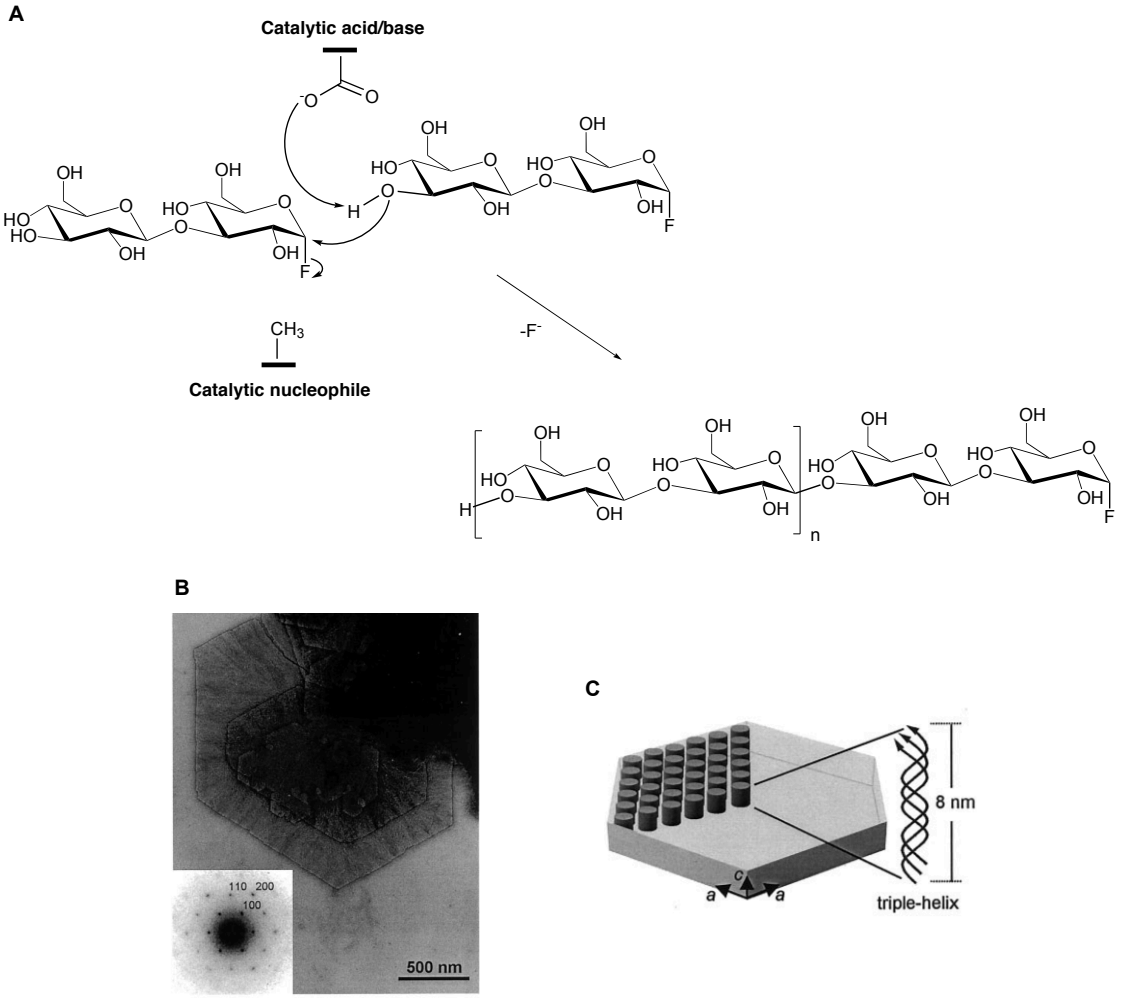


Figure 1.9. Glycosynthase mechanism and its application in β-(1→3)-glucan synthesis. (**A**) Mechanism of the barley glycosynthase. (**B**) Transmission electron micrograph of crystalline β-(1→3)-glucan synthesised by the barley glycosynthase. The diffraction pattern of the crystalline material is present in the inlet. (**C**) Proposed arrangement of triple helices in the hexagonal crystalline material in **B**. Figure **B** and **C** were reproduced from Hrmova *et al.*, 2002[115] © the American Society for Biochemistry and Molecular Biology.

1.5.3. Glycosyltransferases (GTs)

Investigation of β-(1→3)-glucan biosynthetic pathways provide numerous GT candidates that could be used as biocatalysts for the production of β-(1→3)-glucan, especially those in the CAZy family GT2 and 48, because several enzymes in these families are responsible for biosynthesis of natural β-(1→3)-glucans. *In vitro* synthesis

of β-(1→3)-glucan has been demonstrated using partial purified microsomal membrane containing callose synthases from plant extracts as biocatalysts.[117,118] A challenge of using GTs is the requirement for sophisticated membrane protein preparation to maintain the enzyme stability and activity. The challenge of protein preparation has been somewhat overcome by utilising cell-free expression to produce a catalytically-competent β-(1→3)-glucan synthase in a membrane-mimicking environment.[119]

1.5.4. Glycoside phosphorylases (GPs)

GPs catalyse a reversible transfer of sugar moiety from sugar 1-phosphate onto a broad range of acceptors (Figure 1.10). The sugar 1-phosphate donor of GPs is relatively cheap compared to that of nucleotide sugar required for GTs, therefore making GPs attractive as biocatalysts for β-(1→3)-glucan production.

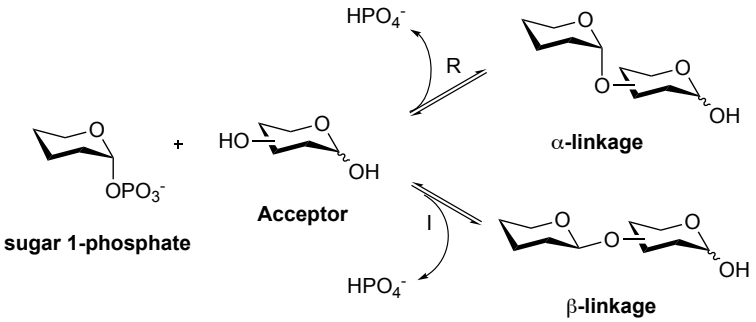


Figure 1.10. GP mechanisms.[120] R =Retaining phosphorylase, I = inverting phosphorylase.

There are 8 reported CAZy families containing GPs, only two families (GH94 and GH112) contain inverting GPs with stereo- and regio-chemistry that is specific for β-(1→3)-linked disaccharide or oligosaccharide (Table 1.3).

Table 1.3. Summary of reported CAZy families containing enzymes with regio- and stereo-chemistry specific for β-(1→3)-linkage. The microalgal enzymes of unknown families are also included.

Family	Linkages	Donor	Acceptor	EC number	References
GH94	β-(1→2)	α-D-Glc1P	D-Glc	2.4.1.333	[121]
	β-(1→3)	α-D-Glc1P	D-Glc	2.4.1.31	[122,123]
	β-(1→4)	α-D-Glc1P	D-Glc/β-D-Glc-(1→4)-β-D-$(Glc)_n$	2.4.1.20/ 2.4.1.49	[124]/[125]
	β-(1→4)	α-D-GlcNAc1P	D-GlcNAc	2.4.1.-	[126]
	β-(1→4)	α-D-Glc1P	D-gluconate	2.4.1.321	[127]
GH112	β-(1→3)	α-D-Gal1P	D-GlcNAc/ D-GalNAc	2.4.1.211	[128]
	β-(1→4)	α-D-Gal1P	L-Rha	2.4.1.247	[129]
Unknown	β-(1→3)	α-D-Glc1P	D-Glc//β-D-Glc-(1→3)-β-D-$(Glc)_n$	2.4.1.30	[107–109]
Unknown	β-(1→3)	α-D-Glc1P	β-D-Glc-(1→3)-β-D-$(Glc)_n$	2.4.1.97	[130]

The GH94 family contains bacterial enzymes operating on β-D-glucopyranosyl-(1→3)-glycopyranose (laminaribiose), but unable to produce longer chain of β-(1→3)-gluco-oligosaccharide. In order to make longer chain of β-(1→3)-gluco-oligosaccharide, the *E. gracilis* β-(1→3)-D-glucan phosphorylases could be used as biocatalyst for the synthesis, but the nucleotide sequences encoding these activities were not known, which prevents further expression and purification of the recombinant proteins for their use as biocatalysts. Nevertheless, semi-purified protein extract containing β-(1→3)-D-glucan phosphorylase activity from *E. gracilis* was used as a source of enzyme to catalyse synthesis of crystalline β-(1→3)-glucan with an average DP of 30,[131] the material of which has similar DP and crystalline morphology to that of the synthetic products from the reaction catalysed by barley glycosynthase. However, the preparation of the semi-purified protein fraction from Euglena culture is time-consuming and laborious.

1.6. Understanding chain length control mechanisms in carbohydrate-active enzymes

Unlike DNA, RNA and proteins, synthesis of carbohydrate is not templated, meaning that the information that is required to precisely control the synthesis is not directly encoded in DNA. Therefore, carbohydrate synthesis relies solely on the controls of the biosynthetic enzymes as well as the degradation of the polymer to control the chain length of the final oligosaccharide products.

Chain length or degree of polymerisation (DP) is one of the contributing factors dictating the physical/chemical/biological activity of β-(1→3)-glucans. For instance, linear β-(1→3)-glucan with DP > 36 is insoluble in water,[132] and interaction of β-(1→3)-glucan with Dectin-1 receptor required β-(1→3)-glucan with DP ≥ 10.[133] Although the precise mechanisms of how polysaccharide chain lengths are controlled by biosynthetic enzymes are still under ongoing investigation, several studies on the subject have led to the proposition of mechanisms that are used by some CAZymes to control and/or recognise oligosaccharide/polysaccharide with different DPs, which are discussed in details below.

1.6.1. Chain length control mechanism relying on the presence of sugar binding subsites

Sugar binding subsites play an important role in determination of substrate chain length that a particular GH can accommodate. The sugar binding subsites are defined based on -*n*, +*n* nomenclature, which labels the sugar binding subsites as -*n* to +*n,* with -*n* and +*n* represent the non-reducing and the reducing end, respectively and the cleavage of glycosidic bond happens between -1 and +1 subsites (Figure 1.11).[134]

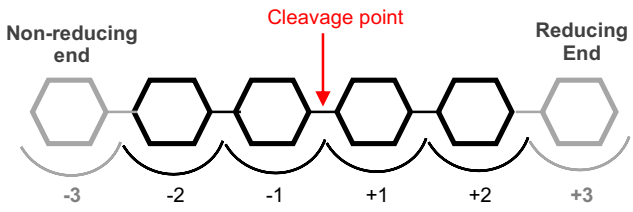


Figure 1.11. Nomenclature for sugar binding subsite.[134]

The numbers of sugar binding subsite can be determined through structural studies of the enzymes in complex with their oligosaccharide ligands and seem to be

important in determining the chain length of cycloglucans produced by transglycosylases. For instance, structural studies of the enzyme involved in production of cycloamylose and cyclodextrin identified sugar binding subsites and possible chain length control mechanism utilised by the enzymes. Cycloamylose is a cyclic α-(1→4)-linked glucan with dozens to hundreds of Glc residues, which are synthesized by 4-α-glucanotransferases (EC 2.4.1.25). A structural study of a 4-α-glucanotransferase belonging to GH77 family from *Thermus aquaticus* in complex with 34-meric cycloamylose led to the identification of 17 sugar binding subsites, and revealed substrate recognition and cycloamylose ring-size determination by the enzyme.[135] Another example is the study of cyclodextrins, cyclic α-(1→4)-linked oligosaccharides consisting of either six, seven or eight Glc residues,[136] which are produced from starch through an intramolecular transglycosylation catalysed by cyclodextrin glycosyltransferases (CGTases) (EC 2.4.1.19) belonging to GH13 family. Structural analysis of CGTase from *Bacillus circulans* strain 251 revealed +2 to -7 sugar binding subsites,[136] with subsite -3 and -7 being important for the chain length specificity of the enzyme,[137–139] and subsite -6 is importance for selective binding of the oligosaccharide with minimal DP of 6.[140]

1.6.2. Chain length control by terminating enzymes

Cyclic β-(1→2)-glucan containing 17-25 Glc residues are osmoregulated periplasmic glucan in *Agrobacterium*,[141] *Rhizobium*,[142] and *Brucella* species.[143] Cyclic β-(1→2)-glucan synthase (CgS) belonging to GT84 family is responsible for initiation, elongation and termination of cyclic β-(1→2)-glucan. The DP of cyclic β-(1→2)-glucan was shown to be dependent on the integrity of C-terminal of the Cgs protein since truncation of the C-terminus led to the production of β-(1→2)-glucan chain with longer DP than that produced by the full length CgS. Bioinformatic analysis of the CgS C-terminal domain identified a cyclic β-(1→2)-glucan phosphorylase, which was shown to catalyse phosphorolysis of the linear glucan chain at the non-reducing end. Therefore, the phosphorylase activity was proposed to be involved in the chain length regulation (Figure 1.12).[144] Although the authors proposed that the cyclisation step was carried out by the CgS, it is reasonable to speculate that the phosphorylase could also perform this reaction.

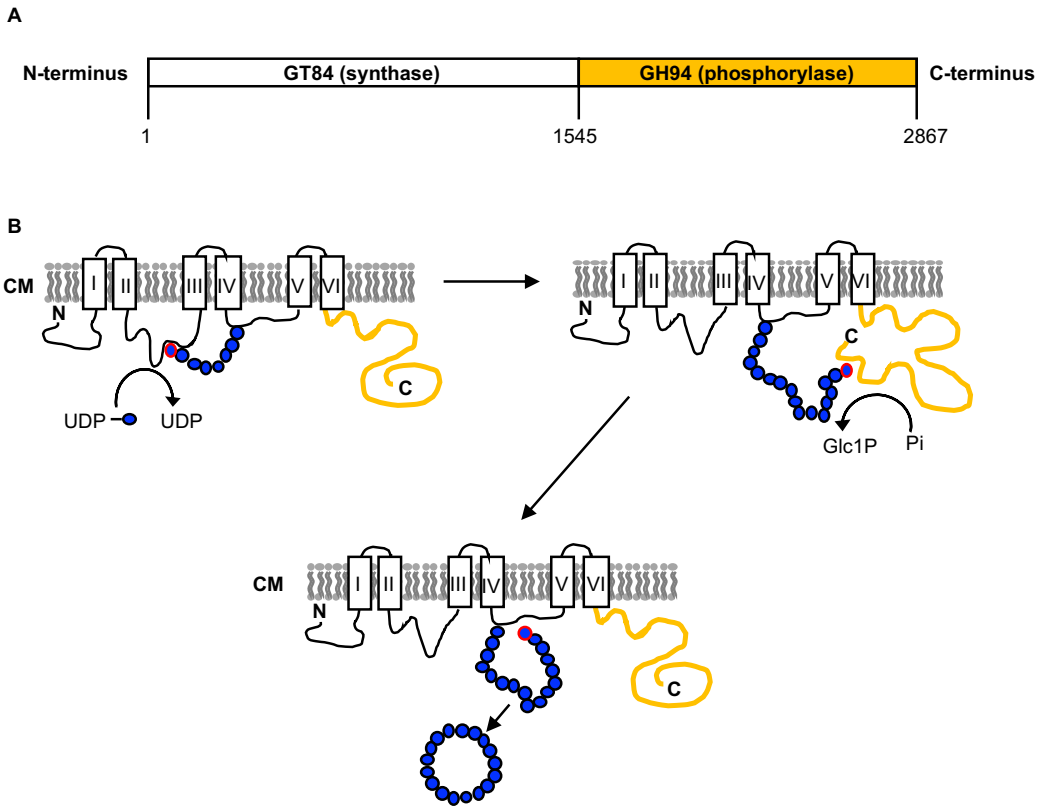


Figure 1.12. Proposed mechanism for chain length control in cyclic β-(1→2)-glucan production.[144] **(A)** Representation of the CgS sequence containing the synthase (white box) and the phosphorylase (yellow box) activities. (**B**) Proposed polymerisation and termination mechanism employed by CgS to produce cyclic β-(1→2)-glucan. N = N-terminus, C = C-terminus, CM = cell membrane. Glc is represented by blue circles. The non-reducing end Glc is indicated by red outline.

Another example for the use of terminating enzyme to control polysaccharide chain length is in the production of *O*-antigen, which is a part of bacterial lipopolysaccharide that is essential for bacterial cell viability[145] and pathogenicity.[146,147] The most detailed study of the *O*-antigen synthesis was performed on *E. coli* O9a antigen (Figure 1.13A). Two hypotheses have been proposed for the regulation of chain length of the repeat domain in O9a structure; (1) variable geometry model and (2) molecular ruler. The variable geometry model proposed that the modal length of O9a relies on the stoichiometry of WbdA-WbdD complex assembles, which in turns depend on the relative enzyme concentrations (Figure 1.13B).[148] In contrast, the molecular ruler model postulates measurement of O9a product against a physical structure in the biosynthetic enzymes. This model was supported by the presence of an extended coiled-coil helical domain in WbdD as the molecular ruler, which effectively controls the length of the O9a product (Figure 1.13C).[145] A similar model has been proposed for O12 antigen in *Raoultella terrigena*, although polymerisation, chain length control and termination were proposed to be regulated by enzyme activities from a single polypeptide rather than separate enzymes.[149] The concept of “molecular ruler” helix

has also been shown in a retaining GT from *Campylobacter jejuni*, which is involved in the synthesis of lipid-linked oligosaccharide as the initial step of *N*-glycosylation.[150]

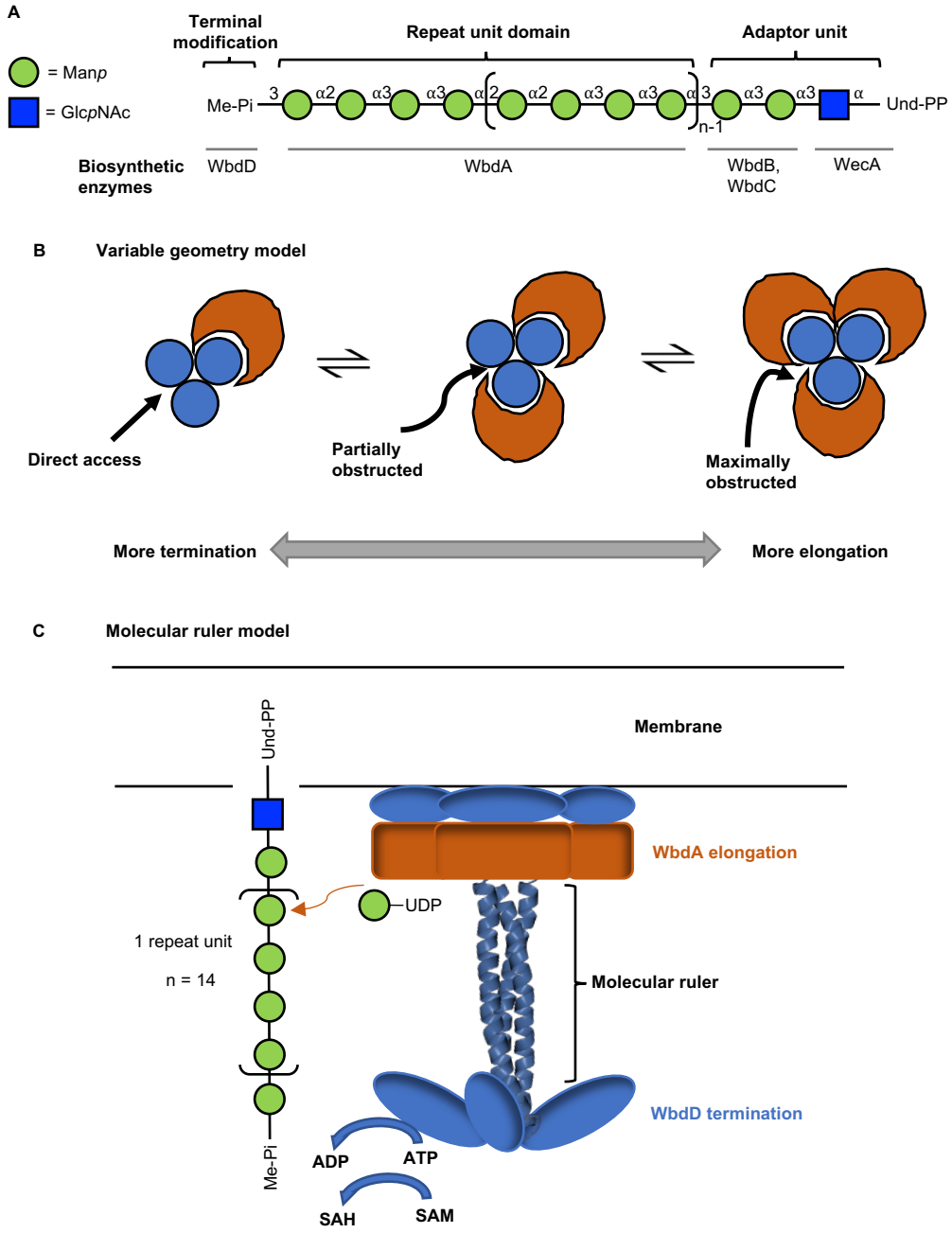


Figure 1.13. Length control mechanism for core lipid A structure of O9a in *E. coli*. (**A**) O9a polysaccharide structure. Biosynthetic enzymes responsible for each unit in the O9a are indicated at the bottom. (**B**) Variable geometry model.[148] The accessibility to the WbdD terminating complex (black arrows) depends on the number of WbdA subunits interacting with WbdD complex. WbdA complex is coloured in orange, and WbdD in blue. (**C**) Molecular ruler model.[145] The coiled-coli domain of WbdD acts as a molecular ruler that determines the length of the repeat unit on O9a. In figure B and C, WbdA and WbdD are coloured orange and blue respectively. SAH = *S*-adenosyl homocysteine, SAM = *S*-adenosyl methionine.

1.6.3. Chain length control by regulating the oligosaccharide binding pocket size

In addition to the number of sugar binding subsites, structural elements neighbouring the enzyme active sites can play important roles in defining the pocket size for oligosaccharide binding. The structural elements defining different chain length preferences have been studied extensively in the GH130 family which contain two sub-families with different substrate preferences; 1) GH130_1 exhibits a very narrow specificity toward β-D-mannopyranosyl-(1→4)-D-glucopyranose disaccharide;[151,152] 2) GH130_2 acting on longer and more complex β-linked manno-oligosaccharides;[152,153] Comparison between the structures of GH130_1 and GH130_2 enzymes have been performed to investigate the structural elements that distinguish the enzyme substrate preferences. The structural analysis showed that a long loop (>50 residues long) in GH130_1 covers the active site and limits the pocket size for acceptor binding, whilst this loop is only 3-residue long in GH130_2, thus enlarging the active site pocket for longer substrate (Figure 1.14).[154,155]

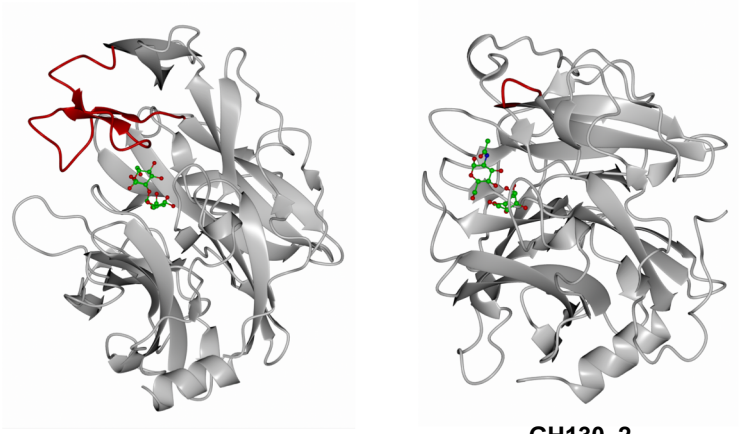


Figure 1.14. Comparison between a GH130_1 β-(1→4)-mannobiose phosphorylase structure from *Listeria innocua (*PDB code*:* 5B0R*)* and a GH130_2 β-(1→4)-mannopyranosyl-[*N*-glycan] phosphorylase (PDB code: 4UDK). The loops that define the pocket size are coloured in red.

Another possible way in which the pocket size can be defined was explained by the presence of an extra α/β N-terminal domain in a GH94 cellodextrin phosphorylase (β-(1→4)-glucan phosphorylase) from *Ruminiclostridium thermocellum* (*Rt*CDP), which is absent in cellobiose phosphorylase (Figure 1.15A). A role of the extended N-terminal domain in *Rt*CDP was proposed to be involved in the interaction of the lower portions of the homodimer, causing the upper portions to move apart, which leads to a widening

of the active site to accommodate a larger acceptor (i.e. cellodextrin) in *Rt*CDP (Figure 1.15B).[156]

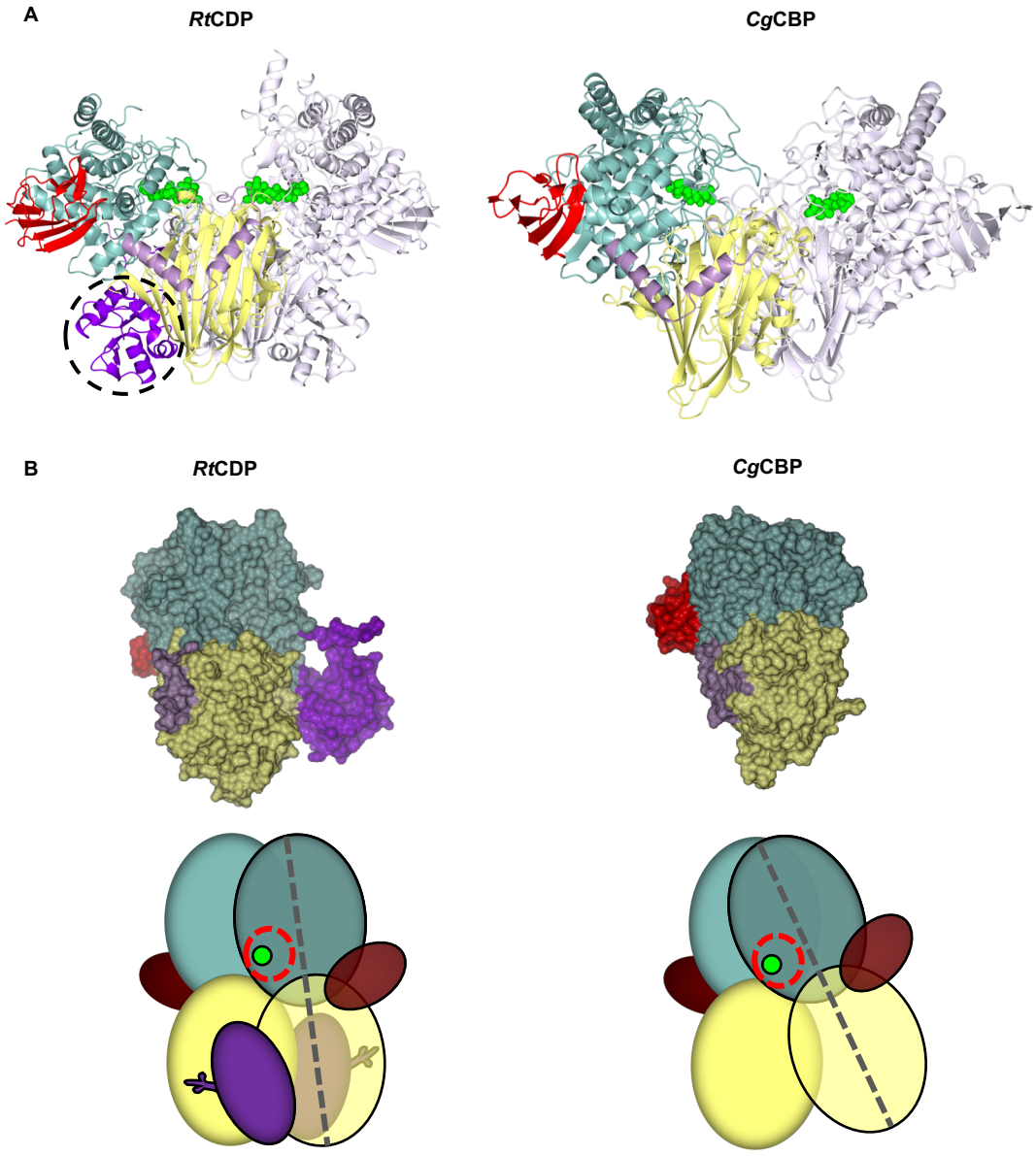


Figure 1.15. Possible mechanism distinguishing substrate chain length preferences in *Rt*CDP and *Cg*CBP. (**A**) Structures of *Rt*CDP and *Cg*CBP in cartoon representation. The domains are coloured in one subunit as followed; N-terminal α/β domain = purple, β-sandwich = lemon, linker helix = lilac, $(\alpha/\alpha)_6$ catalytic domain = green, C-terminal jelly roll = red. (**B**) Molecular surface representation for the monomers of *Rt*CDP and *Cg*CBP (top). Simplified representations of the corresponding views of the dimers, where the various domains are shown as ellipsoids (bottom). The N-terminal α/β domain in *Rt*CDP contains an extension arm. α-helix linkers are not shown. The principal axis of the front subunit in each of the two dimers is shown as a dotted grey line, which emphasises the rotation of this subunit in one dimer relative to the other about an axis perpendicular to the page. Figure reproduced from O'Neill *et al*.[156]

1.6.4. Auxiliary sites and their contributions to CAZymes activity on oligosaccharide and polysaccharide substrates

In addition to the previously discussed mechanisms that CAZymes themselves employ to restrict or distinguish chain length of their substrates/products, binding of long glycan substrate/product in some CAZymes can be assisted by non-catalytic, auxiliary binding sites, namely carbohydrate-binding modules (CBMs) and surface-binding sites (SBSs). These auxiliary sites/modules could have direct impact on the enzyme ability to bind and operate on long substrates/products, or indirect impact on the enzyme efficiency by disrupting the substrate structure to allow more enzyme accessibility. Examples of CBMs and SBSs contribution on their concomitant enzyme activities and efficiency on long glycan substrates are discussed herein.

1.6.4.1. CBMs contribution to CAZymes activity and processivity on polysaccharides

CBMs are non-catalytic modules appended to GHs or GTs through a flexible linker.[157–160] Conservation of amino acid sequences and structural folds permits classification of CBMs into 83 families on CAZy database.[62] The main roles of CBMs include; 1) concentrating the appended enzymes on to the polysaccharide substrates, which leads to faster degradation of the polysaccharide (a proximity effect);[161,162] 2) directing the appended enzyme to substrates, thus increasing the binding specificity towards the targets;[163] 3) disruption of crystalline carbohydrate substrates.[164–167] The supporting evidence for substrate disruption role of CBMs is somewhat questionable since one of the CBM families (CBM33), which was previously shown to promote chitin degradation through substrate binding and disruption,[164] was later characterised and classified as AA10 family containing lytic polysaccharide monooxygenase enzymes.[168] Thus, the enhancement of chitin degradation by CBM33 was a result of its oxygenase activity rather than substrate-CBM binding interaction. CBMs typically functioned independently and their intimate contacts with the catalytic module are normally not required to perform the aforementioned roles.

Deletion of CBMs from its appended enzyme resulted in a significant loss of the enzyme activity on insoluble polysaccharide substrates but had insignificant effect on soluble substrates.[169,170] CBMs have an important contribution to the processivity of their associated enzyme. In processive GHs acting on insoluble polysaccharide substrates, CBMs could have crucial contribution to the enzyme processivity by

impacting the initial binding and/or the decrystallisation of the insoluble substrates (Figure 1.16). In the case of substrate binding, CBM can recognise the quaternary structure of insoluble polysaccharide, such as triple helix crystalline structure of β-(1→3)-glucans, which enable the appended enzyme to operate on insoluble, long chain glucan that adopt the triple helical conformation.[170] Nevertheless, excessive binding of CBMs to crystalline substrates could also result in non-productive binding that subsequently delay substrate hydrolysis. This view has led to the proposal that CBMs contribute more towards decrystallisation of crystalline substrate, which has been supported by recent evidence indicating that decrystallisation of the insoluble materials by CBMs could have a significant impact on the substrate hydrolysis by the appended enzyme despite having weak binding affinity.[171]

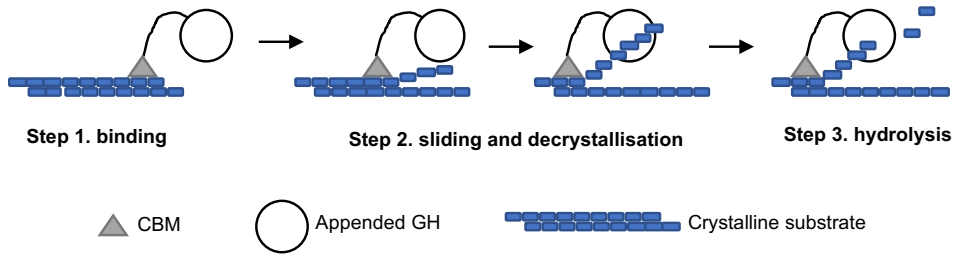


Figure 1.16. A hypothesis for the mode of action of processive GHs on crystalline cellulose. Figure modified from Gao *et al*.[171]

In addition to the contribution on binding and disruption of insoluble polysaccharides, there are examples where the CBMs formed intimate contact with the concomitant enzyme active site and contributed directly to the enzyme activity on substrates/product with different oligosaccharide chain lengths. This is evident in CBM32 appended to a mannanase (GH5) from *Ruminiclostridium thermocellum*, in which the CBM and its appended enzyme were proposed to collectively form a substrate-binding site capable of accommodating a mannopentaose and hydrolysing the substrate primarily into mannose and mannotetraose (Figure 1.17A, wild type). Deletion of CBM32 altered the chain length of the hydrolytic products to mannobiose and mannotriose (Figure 1.17A, GH5 only).[172] This finding is in keeping with another structural and biochemical studies on CBM22 appended to a xylanase (GH10) from *Ruminiclostridium thermocellum*, where the CBM and its concomitant enzyme formed excessive contacts, creating a putative xylan binding tract that accommodate xylo-oligosaccharide chains (Figure 1.17B).[173]

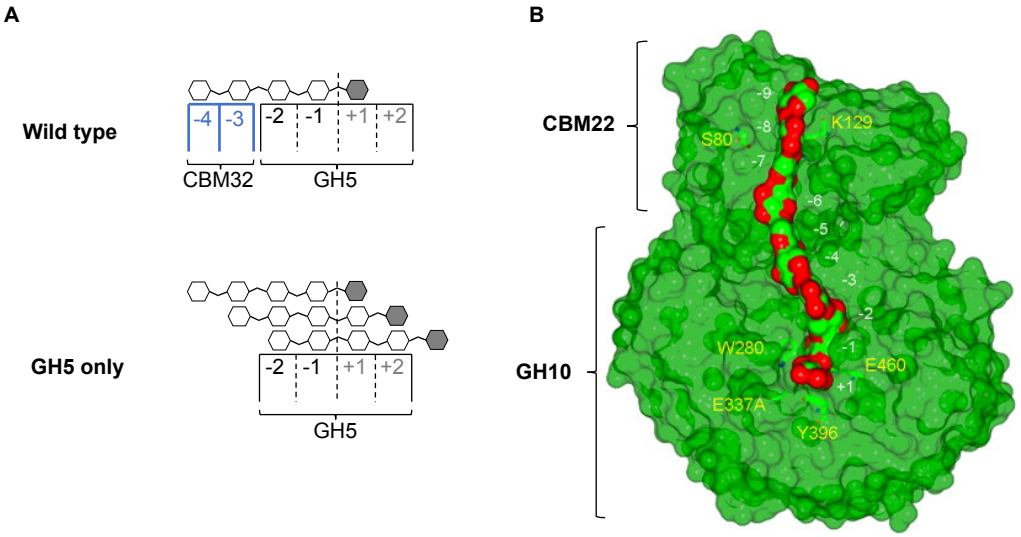


Figure 1.17. Intimate association between CBM and the appended catalytic module determining enzyme substrate/product chain length. (**A**) A proposed model for the adsorption of mannopentaose by the wild type mannanase with appended CBM (GH5-CBM32) or by the GH5 catalytic domain only. White hexagons are mannose residues, with the reducing end coloured in grey. Subsite -3 and -4 have been proposed to be formed by CBM32. Figure modified from Mizutani *et al*.[169] (**B**) The putative xylan binding tract in CBM22-GH10. CBM22–GH10 monomer is shown as a solid green molecular surface In crystal, only the xylotriose moiety is observed in electron density in subsites −1 to −3. Xylose residues labelled as subsite -4 to −9 and +1 were modelled. Figure was reproduced from Najmudin *et al*.[173]

1.6.4.2. Surface binding sites (SBSs) for oligosaccharides and their contributions to enzyme activity on large polysaccharide substrates

SBSs, also called secondary-binding site, are non-catalytic sites found in CAZymes from bacteria,[174,175] fungi,[176,177] plants,[178,179] animals and green alga.[180]. SBSs are normally located in a fixed position at a distant site on the surface of the enzyme catalytic domain or on a domain intimately associated with the catalytic domain.[181] The roles of SBS include; 1) targeting of associated enzymes to their substrates; 2) loading substrates into the active sites; 3) disrupting the structure of substrates; 4) keeping the substrate in close proximity to the enzyme for subsequent reactions; 5) allosteric activation of the enzyme; 6) anchoring of the enzyme to the cell wall of the host microorganism; 7) enhancing enzyme processivity.[182,183] Most SBSs are described in GHs, but an SBS in a glucan synthase has recently been identified.[184] An SBS associated with GP was reported in an *Arabidopsis thaliana* cytoplasmic glucan phosphorylase PHS2 (*At*PHS2) from GT35 family. The *At*PHS2 SBS is proposed to be responsible for the tight binding of the enzyme to insoluble glucan nanoparticle surface.[185] The role of the SBS in binding insoluble long polyglucan was further supported by the absence of the SBS site in a related plastid glucan phosphorylase from barley (*Hv*Pho1), the enzyme of which has low affinity for long α-

(1→4)-glucans (starch and glycogen).[186] Moreover, a glycogen binding site, which is a characteristic feature of GT35, was replaced by an additional negatively-charged loop (L78) in *Hv*Pho1. L78 was proposed to physically occlude the supposedly glycogen binding site in *At*Pho1 (Figure 1.18) and the lack of this loop increased affinity and activity of *Hv*Pho1 on amylopectin and starch.[186]

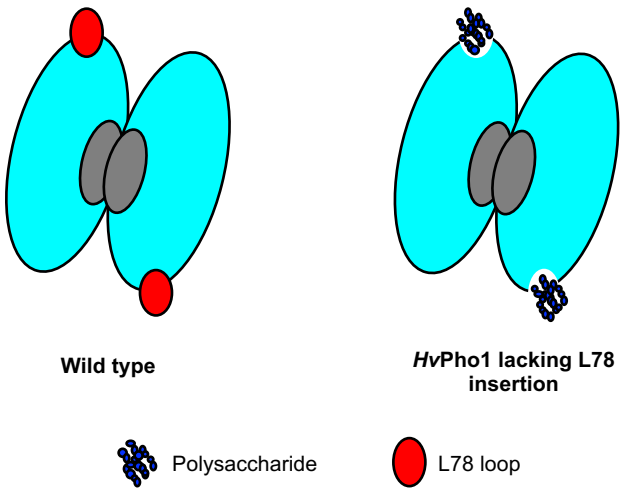


Figure 1.18. Proposed model describing the effect of an additional loop in *Hv*Pho1 on glucan binding.[186] The dimeric form of *Hv*Pho1 is represent by cyan ovals, where the dimer interface is coloured in grey. L78 insertion loops are coloured in red and occluding the supposedly glycogen binding site. The removal of L78 presumably allowed better access to the glycogen binding site, which led to the observed increased enzyme affinity to amylopectin and starch.

Unlike CBMs, identification of SBSs is challenging because the SBS amino acid sequences are rarely conserved. Existing SBSs were typically discovered through structural studies of the enzymes in complex with oligosaccharide substrates or substrate analogues, although complementary techniques such as carbohydrate interaction assays (affinity electrophoresis, insoluble polysaccharide pulldown and surface plasmon resonance) can aid the identification of new SBSs/CBMs.[187] SBSs have been extensively investigated in α-amylases (GH13), possibly due to interests in the industrial applications of the enzymes in the breakdown of water-insoluble starch granules. The presence of SBSs (and CBMs) could aid the adsorption of the enzymes to the surface of starch granules, thus enhancing the enzyme activity on starch. [178,182] In general, the SBS in α-amylases contains two spatially locked aromatic residues, which could provide hydrophobic CH-π interaction to sugar rings of the substrates.[179] Despite the lack of conserved amino acids among SBSs, it was shown that categorisation and prediction of SBSs based on their shapes, residues, charges, hydropathicity and surface accessibility might be possible through computational

analysis of SBS-containing PDB structures.[188] The computational approach could provide the first step towards prediction tool for new SBS discovery.

1.7. Aims of this study

Complexity of carbohydrate structures arises from a variety of their monosaccharide composition, stereo- and regiochemistry of glycosidic bonds, and subsequent modifications of the glycan chains. Structural complexity of carbohydrates enables them to assume various roles in nature. A vast number of CAZymes with precise stereo- and regio-specificity have evolved to cope with the carbohydrate complexity. From biotechnological perspectives, the vast availability of these CAZymes enable their applications as tools for modification and creation of synthetic carbohydrates with defined structures and precise quantity.

Given the importance of β-(1→3)-glucan in nature and their applications in biotechnological settings, this work concentrated on three main aims; 1) investigating the mechanism of existing disaccharide phosphorylase acting on laminaibiose; 2) identifying nucleotide sequences encoding the microalgal oligosaccharide phosphorylases acting on β-(1→3)-glucan; 3) investigating the molecular mechanism distinguishing disaccharide and oligosaccharide phosphorylases acting on β-(1→3)-glucan; 4) possible roles of β-(1→3)-glucan phosphorylases in β-(1→3)-glucan metabolism. This work will briefly explore the application of the β-(1→3)-glucan phosphorylases in oligosaccharide synthesis and outline any translational research opportunities that may arise as a result.

1.8. References

[1] A. Varki, *Glycobiology* **1993**, *3*, 97–130.

[2] H. E. McFarlane, A. Döring, S. Persson, *Annu. Rev. Plant Biol.* **2014**, *65*, 69–94.

[3] P. A. Orlandi, F. W. Klotz, J. D. Haynes, *J. Cell Biol.* **1992**, *116*, 901–909.

[4] C. M. Cabral, Y. Liu, R. N. Sifers, *Trends Biochem. Sci.* **2001**, *26*, 619–624.

[5] E. Weerapana, B. Imperiali, *Glycobiology* **2006**, *16*, 91–101.

[6] A. Varki, *Glycobiology* **2017**, *27*, 3–49.

[7] J. R. Bishop, M. Schuksz, J. D. Esko, *Nature* **2007**, *446*, 1030–1037.

[8] A. Buléon, P. Colonna, V. Planchot, S. Ball, *Int. J. Biol. Macromol.* **1998**, *23*, 85–112.

[9] J. P. Mua, D. S. Jackson, *J. Agric. Food Chem.* **1997**, *45*, 3840–3847.

[10] E. Pilling, A. M. Smith, *Plant Physiol* **2003**, *132*, 365–371.

[11] B. B. Hallac, A. J. Ragauskas, *Biofuels, Bioprod. Biorefining* **2011**, *5*, 215–225.

[12] Y. Nishiyama, J. Sugiyama, H. Chanzy, P. Langan, *J. Am. Chem. Soc.* **2003**, *125*, 14300–14306.

[13] C. Laroche, P. Michaud, *Recent Pat. Biotechnol.* **2007**, *1*, 59–73.

[14] L. Barsanti, V. Passarelli, V. Evangelista, A. M. Frassanito, P. Gualtieri, *Nat. Prod. Rep.* **2011**, *28*, 457–466.

[15] H. Miyakoshi, T. Aoki, *Int. J. Immunopharmacol.* **1984**, *6*, 373–379.

[16] A. M. Neyrinck, A. Mouson, N. M. Delzenne, *Int. Immunopharmacol.* **2007**, *7*, 1497—1506.

[17] J. Y. Lee, Y. J. Kim, H. J. Kim, Y. S. Kim, W. Park, *Molecules* **2012**, *17*, 5404–5411.

[18] E. J. Olson, J. E. Standing, N. Griego-Harper, O. A. Hoffman, A. H. Limper, *Infect. Immun.* **1996**, *64*, 3548–3554.

[19] S. J. Stasinopoulos, P. R. Fisher, B. A. Stone, V. A. Stanisich, *Glycobiology* **1999**, *9*, 31–41.

[20] S. K. Ghai, H. M., A. Amemura, T. Harada, *J. Gen. Microbiol.* **1981**, *122*, 33–40.

[21] W. J. Kenyon, C. S. Buller, *J Ind Microbiol Biotechnol* **2002**, *29*, 200–203.

[22] H. Futatsuyama, T. Yui, K. Ogawa, *Biosci. Biotechnol. Biochem.* **1999**, *63*, 1481–1483.

[23] X.-Y. Chen, J.-Y. Kim, *Plant Signal. Behav.* **2009**, *4*, 489–492.

[24] W. R. Scheible, M. Pauly, *Curr. Opin. Plant Biol.* **2004**, *7*, 285–295.

[25] L. Barsanti, R. Vismara, V. Passarelli, P. Gualtieri, *J. Appl. Phycol.* **2001**, *13*, 59–65.

[26] M. Sletmoen, B. T. Stokke, *Biopolymers* **2008**, *89*, 310–321.

[27] T. Okobira, K. Miyoshi, K. Uezu, K. Sakurai, S. Shinkai, *Biomacromolecules* **2008**, *9*, 783–788.

[28] C. S. Mangolim, T. T. Da Silva, V. C. Fenelon, L. N. Koga, S. B. S. De Ferreira, M. L. Bruschi, G. Matioli, *PLoS One* **2017**, *12*, e0171469.

[29] J. P. Knox, Y. Benitez-Alfonso, *Curr. Opin. Plant Biol.* **2014**, *22*, 93–100.

[30] K. Miyoshi, K. Uezu, K. Sakurai, S. Shinkai, *Chem. Biodivers.* **2004**, *1*, 916–924.

[31] A. Bacic, G. B. Fincher, B. A. Stone, *Chemistry, Biochemistry, and Biology of (1→3)-β-Glucans and Related Polysaccharides*, Academic Press/Elsevier, San Diego, **2009**.

[32] A. Lazaridou, C. G. Biliaderis, *J. Cereal Sci.* **2007**, *46*, 101–118.

[33] S. M. Tosh, Y. Brummer, P. J. Wood, Q. Wang, J. Weisz, *Carbohydr. Polym.* **2004**, *57*, 249–259.

[34] M. A. Caballero, D. Jallet, L. Shi, C. Rithner, Y. Zhang, G. Peers, *Algal Res.* **2016**, *20*, 180–188.

[35] A. Beattie, E. L. Hirst, E. Percival, *Biochem J* **1961**, *79*, 531–537.

[36] B. S. Paulsen, S. Myklestad, *Carbohydr. Res.* **1978**, *62*, 386–388.

[37] B. A. Wustman, M. R. Gretz, K. D. Hoagland, *Plant Physiol.* **1997**, *113*, 1059–1069.

[38] A. Chiovitti, A. Bacic, J. Burke, R. Wetherbee, *Eur. J. Phycol.* **2003**, *38*, 351–360.

[39] B. Gügi, T. Le Costaouec, C. Burel, P. Lerouge, W. Helbert, M. Bardor, *Mar.*

Drugs **2015**, *13*, 5993–6018.

[40] T. R. Størseth, K. Hansen, K. I. Reitan, J. Skjermo, *Carbohydr. Res.* **2005**, *340*, 1159–1164.

[41] P. de S. Bonfim-Mendonça, I. R. G. Capoci, F. K. Tobaldini-Valerio, M. Negri, T. I. E. Svidzinski, *Int. J. Mol. Sci.* **2017**, *18*, 1629.

[42] S. M. Read, G. Currie, A. Bacic, *Carbohydr. Res.* **1996**, *281*, 187–201.

[43] V. Aimanianda, C. Clavaud, C. Simenel, T. Fontaine, M. Delepierre, J. P. Latgé, *J. Biol. Chem.* **2009**, *284*, 13401–13412.

[44] C. M. Douglas, *Med. Mycol.* **2001**, *39*, 55–66.

[45] I. Geoghegan, G. Steinberg, S. Gurr, *Trends Microbiol.* **2017**, *25*, 957–967.

[46] J. J. Volman, J. D. Ramakers, J. Plat, *Physiol. Behav.* **2008**, *94*, 276–284.

[47] R. Russo, L. Barsanti, V. Evangelista, A. M. Frassanito, V. Longo, L. Pucci, G. Penno, P. Gualtieri, *Food Sci. Nutr.* **2017**, *5*, 205–214.

[48] F. Y. Yamamoto, F. Yin, W. Rossi, M. Hume, D. M. Gatlin, *Fish Shellfish Immunol.* **2018**, *77*, 273–279.

[49] J. Skov, P. W. Kania, L. Holten-Andersen, B. Fouz, K. Buchmann, *Fish Shellfish Immunol.* **2012**, *33*, 111–120.

[50] C. Steele, R. R. Rapaka, A. Metz, S. M. Pop, D. L. Williams, S. Gordon, J. K. Kolls, G. D. Brown, *PLoS Pathog.* **2005**, *1*, e42.

[51] P. Kankkunen, L. Teirila, J. Rintahaka, H. Alenius, H. Wolff, S. Matikainen, *J. Immunol.* **2010**, *184*, 6335–6342.

[52] G. D. Brown, S. Gordon, *Nature* **2001**, *413*, 36–37.

[53] G. D. Brown, J. Herre, D. L. Williams, J. A. Willment, A. S. J. Marshall, S. Gordon, *J. Exp. Med.* **2003**, *197*, 1119–1124.

[54] A. Mueller, J. Raptis, P. J. Rice, J. H. Kalbfleisch, R. D. Stout, H. E. Ensley, W. Browder, D. L. Williams, *Glycobiology* **2000**, *10*, 339–346.

[55] S.-H. Young, V. Robinson, M. Barger, D. Frazer, V. Castranova, R. Jacobs, *J. Toxicol. Environ. Heal. Part A* **2003**, *66*, 551–563.

[56] H. Saito, Y. Yoshioka, N. Uehara, **1991**, *7*, 181–190.

[57] G. C.-F. Chan, W. K. Chan, D. M.-Y. Sze, *J. Hematol. Oncol.* **2009**, *2*, 25.

[58] K. Fujimoto, K. Matsuda, K. Aso, *Tohoku J. Agric. Res.* **1962**, *13*, 61–67.

[59] H. He, F. Yang, Y. Du, *Carbohydr. Res.* **2002**, *337*, 1673–1678.

[60] F. Jamois, V. Ferrières, J. P. Guégan, J. C. Yvin, D. Plusquellec, V. Vetvicka, *Glycobiology* **2005**, *15*, 393–407.

[61] V. Lombard, H. Golaconda Ramulu, E. Drula, P. M. Coutinho, B. Henrissat, *Nucleic Acids Res.* **2014**, *42*, D490–D495.

[62] B. I. Cantarel, P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, B. Henrissat, *Nucleic Acids Res.* **2009**, *37*, D233–D238.

[63] B. Henrissat, I. Callebaut, P. Lehn, J. Mornon, G. Davies, *Proc. Natl. Acad. Sci.* **1995**, *93*, 7090–7094.

[64] G. J. Davies, M. L. Sinnott, *Biochem. J.* **2008**, 26–32.

[65] G. J. Davies, S. J. Williams, *Biochem. Soc. Trans.* **2016**, *44*, 79–87.

[66] B. Henrissat, A. Bairoch, *Biochem. J.* **1996**, *316*, 695–696.

[67] B. Henrissat, G. Davies, *Curr. Opin. Struct. Biol.* **1997**, *7*, 637–644.

[68] M. Hess, A. Sczyrba, R. Egan, T.-W. Kim, H. Chokhawala, G. Schroth, S. Luo, D. S. Clark, F. Chen, T. Zhang, R. I. Mackie, L. A. Pennacchio, S. G. Tringe, A. Visel, T. Woyke, Z. Wang, E. M. Rubin, *Science* **2011**, *331*, 463–467.

[69] J. M. Brulc, D. A. Antonopoulos, M. E. Berg Miller, M. K. Wilson, A. C. Yannarell, E. A. Dinsdale, R. E. Edwards, E. D. Frank, J. B. Emerson, P. Wacklin, P. M. Coutinho, B. Henrissat, K. E. Nelson, B. A. White, *Proc. Natl. Acad. Sci.* **2009**, *106*, 1948–1953.

[70] K. A. Barry, I. S. Middelbos, B. M. Vester Boler, S. E. Dowd, J. S. Suchodolski, B. Henrissat, P. M. Coutinho, B. A. White, G. C. Fahey, K. S. Swanson, *J. Proteome Res.* **2012**, *11*, 5924–5933.

[71] S. Abubucker, N. Segata, J. Goll, A. M. Schubert, J. Izard, B. L. Cantarel, B. Rodriguez-Mueller, J. Zucker, M. Thiagarajan, B. Henrissat, O. White, S. T. Kelley, B. Methé, P. D. Schloss, D. Gevers, M. Mitreva, C. Huttenhower, *PLoS Comput. Biol.* **2012**, *8*, e1002358.

[72] D. A. Cecchini, E. Laville, S. Laguerre, P. Robe, M. Leclerc, J. Doré, B. Henrissat,

M. Remaud-Siméon, P. Monsan, G. Potocki-Véronèse, *PLoS One* **2013**, *8*, e72766.

[73] J.-H. Hehemann, G. Correc, T. Barbeyron, W. Helbert, M. Czjzek, G. Michel, *Nature* **2010**, *464*, 908–912.

[74] C. V. Castell-Miller, J. J. Gutierrez-Gonzalez, Z. J. Tu, K. E. Bushley, M. Hainaut, B. Henrissat, D. A. Samac, *PLoS One* **2016**, *11*, e0154122.

[75] R. J. O'Connell, M. R. Thon, S. Hacquard, S. G. Amyotte, J. Kleemann, M. F. Torres, U. Damm, E. A. Buiate, L. Epstein, N. Alkan, J. Altmüller, L. Alvarado-Balderrama, C. A. Bauser, C. Becker, B. W. Birren, Z. Chen, J. Choi, J. A. Crouch, J. P. Duvick, M. A. Farman, P. Gan, D. Heiman, B. Henrissat, R. J. Howard, M. Kabbage, C. Koch, B. Kracher, Y. Kubo, A. D. Law, M.-H. Lebrun, Y.-H. Lee, I. Miyara, N. Moore, U. Neumann, K. Nordström, D. G. Panaccione, R. Panstruga, M. Place, Ro. H. Proctor, D. Prusky, G. Rech, R. Reinhardt, J. A. Rollins, S. Rounsley, C. L. Schardl, D. C. Schwartz, N. Shenoy, K. Shirasu, U. R. Sikhakolli, K. Stüber, S. A. Sukno, J. A. Sweigard, Y. Takano, H. Takahara, F. Trail, H. C. van der Does, L. M. Voll, I. Will, S. Young, Q. Zeng, J. Zhang, S. Zhou, M. B. Dickman, P. Schulze-Lefert, E. V. L. van Themaat, L. -J. Ma, L. J. Vaillancourt, Nat. Genet. 2012, 44, 1060–1065.

[76] S. V. S. Ipcho, J. K. Hane, E. A. Antoni, D. Ahren, B. Henrissat, T. L. Friesen, P. S. Solomon, R. P. Oliver, *Mol. Plant Pathol.* **2012**, *13*, 531–545.

[77] S. DiGuistini, Y. Wang, N. Y. Liao, G. Taylor, P. Tanguay, N. Feau, B. Henrissat, S. K. Chan, U. Hesse-Orce, S. M. Alamouti, C. K. M. Tsui, R. T. Docking, A. Levasseur, S. Haridas, G. Robertson, I. Birol, R. A. Holt, M. A. Marra, R. C. Hamelin, M. Hirst, S. J. M. Jones, J. Bohlmann, C. Breuil, Proc. Natl. Acad. Sci. 2011, 108, 2504–2509.

[78] E. C. O'Neill, M. Trick, L. Hill, M. Rejzek, R. G. Dusi, C. J. Hamilton, P. V. Zimba, B. Henrissat, R. A. Field, *Mol. BioSyst.* **2015**, *11*, 2808–2820.

[79] Y. Yoshida, T. Tomiyama, T. Maruta, M. Tomita, T. Ishikawa, K. Arakawa, *BMC Genomics* **2016**, *17*, 182.

[80] S. S. Macdonald, M. Blaukopf, S. G. Withers, *J. Biol. Chem.* **2015**, *290*, 4887–4895.

[81] S. S. Macdonald, A. Patel, V. L. C. Larmour, C. Morgan-Lang, S. J. Hallam, B. L. Mark, S. G. Withers, *J. Biol. Chem.* **2018**, *293*, 3451–3467.

[82] A. Levasseur, E. Drula, V. Lombard, P. M. Coutinho, B. Henrissat, *Biotechnol. Biofuels* **2013**, *6*, 41.

[83] R. J. Quinlan, M. D. Sweeney, L. Lo Leggio, H. Otten, J.-C. N. Poulsen, K. S. Johansen, K. B. R. M. Krogh, C. I. Jørgensen, M. Tovborg, A. Anthonsen, T. Tryfona, C. P. Walter, P. Dupree, F. Xu, G. J. Davies, P. H. Walton, *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 15079–84.

[84] P. V. Harris, D. Welner, K. C. McFarland, E. Re, J. C. Navarro Poulsen, K. Brown, R. Salbo, H. Ding, E. Vlasenko, S. Merino, F. Xu, J. Cherry, S. Larsen, L. Lo Leggio, *Biochemistry* **2010**, *49*, 3305–3316.

[85] T. Karnezis, V. C. Epa, B. A. Stone, V. A. Stanisich, *Glycobiology* **2003**, *13*, 693–706.

[86] Y. Tanaka, T. Ogawa, T. Maruta, Y. Yoshida, K. Arakawa, *FEBS Lett.* **2017**, *591*, 1360–1370.

[87] X. Cui, H. Shin, C. Song, W. Laosinchai, Y. Amano, M. R. Brown, *Planta* **2001**, *213*, 223–230.

[88] W. Huang, I. Haferkamp, B. Lepetit, M. Molchanova, S. Hou, W. Jeblick, C. Río Bártulos, P. G. Kroth, *Proc. Natl. Acad. Sci.* **2018**, *115*, 4791–4796.

[89] M. Hildebrand, K. Manandhar-Shrestha, R. Abbriano, *Algal Res.* **2017**, *23*, 66–77.

[90] W. Huang, C. Río Bártulos, P. G. Kroth, *J. Eukaryot. Microbiol.* **2016**, *63*, 536–546.

[91] G. Michel, T. Tonon, D. Scornet, J. M. Cock, B. Kloareg, *New Phytol.* **2010**, *188*, 67–81.

[92] S. B. Inoue, N. Takewaki, T. Takasuka, T. Mio, M. Adachi, Y. Fujii, C. Miyamoto, M. Arisawa, Y. Furuichi, T. Watanabe, *Eur J Biochem* **1995**, *231*, 845–854.

[93] C. M. Douglas, F. Foor, J. A. Marrinan, N. Morin, J. B. Nielsen, A. M. Dahl, P. Mazur, W. Baginsky, W. Li, M. El-Sherbeini, *Proc. Natl. Acad. Sci.* **1994**, *91*, 12907–12911.

[94] G. Lesage, H. Bussey, *Microbiol. Mol. Biol. Rev.* **2006**, *70*, 317–343.

[95] A. Gastebois, I. Mouyna, C. Simenel, C. Clavaud, B. Coddeville, M. Delepierre, J. P. Latgé, T. Fontaine, *J. Biol. Chem.* **2010**, *285*, 2386–2396.

[96] L.-M. Papaspyridi, A. Zerva, E. Topakas, *Catalysts* **2018**, *8*, 274.

[97] P. Mazur, N. Morin, W. Baginsky, M. el-Sherbeini, J. A. Clemas, J. B. Nielsen, F. Foor, *Mol. Cell. Biol.* **1995**, *15*, 5671–81.

[98] M. S. Doblin, L. De Melis, E. Newbigin, A. Bacic, S. M. Read, *Plant Physiol.* **2001**, *125*, 2040–2052.

[99] P. G. Kroth, A. Chiovitti, A. Gruber, V. Martin-Jezequel, T. Mock, M. S. Parker, M. S. Stanley, A. Kaplan, L. Caron, T. Weber, U. Maheswari, E. V Armbrust, C. Bowler, *PLoS One* **2008**, *3*, e1426.

[100] M. S. Chauton, P. Winge, T. Brembu, O. Vadstein, A. M. Bones, *Plant Physiol.* **2013**, *161*, 1034–1048.

[101] E. V. Armbrust, J. A. Berges, C. Bowler, B. R. Green, D. Martinez, N. H. Putnam, S. Zhou, A. E. Allen, K. E. Apt, M. Bechner, M. A. Brzezinski, B. K. Chaal, A. Chiovitti, A. K. Davis, M. S. Demarest, J. C. Detter, T. Glavina, D. Goodstein, M. Z. Hadi, U. Hellsten, M. Hildebrand, B. D. Jenkins, J. Jurka, V. V. Kapitonov, N. Kröger, W. W. Y. Lau, T. W. Lane, F. W. Larimer, J. C. Lippmeier, S. Lucas, M. Medina, A. Montsant, M. Obornik, M. S. Parker, B. Palenik, G. J. Pazour, P. M. Richardson, T. A. Rynearson, M. A. Saito, D. C. Schwartz, K. Thamatrakoln, K. Valentin, A. Vardi, F. P. Wilkerson, D. S. Rokhsar, Science 2004, 306, 79–86.

[102] J. M. Cock, L. Sterck, P. Rouzé, D. Scornet, A. E. Allen, G. Amoutzias, V. Anthouard, F. Artiguenave, J.-M. Aury, J. H. Badger, B. Beszteri, K. Billiau, E. Bonnet, J. H. Bothwell, C. Bowler, C. Boyen, C. Brownlee, C. J. Carrano, B. Charrier, G. -Y. Cho, S. M. Coelho, J. Collén, E. Corre, C. Da Silva, L. Delage, N. Delaroque, S. M. Dittami, S. Doulbeau, M. Elias, G. Farnham, C. M. M. Gachon, B. Gschloessl, S. Heesch, K. Jabbari, C. Jubin, H. Kawai, K. Kimura, B. Kloareg, F. C. Küpper, D. Lang, A. Le Bail, C. Leblanc, P. Lerouge, M. Lohr, P. J. Lopez, C. Martens, F. Maumus, G. Michel, D. Miranda-Saavedra, J. Morales, H. Moreau, T. Motomura, C. Nagasato, C. A. Napoli, D. R. Nelson, P. Nyvall-Collén, A. F. Peters, C. Pommier, P. Potin, J. Poulain, H. Quesneville, B. Read, S. A. Rensing, A. Ritter, S. Rousvoal, M. Samanta, G. Samson, D. C. Schroeder, B. Ségurens, M. Strittmatter, T. Tonon, J. W. Tregear, K. Valentin, P. von Dassow, T. Yamagishi, Y. Van de Peer, P. Wincker, Nature 2010, 465, 617–621.

[103] L. R. Marechal, S. H. Goldemberg, *J. Biol. Chem.* **1964**, *239*, 3163–3167.

[104] A. D. Tomos, D. H. Northcote, *Biochem. J.* **1978**, *174*, 283–290.

[105] D. Bäumer, A. Preisfeld, H. G. Ruppel, *J. Phycol.* **2001**, *46*, 38–46.

[106] T. Takeda, Y. Nakano, M. Takahashi, N. Konno, Y. Sakamoto, R. Arashida, Y. Marukawa, E. Yoshida, T. Ishikawa, K. Suzuki, *Phytochemistry* **2015**, *116*, 21–27.

[107] L. R. Maréchal, S. H. Goldemberg, *Biochem. Biophys. Res. Commun.* **1963**, *13*, 2–5.

[108] M. Kitaoka, T. Sasaki, H. Taniguchi, *Arch. Biochem. Biophys.* **1993**, *304*, 508–514.

[109] S. H. Goldemberg, L. R. Maréchal, B. C. De Souza, L. R. Mar, B. C. D. E. Souza, *J. Biol. Chem.* **1966**, *241*, 45–50.

[110] B. Bissaro, P. Monsan, R. Fauré, M. J. O'Donohue, *Biochem. J.* **2015**, *467*, 17–35.

[111] D. E. Koshland, *Biol. Rev.* **1953**, *28*, 416–436.

[112] Z. Qin, S. Yang, L. Zhao, X. You, Q. Yan, Z. Jiang, *J. Biol. Chem.* **2017**, *292*, 1666–1678.

[113] M. Faijes, A. Planas, *Carbohydr. Res.* **2007**, *342*, 1581–1594.

[114] V. M. A. Ducros, C. A. Tarling, D. L. Zechel, A. M. Brzozowski, T. P. Frandsen, I. von Ossowski, M. Schulein, S. G. Withers, G. J. Davies, *Chem. Biol.* **2003**, *10*, 619–628.

[115] M. Hrmova, T. Imai, S. J. Rutten, J. K. Fairweather, L. Pelosi, V. Bulone, H. Driguez, G. B. Fincher, *J. Biol. Chem.* **2002**, *277*, 30102–30111.

[116] J. K. Fairweather, M. Hrmova, S. J. Rutten, G. B. Fincher, H. Driguez, *Chemistry* **2003**, *9*, 2603–2610.

[117] L. Pelosi, T. Imai, H. Chanzy, L. Heux, E. Buhler, V. Bulone, *Biochemistry* **2003**, *42*, 6264–6274.

[118] L. K. Him, L. Pelosi, H. Chanzy, J. Putaux, V. Bulone, *Eur. J. Biochem.* **2001**, *268*, 4628–4638.

[119] A. Periasamy, N. Shadiac, A. Amalraj, S. Garajová, Y. Nagarajan, S. Waters, H. D. T. Mertens, M. Hrmova, *Biochim. Biophys. Acta - Biomembr.* **2013**, *1828*, 743–757.

[120] G. Pergolizzi, S. Kuhaudomlarp, R. A. Field, E. Kalita, *Protein Pept. Lett.* **2017**, *24*, 696–709.

[121] M. Nakajima, H. Toyoizumi, K. Abe, H. Nakai, H. Taguchi, M. Kitaoka, *PLoS One* **2014**, *9*, e92353.

[122] M. Kitaoka, Y. Matsuoka, K. Mori, M. Nishimoto, K. Hayashi, *Biosci. Biotechnol. Biochem.* **2012**, *76*, 343–348.

[123] T. Nihira, Y. Saito, M. Kitaoka, M. Nishimoto, K. Otsubo, H. Nakai, *Carbohydr. Res.* **2012**, *361*, 49–54.

[124] K. Hamura, W. Saburi, S. Abe, N. Morimoto, H. Taguchi, H. Mori, H. Matsui, *Biosci. Biotechnol. Biochem.* **2012**, *76*, 812–818.

[125] Y. H. P. Zhang, L. R. Lynd, *Appl. Environ. Microbiol.* **2004**, *70*, 1563–1569.

[126] M. Hidaka, Y. Honda, M. Kitaoka, S. Nirasawa, K. Hayashi, T. Wakagi, H. Shoun, S. Fushinobu, *Structure* **2004**, *12*, 937–947.

[127] T. Nihira, Y. Saito, M. Nishimoto, M. Kitaoka, K. Igarashi, K. Ohtsubo, H. Nakai, *FEBS Lett.* **2013**, *587*, 3556–3561.

[128] M. Nakajima, M. Kitaoka, *Appl. Environ. Microbiol.* **2008**, *74*, 6333–6337.

[129] M. Nakajima, M. Nishimoto, M. Kitaoka, *J. Biol. Chem.* **2009**, *284*, 19220–19227.

[130] Y. Yamamoto, D. Kawashima, A. Hashizume, M. Hisamatsu, N. Isono, *Biosci. Biotechnol. Biochem.* **2013**, *77*, 1949–1954.

[131] Y. Ogawa, K. Noda, S. Kimura, M. Kitaoka, M. Wada, *Int. J. Biol. Macromol.* **2014**, *64*, 415–419.

[132] K. Ogawa, J. Tsijrugi, T. Watanabe, *Carbohydr. Res.* **1973**, *29*, 397–403.

[133] A. S. Palma, T. Feizi, Y. Zhang, M. S. Stoll, A. M. Lawson, E. Díaz-Rodríguez, M. A. Campanero-Rhodes, J. Costa, S. Gordon, G. D. Brown, W. Chai, *J. Biol. Chem.* **2006**, *281*, 5771–5779.

[134] G. J. Davies, K. S. Wilson, B. Henrissat, *Biochem. J.* **1997**, *321*, 557–559.

[135] C. Roth, N. Weizenmann, N. Bexten, W. Saenger, W. Zimmermann, T. Maier, N. Sträter, *Sci. Adv.* **2017**, *3*, e1601386.

[136] B. Strokopytov, R. M. A. Knegtel, D. Penninga, H. J. Rozeboom, K. H. Kalk, L. Dijkhuizen, B. W. Dijkstra, *Biochemistry* **1996**, *35*, 4241–4249.

[137] B. A. Van Der Veen, J. C. M. Uitdehaag, D. Penninga, G. J. W. M. Van Alebeek, L. M. Smith, B. W. Dijkstra, L. Dijkhuizen, *J. Mol. Biol.* **2000**, *296*, 1027–1038.

[138] K. A. Sin, A. Nakamura, H. Masaki, Y. Matsuura, T. Uozumi, *J. Biotechnol.* **1994**, *32*, 283–288.

[139] G. Parsiegla, A. K. Schmidt, G. E. Schulz, *Eur. J. Biochem.* **1998**, *255*, 710–717.

[140] J. C. Uitdehaag, K. H. Kalk, B. a van Der Veen, L. Dijkhuizen, B. W. Dijkstra, *J. Biol. Chem.* **1999**, *274*, 34868–34876.

[141] J. P. Bohin, *FEMS Microbiol. Lett.* **2000**, *186*, 11–19.

[142] M. W. Breedveld, K. J. Miller, *Microbiol. Rev.* **1994**, *58*, 145–161.

[143] M. S. Roset, A. E. Ciocchini, R. A. Ugalde, N. Iñón De Iannino, *J. Bacteriol.* **2006**, *188*, 5003–5013.

[144] A. E. Ciocchini, L. S. Guidolin, A. C. Casabuono, A. S. Couto, N. I. de Iannino, R. A. Ugalde, *Proc. Natl. Acad. Sci.* **2007**, *104*, 16492–16497.

[145] G. Hagelueken, B. R. Clarke, H. Huang, A. Tuukkanen, I. Danciu, D. I. Svergun, R. Hussain, H. Liu, C. Whitfield, J. H. Naismith, *Nat. Struct. Mol. Biol.* **2015**, *22*, 50–56.

[146] K. A. Joiner, N. Grossman, M. Schmetz, L. Leive, *J. Immunol.* **1986**, *136*, 710–715.

[147] C. Burns, Valorie, E. J. Pishko, A. Preston, D. J. Maskell, E. T. Harvill, *Infect. Immun.* **2003**, *71*, 86–94.

[148] J. D. King, S. Berry, B. R. Clarke, R. J. Morris, C. Whitfield, *Proc. Natl. Acad. Sci.* **2014**, *111*, 6407–6412.

[149] D. M. Williams, O. G. Ovchinnikova, A. Koizumi, I. L. Mainprize, M. S. Kimber, T. L. Lowary, C. Whitfield, *Proc. Natl. Acad. Sci.* **2017**, *114*, E1215–E1223.

[150] A. S. Ramírez, J. Boilevin, A. R. Mehdipour, G. Hummer, T. Darbre, J. L. Reymond, K. P. Locher, *Nat. Commun.* **2018**, *9*, 445.

[151] T. Senoura, S. S. Ito, H. Taguchi, M. Higa, S. Hamada, H. Matsui, T. Ozawa, S. Jin, J. Watanabe, J. Wasaki, S. S. Ito, *Biochem. Biophys. Res. Commun.* **2011**, *408*, 701–706.

[152] R. Kawahara, W. Saburi, R. Odaka, H. Taguchi, S. Ito, H. Mori, H. Matsui, *J. Biol.*

Chem. **2012**, *287*, 42389–42399.

[153] S. Ladevéze, L. Tarquis, D. A. Cecchini, J. Bercovici, I. André, C. M. Topham, S. Morel, E. Laville, P. Monsan, V. Lombard, B. Henrissat, G. Potocki-Véronèse, *J. Biol. Chem.* **2013**, *288*, 32370–32383.

[154] S. Ladevèze, G. Cioci, P. Roblin, L. Mourey, S. Tranier, G. Potocki-Véronèse, *Acta Crystallogr. D. Biol. Crystallogr.* **2015**, *71*, 1335–1346.

[155] T. Tsuda, T. Nihira, K. Chiku, E. Suzuki, T. Arakawa, M. Nishimoto, M. Kitaoka, H. Nakai, S. Fushinobu, *FEBS Lett* **2015**, *589*, 3816–3821.

[156] E. C. O'Neill, G. Pergolizzi, C. E. M. Stevenson, D. M. Lawson, S. A. Nepogodiev, R. A. Field, *Carbohydr. Res.* **2017**, *451*, 118–132.

[157] H. J. Gilbert, J. P. Knox, A. B. Boraston, *Curr. Opin. Struct. Biol.* **2013**, *23*, 669–677.

[158] A. B. Boraston, D. N. Bolam, H. J. Gilbert, G. J. Davies, *Biochem. J.* **2004**, *382*, 769–781.

[159] H. Hashimoto, *Cell. Mol. Life Sci.* **2006**, *63*, 2954–2967.

[160] D. Guillén, S. Sánchez, R. Rodríguez-Sanoja, *Appl. Microbiol. Biotechnol.* **2010**, *85*, 1241–1249.

[161] D. N. Bolam, A. Ciruela, S. McQueen-Mason, P. Simpson, M. P. Williamson, J. E. Rixon, A. Boraston, G. P. Hazlewood, H. J. Gilbert, *Biochem. J.* **1998**, *331*, 775–781.

[162] G. Carrard, A. Koivula, H. Soderlund, P. Beguin, *Proc. Natl. Acad. Sci.* **2000**, *97*, 10342–10347.

[163] C. Herve, A. Rogowski, A. W. Blake, S. E. Marcus, H. J. Gilbert, J. P. Knox, *Proc. Natl. Acad. Sci.* **2010**, *107*, 15293–15298.

[164] G. Vaaje-Kolstad, S. J. Horn, D. M. F. Van Aalten, B. Synstad, V. G. H. Eijsink, *J. Biol. Chem.* **2005**, *280*, 28492–28497.

[165] L. S. Wang, Y. Z. Zhang, P. J. Gao, *Sci. China, Ser. C Life Sci.* **2008**, *51*, 620–629.

[166] N. Din, H. G. Damude, N. R. Gilkes, R. C. Miller, R. a Warren, D. G. Kilburn, *Proc. Natl. Acad. Sci.* **1994**, *91*, 11383–11387.

[167] T. T. Teeri, *Trends Biotechnol.* **1997**, *15*, 160–167.

[168] Z. Forsberg, G. Vaaje-kolstad, B. Westereng, A. C. Bunsæ, Y. Stenstrøm, A. Mackenzie, M. Sørlie, S. J. Horn, V. G. H. Eijsink, *Protein Sci.* **2011**, *20*, 1479–1483.

[169] K. Mizutani, M. Sakka, T. Kimura, K. Sakka, *FEBS Lett.* **2014**, *588*, 1726–1730.

[170] A. Hettle, A. Fillo, K. Abe, P. Massel, B. Pluvinage, D. N. Langelaan, S. P. Smith, A. B. Boraston, *J. Biol. Chem.* **2017**, *292*, 16955–16968.

[171] D. Gao, S. P. S. Chundawat, A. Sethi, V. Balan, S. Gnanakaran, B. E. Dale, *Proc. Natl. Acad. Sci.* **2013**, *110*, 10922–10927.

[172] K. Mizutani, V. O. Fernandes, S. Karita, A. S. Luís, M. Sakka, T. Kimura, A. Jackson, X. Zhang, C. M. G. A. Fontes, H. J. Gilbert, K. Sakka, *Appl. Environ. Microbiol.* **2012**, *78*, 4781–4787.

[173] S. Najmudin, B. A. Pinheiro, J. A. M. Prates, H. J. Gilbert, M. J. Romão, C. M. G. A. Fontes, *J. Struct. Biol.* **2010**, *172*, 353–362.

[174] T. C. Tan, B. N. Mijts, K. Swaminathan, B. K. C. Patel, C. Divne, *J. Mol. Biol.* **2008**, *378*, 850–868.

[175] N. M. Koropatkin, T. J. Smith, *Structure* **2010**, *18*, 200–215.

[176] A. M. Brzozowski, G. J. Davies, *Biochemistry* **1997**, *36*, 10837–10845.

[177] A. Vujičić-Žagar, B. W. Dijkstra, *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **2006**, *62*, 716–721.

[178] M. M. Nielsen, S. Bozonnet, E. N. Seo, J. A. Mótyán, J. M. Andersen, A. Dilokpimol, M. A. Hachem, G. Gyémánt, H. Næsted, L. Kandra, B. W. Sigurskjold, B. Svensson, *Biochemistry* **2009**, *48*, 7686–7697.

[179] S. Bozonnet, M. T. Jensen, M. M. Nielsen, N. Aghajari, M. H. Jensen, B. Kramhøft, M. Willemoës, S. Tranier, R. Haser, B. Svensson, *FEBS J.* **2007**, *274*, 5055–5067.

[180] L. Sim, S. R. Beeren, J. Findinier, D. Dauville, S. G. Ball, A. Henriksen, M. M. Palcic, *J. Biol. Chem.* **2014**, *289*, 22991–23003.

[181] C. Wilkens, D. Cockburn, S. Andersen, B. O. Petersen, C. Ruzanski, R. A. Field, O. Hindsgaul, H. Nakai, B. McCleary, A. M. Smith, M. A. Hachem, B. Svensson, *J. Appl. Glycosci.* **2015**, *62*, 87–93.

[182] U. Baroroh, M. Yusuf, S. D. Rachman, S. Ishmayana, M. R. A. A. Syamsunarno, J. Levita, T. Subroto, *Enzyme Res.* **2017**, *2017*, 4086845.

[183] S. Cuyvers, E. Dornez, J. A. Delcour, C. M. Courtin, *Crit. Rev. Biotechnol.* **2012**, *32*, 93–107.

[184] S. Baskaran, V. M. Chikwana, C. J. Contreras, K. D. Davis, W. A. Wilson, A. A. DePaoli-Roach, P. J. Roach, T. D. Hurley, *J. Biol. Chem.* **2011**, *286*, 33999–34006.

[185] E. C. O'Neill, A. M. Rashid, C. E. M. Stevenson, A.-C. Hetru, A. P. Gunning, M. Rejzek, S. A. Nepogodiev, S. Bornemann, D. M. Lawson, R. A. Field, *Chem. Sci.* **2014**, *5*, 341–350.

[186] J. A. Cuesta-Seijo, C. Ruzanski, K. Krucewicz, S. Meier, P. Hägglund, B. Svensson, M. M. Palcic, *PLoS One* **2017**, *12*, e0175488.

[187] D. Cockburn, C. Wilkens, A. Dilokpimol, H. Nakai, A. Lewińska, M. A. Hachem, B. Svensson, *PLoS One* **2016**, *11*, e0160112.

[188] S. Samaei-Daryan, B. Goliaei, A. Ebrahim-Habibi, *J. Mol. Recognit.* **2017**, *30*, e2624.

Chapter 2

Unravelling the specificity of a GH94 laminaribiose phosphorylase from *Paenibacillus* sp. YM-1 towards donor substrates glucose/mannose 1-phosphate using X-ray crystallography and STD NMR spectroscopy*

*This chapter is an adaptation of work that has been published in ChemBioChem and appears with permission.[1]

2.1. Introduction

Glycoside phosphorylases (GPs) are a group of carbohydrate-active enzymes that catalyse the reversible cleavage of glycosidic linkages in di- or oligo-saccharides by transferring non-reducing end glycosyl residue to inorganic phosphate.[2–5] The reverse reaction (synthetic reaction) of GPs is of practical importance because it can be used as an alternative method of enzymatic glycosylation utilising sugar 1-phosphates as donor substrates (see Figure 2.1 for chemical structures of sugar 1-phosphates mentioned in this study). GPs have been classified based on their sequence identity into glycoside hydrolase (GH) and glycosyltransferase (GT) families, or categorised into retaining and inverting phosphorylases, depending on the anomeric configuration of the *O*-glycoside product with respect to the sugar 1-phosphate substrates. Substrates for GP-catalysed glycosylation are more readily available in comparison to those for GT-catalysed reactions, making GPs attractive biocatalysts for carbohydrate syntheses. The use of GP biocatalysts have been demonstrated in academic research, such as in the synthesis of homogeneous crystalline cellulose;[6] self-assembled structures of alkylated cellulose;[7] cellulose nanoribbon with primary amino groups;[8] and formation of oligo(ethylene glycol)-bearing cellulose hydrogels;[9] and more widely at industrial scale, such as for the synthesis of 2-*O*-(α-D-glucopyranosyl)-*sn*-glycerol, a cosmetic ingredient, by sucrose phosphorylase;[10] kilogram scale synthesis of lacto-*N*-biose, a prebiotic, has been achieved with lacto-*N*-biose phosphorylase;[11] and the synthesis of disaccharide sweetener, kojibiose, has been produced with a sucrose phosphorylase mutant from *Bifidobacterium adolescentis.*[12]

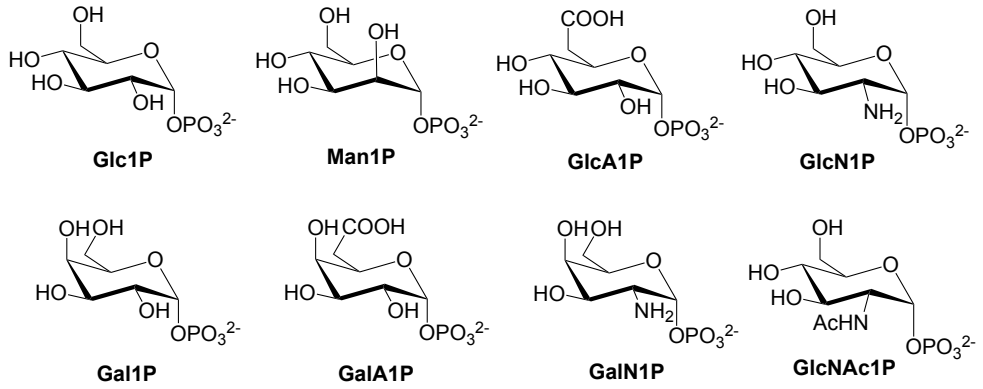


Figure 2.1. Structures of sugar 1-phosphates discussed in this chapter.

One of the most studied GP families is GH94 family, which includes inverting GPs acting on β-(1→2)- (sophorose),[13] β-(1→3)- (laminaribiose),[14] and β-(1→4)-

linked glycans (cellobiose,[15,16] cellodextrin,[17] chitobiose[18] and cellobionic acid[19]). Several characterised GH94 GPs show broad specificity towards non-physiological acceptor substrates. For instance, cellodextrin phosphorylase (CDP) from *Ruminiclostridium stercorarium* has been used to produce cellobiose-containing antioxidants such as glycosylated-resveratrol.[20,21] Cellobiose phosphorylase (CBP) is capable of using xylose as a noncognate acceptor, to produce glucopyranosyl-xylose,[22,23] or using simple alcohols as acceptors to produce alkyl β-glucosides.[24] Whilst relaxed acceptor specificity has been demonstrated for a number of GPs, their specificity for sugar 1-phosphates is relatively narrow. The majority of GH94 GPs use α-D-glucose 1-phosphate (Glc1P) as a donor substrate, with the exception of chitobiose phosphorylase (ChBP), which uses α-D-GlcNAc 1-phosphate (GlcNAc1P) as its natural donor,[18,25] although it can also use Glc1P with 20 times reduction in efficiency.[18] Relaxed donor specificity has also been demonstrated for CDP from *Ruminiclostridium stercorarium* which can use either Glc1P or α-D-galactose 1-phosphate (Gal1P) as its glycosyl donor for glycolipid synthesis, albeit with 10 times less efficiency on Gal1P.[26] Both CBP and CDP from *Ruminiclostridium thermocellum* are capable of using α-D-glucosyl fluoride as a donor for the synthesis of cellobiose and cellodextrin.[27]

Numerous X-ray crystal structures are available for GH94 enzymes, either in the presence of phosphate or sulphate (PDB codes: 2CQS, 3QDE, 3RSY, and 2CQT),[28,29] the acceptors (PDB codes: 3S4B, 1V7X, 5H40, 4ZLG and 5NZ8),[30–33] iminosugar inhibitors (PDB codes: 3QFY, 3QFZ, 3QG0, and 5H41),[32,34] or disaccharide products (PDB codes: 3S4A and 4ZLF).[33] These structures provide valuable resources that can be used to guide the engineering of GPs for noncognate substrates. Structure-guided site-directed mutagenesis has been performed extensively on cellobiose phosphorylase from *Cellvibrio gilvus (Cg*CBP) which enabled its conversion to a lactose phosphorylase with 7.5 times higher specific activity on lactose in comparison to the *Cg*CBP parent.[35] In addition, a single mutation (E649C) in *Cg*CBP created an enzyme variant that is capable of using methyl β-glucoside, ethyl β-glucoside and phenyl β-glucoside as acceptors.[36,37] Another *Cg*CBP variant was created by mutation of five amino acids within and around the entrance to the enzyme active, which broadened the acceptor range to include both β- and α-glucosides.[36] In contrast to the situation of acceptor substrate studies, the number of reported GP structures in complex with the sugar donors are relatively limited, with only a β-(1→2)-glucan phosphorylase from *Lachnoclostridium phytofermentans (Lp*SOGP) in complex with Glc1P being reported (PDB code 5H42),[32] which limits understanding of the recognition of the sugar 1-phosphate donors by the GH94 family.

Whilst crystallographic studies provide valuable 'snapshots' of enzyme active sites, they do not capture the dynamics of the enzyme-ligand interaction in solution. Therefore, other techniques to study protein-ligand interactions in solution are needed to complement the crystallographic data. Saturation transfer difference nuclear magnetic resonance spectroscopy (STD NMR) was developed to study protein-ligand binding interactions in solution, based on the transfer of magnetisation from the protein protons to the protons of the ligand whilst the ligand is bound.[38] Those ligand protons in close contact with the protein exhibit the strongest STD NMR intensities, thus allowing the mapping of the ligand binding epitope.[39] STD NMR can be used to facilitate the study of protein-glycan interactions, which is often difficult to study in solution due to the weak affinity, as well as the flexibility and complexity of the glycan ligands. STD NMR has been used to elucidate the specificity of protein-glycan interactions on different sialoglycan structures,[40] and to reveal the importance of glycan polarity, which determines the interaction and subsequent biological activation of its receptor.[41] This technique has also been used to study enzyme-glycan interactions to elucidate recognition features that can be used for inhibitor design, such as the study of ligand recognition by enzymes that are involved in mycobacterial cell wall biosynthesis, including UDP-galactopyranose mutase[42,43] and galactofuranosyltransferases.[44] The same technique has been used to elucidate the binding of human blood group GTs to their substrates, a process that is crucial to the biosynthesis of human blood group antigen.[45,46]

Following on from previous work on the establishment of GP structure-function relationships and their application in carbohydrate syntheses[31,47], herein the GH94 laminaribiose phosphorylase from *Paenibacillus* sp. YM-1 (*Ps*LBP), which has previously been reported for its specificity towards laminaribiose (β-D-glucopyranosyl-(1→3)-D-glucopyranose)[14], was investigated (Figure 2.2A). The activity of *Ps*LBP on a noncognate donor, α-D-mannose 1-phosphate (Man1P), and cognate acceptor, glucose (Glc), in the production of β-D-mannopyranosyl-(1→3)-D-glucopyranose (disaccharide **1**) was evaluated (Figure 2.2B). Furthermore, X-ray crystallography was used in conjunction with STD NMR to investigate the interaction between *Ps*LBP and its substrates, in order to understand structural features that contribute to its donor substrate specificity.

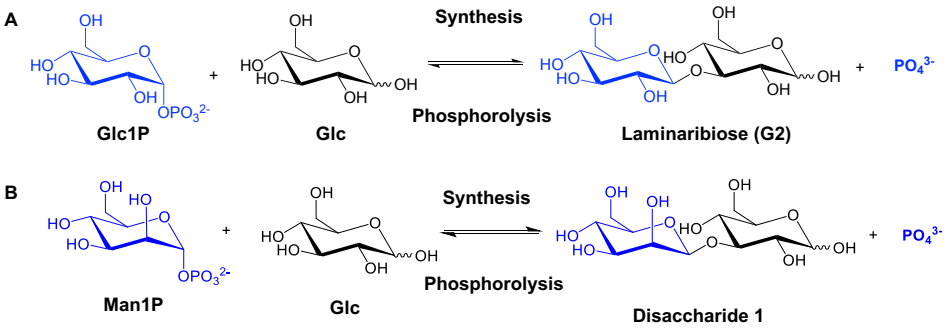


Figure 2.2. Reactions carried out by *Ps*LBP. (**A**) Glc1P and Glc as a donor and acceptor respectively. (**B**) Man1P and Glc as a donor and acceptor respectively.

2.2. Results

2.2.1. Recombinant protein expression and *Ps*LBP activity on the cognate donor and acceptor

To obtain recombinant *Ps*LBP protein for *in vitro* characterisation and X-ray crystallography, the gene-encoding sequence of *Ps*LBP was obtained from GenBank (accession number AB568298.2), codon-optimised for *E. coli* expression and synthesised by Gen9. The gene was amplified by PCR and cloned into a pOPINF expression vector.[48] The recombinant plasmid containing the *PsLBP* gene was introduced into BL21 (DE3) for protein expression. The His_6-tagged recombinant protein was then produced and purified by immobilised metal affinity chromatography (IMAC), followed by gel filtration. The gel filtration trace showed three different main peaks with different elution volumes (Figure 2.3A, peaks a, b and c). The phosphorylase activity of each peak was individually characterised in the synthetic direction (Figure 2.2A) by phosphate release assay. No significant difference in activity was observed between these three peaks (data not shown) and the enzymes in peaks a and b are likely higher oligomeric forms of *Ps*LBP. Therefore, only peak c was used for further experiments, based on its highest protein yield. SDS-PAGE analysis of peak c showed a major band of protein with an approximate mass of 100 kDa, in agreement with the calculated mass of *Ps*LBP monomer (101.6 kDa) (Figure 2.3B). However, gel filtration analysis of peak c against protein standards showed that *Ps*LBP formed a dimer in non-denaturing conditions, with an estimated molecular mass of 240 kDa.

To further confirm that the recombinant *Ps*LBP was active, the enzyme was assayed in the synthetic direction (Figure 2.2A) in the presence of its natural substrates (Glc and Glc1P). The reaction mixture was subjected to High-Performance Anion-Exchange Chromatography with Pulsed Amperometric Detection (HPAEC-PAD) analysis, which showed that laminaribiose (G2) was produced in the reaction (Figure 2.3C).

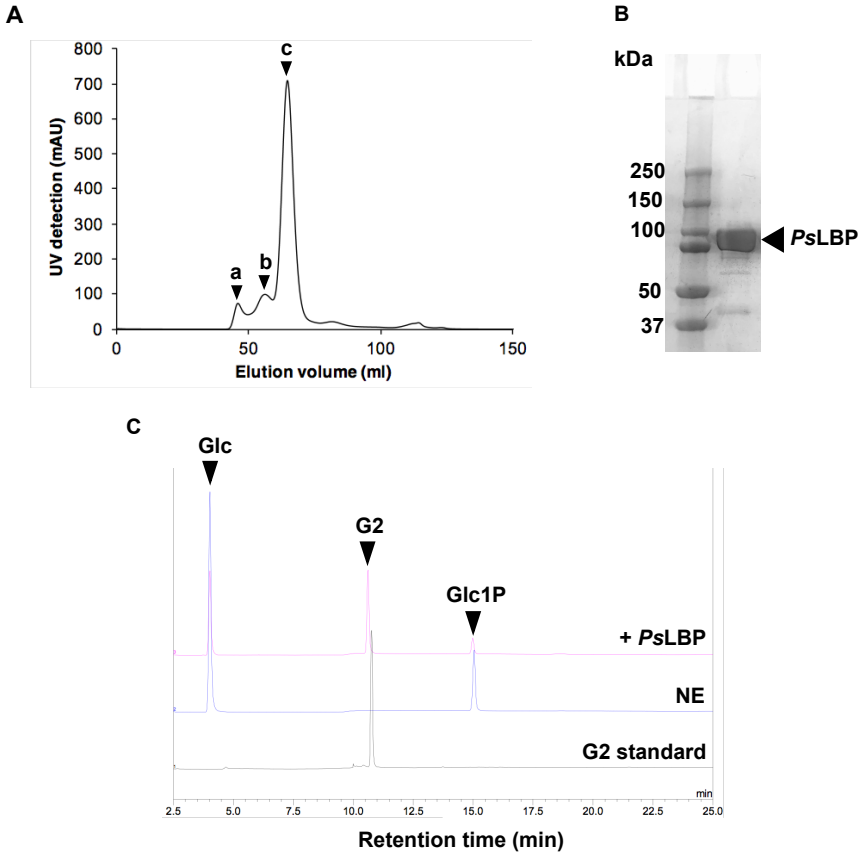


Figure 2.3. Expression and characterisation of the recombinant *Ps*LBP**.** (**A**) Gel filtration trace during the purification of *Ps*LBP. Elution volume of peak c (64.7 ml) was used to estimate the protein molecular mass from a calibration curve constructed from protein standards. (**B**) SDS-PAGE analysis of the recombinant protein after IMAC and gel filtration. (**C**) HPAEC-PAD analysis of the synthetic reaction carried out by *Ps*LBP. The enzyme was incubated with 10 mM Glc and 10 mM Glc1P for 30 minutes at 45℃. NE = no enzyme control. G2 = laminaribiose.

2.2.2. Activity towards noncognate donors

*Ps*LBP activity has previously been screened on several noncognate acceptors, including mannose, methyl β-glucoside, 2-deoxyglucose and 6-deoxyglucose, with 50-100 fold reduction in the activity on these acceptors compared to that of the cognate acceptor Glc.[14] However, the specificity towards noncognate sugar 1-phosphate donors has not been reported. To assess the donor specificity of *Ps*LBP, the enzyme was assayed in the presence of α-D-galactosamine 1-phosphate (GalN1P), α-D-glucosamine 1-phosphate (GlcN1P), α-D-galacturonic-acid 1-phosphate (GalA1P), Gal1P or Man1P as donors and Glc as an acceptor. TLC and HPAEC-PAD analysis of the reactions showed that, of the sugar 1-phosphates tested, the enzyme can only use Man1P as a donor, as indicated by the presence of an additional spot on TLC, corresponding to a generation of disaccharide **1** (Figure 2.4A and B).

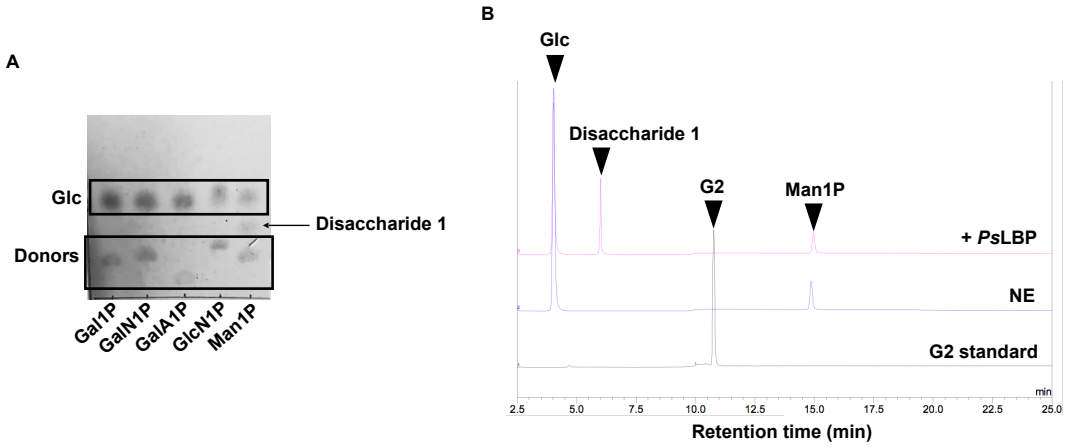


Figure 2.4. *Ps*LBP activity on Man1P and production of disaccharide **1**. (**A**) TLC analysis on the screening reactions of the synthetic activity of *Ps*LBP (8 µg) on 4 different donors (10 mM) in the presence of 10 mM Glc as an acceptor**.** The reactions were incubated at 45 ℃ for 30 minutes. TLC was performed by spotting 0.5 µl of the recovered reaction mixture onto a precoated slides of Silica Gel 60 F254 (Merck) (10 cm x 5 cm), then eluted using a mobile phase containing NH_4OH : H_2O : iso-propanol (3:1:4) in a sealed glass container for 2 hour to allow oligosaccharide separation. (**B**) HPAEC-PAD analysis of the reaction containing Glc and Man1P as substrates.

Kinetic parameters for the synthetic reaction using either Glc1P or Man1P as donors and Glc as an acceptor showed comparable K_M values for Glc1P and Man1P, whilst the k_{cat} for Glc1P is more than 100-fold higher than that for Man1P (Table 2.1).

Table 2.1. Kinetic data of *Ps*LBP acting on Glc1P and Man1P in the presence of 10 mM Glc as an acceptor; for Glc acceptor, experiment was conducted in the presence of 10 mM Glc1P.

Donors/acceptor	k_{cat} (s^{-1})	K_M (mM)	k_{cat}/K_M ($s^{-1}mM^{-1}$)
Glc1P	13.0 ± 1.4	4.20 ± 1.5	3.07
Man1P	0.08 ± 0.01	3.80 ± 1.0	0.02
Glc	15.4 ± 1.3	6.04 ± 1.3	2.55

A large scale enzymatic reaction was carried out in 5 ml reaction containing 700 µg of the enzyme, 10 mM Glc and 20 mM Man1P, incubating for 15 hr to produce milligram quantities of disaccharide **1** (Figure 2.5A). Unreacted Man1P and inorganic phosphate by-product were removed from the reaction mixture by anion exchange chromatography and the disaccharide **1** was subsequently isolated by gel permeation chromatography (GPC) (Figure 2.5B). The isolated disaccharide **1** was analysed on TLC, which showed that only one product was obtained with no Glc contamination (Figure 2.5C). Mass spectrometry analysis of disaccharide **1** on the TLC plate showed a major peak with m/z of 364.9, corresponding to the mass of a disaccharide containing two hexose residues with a sodium adduct (Figure 2.5D). These data are in line with those reported by Awad *et al.* for the same disaccharide, arising from GH130 β-(1→3)-mannan phosphorylase-mediated synthesis.[49]

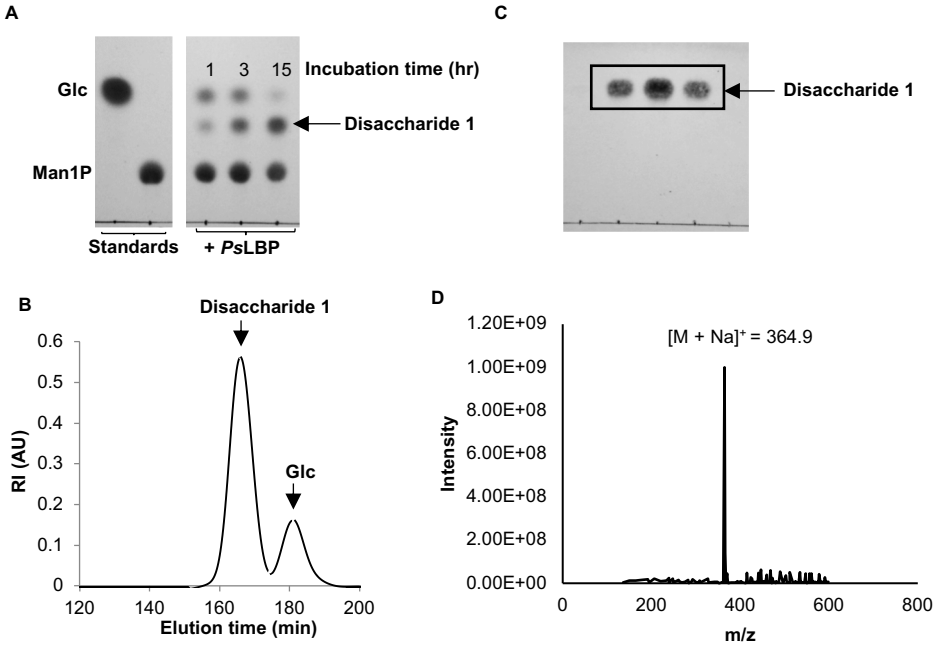


Figure 2.5. Synthesis and purification of disaccharide **1.** (**A**) TLC analysis of the Glc + Man1P large scale reaction at time intervals. (**B**) GPC trace for the purification of disaccharide **1** (Toyopearl HW40S (1,6 x 80 sm), 0.5 mL.min, H_2O). RI = refractive index in arbitrary unit (AU). (**C**) TLC analysis of three fractions from GPC containing purified disaccharide **1**. (**D**) Mass spectrometry analysis of disaccharide **1** on the TLC in (C) by TLC/MS.

Assignment of NMR signals of disaccharide **1** (performed by Dr Sergey Nepogodiev, John Innes Centre, Norwich, UK) (Figure 2.6 and Figure 2.7) was possible with the use of 2D experiments: COSY (Figure 2.8), HSQC (Figure 2.9) and 2D 1H coupled HSQC (Figure 2.10), as well as literature data for model methyl β-mannoside (Appendix 1).[50] Assignment was also helped by simulated spectra of disaccharide **1** generated by CASPER program (Appendix 2).[51] Downfield positions of resonances of C-3 of β-Glc (δ = 84.6 ppm) and C-3 of α-Glc (δ = 82.1 ppm) residues with respect to corresponding signals in Glc (δ = 73.8 and 77.0 for α- and β-anomers respectively)[52] indicated the presence of the 3-*O*-glycosylated glucopyranose unit. Coupled HSQC experiments revealed $^{1}J_{C\text{-}H}$ of 163 Hz the anomeric signals of the mannopyranose residue, a value characteristic of β-mannopyranosides (Figure 2.10).[52] Most of carbon signals of the mannosyl residue in ^{13}C NMR of disaccharide **1** are split into two very close peaks, due to the presence of α/β-anomers of Glc residue (Figure 2.6). Anomeric signals of non-reducing β-Glc residue are expected to appear at δ = 103-104 ppm,[50,53] but there are no peaks in that region. Therefore, the presence β-glucosides can be excluded. By comparison with the previously reported Glc-β-(1→3)-Glc ^{13}C NMR spectra,[54] signals at δ = 95.7 and 92.1 ppm can be assigned to C-1 of reducing β-Glc and α-Glc moieties respectively, whereas signals at δ = 84.6 and 82.1 ppm can be assigned to C-3 β-Glc and C-3 α-Glc respectively.

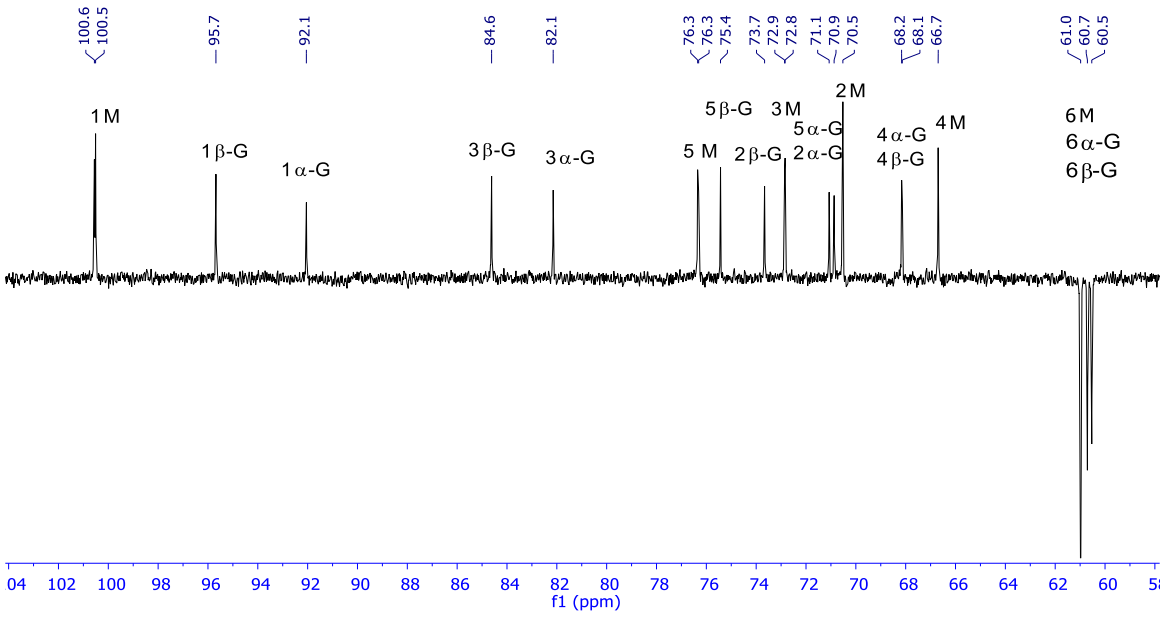


Figure 2.6. ^{13}C NMR (100 MHz, D_2O, 25 °C) of disaccharide **1**.

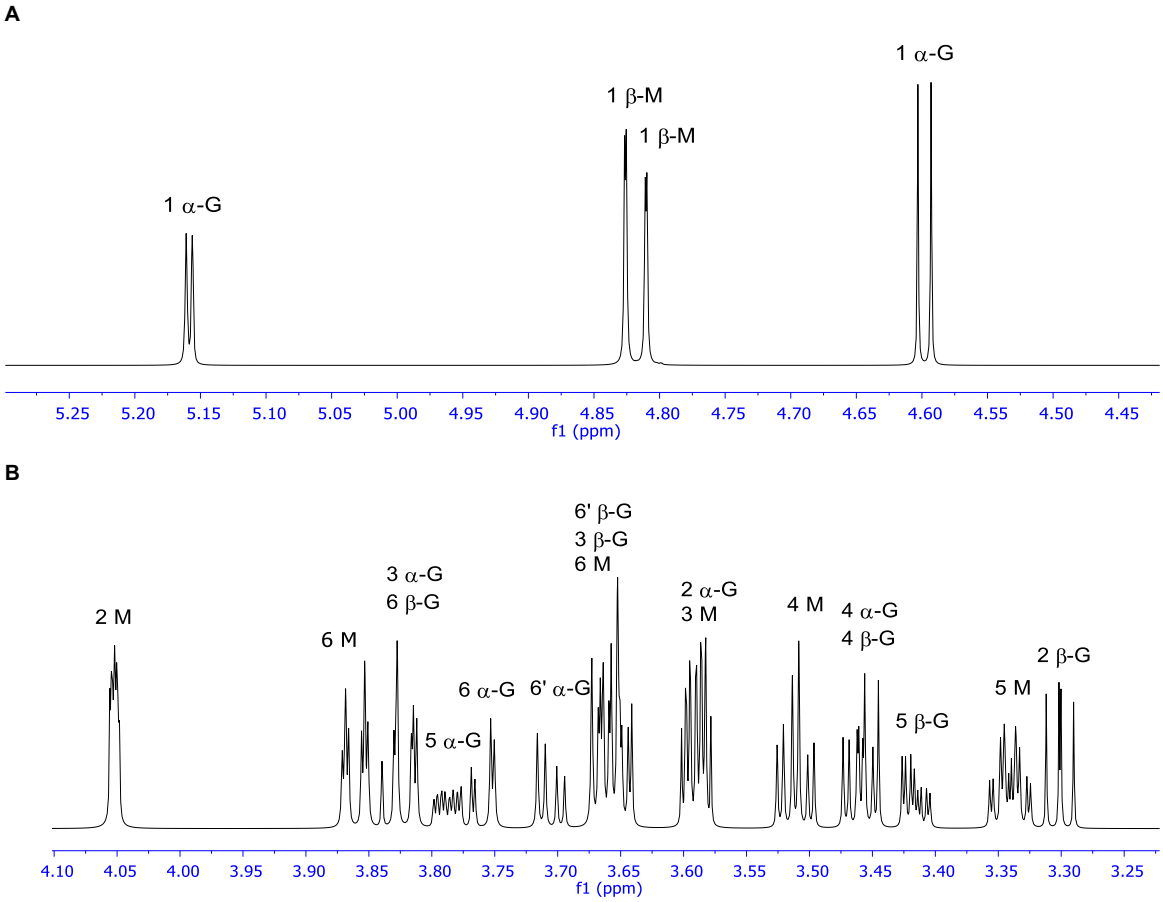


Figure 2.7. ^{1}H NMR (800 MHz, D_2O, 25 °C) of disaccharide **1**. (**A**) anomeric region of ^{1}H NMR (800 MHz, D_2O) of disaccharide **1**. Two signals of H-1 of β-mannopyranosyl residues appear as a result of the presence of two different species of α- and β-anomers of disaccharide **1**. (**B**) non-anomeric regions of the spectrum are characterised by good dispersion of signals corresponding to two species of α- and β-anomers of disaccharide **1**. As a result, signals of β-mannopyranosyl residues appear as two separate resonances for anomeric protons and as overlapping multiplets for ring protons.

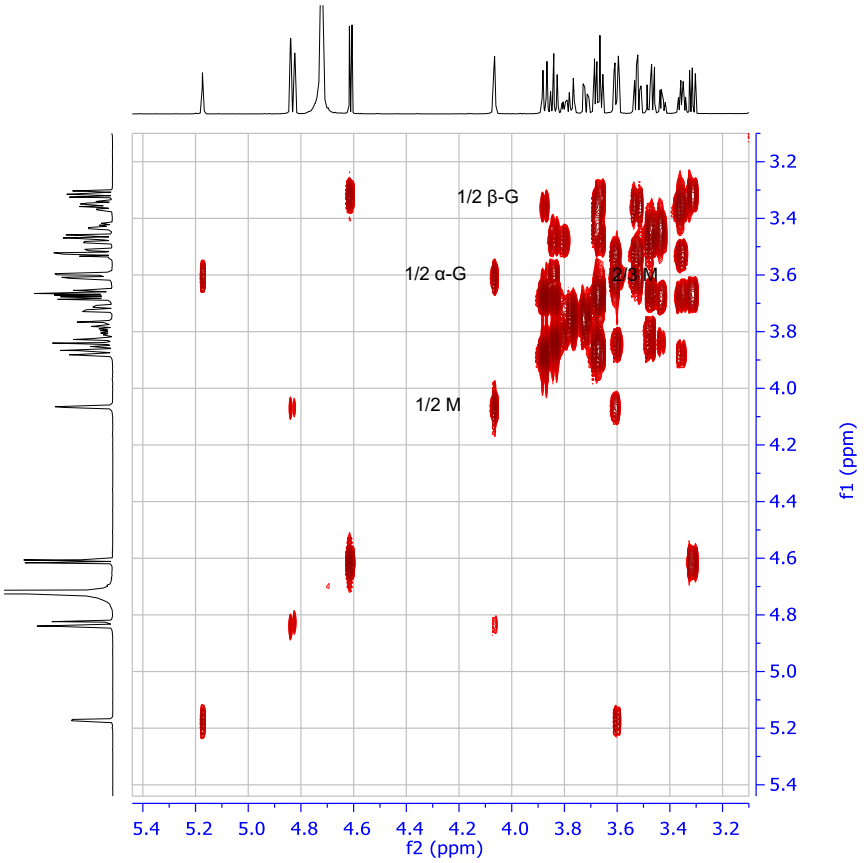


Figure 2.8. COSY spectrum of disaccharide **1**.

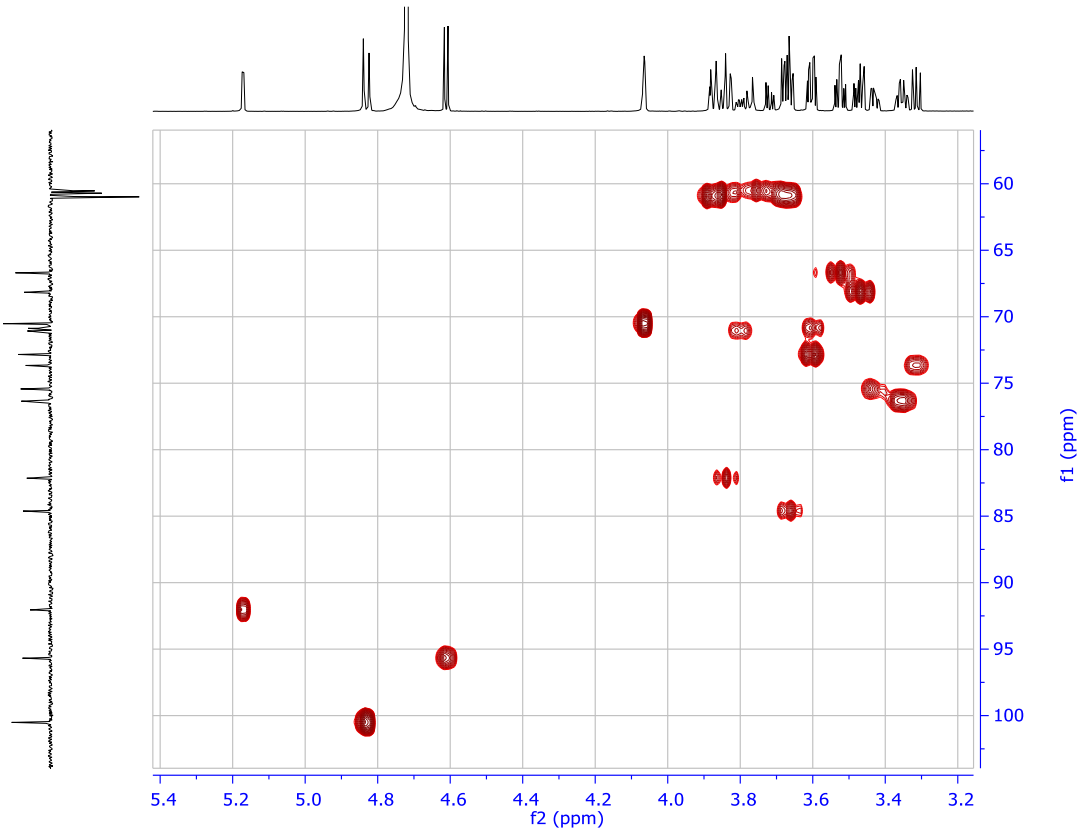


Figure 2.9. HSQC data for disaccharide **1**.

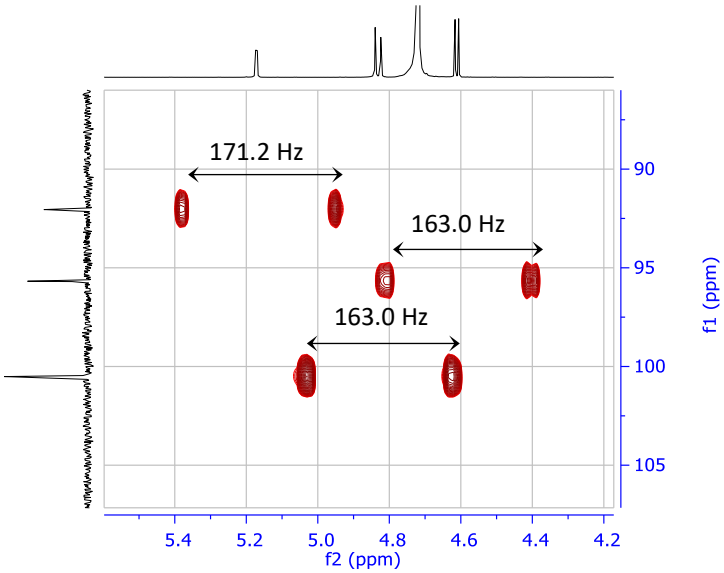


Figure 2.10. 2D 1H coupled HSQC spectrum of disaccharide **1**. Signals of reducing end monosaccharides residues can be clearly identified at δ = 5.17 ($^3J_{C\text{-}H}$ = 171.2 Hz) and δ = 4.61 ($^3J_{C\text{-}H}$ = 163.0 Hz) corresponding to β- and α- glucopyranose residues respectively. The signal at δ 4.90 having typical $^3J_{C\text{-}H}$ value of 163 Hz can be attributed to β-mannopyranosyl residue.

2.2.3. Overall crystal structures of *Ps*LBP

Structural studies of GH94 enzymes provide valuable knowledge on acceptor and phosphate recognition mechanisms by the enzyme, but the knowledge of sugar 1-phosphate recognition is limited because only one structure in complex with Glc1P has been reported so far. In order to investigate the recognition mechanism of Glc1P and Man1P by *Ps*LBP protein, structural studies of the enzyme in complex with the sugar 1-phosphate ligands is clearly required. To obtain protein crystals for structural studies, the recombinant *Ps*LBP (10 mg/ml) was soaked with either Glc1P or Man1P (detailed crystallisation experiment and data processing can be found in Chapter 6, section 6.15.1).

Data collection and processing statistics for *Ps*LBP structures are summarised in Table 2.2. Three *Ps*LBP structures were determined and designated according to the ligands found in their active sites: SO_4^{2-} complex, Glc1P complex and Man1P complex. All structures belong to the same space group ($P4_12_12$) and contain two subunits per asymmetric unit, which are related by a non-crystallographic 2-fold axis that superpose them with an RMSD of 0.7 Å (based on the superimposition of C-α atoms). The two copies of the molecule in the asymmetric unit formed a homodimer with an interfacial area of ~3360 Å^2, as calculated by jsPISA.[55] Homodimer observed in crystal structures is in agreement with the gel filtration analysis, where *Ps*LBP was eluted as a dimer. Other GH94 enzymes also form homodimers, with the exception to *Lp*SOGP which is the only reported monomeric GH94.

Each *Ps*LBP monomer consists of four domains (Figure 2.11A and B); an N-terminal β-sandwich (residues 1-297; yellow), a helical linker region (residues 298-327; lilac), an $(\alpha/\alpha)_6$ catalytic domain (residues 328-808; green) and a C-terminal domain (residues 809-911; red). The domain organisation in *Ps*LBP is similar to that observed in other GH94 disaccharide phosphorylases, which include cellobiose phosphorylases from *Cellulomonas uda* (*Cu*CBP),[56] *Cg*CBP,[28] chitobiose phosphorylase from *Vibrio proteolyticus* (*Vp*ChBP)[30] and cellobionic acid phosphorylase from *Saccharophagus degradans* (*Sd*CBAP).[33] *Ps*LBP lacks an extended N-terminal α/β domain (Figure 2.11B, purple) that has been observed in GH94 oligosaccharide phosphorylases including cellodextrin phosphorylase from *Ruminiclostridium thermocellum* (*Rt*CDP)[31] and *Lp*SOGP.[32] The role of the N-terminal domain in *Rt*CDP was proposed to be involved in the interaction between the lower portions of the homodimer, causing the

upper portions to move apart, which leads to a widening of the active site to accommodate a larger acceptor (i.e. cellodextrin) in *Rt*CDP.[31]

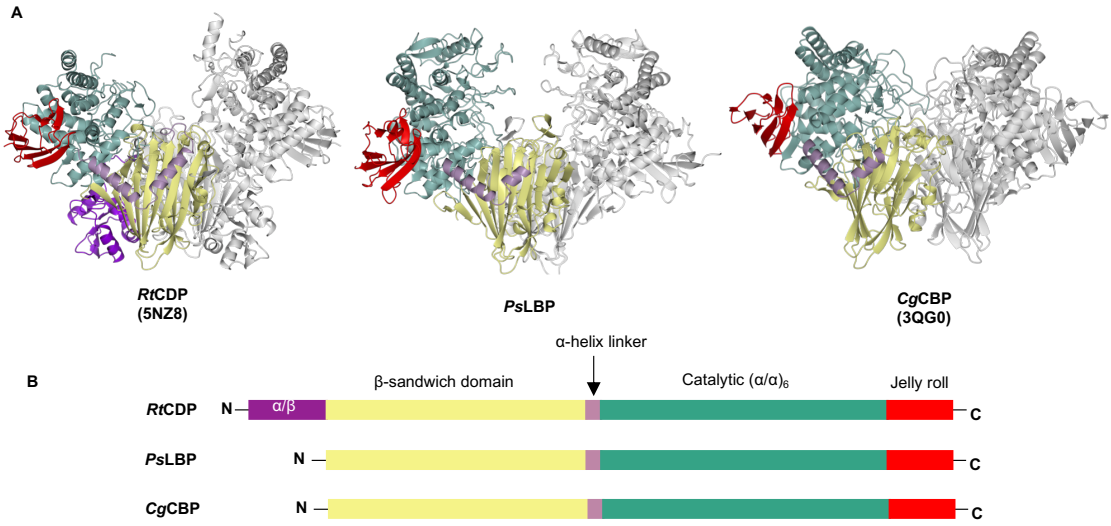


Figure 2.11. Structural comparison between *Cg*CBP, *Rt*CDP and *Ps*LBP. (**A**) Comparison between the GH94 structures. All proteins are in dimeric form. The domains are coloured in one subunit only, whilst the adjacent subunit is coloured in grey. (**B**) Sequences of *Rt*CDP, *Ps*LBP and *Cg*CBP coloured according to the domains present in their structures.

Table 2.2. X-ray data collection and processing of *Ps*LBP structures.

Data set	SO_4^{2-} complex	Mercury derivative	Glc1P complex	Man1P complex
Data Collection				
Beamline	I04	I03/I04	I03	I03
Wavelength (Å)	0.9795	1.0052	0.9800	0.9800
Detector	Pilatus 6M	Pilatus 6M	Pilatus 6M	Pilatus 6M
Resolution range (Å)[a]	103.74-1.95 (1.98-1.95)	88.56-2.90 (2.98-2.90)	75.69-2.50 (2.55-2.50)	73.30-1.82 (1.85-1.82)
Space Group	$P4_12_12$	$P4_12_12$	$P4_12_12$	$P4_12_12$
a, b, c (Å)	146.7,146.7,223.0	146.6,146.6,222.3	146.4,146.4,221.9	146.6,146.6,222.2
α, β, γ (°)	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0
Total observations[a]	3267901 (162701)	4121347 (227678)	1492934 (82381)	1928822 (97105)
Unique reflections[a]	176355 (8640)	54298 (4392)	83874 (4529)	215182 (10580)
Multiplicity[a]	18.5 (18.8)	75.9 (51.8)	17.8 (18.2)	9.0 (9.2)
Mean *I*/σ(*I*)[a]	13.0 (1.6)	15.7 (1.0)	13.6.0 (1.8)	12.7 (1.0)
Completeness (%)[a]	100 (100)	99.8 (99.9)	100 (100)	100 (100)
R_{merge}[a,b]	0.188 (1.913)	0.303 (5.836)	0.217 (1.809)	0.119 (1.992)
R_{meas}[a,c]	0.216 (1.966)	0.305 (5.899)	0.223 (1.859)	0.126 (2.108)
$CC_{½}$[a,d]	0.998 (0.696)	0.999 (0.597)	0.998 (0.624)	1.0 (0.4)
Wilson B value (Å^2)	22.3	68.9	40.6	28.7
Refinement				
Reflections: working/free[e]	176228/8729	-	83781/4278	215057/10878
R_{work}/ R_{free}[f]	0.202/0.223	-	0.179/0.217	0.194/0.218
Ramachandran plot: favoured/allowed/disallowed (%)[g]	97.0/2.8/0.2	-	96.2/3.5/0.3	97.5/2.4/0.1
R.m.s. bond deviations (Å)	0.103	-	0.0091	0.0104
R.m.s. angle deviations (°)	1.40	-	1.34	1.440
No. of protein residues (ranges): Chain A / Chain B	908 (4-911) / 905 (4-908)	-	909 (4-911) / 906 (4-908)	903 (4-877; 884-911) / 900 (4-423; 427-878; 882-908)
No. of ligands[g]**/water molecules/other**[h]	2/893/16	-	2/561/13	2/696/15
Mean *B*-factors: protein/ligand[g]**/ /water/other**[h]**/overall (Å^2)**	39/31/32/42/39	-	50/40/37/59/48	41/37/34/46/40
PDB accession code	6GGY	-	6GH2	6GH3

[a] Values for the outer resolution shell are given in parentheses.

[b] $R_{merge} = \sum_{hkl} \sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl} \sum_i I_i(hkl)$.

[c] $R_{meas} = \sum_{hkl} [N/(N-1)]^{1/2} \times \sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl} \sum_i I_i(hkl)$, where $I_i(hkl)$ is the ith observation of reflection hkl, $\langle I(hkl)\rangle$ is the weighted average intensity for all observations i of reflection hkl and N is the number of observations of reflection hkl.

[d] $CC_{½}$ is the correlation coefficient between symmetry-related intensities taken from random halves of the dataset.

[e] The data set was split into "working" and "free" sets consisting of 95 and 5% of the data, respectively. The free set was not used for refinement.

[f] The R-factors R_{work} and R_{free} are calculated as follows: $R = \sum(|F_{obs} - F_{calc}|)/\sum|F_{obs}|$, where F_{obs} and F_{calc} are the observed and calculated structure factor amplitudes, respectively

[g] refers to ligands bound in the active site i.e. SO_4^{2-}, Glc1P or Man1P depending on the dataset

[h] refers to other ligands, specifically, sulphate bound at a surface site, chloride ions and ethylene glycol molecules from the cryoprotectant

2.2.4. Phosphate recognition by *Ps*LBP

In the SO_4^{2-} complex (Figure 2.12A), SO_4^{2-} which was derived from the precipitant used for crystallisation, occupied a similar position to phosphate and sulphate in other GH94 disaccharide phosphorylase active sites such as in *Cg*CBP (Figure 2.12B). Conserved amino acids within the active sites of other GH94 enzymes can be identified within the *Ps*LBP active site and is in agreement with the conserved amino acid identified from the multiple sequence alignments of GH94 amino acid sequences (Figure 2.13 and Table 2.3). These conserved amino acids are; 1) the WND (W524, N525 and D526) motif locating on the catalytic loop (Figure 2.12, purple); 2) the conserved histidine (H739); 3) conserved arginine and aspartate (R374, D375) located in -1 subsite.

The SO_4^{2-} molecule in *Ps*LBP forms hydrogen bonds with R353, T796 and E782 side chains (Figure 2.12A). From the amino acid sequence alignment of *Ps*LBP, *Cg*CBP and *Lp*SOGP, H739 in *Ps*LBP aligned with the conserved histidine residues in *Cg*CBP and *Lp*SOGP that forms a hydrogen bond with phosphate in the active site of those enzymes (Figure 2.12B). However, the distance between H739 side chain and SO_4^{2-} in *Ps*LBP structure is greater than hydrogen bonding distance, suggesting that the hydrogen bonding interaction between H739 and the phosphate may not be essential for phosphate recognition in *Ps*LBP. This is further supported by three pieces of evidence. Firstly, if the corresponding histidine (H666) in *Cg*CBP was mutated to Asn, a phosphate molecule was still bound to the active site of the enzyme (PDB code 3ACT). Secondly, the phosphate moiety in Glc1P in complex with *Lp*SOGP does not form hydrogen bond with the corresponding histidine (H924) (Figure 2.14B).[32] Lastly, a wild-type CDP from *Ruminococcus albus* has Gln646 instead of the conserved His residue that is found in other GH94 phosphorylases.[57]

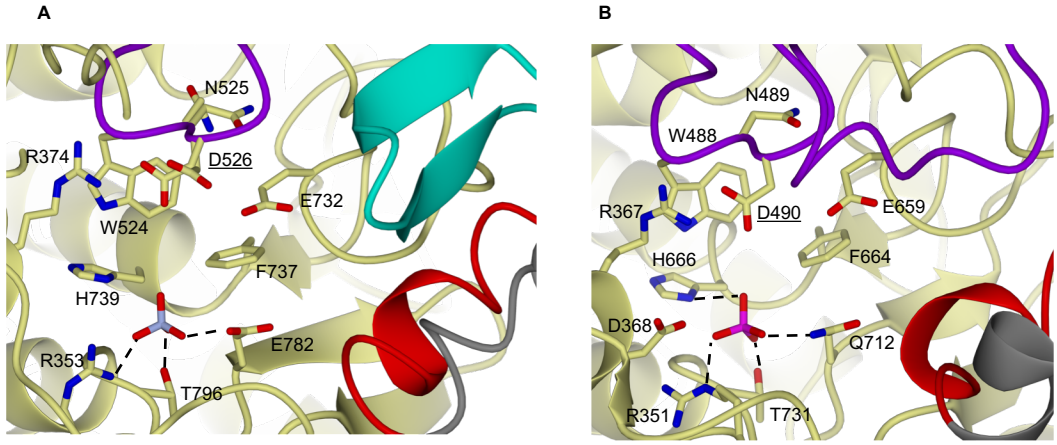


Figure 2.12. Comparison between the active sites of *Ps*LBP and *Cg*CBP. (**A**) *Ps*LBP (SO_4^{2-} complex) and (**B**) *Cg*CBP in complex with phosphate, β-Glc and 1-deoxynojirimycin (PDB code 3QG0). β-Glc and 1-deoxynojirimycin are omitted for clarity. The protein backbones are shown in yellow cartoon representation, whereas the side chains of residues neighbouring of the active sites are shown in stick. Nitrogen and oxygen atoms are coloured in deep blue and red respectively. Sulfur is shown in ice blue in *Ps*LBP and phosphorus in magenta in *Cg*CBP. The conserved catalytic loop is presented in purple and the opposing loop from the adjacent subunit in red. β-hairpin in *Ps*LBP overlapping with the extended catalytic loop of *Cg*CBP is coloured in cyan. In *Ps*LBP, N525 and D526 side chains are presented in two alternate conformations.

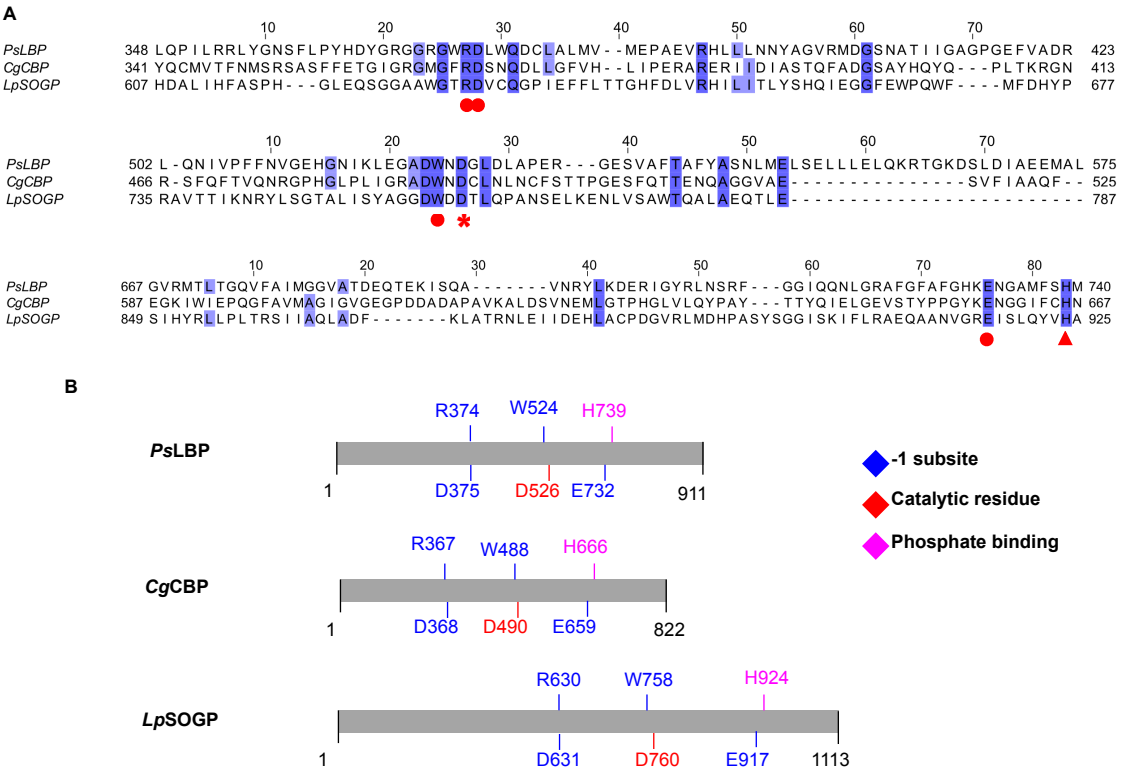


Figure 2.13. Conservation of amino acids constituted -1 subsite in GH94 enzymes. (**A**) Clustal-O multiple sequence alignment of *Ps*LBP, *Cg*CBP, and *Lp*SOGP amino acid sequences. 100% conserved amino acids are highlighted in dark blue. Residues located in -1 subsite are indicated by circles. Catalytic residues are indicated by asterisks. Phosphate binding residues are indicated by triangles. (**B**) Summary of conserved residues found in the amino acid sequence alignment of *Cg*CBP, *Ps*LBP and *Lp*SOGP. The conserved residues are located in -1 subsite (blue), involved in phosphate binding (magenta) or is the predicted catalytic aspartate (red).

Table 2.3. Summary of conserved amino acid residues that are involved in GH94 catalysis, phosphate recognition or located in -1 subsite.

GenBank accession	Catalytic residues	-1 subsite	Phosphate recognition	References	PDB structures
BAA28631.1 (CgCBP)	D490	W488, R367, D368	H666	[28]	2CQS, 2CQT, 3GQ0
ABX41081.1 (LpSOGP)	D760	W758, R630, D631	H924	[32]	5H3Z, 5H40, 5H41, 5H42
BAB71818.1 (RtCDP)	D624	W622, R501, E502	H817	[31]	5NZ7, 5NZ8
BAC87867.1	D492	W490, R349, D350	H644	[30]	1V7V, 1V7W, 1V7X
AAQ20920.1	D490	W488, R367, D368	H666	[56]	3RRS, 3RSY, 3S4A, 3S4B
AAL67138.1	D483	W481, R360, D361	H653	[29]	3QDE
ABD80168.1	D472	W470, R349, (N350)	H626	[33]	4ZLE, 4ZLF, 4ZLG, 4ZLI

2.2.5. Glc1P and Man1P recognition by *Ps*LBP

The overall structure of the SO_4^{2-} complex and the Glc1P complex are very similar (RMSD of 0.23 Å for a dimer on dimer superposition of C-α atoms). Glc1P was bound with the pyranose ring in $^4C_1$ conformation and α-anomeric configuration of phosphate at C1 position, supporting the enzyme specificity for sugar 1-phosphate in α-anomeric configuration (Figure 2.14A). Glc1P is completely buried within a donor subsite (-1 subsite), which is formed entirely within a single subunit of *Ps*LBP. Comparison between the Glc1P complex of *Ps*LBP with that of *Lp*SOGP showed that Glc1P recognition by the two proteins is different. In *Ps*LBP, the hydroxy group on C3 forms a hydrogen bond with the R374 side chain (Figure 2.14A), whereas in *Lp*SOGP, the same hydroxy group forms hydrogen bonds with R630 and D631 (Figure 2.14B). In the *Lp*SOGP structure, D631 is also involved in the recognition of the hydroxy group on C2 via hydrogen bonding (Figure 2.14B), whereas in *Ps*LBP, the hydroxy group forms hydrogen bond with R353. The equivalence of D631 in *Ps*LBP (D375) forms a hydrogen bond with neither the hydroxy groups on C2 or C3, because the distance between the hydroxy groups and D375 is greater than the hydrogen bonding distance (Figure 2.14A). The phosphate moiety in Glc1P also interacts differently with the enzyme active sites. In *Ps*LBP, the phosphate moiety forms hydrogen bonds with the side chains of R353 and E782, similar to those found in SO_4^{2-} complex (Figure 2.14A), whereas in *Lp*SOGP, the phosphate moiety in Glc1P forms hydrogen bonds with S1005 and Y922 (Figure 2.14B).

The Man1P complex (Figure 2.14C) represents the first GP structure in complex with a noncognate sugar 1-phosphate donor. The overall structure of Glc1P complex and Man1P complex are very similar (RMSD of 0.14 Å for a dimer to dimer superposition). Man1P position almost overlaps completely with that of Glc1P, suggesting similar binding mode to Glc1P. However, the hydrogen bond between axial hydroxy group on C2 and R353 side chain cannot form since the distance between the hydroxy group on C2 and R353 side chain is 5.1 Å (in contrast to 3.2 Å between the equatorial hydroxy group on C2 and R353 in Glc1P complex). The loss of hydrogen bond between C2-hydroxy group and R353 side chain in Man1P complex may have a negative impact on the conversion of Man1P and Glc to disaccharide **1**, which is reflected in the reduction in k_{cat}/K_m from 3.07 $s^{-1}mM^{-1}$ (when Glc1P was used as a donor) to 0.02 $s^{-1}mM^{-1}$ (when Man1P was used) that were previously observed in the kinetic studies.

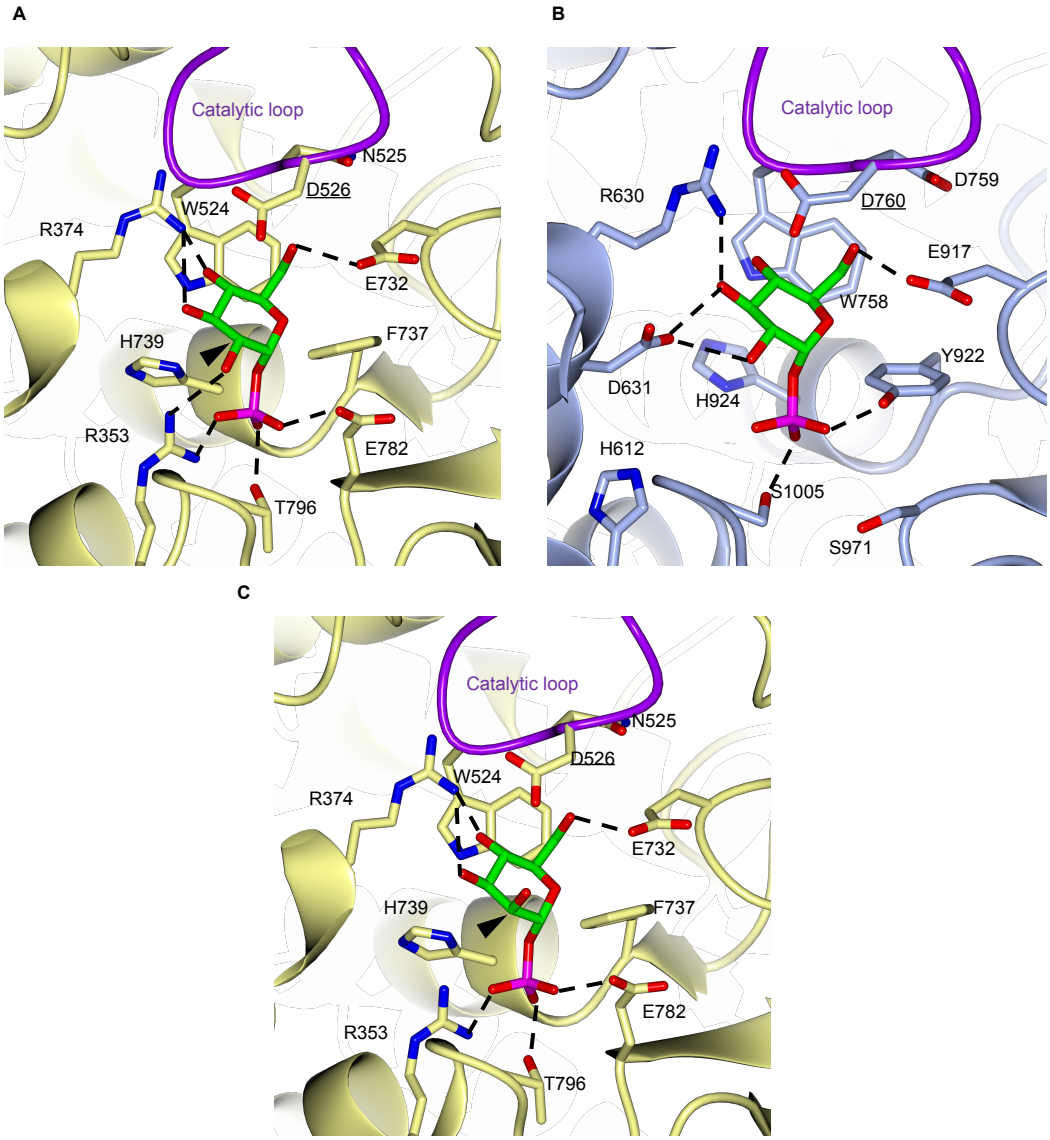


Figure 2.14. Structures of -1 subsites in *Ps*LBP and *Lp*SOGP. **(A)** *Ps*LBP in complex with Glc1P. **(B)** *Lp*SOGP in complex with Glc1P. **(C)** The active site of *Ps*LBP in complex with Man1P. The C2 position on the pyranose ring is indicated with black arrow head. The protein backbone is shown in yellow cartoon representation and the neighbouring side chains in stick. Nitrogen and oxygen atoms are shown in blue and red respectively. The conserved catalytic loop is presented in purple. The catalytic residue (D526) is underlined. Carbon atoms in the ligands are shown in green. Omit *mFobs-dFcalc* difference electron density maps for Glc1P and Man1P can be found in Figure 2.15.

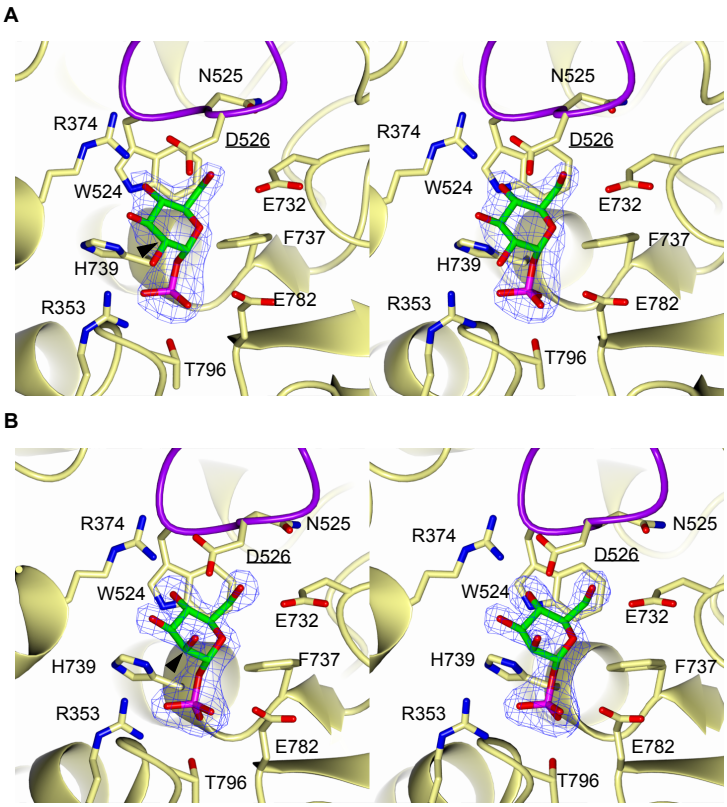


Figure 2.15. Stereo views of the active sites of *Ps*LBP. (**A**) Glc1P complex. (**B**) Man1P complex. C2 positions on the pyranose rings are indicated by black arrow heads. The catalytic loop is presented in purple. The predicted catalytic residue (D526) is underlined. Omit *mFobs-dFcalc* difference electron density maps (~2.0 Å resolution; contoured at ~3σ) are shown in blue mesh and were generated for the bound ligands using phases calculated from the final model without the ligands after the application of small random shifts to the atomic coordinates, re-setting temperature factors, and re-refining to convergence.

2.2.6. Catalytic loop in GH94 enzymes

The catalytic loop (purple, Figure 2.14 and 2.15) is a highly conserved feature among GH94 structures. In *Ps*LBP, this loop consists of a WND motif (W524, N525 and D526), with D526 as a predicted catalytic residue. The tryptophan residue (W524) is structurally conserved among GH94 members and provides a hydrophobic platform for the binding of the donor.[31] Comparison between the *Cg*CBP and *Ps*LBP structures showed that the *Cg*CBP active site is slightly more closed in comparison to *Ps*LBP due to several different structural features. Firstly, the length of the *Ps*LBP catalytic loop is 10-amino acid residues shorter than that of *Cg*CBP, and only partially occludes the active site (Figure 2.16C and D, purple). In contrast, the extended catalytic loop in *Cg*CBP forms a lid-like structure that extends over the active site (Figure 2.16A and B, purple). Secondly, the position of the 'adjacent loop' runs in parallel to a significant proportion of catalytic loop in *Cg*CBP, forming a zipper-like interface that further encloses the *Cg*CBP active site (Figure 2.16A and B, brown). The absence of the extended catalytic loop in *Ps*LBP may be compensated by the presence of a β-hairpin in the β-sandwich domain of the adjacent subunit, which is six residues longer than the equivalent element in *Cg*CBP. This forms a β-hairpin 'gate', which overlaps with the extended catalytic loop in *Cg*CBP (Figure 2.16C and D, cyan). Moreover, the opposing loop in *Cg*CBP projects into the active site more than that of *Ps*LBP (Figure 2.16, red), whilst in *Ps*LBP, the opposing and adjacent loops are located further away from the active site with respect to those in the *Cg*CBP structure. The relatively "open" state of the *Ps*LBP active site may be representative of the initial binding of the sugar 1-phosphate donor before the synthetic reaction occurs or the final stage of phosphorolysis where sugar 1-phosphate is about to be released. On the other hand, the relatively "closed" state observed in *Cg*CBP structure likely represents the intermediate state of the catalysis where the glycosidic bond is either being broken during phosphorolysis or formed during the synthetic reaction.

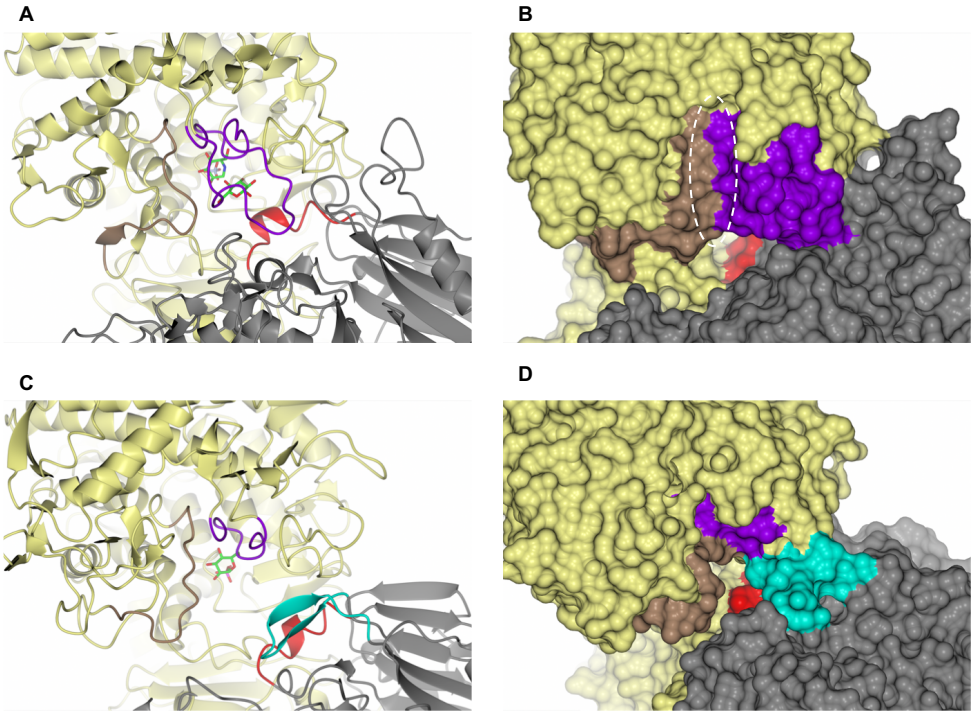


Figure 2.16. Comparison between loops in *Cg*CBP in complex with β-D-Glc, 1-deoxynojirimycin and phosphate (PDB code 3QG0) and Glc1P complex of *Ps*LBP. (**A**) *Cg*CBP structure shown in cartoon representation. (**B**) the equivalent view of *Cg*CBP depicted as a molecular surface. (**C**) *Ps*LBP in complex with Glc1P shown in cartoon representation. (**D**) the equivalent view of *Ps*LBP structure depicted as a molecular surface. The zipper-like interface between the catalytic and the adjacent loops in *Cg*CBP is indicated by white dotted oval. Note that the β-hairpin gate of *Ps*LBP occupies a similar relative space to the end of the longer catalytic loop in *Cg*CBP.

2.2.7. STD NMR studies on the interactions between *Ps*LBP and its ligands (detailed experiments and results are reported in Mr Samuel Walpole's Thesis dissertation at University of East Anglia)

To complement crystallographic data on the binding of Glc1P and Man1P to *Ps*LBP, STD NMR studies were performed to study the enzyme-ligand interaction in solution. In the study of Glc1P, all the ligand protons received strong saturation from the protein in particular at H4 and H6 (Figure 2.17A). This suggests that the H4/H6 area is buried within the binding cavity. This is in very good agreement with the X-ray structure (Figure 2.14A), where H4/H6 form intimate contacts with the enzyme active site via hydrogen bonds. The interaction between *Ps*LBP and Man1P follows a similar epitope to that in Glc1P (Figure 2.17B), suggesting that Man1P binds to the same subsite as Glc1P, with a rather similar binding mode. This is in agreement with the X-ray structure (Figure 2.14C) where the position of Glc1P and Man1P in the active site are highly similar with similar hydrogen bond contacts except at C2-OH position. Since the only difference between the two substrates is the OH configuration at C2, it follows that a favourable interaction, observed for the native Glc1P, is broken by the inversion in Man1P. This conclusion is further supported by the weaken binding epitope at C2 position observed in Man1P (Figure 2.17B).

Although the crystal structure in complex with other ligands (Glc, G2 and disaccharide **1**) could not be determined, STD NMR studies were performed on these ligands to determine their interaction with *Ps*LBP. In the study of Glc binding (Figure 2.17C), the STD intensities from the α-anomer are very much weaker than those from the β-anomer (data not shown), although similar concentrations of α and β anomers are present. This indicates that the β-anomer is preferentially recognised by the enzyme. This suggests that β-Glc cannot bind to the sugar 1-phosphate (donor) subsite, which accommodates the α-anomer of the sugar 1-phosphate and that β-Glc must occupy a separate subsite (acceptor subsite). This disagrees with the report suggesting competitive inhibition of the substrate, in which Glc inhibits the binding of Glc1P.[14]

STD NMR study of G2 showed that β-anomer is preferentially recognised by the enzyme, albeit both of its α- and β-anomers exists in equilibrium. The study showed that the binding of G2 to the enzyme is directional, with the non-reducing ring of G2 (Glc1) binds to the same subsite as Glc1P/Man1P, whilst the corresponding reducing sugar (Glc2β) binds to the same subsite as Glc. The saturation transferred to Glc2β of G2 is much stronger than that received by Glc1, suggesting that the key interaction

between the ligand and the enzyme is formed with the reducing end (Figure 2.17D). STD NMR performed on disaccharide **1** revealed similar magnitude of the STD intensities to that in G2, suggesting that the affinity of disaccharide **1** and G2 is similar (Figure 2.17D and E). β-anomer of disaccharide **1** is preferentially recognised by the enzyme. The main contacts between disaccharide **1** and *Ps*LBP appears to be with Glc2β rather than the non-reducing end Man (Man1) (Figure 2.17E), suggesting that the Glc2β, which is common to both G2 and disaccharide **1**, is essential for the recognition of disaccharides by *Ps*LBP.

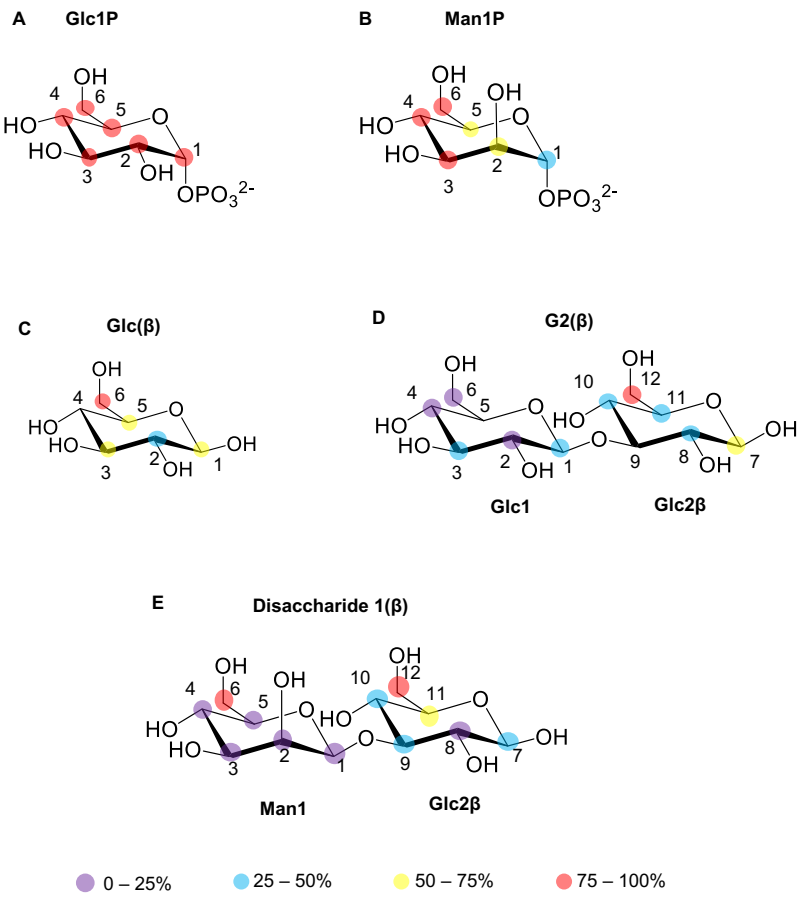


Figure 2.17. STD NMR binding epitopes of ligands in the presence of *Ps*LBP. (**A**) Binding epitopes of Glc1P. (**B**) Binding epitopes for Man1P. (**C**) Binding epitopes for Glc. (**D**) Binding epitopes for laminaribiose (G2). (**E**) Binding epitopes for disaccharide **1**. Figure C D and E show only β-anomer of the ligands. Colours represent normalised values of STD_0 at each position depicted in the structure. For each ligand, the values of STD_0 are normalised against the largest value of binding epitope in the respective molecule.

2.3. Discussion

GPs are attractive biocatalysts for oligo- and poly-saccharide syntheses due to their broad specificity towards acceptor substrates and relatively low cost of donors compared to other enzymes used for glycan syntheses. Therefore, understanding the mechanism of GP action on both natural and noncognate substrates would provide background knowledge that could underpin applications of GPs in carbohydrate synthesis, both in academic and industrial settings. Unlike other conventional substrate screening experiments which have been conducted by various groups on GPs,[14–16,27,31,36,49,58,59] the aim of this study was to pinpoint the mechanism by which *Ps*LBP recognised and utilised Man1P as its noncognate donor, using X-ray crystallography and STD NMR spectroscopy.

The use of GPs for β-(1→3)-mannosylation has been previously conducted using a GH130 β-(1→3)-mannoside phosphorylase (Zg0232) from *Zobellia galactanivorans* DSM 12802, which transfers mannose from Man1P to a variety of sugar acceptors, including a noncognate acceptor, Glc.[49] In contrast, this study demonstrated the relaxed specificity of *Ps*LBP towards the sugar 1-phosphate non cognate donor, Man1P, from which mannose was transferred onto a cognate Glc acceptor, resulting in the production of disaccharide **1**.

Significant interactions between Glc1P and *Ps*LBP, as indicated by the crystal structure and STD NMR analysis, suggests that the specificity of this enzyme towards the donor substrate is more restricted compared to the acceptor site. Therefore, any manipulation to broaden the donor specificity may be challenging. Nevertheless, this work demonstrated a relaxed specificity of *Ps*LBP towards Man1P, which indicates that the alternative configuration of the hydroxy group at C2 on the pyranose ring is tolerated. A strengthening of the interaction of Man1P with the active site could be the strategy to make the production of disaccharide **1** more efficient. However, the enzyme crystal structure showed that the axial configuration of the C2-OH of Man1P is pointing into empty space between the dimer interface, suggesting that a simple mutation approach is unlikely to restore a hydrogen bond with C2-OH on Man1P. Interestingly, GlcN1P was not a substrate for *Ps*LBP, despite having the same configuration as Glc1P at C2. In this case, the C2-OH group is substituted by an NH_2, which likely causes steric and/or electrostatic clashes with R353, thus disfavouring the binding of GlcN1P to the -1 subsite. Size restriction of the substituent at C2 has been reported in *Cg*CBP, which cannot accommodate GlcNAc1P (OH is replaced by CH_3CONH at C2), whereas

*Vp*ChBP can accommodate both GlcNAc1P and Glc1P, despite having the same amino acid for interaction with the hydroxy group at C2.[18] The difference is only in the placement of the Arg side chain that interacts with the C2 substituent, which is more distant in *Vp*ChBP to accommodate a larger group.[28]

Crystallographic structures and STD NMR data generated in this study can be used to explain the *Ps*LBP inactivity on other sugar 1-phosphates that were screened in this study. For instance, a hydrogen bond formed between C4-OH and the side chains of R374 and the strong saturation transfer signal from the protein to the hydrogen on C4 in the STD NMR experiment indicate that C4-OH and its configuration might be crucial for the recognition of Glc1P by *Ps*LBP and therefore any modification at this position may compromise the enzyme activity on the donor. This hypothesis is supported by the fact that *Ps*LBP could not use Gal1P as a sugar donor (Figure 2.4A), which has the C4-OH in an axial rather than equatorial position. The same explanation can be used to explain the lack of *Ps*LBP activity towards GalN1P and GalA1P, both of which are derivatives of Gal1P. *Ps*LBP activity on other donors that have not been screened can be predicted based on the structural information from this study. For instance, glucuronic-acid 1-phosphate (GlcA1P) which has carboxyl group substitution at C6 would likely cause steric and electrostatic clashes with E732 and therefore may not permit binding of GlcA1P to *Ps*LBP. The importance of C6 for binding to *Ps*LBP was evident from the strong STD intensities at this position in both Glc1P and Man1P (Figure 2.17A and B).

Both *Ps*LBP and *Cg*CBP work on disaccharides, but with different linkage specificity (i.e. β-(1→3) vs β-(1→4)). However, each enzyme may employ a different mechanism to restrict the length of the substrate/product. The extended catalytic loop is a unique characteristic to *Cg*CBP that was not found in *Ps*LBP structure, nor in any other characterised disaccharide phosphorylases in GH94 family. It is likely to be involved in the substrate specificity of *Cg*CBP with regards to the degree of polymerisation of the product.[28] In contrast, *Ps*LBP contains a unique β-hairpin 'gate', which when superposed with the *Cg*CBP structure, overlaps with the position of the extended catalytic loop of the latter. Therefore, the β-hairpin gate in *Ps*LBP may perform similar role to that of the *Cg*CBP extended catalytic loop in restricting the degree of polymerisation of the synthetic product.

2.4. Conclusion and future perspectives

In summary, this work provides structural insight into the mechanistic detail of sugar 1-phosphate recognition by GH94 *Ps*LBP, which explains the limited promiscuity toward noncognate sugar 1-phosphates by the enzyme. These findings provide a stepping stone towards the design and engineering of GPs for tolerance towards other noncognate sugar donors, which will help expand the range of GP applications in carbohydrate synthesis. Furthermore, the structural studies of *Ps*LBP and *Cg*CBP unravelled small secondary structures neighbouring the enzyme active sites which are absent from the GH94 oligosaccharide phosphorylases, thus highlighting the features that likely distinguish the substrate preferences of the two groups of enzymes.

2.5. References

[1] S. Kuhaudomlarp, S. Walpole, C. E. M. Stevenson, S. A. Nepogodiev, D. M. Lawson, J. Angulo, R. A. Field, *ChemBioChem* **2018**, *19*, in press.

[2] V. Puchart, *Biotechnol. Adv.* **2015**, *33*, 261–276.

[3] E. C. O'Neill, R. A. Field, *Carbohydr. Res.* **2015**, *403*, 23–37.

[4] G. Pergolizzi, S. Kuhaudomlarp, R. A. Field, E. Kalita, *Protein Pept. Lett.* **2017**, *24*, 696–709.

[5] H. Nakai, M. Kitaoka, B. Svensson, K. Ohtsubo, *Curr. Opin. Chem. Biol.* **2013**, *17*, 301–309.

[6] M. Hiraishi, K. Igarashi, S. Kimura, M. Wada, M. Kitaoka, M. Samejima, *Carbohydr. Res.* **2009**, *344*, 2468–2473.

[7] Y. Yataka, T. Sawada, T. Serizawa, *Langmuir* **2016**, *32*, 10120–10125.

[8] T. Nohara, T. Sawada, H. Tanaka, T. Serizawa, *J. Biomater. Sci. Polym. Ed.* **2017**, *28*, 925–938.

[9] T. Nohara, T. Sawada, H. Tanaka, T. Serizawa, *Langmuir* **2016**, *32*, 12520–12526.

[10] C. Goedl, T. Sawangwan, M. Mueller, A. Schwarz, B. Nidetzky, *Angew. Chemie - Int. Ed.* **2008**, *47*, 10086–10089.

[11] M. Nishimoto, M. Kitaoka, *Biosci. Biotechnol. Biochem.* **2007**, *71*, 2101–2104.

[12] K. Beerens, K. De Winter, D. Van De Walle, C. Grootaert, S. Kamiloglu, L. Miclotte, T. Van De Wiele, J. Van Camp, K. Dewettinck, T. Desmet, *J. Agric. Food Chem.* **2017**, *65*, 6030–6041.

[13] M. Nakajima, H. Toyoizumi, K. Abe, H. Nakai, H. Taguchi, M. Kitaoka, *PLoS One* **2014**, *9*, e92353.

[14] M. Kitaoka, Y. Matsuoka, K. Mori, M. Nishimoto, K. Hayashi, *Biosci. Biotechnol. Biochem.* **2012**, *76*, 343–348.

[15] M. Kitaoka, T. Sasaki, H. Taniguchi, *J. Biochem.* **1992**, *112*, 40–44.

[16] Y.-K. Kim, M. Kitaoka, M. Krishnareddy, Y. Mori, K. Hayashi, *J. Biochem.* **2002**, *132*, 197–203.

[17] M. Reichenbecher, F. Lottspeich, K. Bronnenmeier, *Eur. J. Biochem.* **1997**, *247*, 262–267.

[18] Y. Honda, M. Kitaoka, K. Hayashi, *Biochem. J.* **2004**, *377*, 225–232.

[19] T. Nihira, Y. Saito, M. Nishimoto, M. Kitaoka, K. Igarashi, K. Ohtsubo, H. Nakai, *FEBS Lett.* **2013**, *587*, 3556–3561.

[20] K. De Winter, G. Dewitte, M. E. Dirks-Hofmeister, S. De Laet, H. Pelantová, V. Křen, T. Desmet, *J. Agric. Food Chem.* **2015**, *63*, 10131–10139.

[21] K. De Winter, L. Van Renterghem, K. Wuyts, H. Pelantova, V. Kren, W. Soetaert, T. Desmet, *Adv. Synth. Catal.* **2015**, *357*, 1961–1969.

[22] K. Hamura, W. Saburi, S. Abe, N. Morimoto, H. Taguchi, H. Mori, H. Matsui, *Biosci. Biotechnol. Biochem.* **2012**, *76*, 812–818.

[23] K. Chomvong, V. Kordić, X. Li, S. Bauer, A. E. Gillespie, S.-J. Ha, E. Oh, J. M. Galazka, Y.-S. Jin, J. H. D. Cate, *Biotechnol. Biofuels* **2014**, *7*, 85.

[24] K. Kino, R. Satake, T. Morimatsu, S. Kuratsu, Y. Shimizu, M. Sato, K. Kirimura, *Biosci. Biotechnol. Biochem.* **2008**, *72*, 2415–2417.

[25] J. K. Park, N. O. Keyhani, S. Roseman, *J. Biol. Chem.* **2000**, *275*, 33077–33083.

[26] H. G. Tran, T. Desmet, K. Saerens, H. Waegeman, S. Vandekerckhove, M. D'hooghe, I. Van Bogaert, W. Soetaert, *Bioresour. Technol.* **2012**, *115*, 84–87.

[27] H. Nakai, M. A. Hachem, B. O. Petersen, Y. Westphal, K. Mannerstedt, M. J. Baumann, A. Dilokpimol, H. a. Schols, J. Ø. Duus, B. Svensson, *Biochimie* **2010**, *92*, 1818–1826.

[28] M. Hidaka, M. Kitaoka, K. Hayashi, T. Wakagi, H. Shoun, S. Fushinobu, *Biochem. J.* **2006**, *398*, 37–43.

[29] C. M. Bianchetti, N. L. Elsen, B. G. Fox, G. N. Phillips, *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **2011**, *67*, 1345–1349.

[30] M. Hidaka, Y. Honda, M. Kitaoka, S. Nirasawa, K. Hayashi, T. Wakagi, H. Shoun, S. Fushinobu, *Structure* **2004**, *12*, 937–947.

[31] E. C. O'Neill, G. Pergolizzi, C. E. M. Stevenson, D. M. Lawson, S. A. Nepogodiev, R. A. Field, *Carbohydr. Res.* **2017**, *451*, 118–132.

[32] M. Nakajima, N. Tanaka, N. Furukawa, T. Nihira, Y. Kodutsumi, Y. Takahashi, N. Sugimoto, A. Miyanaga, S. Fushinobu, H. Taguchi, H. Nakai, *Sci. Rep.* **2017**, *7*, 42671.

[33] Y.-W. Nam, T. Nihira, T. Arakawa, Y. Saito, M. Kitaoka, H. Nakai, S. Fushinobu, *J. Biol. Chem.* **2015**, *290*, 18281–18292.

[34] S. Fushinobu, M. Hidaka, A. M. Hayashi, T. Wakagi, H. Shoun, M. Kitaoka, *J. Appl. Glycosci.* **2011**, *58*, 91–97.

[35] M. R. M. De Groeve, M. De Baere, L. Hoflack, T. Desmet, E. J. Vandamme, W. Soetaert, *Protein Eng. Des. Sel.* **2009**, *22*, 393–399.

[36] M. R. M. De Groeve, G. H. Tran, A. Van Hoorebeke, J. Stout, T. Desmet, S. N. Savvides, W. Soetaert, *Anal. Biochem.* **2010**, *401*, 162–167.

[37] M. R. M. De Groeve, T. Desmet, W. Soetaert, *J. Biotechnol.* **2011**, *156*, 253–260.

[38] M. Mayer, B. Meyer, *Angew. Chemie - Int. Ed.* **1999**, *38*, 1784–1788.

[39] M. Mayer, B. Meyer, *J. Am. Chem. Soc.* **2001**, *123*, 6108–6117.

[40] C. D. Owen, L. E. Tailford, S. Monaco, T. Šuligoj, L. Vaux, R. Lallement, Z. Khedri,

H. Yu, K. Lecointe, J. Walshaw, S. Tribolo, M. Horrex, A. Bell, X. Chen, G. L. Taylor, A. Varki, J. Angulo, N. Juge, *Nat. Commun.* **2017**, *8*, 2196.

[41] J. C. Muñoz-García, M. J. García-Jiménez, P. Carrero, A. Canales, J. Jiménez-Barbero, M. Martín-Lomas, A. Imberty, J. L. de Paz, J. Angulo, H. Lortat-Jacob, P. M. Nieto, *Glycobiology* **2014**, *24*, 1004–1009.

[42] Y. Yuan, X. Wen, D. A. R. Sanders, B. M. Pinto, *Biochemistry* **2005**, *44*, 14080–14089.

[43] Y. Yuan, D. W. Bleile, X. Wen, D. A. R. Sanders, K. Itoh, H. W. Liu, B. M. Pinto, *J. Am. Chem. Soc.* **2008**, *130*, 3157–3168.

[44] M. G. Szczepina, R. B. Zheng, G. C. Completo, T. L. Lowary, B. M. Pinto, *ChemBioChem* **2009**, *10*, 2052–2059.

[45] J. Angulo, B. Langpap, A. Blume, T. Biet, B. Meyer, N. Rama Krishna, H. Peters, M. M. Palcic, T. Peters, *J. Am. Chem. Soc.* **2006**, *128*, 13529–13538.

[46] S. M. L. Gagnon, M. S. G. Legg, N. Sindhuwinata, J. A. Letts, A. R. Johal, B. Schuman, S. N. Borisova, M. M. Palcic, T. Peters, S. V Evans, *Glycobiology* **2017**, *27*, 966–977.

[47] E. C. O'Neill, A. M. Rashid, C. E. M. Stevenson, A.-C. Hetru, A. P. Gunning, M. Rejzek, S. A. Nepogodiev, S. Bornemann, D. M. Lawson, R. A. Field, *Chem. Sci.* **2014**, *5*, 341–350.

[48] N. S. Berrow, D. Alderton, S. Sainsbury, J. Nettleship, R. Assenberg, N. Rahman, D. I. Stuart, R. J. Owens, *Nucleic Acids Res.* **2007**, *35*, e45.

[49] F. N. Awad, P. Laborda, M. Wang, A. M. Lu, Q. Li, Z. P. Cai, L. Liu, J. Voglmeir, *Biochim. Biophys. Acta - Gen. Subj.* **2017**, *1861*, 3231–3237.

[50] P.-E. Jansson, J. Lindberg, G. Widmalm, *Acta Chem. Scand.* **1993**, *47*, 711–715.

[51] R. Stenutz, in *Ioinformatics for Glycobiology and Glycomics: An Introduction* (Eds.: C.-W. von der Lieth, T. Lütteke, M. Frank), Wiley-Blackwell, Hoboken, **2009**, pp. 311–320.

[52] P. Colson, R. R. King, *Carbohydr. Res.* **1976**, *47*, 1–13.

[53] M. U. Roslund, E. Säwén, J. Landström, J. Rönnols, K. H. M. Jonsson, M. Lundborg, M. V. Svensson, G. Widmalm, *Carbohydr. Res.* **2011**, *346*, 1311–1319.

[54] M. U. Roslund, P. Tähtinen, M. Niemitz, R. Sjöholm, *Carbohydr. Res.* **2008**, *343*, 101–112.

[55] E. Krissinel, *Nucleic Acids Res.* **2015**, *43*, W314–W319.

[56] A. Van Hoorebeke, J. Stout, J. Kyndt, M. De Groeve, I. Dix, T. Desmet, W. Soetaert, J. Van Beeumen, S. N. Savvides, *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **2010**, *66*, 346–351.

[57] T. Sawano, W. Saburi, K. Hamura, H. Matsui, H. Mori, *FEBS J.* **2013**, *280*, 4463–4473.

[58] D. Aerts, T. F. Verhaeghe, B. I. Roman, C. V. Stevens, T. Desmet, W. Soetaert, *Carbohydr. Res.* **2011**, *346*, 1860–1867.

[59] T. Nihira, Y. Saito, M. Kitaoka, M. Nishimoto, K. Otsubo, H. Nakai, *Carbohydr. Res.* **2012**, *361*, 49–54.

Chapter 3

Identification of *Euglena gracilis* β-(1→3)-glucan phosphorylase and establishment of a new glycoside hydrolase (GH) family GH149*

*A major part of this chapter is an adaptation of work that has been originally published in the Journal of Biological Chemistry[1] and appears with permission.

3.1. Introduction

Glycoside phosphorylases (GPs) (EC 2.4.x.x) carry out the phosphorolysis of sugar polymers to produce the corresponding sugar 1-phosphates and shortened glycan polymer chains. However, the reaction is freely reversible *in vitro*, enabling the production of oligosaccharides from a simple sugar 1-phosphate and suitable carbohydrate acceptors with strict regio-, stereo- and chain length specificity.[2,3] Despite the potential use of GPs for oligosaccharide production, a relatively small number of these enzymes have been identified, limiting their range of applications. Nevertheless, the number of enzyme accessions are gradually increasing due to the substantial increase in available genome sequences and research efforts into characterisation of novel GP activities.

Linear β-(1→3)-D-glucan polysaccharides are found across the prokaryotes and eukaryotes.[4] The most well-known examples are bacterial curdlan,[5] plant callose and microalgal (Euglena) paramylon, all of which are high molecular weight (MW) and water insoluble. Paramylon, curdlan and related compounds have been used in a wide range of applications: as part of anti-tumour and anti-HIV treatments,[6,7] as immune stimulants,[8] and as alternative materials to fossil fuel products.[9] While the water insolubility of paramylon enables it to accumulate to up to 90% of Euglena cell mass,[10] it creates complications for subsequent chemical modification, such as sulfation.[11] Therefore, research into the generation of linear, soluble β-(1→3)-D-glucan polymers *in vitro* is of interest. Enzymatic transglycosylation using a mutant β-(1→3)-D-endoglucanase[12] or a partially purified β-(1→3)-D-glucan phosphorylase preparation from *Euglena gracilis* have been used for β-(1→3)-D-glucan production.[13] The latter approach can be coupled with sucrose phosphorylase to generate β-(1→3)-D-glucan disaccharide, laminaribiose, from sucrose.[14] Enzymatic synthesis of β-(1→3)-D-glucan using Euglena phosphorylases is attractive because of the relatively cheap substrates required (D-glucose (Glc) and α-D-glucose 1-phosphate (Glc1P)). However, the generation of the Euglena phosphorylase catalyst is time consuming and laborious because it requires large volume of Euglena cell culture that takes several days to grow. Moreover, the partial purification of the phosphorylase may result in loss of protein yield and attenuation of the activity.

Continuing the efforts to explore the glycobiology of *Euglena gracilis*[15–19] and the investigation into GPs and their applications in enzymatic synthesis [20,21] from the

previous works in Rob Field's group, I focused on the established capability of the organism for β-(1→3)-D-glucan biochemistry, in particular its known β-(1→3)-D-glucan phosphorylase activities (EC 2.4.1.31) (Figure 3.1). The enzymes have been categorised into two subgroups, based on the chain length specificity for their acceptor substrates; laminaribiose (β-(1→3)-glucobiose) phosphorylases (LBPs) or laminarin (laminaridextrin) phosphorylases (LDPs). LBP was previously considered to catalyse only a reversible phosphorolysis of laminaribiose,[22] whereas LDPs have strong substrate preferences for β-(1→3)-D-glucans with degree of polymerisation (DP) 3 or greater.[23] LBP activities were originally discovered in *Euglena gracilis* [24,25] and *Astasia ocellate,*[26] and subsequently in bacteria, *Paenibacillus* sp. YM-1 [27] and *Acholeplasma laidlawii.*[28] In contrast, LDP activities have only been described in the Euglenozoan *E. gracilis* [29] and in the heterokonts *Ochromonas malhemensis* [30] and *Ochromonas danica.*[31] Despite the various reported β-(1→3)-D-glucan phosphorylase activities, LBP sequences are the only β-(1→3)-D-glucan phosphorylases that have been successfully cloned to date; they have been classified as members of glycoside hydrolase family GH94 in the CAZy database (www.cazy.org).[32] However, the eukaryotic phosphorylases have not, to date, been classified due to the lack of their sequence information, which compromises further biochemical and structural studies of this fascinating group of enzymes.

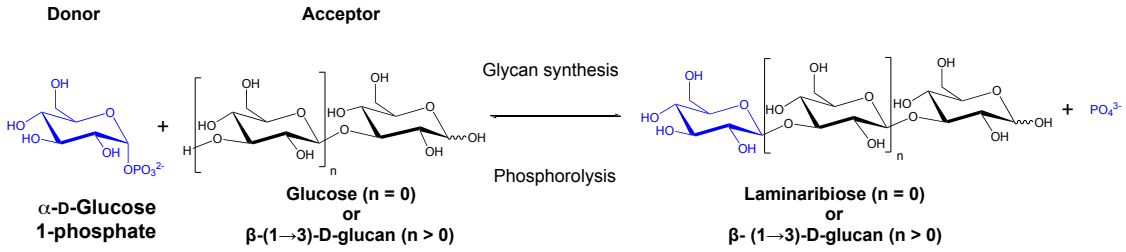


Figure 3.1. Reaction carried out by β-(1→3)-D-glucan phosphorylase.

The physiological role of eukaryotic β-(1→3)-D-glucan phosphorylases remains elusive. In contrast, the biological function of the bacterial phosphorylases can be predicted from the identity of other genes with which they are clustered within the genome. LBP from *Paenibacillus* sp. YM-1 (*Ps*LBP) was hypothesised to be involved in degradation of laminaribiose, based on the position of LBP genetic locus next to ATP-binding cassette (ABC) sugar transporter components.[27] Organisation of GP encoding genes with other sugar transporters have also been observed. For instance, a sucrose phosphorylase gene in human gut *Bifidobacterium lactis*, a sucrose transporter and a transcriptional regulator from a sucrose-utilisation gene cluster which are

transcriptionally upregulated upon induction by sucrose and raffinose.[33] A β-(1→4)-D-mannosyl-*N*-acetyl-D-glucosamine phosphorylase was reported as part of a gene cluster involved in *N*-glycan metabolism in *Bacteroidetes thetaiotaomicron* VPI-5482 and was shown to be involved in the phosphorolysis of β-(1→4)-D-mannosyl-*N*-acetyl-D-glucosamine into α-D-mannose 1-phosphate and *N*-acetyl-D-glucosamine.[34] Two GH130 β-(1→2)-mannoside phosphorylases (Teth514_1788 and Teth514_1789) from *Thermoanaerobacter* sp. X-514 showed different chain length specificities on phosphorolysis of β-(1→2)-oligomannan. The genes encoding Teth514_1788 and Teth514_1789 are located in a predicted GDP-D-mannose biosynthetic gene cluster. This finding demonstrates a possible role of Teth514_1788 and Teth514_1789 in a novel GDP-D-mannose salvage pathway, in contrast to the simple glycan degradative roles that have been predicted for other GPs.[35]

In this chapter, a new family of GPs was uncovered, designated GH149. A eukaryotic *Euglena gracilis* phosphorylase 1 (EgP1) and its bacterial orthologue (Pro_7066) have been characterised and confirmed as β-(1→3)-D-glucan phosphorylase, albeit with distinct substrate chain length preferences. Structural studies of EgP1 and Pro_7066 were performed to study the active site architecture governing their activities. The crystal structures of Pro_7066 were solved, which revealed a conservation of key amino acid residues previously identified in GH94 enzymes and conservation of overall domain organisation. However, Pro_7066 contained two additional domains (Dom 1 and 2) flanking the catalytic domains. Pro_7066 structure in complex with laminarihexaose (G6) revealed a unique oligosaccharide surface binding site (SBS) that is likely involved in substrate targeting and disruption. Although crystallisation of EgP1 was unsuccessful, homology modelling of EgP1 was performed to reveal the same conservation of the active site and key amino acid residues in the sugar donor subsite, and revealed the lack of SBS site, which may explain the difference in the substrate chain length preferences observed in Pro_7066 and EgP1. GH149 family contains several hundred sequences, most of which belong to Gram-negative marine bacteria from the phyla Proteobacteria and Bacteroidetes. The Bacteroidetes *GH149* genes map to previously predicted polysaccharide utilisation loci (PULs),[36,37] strongly suggesting a role for these GH149 enzymes in polysaccharide degradation by marine bacteria. Possible evolutionary relationships between the bacterial GH149 and the eukaryote GH149 are discussed based on phylogenetic analysis of the GH149 amino acid sequences and GC content analysis of the *GH149* genes.

3.2. Results

3.2.1. Identification of Euglena phosphorylase candidates through proteomic analysis.

It has previously been hypothesised that Euglena β-(1→3)-D-glucan phosphorylase sequences may belong to GH94 family based on an activity somewhat similar to the GH94 bacterial LBPs.[27,28] To test this hypothesis, 20 characterised GH94 amino acid sequences were used as queries (Table 3.1) to interrogate the translated Euglena transcriptome[15] using tBLASTn (threshold = 0.0001). No sequences were recovered, suggesting that the Euglena phosphorylases may not be closely related to the members of GH94 family as previously hypothesised. This is supported by previous comprehensive carbohydrate- active enzyme (CAZyme) annotations on Euglena transcriptomes,[15,38] which did not identify any GH94 candidates.

Table 3.1. Characterised GH94 sequences that were used as queries for BLAST analysis

Sequence names*	GenBank accessions
CBP from *Cellvibrio gilvus* ATCC 13127	BAA28631.1
ChBP from *Vibrio proteolyticus*	BAC87867.1
CBP from *Cellulomonas uda*	AAQ20920.1
CBP from *Ruminiclostridium thermocellum*	AAL67138.1
CBP from *Saccharophagus degradans*	ABD80168.1
CDP from *Ruminiclostridium thermocellum*	BAB71818.1
LBP from *Halorhabdus tiamatea*	WP_020936056.1
CBP from *Clostridium stercorarium*	AAC45510.0
CDP from *Clostridium stercorarium*	AAC45511.1
CBP from *Cellvibrio gilvus* ATCC 13127	BAA28631.1
CBP from *Thermotoga maritima* MSB8	AAD36910.1
CBP from *Thermotoga neapolitana*	AAB95491.2
N,N'-diacetylchitobiose phosphorylase from *Vibrio furnissii*	AAG23740.1
ChBP from *Vibrio proteolyticus*	BAC87867.1
LBP from *Acholeplasma laidlawii* PG-8A	ABX81345.1
LBP from *Paenibacillus* sp. YM1	BAJ10826.1
N,N'-diacetylchitobiose phosphorylase from *Ruminococcus albus*	WP_013499018.1
CBP from *Halorhabdus tiamatea* SARL4B	CCQ33375.1
CBAP from *Neurospora crassa* OR74A	EAA28929.1
CBAP from *Xanthomonas campestris*	WP_011039146.1

*CBP = cellobiose phosphorylase, ChBP = chitobiose phosphorylase, CDP = cellodextrin phosphorylase, LBP = laminaribiose phosphorylase, CBAP = cellobionic acid phosphorylase

To identify the Euglena proteins with β-(1→3)-D-glucan phosphorylase activity, a fractionation of Euglena protein extracts was performed followed by proteomic analysis of the fractions with active phosphorylase activity. Cell-free extracts of 7-day old, dark-grown *E. gracilis* were fractionated through Anion Exchange Chromatography (AIEX) as the first step of partial purification of β-(1→3)-D-glucan phosphorylase, based on previous publications.[22,25] The AIEX fractions were screened for phosphorylase activity in a glycan synthetic reaction (Figure 3.1) using a phosphate release assay with

Glc and Glc1P as an acceptor and a donor respectively. The activity of the phosphorylase was detected in two sets of fractions, eluted at 18% w/v NaCl (Peak I) and 24-34 % w/v NaCl (Peak II) (Figure 3.2A). The oligosaccharide products of the glycan synthetic reactions were analysed by TLC, which showed production of a disaccharide only from the Peak II-catalysed reactions (Figure 3.2B). The phosphorylase activity in Peak II could use laminaribiose (G2) and laminaritriose (G3) as acceptors (Figure 3.2C), forming higher DP oligosaccharides, which would suggest that this fraction contained an LDP activity.

Iminosugars such as 1-deoxynojirimycin (DNJ) show weak competitive inhibition (K_i = 570 μM, where K_i represent the inhibition against cellobiose) against the phosphorylase activity of a GH94 cellobiose phosphorylase (CBP) from *Cellvibrio gilvus*.[39] To investigate whether there is an inhibition of the Euglena phosphorylase(s) by DNJ, the partially purified enzyme activity in Peak II was incubated with Glc and Glc1P in the presence of DNJ. DNJ showed weak inhibition in 100-1000 μM range in the glycan synthetic reaction (Figure 3.2C, and D), in keeping with the inhibition reported for the *C. gilvus* CBP.

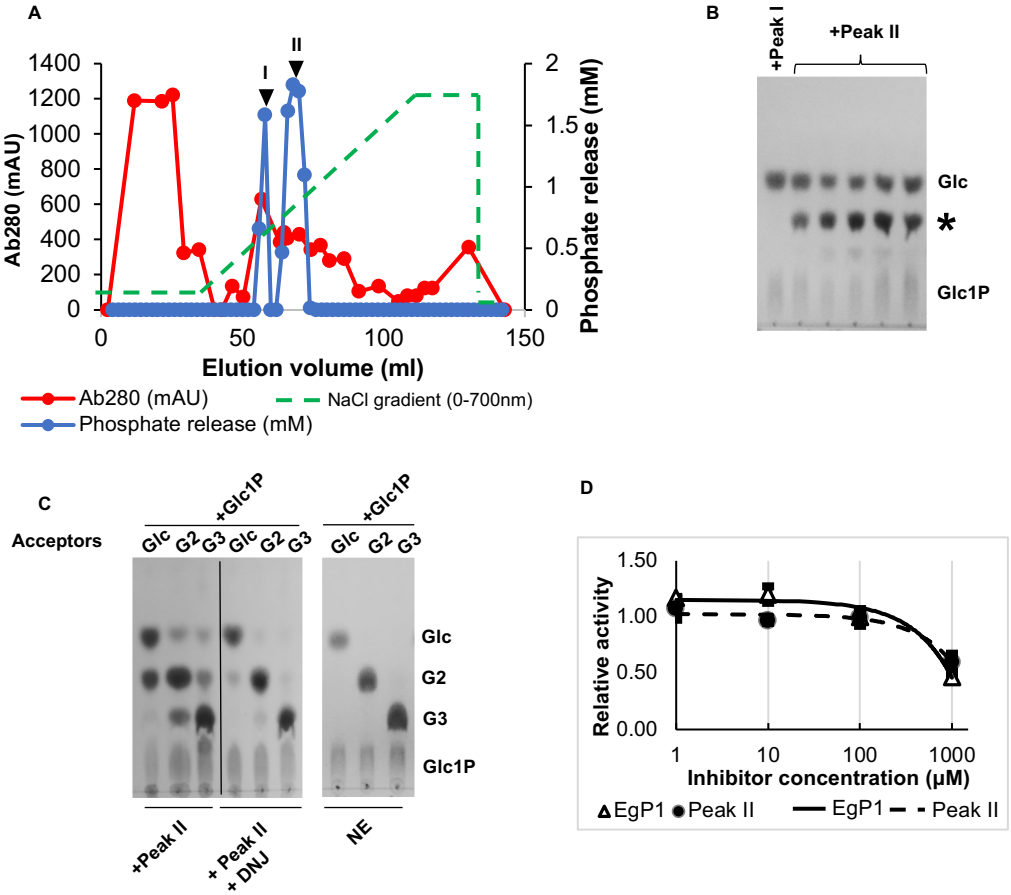


Figure 3.2. Partial purification of Euglena phosphorylase from the Euglena cell-free extract. (**A**) An overlay of AIEX elution profile (red) and the amount of phosphate release by the phosphorylase activity in each fraction in the phosphate release assay (blue). (**B**) TLC analysis of Euglena phosphorylase mediated reaction between Glc and Glc1P. Disaccharide product was detected (asterisk). (**C**) TLC analysis of Euglena phosphorylase (Peak II) mediated reactions between Glc1P and laminaribiose (G2), or laminaritriose (G3) as substrates in the presence (+DNJ, 1000 μM) and absence of DNJ. (**D**) Relative glycan synthetic activity carried out by partially purified phosphorylase (Peak II) using Glc and Glc1P as substrates (dotted line) or recombinant EgP1 (solid line) in the presence of DNJ. Glc = D-glucose, Glc1P = α-D-glucose 1-phosphate, DNJ = 1-deoxynojirimycin. NE = no enzyme control.

Immobilised glycomimetic molecules such as *N*-butyldeoxynojirimycin (NB-DNJ) have been used in affinity-enrichment proteomics to identify iminosugar-interacting proteins from mammalian tissue.[40] Affinity-enrichment proteomics was performed using *N*-5-carboxypentyl-DNJ immobilised on agarose beads as a glycomimetic probe to identity DNJ-interacting phosphorylase candidates in the Euglena Peak II sample (Figure 3.3). In brief, the sample was applied to the *N*-5-carboxypentyl-DNJ matrix (the matrix was synthesised by Dr Martin Rejzek, John Innes Centre) followed by washing with buffer to remove unbound proteins. The binding proteins were eluted with DNJ-containing buffer.

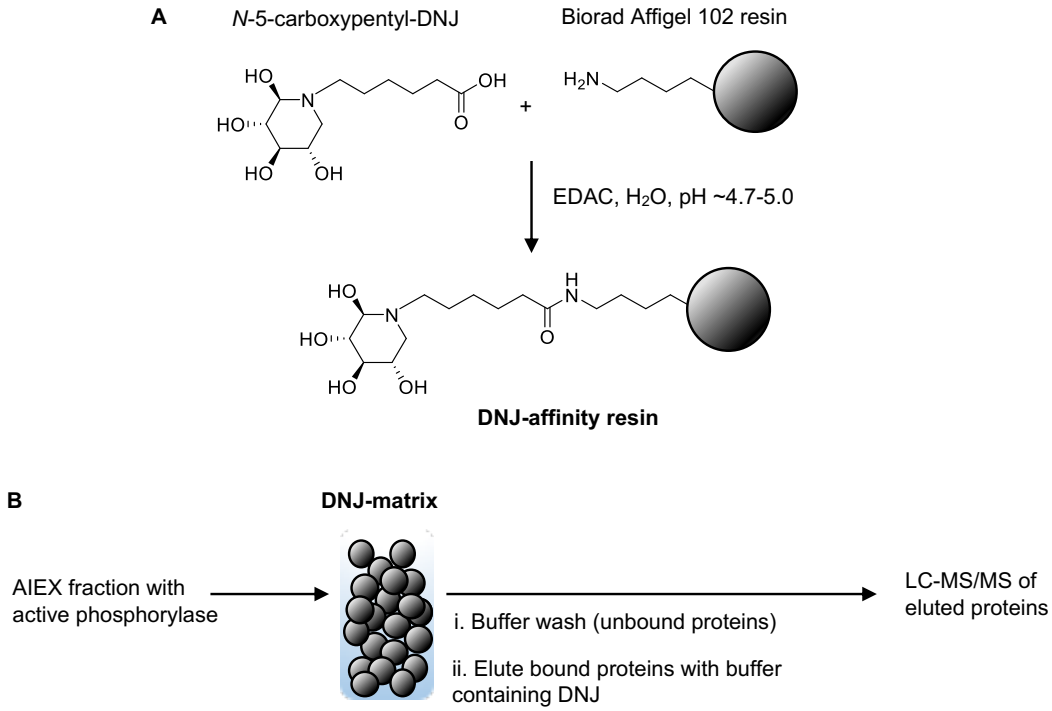


Figure 3.3. Glycomimetic probe for identification of DNJ-interacting phosphorylase candidates in Euglena protein lysate. (**A**) Synthesis of DNJ-affinity resin (performed by Dr Martin Rejzek, John Innes Centre). (**B**) DNJ-affinity enrichment of Peak II sample for proteomic analysis.

The eluted protein fractions were subjected to proteomic LC-MS/MS analysis (performed by Dr Gerhard Saalbach, John Innes Centre). From a total of 253 proteins with unique peptide counts > 3 and 100% protein identification probability, 239 were identified in the translated transcriptome database of *Euglena gracilis*[15] using Mascot (Matrix Science, Boston MA, USA). Putative functions of these proteins were assigned by comparison to protein sequences with annotated functions in the non-redundant database NCBI by BLASTP pairwise alignment (Table 3.2 and Appendix 3). One protein (m. 14570_dark) showed similarity to a deposited sequence, which was annotated by the depositor as a laminaridextrin phosphorylase-like protein from *Euglena gracilis* (BAV19384.1, E-value = 0, %identity = 53). The calculated MW of m.14570_dark is 130 kDa, which is similar to what has been reported for the partially purified Euglena LBP.[25] The translated Euglena transcriptome was interrogated using the m.14570_dark sequence (designated *Euglena gracilis* phosphorylase 1; EgP1) as a BLASTP query and three further sequences were identified with >70% sequence identity to m.14570_dark (m.14571_dark, 14576_dark and 14578_dark). These three candidate phosphorylases were designated EgP2, EgP3 and EgP4, respectively.

Table 3.2. Example of protein hits from Mascot search on the DNJ-eluted protein fraction. EgP1 candidate is coloured in red.

Euglena sequences identified from Mascot search	MW (kDa)	Number of exclusive unique peptide count	BLAST hit
m.12864_dark	107	17	hypothetical protein [*Paramecium tetraurelia* strain d4-2]
m.6710_dark	107	13	glycine dehydrogenase [*Spizellomyces punctatus* DAOM BR117]
m.15695_dark	114	8	uncharacterized protein LOC105157340 [*Sesamum indicum*]
m.12386_dark	123	4	PREDICTED: isoleucine--tRNA ligase, cytoplasmic [*Equus caballus*]
m.1626_dark	127	8	hybrid cluster protein, putative [*Bodo saltans*]
m.14570_dark	127	36	laminaridextrin phosphorylase-like protein [*Euglena gracilis*]
m.13041_light	128	14	Pyrophosphate-fructose 6-phosphate 1-phosphotransferase subunit beta 2 [*Symbiodinium microadriaticum*
m.11124_light	129	21	PREDICTED: pyruvate carboxylase, mitochondrial-like [*Limulus polyphemus*]
m.14_light	130	50	malate synthase-isocitrate lyase [*Euglena gracilis*]

3.2.2. Recombinant protein production and biochemical characterisation of EgP1

To confirm the enzymatic activity of the EgP1-4, the recombinant proteins were produced for biochemical characterisation. It was anticipated that heterologous expression of Euglena CAZymes might not be straightforward, given the problems encountered by the Suzuki group when trying to achieve productive expression of Euglena β-(1→3)-glucanase.[41] An *E. coli* protein expression system was chosen for its simplicity and relatively short protein production time. Coding sequences were amplified from synthetic DNA of EgP1-4 and individually cloned, into the pOPINF plasmid vectors,[42] which were transformed into *E. coli* (Lemmo (DE3)) and expression induced using 0.2 mM IPTG. Only expression of EgP1 was detectable. Initially, EgP1 formed insoluble aggregates (IS-EgP1), which were purified by washing with buffer containing 6 M urea (Figure 3.4A). The washed IS-EgP1 suspension (Figure 3.4B) unexpectedly showed phosphorylase activity when assayed in the glycan synthetic reaction in the presence of Glc and Glc1P as substrates (Figure 3.4C).

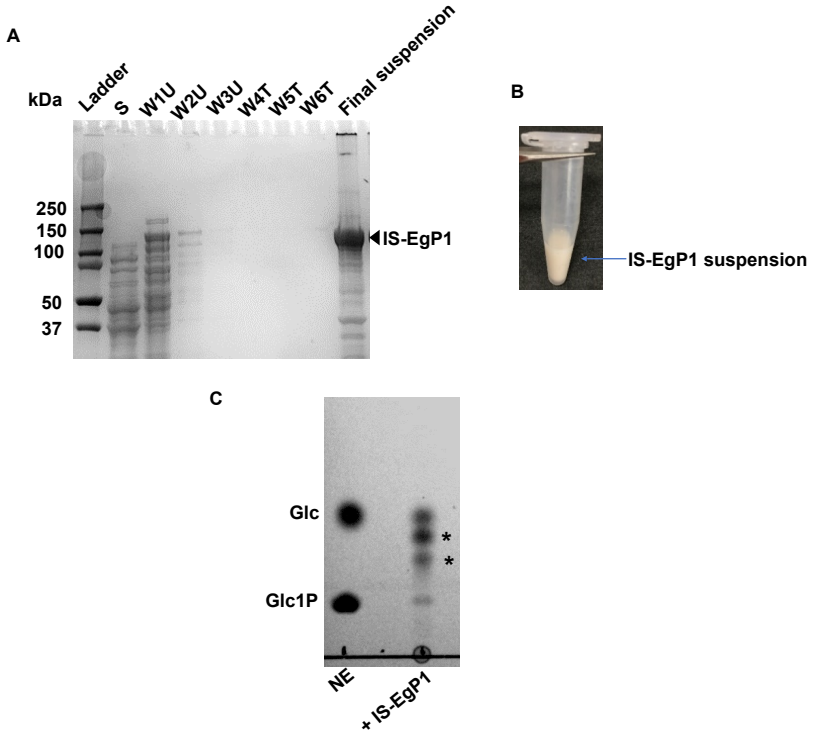


Figure 3.4. Expression and characterisation of insoluble recombinant EgP1 (IS-EgP1). (**A**) SDS-PAGE analysis of IS-EgP1 after purification by washing with 50 mM Tris-HCl pH 7.5 containing 6 M urea (W1U-W3U) followed by washing with buffer alone (W4T-W6T, 50 mM Tris-HCl pH 7.5). (**B**) The final suspension of IS-EgP1 in 50 mM Tris-HCl pH 7.5 buffer. (**C**) TLC analysis of the glycan synthetic reaction carried out by IS-EgP1 in the presence of 10 mM Glc and 20 mM Glc1P. * = oligosaccharide products from the glycan synthetic reaction.

Analysis of the EgP1 amino acid sequence using TargetP version 1.1 (http://www.cbs.dtu.dk/services/TargetP/) predicted the presence of an N-terminal mitochondrial targeting peptide at residues 1-31 (Reliability class = 2), which was hypothesised to be the cause of protein aggregation. A truncated version of the EgP1 coding sequence was re-cloned without the predicted target peptide, but with an N-terminal His_6-tag and the recombinant protein that expressed as previously described. A soluble EgP1 recombinant protein was detected (EgP1), which was purified by IMAC and gel filtration (Figure 3.5) with a final yield of ~30 mg per a litre *E. coli* culture. Gel filtration analysis of EgP1 against protein standards with known molecular masses revealed that EgP1 formed a dimer in solution, with an estimated molecular mass of 286 kDa.

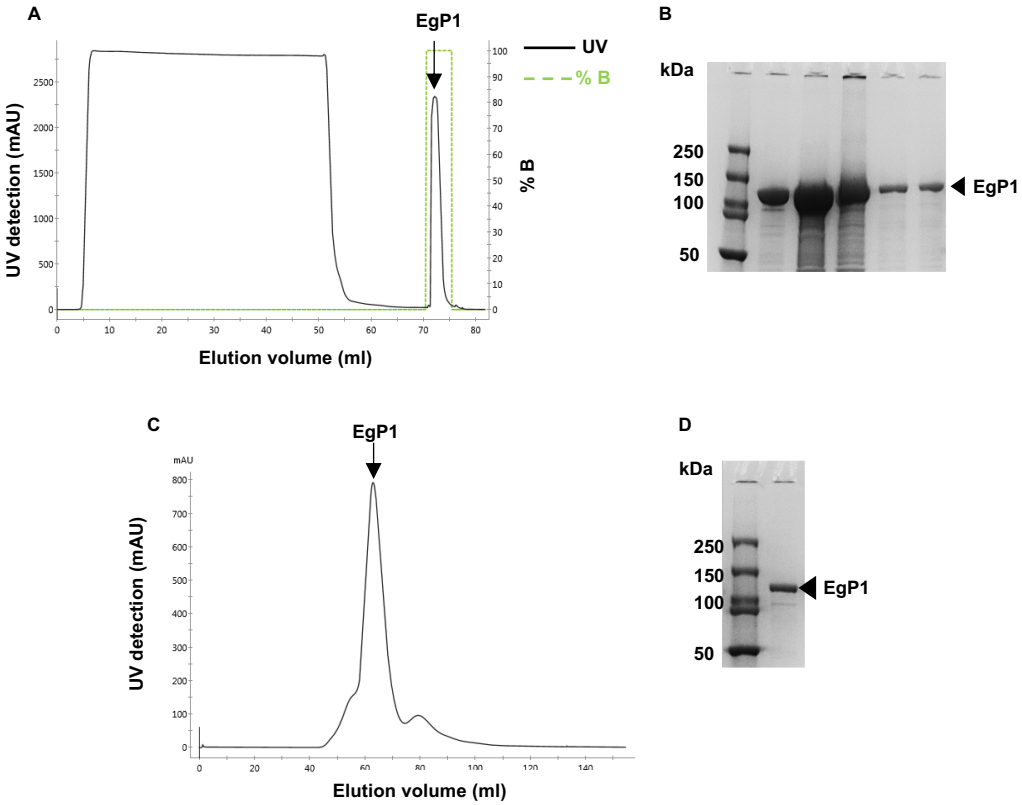


Figure 3.5. Expression and purification of soluble EgP1 recombinant protein (EgP1). (**A**) IMAC purification with UV detection. EgP1 was eluted in 100% of buffer B (10 mM HEPES pH 7.5, 250 mM NaCl, 500 mM imidazole). (**B**) SDS-PAGE of EgP1 after IMAC purification. (**C**) Gel filtration of EgP1. The protein was eluted at 63 ml elution volume in 20 mM HEPES pH 7.5, 150 mM NaCl. (**D**) SDS-PAGE analysis of purified recombinant EgP1 after IMAC and gel filtration.

The soluble EgP1 protein was able to catalyse the conversion of Glc and Glc1P into oligosaccharides with DP up to 12 Glc units that could be detected by MALDI-ToF after 1 hour (Figure 3.6A, B and C and Table 3.3). The EgP1 activity was dependent on both Glc and Glc1P as substrates: excluding either from the reaction mixture completely abolished oligosaccharide production (Figure 3.6A).

To investigate whether EgP1 activity is specific to β-(1→3)-linked glucan, phosphorolysis assays (Figure 3.1) were performed against a variety of disaccharides containing solely glucose with various glycosidic linkages and anomeric configurations. Analysis of these reactions by TLC and High-Performance Anion-Exchange Chromatography with Pulsed Amperometric Detection (HPAEC-PAD) showed that EgP1 was active exclusively on the disaccharide with β-(1→3) linkage, thus confirming its function as a β-(1→3)-D-glucan phosphorylase (Figure 3.6D and E).

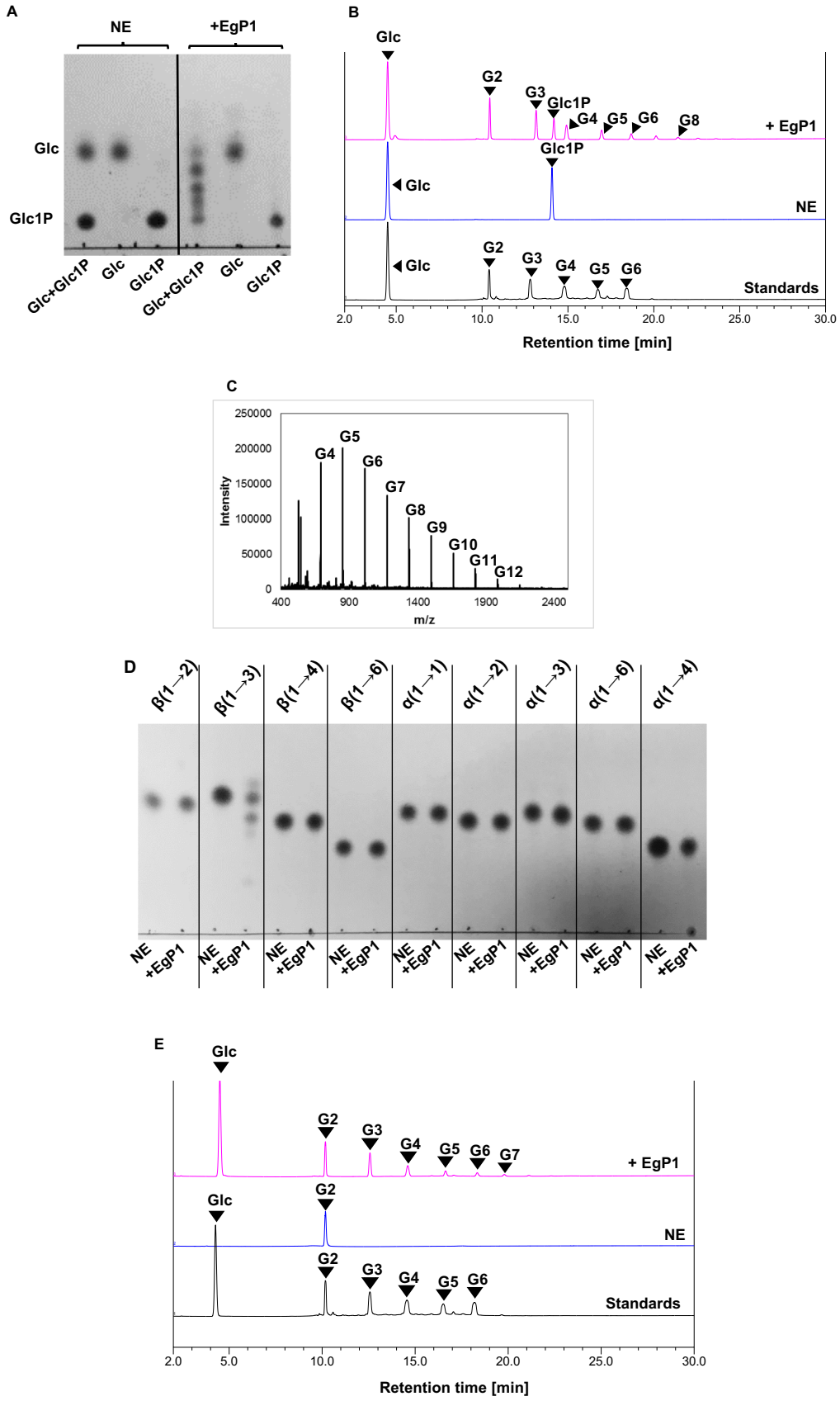


Figure 3.6. Characterisation of the soluble recombinant EgP1. (**A**) TLC analysis of the glycan synthetic reaction catalysed by EgP1. (**B**) HPAEC-PAD analysis of the glycan synthetic carried out by EgP1 in the presence of Glc and Glc1P as substrates. (**C**) MALDI-ToF analysis of the reaction after 1 hour. (**D**) TLC analysis of the phosphorolysis reactions with glucose-containing disaccharides linked by various glycosidic bonds. The linkages of the substrates are indicated at the top. (**E**) HPAEC-PAD analysis of the

phosphorolysis of G2 (Glc-β-(1→3)-Glc) to confirm the breakdown of G2 to Glc by EgP1. However, due to the reversibility of the reaction, longer oligosaccharides (G3 to G7) were detected. NE = no enzyme control, G2 = laminaribiose, G3 = laminaritriose, G4 = laminaritetraose, G5 = laminaripentaose, G6 = laminarihexaose.

Table 3.3. Summary of the observed $[M+Na]^+$ of glucans in MALDI-ToF analysis.

DP of glucan	Calculated $[M+Na]^+$	The closest observed $[M+Na]^+$
G2	365.115	365.148
G3	527.168	527.273
G4	689.221	689.203
G5	851.274	851.223
G6	1013.326	1013.287
G7	1175.379	1175.330
G8	1337.432	1337.374
G9	1499.485	1499.483
G10	1661.538	1661.517
G11	1824.581	1823.591
G12	1986.723	1985.649

The recombinant EgP1 showed similar catalytic capability over a range of pH values (5-8) (Figure 3.7A), and a range of temperature (30-40 °C) (Figure 3.7B).

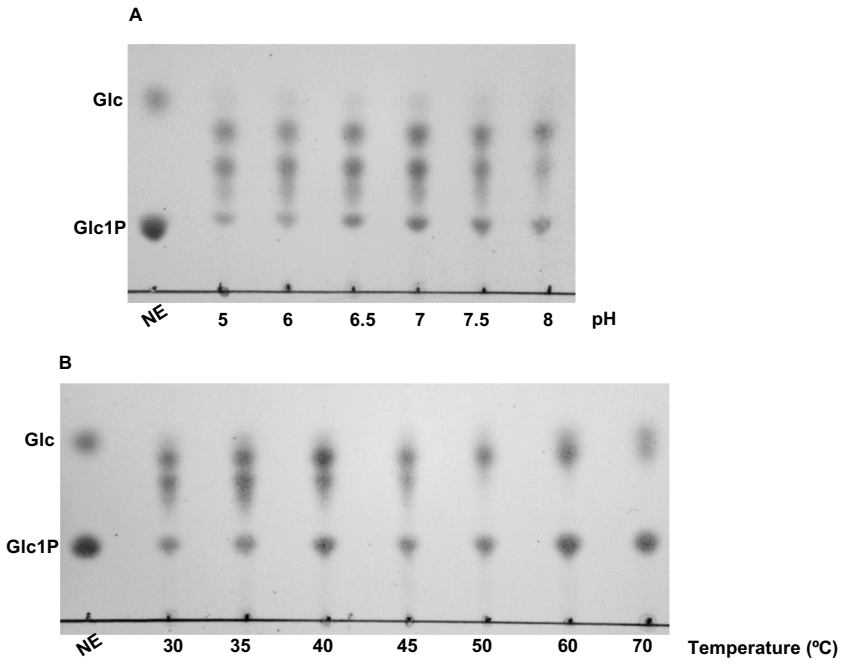


Figure 3.7. TLC analysis of the glycan synthesis carried out by EgP1. (**A**) EgP1 activity over a range of pH values at 30°C. (**B**) EgP1 activity over a range of temperatures at pH 7.0.

Kinetic data of the glycan synthesis with Glc, laminaribiose (G2), laminaritriose (G3), laminaritetraose (G4), laminaripentaose (G5) or laminarihexaose (G6) as acceptors and Glc1P as a donor indicated that the enzyme preferred Glc and G2 as acceptors with similar catalytic efficiency (k_{cat}/K_M = 1.99 and 1.62 for Glc and G2

respectively). There was nearly a two-fold reduction in k_{cat}/K_M as the chain length increased from G2 to G3, which is reflected in a two-fold increase in the K_M values (0.67 for G2 and 1.26 for G3). The catalytic efficiency continued to decrease with increasing chain length from G4 to G6, accompanied by an increase in the K_M values for the respective acceptors (Table 3.4).

Table 3.4. Kinetic parameters of EgP1 for the glycan synthesis using 10 mM Glc1P as a donor.

Acceptors	k_{cat} (s^{-1})	K_M (mM)	k_{cat}/K_M ($s^{-1}mM^{-1}$)
Glc	1.10 ± 0.03	0.56 ± 0.06	1.99
G2	1.08 ± 0.03	0.67 ± 0.08	1.62
G3	1.12 ± 0.02	1.26 ± 0.09	0.89
G4	1.12 ± 0.03	1.41 ± 0.13	0.79
G5	1.13 ± 0.03	2.29 ± 0.19	0.50
G6	1.10 ± 0.03	2.88 ± 0.23	0.38

3.2.3. Identification, recombinant protein production and characterisation of a bacterial orthologue of EgP1

A proprietary metagenome database was interrogated with tBLASTN using EgP1 as a query (communication with Prozomix Limited). The search identified a putative orthologue, designated Pro_7066, with 45% sequence identity to EgP1 (E-value = 0). To investigate whether Pro_7066 has similar activity to EgP1, a recombinant form of Pro_7066 was produced in *E. coli* in the same manner as EgP1 (Figure 3.8). Gel filtration analysis of Pro_7066 against standard proteins with known molecular masses indicated that Pro_7066 formed a dimer in solution, with an estimated molecular mass of 323 kDa.

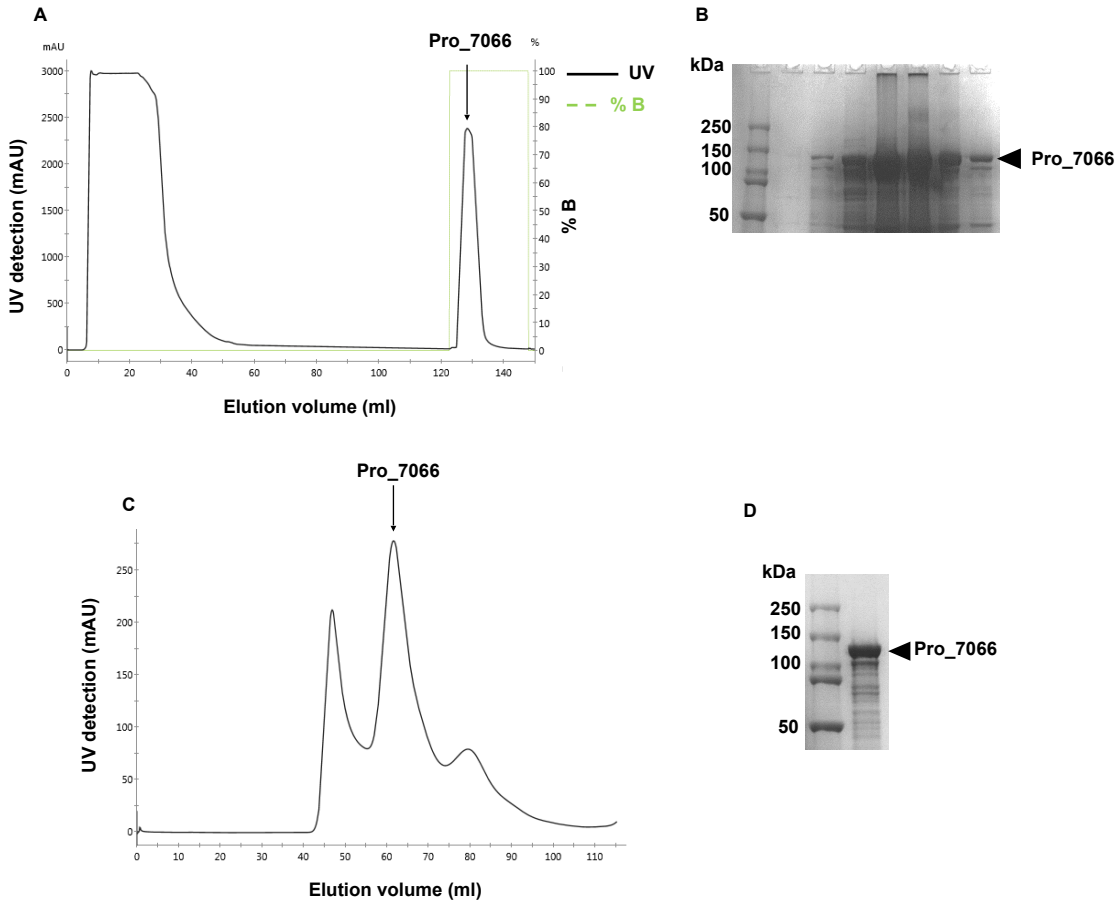


Figure 3.8. Expression of the recombinant Pro_7066 in *E. coli* and purification of the soluble protein. (**A**) IMAC purification with UV detection. Pro_7066 was eluted in 100% of buffer B (10 mM HEPES pH 7.5, 250 mM NaCl, 500 mM imidazole). (**B**) SDS-PAGE of Pro_7066 after IMAC purification. (**C**) Gel filtration of Pro_7066. The protein was eluted at 62 ml elution volume in 20 mM HEPES pH 7.5, 150 mM NaCl. (**D**) SDS-PAGE analysis of Pro_7066 after gel filtration.

Characterisation of Pro_7066 was performed as previously described for EgP1, which showed that the enzyme was also active as a β-(1→3)-D-glucan phosphorylase (Figure 3.9D and E), generating oligosaccharides detectable by MALDI-ToF up to DP 11 (Figure 3.9A-C). Pro_7066 showed similar pH and temperature preferences to EgP1 (Figure 3.10). Interestingly, the kinetic parameters for the glycan synthetic reaction with Glc-G6 as acceptors and Glc1P as a donor showed similar catalytic efficiency for all acceptors investigated, which is strikingly different from that observed in EgP1 (Table 3.5).

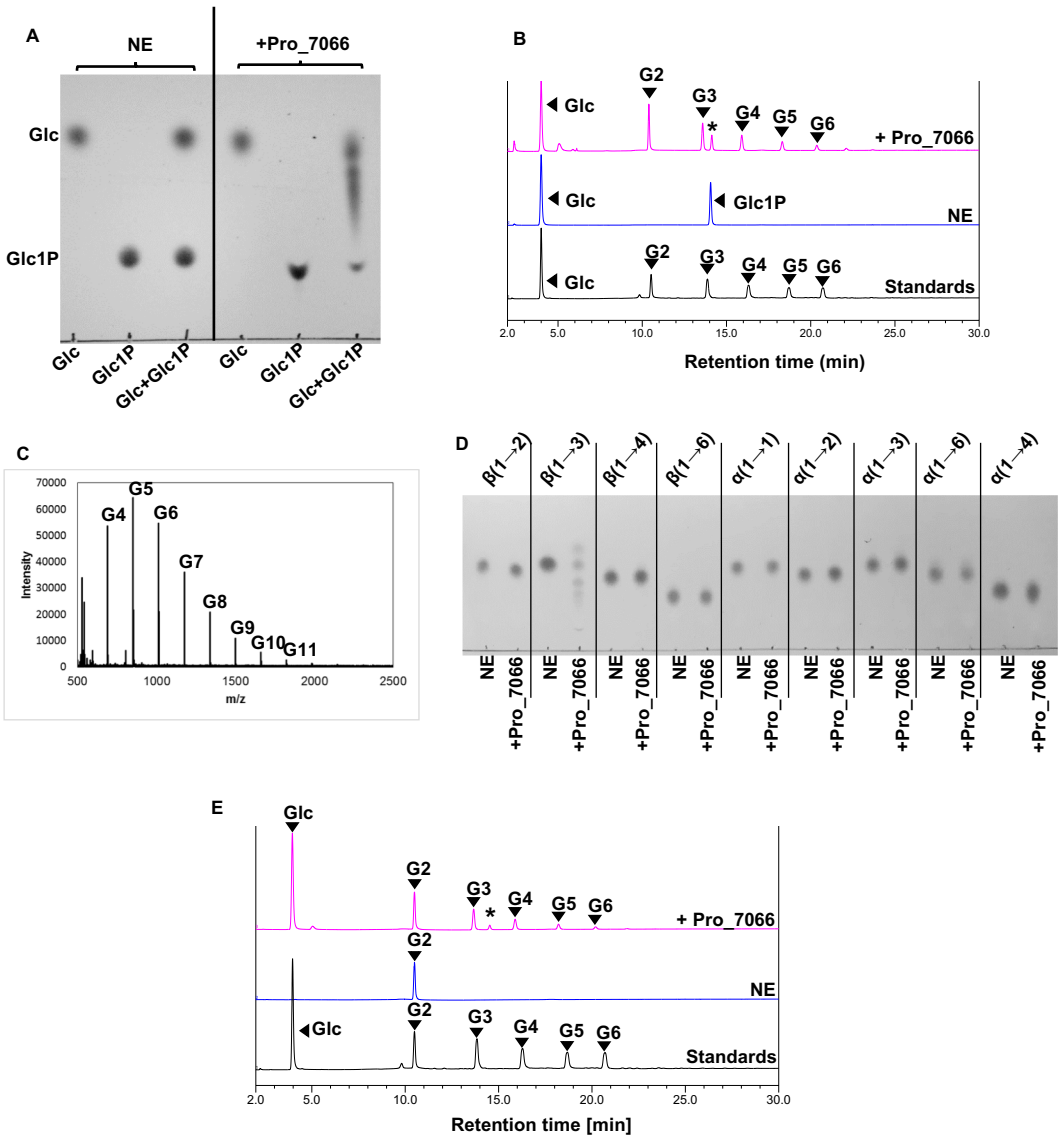


Figure 3.9. Characterisation of Pro_7066. (**A**) TLC analysis of the glycan synthesis carried out by Pro_7066. (**B**) HPAEC-PAD analysis of the glycan synthetic reaction. (**C**) MALDI-ToF analysis of the reverse phosphorolysis after 1 hr. Asterisk = Glc1P. (**D**) TLC analysis of the phosphorolysis reactions with glucose-glucose disaccharides with different linkages. The linkages of the substrates are indicated at the top. (**E**) HPAEC-PAD analysis of G2 (Glc-β-(1→3)-Glc) phosphorolysis by Pro_7066 to confirm the breakdown of G2 to Glc. However, due to the reversibility of the reaction, the formation of G3 to G6 oligosaccharides was also detected. * = small trace of Glc1P due to incomplete desalting by mixed bed ion exchange resin. NE = no enzyme control, G2 = laminaribiose, G3 = laminaritriose, G4 = laminaritetraose, G5 = laminaripentaose, G6 = laminarihexaose.

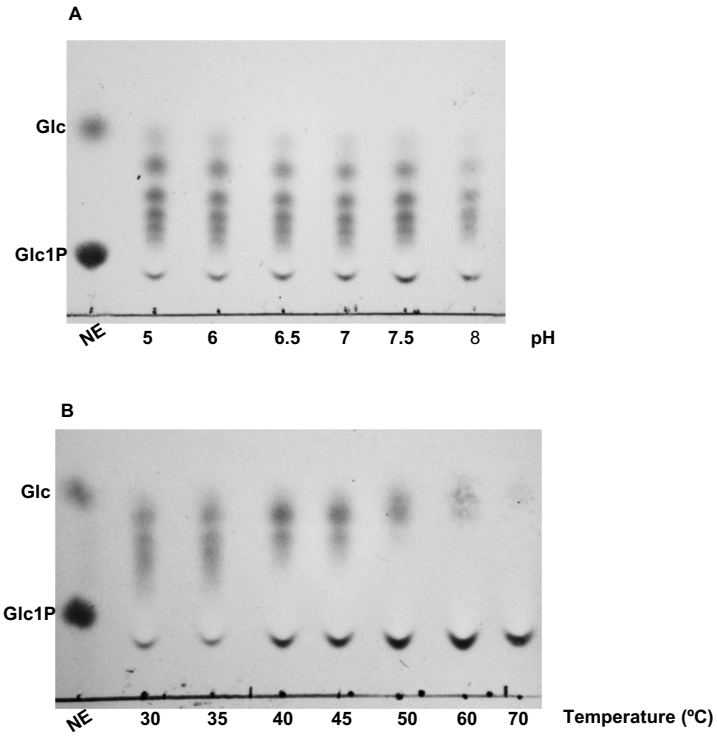


Figure 3.10. TLC analysis of the glycan synthesis carried out by Pro_7066. (**A**) Activity over a range of pH at 30°C. (**B**) Activity over a range of temperatures at pH 7.0.

Table 3.5. Kinetic parameters of Pro_7066 for the glycan synthesis using 10 mM Glc1P as a donor.

Acceptors	k_{cat} (s^{-1})	K_M (mM)	k_{cat}/K_M ($s^{-1}mM^{-1}$)
Glc	1.66 ± 0.04	0.29 ± 0.03	5.79
G2	1.54 ± 0.01	0.25 ± 0.02	6.03
G3	1.53 ± 0.02	0.37 ± 0.03	4.16
G4	1.39 ± 0.01	0.36 ± 0.02	3.89
G5	1.27 ± 0.01	0.32 ± 0.02	4.04
G6	1.18 ± 0.04	0.26 ± 0.04	4.62

3.2.4. Multiple amino acid sequence alignment and phylogeny of EgP1 and its orthologues

The non-redundant protein sequence database (https://www.ncbi.nlm.nih.gov/protein/), the Microbial Eukaryote Transcriptome Sequencing Project (MMET, http://marinemicroeukaryotes.org/) (7.8×10^6 sequences), and the *E. gracilis* strain Z transcriptome (http://euglenadb.org/) have been interrogated with PSI-BLASTP and tBLASTN (threshold 0.0001) using the EgP1 amino acid sequence as a query. Several hundred sequences of bacterial origin with >30% sequence identity to EgP1 were identified in the non-redundant protein sequence NCBI database (Appendix 4 and 5). Protein sequences of eukaryotic origin with 50-62% sequence identity to EgP1 (93-97% query coverage, E-value = 0) were identified in the MMET database and *E. gracilis* Z strain transcriptome.

A reconstructed molecular phylogeny showed that these EgP1 orthologous sequences formed a discrete clade (Figure 3.11A, red branches) that is separated from GH94 sequences (Figure 3.11A, black branches), thus forming a new GH family, denominated GH149. Within the GH149 clade, all sequences from Euglenophyceae share a common node (Figure 3.11B, red label, branch support value = 1) comprising EgP1-4, two sequences from *E. gracilis* strain Z (EuglenaDB IDs 750 and 1135, with 50 and 53% sequence identity to EgP1 respectively; E-value = 0), the laminaridextrin phosphorylase-like protein from *E. gracilis* strain NIES-47 and three sequences from *Eutreptiella gymnastica* CCMP1594 identified from MMET (CCMP1594 IDs 200380720, 200416338 and 200418968 with 62, 50 and 51% sequence identity to EgP1, E-value = 0).

Most bacterial GH149 family members were identified from sequenced and translated genomes of phylum Proteobacteria, dominated by class γ-proteobacteria (Figure 3.11B, 134 sequences, purple labels), and phylum Bacteroidetes (Figure 3.11B, 125 sequences, light green labels). The majority of these sequences are currently annotated as predicted hypothetical proteins with unknown function. Pro_7066 was found to be most closely related to sequences from Bacteroidetes species (Figure 3.11B, a black arrow head with Pro_7066 label), in particular *Flavobacteria bacterium* GWF1_32_7 (OGS61795.1, %identity = 76, E-value = 0), suggesting that Pro_7066 might belong to a *Flavobacteria* species. The Bacteroidetes species from which GH149 were found were predominantly isolated from marine environments; 13 have been described to be associated with algae and phytoplankton.

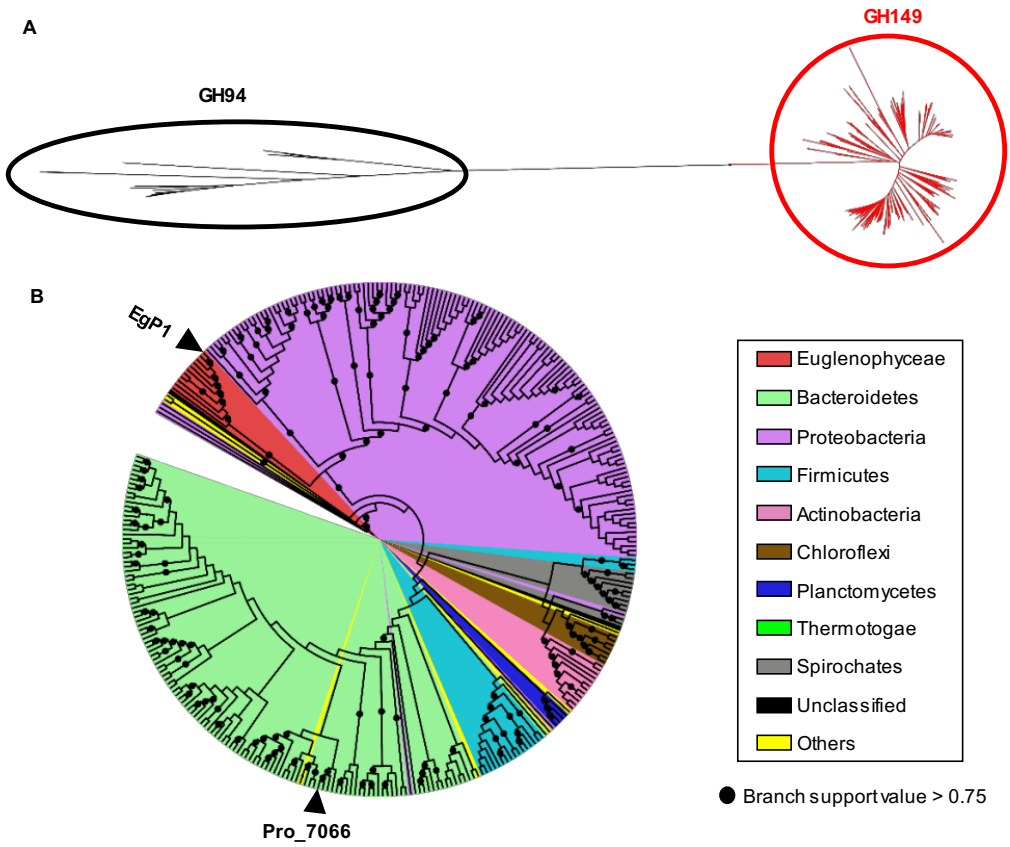
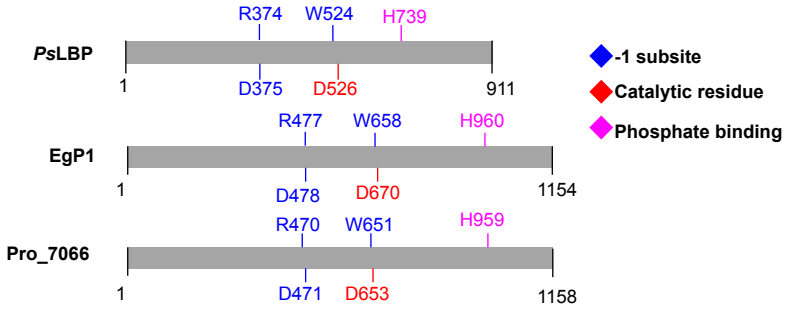


Figure 3.11. Phylogenetic analysis of GH149. (**A**) An unrooted tree representing the phylogenetic relationship between GH94 queries (black branches) and GH149 (red branches). (**B**) Phylogenetic analysis of GH149. The colours represent the taxonomic phyla of the species containing GH149 as indicated in the legend. The positions of EgP1 and Pro_7066 are indicated by arrow heads.

In order to generate hypotheses about the active site residues of GH149, the amino acid sequences of EgP1-4 and Pro_7066 were aligned with those of GH94 family. Interestingly the amino acids that constitute the active site of GH94, including the -1 subsite (sugar donor subsite), Asp catalytic residues and a His residue for phosphate recognition (Figure 3.12) appeared conserved in GH149, suggesting that GH149 and GH94 form a clan of related families that is termed GH-Q (communication from Professor Bernard Henrissat, Head of the Carbohydrate-active enzyme database (CAZy)).

A



B

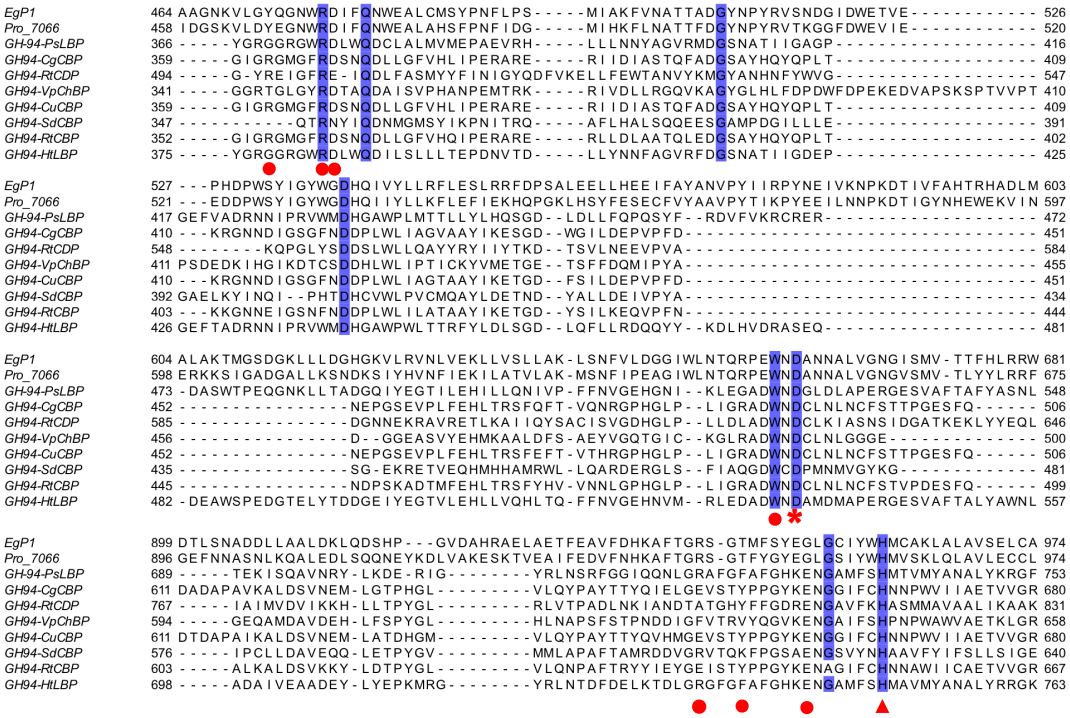


Figure 3.12. Conservation of key amino acids in GH94 and GH149 sequences. (**A)** Summary of the conserved amino acids located in the enzyme active sites. EgP1 sequence represents the truncated version where the predicted target peptide (residue 1-31) has been removed. *Ps*LBP = laminaribiose phosphorylase from *Paenibacillus* sp. YM-1. (**B**) Detail of the multiple alignment of EgP1, and Pro_7066 and GH94 enzymes revealed conservation of key amino acids**.** 100% conserved amino acid residues are highlighted in dark blue. Red circles indicated the residues involved in the formation for -1 subsite. Red triangle represents the residue involved in phosphate binding. Red asterisk represents the predicted catalytic residue.

3.2.5. Structural studies of Pro_7066

3.2.5.1. Crystallisation of Pro_7066

In order to confirm the structural conservation of the amino acids predicted from the alignment of GH149 and GH94 sequences, X-ray crystallography was used to determine the structure of Pro_7066. The substrate-free structure was solved to 2.05 Å resolution. Structures with bound oligosaccharide ligands, were then sought through co-crystallisation with laminaribiose (G2), -triose (G3), -tetraose (G4), -pentaose (G5) or -hexaose (G6) (at 8 mM). However, useful data could only be obtained from the G6 and G2 co-crystallisations.

3.2.5.2. Overall structure of Pro_7066

All structures belong to space group $P2_12_12_1$ and contain two subunits per asymmetric unit. The two subunits superposed with an RMSD of 0.23 Å (Cα- Cα superposition). The two copies of the molecule in the asymmetric unit form a biologically relevant homodimer with an interfacial area of ~4230 $Å^2$ as calculated by jsPISA.[43] The formation of a homodimer in the crystal structure is consistent with the gel filtration analysis, where Pro_7066 was eluted as a dimer. The domains present within each subunit can be defined as follows: an N-terminal β-sandwich (residues 1-303; yellow), a helical linker region (residues 304-341; lilac), an $(α/α)_6$ catalytic domain (residues 342-1045; green) and a C-terminal jelly-roll domain (residues 1046-1156; red) (Figure 3.13). These domains are similar to those observed in *Ps*LBP and other GH94 enzymes.[20,44–49] Cellodextrin phosphorylase from *Ruminiclostridium thermocellum* (*Rt*CDP)[20] and β-(1→2)-oligoglucan phosphorylase from *Lachnoclostridium phytofermentans (Lp*SOGP)[48], both belonging to GH94 family, have additional, but dissimilar, N-terminal domains (Figure 3.13, purple in *Rt*CDP), but neither is present in Pro_7066.

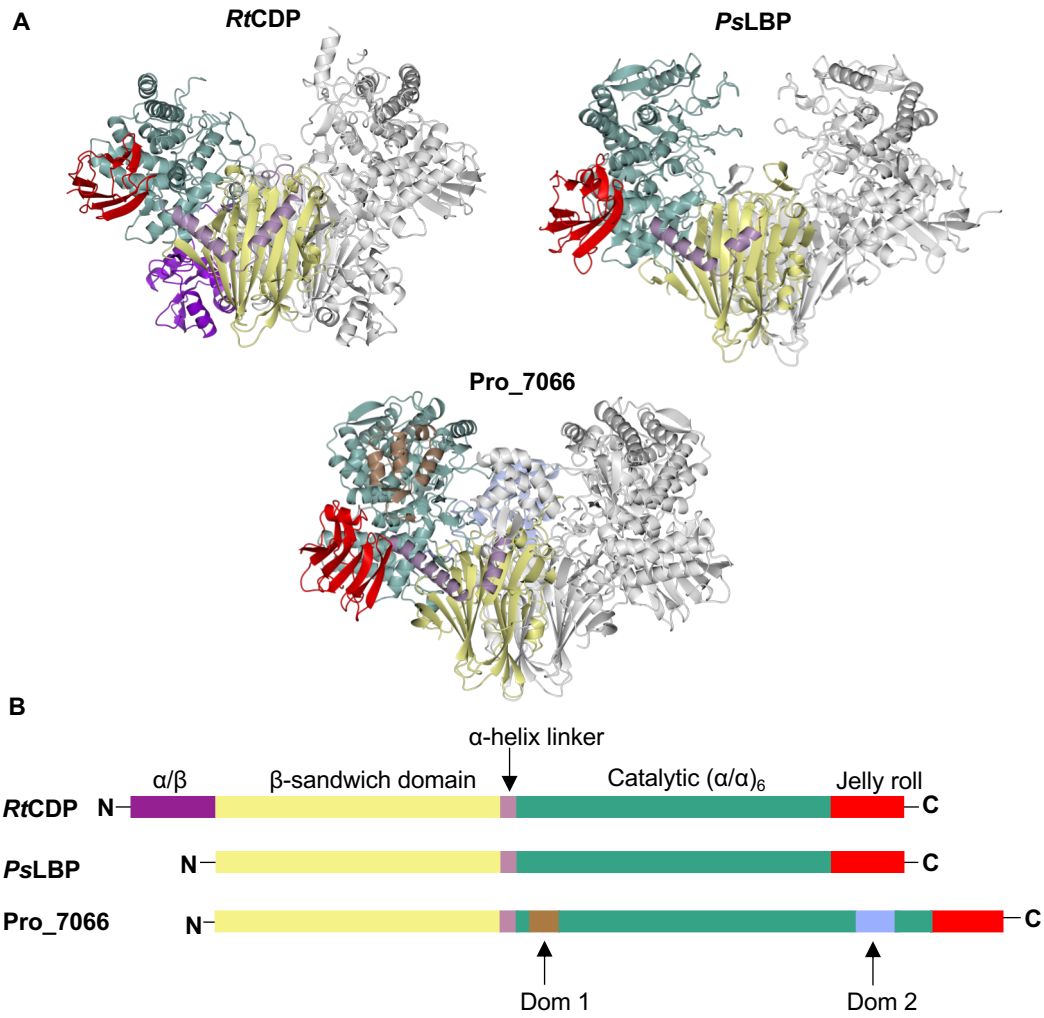


Figure 3.13. Structural comparison between *Rt*CDP, *Ps*LBP and Pro_7066. (**A**) Overall structures of Pro_7066 in comparison to *Ps*LBP (PDB code; 6GH2) and *Rt*CDP (PDB code; 5NZ8). (**B**) Schematic representation of the primary sequences of *Rt*CDP, *Ps*LBP and Pro_7066 coloured according to the domains.

Interestingly, there are two extra domains inserted within the catalytic domain, designated Dom 1 (residues 370-428, brown) and Dom 2 (residues 852-938, ice blue) (Figure 3.14B). The structures of Dom 1 and 2 were investigated further by DALI analysis[50] (http://ekhidna2.biocenter.helsinki.fi/dali/), which is a server for comparing 3-D protein structures to reported protein structures in PDB database and lists structural similarity based on Z-score ranking. Dom 1 and 2 are not significantly similar to domains in any known protein structures according to the DALI analysis (Z-scores < 5). Analysis of Dom 1 and 2 sequences by Pfam[51] and an NCBI conserved domain search[52] did not reveal any significant hits. However, Dom 1 and 2 are predominantly α-helical and augmented by one or two β-hairpins, somewhat reminiscent of a CBM43 associated with a transglycosidase acting on β-(1→3)-glucan chains (Figure 3.14C).[53]

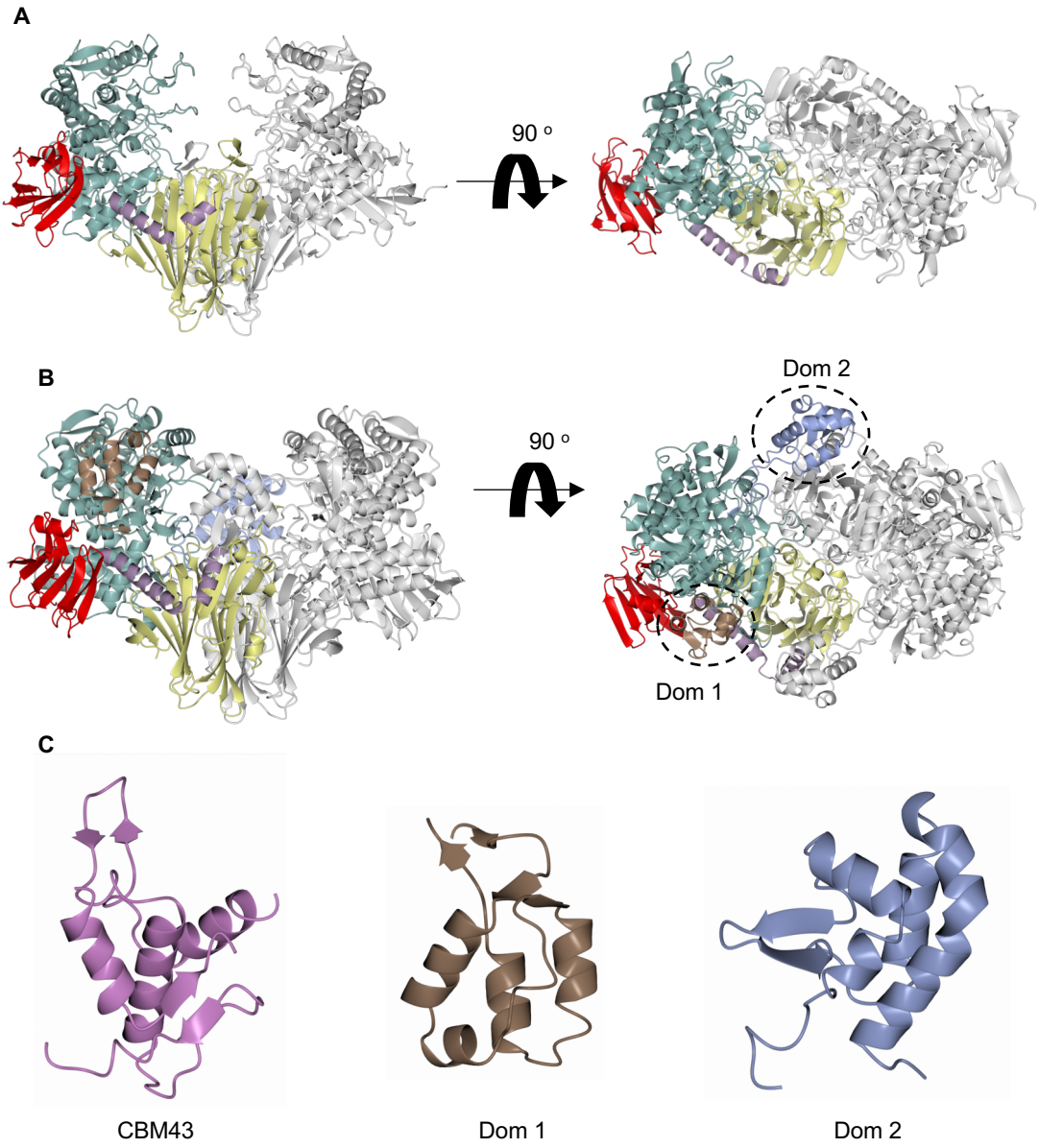


Figure 3.14. Additional domains in Pro_7066. (**A**) *Ps*LBP structure. (**B**) Pro_7066 structure. Dom 1 and 2 in Pro_7066 are coloured in brown and ice blue respectively and indicated by dotted circles. (**C**) Structures of Dom 1 and 2 in Pro_7066 in comparison to CBM43 (pink) associated with GH72 (PDB code; 2W63).[53]

3.2.5.3. Pro_7066 active site

The architecture of the Pro_7066 active site is highly similar to that observed in GH94 enzymes (Figure 3.15). The conserved structural characteristics include; 1) the presence of a WND motif in the catalytic loop, with D653 as the predicted catalytic residue, although the D653 side chain is rotated 100° around Cα-Cβ bond in comparison to that in *Ps*LBP structure; 2) a conserved RD motif (R470 and D471) which is involved in the recognition of sugar 1-phosphate (with the exception of *Ps*LBP where the distance between the aspartate residue (D375) and Glc1P is greater than the hydrogen bonding distance); 3) the conserved histidine (H959) which was predicted to be involved in phosphate recognition, although this residue does not form a hydrogen

bond with the sulphate molecule in Pro_7066 or *Ps*LBP structures. These structurally-conserved amino acids are in agreement with the sequence alignment which predicted the presence of these residues in the enzyme active site (Figure 3.12). Bicine (BCN) derived from the crystallisation precipitants was bound via hydrogen bonding contacts with R470 and E951, possibly mimicking the interaction of the enzyme active site with the hydroxy groups on C3 and C6 of Glc1P donor (Figure 3.15B). A sulphate molecule was observed (ice blue, Figure 3.15B) occupying similar position to phosphate moiety in *Ps*LBP structure (magenta, Figure 3.15C).

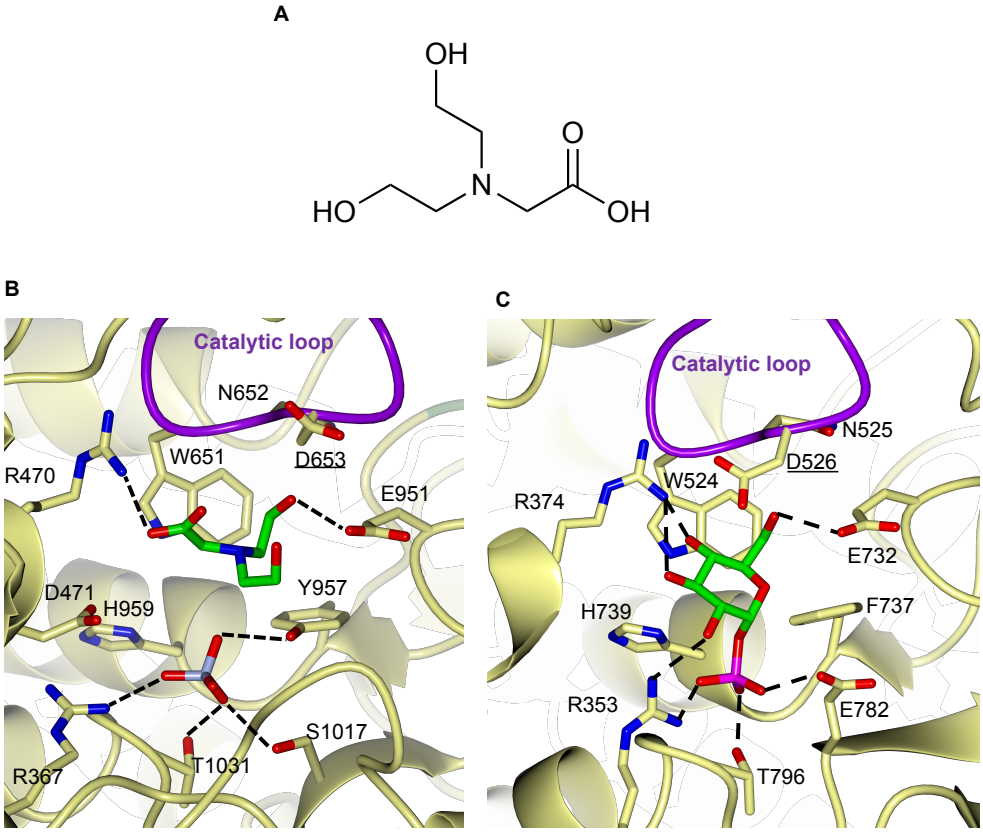


Figure 3.15. The active site of Pro_7066 in comparison to that of *Ps*LBP. (**A**) Molecular structure of BCN. (**B**) The active site of substrate-free Pro_7066. (**C**) The active site of *Ps*LBP in complex with Glc1P.

3.2.5.4. Structural elements surrounding Pro_7066 active site

In order to rationalise the difference in substrate preferences between *Ps*LBP and Pro_7066, the structures of *Ps*LBP (PDB code 6GH2), *Rt*CDP (PDB code 5NZ8) and substrate-free Pro_7066 were superimposed. Although the RMSD values suggest that these structures are not highly similar (RMSD > 1 Å), the structural elements neighbouring the active site are relatively well conserved (Figure 3.16); 1) the gate (blue) described in *Ps*LBP can be identified in Pro_7066, although the gate in Pro_7066 is a disordered loop; 2) the opposing (red) and the adjacent loops (brown), which are

present in all of the three structures. Despite the structural conservation, there is no conservation in the amino acid compositions in these loops. In Pro_7066 structure, Dom 2 (Figure 3.16, green) is found in proximity to the gate. The average B-factor of Dom 2 (~70 $Å^2$) is higher than the average B-factor of the protein subunit (48 $Å^2$), suggesting Dom 2 is more dynamic compared to the rest of the subunit.

In addition to the conserved structural elements, there is an additional structural element, designated as an 'upper' strand (Figure 3.16, pink). This element is located above the adjacent loop in *Ps*LBP (residue 465-480) and Pro_7066 (residue 568-603), but absent in *Rt*CDP. In Pro_7066, the upper strand is 20-residue longer than that in *Ps*LBP and contains a helical structure connected to a small β-hairpin, whereas the portion connected to the hairpin is disordered in *Ps*LBP.

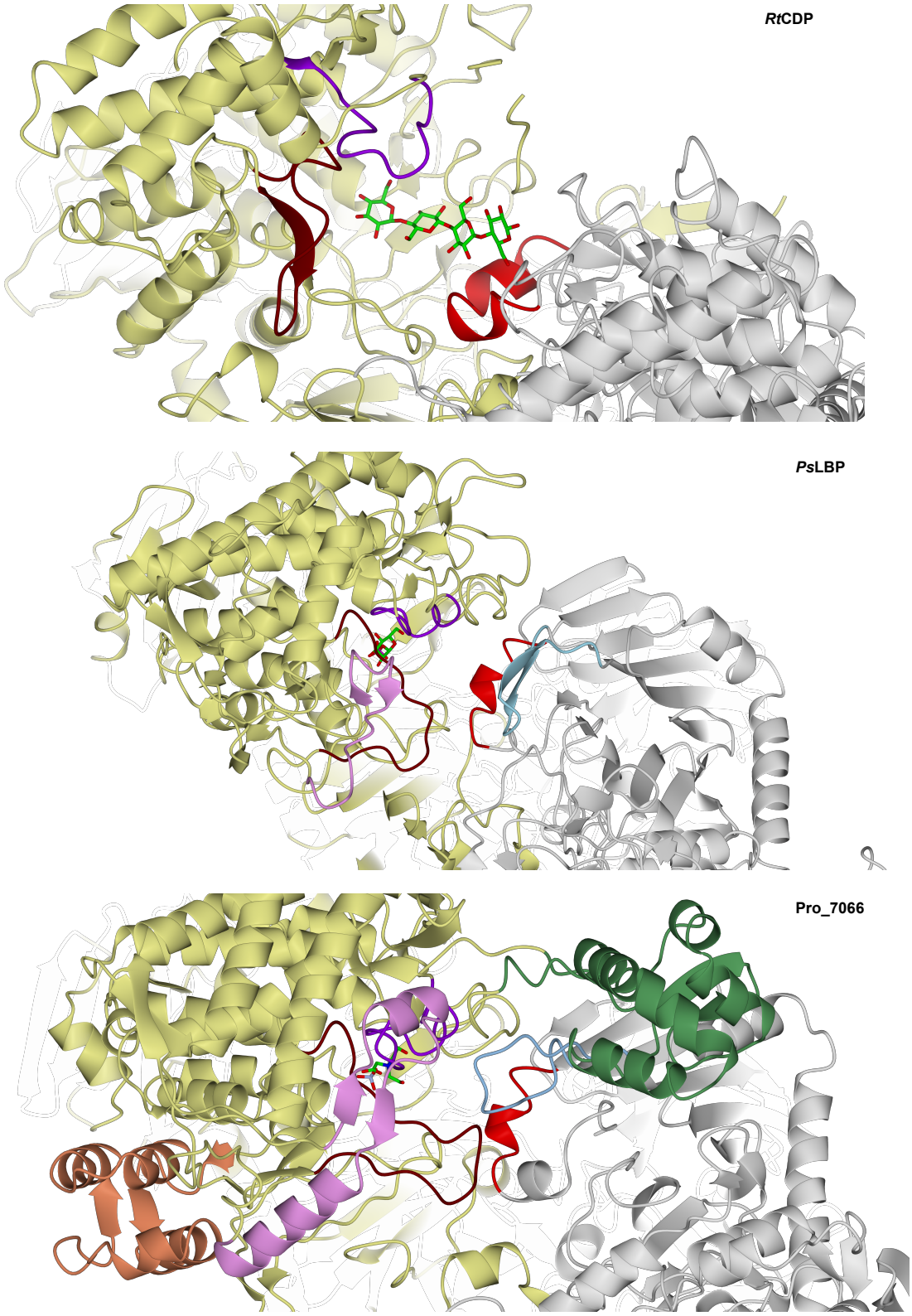


Figure 3.16. Structural elements surrounding the active sites of *Rt*CDP, *Ps*LBP and Pro_7066. Purple = catalytic loop, red = opposing loop, blue = gate, brown = adjacent loop, pink = upper strand, orange = Dom 1, green = Dom 2. The gate and opposing loops are located the adjacent subunit.

3.2.5.5. Structure of Pro_7066 in complex with G6 revealed an oligosaccharide SBS on the surface of the catalytic domain

The overall structure of the G6 complex is similar to the substrate-free complex (RMSD = 0.35 Å for a dimer on dimer superposition). The active site of the G6 complex is occupied by BCN and SO_4^{2-}, identical to that in the substrate-free complex. There was no detectable binding of G6 within the active site. However, G6 bound to a potential SBS located on the solvent-exposed surface of the catalytic domain and is ~30 Å away from the active site (Figure 3.17A and C). Although structures with G2-G5 bound could not be refined due to the poor quality of the diffraction data, the auto-processed data from the G2-G5 co-crystallisations were analysed by Dimple[54] which showed weak residual electron density consistent with sugar pyranose rings near the same SBS site.

The SBS consists of two aromatic residues (Y841 and Y993) that likely form stacking interactions with sugar residue B and D of G6 respectively. Several hydrogen bonds are formed between the hydroxy groups (mainly C2-OH) of G6 and the SBS amino acid side chains (K831, E838, E990, E994 E982 and R1062). Comparison between the G6 and substrate-free complexes revealed dynamic conformations of the side chains that make hydrogen bonding contacts with G6 (E838, E982, E990, Y993, and E994) (Figure 3.17B). Molecular surface representation of the SBS (Figure 3.17D) revealed a shallow cleft with a U-shaped topology, which complements the curvature of G6, the conformation commonly adopted by β-(1→3)-linked gluco-oligosaccharides.[55,56]

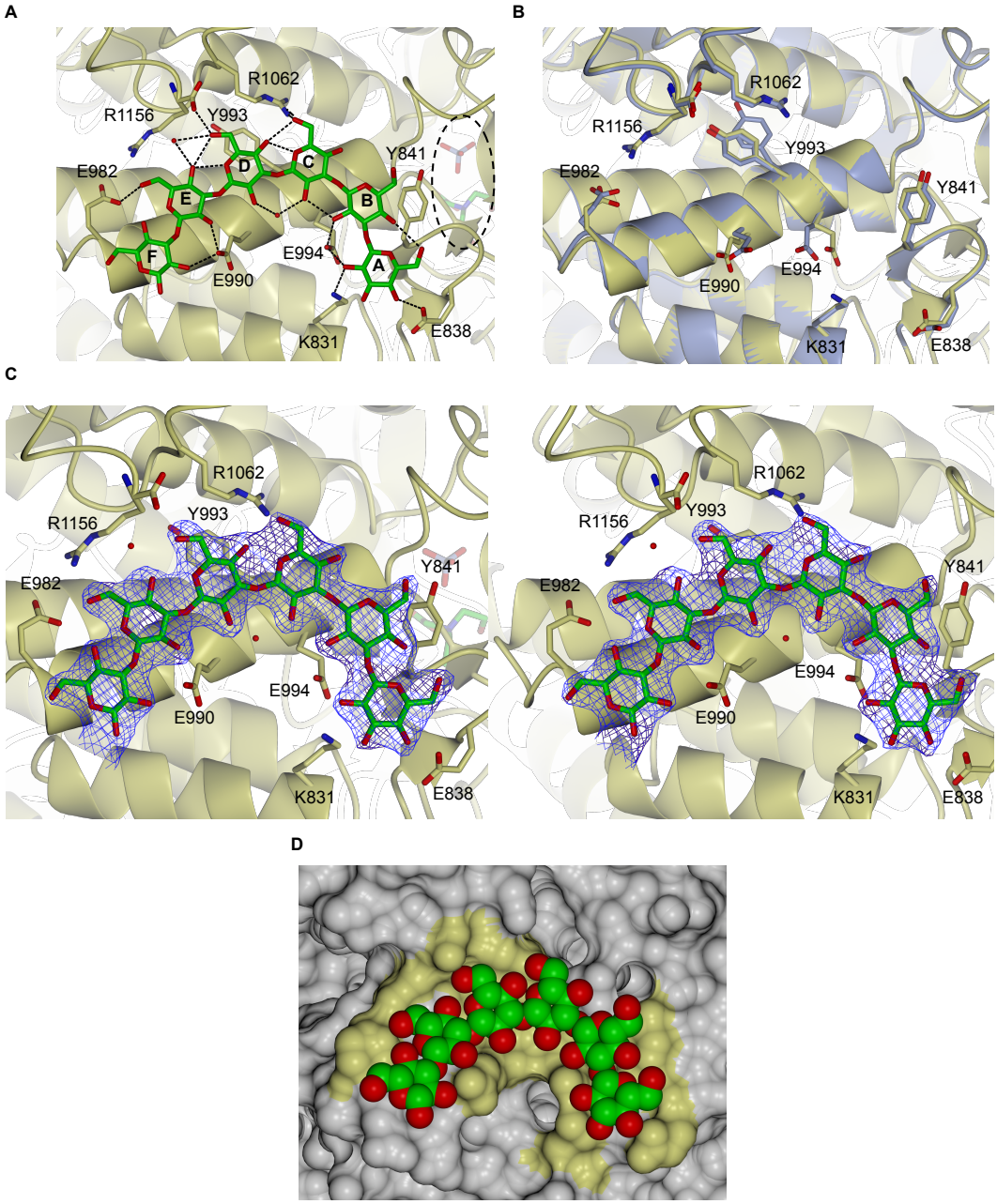


Figure 3.17. The SBS site on the surface of Pro_7066 catalytic domain. (**A**) SBS site with G6 bound. The enzyme backbone is presented as ribbon (yellow). The active site with BCN and SO_4^{2-} bound is buried within the catalytic domain and is on the opposite site of the SBS (dotted oval). Carbon atoms of G6 are presented in green and oxygen in red. The Glc residues in G6 are designated as A-F from the non-reducing to the reducing end. (**B**) Superposition of the SBS site in the G6 complex (yellow) with the substrate-free complex (blue). Residual electron density indicates the presence of Y993 side chain in two alternate conformations in the substrate-free complex. (**C**) A stereo image of the SBS with an omit *mFobs-dFcalc* difference electron density map (~2.5 Å resolution; contoured at ~3σ) shown in blue mesh, which was generated for the bound ligand using phases calculated from the final model without the ligand after the application of small random shifts to the atomic coordinates, re-setting temperature factors, and re-refining to convergence. (**D**) Molecular surface representation of the SBS, with yellow patches representing the location of side chains that are involved in the interaction with G6. G6 is represented in space-filled model in D.

To determine whether the SBS observed in Pro_7066 is present in GH94 enzymes, superposition of Pro_7066 with *Rt*CDP, *Ps*LBP and *Lp*SOGP structures was performed. The analysis concentrates on the structural conservation of E990, Y993 and

E994, which form intimate contacts with G6 in Pro_7066. Superposition of Pro_7066 with *Ps*LBP, *Rt*CDP and *Lp*SOGP structures showed that the residues constructing the Pro_7066 SBS are not well-conserved in the GH94 structures (Figure 3.18).

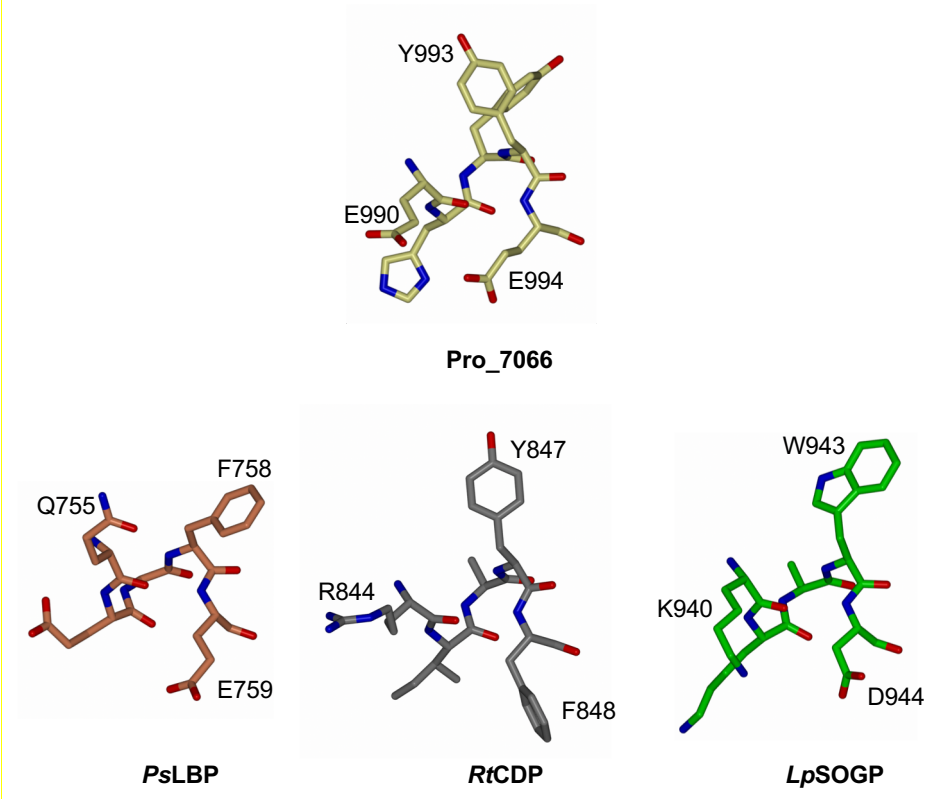


Figure 3.18. Structures of the residues in GH94 enzymes that superposed with E990, Y993 and E994 in Pro_7066 structure.

In order to investigate whether the SBS is conserved among enzymes in GH149 family, multiple amino acid sequence alignment of GH149 members was performed. The alignment revealed that Y993 and E994 (YE motif) is relatively conserved in Euglenophyceae sequences, with two of the Eutreptiella sequences (200418968 and 200416338) having the glutamate substituted by aspartate, whilst E990 is less conserved (Figure 3.19). In contrast, the YE motif and E990 are relatively conserved in bacterial GH149 sequences that shared >60% overall sequence identity to Pro_7066, with an exception to WP_035128081.1 from *Flavobacterium aquatile*, in which YE motif and E990 are not conserved. The YE motif and E990 are less conserved in other bacterial GH149 sequences that shared lower sequence homology to Pro_7066.

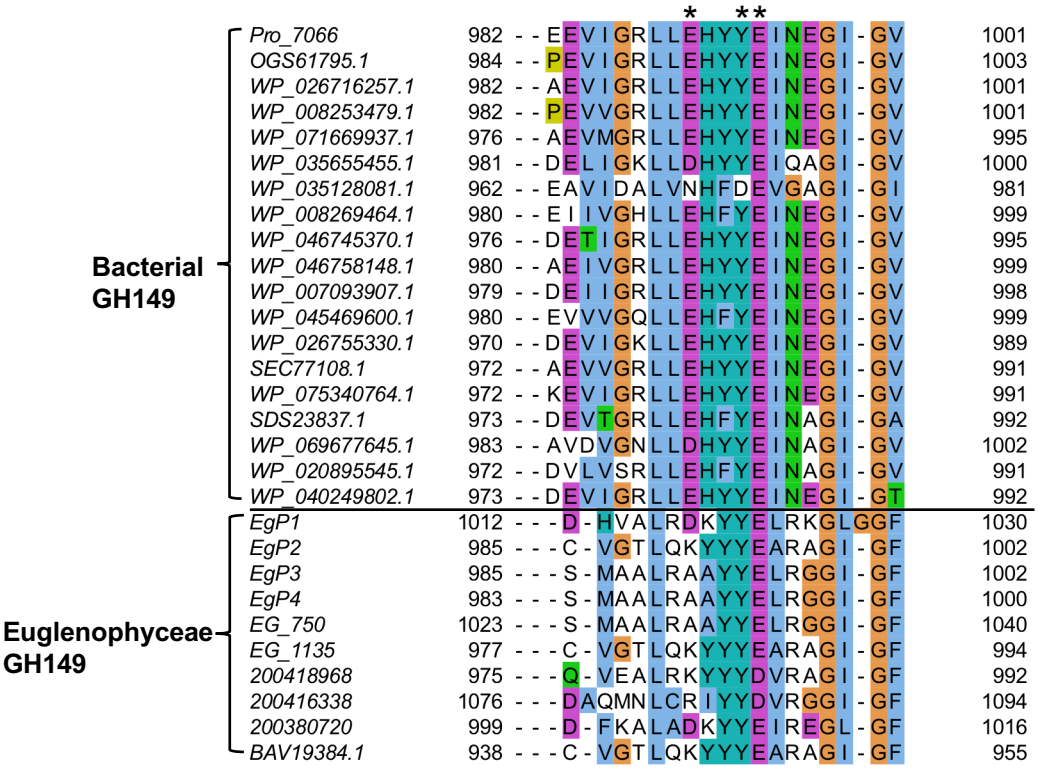


Figure 3.19. Multiple amino acid sequence alignment of Pro_7066 with Euglenophyceae GH149 and bacterial GH149 sequences. The presented bacterial GH149 sequences shared >60% sequence identity to Pro_7066. The amino acids are coloured according to Clustal amino acid colour scheme. Asterisks indicate the amino acids that are in intimate contacts with G6 in Pro_7066. The organism sources of the sequences are as followed; OGS61795.1 [*Flavobacteria bacterium* GWF1_32_7], WP_026716257.1 [*Flavobacterium gelidilacus*], WP_008253479.1 [*Flavobacteria bacterium* BAL38], WP_071669937.1 [*Flavobacteriaceae* bacterium UJ101], WP_035655455.1 [*Flavobacterium filum*], WP_035128081.1 [*Flavobacterium aquatile*], WP_008269464.1 [*Flavobacteriales* bacterium ALC-1], WP_046745370.1 [*Kordia zhangzhouensis*], WP_046758148.1 [*Kordia jejudonensis*], WP_007093907.1 *[Kordia algicida*], WP_045469600.1 [*Winogradskyella* sp. PG-2], WP_026755330.1 [*Sediminibacter* sp. Hel_I_10], SEC77108.1 [*Tenacibaculum* sp. MAR_2009_124], WP_075340764.1 [*Tenacibaculum* sp. HZ1], SDS23837.1 [*Formosa* sp. Hel1_31_208], WP_069677645.1 [*Formosa* sp. Hel1_33_131], WP_020895545.1 [*Winogradskyella psychrotolerans*], WP_040249802.1 [*Psychroserpens mesophilus*], EgP1-4 [*Euglena gracilis*], EG_750 [*Euglena gracilis*], EG_1135 [*Euglena gracilis*], BAV19384.1 [*Euglena gracilis*], 200418968 [*Eutreptiella gymnastica*], 200416338 [*Eutreptiella gymnastica*], 200380720 [*Eutreptiella gymnastica*].

3.2.5.6. Conformation of G6 ligand at the Pro_7066 SBS

Molecular dynamics (MD) simulations of β-(1→3)-linked gluco-oligosaccharides revealed the global energy minima of the oligosaccharide is at *φ*/*ψ* torsion angles -72°/108° (see definition of the torsion angles in Figure 3.20A)[55] and a triple-helical conformation is the most favourable with pitch 22 Å and 6 Glc residues per turn, suggesting that free G6 is capable of forming a triple helical structure in solution.[57] The structure of free G6 based on the calculated MD torsion angles was built in Polys Glycan Builder (http://glycan-builder.cermav.cnrs.fr/) and compared with the conformation of the G6 ligand in complex with Pro_7066, which showed that the G6 ligand is relatively flat compared to the free G6 structure that is more twisted at residues A, B, E and F (Figure 3.20C and D).

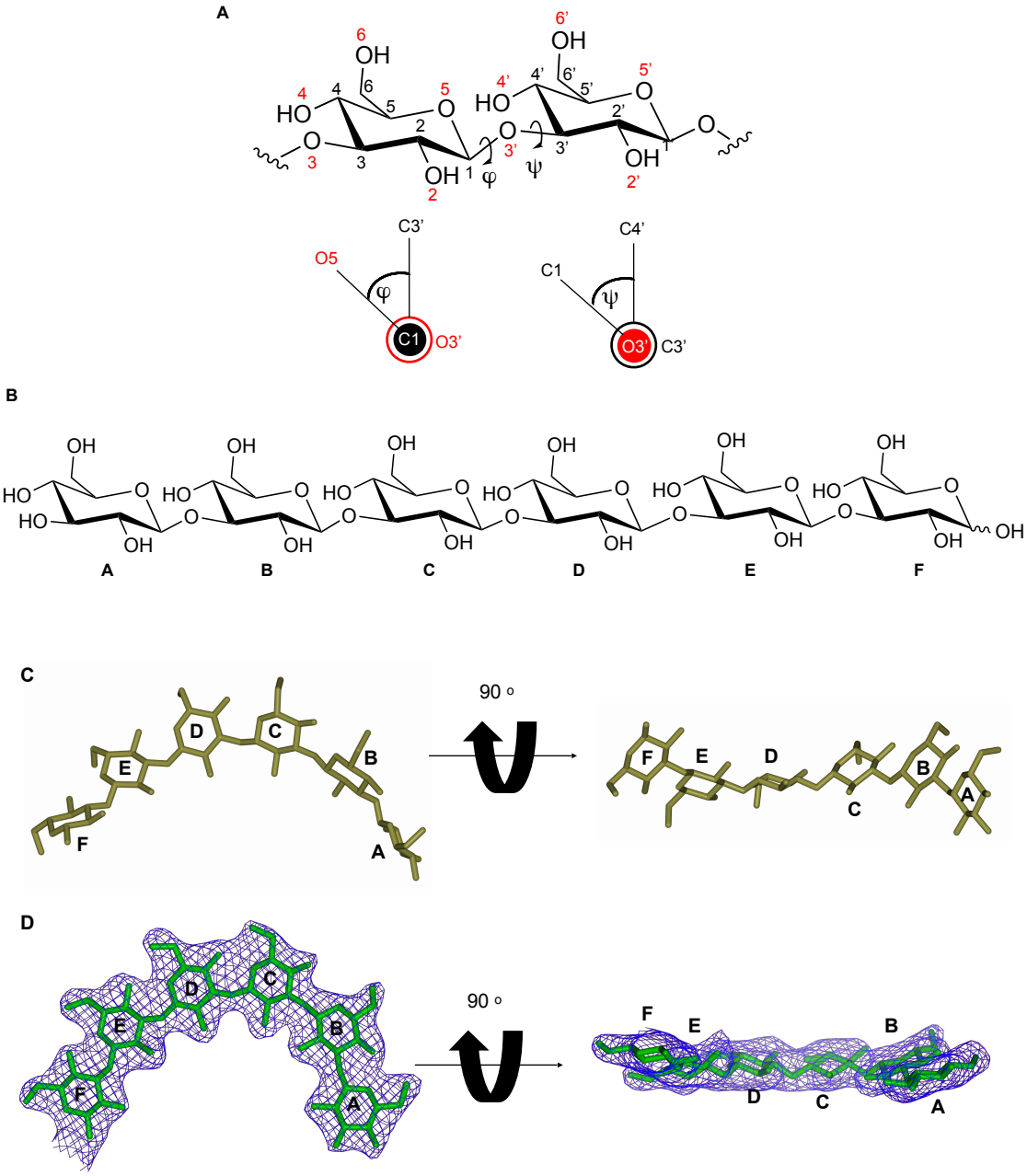


Figure 3.20. G6 conformations in solution and at the Pro_7066 SBS. (**A**) φ and ψ definitions for β-(1→3)-glycosidic bond. (**B**) Glc residues in G6 are designated as A to F, which is used to define the position of torsion angles in Table 3.6. (**C**) Free G6 structure built by Polys Glycan Builder (http://glycan-builder.cermav.cnrs.fr) with φ/ψ torsion angles = -72°/108°. (**D**) G6 ligand at the SBS of Pro_7066 with the omit *mFobs-dFcalc* difference electron density (blue mesh).

The triple helical structure of β-(1→3)-glucan is stabilised by inter- and intra-strand hydrogen bonds. The inter-strand hydrogen bonds are formed between OH groups on C2 (C2-OH),[58] whilst intra-strand hydrogen bonds are formed between O4 and O5 of adjacent Glc subunits.[58–60] In the Pro_7066 G6 complex, the inter-strand hydrogen bonds are likely broken because the C2-OH groups are predominantly involved in the interaction with the SBS. The number of intra-strand hydrogen bonds was substantially reduced in G6 complex, where only one hydrogen bond was detected

between O5 of residue D and OH on C4 of residue E. The loss of inter-strand hydrogen bonds and the reduced number of intra-strand hydrogen bonds likely prevent the G6 ligand from adopting a triple helical conformation.

Several β-(1→3)-linked gluco-oligosaccharide-interacting proteins have been reported to recognise the triple helical form of β-(1→3)-linked gluco-oligosaccharide.[56,60–63] In order to determine whether the G6 ligand in Pro_7066 adopts similar conformations to other β-(1→3)-linked gluco-oligosaccharide ligands, the torsion angles of glycosidic bonds in β-(1→3)-linked gluco-oligosaccharide ligands in other PDB structures were measured and compared to those from the G6 in Pro_7066 (Table 3.6). There is a substantial variation in the torsion angles of β-(1→3)-linked gluco-oligosaccharides among different reported PDB structures and these values differ from those measured from the G6 in Pro_7066 structure.

Table 3.6. Summary of dihedral angles from published PDB structures in complex with β-(1→3)-gluco-oligosaccharides.

Bonds / PDB	Binding site	φ (O5 – C1 – O3' – C3')					ψ (C1 – O3' – C3' – C4')				
		A/B	B/C	C/D	D/E	E/F	A/B	B/C	C/D	D/E	E/F
G6 complex (chain A)*	SBS	-95	-111	-88	-73	-98	86	116	86	103	81
G6 complex (chain B)*	SBS	-88	-112	-80	-76	-96	85	120	83	102	78
3AQX[63]	β-(1→3)-glucan receptor	-82	-73	-84	-79	-70	127	106	137	115	127
5T4C[62]	Ancillary binding site (parallel to active site)	-47	-66	-84	-79	N/A	110	110	81	100	N/A
	Active site	-110	-71	-132	-74	-60	120	99	97	105	128
4PF0[61]	GH55 active site	-61	-84	-69	-57	-62	108	102	122	113	107
1GUI[56]	CBM4	-82	-94	-120	-63	-116	175	109	120	122	11
2WNE[57]	GH16 active site	-89	-117	-124	-130	-107	95	121	113	86	143
5H9Y[60]	GH64 active site	-60	-82	-104	-97	N/A	135	123	143	135	N/A

*in Pro_7066 structure

3.2.5.7. Effect of BCN on Pro_7066 activity

The BCN ligand occupying the Pro_7066 active site might prevent G6 ligand binding and inhibit the enzyme activity. In order to investigate whether BCN has an inhibitory effect on the enzyme activity, the glycan synthetic activity of Pro_7066 was assayed in the presence of 10 mM Glc and 10 mM Glc1P with varying concentrations

of BCN (0-200 mM), which showed that there was no detectable inhibition of Pro_7066 activity by BCN (Figure 3.21). This is surprising considering the interactions between BCN and the enzyme active site, which likely mimics the interaction of the enzyme with the hydroxy groups on Glc1P.

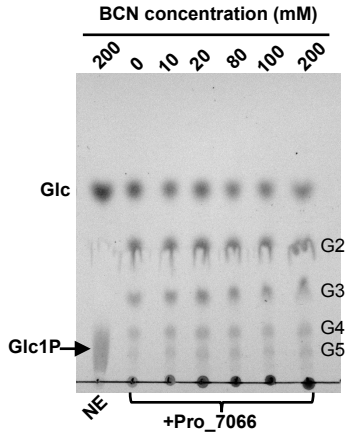


Figure 3.21. TLC analysis of Pro_7066-catalysed glycan synthetic reactions in the presence of BCN. The reactions were performed in the presence of 10 mM Glc and 10 mM Glc1P as substrates and BCN as buffer at the indicated concentrations. TLC was developed using a mobile phase consisting of 1-butanol:acetic acid:water (4:8:1).

In attempting to exclude BCN from the active site, further crystallisation was performed as previously described, but in the absence of BCN. Single crystals were obtained and sent for data collection. However, the exclusion of BCN resulted in crystals with no detectable X-ray diffraction. Since BCN appeared to be important for crystal diffraction, reducing the concentration of BCN may prevent competitive binding of BCN to the active site, whilst maintaining crystal diffraction. Optimisation was performed in the presence of lower concentration of BCN (15 mM instead of 100 mM) and the ligand concentration was increased (30 mM) to conceivably outcompete BCN binding. Laminaribiose (G2) was chosen as a co-crystallised ligand in this experiment due to its shorter chain length (DP 2), which may have better access to the active site in comparison to G6. The crystallisation under these conditions resulted in good quality crystals that diffracted to 2.15 Å resolution. A structure solved from the collected data was designated as G2 complex. Analysis of the active site of the G2 complex revealed that BCN and sulphate molecules still occupy the active site. Surprisingly, residual density near the SBS in the G2 complex resembles a G6 rather than a G2 structure, and the amino acids making contacts with the oligosaccharide are identical to those found in the G6 complex. It is possible that the G2 ligand was polymerised by Pro_7066 during co-crystallisation into longer oligosaccharides which subsequently bound to the SBS. This hypothesis is supported by the presence of longer products when the

enzyme was incubated with G2 in the absence of inorganic phosphate (Figure 3.22), suggesting that residual from phosphate in the enzyme preparation is sufficient to support catalysis.

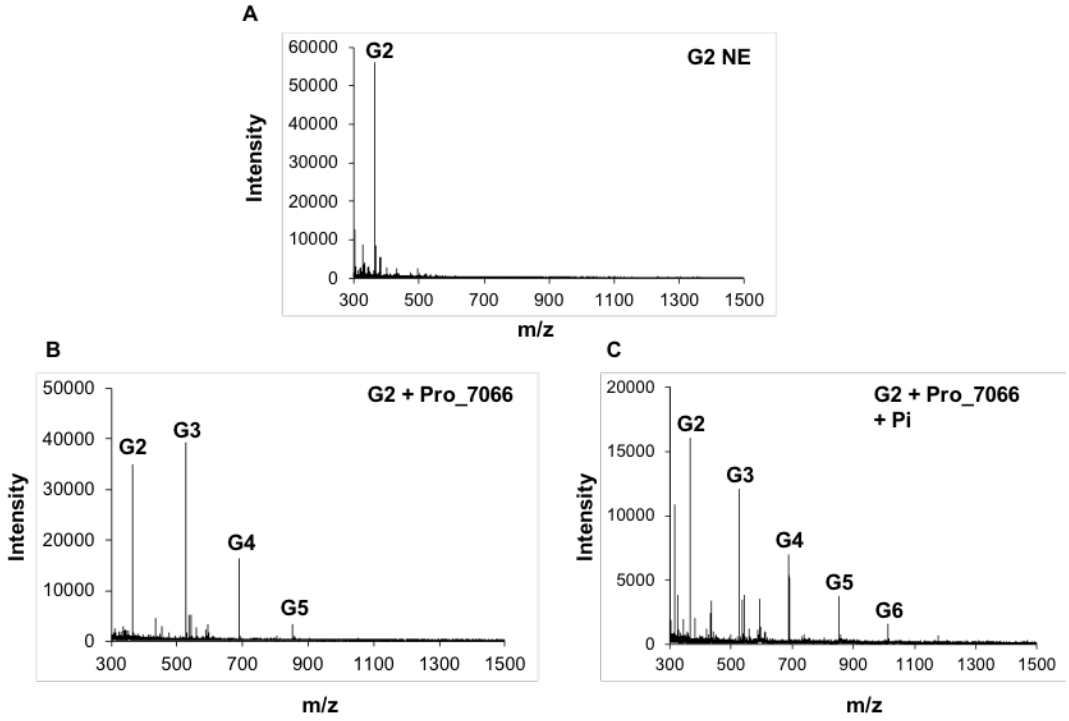


Figure 3.22. MALDI-ToF analysis of the reaction carried out by Pro_7066 in the presence of G2 and either the presence or absence of inorganic phosphate. (**A**) G2 only. NE = no enzyme control. (**B**) Products of the reaction in the absence of inorganic phosphate. (**C**) Products of the reaction in the presence of inorganic phosphate. G2-G5 indicates the DP of the glucan products associated with the peaks.

Superposition of the G6 and G2 complexes showed that the two structures are essentially identical (dimer on dimer superposition, RMSD = 0), therefore the G2 complex was not analysed in further detail. Data collection and processing statistics for all Pro_7066 structures are summarised in Table 3.7. Detailed experiments for crystallisation optimisation can be found in chapter 6, section 6.15.2.

Table 3.7. X-ray data collection and processing of Pro_7066 structures.

Data set	Mercury derivative	Substrate-free complex	G6 complex	G2 complex
Data Collection				
Beamline	I04	I04	I04	I03
Wavelength (Å)	1.0052	0.9795	0.9795	0.9800
Detector	Pilatus 6M	Pilatus 6M	Pilatus 6M	Pilatus 6M
Resolution range (Å)[a]	62.26-2.55 (2.59-2.55)	79.49-2.05 (2.09-2.05)	66.41-2.25 (2.29-2.25)	59.56-2.15 (2.19-2.15)
Space Group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
a, b, c (Å)	99.8 159.3 180.9	100.2,159.0,181.6	99.1,158.8,178.9	99.6, 159.1, 178.7
α, β, γ (°)	90.0, 90.0, 90.0	90.0,90.0,90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0
Total observations[a]	2559459 (127918)	2447545 (123380)	1808389 (88317)	2057640 (105657)
Unique reflections[a]	94654 (4584)	181781 (8926)	134201 (6512)	154486 (7557)
Multiplicity[a]	27.0 (27.9)	13.5 (13.8)	13.5 (13.6)	13.3 (14.0)
Mean *I*/σ(*I*)[a]	11.3 (1.4)	14.7 (1.1)	13.5 (1.4)	8.1 (1.3)
Completeness (%)[a]	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)
R_{merge}[a,b]	0.240 (2.758)	0.107 (2.387)	0.132 (2.024)	0.214 (2.146)
R_{meas}[a,c]	0.244 (2.808)	0.111 (2.478)	0.138 (2.103)	0.223 (2.227)
$CC_{½}$[a,d]	0.998 (0.741)	0.999 (0.549)	0.999 (0.637)	0.996 (0.535)
Wilson B value (Å^2)	44.2	41.4	45.0	36.8
Refinement				
Reflections: working/free[e]	-	181679/9118	134107/6752	154377/7783
R_{work}/ R_{free}[f]	-	0.182/0.211	0.180/0.217	0.195/0.232
Ramachandran plot: favoured/allowed/disallowed (%)[g]	-	97.1/99.9/0.1	96.7/99.7/0.3	96.5/99.7/0.3
R.m.s. bond deviations (Å)	-	0.0119	0.0084	0.0124
R.m.s. angle deviations (°)	-	1.435	1.250	1.486
No. of protein residues (ranges): Chain A / Chain B	-	1138/1138	1138/1135	1138/1135
No. of ligands[g]**/water molecules/other**[h]	-	3/752/6	6/649/4	6/692/5
Mean *B*-factors: protein/ligand[g]**/ /water/other**[h]**/overall (Å^2)**	-	63/60/48/54/62	61/59/48/50/60	48/54/41/43/47

[a] Values for the outer resolution shell are given in parentheses.
[b] $R_{merge} = \sum_{hkl} \sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl} \sum_i I_i(hkl)$.
[c] $R_{meas} = \sum_{hkl} [N/(N-1)]^{1/2} \times \sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl} \sum_i I_i(hkl)$, where $I_i(hkl)$ is the ith observation of reflection hkl, $\langle I(hkl)\rangle$ is the weighted average intensity for all observations i of reflection hkl and N is the number of observations of reflection hkl.
[d] $CC_{½}$ is the correlation coefficient between symmetry-related intensities taken from random halves of the dataset.
[e] The data set was split into "working" and "free" sets consisting of 95 and 5% of the data, respectively. The free set was not used for refinement.
[f] The R-factors R_{work} and R_{free} are calculated as follows: $R = \sum(|F_{obs} - F_{calc}|)/\sum|F_{obs}|$, where F_{obs} and F_{calc} are the observed and calculated structure factor amplitudes, respectively
[g] refers to ligands bound in the active site and potential surface binding sites i.e. sulphate, BCN and G6 depending on the dataset
[h] refers to other ligands, specifically, sulphate bound at a surface site, chloride ions and ethylene glycol molecules from the cryoprotectant

3.2.6. Homology modelling of EgP1

Kinetic studies on the glycan synthetic reaction carried out by EgP1 and Pro_7066 revealed different kinetic parameters, in which EgP1 preferred Glc and G2 as its substrates, whereas the Pro_7066 substrate preferences are more relaxed towards longer substrates. In order to investigate possible molecular mechanism distinguishing the substrate preferences of Pro_7066 and EgP1, crystallographic study was attempted on EgP1 protein. However, crystallisation screens for the EgP1 protein did not result in any crystal hits.

Since EgP1 structure could not be solved experimentally, homology modelling of EgP1 using Pro_7066 structure as a template was performed by Protein Homology/analogy Recognition Engine version 2.0 (Phyre^2).[64] A model of EgP1 monomer was obtained, which is similar to the Pro_7066 template (RMSD = 0.4 based on monomer on monomer superposition). The monomer can be overlaid onto the dimeric structure of Pro_7066, which showed a possible formation of a dimeric EgP1 structure (Figure 3.23A), in agreement with gel filtration analysis of the EgP1 recombinant protein. The overall structure of EgP1 is similar to Pro_7066, comprising an N-terminal β-sandwich (residue 1-308; yellow), a helical linker region (residue 309-336; lilac), an $(\alpha/\alpha)_6$ catalytic domain (residue 337-1042; green), a C-terminal jelly-roll domain (residues 1043-1153; red), Dom 1 (residue 376-442; brown) and Dom 2 (residue 858-940; blue). Inspection on the EgP1 active site showed a complete structural conservation of the residues in the active site, identical to that in Pro_7066 (Figure 3.23B), which is in agreement with the conserved residues that were predicted from the sequence alignment (Figure 3.12). The conserved residues include; 1) R477 and D478; 2) WND motif located on the catalytic loop (W658, N659 and D670); and 3) histidine residue (H960). Superposition of EgP1 on Pro_7066 allow identification of the conserved structural elements surrounding the active sites (Figure 3.23B and C).

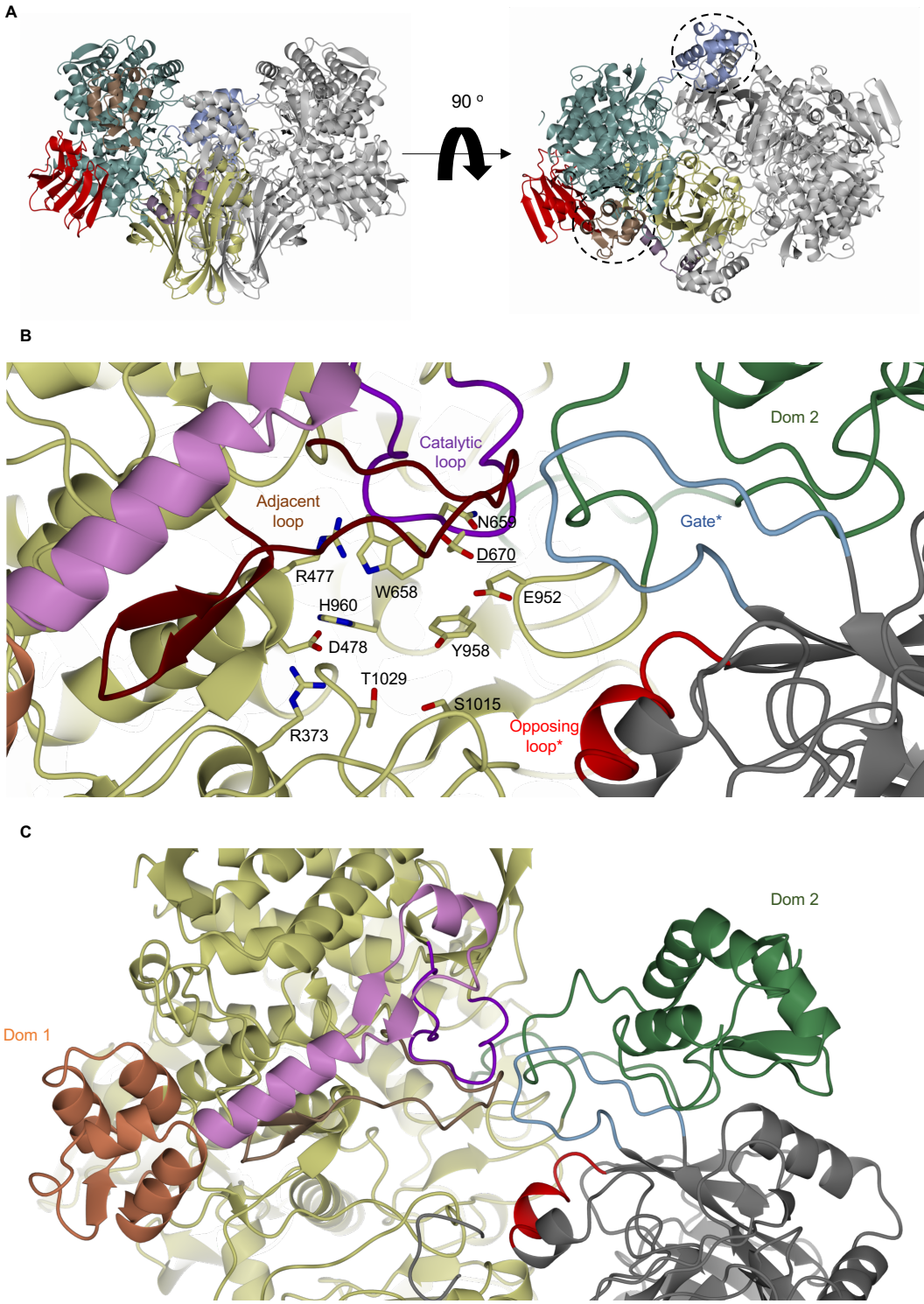


Figure 3.23. EgP1 homology model based on Pro_7066 structure**.** (**A**) The monomer of EgP1 coloured by domains. The monomer EgP1 was superposed onto the dimeric structure of Pro_7066, with the adjacent subunit coloured in grey. (**B**) The active site with side chains of the conserved amino acids involving in the binding of BCN and SO_4^{2-}. Asterisks indicate loops that derived from the superposed Pro_7066 structure. (**C**) Conserved structural elements surrounding the active sites; catalytic loop (purple), Dom 1 (coral), Dom 2 (green), adjacent loop (brown), the upper strand (pink), the gate (blue) and the opposing loop (red).

A potential SBS in EgP1 model was also identified by superposition of EgP1 model on Pro_7066 structure. Some part of the helix that contain YE motif was not

completely modelled (dotted circle, Figure 3.24B), which affected further interpretation of the structure at this site. The YE (Y990 and E991) motif and the equivalence of E990 (E987) could be identified and are structurally conserved in EgP1, in agreement with the sequence alignment (Figure 3.19), whilst the other amino acids are less conserved (Figure 3.24A and B). Other EgP1 orthologues from *E. gracilis* (EgP2, EgP3 and EgP4) were also modelled, which also showed that the SBS sites are not completed in all of the structures and the amino acids constituting the SBSs are not conserved except for the YE motif (Figure 3.24 C-E). The incomplete modelling of the SBSs in EgP structures is probably due to an indel within the sequences in this region.

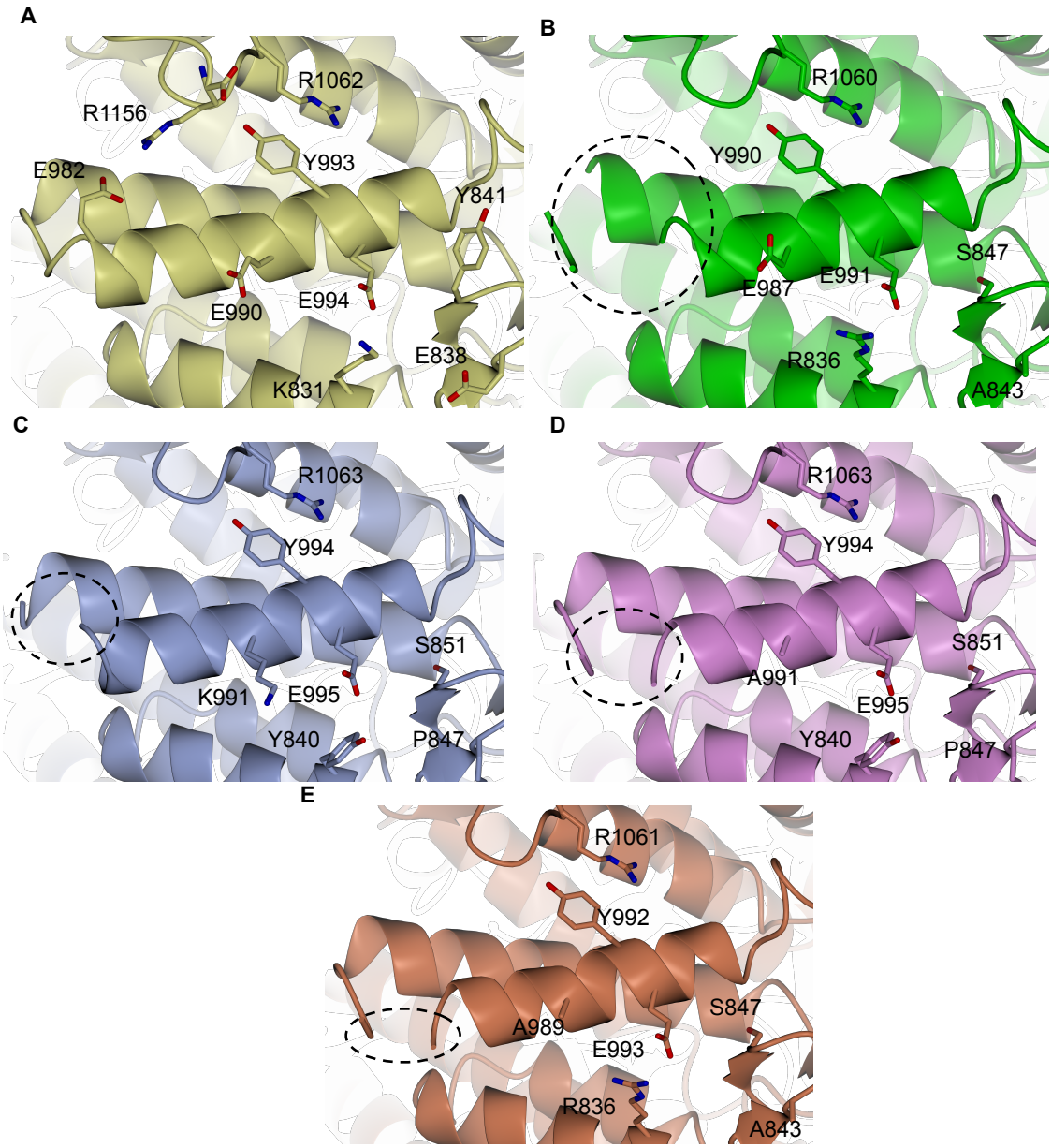


Figure 3.24. Potential SBS sites on EgP1 and other EgP candidates. (**A**) Pro_7066. (**B**) EgP1 model. (**C**) EgP2 model. (**D**) EgP3 model. (**E**) EgP4 model. Dotted ovals indicate incomplete modelling of the SBS in EgP1, 2, 3 and 4.

3.3. Discussion

β-(1→3)-D-Glucan phosphorylase activities have previously been detected in both bacterial LBPs and eukaryote microalgal LDPs. However, only bacterial LBPs were fully characterised; their sequences identified and categorised into the GH94 family. The lack of sequence information for eukaryotic β-(1→3)-D-glucan phosphorylases prevented heterologous expression and further characterisation. It also prevented enzyme classification based on CAZy criteria and prohibited investigation of their relationship to other GH families. Identification of nucleotide sequences encoding proteins with β-(1→3)-D-glucan phosphorylase activity is therefore a crucial step in this study. DNJ-affinity enrichment of *Euglena* protein extract with active β-(1→3)-D-glucan phosphorylase activity coupled with proteomic analysis was employed to identify a sequence candidate, designated EgP1. The EgP1 sequence was then used to interrogate a propriety metagenome database, identifying a second bacterial β-(1→3)-D-glucan phosphorylase candidate, designated Pro_7066. Both proteins were successfully expressed as recombinant proteins in *E. coli*, which allowed us to characterise their functions. Although more closely related to GH94 than to other GH families, EgP1 and Pro_7066 have no significant overall sequence identity to any characterised GH94 sequences, which suggests that they belong to a new CAZyme family, namely GH149 (communication with Professor Bernard Henrissat, Head of the CAZy database).

In vitro characterisation confirmed the function of EgP1 and Pro_7066 as β-(1→3)-D-glucan phosphorylases. EgP1 and Pro_7066 used Glc and longer β-(1→3)-linked gluco-oligosaccharides as acceptors in the glycan synthesis. This substrate specificity is slightly different from the previously described β-(1→3)-D-glucan phosphorylases where substrate chain length was used to distinguish between different groups of the phosphorylases (i.e. LBPs and LDPs). The preference of EgP1 for Glc resembles the activity reported for LBP from semi-purified *E. gracilis* extracts.[24,25] However, the findings herein are in agreement with more recent works by Ogawa *et al.* [13] and Muller *et al.,*[14] which described a more relaxed chain length specificity of *E. gracilis* LBP. Therefore, the chain length specificity of β-(1→3)-D-glucan phosphorylases might be more flexible than previously observed. This has also been observed in a recently characterised thermophilic cellodextrin phosphorylase (β-(1→4)-D-glucan phosphorylase), which uses Glc and longer cello-oligosaccharide as

acceptors[65] and is distinct from GH94 cellodextrin phosphorylases, which have no detectable activity on Glc as an acceptor.[20,66]

Comparisons of the kinetic parameters of EgP1 and Pro_7066 showed distinct preferences for the chain length of sugar acceptors; EgP1 preferred Glc and G2 and the k_{cat}/K_M values declined as the chain length rose to G6. Increasing acceptor chain length led to an increase in K_M values while k_{cat} values remained comparable. This result indicates that substrate binding (K_M) dictates the acceptor chain length preference in EgP1 rather than the turnover (k_{cat}) of the enzyme-substrate complex to product. In contrast, Pro_7066 showed comparable catalytic efficiency as well as k_{cat} and K_M towards all acceptors. The contrasting kinetic parameters of EgP1 and Pro_7066 were surprising considering that the two proteins are 45% identical.

Overall k_{cat}/K_M for the glycan synthetic reaction of EgP1 and Pro_7066 are below 10 $mM^{-1}s^{-1}$, which are within the similar range to what has been reported for *Rt*CDP.[20] However, these values are several folds smaller than those reported for GH94 CBPs (Table 3.8), highlighting the difference between disaccharide and oligosaccharide phosphorylases in their kinetic parameters.

Table 3.8. Summary of k_{cat}/K_M values for the synthetic reaction carried out by GPs from GH94.

Protein acronym	Organisms	k_{cat}/K_M ($mM^{-1}s^{-1}$)
Oligosaccharide phosphorylases		
***Rt*CDP[20]**	*Ruminiclostridium thermocellum*	Glc = N/A (β-(1→4)-Glc)$_2$ = 6.5 (β-(1→4)-Glc)$_3$ = 14 (β-(1→4)-Glc)$_4$ = 9.3 (β-(1→4)-Glc)$_5$ = 12 (β-(1→4)-Glc)$_6$ = 4
***Lp*SOGP[48]**	*Lachnoclostridium phytofermentans*	Glc = N/A (β-(1→2)-Glc)$_2$ = 7.3 (β-(1→2)-Glc)$_3$ = 20 (β-(1→2)-Glc)$_4$ = 21 (β-(1→2)-Glc)$_5$ = N/A Glc1P = 11
Disaccharide phosphorylases		
***Cg*CBP[46]**	*Cellvibrio gilvus*	Glc = 46.8
***Cu*CBP[67]**	*Cellulomonas uda*	Glc = 19.6
***Tm*CBP[68]**	*Thermotoga maritima* MSB8	Glc = 110
***Ct*CBP[69]**	*Ruminiclostridium thermocellum* YM4	Glc = 14.6
***Sd*CBAP[47]**	*Saccharophagus degradans*	D-Gluconic acid = 76 D-GlcA = 2.58
XCC4077 (CBAP)[70]	*Xanthomonas campestris*	D-GlcA = 87 D-GlcU = 18 Glc1P = 605
***Ra*CBP[71]**	*Ruminococcus albus* NE1	Glc = 63.2

GlcA = glucuronic acid, Glc = D-glucose, Glc1P = α-D-glucose 1-phosphate. N/A indicates where the specific activity could not be detected or determined. CBP = cellobiose phosphorylase, CDP = cellodextrin phosphorylase, CBAP = cellobionic acid phosphorylase.

GH149 and GH94 members share conserved amino acids that are required for the phosphorylase activity but, overall, are not significantly similar. It is likely that phosphorylase activity in EgP1 and Pro_7066 proceeds via a similar mechanism to that deciphered in GH94 enzymes.[44–47,72] Therefore, GH149 and GH94 form a clan of related families, likely to have evolved from a common ancestor and diversified while retaining the key catalytic apparatus and substrate binding sites. This hypothesis is further supported by the structural studies of Pro_7066 (GH149 family) and the homology model of EgP1 which confirmed structural conservation of the Asp catalytic residue, the residues forming the -1 donor subsite and the histidine previously predicted for phosphate recognition.

On the other hand, the conservation of the acceptor binding subsite could not be determined in this work due to the lack of oligosaccharide acceptor binding to the Pro_7066 and *Ps*LBP active sites. In Pro_7066, the lack of oligosaccharide binding could be due to two possible reasons. Firstly, the presence of BCN may interfere with the oligosaccharide binding, although this interaction did not inhibit the enzyme activity. Secondly, the enzyme was active during the crystallisation process and elongated the short oligosaccharide to polysaccharide. The interaction of long oligosaccharide with

the active site may be too dynamic to be captured during crystallisation. This hypothesis is supported by the presence of residual density that resembled G6 rather than G2 at the SBS when G2 was used as a co-crystallised ligand and the detection of oligosaccharide production in the synthetic reaction carried out without inorganic phosphate addition. The oligosaccharide production without inorganic phosphate would suggest the presence of transglycosylation by the enzyme, although it is highly possible that a trace amount of inorganic phosphate may be present in the crystallisation solution, which could trigger the observed glycan synthesis. In order to avoid the elongation of the co-crystallised ligands, further crystallisation could be performed using Pro_7066 with mutated catalytic residue to inhibit the conversion of the oligosaccharide ligand to longer products.

A unique N-terminal α/β domain containing an extended arm located at the lower portion of GH94 *Rt*CDP is closely associated with the β-sandwich domain of the adjacent subunit. Such interaction has been predicted to draw the lower portion of the subunits together whilst moving the upper portion apart, effectively widening the active site to accommodate larger acceptor.[20] The lack of the N-terminal domain in the Pro_7066 structure suggests that Pro_7066 employs different mechanism to widening the acceptor binding site to accommodate longer substrates.

Conserved structural elements neighbouring the active site of *Ps*LBP and Pro_7066 can be identified. These elements are the gate, the opposing, the adjacent and the catalytic loops. There is a noticeable difference in the gate conformation, which is disordered in Pro_7066 but formed a β-hairpin in *Ps*LBP. The highly-structured β-hairpin gate in *Ps*LBP likely occludes the acceptor binding subsite which restricts the chain length of the enzyme substrate/product, whilst the disordered gate in Pro_7066 may provide more flexibility to allow accommodation of longer substrates. Mutagenesis of the gate should be performed in order to study the role of this structure in defining the enzyme substrate/product chain length preference.

In addition to the structural elements within/surrounding the active sites, remote auxiliary sites were identified in Pro_7066 structure. Two potential CBM candidates (Dom 1 and 2) were found flanking near the N- and the C-terminal of the catalytic domain. The position of Dom 1 and 2 in Pro_7066 are similar to those described for a CBM4 modules flanking a catalytic domain in an endo-β-glucanase acting on β-(1$\rightarrow$3;1$\rightarrow$4) mixed-linked glucan.[73] The location of Dom 1 and 2 is unusual for CBM,

considering the common reported locations of CBMs at the N- or C-terminal of GHs.[74] Moreover, Dom 1 and 2 contain predominantly α-helical structures augmented by small β-sheets, which is unexpected for β-(1→3)-glucan-recognising CBMs that normally adopt β-sandwich folds. A CBM43 containing α-helical chains and small β-sheet has been reported to be associated with a GH72 enzyme, which is a transglycosidase that elongates β-(1→3)-glucan chain.[53] However, there is no report on interaction of the CBM43 to β-(1→3)-glucan substrate, raising the possibility that the CBM may recognise other carbohydrates. The recognition of β-(1→3)-glucan by CBMs often involves several surface-exposed aromatic residues.[75] However, in the CBM43 associated with GH72, the surface-exposed aromatic residues have negligible effect on the GH72 enzyme activity.[53] In Dom 1 and 2, all aromatic residues are not surface-exposed, thus unlikely to participate in β-(1→3)-glucan binding. Therefore, the exact roles of Dom 1 and 2 cannot be concluded based on these findings. Further investigation on the roles of Dom 1 and 2, in particular on their interaction with oligosaccharide ligand, must be performed to confirm their contribution to the activity of Pro_7066 on oligosaccharide substrates.

The oligosaccharide SBS discovered in Pro_7066 is located on the solvent-exposed surface of the catalytic domain. Since the active site is deeply buried within the catalytic domain, there is no obvious direct path connecting the active site to the SBS. Thus, it is unlikely that the SBS is involved in direct loading of long glucan chain to the enzyme active site. The presence of SBS on the opposite site of the active site has been observed in a barley α-amylase, of which the SBS is predicted to be involved in substrate targeting.[76,77] Thus, it is possible that the SBS in Pro_7066 may perform similar substrate targeting function.[78] The flat conformation of G6 ligand at the SBS and the lack of intra- and inter-strand hydrogen bonds in the ligand strongly indicate that the G6 ligand assumed a linear rather than triple helical conformation while interacting with the Pro_7066 SBS. Therefore, the Pro_7066 SBS may also perform a substrate disruption role by breaking the triple helical structure of G6 into a linear chain.

Comparisons between Pro_7066 and other GH94 enzymes showed that there is no conservation of the amino acid residues comprising the SBS. Therefore, the Pro_7066 SBS appears to be a unique oligosaccharide binding site which may contribute to Pro_7066 activity on long glucan substrates. In contrast, the absence of SBS in PsLBP may contribute to the lack of activity of PsLBP on longer oligosaccharides. The absence of SBS in RtCDP may be explained by the fact that β-

(1→4)-glucan with DP greater than 9 precipitates in solution (unpublished result, communication with Dr Giulia Pergolizzi, John Innes Centre), thus preventing the extension of the oligos beyond this DP, therefore the SBS for RtCDP is probably not required for recognition of long glucan substrates. The lack of SBS in LpSOGP despite the enzyme being able to synthesis soluble β-(1→2)-oligoglucan with an average DP of 39 [79] may imply that LpSOGP utilises unique structural features that were not identified in the previously reported structure.[48] Determination of LpSOGP structure in complex with long β-(1→2)-oligoglucan could be performed in order to confirm this hypothesis.

The GH149 family contains several hundred EgP1 orthologues from bacteria in phyla Bacteroidetes and Proteobacteria and from eukaryotes in phylum Euglenophyta, class Euglenophyceae. Phylogenetic analysis of GH149 sequences revealed a clade of Euglenophyceae sequences from *E. gracilis* and *Eutreptiella gymnastica* under a common node. This finding suggests a common origin for GH149 sequences found in *E. gracilis* (fresh water living) and *E. gymnastica* (sea water living), indicated that these sequences were inherited from a common Euglenophyceae ancestor before topographic isolation of *Euglena* and *Eutreptiella*. The origin of the Euglenophyceae ancestral GH149 is unclear, however, the position of the Euglenophyceae clade within bacteria, might suggest a lateral gene transfer from bacteria to the Euglenophyceae. A similar horizontal gene transfer from a bacterium to *E. gracilis* was predicted for mitochondrial trans-2-enoyl-CoA reductase.[80] Nevertheless, it is important to note that bacteria genomes are overrepresented in most available sequence databases compared to eukaryote counterparts, therefore the prevalence of the bacterial orthologues in our study is unsurprising. Increasing the availability of sequences from the diversity of unicellular eukaryotes is essential for investigating evolutionary origins.

A predicted targeting peptide in EgP1 caused aggregation of the EgP1 recombinant protein when expressed in *E. coli.* Removing this peptide from the sequence enabled solubility without affecting the expression level or activity of the protein. The targeting peptide length is consistent with the average length of predicted mitochondrial targeting peptides in *Euglena* previously reported by Krnáčová *et al*.[81] but it is still not known whether the targeting peptide is functional *in vivo*. Further investigation of the sub-cellular localisation of the full length EgP1 in *E. gracilis* cells is required to elucidate its biological function. In contrast, no target peptide was predicted for Pro_7066 and the length of this sequence was similar to the putative mature EgP1 protein (Figure 3.12). Similarly, no targeting peptide was predicted for the GH149 sequences found in *Eutreptiella*.

Examination of the genomic location of *GH149* open reading frames (ORF) in Proteobacteria revealed their presence in putative gene clusters; Type I clusters containing ORFs for an ABC transporter cassette [a substrate binding protein (COG0747), two permeases (COG0601 and 1173) and two ATP-binding proteins (COG0444 and 4608)] and a transcriptional regulator LacI (COG1609) or AraC (COG2207)) (Figure 3.25A). The same type of ABC transporter has been reported to be used by a thermophilic Gram-negative bacteria *Thermotoga maritima* for the utilisation of extracellular laminarin after degradation by an extracellular laminarinase. Oligosaccharide products from laminarinase-catalysed reactions are predicted to be taken up by the inner membrane ABC transporter into the cytoplasm, where the oligosaccharides could be broken down into glucose by a laminaribiase.[82] The lack of genes encoding extracellular laminarinase and an outer membrane transporter in Type I clusters suggests that this cluster may be involved in scavenging low MW β-(1→3)-gluco-oligosaccharides such as laminaritriose/biose using the putative GH149 phosphorylase as a catalyst (Figure 3.26A). Moreover, the lack of a gene predicted for an outer membrane transporter suggests that the gene encoding the transporter might be located elsewhere in the genomes. In our analysis, GH149 sequences were found in many *Vibrio* species including *Vibrio campbellii* HY01, which has recently been reported to express a novel GH3 family β-glucosidase (LamN) capable of digesting laminaribiose.[83] The co-occurrence of a GH149 and LamN in *V. campbellii* supports the utilisation of β-(1→3)-D-glucan by *Vibrio* spp., reinforcing its importance in the carbon cycle of marine environments.

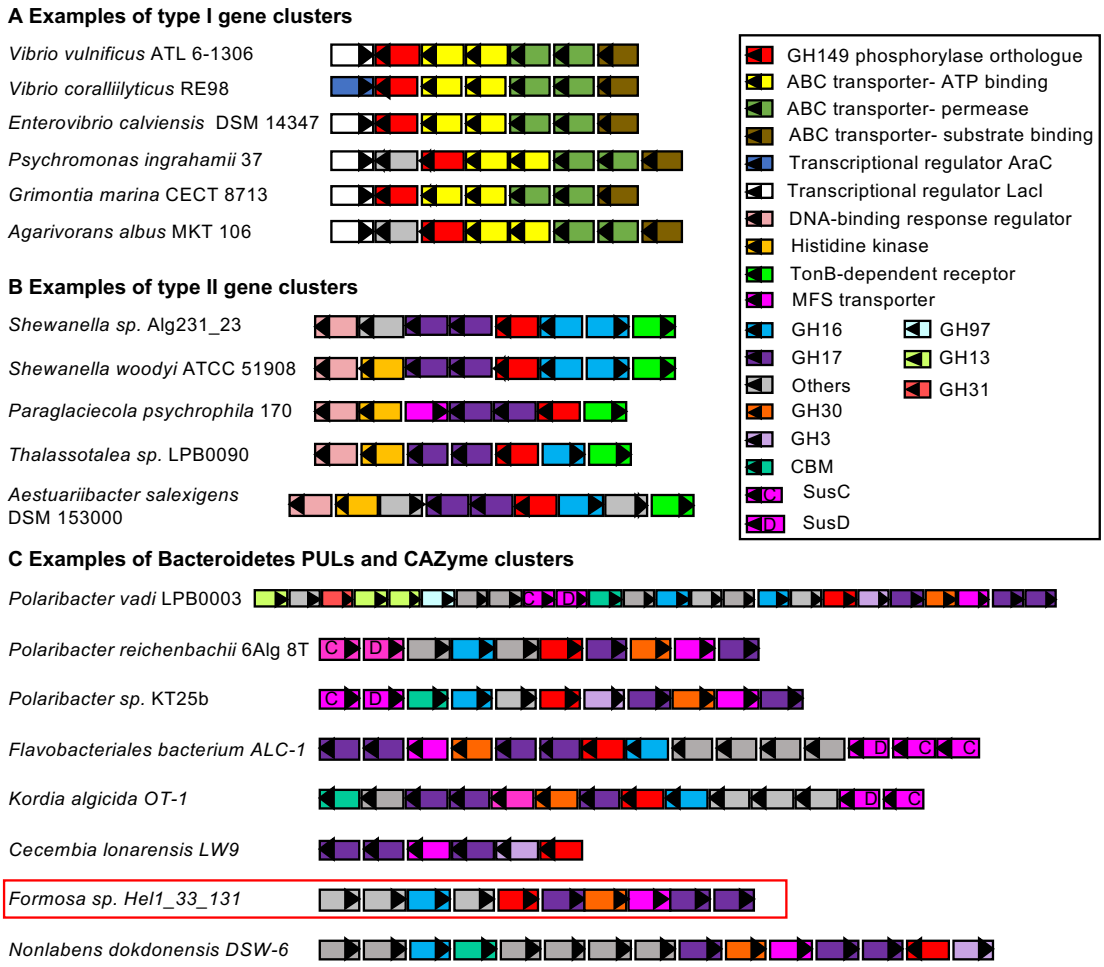


Figure 3.25. Genetic organisation of putative gene clusters containing the bacterial *GH149* genes. (**A**) Type I gene clusters contain genes encoding ABC-transporter cassette, the phosphorylase, and a transcriptional regulator. (**B**) Type II gene clusters contain genes encoding a TonB-dependent receptor together with GHs and transcriptional regulators. (**C**) Predicted PULs and CAZyme clusters containing the putative GH149s from *Bacteroidetes* species.

A second genetic co-localisation pattern was also observed (designated Type II) mainly in genomes of bacteria in class γ-proteobacteria, order Alteromonadales. Type II clusters contain the respective *GH149* genes co-localised with gene encoding a TonB-dependent receptor (cd01347), which is an outer membrane transporter involved in the transport of iron siderophore, vitamin B_{12}, nickel complexes, and carbohydrates into the periplasmic space.[84] Type II clusters also contain genetic components encoding other CAZymes [extracellular GH16 (Pfam11721), cytoplasmic GH17 (COG5309)], regulators [histidine kinase (Pfam06580) and DNA-binding response regulator (COG3279)], and an inner membrane Major Facilitator Superfamily (MFS) transporter (Pfam13347) (Figure 3.25B). A more complicated organisation of type II cluster suggests that they might be involved in degrading more complex β-(1→3)-glucan structures than that utilised by type I cluster (Figure 3.26B). Most importantly, the majority of *GH149* genes from Bacteroidetes map to the previously

predicted PULs (clusters containing *susCD*-like gene pairs) and CAZyme clusters (clusters containing CAZymes but lack *susCD*-like gene pairs) (Figure 3.25C).[36,85] The common features that are found in these PULs and CAZyme clusters are the presence of genes encoding β-glucosidases (GH3 and/or GH30_1) and glycanases able to hydrolyse β-(1→3)-glucans (GH17 and/or GH16), thus emphasize the potential contribution of the GH149 in the degradation of complex β-(1→3)-glucan structures (Figure 3.26C). The importance of PULs containing GH149 in laminarin degradation was further highlighted by Unfried et al., who have performed a comprehensive study of PUL from *Formosa* sp. Hel1_33_13 (Figure 3.25C, red box) and its role in laminarin turnover during diatom-dominated spring bloom.[86] Expression of the *Formosa* PUL was detected in metagenome and metaproteome analyses of the water sample collected during the bloom, suggesting a correlation between the upregulation of PUL expression and the utilisation of laminarin by the bacterium. Moreover, a GH30 and two GH17 enzymes in the Formosa PUL were expressed as recombinant proteins and biochemically characterised to confirm their hydrolytic activity on laminarin, strongly indicating the roles of these enzymes and their respective PUL in laminarin degradation.

The phylogenetic clustering of the bacterial GH149 amino acid sequences in Figure 3.11B agreed with the taxonomic relationships of the species, suggesting that these bacterial *GH149* genes are likely inherited through a vertical gene transfer within each species. Comparison of the GC content of 52 bacterial *GH149* genes (Appendix 6) showed limited deviation from the median GC content of the corresponding complete genome sequences, supporting the vertical gene transfer hypothesis.

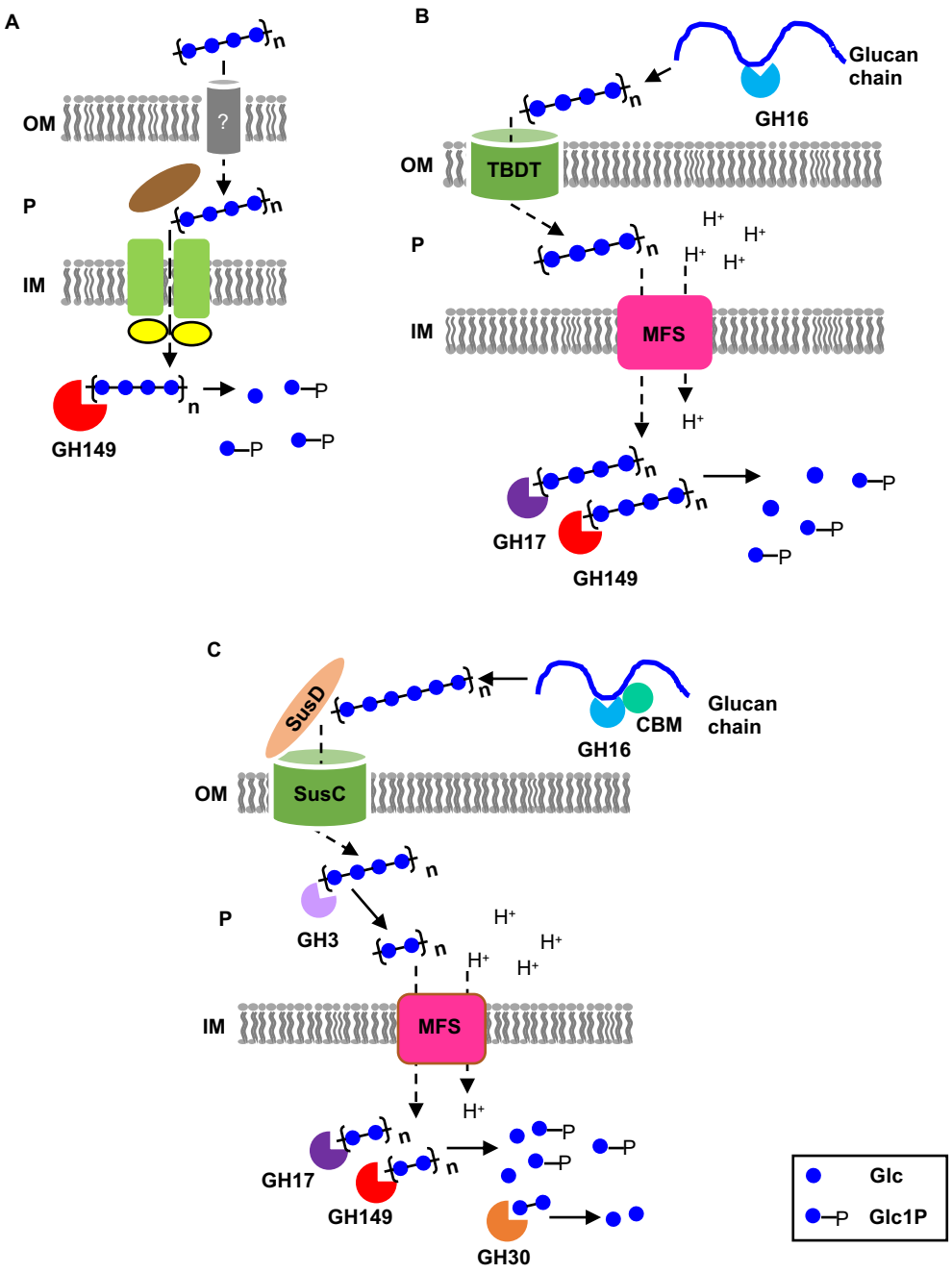


Figure 3.26. Proposed mechanisms of β-(1→3)-glucan utilisation in bacteria containing *GH149* gene clusters. (**A**) Possible mechanism for Type I gene cluster in *Vibrio vulnificus* ATL 6-13036. The components of ABC transporters are coloured as followed; green (permeases), yellow (ATP binding) and brown (substrate binding). (**B**) Possible mechanism for Type II gene cluster in *Shewanella* sp. Alg231_23. (**C**) Possible mechanism for PUL in *Polaribacter* sp. KT25b. OM = outer membrane, P = periplasmic space, IM = inner membrane, MFS = Major Facilitator Superfamily transporter, TBDT = TonB-dependent transporter. Glc and Glc1P are released from the oligosaccharide degradation out by GH149. Glc1P can be converted to Glc6P by phosphoglucomutase and used in glycolysis. The subcellular localisation prediction of GHs and transporters was performed by PSORTb version 3.0.2 (http://psort.org/psortb/index.html) using default settings for Gram-negative bacteria.

3.4. Conclusion and future perspectives

In summary, the genetic loci containing *GH149* genes revealed co-localisation of *GH149* with genes encoding several families of GH as well as sugar transporters, suggesting the biological roles of bacterial GH149 protein in polysaccharide degradation in particular β-(1→3)-glucan from external sources. Similar GC contents of the *GH149* genes and their associated bacterial genomes suggests that the bacterial *GH149* genes were inherited through vertical gene transfer. Variation in the architecture of gene clusters containing the *GH149* genes suggests that the inherited *GH149* genes were independently associated into the gene clusters and PULs of individual species. The involvement of novel phosphorylases in β-(1→3)-glucan degradation emphasises the importance of gene clusters containing CAZymes and their roles in carbohydrate metabolism in marine Gram-negative bacteria. The discovery of the eukaryotic GH149 sequences enables further assessments of their physiological function in Euglenophyceae using genetic manipulation. Alternatively, the expression level of *EgP1* could be assessed to determine whether the gene expression is associated with paramylon accumulation or breakdown, the approach of which has recently been used to investigate possible roles of other predicted GTs and GHs associated with paramylon metabolic pathways.[87]

The presence of Dom 1, Dom 2 and oligosaccharide SBS provides preliminary explanation for Pro_7066's ability to produce long oligosaccharide chains. However, further investigation on the roles of these elements in this enzyme is needed to understand the precise molecular mechanisms by which these non-catalytic sites operate and their contributions, if any, to the enzyme chain length preference. Mutagenesis of the non-catalytic sites complemented by biochemical and biophysical assays such as isothermal titration calorimetry and surface plasmon resonance of the mutants would provide evidence for the contribution of these non-catalytic sites on the enzyme specificity. Finally, this study revealed for the first time the SBS associated with GPs acting on β-(1→3)-glucan, providing a major stepping stone for future research on GPs and their non-catalytic auxiliary domains.

3.5. References

[1] S. Kuhaudomlarp, N. J. Patron, B. Henrissat, M. Rejzek, G. Saalbach, R. A. Field, *J. Biol. Chem.* **2018**, *293*, 2865–2876.

[2] E. C. O'Neill, R. A. Field, *Carbohydr. Res.* **2015**, *403*, 23–37.

[3] G. Pergolizzi, S. Kuhaudomlarp, R. A. Field, E. Kalita, *Protein Pept. Lett.* **2017**, *24*, 696–709.

[4] A. Bacic, G. B. Fincher, B. A. Stone, *Chemistry, Biochemistry, and Biology of (1→3)-β-Glucans and Related Polysaccharides*, Academic Press/Elsevier, San Diego, **2009**.

[5] C. S. Buller, *Water Insoluble Polysaccharide Polymer and Method Thereof*, **1990**, US4908310 A.

[6] P. P. Jagodzinski, R. Wiaderkiewicz, G. Kurzawski, M. Kloczewiak, H. Nakashima, E. Hyjek, N. Yamamoto, T. Uryu, Y. Kaneko, M. R. Posner, D. Kozbor, *Virology* **1994**, *202*, 735–745.

[7] T. Watanabe, R. Shimada, A. Matsuyama, M. Yuasa, H. Sawamura, E. Yoshida, K. Suzuki, *Food Funct.* **2013**, *4*, 1685–1690.

[8] P. Kankkunen, L. Teirila, J. Rintahaka, H. Alenius, H. Wolff, S. Matikainen, *J. Immunol.* **2010**, *184*, 6335–6342.

[9] M. Shibakami, G. Tsubouchi, M. Sohma, M. Hayashi, *Carbohydr. Polym.* **2015**, *119*, 1–7.

[10] L. Barsanti, R. Vismara, V. Passarelli, P. Gualtieri, *J. Appl. Phycol.* **2001**, *13*, 59–65.

[11] S. Trouvelot, A.-L. Varnier, M. Allègre, L. Mercier, F. Baillieul, C. Arnould, V. Gianinazzi-Pearson, O. Klarzynski, J.-M. Joubert, A. Pugin, X. Daire, *Mol. Plant. Microbe. Interact.* **2008**, *21*, 232–243.

[12] M. Hrmova, T. Imai, S. J. Rutten, J. K. Fairweather, L. Pelosi, V. Bulone, H. Driguez, G. B. Fincher, *J. Biol. Chem.* **2002**, *277*, 30102–30111.

[13] Y. Ogawa, K. Noda, S. Kimura, M. Kitaoka, M. Wada, *Int. J. Biol. Macromol.* **2014**, *64*, 415–419.

[14] C. Müller, T. Ortmann, A. Abi, D. Hartig, S. Scholl, H.-J. Jördening, *Appl. Biochem. Biotechnol.* **2017**, *182*, 197–215.

[15] E. C. O'Neill, M. Trick, L. Hill, M. Rejzek, R. G. Dusi, C. J. Hamilton, P. V. Zimba, B. Henrissat, R. A. Field, *Mol. BioSyst.* **2015**, *11*, 2808–2820.

[16] E. C. O'Neill, M. Trick, B. Henrissat, R. A. Field, *Perspect. Sci.* **2015**, *6*, 84–93.

[17] I. M. Ivanova, S. A. Nepogodiev, G. Saalbach, E. C. O'Neill, M. D. Urbaniak, M. A. J. Ferguson, S. S. Gurcha, G. S. Besra, R. A. Field, *Carbohydr. Res.* **2017**,

438, 26–38.

[18] M. Rejzek, L. Hill, E. S. Hems, S. Kuhaudomlarp, B. A. Wagstaff, R. A. Field, *Methods Enzymol.* **2017**, *597*, 209–238.

[19] E. O'Neill, S. Kuhaudomlarp, M. Rejzek, J. Fangel, K. Alagesan, D. Kolarich, W. Willats, R. Field, *Biology* **2017**, *6*, 45.

[20] E. C. O'Neill, G. Pergolizzi, C. E. M. Stevenson, D. M. Lawson, S. A. Nepogodiev, R. A. Field, *Carbohydr. Res.* **2017**, *451*, 118–132.

[21] E. C. O'Neill, A. M. Rashid, C. E. M. Stevenson, A.-C. Hetru, A. P. Gunning, M. Rejzek, S. A. Nepogodiev, S. Bornemann, D. M. Lawson, R. A. Field, *Chem. Sci.* **2014**, *5*, 341–350.

[22] S. H. Goldemberg, L. R. Maréchal, B. C. De Souza, L. R. Mar, B. C. D. E. Souza, *J. Biol. Chem.* **1966**, *241*, 45–50.

[23] M. Kitaoka, *Appl. Microbiol. Biotechnol.* **2015**, *99*, 8377–8390.

[24] L. R. Maréchal, S. H. Goldemberg, *Biochem. Biophys. Res. Commun.* **1963**, *13*, 2–5.

[25] M. Kitaoka, T. Sasaki, H. Taniguchi, *Arch. Biochem. Biophys.* **1993**, *304*, 508–514.

[26] D. J. Manners, D. Taylor, *Arch. Biochem. Biophys.* **1967**, *121*, 443–451.

[27] M. Kitaoka, Y. Matsuoka, K. Mori, M. Nishimoto, K. Hayashi, *Biosci. Biotechnol. Biochem.* **2012**, *76*, 343–348.

[28] T. Nihira, Y. Saito, M. Kitaoka, M. Nishimoto, K. Otsubo, H. Nakai, *Carbohydr. Res.* **2012**, *361*, 49–54.

[29] L. R. Marechal, *Biochem. Biophys. Acta* **1967**, *146*, 417–430.

[30] G. J. Albrecht, H. Kauss, *Phytochemistry* **1971**, *10*, 1293–1298.

[31] Y. Yamamoto, D. Kawashima, A. Hashizume, M. Hisamatsu, N. Isono, *Biosci. Biotechnol. Biochem.* **2013**, *77*, 1949–1954.

[32] V. Lombard, H. Golaconda Ramulu, E. Drula, P. M. Coutinho, B. Henrissat, *Nucleic Acids Res.* **2014**, *42*, D490–D495.

[33] M. I. Trindade, V. R. Abratt, S. J. Reid, *Appl. Environ. Microbiol.* **2003**, *69*, 24–32.

[34] T. Nihira, E. Suzuki, M. Kitaoka, M. Nishimoto, K. Ohtsubo, H. Nakai, *J. Biol. Chem.* **2013**, *288*, 27366–27374.

[35] K. Chiku, T. Nihira, E. Suzuki, M. Nishimoto, M. Kitaoka, K. Ohtsubo, H. Nakai, *PLoS One* **2014**, *9*, e114882.

[36] N. Terrapon, V. Lombard, H. J. Gilbert, B. Henrissat, *Bioinformatics* **2015**, *31*, 647–655.

[37] J. M. Grondin, K. Tamura, G. Déjean, D. W. Abbott, H. Brumer, *J. Bacteriol.*

2017, *199*, e00860-16.

[38] Y. Yoshida, T. Tomiyama, T. Maruta, M. Tomita, T. Ishikawa, K. Arakawa, *BMC Genomics* **2016**, *17*, 182.

[39] S. Fushinobu, M. Hidaka, A. M. Hayashi, T. Wakagi, H. Shoun, M. Kitaoka, *J. Appl. Glycosci.* **2011**, *58*, 91–97.

[40] I. N. Cruz, C. S. Barry, H. B. Kramer, C. C. Chuang, S. Lloyd, A. C. van der Spoel, F. M. Platt, M. Yang, B. G. Davis, *Chem. Sci.* **2013**, *4*, 3442–3446.

[41] T. Takeda, Y. Nakano, M. Takahashi, N. Konno, Y. Sakamoto, R. Arashida, Y. Marukawa, E. Yoshida, T. Ishikawa, K. Suzuki, *Phytochemistry* **2015**, *116*, 21–27.

[42] N. S. Berrow, D. Alderton, S. Sainsbury, J. Nettleship, R. Assenberg, N. Rahman, D. I. Stuart, R. J. Owens, *Nucleic Acids Res.* **2007**, *35*, e45.

[43] E. Krissinel, *Nucleic Acids Res.* **2015**, *43*, W314–W319.

[44] A. Van Hoorebeke, J. Stout, J. Kyndt, M. De Groeve, I. Dix, T. Desmet, W. Soetaert, J. Van Beeumen, S. N. Savvides, *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **2010**, *66*, 346–351.

[45] M. Hidaka, Y. Honda, M. Kitaoka, S. Nirasawa, K. Hayashi, T. Wakagi, H. Shoun, S. Fushinobu, *Structure* **2004**, *12*, 937–947.

[46] M. Hidaka, M. Kitaoka, K. Hayashi, T. Wakagi, H. Shoun, S. Fushinobu, *Biochem. J.* **2006**, *398*, 37–43.

[47] Y.-W. Nam, T. Nihira, T. Arakawa, Y. Saito, M. Kitaoka, H. Nakai, S. Fushinobu, *J. Biol. Chem.* **2015**, *290*, 18281–18292.

[48] M. Nakajima, N. Tanaka, N. Furukawa, T. Nihira, Y. Kodutsumi, Y. Takahashi, N. Sugimoto, A. Miyanaga, S. Fushinobu, H. Taguchi, H. Nakai, *Sci. Rep.* **2017**, *7*, 42671.

[49] S. Kuhaudomlarp, S. Walpole, C. E. M. Stevenson, S. A. Nepogodiev, D. M. Lawson, J. Angulo, R. A. Field, *ChemBioChem* **2018**, *19*, in press.

[50] L. Holm, L. M. Laakso, *Nucleic Acids Res.* **2016**, *44*, W351–W355.

[51] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, A. Bateman, *Nucleic Acids Res.* **2016**, *44*, D279–D285.

[52] A. Marchler-Bauer, Y. Bo, L. Han, J. He, C. J. Lanczycki, S. Lu, F. Chitsaz, M. K. Derbyshire, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, F. Lu, G. H. Marchler, J. S. Song, N. Thanki, Z. Wang, R. A. Yamashita, D. Zhang, C. Zheng, L. Y. Geer, S. H. Bryant, *Nucleic Acids Res.* **2017**, *45*, D200–D203.

[53] R. Hurtado-Guerrero, A. W. Schüttelkopf, I. Mouyna, A. F. M. Ibrahim, S. Shepherd, T. Fontaine, J. P. Latgé, D. M. F. van Aalten, *J. Biol. Chem.* **2009**,

284, 8461–8469.

[54] R. Keegan, M. Wojdyr, G. Winter, A. Ashton, *Acta Crystallogr. Sect. A Found. Adv.* **2015**, *71*, s18.

[55] V. Frecer, R. Rizzo, S. Miertus, *Biomacromolecules* **2000**, *1*, 91–99.

[56] A. B. Boraston, D. Nurizzo, V. Notenboom, V. Ducros, D. R. Rose, D. G. Kilburn, G. J. Davies, *J. Mol. Biol.* **2002**, *319*, 1143–1156.

[57] J. Vasur, R. Kawai, K. H. M. Jonsson, G. Widmalm, Å. Engström, M. Frank, E. Andersson, H. Hansson, Z. Forsberg, K. Igarashi, M. Samejima, M. Sandgren, J. Ståhlberg, *J. Am. Chem. Soc.* **2010**, *132*, 1724–1730.

[58] K. Miyoshi, K. Uezu, K. Sakurai, S. Shinkai, *Chem. Biodivers.* **2004**, *1*, 916–924.

[59] K. Takahasi, M. Ochiai, M. Horiuchi, H. Kumeta, K. Ogura, M. Ashida, F. Inagaki, *Proc. Natl. Acad. Sci.* **2009**, *106*, 11679–84.

[60] Z. Qin, D. Yang, X. You, Y. Liu, S. Hu, Q. Yan, S. Yang, Z. Jiang, *Chem. Commun.* **2017**, *53*, 9368–9371.

[61] C. M. Bianchetti, T. E. Takasuka, S. Deutsch, H. S. Udell, E. J. Yik, L. F. Bergeman, B. G. Fox, *J. Biol. Chem.* **2015**, *290*, 11819–11832.

[62] B. Pluvinage, A. Fillo, P. Massel, A. B. Boraston, *Structure* **2017**, *25*, 1348–1359.

[63] M. Kanagawa, T. Satoh, A. Ikeda, Y. Adachi, N. Ohno, Y. Yamaguchi, *J. Biol. Chem.* **2011**, *286*, 29158–29165.

[64] L. A. Kelly, S. Mezulis, C. Yates, M. Wass, M. Sternberg, *Nat. Protoc.* **2015**, *10*, 845–858.

[65] Y. Wu, G. Mao, H. Fan, A. Song, Y.-H. P. Zhang, H. Chen, *Sci. Rep.* **2017**, *7*, 4849.

[66] T. Sawano, W. Saburi, K. Hamura, H. Matsui, H. Mori, *FEBS J.* **2013**, *280*, 4463–4473.

[67] B. Nidetzky, C. Eis, M. Albert, *Biocatal. Biotransform.* **2000**, *659*, 649–659.

[68] E. Rajashekhara, M. Kitaoka, Y.-K. Kim, K. Hayashi, *Biosci. Biotechnol. Biochem.* **2002**, *66*, 2578–2586.

[69] Y.-K. Kim, M. Kitaoka, M. Krishnareddy, Y. Mori, K. Hayashi, *J. Biochem.* **2002**, *132*, 197–203.

[70] T. Nihira, Y. Saito, M. Nishimoto, M. Kitaoka, K. Igarashi, K. Ohtsubo, H. Nakai, *FEBS Lett.* **2013**, *587*, 3556–3561.

[71] K. Hamura, W. Saburi, S. Abe, N. Morimoto, H. Taguchi, H. Mori, H. Matsui, *Biosci. Biotechnol. Biochem.* **2012**, *76*, 812–818.

[72] C. M. Bianchetti, N. L. Elsen, B. G. Fox, G. N. Phillips, *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **2011**, *67*, 1345–1349.

[73] A. Angelov, V. T. T. Pham, M. Übelacker, S. Brady, B. Leis, N. Pill, J. Brolle, M.

Mechelke, M. Moerch, B. Henrissat, W. Liebl, *Sci. Rep.* **2017**, *7*, 17306.

[74] D. Guillén, S. Sánchez, R. Rodríguez-Sanoja, *Appl. Microbiol. Biotechnol.* **2010**, *85*, 1241–1249.

[75] A. B. Boraston, D. N. Bolam, H. J. Gilbert, G. J. Davies, *Biochem. J.* **2004**, *382*, 769–781.

[76] D. Cockburn, M. M. Nielsen, C. Christiansen, J. M. Andersen, J. B. Rannes, A. Blennow, B. Svensson, *Int. J. Biol. Macromol.* **2015**, *75*, 338–345.

[77] X. Robert, R. Haser, H. Mori, B. Svensson, N. Aghajari, *J. Biol. Chem.* **2005**, *280*, 32968–32978.

[78] M. M. Nielsen, S. Bozonnet, E. N. Seo, J. A. Mótyán, J. M. Andersen, A. Dilokpimol, M. A. Hachem, G. Gyémánt, H. Næsted, L. Kandra, B. W. Sigurskjold, B. Svensson, *Biochemistry* **2009**, *48*, 7686–7697.

[79] M. Nakajima, H. Toyoizumi, K. Abe, H. Nakai, H. Taguchi, M. Kitaoka, *PLoS One* **2014**, *9*, e92353.

[80] M. Hoffmeister, M. Piotrowski, U. Nowitzki, W. Martin, *J. Biol. Chem.* **2005**, *280*, 4329–4338.

[81] K. Krnáčová, M. Vesteg, V. Hampl, Č. Vlček, A. Horváth, *J. Mol. Evol.* **2012**, *75*, 119–129.

[82] S. B. Conners, C. I. Montero, D. A. Comfort, K. R. Shockley, M. R. Johnson, S. R. Chhabra, R. M. Kelly, *J. Bacteriol.* **2005**, *187*, 7267–7282.

[83] Z. Wang, K. L. Robertson, C. Liu, J. L. Liu, B. J. Johnson, D. H. Leary, J. R. Compton, V. Vuddhakul, P. M. Legler, G. J. Vora, *FEMS Microbiol. Ecol.* **2015**, *91*, fiv087.

[84] N. Noinaj, M. Guillier, T. J. Barnard,, S. K. Buchanan, *Annu. Rev. Microbiol.* **2010**, *64*, 43–60.

[85] N. Terrapon, V. Lombard, É. Drula, P. Lapébie, S. Al-Masaudi, H. J. Gilbert, B. Henrissat, *Nucleic Acids Res.* **2018**, *46*, D677–D683.

[86] F. Unfried, S. Becker, C. S. Robb, J.-H. Hehemann, S. Markert, S. E. Heiden, T. Hinzke, D. Becher, G. Reintjes, K. Krüger, B. Avci, L. Kappelmann, R. L. Hahnke, T. Fischer, J. Harder, H. Teeling, B. Fuchs, T. Barbeyron, R. I. Amann, T. Schweder, *ISME J.* **2018**, in press.

[87] A. Sun, M. T. Hasan, G. Hobba, H. Nevalainen, J. Te'o, *J. Phycol.* **2018**, *54*, 529–538.

Chapter 4

A new glycoside hydrolase (GH) family, GHyyy, comprising β-(1→3)-glucan phosphorylases with different acceptor preferences to GH94 and GH149

4.1. Introduction

Glycoside phosphorylases (GPs) acting on β-D-glucopyranosyl-(1→3)-D-glucopyranose (laminaribiose) have previously been described in bacteria, *Paenibacillus* sp. YM-1 (*Ps*LBP)[1] and *Acholeplasma laidlawii* PG-8A,[2] whereas GPs acting on longer β-(1→3)-D-gluco-oligosaccharides have been described in eukaryotic microalga (*Euglena gracilis*)[3–5] and heterokonts *(Ochromonas danica*[6] and *Ochromonas malhemensis*[7]). The specificity of *Ps*LBP has been explored in detail in chapter 2, which revealed a somewhat relaxed donor specificity toward a noncognate donor, Man1P. Structural studies of *Ps*LBP revealed the architecture of the *Ps*LBP active site and -1 donor subsite, which is similar to other GH94 enzymes.[8–13] An additional β-hairpin gate near the active site was identified in the *Ps*LBP structure, which likely defines the preference of *Ps*LBP for disaccharide substrates.[1] In chapter 3, the nucleotide sequences of β-(1→3)-glucan phosphorylase from a microalga *Euglena gracilis* (EgP1) and its bacterial orthologue (Pro_7066) have been identified and characterised. EgP1, Pro_7066 and their orthologues constituted a new family of GH-like GPs (GH149). Structural studies of the GH94 *Ps*LBP and GH149 Pro_7066 have confirmed the conservation of key amino acid residues that are involved in the catalysis and validated the placement of both families within the same GH-clan (GH-Q). Pro_7066 contains additional domains and an oligosaccharide surface binding site, which are unique to the GH149 enzyme and likely contribute to defining the enzyme preference for longer oligosaccharide substrates.

The remaining β-(1→3)-D-glucan phosphorylase activity for which the nucleotide sequence(s) remained unknown are the heterokont β-(1→3)-D-glucan phosphorylases. These enzymes have not been classified into a CAZy family due to the lack of sequence information. The *Ochromonas* GPs are likely involved in the metabolism of chrysolaminarin (previously known as leucosin), a soluble β-(1→3)-glucan with a limited degree of β-(1→6)-branches[14–20] that accumulates within the vacuole of photosynthetic heterokonts.[21] A similar carbohydrate, called laminarin, has been identified in brown algae,[22] comprising (1→3)-linked β-D-glucose with different degrees of (1→6)-linked branching depending on the species of which laminarin was isolated.[22–27]

Availability of genomes and transcriptomes of heterokont model, *Phaeodactylum tricornutum* and *Thalassiosira pseudonana,* enables the study of

chrysolaminarin metabolism. Genome analysis of *P. tricornutum* has led to the identification of a UDP-glucose pyrophosphorylase, a β-(1→3)-D-glucan synthase and enzymes generating β-(1→6) branches,[28,29] supporting the presence of chrysolaminarin synthesis in this heterokont. The physiological role of UDP-glucose pyrophosphorylase has been confirmed, by a gene silencing experiment, to be an enzyme catalysing a rate-limiting step in chrysolaminarin biosynthesis.[30] Mutation of the UDP-glucose pyrophosphorylase gene led to enhanced lipid accumulation, indicating mobilisation of carbohydrate storage into lipid biosynthesis.[31] These findings were augmented by transcriptomic analysis, which revealed a diurnal expression pattern of the β-(1→3)-D-glucan synthase and the β-(1→6)-branching enzymes, with the highest gene expression level at the beginning of the light period and the lowest at the end, whilst the expression of endo-(1→3)-glucanase genes reached a maximum during the dark period. These findings have led to the proposition of chrysolaminarin metabolic pattern, where chrysolaminarin synthesis happens during the day and the equilibrium shifts toward degradation at night,[32] a situation that parallels leaf starch diurnal cycling in plants.[33–35]

Likewise, analysis of the *T. pseudonana* genome identified several sequence candidates involving in chrysolaminarin metabolism.[36] A *T. pseudonana* chrysolaminarin synthase was shown by a gene knockdown experiment to be important for chrysolaminarin accumulation.[37] Two of the *T. pseudonana* genes encoding β-(1→6)-branching enzymes were shown to partially complement mutation of the corresponding functional orthologues in yeast, strongly suggesting a role of the heterokont enzymes as β-(1→6)-branching enzymes.[38] Laminarin metabolic pathway was also proposed in brown algae through bioinformatic analysis of genomes and transcriptomes. For example, analyses of genome and transcriptome of a brown alga (*Ectocarpus siliculosus*) revealed candidate genes for UDP-glucose pyrophosphorylases, β-(1→3)-D-glucan synthases (GT48) as well as candidate genes for the degradation of laminarin, including endo-glucanases (GH16, 17, 81), exo-glucanases (GH5) and β-glucosidases (GH1 and 3).[39]

None of the aforementioned genome and transcriptome analyses described the discovery of any GPs related to laminarin/chrysolaminarin metabolism, possibly due to the lack of established nucleotide sequence encoding the phosphorylase activity. Therefore, sequence identification for the heterokont β-(1→3)-D-glucan

phosphorylase(s) is a key step towards investigation of the enzyme activity and its physiological role in laminarin/chrysolaminarin metabolism.

In continuation of the previous work on the identification of new β-(1→3)-D-glucan phosphorylases and the understanding of their mechanism and potential application in carbohydrate synthesis, a heterokont phosphorylase sequence from *Ochromonas* spp. (OcP1) was uncovered together with its orthologues, which subsequently led to an establishment of another new GH family designated GHyyy. In contrast to GH149, which contains mainly sequences from Gram-negative bacteria and eukaryotic organisms in class Euglenophyceae, the majority of GHyyy sequences were identified in Gram-positive bacteria and eukaryotic heterokonts, highlighting the possible difference in evolution of these two enzyme families. Unfortunately, expression of the recombinant OcP1 protein in *E. coli* was not successful. In order to establish the possible activity of GHyyy members, a bacterial *GHyyy* gene sequence (*Pap*P) from a Gram-positive bacterium *Paenibacillus polymyxa* ATCC 842 was cloned and expressed in *E. coli* for recombinant protein production. Biochemical characterisation of the recombinant *Pap*P protein showed that the enzyme operated on β-(1→3)-gluco-oligosaccharide acceptors with degree of polymerisation (DP) ≥ 2 and could not use Glc as its acceptor substrate. The acceptor length specificity of *Pap*P was clearly distinct from that of the characterised GH149 enzymes, which could act on Glc. Bacterial *GHyyy* genes co-localised with genes encoding glucosidases (GH1, GH3 or GH30) and ATP-binding cassette (ABC) transporters, highlighting the probable involvement of such gene clusters in carbohydrate degradation in Gram-positive bacteria. More importantly, some *GHyyy* genes co-localised with *GH94* genes in a tandem manner, providing the first example of two GPs from different CAZy families possibly working together in sequential degradation of oligosaccharides.

4.2. Results

4.2.1. Analysis of *Ochromonas* transcriptome revealed candidate β-(1→3)-D-glucan phosphorylases

The investigation began with the premise that the *Ochromonas* phosphorylases may share significant sequence identity to either GH94 or GH149, considering their similar specificity for β-(1→3)-glycosidic linkage. To investigate this hypothesis, the translated transcriptome of *Ochromonas* sp. BG-1 from the MMET database was interrogated by BLASTP with GH94 sequences (21 sequences) and with EgP1 queries. Three hits were recovered (CAMPEP_0173148212, CAMPEP_0173133844 and CAMPEP_0173155066), with 21% identity to a GH94 laminaribiose phosphorylase (LBP) sequence from *Halorhadus tiamatea* and 24% to EgP1 sequence (Table 4.1). BLASTP analysis of CAMPEP_0173148212, CAMPEP_0173133844 and CAMPEP_0173155066 showed that they shared 99% sequence identity. Due to extremely high sequence identity of the three *Ochromonas* candidates, only CAMPEP_0173155066 was taken as the representative of the three sequences and designated as OcP1.

Table 4.1. BLASTP analyses of *Ochromonas* sp. transcriptomic sequences using 21 sequences of GH94 family and EgP1 of GH149 as queries.

Query	Subject	Hits ID	E-value	%identity
WP_020936056.1|LBP family GH94 [*Halorhabdus tiamatea*]	*Ochromonas sp.* strain BG-1 transcriptome (MMETSP1105)	CAMPEP_0173148212	3e-06	21
		CAMPEP_0173133844	8e-06	21
		CAMPEP_0173155066 (OcP1)	8e-06	21
EgP1|GH149 [*Euglena gracilis*]	*Ochromonas sp.* strain BG-1 transcriptome (MMETSP1105)	CAMPEP_0173148212	3e-08	24
		CAMPEP_0173133844	3e-08	24
		CAMPEP_0173155066 (OcP1)	3e-08	24

PSI-BLAST searches against the NCBI and MMET database were performed using the translated amino acid sequence of OcP1 as a query, which uncovered several hundred orthologous sequences from heterokonts and bacteria (Appendix 7). 67 orthologous sequences (32-58% sequence identity to OcP1, Appendix 8) retrieved from the MMET database originated from eukaryotes in phyla Bacillariophyta, Ochrophyta, Haptophyta and Miozoa, all of which belong to infrakingdom Heterokonta (collectively termed heterokonts) (Figure 4.1, orange). One of the heterokont sequence is a

candidate from *Ochromonas danica* recovered from the NCBI database (GenBank ID: BAU78234.1; %identity = 58, E-value = 0). The bacterial orthologues (332 sequences) shared 31-41% sequence identity to OcP1 (Appendix 8). The majority of these bacterial hits belong to Gram-positive bacteria in phylum Firmicutes (Figure 4.1, cyan, 223 sequences), 85 of which were from *Paenibacillus* spp.

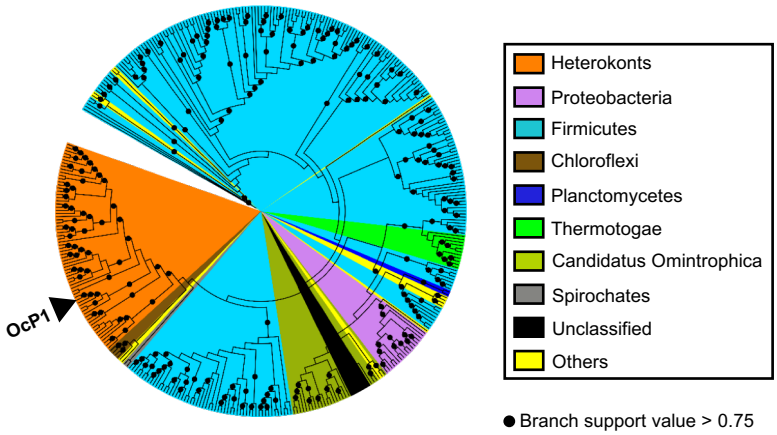


Figure 4.1. Phylogenetic analysis of OcP1 and its orthologues. The position of OcP1 sequence is indicated by an arrow head.

In order to establish the relationship between OcP1 orthologues and GH families containing β-(1→3)-D-glucan phosphorylases (GH94 and GH149), phylogenetic analysis of GH94, GH149 and the OcP1 orthologous sequences was performed, which indicated that the OcP1 orthologues form a distinct clade away from GH94 and GH149 (Figure 4.2A, blue). Thus, OcP1 and its orthologues constitute a new family, designated GHyyy. However, multiple sequence alignment of amino acid sequences from GH94, GH149 and GHyyy revealed conservation of aspartate catalytic residue, residues located in the -1 subsite and a histidine involved in phosphate recognition that have been previously identified in GH94 and GH149 (Figure 4.2B), despite low overall sequence homology (~20%). The conservation of the active site and -1 subsite in all three families strongly support a placement of GHyyy family into the same GH clan as GH149 and GH94 (clan GH-Q).

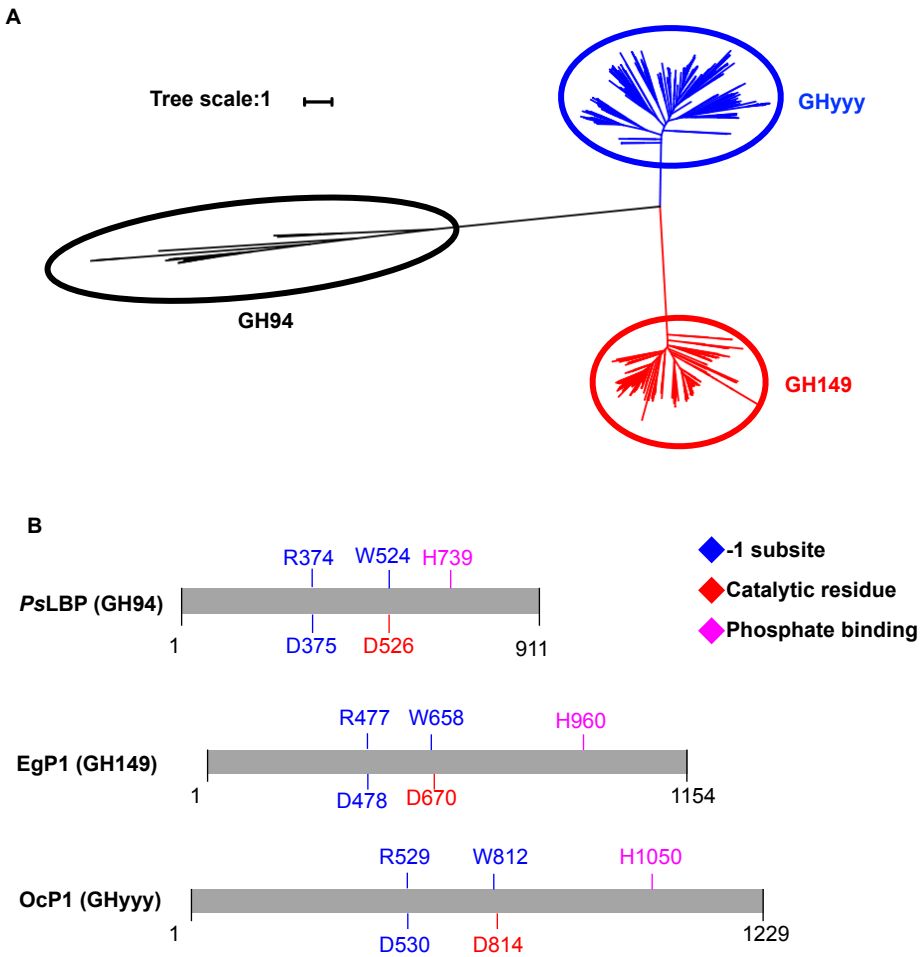


Figure 4.2. Relationship between GH94, GH149 and GHyyy**.** (**A**) An unrooted tree representing the phylogenetic relationship between GH94 queries (black branches), GH149 (red branches) and GHyyy (blue branches). (**B**) Summary of sequence alignment of GH94 (*Ps*LBP), GH149 (EgP1) and GHyyy (OcP1).

4.2.2. Cloning, heterologous expression and biochemical characterisation of a GHyyy from *Paenibacillus polymyxa* ATCC 842

Although OcP1 was discovered as the first sequence of GHyyy family in the initial bioinformatic study, the recombinant OcP1 protein could not be produced in *E. coli*, likely due to its high GC content of its encoding nucleotide sequence. Another GHyyy candidate from a heterokont *Prymnesium parvum* (*Prym*P), which shared 41% amino acid sequence identity to OcP1, was also cloned into an expression vector and transformed into *E. coli* for recombinant protein expression. However, *E. coli* cell culture containing the *PrymP* recombinant plasmid did not grow to high density in liquid culture after the protein expression was induced, suggesting that the recombinant *Prym*P protein may be toxic to the *E. coli* cells. Due to the unsuccessful expression of the heterokont sequence candidates in *E. coli*, bacterial GHyyy sequences were chosen as an alternative target to provide evidence for the enzyme activity of GHyyy family. GHyyy candidates from *Paenibacillus* spp. were selected as a representative of this family for several reasons; 1) genus Paenibacillus consists of more than 150 species, many of which are plant-growth promoting rhizobacteria which are important in agriculture and horticulture.[40] Therefore, the study of *Paenibacillus* spp. carbohydrate metabolisms may provide useful evidence for utilisation of plant carbohydrates by the bacteria which is crucial for establishment of plant-microbial interaction. Also, from a biomass utilisation perspective, study of carbohydrate-active enzymes (CAZymes) from *Paenibacillus* may uncover a repertoire of novel plant cell wall degradative enzymes; 2) analyses of transcriptomes and genomes of *Paenibacillus* spp. revealed several gene candidates for CAZymes that operate on β-glucans;[41] 3) several reports detailed *in vitro* characterisation of *Paenibacillus* CAZymes acting on β-(1→3)-glucan and β-(1→3),(1→4)-glucan which suggest that β-glucans could be significant carbon sources for *Paenibacillus*. For example, identification of a β-(1→3)-endoglucanase (LamA) in *Paenibacillus* sp. CCRC 17245 which digested β-(1→3),(1→4)-glucan,[42] laminarin-degrading β-glucanase ($BglA_1$) and a β-(1→3),(1→4)-endoglucanase ($BglA_{2)}$ from *Paenibacillus* sp. JDR-2,[43] and β-(1→3),(1→4)-endoglucanase from *Paenibacillus* sp. F-40;[44] 4) the majority of bacterial GHyyy sequences were identified in *Paenibacillus* spp., implying the importance of these enzymes from this genus; 5) an LBP from *Paenibacillus* sp. YM-1 (*Ps*LBP) in GH94 family has been described,[1] therefore characterisation of any GHyyy from *Paenibacillus* spp. would enable comparison between the GH94 and GHyyy β-(1→3)-D-glucan phosphorylases that came from the same bacterial genus.

A GHyyy sequence from *Paenibacillus polymyxa* ATCC 842 (GenBank accession: WP_019688419.1, 37% identity to OcP1), designated *Pap*P, was chosen as a target for recombinant protein expression and *in vitro* biochemical characterisation. *P. polymyxa* ATCC 842 is available from commercial culture collection (ATCC®) and can be cultured easily in the laboratory, enabling genomic DNA isolation and subsequent cloning of the *PapP* gene from the bacterial genome. Primers were designed for amplification of the *PapP* gene (gene locus: PPT_RS0121460) from *P. polymyxa* ATCC 842 (genome accession: NZ_GL905390.1). The amplified gene was then cloned into a pOPINF vector and transformed into *E. coli* (Rosetta PLysS) for recombinant protein expression with an N-terminal His_6-tag. The recombinant *Pap*P was purified by IMAC (Figure 4.3A and B) and subsequently by gel filtration (Figure 4.3C) to obtain the purified protein with an approximate molecular mass of 120 kDa based on SDS-PAGE analysis (Figure 4.3D). Gel filtration analysis of *Pap*P against protein standards revealed that *Pap*P formed a dimer in solution under non-denaturing conditions, with an estimated molecular mass of 241 kDa. The dimeric forms of *Pap*P is similar to that described of *Ps*LBP, EgP1 and Pro_7066 (Table 4.2).

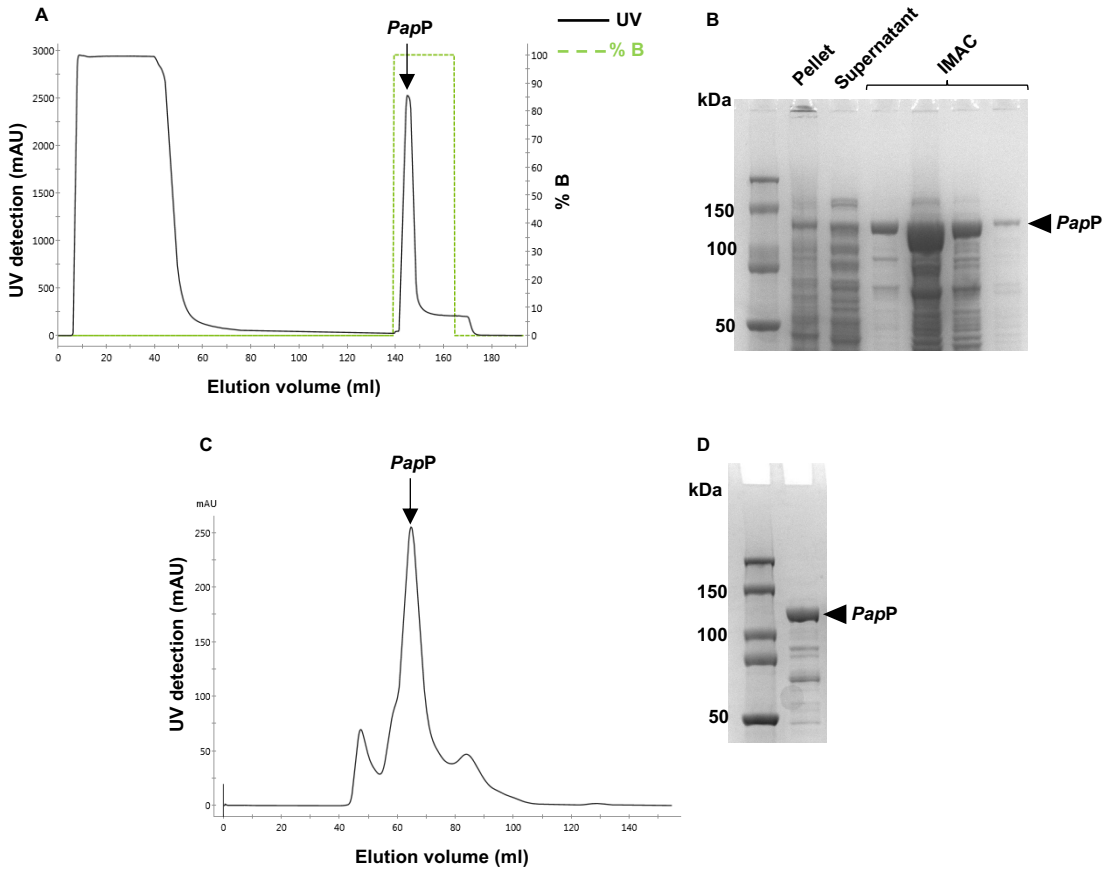


Figure 4.3. Production of recombinant *Pap*P protein in *E. coli* and purification of the soluble protein. (**A**) IMAC purification with UV detection trace (black line). *Pap*P was eluted in 100% of buffer B (10 mM HEPES

pH 7.5, 250 mM NaCl, 500 mM imidazole). (**B**) SDS-PAGE of *Pap*P after IMAC purification. (**C**) Gel filtration of *Pap*P. The protein was eluted at 65-ml elution volume in 20 mM HEPES pH 7.5, 150 mM NaCl. (**D**) SDS-PAGE analysis of the purified recombinant *Pap*P after IMAC and gel filtration.

Table 4.2. Summary of the estimated molecular masses of *Ps*LBP, Pro_7066, EgP1 and *Pap*P based on their elution volumes from gel filtration column (Superdex S200 16/600, flow rate: 1 ml/min). See Chapter 6, section 6.10 for detailed method of molecular mass estimation.

Protein	Elution volume (ml)	Estimated molecular mass (kDa)	Calculated mass of monomer (kDa)
***Ps*LBP**	65	240	102
Pro_7066	62	323	130
EgP1	63	286	127
PapP	65	241	124

To investigate the function of the recombinant *Pap*P, the enzyme was assayed *in vitro* in the glycan synthetic reaction using Glc and Glc1P as an acceptor and a donor, respectively. The reaction mixture was analysed by TLC, which showed that there was no detectable activity on these substrates (Figure 4.4A). Assays in the phosphorolysis direction were performed in the presence of Glc-Glc disaccharide with varying glycosidic linkages as substrates. The reactions were analysed by TLC, which did not show any detectable phosphorolysis of the tested disaccharides (Figure 4.4B). These initial experiments have led to the hypothesis that *Pap*P may not recognise disaccharides as its substrate and might prefer substrates with longer degree of polymerisation for the phosphorolysis reaction.

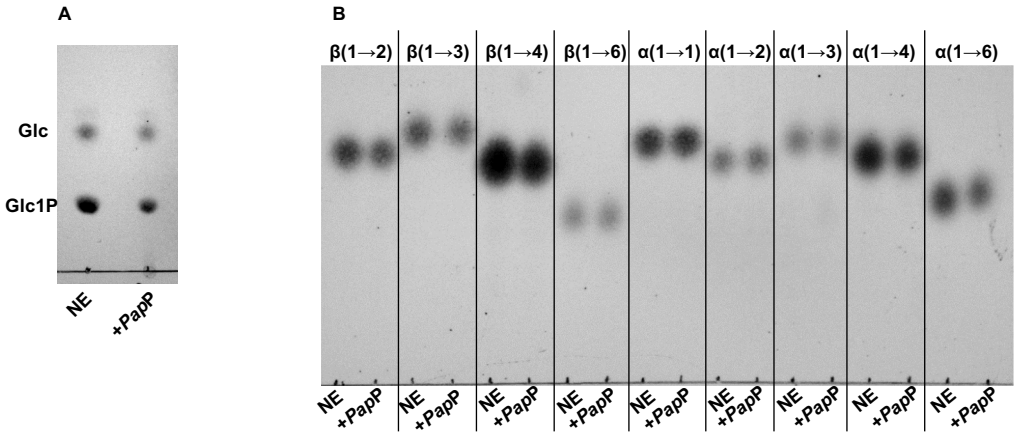


Figure 4.4. TLC analysis of *Pap*P activity on Glc and disaccharides as acceptors. (**A**) Glycan synthetic reaction in the presence of Glc (10 mM) and Glc1P (20 mM) as substrates. (**B**) Phosphorolysis reactions using Glc-Glc disaccharides with varying glycosidic linkages (10 mM) (as indicated at the top) as substrate in the presence of inorganic phosphate (10 mM). NE = no enzyme control.

To determine whether *Pap*P can operate on longer oligosaccharide substrates, *Pap*P was assayed in the phosphorolysis reaction in the presence of a trisaccharide

[either β-(1→3)- (laminaritriose, G3) or β-(1→4)-oligosaccharides (cellotriose)] and inorganic phosphate as substrates, which showed that *Pap*P could phosphorylyse laminaritriose but not cellotriose, suggesting that *Pap*P prefers β-(1→3)-glycosidic linkages (Figure 4.5A). To further confirm this finding, an assay in the phosphorolysis direction was conducted in the presence of laminarihexaose (G6) and inorganic phosphate, which showed that G6 was broken down by *Pap*P into shorter oligosaccharides (G2-G5), but could not phosphorylyse disaccharide (G2) to Glc (Figure 4.5B). Since the phosphorolysis is reversible, peaks corresponding to oligosaccharide products longer than the G6 substrate were also detected (G7-G13, Figure 4.5B and C). Altogether, the results strongly indicate that *Pap*P is a β-(1→3)-D-glucan phosphorylase which can phosphorylyse linear oligosaccharide chain with DP ≥ 3, distinguishing this enzyme as distant from GH94 and GH149 relatives.

Previously, the GH149 family was discovered and two GH149 proteins (EgP1 and Pro_7066) were confirmed as β-(1→3)-D-glucan phosphorylases that can carry out the phosphorolysis of β-(1→3)-oligosaccharides to Glc, representing a different substrate preference from the GHyyy *Pap*P. To emphasise the difference in substrate chain length preferences between the GH149 and GHyyy enzymes, assays were carried out in the presence of G3 and inorganic phosphate as substrates using either *Pap*P or Pro_7066 as an enzyme catalyst. The reactions were analysed by HPAEC-PAD, which showed that Pro_7066 was capable of degrading G3 to Glc, whereas *Pap*P degraded G3 to G2 only (Figure 4.5D).

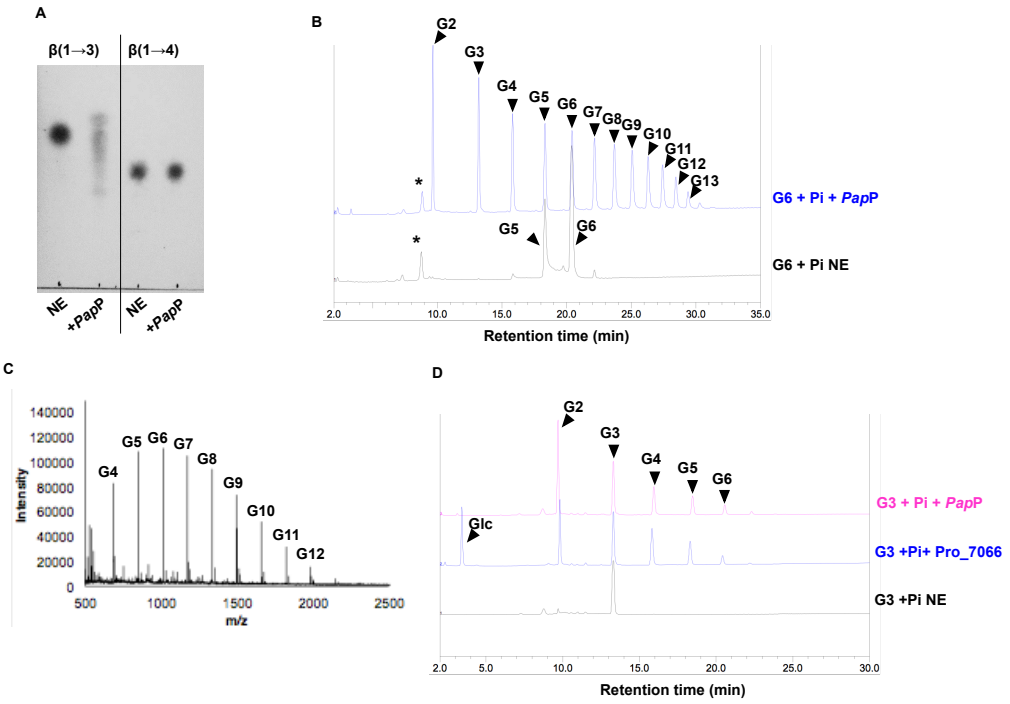


Figure 4.5. Characterisation of *Pap*P and comparison of *Pap*P substrate chain length preference to that of GH149 enzyme. (**A**) TLC analysis of the phosphorolysis reaction carried out by *Pap*P in the presence of 10 mM laminaritriose (β-(1→3)-linkage) or cellotriose (β-(1→4)-linkage). (**B**) HPAEC-PAD analysis of the phosphorolysis reaction of 10 mM laminarihexaose (G6) carried out by *Pap*P in the presence of 10 mM inorganic phosphate. Laminaripentaose (G5) was detected as a contaminant in the starting material. (**C**) MALDI-ToF analysis of G6+Pi+*Pap*P reaction in **B**. G2-12 indicates the DP of the β-(1→3)-gluco-oligosaccharide products. NE = no enzyme control. All assays were performed in 100 mM HEPES pH 7.0 at 25℃ for 1 hr. Asterisks = unknown peaks. (**D**) Comparison between the phosphorolysis carried out by either *Pap*P (magenta) or Pro_7066 from GH149 family (blue) in the presence of 10 mM laminaritriose (G3) and 10 mM inorganic phosphate as substrates.

To further probe the acceptor chain length specificity of *Pap*P, a kinetic study was performed in the synthetic direction in the presence of laminaribiose (G2), laminaritriose (G3), laminaritetraose (G4), laminaripentaose (G5) or laminarihexaose (G6) as acceptors and Glc1P as a donor. *Pap*P showed similar catalytic efficiency (k_{cat}/K_m^{app}) towards all acceptors (10 mM) when 10 mM Glc1P was used as a donor (Table 4.3).

Table 4.3. Kinetic parameters of the glycan synthesis reaction catalysed by *Pap*P. K_M^{app} is an apparent K_M when 10 mM Glc1P was used as a donor.

Acceptors	k_{cat} (s^{-1})	K_M^{app} (mM)	k_{cat}/K_M^{app} ($s^{-1}mM^{-1}$)
G2	31.6 ± 1.0	1.58 ± 0.17	20.0
G3	33.3± 1.6	1.05 ± 0.19	31.6
G4	33.3 ± 1.7	1.82 ± 0.29	18.3
G5	27.4 ± 0.7	1.61 ± 0.13	17.0
G6	30.0 ± 0.6	2.31 ± 0.56	13.0

*Pap*P required substrate with a minimum DP 3 for phosphorolysis, implying that the enzyme must be able to use a disaccharide (DP = 2) for the glycan synthetic reaction. To test this hypothesis, an assay in the glycan synthetic direction was performed in the presence of Glc-Glc disaccharides with varying glycosidic linkages as acceptors and Glc1P as a donor. Interestingly, not only *Pap*P could use G2, but it could also use other β-linked disaccharides as acceptors, regardless of the linkage regiochemistry [(1→2), (1→3), (1→4) or (1→6)], whilst there was no detectable activity on α-linked glucose-glucose disaccharides (Figure 4.6 and 4.7).

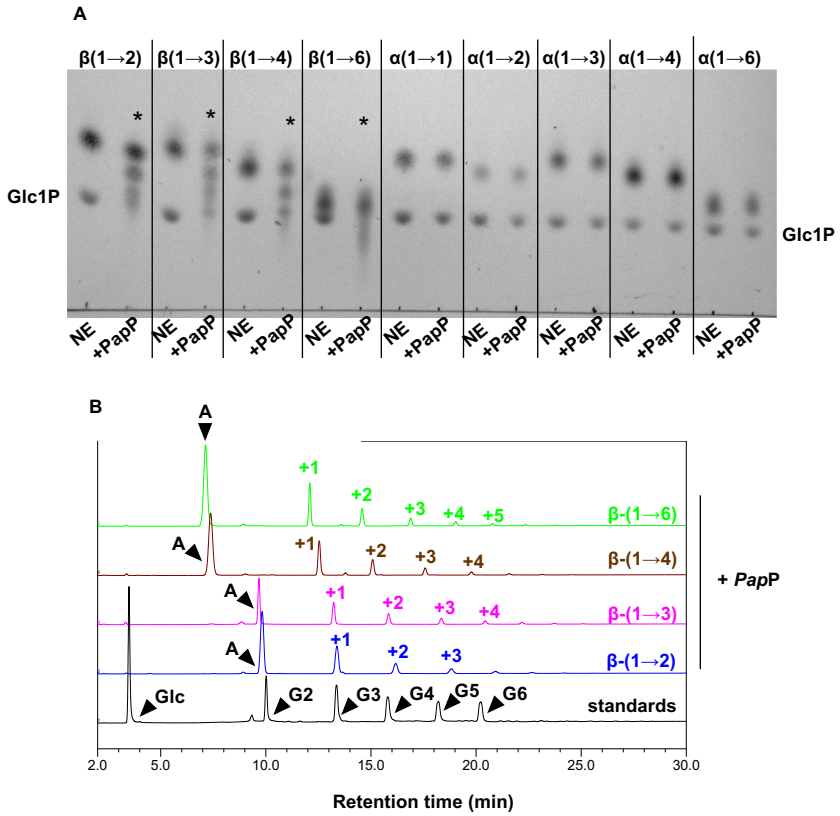


Figure 4.6. Glycan synthetic reactions carried out by *Pap*P in the presence of β-linked disaccharide acceptors and Glc1P**.** (**A**) TLC analysis of the glycan synthesis in the presence of glucose-glucose disaccharides (linkages as indicated above the figure). Asterisk indicates the detectable extension of the acceptors carried out by *Pap*P. (**B**) HPAEC-PAD analysis of the glycan synthetic reaction in the presence of β-linked glucose-glucose disaccharide. A = peaks that represents the acceptors in each reaction. The linkages of the acceptors are indicated on the right. +1-+5 = the number of Glc that have been transferred from Glc1P donor onto the acceptors, thus creating oligomers that are represented by the peaks. For example, peak labelled with +1 represents an acceptor that has been extended by 1 Glc unit, therefore has the total DP of 3. G2 = laminaribiose, G3 = laminaritriose, G4 = laminaritetraose, G5 = laminaripentaose, G6 = laminarihexaose.

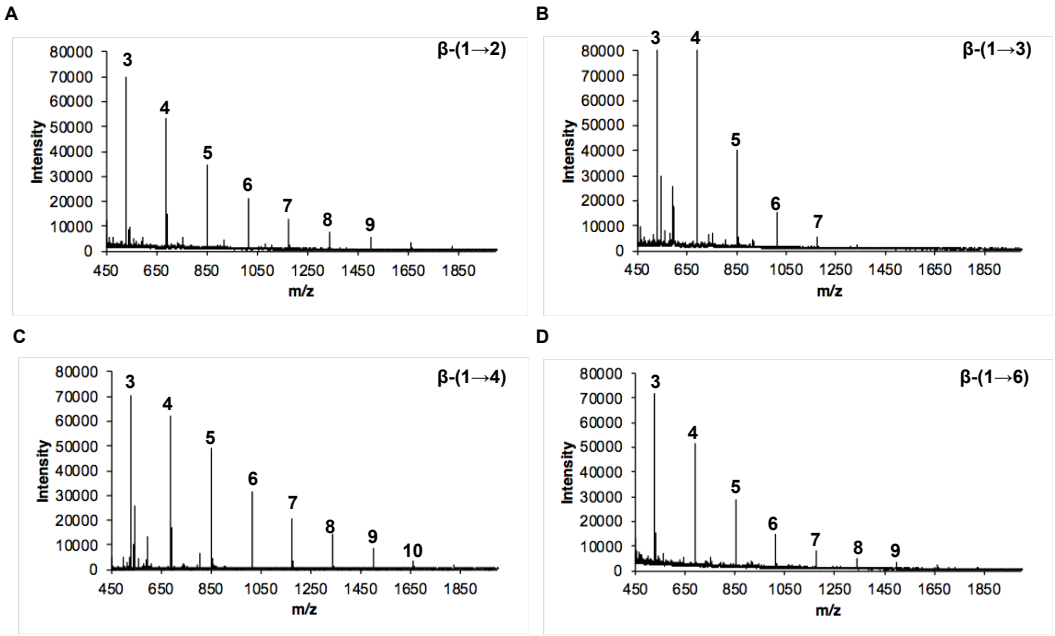


Figure 4.7. MALDI-ToF analyses of the glycan synthetic reactions catalysed by *Pap*P in the presence of β-linked disaccharide acceptors and Glc1P**.** The glycosidic linkages of the acceptors are indicated in the figures. The DP of glucan products corresponding to the peaks are indicated by numbers (4-11).

The relaxed acceptor regiochemistry observed in *Pap*P led to the hypothesis that this property could be conserved in its GH149 relative. To investigate this hypothesis, Pro_7066 (GH149) was assayed in the glycan synthetic reactions in the presence of glucose-glucose disaccharides with varying glycosidic linkages, which also showed the selectivity towards the β-linked acceptors and tolerance towards variation of the linkage regiochemistry in the same manner as for *Pap*P (Figure 4.8 and 4.9).

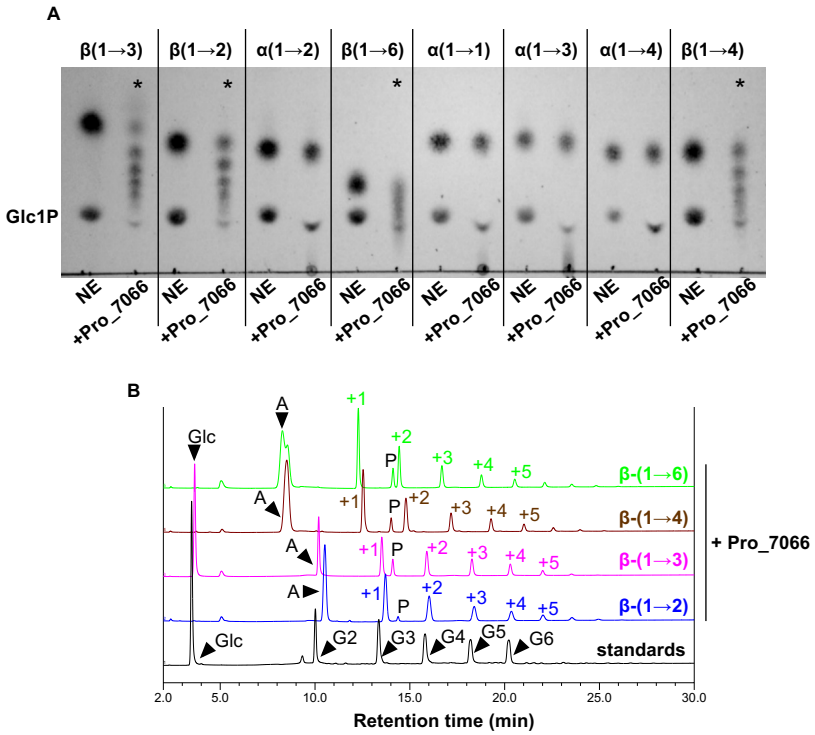


Figure 4.8. Glycan synthetic reactions carried out by Pro_7066 from GH149 family in the presence of β-linked disaccharide acceptors (10 mM) and Glc1P (10mM). (**A**) TLC analysis of the glycan synthesis in the presence of glucose-glucose disaccharides (linkages as indicated above the figure). Asterisks indicate the detectable extension of the acceptors carried out by Pro_7066. (**B**) HPAEC-PAD analysis of the glycan synthetic reaction in the presence of β-linked Glc-Glc disaccharide. A = peaks that represents the acceptors in each reaction. The linkages of the acceptors are indicated on the right. +1-+5 = the number of Glc that have been transferred from Glc1P donor onto the acceptors, thus creating oligomers that are represented by the peaks. For example, peak labelled with +1 represent an acceptor that has been extended by 1 Glc, therefore has the total DP of 3. P = trace amount of Glc1P that was not completely removed before HPAEC-PAD analysis. G2 = laminaribiose, G3 = laminaritriose, G4 = laminaritetraose, G5 = laminaripentaose, G6 = laminarihexaose.

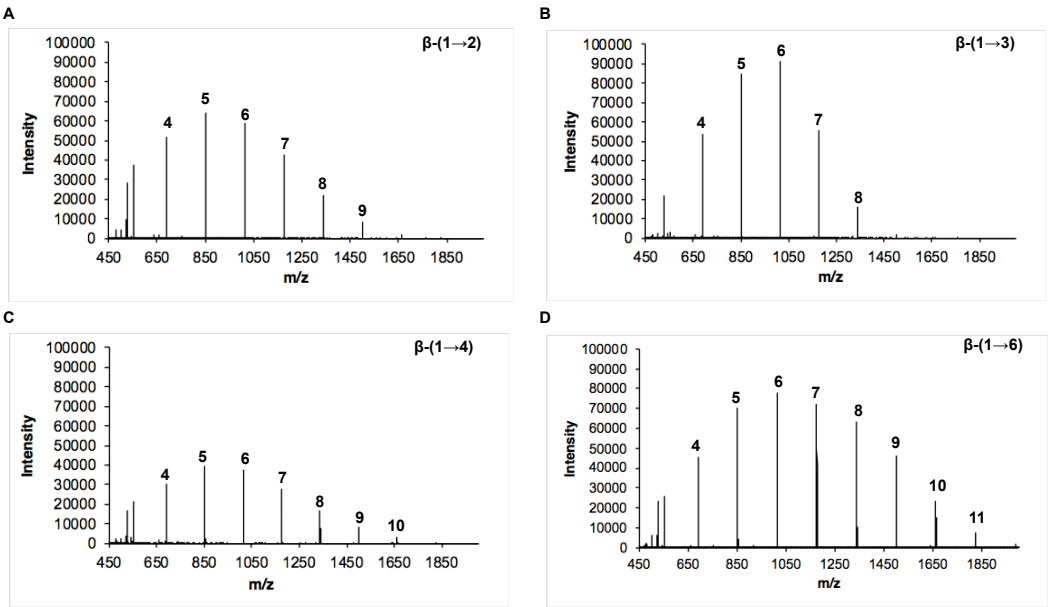


Figure 4.9. MALDI-ToF analysis of glycan synthesis using Pro_7066 as a catalyst and with β-linked disaccharides and Glc1P as substrates. Glycosidic linkages in the disaccharide acceptors are indicated. The DP of glucan products corresponding to the peaks are indicated by numbers (4-11).

4.2.3. Gene clusters containing *GHyyy* genes in bacterial genomes suggesting a role of bacterial GHyyy in oligosaccharide degradation

The majority of identified GPs have been predicted to be involved in carbohydrate degradation, largely due to their co-localisation with sugar transporters and other GHs.[1,45–47] Based on this assumption, the lack of phosphorolytic activity on disaccharides in GHyyy enzyme would have an important physiological implication, because the disaccharide product created by GHyyy cannot be used in downstream carbon metabolic processes. Therefore, the GHyyy-possessing organisms must have additional GHs or GPs that can degrade the disaccharide product of GHyyy to Glc.

In order to predict a possible role of GHyyy in bacteria possessing *GHyyy* genes, locations of *GHyyy* genes within the corresponding bacterial genomes were investigated. *PapP* gene co-localises with genes encoding β-glucosidase (GH1) and ABC transporter (Figure 4.10A). The subcellular localisation of *Pap*P and GH1 proteins was predicted to be in cytoplasm, suggesting a role of *Pap*P enzyme for oligosaccharide degradation to disaccharide, which is subsequently broken down to Glc by GH1 (Figure 4.10B). The simple genetic organisation would suggest that the *Pap*P gene cluster might be involved in degradation of simple linear oligosaccharide substrate.

A gene encoding GHyyy from *Bacillus weihaiensis* Alg07 (WP_072580822.1) was mapped to a previously identified laminarin utilisation locus (Figure 4.10A), which contains GH30, a β-glucosidase (GH1) and a laminarinase (GH16), all of which are highly expressed upon exposure to laminarin.[48] The location of *GHyyy* in the laminarin utilisation locus raises the possibility that the GHyyy enzyme may be involved in laminarin degradation. The *B. weihaiensis* GH30 protein was predicted to be secreted into extracellular space. GH30 family contains enzymes with β-(1→6)-endoglucanase activity, which is required for degradation of β-(1→6)-linkages in β-glucans. Since laminarin contains β-(1→6)-linked branches, it is likely that the *B. weihaiensis* GH30 protein is involved in the debranching of β-(1→6)-linked glucan, to enable the import of linear β-(1→3)-linked oligosaccharides into the bacterial cell (Figure 4.10C). In contrast, the *B. weihaiensis* GH1 and GHyyy proteins were predicted to localise in cytoplasm, thus they are likely to be involved in sequential depolymerisation of linear β-(1→3)-linked oligosaccharides into Glc (Figure 4.10C).

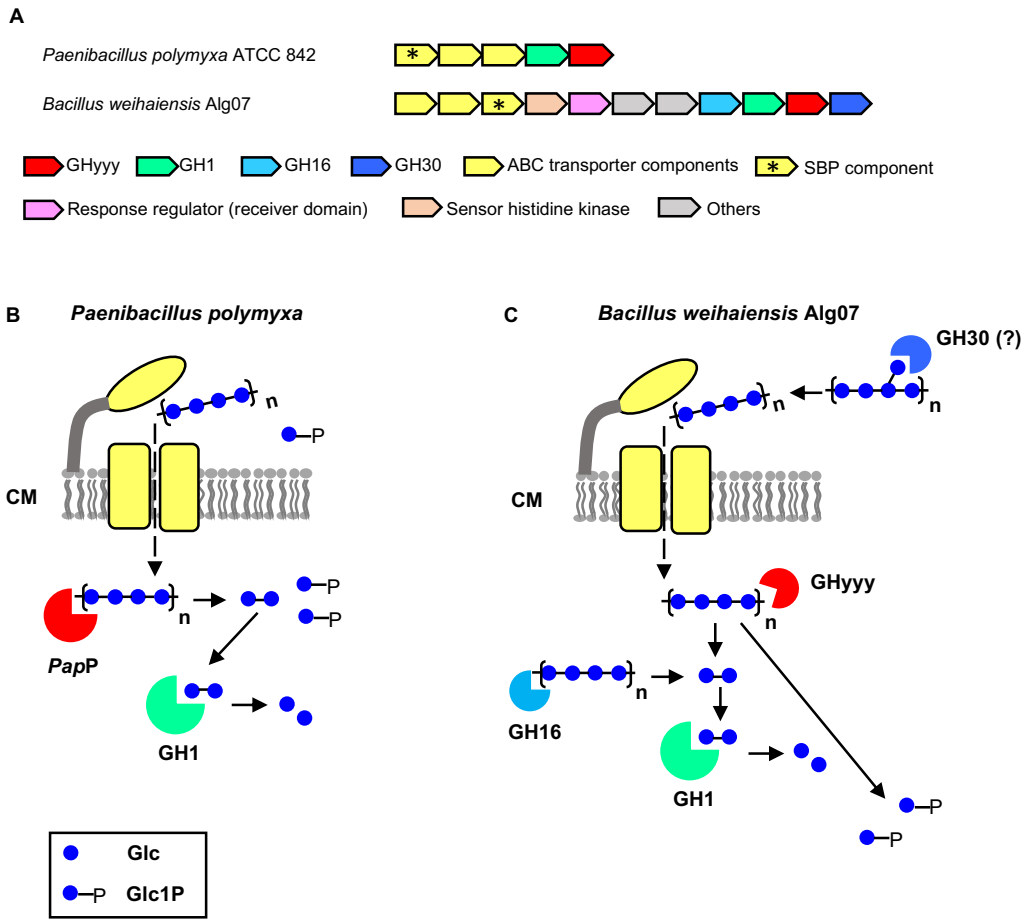


Figure 4.10. Gene clusters containing *GHyyy* genes from *Paenibacillus polymyxa* and *Bacillus weihanensis*. (**A**) Genetic loci containing *GHyyy* genes; *Bacillus weihanensis* Alg07[48] and *Paenibacillus polymyxa* ATCC 842. (**B and C**) Predicted roles of the enzymes encoded by the gene clusters in Figure **A.** ABC transporters are coloured in yellow and located in CM. The enzymes are coloured according to the colours of the corresponding genes presented in the gene clusters. The subcellular localisation prediction of GHs and transporters was performed by PSORTb version 3.0.2 (http://psort.org/psortb/index.html) using default settings for Gram-positive bacteria. CM = cell membrane. Substrate-binding protein (SBP) component of ABC transport system is indicated by asterisks. Glc1P and disaccharide are released from the oligosaccharide degradation out by GHyyy. Glc1P can be converted to Glc6P by phosphoglucomutase and used in glycolysis.

Analysis of other bacterial genomes containing *GHyyy* genes revealed two distinctive patterns of GHyyy-containing gene clusters in these bacteria;

1) Co-localisation of GHyyy with other GHs, in particular with genes encoding glucosidases from GH1, GH3 and GH30 families (Figure 4.11). Again, this organisation likely suggests a role for these GHyyy enzymes in the depolymerisation of oligosaccharide substrate, similar to that described for the *PapP* gene cluster. For example, a gene cluster from *Acholeplasma laidlawii* PG-8A contains genes encoding putative GHyyy, two permease components of ABC transporter, a laminarinase (GH16) and a glucosidase (GH3) (Figure 4.11, red-dotted box). The substrate-binding protein (SBP) component of the ABC transporter, which is important for substrate specificity

and high affinity of substrate uptake,[49] is missing from the *A. laidlawii* GHyyy gene cluster. Interestingly, a gene cluster containing a gene encoding the previously characterised *A. laidlawii* GH94 LBP,[2] was found located next to the putative GHyyy gene cluster (Figure 4.11, black-dotted box). The *A. laidlawii* GH94 gene cluster also contains genes encoding ABC transporter components, including SBP. The close proximity of the *A. laidlawii* GHyyy and GH94 gene clusters and the lack of SBP gene in the former imply that the two clusters may share the SBP component for substrate uptake and that the two clusters encode enzymes that work cooperatively in degrading the same β-glucan substrate. The same organisation of GH94 and GHyyy gene cluster was also observed in *A. oculi* 19L (Figure 4.11).

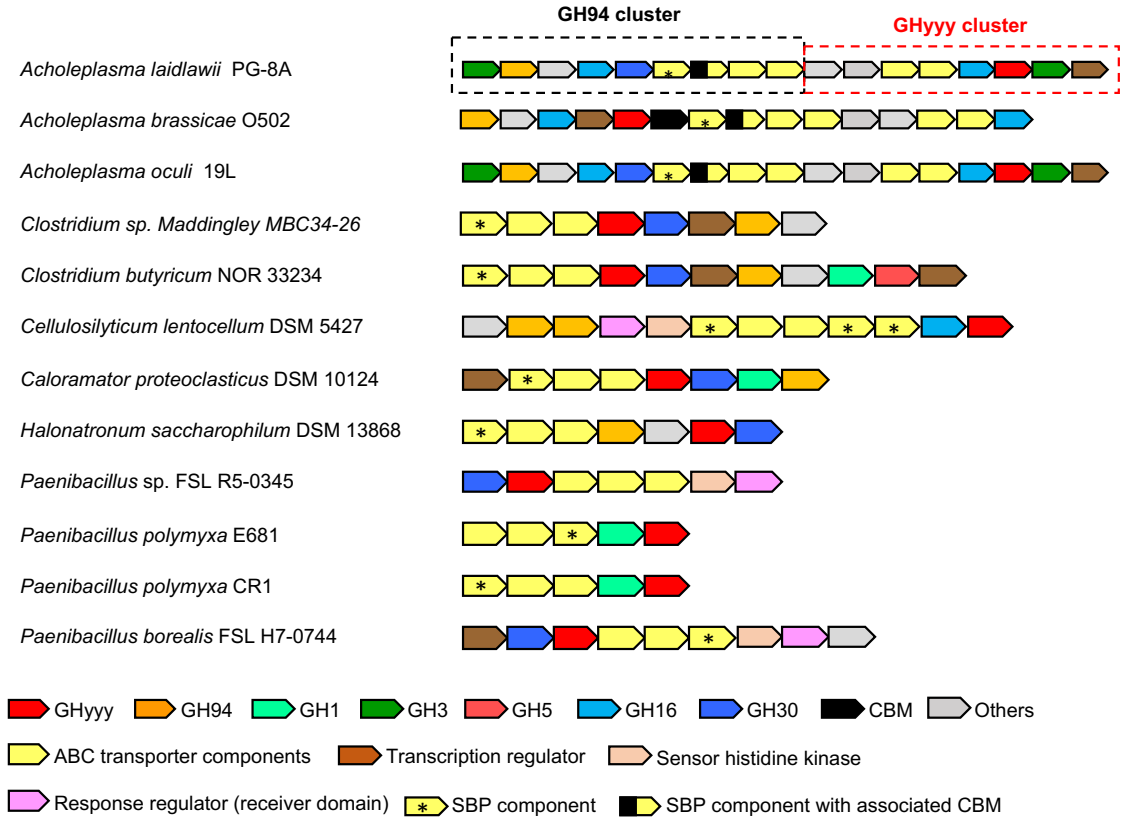


Figure 4.11. Genetic clusters containing genes encoding *GHyyy* and other glucosidases**.** The gene cluster containing the previously characterised GH94 LBP from *A. laidlawii* PG-8A is indicated by black-dotted box,[2] whilst the adjacent GHyyy gene cluster is indicated by red-dotted box. Substrate-binding protein (SBP) components of ABC transport system is indicated by asterisks.

2) Co-localisation of putative *GHyyy* with *GH94* genes in a tandem manner (Figure 4.12A) illustrates the first example of two different GP families co-existing in this pattern. Based on the knowledge of *Pap*P activity that is only capable of digesting oligosaccharides into disaccharides, and the prevalence of disaccharide phosphorylases in GH94 family, it is likely that the GHyyy and GH94 enzymes encoded

by the tandem genes may act sequentially on the same oligosaccharide substrate (Figure 4.12B). A similar tandem arrangement of GP-encoding genes has also been observed in a gene cluster containing two GH130 β-(1→2)-mannoside phosphorylases from *Thermoanaerobacter* sp. X-513, which was proposed to be involved in degradation of β-(1→2)-mannoside to supply Man1P for GDP-D-mannose biosynthesis.[50] Based on this notion, it is possible that the purpose of gene clusters containing the *GH94-GHyyy* tandem is to degrade β-glucan from external sources to supply Glc1P for other metabolic pathways such as glycogen synthesis and glycolysis. The occurrence of *GH94-GH149* tandem genes points towards genetic co-inheritance of the two families, the enzyme activities of which have evolved to operate on glucan substrates with different lengths (disaccharide for GH94 versus oligosaccharide for GHyyy), but in a functionally coordinated manner to ensure complete oligosaccharide substrate degradation.

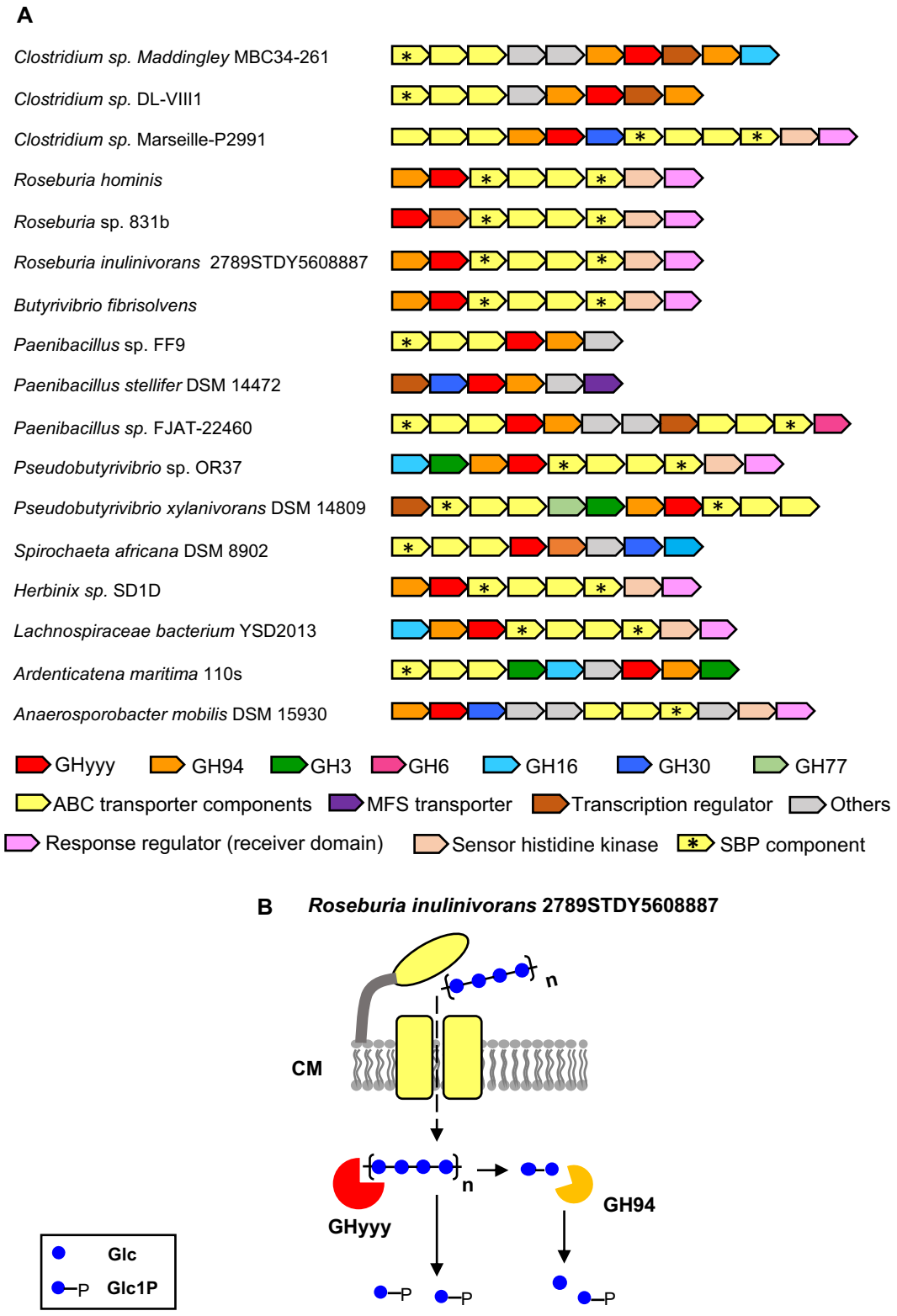


Figure 4.12. Genetic loci containing *GH94-GHyyy* genes in tandem and possible roles of the enzymes encoded by the genetic loci. (**A**) Gene clusters containing putative *GH94-GHyyy* genes in tandem arrangement. **(B)** Possible roles of GH94 and GHyyy in sequential degradation of oligosaccharide substrate. ABC transporter is coloured in yellow and located in CM. The subcellular localisation prediction of GHs and transporters was performed by PSORTb version 3.0.2 (http://psort.org/psortb/index.html) using default settings for Gram-positive bacteria. CM = cell membrane. Substrate-binding protein (SBP) components of ABC transport system is indicated by asterisks. Glc1P and disaccharide are released from the oligosaccharide degradation out by GHyyy. Glc1P can be converted to Glc6P by phosphoglucomutase and used in glycolysis.

ABC transporters were found prominently in the GHyyy-containing genetic loci that were analysed in this study, highlighting the importance of these transporters in carbohydrate transport in bacteria. Components of the identified ABC transporters include transmembrane permeases (COG1175 and COG0395), and SBPs (COG1653 and cd13585). Surprisingly, the majority of the ABC transporter genes in this study lacks the nucleotide binding domains (NBDs), which is required for ATP binding. The binding and hydrolysis of ATP typically provide the energy needed to drive the conformational changes of the ABC transporter complex that subsequently enables substrate uptake.[51] It is possible that incomplete ABC transporters found in GHyyy gene clusters may share the NBD component with other ABC transport systems in the corresponding bacteria. This hypothesis is reasonable considering the fact that NBD components are highly conserved in all ABC transporter systems.[51,52] A similar hypothesis has been proposed for a recently identified ABC transporter for β-(1→2)-glucan uptake in *Listeria innocua,* in which the gene encoding a putative NBD was found in a distant location in the genome and did not cluster with the genes encoding the remaining ABC transporter components.[53]

Some GHyyy-containing genetic loci have more than one copy of the SBP component in an ABC transporter (Figure 4.11 and 4.12A, SBP components are indicated by asterisks). The multiple SBPs may be functionalised with different ligand affinities and/or specificities, thus allowing the transporter to cope with a broad range of ligand concentrations, which would match with the constantly changing nutrient flux under various environmental conditions. Functionalisation of SBPs has been proposed for SBPs in an ABC mannose transport system from *Thermotoga maritima,* which contains two copies of the mannose-binding protein (48% sequence identity) with similar ligand specificity but approximately 1000-fold difference in ligand affinity, thus allowing this transport system to operate over a wide range of ligand concentrations.[54] Recently, an SBP in *Listeria innocua* has been shown to prefer oligosaccharide substrates with DP longer than 3 and its encoding gene co-localised with genes encoding β-(1→2)-glucanase and β-glucosidase, suggesting a role of the SBP in β-(1→2)-oligoglucan import.[53] SBP components identified in the GHyyy gene clusters could perform a similar function, although the *L. innocua* SBP and the SBPs from GHyyy-containing clusters did not share high amino acid sequence identity (<30%). Interestingly, gene clusters from *A. laidlawii, A. brassicae* and *A. oculi* contain SBPs associated with carbohydrate binding module (CBM)-like domain (Figure 4.11, yellow and black), which presumably assist the binding of oligosaccharide substrate. Association of CBM with an ABC transporter component was described for export of *O*-

antigenic polysaccharides in *E. coli* [55] and *Klebsiella pneumoniae*,[56] although CBM was found associated with the NBD component of the ABC transporter rather than the SBP.

4.2.4. Subcellular localisation of the heterokont GHyyy candidates

GHyyy candidates were discovered in the heterokonts in which chryslolaminarin metabolic enzymes have been previously reported through transcriptome and genome analyses, such as in *P. tricornutum* and *T. pseudonana*. Availability of the heterokont genomes enabled further analysis of the full-length GHyyy genes in the corresponding genomes, which showed that all of the analysed genes contain at least one intron (Table 4.4). The presence of introns in the GHyyy candidates coincides with the presence of introns in other β-(1→3)-glucan metabolic enzymes, such as laminarin synthase from *P. tricornutum* which contains 4 introns (ans_1_AGR_contig977, JGI database).[57] The presence of introns within these GHyyy genes strongly excludes the possibility that these sequences derive from bacterial contamination.

Table 4.4. Intron analysis of the *GHyyy* genes from heterokonts of which their genomes have been reported on GenBank.

Species and strain	RefSeq ID	Nucleotide ID	Gene loci	Introns
Phaeodactylum tricornutum CCAP 1055/1	XP_002177898.1	NC_011670.1	PHATRDRAFT_43370	4
Phaeodactylum tricornutum CCAP 1055/1	XP_002181175.1	NC_011680.1	PHATRDRAFT_54686	1
Thalassiosira pseudonana CCMP1335	XP_002290006.1	NC_012067.1	THAPSDRAFT_22483	4
Thalassiosira pseudonana CCMP1335	XP_002297139.1	NW_002243509.1	THAPSDRAFT_25518	9
Thalassiosira oceanica CCMP1005	EJK69967.1	AGNL01009288.1	THAOC_08721	6
Aureococcus anophagefferens	XP_009032155.1	NW_008705561.1	AURANDRAFT_10068	2
Symbiodinium microadriaticum	OLP88051.1	LSRX01000830.1	AK812_SmicGene30670	2
Fragilariopsis cylindrus CCMP1102	OEU10145.1	KV784373.1	FRACYDRAFT_194280	1
Fragilariopsis cylindrus CCMP1102	OEU15658.1	KV784359.1	FRACYDRAFT_170346	2

Recently, a β-(1→3)-D-glucan synthase (*Pt*BGS) and two β-(1→6)-transglycosylases (*Pt*TGS1 and 2) from *P. tricornutum* have been demonstrated to localise in the tonoplast (vacuolar membrane) of the heterokont.[38,57] Based on the subcellular localisation data of *Pt*BGS and *Pt*TGSs, it is likely that chrysolaminarin metabolism might be centred in the vacuole and that other metabolic enzymes including the heterokont GHyyy candidates may be located in this organelle.

Sequence prediction for vacuolar targeting signal is possible, although it was anticipated that the prediction will not be straightforward because vacuolar sorting is complex and requires three different types of targeting signals; 1) N-terminal signals for targeting to the ER, 2) C-terminal motifs for exit from the ER, and 3) C-terminal motifs

that promote sorting to the tonoplast.[58] *Pt*TGS1 and *Pt*TGS2 were reported to have Tyr-based sorting signals (YxxΦ, where xx are any amino acids and Φ is a bulky hydrophobic acid) and di-Leu-type sorting signals ([D/E]XXXL[L/I]), both of which are typical signals for ER exit and final targeting to destination (e.g. tonoplast and lysosomes) in plants,[59,60] animals and yeasts.[61] These two motifs are found at the C-termini of *Pt*TGS1 and *Pt*TGS2 proteins.[38] In order to determine whether the same C-terminal signals are present in the heterokont GHyyy, the *Pt*TGS1 and the heterokont GHyyy sequences were aligned to identify potential di-Leu-type and Tyr-based sorting signals. The sequence alignment did not reveal any obvious consensus of these C-terminal sorting signals in the GHyyy sequences (Figure 4.13). Therefore, further *in vivo* experiments are required to confirm the subcellular localisation of the GHyyy sequences.

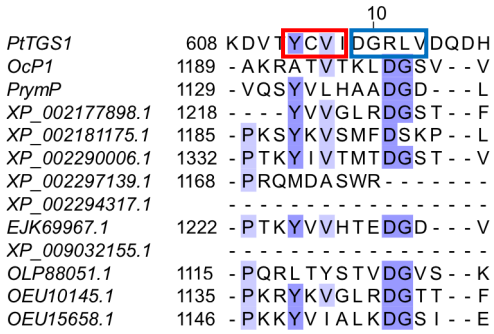


Figure 4.13. Multiple sequence alignment of GHyyy amino acid sequences against *Pt*TGS1 to identify the C-terminal sorting signals. The predicted Tyr-based and di-Leu-type sorting signals in *Pt*TGS1 are indicated by red and blue boxes, respectively. The organism sources of the sequences are as followed; OcP1 [*Ochromonas* sp. BG-1]. *Prym*P [*Prymnesium parvum*], XP_002177898.1 [*Phaeodactylum tricornutum* CCAP 1055/1], XP_002181175.1[*Phaeodactylum tricornutum* CCAP 1055/1], XP_002290006.1 [*Thalassiosira pseudonana* CCMP1335], XP_002297139.1 *Thalassiosira pseudonana* CCMP1335, EJK69967.1 [*Thalassiosira oceanica* CCMP1005], XP_009032155.1 [*Aureococcus anophagefferens*], OLP88051.1 [*Symbiodinium microadriaticum*], OEU10145.1 [*Fragilariopsis cylindrus* CCMP1102], OEU15658.1 [*Fragilariopsis cylindrus* CCMP1102].

4.3. Discussion

From the bioinformatic analysis of sequences from GH94, GH149 and GHyyy families, the conservation of amino acids involved in catalysis and recognition of sugar donor and inorganic phosphate are highly conserved, suggesting possible divergent evolution of these families whilst maintaining the catalytic apparatus. However, the characterised enzymes from the three families showed different substrate preferences, in which GH94 LBP only operates on G2, whilst the characterised GH149 and GHyyy are oligosaccharide phosphorylases catalysing phosphorolysis of longer substrates (Figure 4.14). Among GH149 and GHyyy, the minimum chain lengths required for phosphorolysis were also different, in which GH149 required DP2, whilst GHyyy required DP3. The difference between the substrate chain length preferences in GH94, GH149 and GHyyy indicates less conservation of the acceptor binding subsites among these enzymes. It is interesting to note that the tolerance towards β-linked disaccharide acceptors with different linkage regiochemistry observed in both Pro_7066 and *Pap*P indicates a somewhat conserved acceptor recognition mechanism among GH149 and GHyyy enzymes (Figure 4.6-4.9).

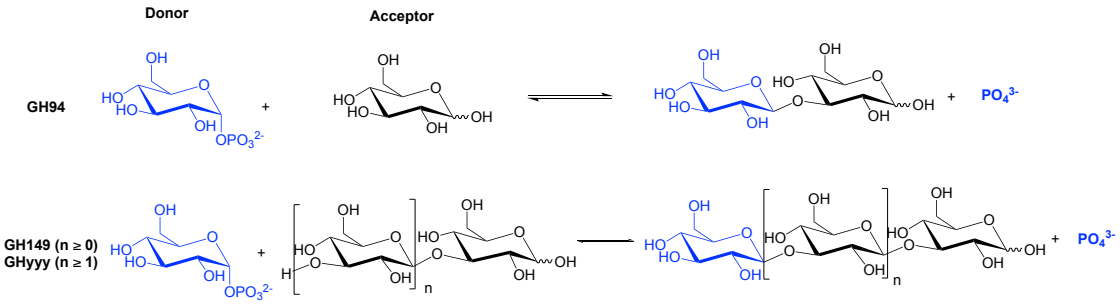


Figure 4.14. Summary of characterised GH94, GH149 and GHyyy activities.

It is highly likely that the donor subsite of *Pap*P will be similar to that of *Ps*LBP and Pro_7066 because the amino acids in this subsite are conserved in all three enzymes. To confirm this hypothesis, crystallisation experiments were performed on the recombinant *Pap*P protein in order to obtain crystals for X-ray crystallographic study (Chapter 6 section 6.15.4). However, the diffraction data was too poor for structure determination (>5 Å). In order to improve the diffraction resolution, post-crystallisation treatments such as crystal dehydration could be performed.[62–64] Homology modelling of *Pap*P based on either *Ps*LBP or Pro_7066 structure is possible, although this may not give a reliable model because the overall sequence identity of *Pap*P to *Ps*LBP and Pro_7066 is very low (20%).

*Pap*P exhibited preference for acceptor with DP ≥ 2, similar to that described for *Lp*SOGP which is a GH94 β-(1→2)-oligoglucan phosphorylase.[65] The crystal structure of *Lp*SOGP has been solved with sophorose (β-(1→2)-linked disacccahride) occupying +1 and +2 acceptor subsites. The Glc residue at the +1 subsite is rotated by more than 120˚about its C2-O2 bond, in comparison to the Glc moiety in *Cg*CBP and *Vp*ChBP at the same subsite. This distortion disfavoured the interaction between the Glc moiety at the +1 subsite in the *Lp*SOGP structure. However, the unfavourable interaction at +1 subsite is compensated by the binding of the second Glc residue at +2 subsite. Therefore, the interaction at the +2 subsite in the *Lp*SOGP could be essential for the recognition of sophorose and that Glc alone would not be able to bind stably to the +1 subsite.[65] A similar mechanism may be utilised by *Pap*P to recognise acceptors with DP ≥ 2. Alignment of *Pap*P and *Lp*SOGP amino acid sequences did not reveal any significant conservation of key amino acids that are present in *Lp*SOGP +1 and +2 subsites (R907, Y141, G623, G624, A625 and A626).[65] Thus, whether *Pap*P and *Lp*SOGP use a similar acceptor recognition mechanism cannot be deduced from sequence alignment alone.

The presence of GHyyy in the laminarin utilisation locus from *B. weihaiensis* (Figure 4.10A) and the *Pap*P enzyme specificity for β-(1→3)-glucan would indicate that the GHyyy family contains enzymes acting specifically on β-(1→3)-gluco-oligosaccharides. However a recently characterised cellodextrin (β-(1→4)-glucan) phosphorylase from a Gram-negative thermophilic bacterium, *Thermosipho africanus* TCF52B[66] (*Ta*CDP, GenBank ID = ACJ76363.1) was discovered as an OcP1 orthologue during the PSI-BLAST analysis in this study. This may suggest that the GHyyy family contains both β-(1→3)- and β-(1→4)-glucan-specific enzymes. *Ta*CDP has previously been classified as a GH94 enzyme due to its 11% sequence identity to a GH94 *Ruminiclostridium thermocellum* CBP sequence (GenBank ID: ABN51514.1).[66] However, BLASTP analysis of *Ta*CDP against GHyyy sequences strongly suggests a reclassification of *Ta*CDP to GHyyy family, because it shares greater sequence identity to GHyyy members (36% and 46% to OcP1 and *Pap*P, respectively). Although *Ta*CDP has been characterised as a phosphorylase acting on β-(1→4)-glucan, it is worth noting that the activity of *Ta*CDP on β-(1→3)-gluco-oligosaccharides has not been investigated,[66] and this activity cannot be ruled out.

The *Ta*CDP-encoding gene in the *T. africanus* genome co-localised with genes encoding β-glucosidase (GH3), β-glucanase (GH16) and ABC transporter components (Figure 4.15A), implying a role of *Ta*CDP in degradation of β-(1→4)-glucan. However,

*Ta*CDP showed much greater k_{cat}/K_M values for the synthetic reactions compared to the phosphorolysis reactions (Table 4.5) and has the highest catalytic efficiency among oligosaccharide phosphorylases reported thus far.[66] Such high catalytic efficiency for the glycan synthesis has led to a speculation that the enzyme may be involved in oligosaccharide formation rather than degradation,[66] although further investigation of the role of this enzyme is required since the catalytic efficiency was determined *in vitro* which may not reflect the physiological conditions where the enzyme normally operates.

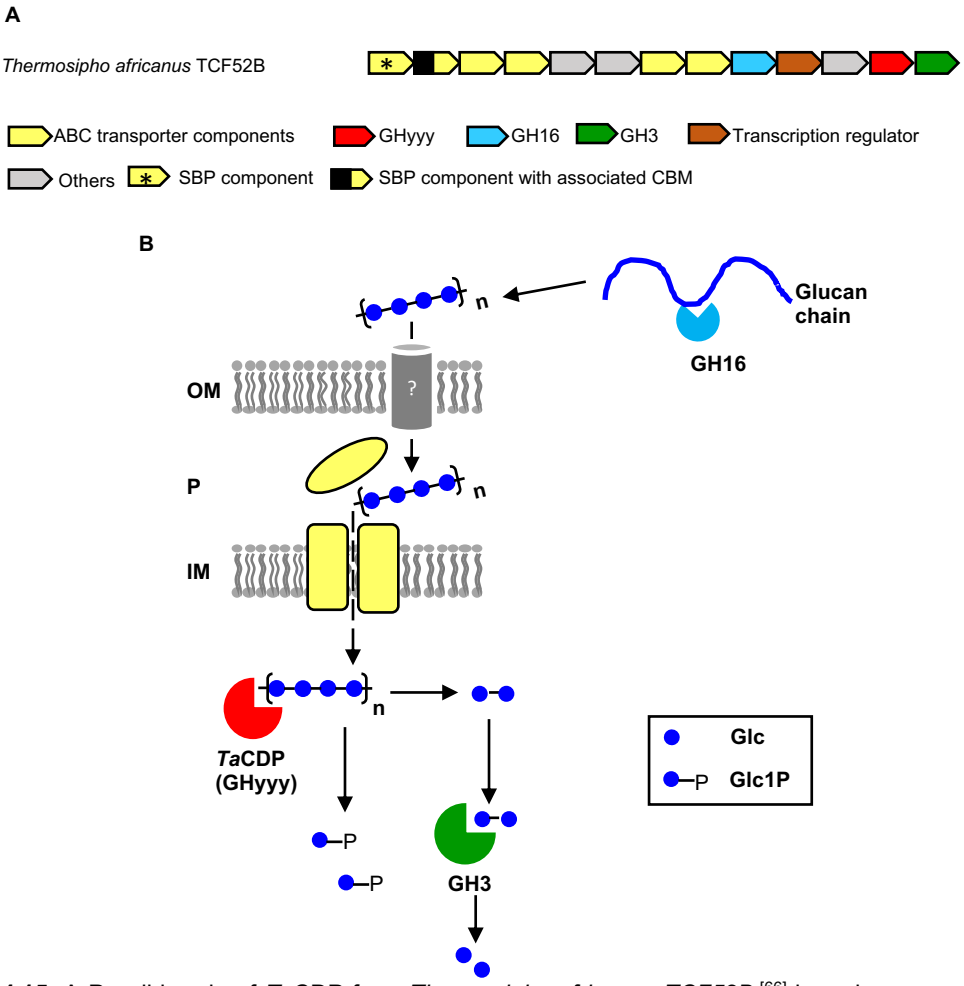


Figure 4.15. A Possible role of *Ta*CDP from *Thermosipho africanus TCF52B,*[66] based on gene cluster analysis. (**A**) gene cluster containing *Ta*CDP. (**B**) Probable roles of the enzymes encoded by the gene cluster in **A**. ABC transporter are depicted in yellow located in IM. The subcellular localisation prediction of GHs and transporters was performed by PSORTb version 3.0.2 (http://psort.org/psortb/index.html) using default settings for Gram-negative bacteria. OM = outer membrane, P = periplasmic space, IM = inner membrane. SBP = substrate-binding component of ABC transporter. Glc1P and disaccharide are released from the oligosaccharide degradation out by GHyyy. Glc1P can be converted to Glc6P by phosphoglucomutase and used in glycolysis.

Table 4.5. Reported catalytic efficiency [k_{cat}/K_m ($s^{-1}mM^{-1}$)] of *Ta*CDP.[66]

Substrates	Synthesis	Phosphorolysis
Glc	2.12	N/A
Cellobiose	48.87	1.92
Cellotriose	148.90	2.98
Cellotetraose	165.46	7.91
Cellopentaose	203.50	5.67

GHyyy candidates were identified predominantly in *Paenibacillus* spp., suggesting an important role of GHyyy in this genus. Comparison between GHyyy-containing regulons from *Paenibacillus* spp. showed that different *Paenibacillus* species adopt different genetic organisations of the regulons. However, the inheritance of the entire regulon seems to be conserved among different strains of the same species (i.e. *P. polymyxa* ATCC842, E681 and CR1 all share the same GHyyy-containing regulon) (Figure 4.10 and 4.11). It is worth noting that the prevalence of GHyyy sequences in *Paenibacillus* spp. could be due to the overrepresentation of the sequencing data from the databases that were analysed in this study.

The GC content analysis of *GHyyy* genes showed no significant deviation from the median GC content of the corresponding genomes (Appendix 9) with the exception of a *GHyyy* gene (AK812_SmicGene30670) from a dinoflagellate, *Symbiodinium microadriaticum,* which is a symbiont of the jelly fish *Cassiopea xamachana.*[67] The *S. microadriaticum GHyyy* gene contains 8% higher GC content compared to its associated genome (GenBank ID: LSRX00000000.1). It is possible that the *S. microadriaticum GHyyy* gene has been acquired through endosymbiotic/horizontal gene transfer from other organism(s), the mechanism of which has been considered as an important source of genetic variation in the dinoflagellates.[68] In contrast, the *GHyyy* genes that have similar GC content to their associated genomes may have been inherited via vertical gene transfer within the same species.

4.4. Conclusion and future perspectives

Identification of a new GH family GHyyy and characterisation of a GHyyy member, *Pap*P, extend the known repertoire of GPs acting on β-(1→3)-D-glucans, albeit with different acceptor substrate preferences from GH94 and GH149. The remaining challenge is characterisation of the heterokont GHyyy sequences in order to confirm their functions. The expression of heterokont GHyyy in Eukaryotic hosts such as yeast could provide a solution to recombinant protein production. Structural studies of *Pap*P should be further pursued in order to study the enzyme structural features governing its acceptor specificity. Analysis of gene clusters containing GHyyy genes provides the first evidence of possible coordination between GH94 and GH149 enzymes in sequential degradation of oligosaccharide substrates. Further characterisation of other co-localised GHs and the transporters encoded in the GHyyy-containing gene clusters could provide further support to the significance of these gene clusters and their encoded proteins in oligosaccharide degradation. Finally, the distinct distribution of GH149 and GHyyy sequences in Gram-negative and Gram-positive bacteria, respectively, suggests that the two bacterial groups acquired different enzymatic and transport machineries through the course of evolution for degradation of oligosaccharide substrates.

4.5. References

[1] M. Kitaoka, Y. Matsuoka, K. Mori, M. Nishimoto, K. Hayashi, *Biosci. Biotechnol. Biochem.* **2012**, *76*, 343–348.

[2] T. Nihira, Y. Saito, M. Kitaoka, M. Nishimoto, K. Otsubo, H. Nakai, *Carbohydr. Res.* **2012**, *361*, 49–54.

[3] L. R. Maréchal, S. H. Goldemberg, *Biochem. Biophys. Res. Commun.* **1963**, *13*, 2–5.

[4] L. R. Marechal, *Biochem. Biophys. Acta* **1967**, *146*, 417–430.

[5] M. Kitaoka, T. Sasaki, H. Taniguchi, *Arch. Biochem. Biophys.* **1993**, *304*, 508–514.

[6] Y. Yamamoto, D. Kawashima, A. Hashizume, M. Hisamatsu, N. Isono, *Biosci. Biotechnol. Biochem.* **2013**, *77*, 1949–54.

[7] G. J. Albrecht, H. Kauss, *Phytochemistry* **1971**, *10*, 1293–1298.

[8] M. Hidaka, M. Kitaoka, K. Hayashi, T. Wakagi, H. Shoun, S. Fushinobu, *Biochem. J.* **2006**, *398*, 37–43.

[9] M. Hidaka, Y. Honda, M. Kitaoka, S. Nirasawa, K. Hayashi, T. Wakagi, H. Shoun, S. Fushinobu, *Structure* **2004**, *12*, 937–947.

[10] A. Van Hoorebeke, J. Stout, J. Kyndt, M. De Groeve, I. Dix, T. Desmet, W. Soetaert, J. Van Beeumen, S. N. Savvides, *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **2010**, *66*, 346–351.

[11] C. M. Bianchetti, N. L. Elsen, B. G. Fox, G. N. Phillips, *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **2011**, *67*, 1345–1349.

[12] Y.-W. Nam, T. Nihira, T. Arakawa, Y. Saito, M. Kitaoka, H. Nakai, S. Fushinobu, *J. Biol. Chem.* **2015**, *290*, 18281–18292.

[13] E. C. O'Neill, G. Pergolizzi, C. E. M. Stevenson, D. M. Lawson, S. A. Nepogodiev, R. A. Field, *Carbohydr. Res.* **2017**, *451*, 118–132.

[14] M. A. Caballero, D. Jallet, L. Shi, C. Rithner, Y. Zhang, G. Peers, *Algal Res.* **2016**, *20*, 180–188.

[15] A. Beattie, E. L. Hirst, E. Percival, *Biochem J* **1961**, *79*, 531–537.

[16] B. S. Paulsen, S. Myklestad, *Carbohydr. Res.* **1978**, *62*, 386–388.

[17] B. A. Wustman, M. R. Gretz, K. D. Hoagland, *Plant Physiol.* **1997**, *113*, 1059–1069.

[18] A. Chiovitti, A. Bacic, J. Burke, R. Wetherbee, *Eur. J. Phycol.* **2003**, *38*, 351–360.

[19] B. Gügi, T. Le Costaouec, C. Burel, P. Lerouge, W. Helbert, M. Bardor, *Mar. Drugs* **2015**, *13*, 5993–6018.

[20] T. R. Størseth, K. Hansen, K. I. Reitan, J. Skjermo, *Carbohydr. Res.* **2005**, *340*, 1159–1164.

[21] E. Granum, E. Granum, S. M. Myklestad, S. M. Myklestad, *Hydrobiologia* **2002**, *477*, 155–161.

[22] L. E. Rioux, S. L. Turgeon, M. Beaulieu, *Carbohydr. Polym.* **2007**, *69*, 530–537.

[23] T. E. Nelson, B. A. Lewis, *Carbohydr. Res.* **1974**, *33*, 63–74.

[24] L. E. Rioux, S. L. Turgeon, M. Beaulieu, *Phytochemistry* **2010**, *71*, 1586–1595.

[25] S. U. Kadam, B. K. Tiwari, C. P. O'Donnell, *Int. J. Food Sci. Technol.* **2015**, *50*, 24–31.

[26] A. Bacic, G. B. Fincher, B. A. Stone, *Chemistry, Biochemistry, and Biology of (1→3)-β-Glucans and Related Polysaccharides*, Academic Press/Elsevier, San Diego, **2009**.

[27] Z. Liu, Y. Xiong, L. Yi, R. Dai, Y. Wang, M. Sun, X. Shao, Z. Zhang, S. Yuan, *Carbohydr. Polym.* **2018**, *194*, 339–349.

[28] P. G. Kroth, A. Chiovitti, A. Gruber, V. Martin-Jezequel, T. Mock, M. S. Parker, M. S. Stanley, A. Kaplan, L. Caron, T. Weber, U. Maheswari, E. V Armbrust, C. Bowler, *PLoS One* **2008**, *3*, e1426.

[29] M. Fabris, M. Matthijs, S. Rombauts, W. Vyverman, A. Goossens, G. J. E. Baart, *Plant J.* **2012**, *70*, 1004–1014.

[30] B. H. Zhu, H. P. Shi, G. P. Yang, N. N. Lv, M. Yang, K. H. Pan, *N. Biotechnol.* **2016**, *33*, 237–244.

[31] F. Daboussi, S. Leduc, A. Maréchal, G. Dubois, V. Guyot, C. Perez-Michaut, A. Amato, A. Falciatore, A. Juillerat, M. Beurdeley, D. F. Voytas, L. Cavarec, P. Duchateau, *Nat. Commun.* **2014**, *5*, 3831.

[32] M. S. Chauton, P. Winge, T. Brembu, O. Vadstein, A. M. Bones, *Plant Physiol.* **2013**, *161*, 1034–1048.

[33] A. Graf, A. Schlereth, M. Stitt, A. M. Smith, *Proc Natl Acad Sci* **2010**, *107*, 9458–9463.

[34] A. Graf, A. M. Smith, *Trends Plant Sci.* **2011**, *16*, 169–75.

[35] S. C. Zeeman, S. M. Smith, A. M. Smith, *Biochem. J.* **2007**, *401*, 13–28.

[36] E. V. Armbrust, J. A. Berges, C. Bowler, B. R. Green, D. Martinez, N. H. Putnam, S. Zhou, A. E. Allen, K. E. Apt, M. Bechner, M. A. Brzezinski, B. K. Chaal, A. Chiovitti, A. K. Davis, M. S. Demarest, J. C. Detter, T. Glavina, D. Goodstein, M. Z. Hadi, U. Hellsten, M. Hildebrand, B. D. Jenkins, J. Jurka, V. V. Kapitonov, N. Kröger, W. W. Y. Lau, T. W. Lane, F. W. Larimer, J. C. Lippmeier, S. Lucas, M. Medina, A. Montsant, M. Obornik, M. S. Parker, B. Palenik, G. J. Pazour, P. M. Richardson, T. A. Rynearson, M. A. Saito, D. C. Schwartz, K. Thamatrakoln, K.

Valentin, A. Vardi, F. P. Wilkerson, D. S. Rokhsar, *Science* **2004**, *306*, 79–86.

[37] M. Hildebrand, K. Manandhar-Shrestha, R. Abbriano, *Algal Res.* **2017**, *23*, 66–77.

[38] W. Huang, C. Río Bártulos, P. G. Kroth, *J. Eukaryot. Microbiol.* **2016**, *63*, 536–546.

[39] G. Michel, T. Tonon, D. Scornet, J. M. Cock, B. Kloareg, *New Phytol.* **2010**, *188*, 67–81.

[40] R. Anand, S. Grayston, C. Chanway, *Microb. Ecol.* **2013**, *66*, 369–374.

[41] N. Sawhney, C. Crooks, V. Chow, J. F. Preston, F. J. St John, *BMC Genomics* **2016**, *17*, 131.

[42] Y.-M. Cheng, T.-Y. Hong, C.-C. Liu, M. Meng, *Appl. Microbiol. Biotechnol.* **2009**, *81*, 1051–1061.

[43] V. Chow, Y. S. Kim, M. S. Rhee, N. Sawhney, F. J. St. John, G. Nong, J. D. Rice, J. F. Preston, *Appl. Environ. Microbiol.* **2016**, *82*, 1789–1798.

[44] Y. Peilong, P. Shi, Y. Wang, Y. Bai, K. Meng, H. Luo, T. Yuan, B. Yao, *J. Microbiol. Biotechnol.* **2006**, *17*, 58–66.

[45] M. I. Trindade, V. R. Abratt, S. J. Reid, *Appl. Environ. Microbiol.* **2003**, *69*, 24–32.

[46] T. Nihira, E. Suzuki, M. Kitaoka, M. Nishimoto, K. Ohtsubo, H. Nakai, *J. Biol. Chem.* **2013**, *288*, 27366–27374.

[47] S. Kuhaudomlarp, N. J. Patron, B. Henrissat, M. Rejzek, G. Saalbach, R. A. Field, *J. Biol. Chem.* **2018**, *293*, 2865–2876.

[48] Y. Zhu, P. Chen, Y. Bao, Y. Men, Y. Zeng, J. Yang, J. Sun, Y. Sun, *Sci. Rep.* **2016**, *6*, 38248.

[49] A. Maqbool, R. S. P. Horler, A. Muller, A. J. Wilkinson, K. S. Wilson, G. H. Thomas, *Biochem. Soc. Trans.* **2015**, *43*, 1011–1017.

[50] K. Chiku, T. Nihira, E. Suzuki, M. Nishimoto, M. Kitaoka, K. Ohtsubo, H. Nakai, *PLoS One* **2014**, *9*, e114882.

[51] A. L. Davidson, E. Dassa, C. Orelle, J. Chen, *Microbiol. Mol. Biol. Rev.* **2008**, *72*, 317–364.

[52] Y. Quentin, G. Fichant, F. Denizot, *J. Mol. Biol.* **1999**, *287*, 467–484.

[53] K. Abe, N. Sunagawa, T. Terada, Y. Takahashi, T. Arakawa, K. Igarashi, M. Samejima, H. Nakai, H. Taguchi, M. Nakajima, S. Fushinobu, *J. Biol. Chem.* **2018**, *293*, 8812–8828.

[54] S. Ghimire-Rijal, X. Lu, D. A. Myles, M. J. Cuneo, *J. Biol. Chem.* **2014**, *289*, 30090–30100.

[55] L. Cuthbertson, M. S. Kimber, C. Whitfield, *Proc. Natl. Acad. Sci.* **2007**, *104*,

19529–19534.

[56] E. Mann, E. Mallette, B. R. Clarke, M. S. Kimber, C. Whitfield, *J. Biol. Chem.* **2016**, *291*, 9748–9761.

[57] W. Huang, I. Haferkamp, B. Lepetit, M. Molchanova, S. Hou, W. Jeblick, C. Río Bártulos, P. G. Kroth, *Proc. Natl. Acad. Sci.* **2018**, *115*, 4791–4796.

[58] E. Pedrazzini, N. Y. Komarova, D. Rentsch, A. Vitale, *Traffic* **2013**, *14*, 622–628.

[59] S. Wolfenstetter, P. Wirsching, D. Dotzauer, S. Schneider, N. Sauer, *Plant Cell* **2012**, *24*, 215–232.

[60] S. Isayenkov, J.-C. Isner, F. J. M. Maathuis, *Plant Cell* **2011**, *23*, 756–768.

[61] J. S. Bonifacino, L. M. Traub, *Annu. Rev. Biochem.* **2003**, *72*, 395–447.

[62] B. Heras, J. L. Martin, *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2005**, *61*, 1173–1180.

[63] C. Abergel, *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2004**, *60*, 1413–1416.

[64] J. Newman, *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2006**, *62*, 27–31.

[65] M. Nakajima, N. Tanaka, N. Furukawa, T. Nihira, Y. Kodutsumi, Y. Takahashi, N. Sugimoto, A. Miyanaga, S. Fushinobu, H. Taguchi, H. Nakai, *Sci. Rep.* **2017**, *7*, 42671.

[66] Y. Wu, G. Mao, H. Fan, A. Song, Y.-H. P. Zhang, H. Chen, *Sci. Rep.* **2017**, *7*, 4849.

[67] B. S. Holland, M. N. Dawson, G. L. Crow, D. K. Hofmann, *Mar. Biol.* **2004**, *145*, 1119–1128.

[68] J. H. Wisecaver, M. L. Brosnahan, J. D. Hackett, *Genome Biol. Evol.* **2013**, *5*, 2368–2381.

Chapter 5

Summary and translational opportunities

5.1. Introduction

Research into the generation of β-(1→3)-D-glucan polymers *in vitro* is of interest due to a wide range of β-(1→3)-D-glucan applications in the food and pharmaceutical industries. Enzymatic synthesis is an attractive route for β-(1→3)-D-glucan production due to its regio- and stereo-specificity. Glycoside phosphorylases (GPs) with specificity for β-(1→3)-glycosidic bonds are attractive biocatalytic candidates for the enzymatic synthesis because of the relatively cheap substrates required (D-glucose (Glc) and α-D-glucose 1-phosphate (Glc1P)). GPs acting on β-D-glucopyranosyl-(1→3)-D-glucopyranose (laminaribiose) have previously been described in bacteria *Paenibacillus* sp. YM-1 (*Ps*LBP)[1] and *Acholeplasma laidlawii* PG-8A,[2] and the nucleotide sequences responsible for their activities are known, which allow their classification into the GH94 family. On the other hand, GPs acting on longer β-(1→3)-D-gluco-oligosaccharides have been described in Eukaryotic microalga (*Euglena gracilis*)[3–5] and heterokonts *(Ochromonas danica*[6] and *Ochromonas malhemensis*[7]). The microalgal enzymes are the desirable biocatalysts for production of long chain β-(1→3)-D-glucan, but nucleotide sequence encoding the activities was not available prior to this investigation, making it difficult to produce these enzymes for use in biocatalysis. Therefore, identification of the nucleotide sequences encoding the microalgal β-(1→3)-D-glucan phosphorylase activities was the primary aim of this work, which would enable production of these prospective biocatalysts for application in enzymatic synthesis of linear β-(1→3)-D-gluco-oligosaccharides.

The molecular mechanism underlying the activity of GH94 *Ps*LBP was investigated in chapter 2, which revealed a somewhat relaxed donor specificity toward a noncognate donor, Man1P. Structural studies of *Ps*LBP revealed the architecture of the *Ps*LBP active site and -1 donor subsite, which are similar to that previously observed in other GH94 enzymes. However, an additional β-hairpin gate near the active site was identified in the *Ps*LBP structure, which likely defines the preference of *Ps*LBP for disaccharide substrates/products.[1]

In chapter 3, the nucleotide sequence of laminaridextrin phosphorylase from a microalga *Euglena gracilis* (EgP1) was identified from proteomic analysis of a partially purified protein fraction containing β-(1→3)-D-glucan phosphorylase activity. A bioinformatics search for EgP1 orthologues identified several hundred sequences, which altogether constituted a new family, designated GH149. EgP1 and its bacterial

orthologue (Pro_7066) were produced as recombinant proteins and characterised to confirm their function as β-(1→3)-D-glucan phosphorylases. Structural studies of the GH94 *Ps*LBP and GH149 Pro_7066 confirmed the conservation of key amino acid residues that are involved in catalysis and validated the placement of both families within the same GH-clan (GH-Q). However, Pro_7066 contains additional domains flanking its catalytic domain and an oligosaccharide surface binding site, which are unique to the GH149 enzyme and likely contribute to defining the enzyme's ability to operate on longer oligosaccharide substrates than *Ps*LBP.

In chapter 4, the *Ochromonas* β-(1→3)-D-glucan phosphorylase sequence candidate (OcP1) was identified through BLAST analysis of the diatom's transcriptome using GH94 query sequences. Further BLAST searches for OcP1 orthologues identified several hundred sequences from heterokonts and bacteria. OcP1 and its orthologues are distinct from GH94 and GH149 sequences, thus forming a new enzyme family designated GHyyy. Multiple sequence alignment of OcP1, GH94 and GH149 enzymes revealed conservation of the key amino acids that were previously identified in the active site and -1 subsite of GH94 and GH149 enzymes, despite very low or no overall sequence homology among these sequences. The conservation of key catalytic residues suggests that the catalytic mechanism in these three groups of enzymes are conserved and that they can be classified into the same GH clan (GH-Q). A bacterial GHyyy candidate from *Paenibacillus polymyxa*, *Pap*P, was produced as a recombinant protein, characterisation of which showed that the enzyme operated on β-(1→3)-gluco-oligosaccharide acceptors with degree of polymerisation (DP) ≥ 2 and could not use Glc as its acceptor substrate. The acceptor length specificity of *Pap*P was clearly distinct from the characterised GH149 enzymes, which could act on Glc.

In order to postulate possible physiological roles of GH149 and GHyyy in the organisms possessing the enzymes, the location of the bacterial genes encoding either GH149 or GHyyy were mapped in their corresponding genomes. This revealed the co-localisation of the GH149 or GHyyy genes to genetic clusters containing genes encoding sugar transporters and other putative GH enzymes such as glucanases, glucosidases and/or GH94 GPs, which strongly suggests a role for GH149 and GHyyy enzymes in oligosaccharide degradation. However, the architecture of the gene clusters varied between different bacterial species and the distribution of these enzymes in bacteria are different; GH149 genes were identified mostly from Gram-negative bacteria whilst GHyyy were identified from Gram-positive species. This finding implies separate

acquisition of GH149 and GHyyy genes by different groups of bacteria during the course of evolution. On the other hand, the role of microalgal GH149 and heterokont GHyyy have not been investigated during the time of this study due to the lack of expertise in knockdown experiments to study possible functions of these sequence *in vivo*. Regardless of their physiological roles in oligosaccharide degradation, characterisation of GH149 and GHyyy enzymes in the glycan synthetic reactions have proven their potential use as biocatalysts in the *in vitro* synthesis of β-(1→3)-gluco-oligosaccharides.

Building upon the work I have presented in the previous chapters, the following sections outline preliminary data regarding the applications of the newly identified GPs in enzymatic synthesis of oligosaccharides. Furthermore, this chapter highlights possible strategies in controlling chain length of the oligosaccharide products, strategies for GP improvement via protein engineering and directed evolution, and development of high-throughput functional screens for identification of new GP activities.

5.2. Application of β-(1→3)-D-glucan phosphorylases in oligosaccharide synthesis

As have recently been reviewed,[8] GPs have been used for oligosaccharide synthesis in one-pot enzymatic approaches, where two or more enzymes are combined in one reaction to produce a final product through a series of reactions catalysed by different enzymes, providing that suitable substrates are supplied. This approach has been used for a large scale production of prebiotic and probiotic oligosaccharides, lacto-*N*-biose (LNB)[9] and galacto-*N*-biose (GNB) (Figure 5.1A).[10] The one-pot enzymatic approach can be further complemented by improving the stability and reusability of biocatalysts by enzyme immobilisation on solid support.[11] This has been demonstrated for the production of laminaribiose, which was accomplished by a bi-enzymatic system containing individually immobilised *Euglena gracilis* protein extract with laminaribiose phosphorylase (LBP) activity and recombinant SP onto Sepabeads (Figure 5.1B).[12] The immobilised LBP retained more than 50% activity when used seven times, providing a proof-of-concept that could potentially be implemented for other GPs.

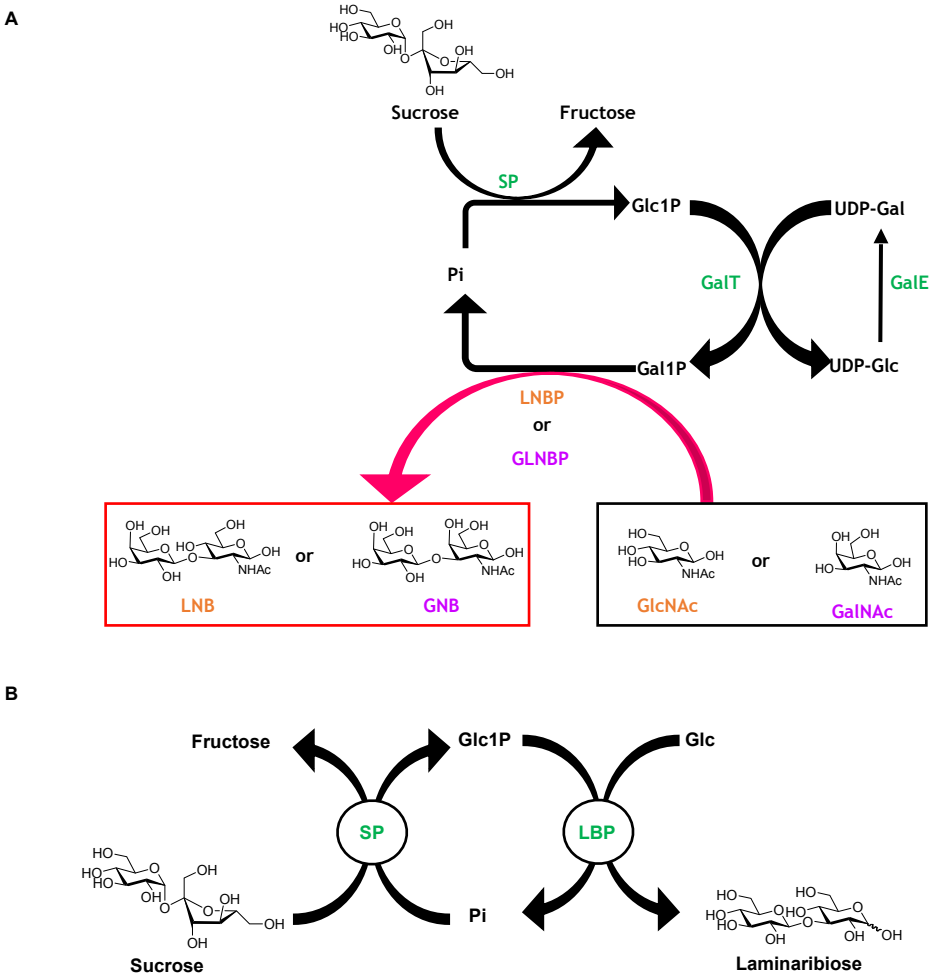


Figure 5.1. The use of GPs as biocatalysts in one-pot enzymatic reactions for oligosaccharide production. (**A**) Two biosynthetic routes for GNB and LNB production proposed by Nishimoto and Kitaoka.[9,10] Commonly supplied biocatalysts (SP, GalT and GalE) are highlighted in green. For LNB production, Lacto-*N*-biose phosphorylase (LNBP) and D-GlcNAc were used as the final biocatalyst and substrate, respectively (in orange). Galacto-*N*-biose/lacto-*N*-biose I phosphorylase (GLNBP) and GalNAc were used as the final biocatalyst and substrate, respectively, for the production of GNB (in magenta). (**B**) Immobilised partially purified laminaribiose phosphorylase (LBP) from *E. gracilis* was used in combination with immobilised SP for the production of laminaribiose in the presence of sucrose as a starting material.[12] LBP = laminaribiose phosphorylase from *E. gracilis*. SP = sucrose phosphorylase.

To illustrate the use of the one-pot enzymatic approach in β-(1→3)-D-glucan production, an enzymatic reaction containing a commercial sucrose phosphorylase (SP) and Pro_7066 from GH149 was implemented (Figure 5.2). The presence of SP as the first biocatalyst of the process enabled the use of sucrose as a substrate. Glc1P and Glc could be obtained simultaneously *in situ* from sucrose because SP possesses both phosphorylase and hydrolase activities (Figure 5.3A).[13] Glc and Glc1P were then used as substrates by Pro_7066, which catalysed the synthesis of β-(1→3)-gluco-oligosaccharides with DP up to 22 after 25 hr incubation in the presence of 20 mM sucrose as a starting substrate (Figure 5.3B and C). Inorganic phosphate generated during the glycan synthesis catalysed by Pro_7066 can be reused by SP for the

phosphorolysis of sucrose in the first step. In principle, the enzymatic cycle should continue, providing that sucrose substrate is continuously supplied.

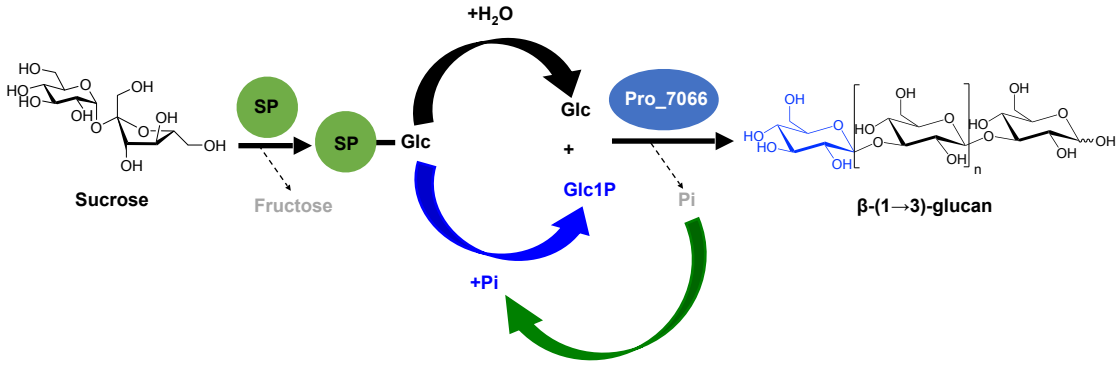


Figure 5.2. Schematic representation of a one-pot enzymatic reaction using sucrose phosphorylase (SP) and Pro_7066 from GH149 as biocatalysts and sucrose and inorganic phosphate as starting materials.

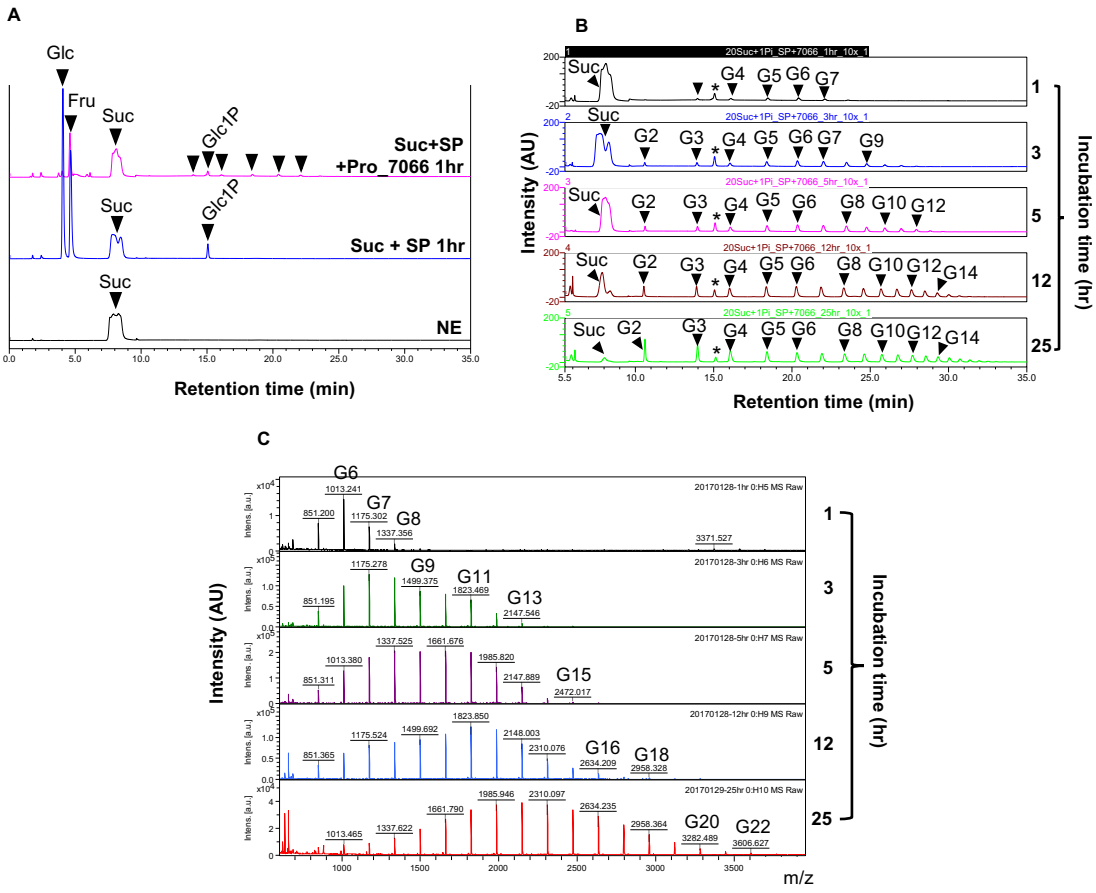


Figure 5.3. Analysis of oligosaccharide products obtained from the one-pot enzymatic reaction. (**A**) HPAEC-PAD analysis of the reactions either in the presence of sucrose phosphorylase (SP) alone or the coupled reaction in the presence of SP and Pro_7066 after 1 hr. Arrow heads without labels represent the peaks corresponding to oligosaccharide products. NE = no enzyme control. (**B**) HPAEC-PAD analysis of the coupled reaction at different time points. DP of the products are indicated (G2-G14). (**C**) MALDI-ToF analysis of the time course reactions in **B**. Suc = sucrose, Fru = fructose.

5.3. Activity of GH149 phosphorylases on a noncognate sugar donor Man1P

The relaxed donor specificity of GPs has been observed in GH94 enzymes such as *Rt*CDP, which the Field group has reported to be able to use noncognate donors, Gal1P and GlcN1P, as substrates with relatively good turnover for glycan synthetic reactions in the presence of a fluorescently labelled cellotriose as an acceptor.[14] The relaxed donor specificity of another GH94 enzyme (*Ps*LBP) was also explored in chapter 2, which showed that a non-cognate sugar donor Man1P could be utilised by the enzyme for production of disaccharide **1** (β-D-mannopyranosyl-(1→3)-D-glucopyranose). Based on the notion that the amino acids constituting the active sites of GH94 and GH149 are well-conserved, it is highly likely that GH149 may be able to use the noncognate donor Man1P.

To further investigate whether GH149 phosphorylases can utilise Man1P as a donor for glycan synthesis, Pro_7066 and EgP1 enzymes were assayed in the presence of Man1P as a sugar donor and Glc as an acceptor, which showed that both enzymes could transfer mannose from Man1P onto Glc to produce disaccharide **1** (Figure 5.4). However, only a single transfer was detected and there was no further polymerisation of disaccharide **1**. This finding suggests that the sugar donor subsite in GH149 and its interaction with Man1P might be similar to that in GH94, which is in agreement with the conclusion from the structural studies of these enzymes. However, the lack of further extension of disaccharide **1** suggests that GH149 enzymes cannot tolerate the configuration change of 2-OH on the mannopyranosyl ring at the acceptor binding subsite, which presumably prohibits the binding and subsequent extension of disaccharide **1**. This finding could have useful applications in controlling the chain length of oligosaccharide product by using Man1P as a ‘stopper’ residue, whereby Man1P can be added to a prolonged glycan synthetic reaction catalysed by Pro_7066 to terminate the reaction and simultaneously modify the non-reducing end of the oligosaccharide product with Man residue.

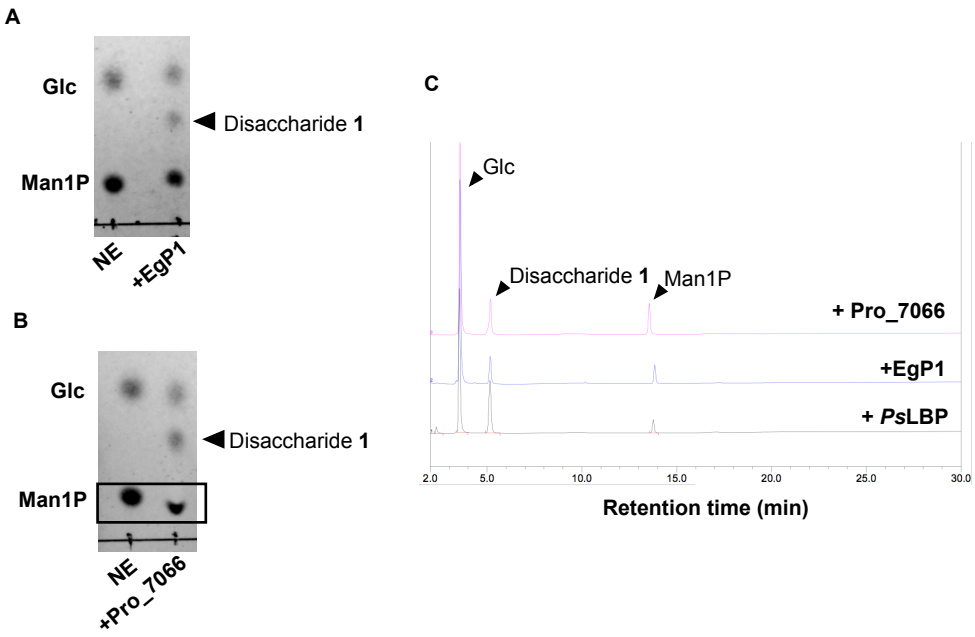


Figure 5.4. Activity of EgP1 and Pro_7066 on a noncognate donor Man1P. (**A**). TLC analysis of the reaction catalysed by EgP1. (**B**) TLC analysis of the reaction catalysed by Pro_7066. (**C**) HPAEC-PAD analysis of the reactions. The reaction carried out by *Ps*LBP was used as a positive control.

5.4. Chain length control strategies

An important challenge in oligosaccharide synthesis using GPs is the ability to control the DP of oligosaccharide products. This is particularly important for β-(1→3)-glucans because their physical/chemical/biological properties are influenced by their DP. For instance, successful interaction of β-(1→3)-glucan with its receptor Dectin-1[15] depended on the oligosaccharide chain length.[16] Therefore, designing optimal enzymatic reaction and/or modification of GP biocatalysts to enable regulation of the chain length of the oligosaccharide products are of interest.

Polymerisation of oligosaccharide by GPs results in products with different ranges of DP depending on the reaction times. Therefore, monitoring the reaction progress at different time points is crucial in order to determine when the product with a desired DP has been produced. Purification of the desired product can then be performed at the identified time point. However, this approach requires continuous motoring of the enzyme reaction, which is time-consuming and laborious. Therefore, chain length control strategies could be developed in order to overcome this barrier. Manipulation of sugar binding subsites in GPs to control oligosaccharide product chain length is one of the potential approach that could be used. This approach has been demonstrated in starch debranching enzymes, whereby R456 was mutated in order to control the degree of branching and the chain length of the glycan backbone.[17] Another prime example is the modification of the putative sugar binding sites in GH70 dextransucrase, which led to the reduction in the molar mass of dextran produced by the enzyme variants in comparison to the wild type enzyme.[18] In order to apply this methodology to GH149 and GHyyy, identification of the sugar binding subsites in GH149 enzymes through structural studies are required in order to predict the residues participating in the chain length control mechanism. Moreover, the presence of oligosaccharide surface binding site and two additional domains (Dom1 and 2) in Pro_7066 (GH149) may contribute to substrate/product chain length preference of the enzyme. Therefore, the precise roles of these non-catalytic sites need to be investigated to provide evidence for their contribution in chain length preference of the enzyme.

5.5. Protein engineering and directed evolution strategy for GP improvement

GPs are relatively well-known for their broad acceptor specificities, which enable their exploitation in synthesis of new carbohydrate structures and glycosylation of aglycone molecules. For example, the use of sucrose phosphorylase (SP) from *Leuconostoc mesenteroides* for production of D-glucosyl-ketohexoses from eight different ketohexose acceptors and Glc1P as a donor.[19] The acceptor promiscuity of the GPs, in particular of SP, can be exploited in this respect to glycosylate non-carbohydrate molecules, such as aromatics, sugar alcohols, phenolic compounds and carboxylic acids to improve the physical/chemical/biological properties of the molecules.[20,21] A prime example of the use of GPs to glycosylate aglycone molecule is an enzymatic production of 2-*O*-α-D-glucopyranosyl-*sn*-glycerol (αGG), an additive for cosmetic agents (Glycoin®), using SP from *Leuconostoc mesenteroides* as a biocatalyst and glycerol as an acceptor, with the final yield of 10.2 g in 90-ml reaction mixture.[22]

Despite the reported broad acceptor preferences of GPs, there is still scope for enzyme improvement by protein engineering and directed evolution. Site-directed mutagenesis has already been employed to tailor GP activity toward non-native substrates. This is evident for the engineering SP from *Bifidobacterium adolescentis* by exchanging Gln345 for Phe to improve SP activity towards glucosylation of resveratrol, (+)-catechin and (-)-epicatechin.[23] Glycosylation yields of up to 97% was achieved by the SP variant, which was a dramatic improvement with respect to those achievable with the wild type enzyme. Mutagenesis has also been applied to CBP to alter its donor and acceptor specificities.[24]

GP specificity towards sugar donors is relatively narrow, possibly due to the conservation of the sugar 1-phosphate binding subsite, which is evident from the structural analyses of GH94 and GH149 enzymes in this study. Nevertheless, relaxed specificity towards a noncognate donor has been demonstrated herein for *Ps*LBP. The relaxed donor specificity and structural evidence provide a good starting point for developing a protein engineering strategy to further improve specificity towards noncognate donors.

Directed evolution can be considered as an alternative or a complementary approach to rational mutagenesis for enzyme improvement. It requires a library containing diverse genes stemming from either mutation of a single parental gene, a set of homologues or combination thereof. The library is initially screened for improved enzyme activity. A set of selected genes with beneficial mutations from the first screen are then recombined and shuffled, then rescreened to identify the combination of beneficial mutations that give the optimal desired enzymatic activity.[25] Directed evolution experiments have been applied to glycosidases, glycosynthases and GTs,[26,27] but less so to GPs, with the only directed evolution experiment being reported of CBP from *Cellvibrio uda,* to create a variant with increased lactose phosphorylase activity.[28]

5.6. Strategies for high-throughput functional screening of new GP activities

Identification of two new families of the GH-like phosphorylases (GH149 and GHyyy) in chapter 3 and 4 provides several hundred potential biocatalysts for enzymatic synthesis of β-(1→3)-glucans. Two GH149 enzymes and a GHyyy enzyme have been biochemically characterised to provide evidence for sequence-activity relationships for these two families. However, there is an enormous number of remaining sequences that have not been characterised and thus their activities are unknown. This highlights a major drawback of the CAZy classification system, which categorises enzymes based on sequence similarity rather than their activities,[29] and often leads to the presence of enzymes with different activities being classified into the same family. This has been highlighted in the recent discoveries of two new GPs in GH3 family,[30,31] which previously known for GH activities only. Due to the vast number of uncharacterised sequence candidates in each CAZy family, development of fast and effective functional screening for GP activities is clearly required to pinpoint the activities of all sequences and to establish a catalogue of usable biocatalysts for oligosaccharide synthesis.

To facilitate characterisation and identification of new GP activities, high-throughput functional screening should be developed. A recently developed GP screen in a 96-well format provides a good starting platform.[31] Other existing high-throughput platforms can be adapted for the discovery of new GPs such as droplet microfluidics (Figure 5.5).[32] Such approach can be made possible by incorporating assays containing fluorogenic glycosides and orthophosphate as substrates for GP phosphorolysis in the droplet (Figure 5.6A). Alternatively, coupled enzyme systems can be used in the droplet to detect orthophosphate released by GP glycan synthetic activities (Figure 5.6B).

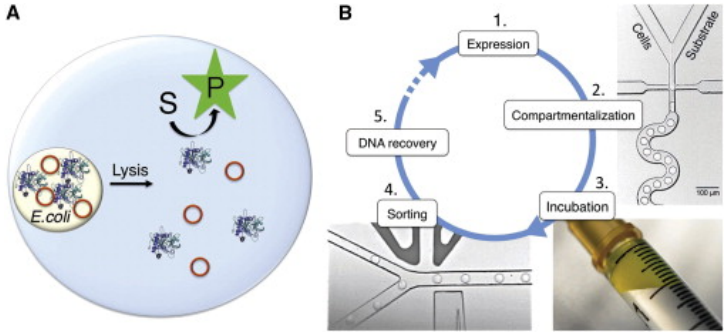


Figure 5.5. Schematic of ultrahigh-throughput droplet system for metagenomic libraries screening (Figure reproduced from Kintses *et al.*[32]). **(A)** Each droplet contains a single cell expressing a gene of interest. The cell is lysed, and the enzyme (blue) is assayed. Active enzymes convert a non-fluorescent substrate (S) into a fluorescent product (P). The enzyme-encoding plasmid (red) can be retrieved, amplified and sequenced (genotype-phenotype linkage). **(B)** Overall workflow. 1) Expression of enzymes *in vivo*; (2) compartmentalisation of single cells into droplets; (3) droplets assayed, generating fluorescent/chromogenic product; (4) droplets are sorted based on their fluorescence/absorbance; (5) plasmid DNA from selected library members are recovered.

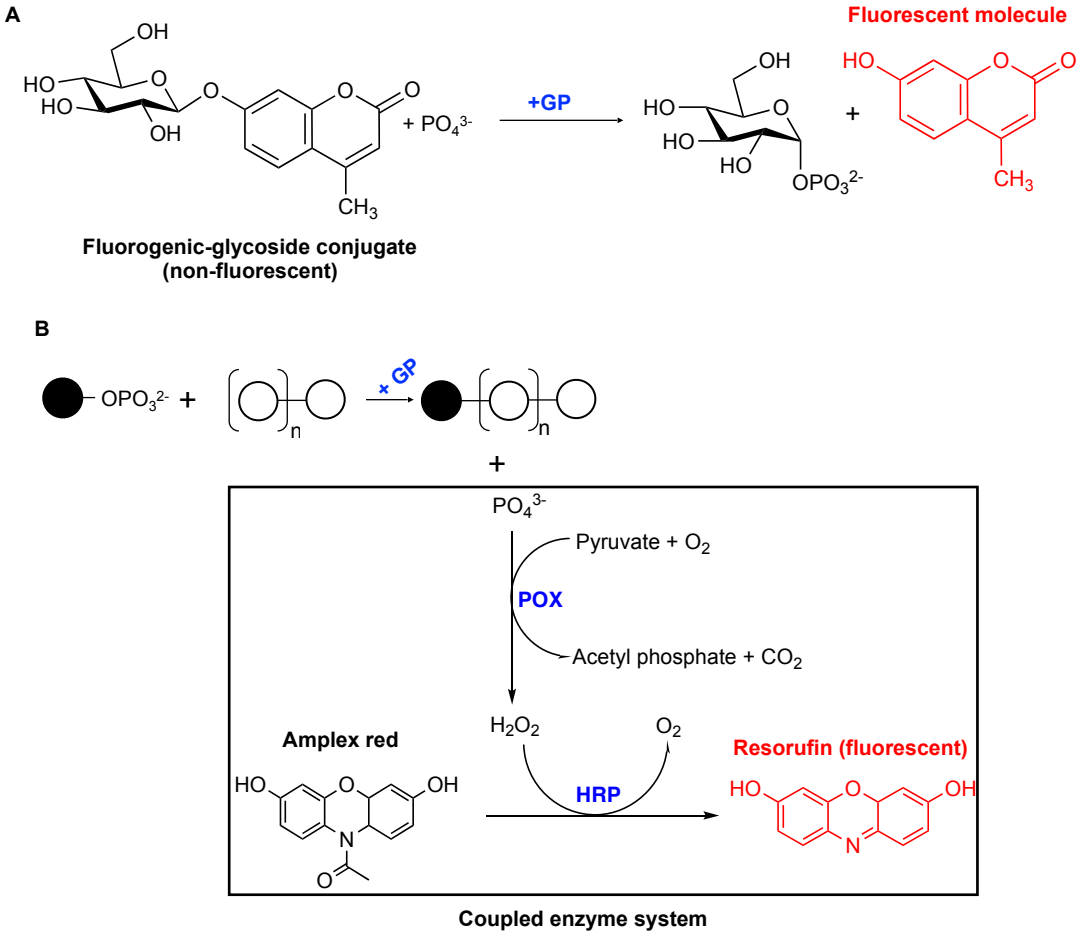


Figure 5.6. Assays for GP activity that can be cooperated into the droplet screening platform. **(A)** Phosphorolysis of fluorogenic-glycoside conjugate by GP, releasing a detectable fluorescent molecule. **(B)** Coupled enzyme system for orthophosphate detection. POX = pyruvate oxidase, HRP = horseradish peroxidase.

5.7. Conclusion

In summary, this work investigated the identification, characterisation and structural elucidation of new GPs acting on β-(1→3)-glucans from microalgae and bacteria and shed some light on the possible roles of these enzymes in bacteria through gene cluster analysis. Functional characterisation of these enzymes enabled subsequent proof-of-principle use of these enzymes as biocatalysts for β-(1→3)-glucan production through one-pot enzymatic synthesis, the approach of which can be further optimised for large scale synthesis of the oligosaccharides. Structural knowledge of the enzymes provided herein could enable future development of a protein engineering strategy, which can be complemented by directed evolution approaches, to improve the catalytic efficiency, broaden or further define enzyme specificity towards desirable substrates and products. Identification of new CAZymes have been vastly accelerated by database mining and CAZy classification, but experimental evidence for these prospective enzymes are lacking. Thus, development of high-throughput functional screens outlined in this work is a crucial step forward in an establishment of a catalogue of enzymes with known specificity and efficiency as biocatalyst for oligosaccharide production in the future.

5.8. References

[1] M. Kitaoka, Y. Matsuoka, K. Mori, M. Nishimoto, K. Hayashi, *Biosci. Biotechnol. Biochem.* **2012**, *76*, 343–348.

[2] T. Nihira, Y. Saito, M. Kitaoka, M. Nishimoto, K. Otsubo, H. Nakai, *Carbohydr. Res.* **2012**, *361*, 49–54.

[3] L. R. Maréchal, S. H. Goldemberg, *Biochem. Biophys. Res. Commun.* **1963**, *13*, 2–5.

[4] L. R. Marechal, *Biochem. Biophys. Acta* **1967**, *146*, 417–430.

[5] M. Kitaoka, T. Sasaki, H. Taniguchi, *Arch. Biochem. Biophys.* **1993**, *304*, 508–514.

[6] Y. Yamamoto, D. Kawashima, A. Hashizume, M. Hisamatsu, N. Isono, *Biosci. Biotechnol. Biochem.* **2013**, *77*, 1949–54.

[7] G. J. Albrecht, H. Kauss, *Phytochemistry* **1971**, *10*, 1293–1298.

[8] G. Pergolizzi, S. Kuhaudomlarp, R. A. Field, E. Kalita, *Protein Pept. Lett.* **2017**, *24*, 696–709.

[9] M. Nishimoto, M. Kitaoka, *Biosci. Biotechnol. Biochem.* **2007**, *71*, 2101–2104.

[10] M. Nishimoto, M. Kitaoka, *Carbohydr. Res.* **2009**, *344*, 2573–2576.

[11] A. A. Homaei, R. Sariri, F. Vianello, R. Stevanato, *J. Chem. Biol.* **2013**, *6*, 185–205.

[12] C. Müller, T. Ortmann, A. Abi, D. Hartig, S. Scholl, H.-J. Jördening, *Appl. Biochem. Biotechnol.* **2017**, *182*, 197–215.

[13] J. G. Voet, R. H. Abeles, *J. Biol. Chem.* **1970**, *245*, 1020–1031.

[14] E. C. O'Neill, G. Pergolizzi, C. E. M. Stevenson, D. M. Lawson, S. A. Nepogodiev, R. A. Field, *Carbohydr. Res.* **2017**, *451*, 118–132.

[15] A. Mueller, J. Raptis, P. J. Rice, J. H. Kalbfleisch, R. D. Stout, H. E. Ensley, W. Browder, D. L. Williams, *Glycobiology* **2000**, *10*, 339–346.

[16] G. D. Brown, S. Gordon, *Nature* **2001**, *413*, 36–37.

[17] C. Li, A. C. Wu, R. M. Go, J. Malouf, M. S. Turner, A. K. Malde, A. E. Mark, R. G. Gilbert, *PLoS One* **2015**, *10*, 1–21.

[18] M. Claverie, G. Cioci, M. Vuillemin, N. Monties, P. Roblin, G. Lippens, M. Remaud-Simeon, C. Moulis, *ACS Catal.* **2017**, *7*, 7106–7119.

[19] K. Morimoto, A. Yoshihara, T. Furumoto, G. Takata, *J. Biosci. Bioeng.* **2015**, *119*, 652–656.

[20] T. Desmet, W. Soetaert, P. Bojarová, V. Křen, L. Dijkhuizen, V. Eastwick-Field, A. Schiller, *Chem. - A Eur. J.* **2012**, *18*, 10786–10801.

[21] D. Aerts, T. F. Verhaeghe, B. I. Roman, C. V. Stevens, T. Desmet, W. Soetaert, *Carbohydr. Res.* **2011**, *346*, 1860–1867.

[22] C. Goedl, T. Sawangwan, M. Mueller, A. Schwarz, B. Nidetzky, *Angew. Chemie - Int. Ed.* **2008**, *47*, 10086–10089.

[23] M. Kraus, C. Grimm, J. Seibel, *ChemBioChem* **2016**, *17*, 33–36.

[24] M. R. M. De Groeve, T. Desmet, W. Soetaert, *J. Biotechnol.* **2011**, *156*, 253–260.

[25] F. H. Arnold, *Angew. Chemie Int. Ed.* **2018**, *57*, 4143–4148.

[26] R. Kittl, S. G. Withers, *Carbohydr. Res.* **2010**, *345*, 1272–1279.

[27] F. A. Shaikh, S. G. Withers, *Biochem. Cell Biol.* **2008**, *86*, 169–177.

[28] M. R. M. De Groeve, M. De Baere, L. Hoflack, T. Desmet, E. J. Vandamme, W. Soetaert, *Protein Eng. Des. Sel.* **2009**, *22*, 393–399.

[29] B. I. Cantarel, P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, B. Henrissat, *Nucleic Acids Res.* **2009**, *37*, 233–238.

[30] S. S. Macdonald, M. Blaukopf, S. G. Withers, *J. Biol. Chem.* **2015**, *290*, 4887–4895.

[31] S. S. Macdonald, A. Patel, V. L. C. Larmour, C. Morgan-Lang, S. J. Hallam, B. L. Mark, S. G. Withers, *J. Biol. Chem.* **2018**, *293*, 3451–3467.

[32] B. Kintses, C. Hein, M. F. Mohamed, M. Fischlechner, F. Courtois, C. Lainé, F. Hollfelder, *Chem. Biol.* **2012**, *19*, 1001–1009.

Chapter 6

Materials and methods

6.1. Euglena culture

E. gracilis was grown in two litres of EG:JM media (replacing peptone and tryptone with casein hydrolysate (Sigma-Aldrich)) supplemented with 1% glucose for seven days at 30°C with gentle agitation (150 rpm) in the dark.

6.2. Anion exchange chromatography (AIEX)

Approximately 20 g wet weight of 7-day old Euglena cells was harvested by centrifugation (500 x g, 5 minutes). The cell pellet was washed twice with MilliQ water and re-suspended in a lysis buffer (50 ml, 20 mM HEPES pH 7.0, 1 mg ribonuclease A (Sigma), 1 tablet complete protease inhibitor cocktail (Roche)). The cells were lysed on ice by sonication, followed by centrifugation (32,914 x g, 30 minutes) to remove the cell debris. The supernatant was collected and filtered through 0.2 µm disc filter (Millipore). A 5 ml AIEX column (Hitrap Sepharose Q, GE Healthcare) was pre-equilibrated with AIEX buffer (25 mM Tris-HCl, pH 8.5). The supernatant containing Euglena proteins (5 ml) was mixed with an equal volume of the AIEX buffer, then loaded onto the pre-equilibrated AIEX column. The proteins were eluted with NaCl gradient (0-700 mM) in the AIEX buffer over 20 column volumes and collected as 2 ml fractions.

6.3. DNJ affinity column preparation (performed by Dr Martin Rejzek, John Innes Centre)

The protocol was modified from;[1] *N*-5-carboxypentyl-DNJ (50 µmol, Toronto Research Chemicals) was conjugated to resin (1 ml, Bio-Rad Affidex 102 resin) overnight in anhydrous methanol using 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDAC, 10 mg, Bio-Rad) as a catalyst. The reaction was performed at room temperature with gentle agitation. The conjugated resin was poured into a glass column and washed with 10 x 4 ml of water to remove methanol. The column was stored at 4°C in aqueous 0.02 % sodium azide solution before use.

6.4. DNJ affinity chromatography

The DNJ column was pre-equilibrated with 4 x 4 ml water and 4 x 2 ml washing buffer (20 mM HEPES, pH 7.0). Protein solution (800 µl) was loaded onto the column. Unbound proteins were washed with 2 ml of the washing buffer and 2 ml of 100 mM

NaCl in the washing buffer. Bound proteins were eluted with an increasing concentration of DNJ (50 nM, 1 µM, 50 µM and 1 mM) in the washing buffer (2 ml per each concentration) and collected as 2 ml fractions. The protein fractions were concentrated by freeze-drying. The concentrated fractions were pooled and analysed by LC-MS/MS.

6.5. Proteomic analysis (performed by Dr Gerhard Saalbach, John Innes Centre)

Proteins were purified from the concentrated fraction in 6.4 using OMIX C4 tips (Agilent, Cheadle, UK). The amount of protein was determined using the Direct Detect® Spectrometer (Merck Millipore, Watford, UK). Protein mixture (8 µg) was dissolved in 30 µl of 0.1 M TEAB buffer (Sigma Aldrich, Dorset, UK) and 0.2% Rapigest (Waters, Manchester, UK), reduced with DTT and alkylated with iodoacetamide, and digested with 0.4 µg Sequencing Grade Trypsin (Promega, Southampton, UK) at 37 °C for 16 hours. The reaction was stopped by adding TFA, and aliquots were analysed by nanoLC-MS/MS on an Orbitrap Fusion™ Tribrid™ Mass Spectrometer coupled to an UltiMate® 3000 RSLCnano LC system (Thermo Scientific, Hemel Hempstead, UK). The mixtures were separated on a PepMap™ 100 C18 LC Column (C18, 2 µm, 500 x 0.75 mm, Thermo) using a gradient of 0.75% min^{-1} acetonitrile from 6% to 40% in water/0.1% formic acid at a flow rate of 0.3 µl min^{-1} and infused directly into the mass spectrometer. The mass spectrometer was run in positive ion mode, with no quadruple isolation, at 120 K resolution over the mass range 350-1800 (m/z) for the precursor scans (orbitrap). One microscan of 50 ms with an AGC target of $2e^5$ was used. MS2 threshold was set to $1.5e^4$ and precursors fragmented by both CID and HCD with CE=30 and an isolation window of 1.6 Da (quadrupole) using the automatic maximum speed option with ion injection for all available parallelizable time. Dynamic exclusion was set to 1 count and 30 s. Recalibrated peak lists were generated using MaxQuant 1.5.2.8 (www.MaxQuant.org) and the database search was performed with the merged HCD and CID peak lists using Mascot 2.4 (Matrixscience, London, UK). The search was performed on a *Euglena gracilis* protein database[2] with a precursor tolerance of 6 ppm and a fragment tolerance of 0.6 Da. The enzyme was set to trypsin/P with a maximum of 2 allowed missed cleavages. Carbamidomethyl (C) was set as fixed modification, and oxidation (M) and acetylation (Protein N-terminus) were used as variable modifications. The Mascot search results were imported into Scaffold 4.4.1.1 (www.proteomsoftware.com) using identification probabilities of 99% and 95% for proteins and peptides.

6.6. Generation of *E. coli* competent cells for transformation by heat shock

E. coli cells were grown in fresh lysogeny broth (LB) medium (100 ml) at 37°C with shaking (220 rpm) until OD_{600} of 0.4-0.5 was reached. The culture was chilled on ice (10 min) and the cells were collected by centrifugation (4000 rpm, 10 min, 4°C). The cell pellet was resuspended in ice cold $MgCl_2$ solution (100 mM, 50 ml) and left on ice (10 min). The cells were pelleted by centrifugation as previously described, then resuspended in ice cold $CaCl_2$ solution (100 mM, 50 ml) and left on ice (10 min). The cells were pelleted and resuspended in 2.5 ml of 100 mM $CaCl_2$, 20% glycerol solution and aliquoted into pre-chilled Eppendorf tubes (50 μl), flashed frozen in liquid nitrogen and stored at -80°C.

6.7. General cloning protocol

A pOPINF expression vector which contain a cleavable His_6-tag was cut by *Hin*dIII (15U) and *Kpn*I (10U). The fragments were separated by gel electrophoresis (1% agarose, 120 V, 40 minutes) and the gel section containing the linearised vector was cut and purified by gel extraction (gel extraction kit, QIAGEN). Ligation of the PCR product and the vector was performed by In-Fusion™ (TakaraBio, Mountain View, CA, USA) following the manufacturer's protocol using In-fusion HD enzyme premix (Clontech Laboratories). The recombinant vector was transformed into Stella Competent Cells (Clontech Laboratories) and plated on LB agar containing 100 μg/ml carbenicillin for transformant selection. The transformant was then inoculated into 5 ml LB containing 100 μg/ml carbenicillin overnight for plasmid preparation (QIAprep Miniprep kit, QIAGEN). The plasmid was then sent for sequencing (GATC biotech) to confirm the presence of the insert and the correct sequence.

6.8. Recombinant protein production in *E. coli*

6.8.1. EgP1

The *EgP1* DNA sequence was synthesised and optimised for *E. coli* expression (custom DNA synthesis by Gen9, Inc.). The sequence was amplified by PCR and ligated into the pOPINF plasmid vector[3] as previously described in general cloning protocol. The recombinant *pOPINF-EgP1* plasmid was transformed into chemically competent *E. coli* (Lemmo21) cells and a 1 l culture of the transformant was grown at 37°C in LB media with agitation (180 rpm) until OD_{600} reached 0.7. Heterologous protein expression was induced by adding IPTG to a final concentration of 0.2 mM and

incubated for 2 days at 18°C. For the soluble EgP1 protein, the DNA encoding predicted N-terminal signal peptide was truncated during PCR amplification, whilst insoluble EgP1 was produced from the full-length DNA.

6.8.2. Pro_7066

A recombinant plasmid pET28a containing *Pro_7066* DNA (Prozomix Limited) was transformed into chemically competent *E. coli* (BL21 (DE3)) cells. The transformant was selected on LB agar containing 50 μg/ml kanamycin and grown in 1 l Terrific broth (TB) containing the same antibiotic at 22°C overnight. The expression was induced with the addition of IPTG to the final concentration of 0.2 mM. The culture was further incubated overnight at 18°C.

6.8.3. *Ps*LBP

The *PsLBP* DNA sequence (GenBank accession number; AB568298.2) was synthesised and optimised for *E. coli* expression (custom DNA synthesis by Gen9, Inc.). The sequence was amplified by PCR and cloned into the pOPINF plasmid vector[3] as previously described in general cloning protocol. The recombinant *pOPINF-PsLBP* was transformed into chemically competent *E. coli* (BL21(DE3) cells by heat shock and a 1 l culture of the transformant was grown at 37°C in LB media containing 100 μg/ml carbenicillin with agitation (180 rpm) until OD_{600} ~ 0.7. The expression of *Ps*LBP was induced with the addition of IPTG to the final concentration of 1 mM. The culture was incubated further at 18°C for 14 hr.

6.8.4. *Pap*P

PapP gene sequence was amplified from *P. polymyxa* ATCC 842 genome. The sequence was ligated into pOPINF plasmid vector[3] as previously described. The recombinant *pOPINF-PapP* was transformed into Stella competent cells for plasmid propagation. The recombinant plasmid was transformed into chemically competent *E. coli* (Rosetta PLysS) cells and a 1 l culture of the transformant was grown at 18°C in LB media containing 100 μg/ml carbenicillin with agitation (180 rpm) overnight. Heterologous protein expression was induced by adding IPTG to the final concentration of 0.2 mM and incubated for 1 day at 18°C.

6.9. Recombinant protein purification

The *E. coli* cells expressing the recombinant proteins were harvested (6,721 x g, 10 min) and lysed by sonication in buffer A (10 mM HEPES pH 7.5, 250 mM NaCl) supplemented with 1 mg/ml DNase (Sigma) and 1 tablet complete protease inhibitor cocktail (Roche). Supernatant containing the recombinant proteins was separated from cell debris by centrifugation (32,914 x g, 30 min). Proteins were purified with ÄKTA pure FPLC system (GE Healthcare) at 4°C. The supernatant containing His_6-tagged recombinant protein was loaded to a 1-ml HisTrap™ HP column (GE Healthcare) pre-equilibrated with buffer A. The column was washed with buffer A and bound proteins were eluted in one step with buffer B (10 mM HEPES pH 7.5, 250 mM NaCl, 500 mM imidazole). The proteins were further purified by gel filtration using a Superdex S200 16/600 column (GE Healthcare), eluted with 20 mM HEPES pH 7.5, 150 mM NaCl, 1 ml/min. Fractions containing the proteins were pooled and concentrated to 10 mg/ml using Amicon Ultra-15 30 kDa MW cut off concentrator. The proteins were stored in 30 μl aliquots at -80°C until required.

6.10. Determination of protein molecular weights by gel filtration analysis

Protein standards (Gel Filtration Calibration Kit HMW, GE Healthcare) were diluted in gel filtration buffer (20 mM HEPES pH 7.5, 150 mM NaCl) to the final concentration of 2 mg/ml and individually injected into a Superdex S200 16/600 column (GE Healthcare) and eluted with the gel filtration buffer (1 ml/min) to determine their elution volumes. The elution volumes (V_e) are used to calculate K_{av} of the individual protein standards, which is defined as:

$$K_{av} = (V_e - V_0)/ (V_c - V_0) \quad (1)$$

where V_0 = column void volume (48.1 ml), V_e = elution volume of the protein, and V_c = geometric column volume (120 ml).

A calibration curve was prepared by plotting log molecular weight versus K_{av} values in Excel to obtain a linear regression ($R^2 = 0.9601$) with an equation:

$$y = -2.9556x + 6.0687 \quad (2)$$

where $x = K_{av}$ and y = log molecular weight.

K_{av} values of the proteins under investigation (*Ps*LBP, Pro_7066, EgP1 and *Pap*P) were calculated from (1) using V_e obtained during gel filtration purification in 6.9. K_{av} values of these proteins were then used to calculate molecular weights from (2).

6.11. SDS-PAGE analysis

The analysis was performed on a 12% precast SDS-PAGE gel cassette (RunBlue™, Expedeon). The AIEX fractions (10 µl) was mixed with an equal volume of 4X Laemmli sample buffer (Bio-Rad) containing β-mercaptoethanol. The mixture was boiled for 20 minutes to denature the proteins. 20 µl of the mixture were loaded onto the gel and run at 180 V for 60 minutes.

6.12. Enzyme assays

6.12.1. The glycan synthetic assays

6.12.1.1. Assay on GH149 enzymes

To screen for phosphorylase activity in Euglena AIEX fractions from section 6.2, phosphate release assay[4] was carried out in 20 µl of an assay buffer (100 mM HEPES pH 7.0, 20 mM α-glucose 1-phosphate (Glc1P), 10 mM acceptor (glucose (Glc), laminaribiose (G2), or laminaritriose (G3)) and 200 mM sodium molybdate) (all concentrations are final concentrations). The Euglena AIEX fractions (10 µl) were incubated with the assay buffer for 2 hours at 30°C. The reaction was stopped by boiling (5 minutes) and left to cool to room temperature. A colour solution (90 µl, 13.6 mM sodium ascorbate in 0.1 M HCl) was added to the boiled reaction mixture and incubated for 30 minutes at room temperature to allow colour development. A stop solution (90 µl, 68 mM sodium citrate tribasic dihydrate in 2% acetic acid) was added to the mixture to stop the colour development. The absorbance of final solution was measured at 620 nm on a 96-well plate reader. The amount of phosphate released was calculated from the absorbance by comparing to a phosphate standard curve ranging between 0-10 mM. All assays were performed in triplicate. Kinetic parameters of EgP1 and Pro_7066 were determined using the phosphate release assay (20 µl) with the enzymes (25 µg/ml) in the presence of 0.2-10 mM of Glc, G2, G3, laminaritetraose (G4), laminaripentaose (G5), or laminarihexaose (G6) and 10 mM Glc1P (all concentrations are final concentrations). The amount of phosphate released from the assays was measured and the values were fitted on non-linear regression with Michaelis-Menten model using GraphPad Prism to determine V_{max} and K_M.

6.12.1.2. *Ps*LBP enzymatic assays

Phosphate release assay[4] was carried out in 20 µl of an assay buffer (100 mM HEPES pH 7.0, 10 mM Glc1P, 10 mM Glc, 200 mM sodium molybdate) (all concentrations are final concentrations) and 1µL of fractions from peak a, b or c. The reactions were incubated at 45°C for 30 minutes. A colour solution (90 µl, 13.6 mM sodium ascorbate in 0.1 M HCl) was added to the boiled reaction mixture and incubated for 30 minutes at room temperature to allow colour development. A stop solution (90 µl, 68 mM sodium citrate tribasic dihydrate in 2% acetic acid) was added to the mixture to stop the colour development. The absorbance of the final solution was measured at 620 nm on a 96-well plate reader. The amount of phosphate released was calculated from the absorbance by comparing to a phosphate standard curve ranging between 0-10 mM. All assays were performed in triplicate. Screening for *Ps*LBP activity on sugar 1-phosphate donors was performed in the assay buffer containing 10 mM of sugar 1-phosphates and 10 mM Glc in the presence of 8 µg of the recombinant *Ps*LBP. Kinetic parameters of *Ps*LBP were determined using the phosphate release assay (20 µl) with the enzymes (25 µg/ml) in the presence of 0.2-10 mM of Glc and 10 mM Glc1P or Man1P (all concentrations are final concentrations). The amount of phosphate released from the assays was measured and the values were fitted on non-linear regression with Michaelis-Menten model using GraphPad Prism to determine V_{max} and K_M.

6.12.2. The phosphorolysis assays

The assay was carried out in 20 µl of an assay buffer (20 mM oligosaccharides, 10 mM KH_2PO_4 in 20 mM HEPES pH 7.0, all concentrations are final concentrations) in the presence of 10 µl of AIEX Euglena protein fractions, which was desalted to remove any endogenous phosphate by passing through a PD-10 column (GE Healthcare), or 1 µl of recombinant protein (EgP1, Pro_7066, or *Pap*P, 10 mg/ml stock solution). The reaction mixture was incubated for 1 hour at 30°C. The reaction was stopped by boiling (5 minutes) and oligosaccharide products were analysed by TLC analysis or High-Performance Anion-Exchange Chromatography with Pulsed Amperometric Detection (HPAEC-PAD).

6.13. Oligosaccharide analyses

6.13.1. Thin layer chromatography (TLC)

Unless otherwise stated, TLC was performed by spotting 0.5 μl of the recovered reaction mixture onto a precoated slides of Silica Gel 60 F254 (Merck) (10 cm x 5 cm), then eluted using a mobile phase containing NH_4OH : H_2O : *iso*-propanol (3:1:4) in a sealed glass container for 2 hour to allow oligosaccharide separation. The plate was air-dried and stained with orcinol, which was prepared by adding concentrated sulfuric acid (20 ml) to ice cold solution of 3,5-dihydroxytoluene (360 mg) in ethanol (150 ml) and water (10 ml). The stained plate was then heated until oligosaccharide spots were visible.

6.13.2. HPAEC-PAD analyses

HPAEC-PAD analyses were performed by diluting the reaction mixtures in MilliQ water to a final volume of 500 µl and desalted by mixed bed ion exchange resin (Sigma). The desalted mixtures were filtered through a disposable PTFE 0.45 µm filter disc (Merck Millipore), and subjected to HPAEC-PAD analysis using a Dionex ICS3000 chromatography system equipped with PAD and controlled by Chromeleon® software. A PA100 CarboPac column (analytical: 4 x 250 mm, guard: 4 x 50 mm) was used for all analyses with a mobile phase composed of 100 mM sodium hydroxide (solution A), and 100 mM sodium hydroxide + 400 mM sodium acetate (solution B). The separation was achieved by gradient elution; 0-100% solution B) over 30 min, followed by 20 min of 100% B, then 10 min re-equilibration of the column with 100% solution A. The solutions were delivered to the column at the rate of 0.25 ml/min.

6.14. Specific experiments for disaccharide 1

6.14.1. Preparation of disaccharide 1

A large scale enzymatic reaction was carried out in 5 ml reaction containing 700 µg of the recombinant *Ps*LBP in 100 mM HEPES pH 7.0, 10 mM Glc and 20 mM Man1P, incubating for 15 hr.

6.14.2. Purification of disaccharide 1

The reaction mixture from 6.14.1 was diluted 2-fold by water then desalted by passing through 5-ml Bio-Scale™ Mini Macro-Prep® High Q Cartridge (Bio-Rad Laboratories) and eluted with water (10 ml). The eluent was freeze-dried, re-dissolved in water (2 ml) and purified by gel permeation chromatography on Toyopearl HW40S (1.6 x 80 sm) column. Disaccharide 1 was eluted with water at the flow rate of 0.5 ml/min and the fractions containing disaccharide **1** were combined and freeze-dried.

6.14.3. Optical rotation analysis

Disaccharide **1** was dissolved in water to the final concentration of 2.2 mg/ml. The specific rotation of disaccharide 1 was recorded on a polarimeter model 341 polarimeter (PerkinElmer) at 20°C, 589 nm.

6.14.4. NMR analysis (performed by Dr Sergey A. Nepogodiev, John Innes Centre)

^{1}H NMR spectra were recorded at 298 K on a Bruker at 800 MHz. ^{13}C NMR spectra were recorded at 298 K on a Bruker Avance III 400 spectrometer at 100 MHz. Chemical shifts (δ) are reported in parts per million (ppm) with respect to residual HOD signal in D2O (δ_H 4.79). Coupling constants (*J*) are reported in Hz. NMR signal assignments were made with the aid of COSY and HSQC experiments.

6.14.5. TLC/MS analysis

GPC-purified fractions containing disaccharide **1** were analysed by spotting 1 µl of 2 mg/ml in water onto a silica gel plate. The spot was analysed by TLC/MS (Plate Express™, Advion BioSciences, Ithaca, NY, USA), which subjects the compound to electrospray ionisation using spray voltage and sample delivery pressure of 3.5 kV and 3,000 psi for positive ion mode, with the flow rate of 0.3 ml/min. The sample was analysed at the capillary temperature of 250°C, collision energy and the scan time of 1799 ms.

6.14.6. Physical data for disaccharide 1 (NMR analyses were performed by Dr Sergey A. Nepogodiev and HRMS by Dr Gerhard Saalbach, John Innes Centre, Norwich, UK)

$[\alpha]_D^{20}$ +7° (*c* 0.2, H_2O); ^{1}H NMR (800 MHz, D_2O) δ 5.172 (d, *J*=3.7, 1H, H-1 α-Glc), 4.839 (d, *J*=1.1, 1H, H-1 β-Man), 4.823 (d, *J*=1.0, 1H, H-1 β-Man), 4.611 (d, *J*=8.1, 1H, H-1 β-Glc), 4.081 – 4.049 (m, 2H, H-2 β-Man), 3.897 – 3.851 (m, 2H, H-6 β-Man), 3.865 – 3.808 (m, 2H, H-3 α-Glc, H-6 β-Glc), 3.801 (dddd, *J*=10.0, 5.0, 2.3, 0.6, 1H, H-5 α-Glc), 3.772 (dd, *J*=12.3, 2.3, 1H, H-6 α-Glc), 3.718 (dd, *J*=12.3, 5.0, 1H, H-6' α-Glc), 3.698 – 3.642 (m, 4H, H-6 β-Man, H-3 β-Glc, H-6' β-Glc), 3.633 – 3.579 (m, 3H, H-3 Man, H-2 α-Glc), 3.527 (t, *J*=9.7, 1H, H-4 β-Man), 3.522 (t, *J*=9.7, 1H, H-4 β-Man), 3.491 – 3.454 (m, 2H, H-4 α-Glc, H-4 β-Glc), 3.429 (ddd, *J*=10.0, 5.5, 2.2, 1H, H-5 β-Glc), 3.374 – 3.334 (m, 2H, H-5 β-Man), 3.314 (dd, *J*=9.3, 8.1, 1H, H-2 β-Glc); ^{13}C NMR (101 MHz, D_2O) δ 100.56 and 100.51 C-1 Man), 95.7 (C-1 β-Glc), 92.1 (C-1 α-Glc), 84.6 (C-3 β-Glc), 82.1 (C-3 α-Glc), 76.35 and 76.31 (C-5 Man), 75.4 (C-5 β-Glc), 73.7 (C-2 β-Glc), 72.9 and 72.8 (C-3 Man), 71.1 (C-5 α-Glc), 70.9 (C-2 α-Glc, 70.5 (C-2 Man), 68.2 and 68.1 (C-4 α-Glc and C-4 β-Glc), 66.7 (C-4 Man), 61.0 (C-6 Man), 60.7, 60.53 (C-6 Glc); HRMS (ESI): calculated for $C_{12}H_{22}NaO_{11}^+$ *m/z* 365.1054, found *m/z* 365.1062.

6.15. Crystallographic methods (with help from Dr Clare E. M. Stevenson and Dr David M. Lawson, John Innes Centre)

6.15.1. *Ps*LBP

Crystallisation trials were set up for purified *Ps*LBP (~10 mg/ml in 20 mM HEPES pH 7.0, 150 mM NaCl) using a range of commercial crystallisation screens (Molecular Dimensions) in MRC2 96-well sitting-drop vapour diffusion crystallisation plates (Swissci) with a mixture of 0.3 µl well solution and 0.3 µl protein solution using an OryxNano robot (Douglas Instruments) (Figure 6.1A). After several rounds of optimisation (Figure 6.1B), the best crystals (Figure 6.1C) were obtained from drops containing 0.4 µl of protein and 0.2 µl of a crystallisation solution comprised of 18% (w/v) polyethylene glycol (PEG) 3350 (Molecular Dimensions), 0.1 M Tris-citrate buffer pH 6.0, 0.3 M ammonium sulfate. Crystals were cryo-protected with well solution containing 20% (v/v) ethylene glycol (EG) and flash-cooled in liquid nitrogen. For phasing, crystals were soaked for 30 minutes in a saturated solution of either mercury (II) chloride (100 mM) or potassium iodide (100 mM) made up in the cryoprotectant

solution; the ligand bound complexes were obtained by soaking crystals for 5 minutes in cryoprotectant containing 20 mM of the compound (Glc1P or Man1P).

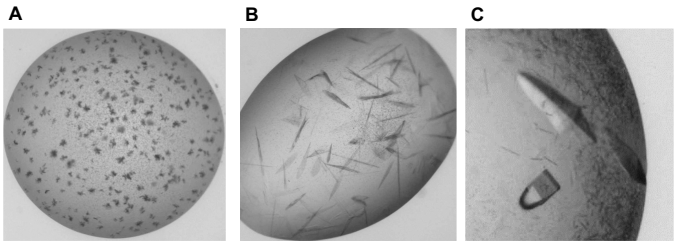


Figure 6.1. Crystallisation of *Ps*LBP. **(A)** Crystals obtained from commercial crystallisation screens. **(B)** Crystals obtained from optimisations. **(C)** The best crystals for data collection and processing.

The pre-cooled crystals were transferred robotically to the goniostat on either beamline I03 or I04 at Diamond Light Source (Oxfordshire, UK) and maintained at -173°C with a Cryojet cryocooler (Oxford Instruments). X-ray diffraction data were recorded using a Pilatus 6M hybrid photon counting detector (Dectris), then integrated and scaled using XDS[5] via the XIA2 expert system[6] and merged using AIMLESS.[7] All crystals belonged to space group $P4_12_12$ with approximate cell parameters of a = b = 147 Å, c = 222 Å.

Analysis of the likely composition of the asymmetric unit (ASU) suggested that it would contain two copies of the 102 kDa protein chain, giving an estimated solvent content of 58%. The structure was solved at 2.9 Å resolution by SAD phasing using the CRANK2 pipeline[8] by combining data collected from two mercury (II) chloride-soaked crystals at the L_{III} X-ray absorption edge of mercury (wavelength = 1.0052 Å). SHELXD[9] located eleven sites in the ASU with occupancies >0.25 and BUCCANEER[10] went on to build a model in which 59% of the sequence was fitted with R_{work} and R_{free} values of 0.342 and 0.400, respectively. This was then edited in COOT[11] before refining in REFMAC5[12] against native data processed to 1.95 Å resolution. Phases calculated from this model were used as input to second BUCCANEER job, which produced a model with 97% of the sequence fitted and R_{work} and R_{free} values of 0.282 and 0.323, respectively. The model was finalised by further iterations of manual rebuilding in COOT and restrained refinement in REFMAC5 using isotropic thermal parameters and TLS group definitions obtained from the TLSMD server (http://skuld.bmsc.washington.edu/~tlsmd/).[13] In each of the expected active sites, residual density consistent with an oxyanion was present. This was interpreted as sulfate derived from the precipitant solution. This sulfate-bound structure was used as the starting model for the Glc1P and Man1P complexes, which were built and refined as above.

The geometries of the final models were validated with MOLPROBITY[14] and wwPDB validation service (https://validate-rcsb-1.wwpdb.org/) before submission to the Protein Data Bank. Omit *mFobs-dFcalc* difference electron density maps were generated for the bound ligands using phases from the final model without the ligands after the application of small random shifts to the atomic coordinates, re-setting temperature factors, and re-refining to convergence. All structural figures were prepared using CCP4MG.[15]

6.15.2. Pro_7066

Crystallisation trials were set up for purified Pro_7066 (~15 mg/ml in 20 mM HEPES pH 7.0, 150 mM NaCl) on a BCS HT-96 screen (Molecular Dimensions) in MRC2 96-well sitting-drop vapour diffusion crystallisation plates (Swissci) with a mixture of 0.3 µl well solution and 0.3 µl protein solution in drop 1 and with 8 mM G6 in drop 2 using an OryxNano robot (Douglas Instruments). The best crystals hits were detected in the well containing 0.2 M Ammonium sulfate, 0.05 M magnesium sulfate heptahydrate, 0.1 M bicine (BCN) pH 9.0 and 20% (v/v) PEG Smear Medium (Molecular Dimensions) (Figure 6.2A). These conditions were then optimised (19.5-20.5% PEG Smear Medium (Molecular Dimensions), 0.15-0.25 M ammonium sulfate, 0.045-0.055 magnesium sulfate heptahydrate and 0.1 M BCN, pH 9.0) to obtain better quality crystals (Figure 6.2B). The best crystals obtained from the optimisation were cryoprotected with well solution containing 20% (v/v) EG and flash-cooled in liquid nitrogen. (Figure 6.2C). For phasing, crystals were soaked for 30 minutes in a saturated solution of mercury (II) chloride (100 mM) made up in the cryoprotectant solution; the ligand bound complexes were obtained by co-crystallisation of β-(1→3)-gluco oligosaccharides (8 mM of G6 or 30 mM of G2 from Megazyme) with the protein during the crystal optimisation.

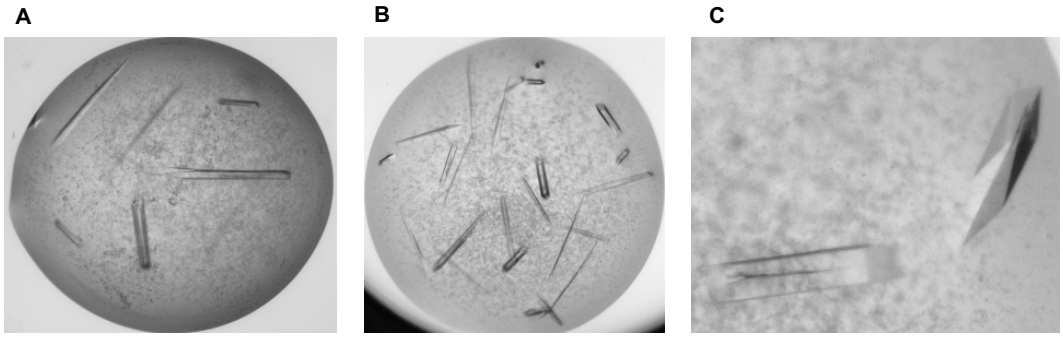


Figure 6.2. Crystal trials and optimisation for Pro_7066. (**A**) Initial hits from BCS screen. (**B**) Crystals after first optimisation. (**C**) Crystals for harvesting and data collection.

The pre-cooled crystals were transferred robotically to the goniostat on either beamline I03 or I04 at Diamond Light Source (Oxfordshire, UK) and maintained at -173°C with a Cryojet cryocooler (Oxford Instruments). X-ray diffraction data were recorded using a Pilatus 6M hybrid photon counting detector (Dectris), then integrated and scaled using XDS[5] via the XIA2 expert system[6] and merged using AIMLESS[7] All crystals belonged to space group $P2_12_12_1$ with approximate cell parameters of a = 100.0, b = 159.0, c = 181.6 Å.

Analysis of the likely composition of the asymmetric unit (ASU) suggested that it would contain two copies of the 130 kDa protein chain, giving an estimated solvent content of 55%. The structure was solved at 2.6 Å resolution by SAD phasing using the CRANK2 pipeline[8] with data collected from a mercury chloride-soaked crystal at the L_{III} X-ray absorption edge of mercury (wavelength = 1.0052 Å). SHELXD[9] located 17 sites in the ASU with occupancies >0.25 and BUCCANEER[10] went on to build a model in which 97% of the sequence was fitted with R_{work} and R_{free} values of 0.288 and 0.340, respectively. This was then edited in COOT[11] before refining in REFMAC5[12] against native data processed to 2.05 Å resolution. The model was finalised by further iterations of manual rebuilding in COOT and restrained refinement in REFMAC5 using isotropic thermal parameters and TLS group definitions obtained from the TLSMD server (http://skuld.bmsc.washington.edu/~tlsmd/).[13] In each of the expected active sites, residual density consistent with sulfate and BCN derived from the precipitant solution was present. This structure is referred to as the substrate-free complex and was used as the starting point for the other two structures described in chapter 3, which were refined using similar protocols. The geometries of the final models were validated as previously described in *Ps*LBP. *Omit mFobs-dFcalc* difference electron density maps (Figure 6.3) and figures were prepared as previously described for *Ps*LBP.

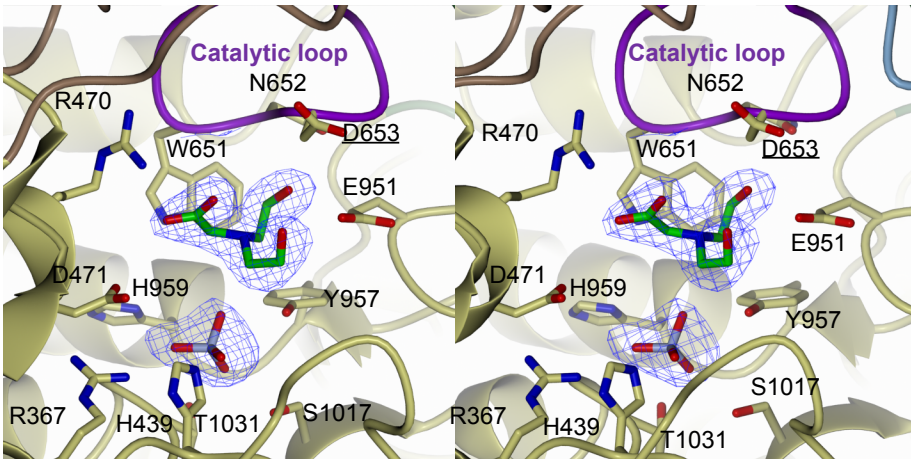


Figure 6.3 Stereo view of the Pro_7066 active site. Omit *mFobs-dFcalc* difference electron density maps (~2.0 Å resolution; contoured at ~3σ) are shown in blue mesh and were generated for the BCN and sulfate using phases calculated from the final model without the ligands after the application of small random shifts to the atomic coordinates, re-setting temperature factors, and re-refining to convergence.

6.15.3. EgP1

Initial crystallisation screens were performed using vapour diffusion method on the following screens; JCSG-plus (Molecular Dimensions, BCS HT-96 (Molecular Dimensions), Structure (Molecular Dimensions) and Kiss (Developed by Dr Clare Stevenson, John Innes Centre, Norwich, UK). A mixture of 0.3 µl of protein (~15 mg/ml in 20 mM HEPES pH 7, 150 mM NaCl) and 0.3 µl of the screen solution was aliquoted in MRC2 96-well sitting-drop vapour diffusion crystallisation plates (Swissci). There were no crystals formation in any of the screens, which prevented further optimisation to obtain crystals for data collection.

6.15.4. *Pap*P

Initial crystallisation screenings were performed using JCSG-plus and BCS HT-96 (Molecular Dimensions) using vapour diffusion method with a mixture of 0.3 µl of protein (~10 mg/ml stock solution in 20 mM HEPES pH 7.0, 150 mM NaCl) and 0.3 µl of the screen solution. Several crystals hits were detected in the screen solution. Further optimisations were performed in order to improve the crystal quality. The optimisation was set up using a solution mixture containing 0.15-0.25 M ammonium sulfate, 0.1 M Tris pH 8.0 and 10-20 % (v/v) PEG3350. The best crystals from optimisation were obtained from drops containing 0.2 M ammonium sulfate, 0.1 M Tris pH 8, 10% (v/v) PEG 3350 (Figure 6.4). The crystal was cryoprotected in the well solution supplemented with 20% (v/v) EG and sent for data collection. Diffraction was approximately 5 Å, which was too poor for solving structures.

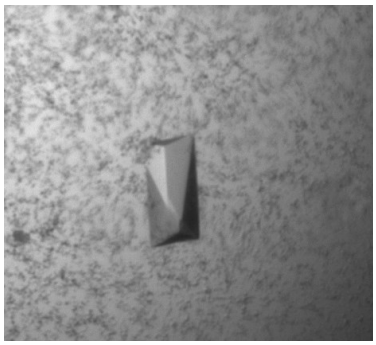


Figure 6.4. *Pap*P crystal after optimisation.

6.16. Bioinformatics analyses

6.16.1. GH149

Orthologous sequences to EgP1 were obtained from the non-redundant protein sequence database (https://www.ncbi.nlm.nih.gov/protein/), the Marine Microbial Eukaryote Transcriptome (MMET) Sequencing Project (http://marinemicroeukaryotes.org/), and the *E. gracilis Z* strain transcriptome (http://euglenadb.org/) using BLASTP or tBLASTn with E-value score of 0.0001 or more. Multiple sequence alignments of amino acid sequences were performed using Clustal Omega[16] (www.clustal.org/omega, version 1.2.2) with the default settings and edited with trimAl v.1.2 using a heuristic automated method.[17] The alignments were visualised by Jalview (version 14.6.4).[18] Phylogenetic trees were reconstructed from a matrix of 331 unambiguously aligned amino acids from 325 species using PhyML 3.0[19] with the best fit model as inferred by a smart model selection (SMS). Bootstrap values were determined from a population of 100 replicates. Tree annotation and visualisation were performed using iTOL v.3.4.3.[20] The GenBank IDs of the GH149 members can be found in Appendix 4.

6.16.2. GHyyy

Orthologous sequences to OcP1 were obtained from the non-redundant protein sequence database (https://www.ncbi.nlm.nih.gov/protein/) and the Marine Microbial Eukaryote Transcriptome (MMET) Sequencing Project (http://marinemicroeukaryotes.org/), using BLASTP or tBLASTn with E-value score of 0.0001 or more. Multiple sequence alignments of amino acid sequences were performed using Clustal Omega[16] (www.clustal.org/omega, version 1.2.2) with the default settings and edited with trimAl v.1.2 using a heuristic automated method.[17] The alignments were visualised by Jalview (version 14.6.4).[18] Phylogenetic trees were

reconstructed from a matrix of 306 unambiguously aligned amino acids using PhyML 3.0[19] with the best fit model as inferred by a smart model selection (SMS). Bootstrap values were determined from a population of 100 replicates. Tree annotation and visualisation were performed using iTOL v.3.4.3.[20] The GenBank IDs of the GHyyy members can be found in Appendix 7.

6.17. References

[1] I. N. Cruz, C. S. Barry, H. B. Kramer, C. C. Chuang, S. Lloyd, A. C. van der Spoel, F. M. Platt, M. Yang, B. G. Davis, *Chem. Sci.* **2013**, *4*, 3442–3446.

[2] E. C. O'Neill, M. Trick, L. Hill, M. Rejzek, R. G. Dusi, C. J. Hamilton, P. V. Zimba, B. Henrissat, R. A. Field, *Mol. BioSyst.* **2015**, *11*, 2808–2820.

[3] N. S. Berrow, D. Alderton, S. Sainsbury, J. Nettleship, R. Assenberg, N. Rahman, D. I. Stuart, R. J. Owens, *Nucleic Acids Res.* **2007**, *35*, e45.

[4] M. R. M. De Groeve, G. H. Tran, A. Van Hoorebeke, J. Stout, T. Desmet, S. N. Savvides, W. Soetaert, *Anal. Biochem.* **2010**, *401*, 162–167.

[5] W. Kabsch, *Acta Crystallogr. Sect. D* **2010**, *66*, 125–132.

[6] G. Winter, *J. Appl. Crystallogr.* **2010**, *43*, 186–190.

[7] P. R. Evans, G. N. Murshudov, *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2013**, *69*, 1204–1214.

[8] P. Skubák, N. S. Pannu, *Nat. Commun.* **2013**, *4*, 2777.

[9] G. M. Sheldrick, *Acta Crystallogr. Sect. A Found. Crystallogr.* **2007**, *64*, 112–122.

[10] K. Cowtan, *Acta Crystallogr. Sect. D* **2006**, *62*, 1002–1011.

[11] P. Emsley, K. Cowtan, *Acta Crystallogr. Sect. D* **2004**, *60*, 2126–2132.

[12] G. N. Murshudov, A. A. Vagin, E. J. Dodson, *Acta Crystallogr. Sect. D* **1997**, *53*, 240–255.

[13] J. Painter, E. A. Merritt, *J. Appl. Crystallogr.* **2006**, *39*, 109–111.

[14] I. W. Davis, A. Leaver-Fay, V. B. Chen, J. N. Block, G. J. Kapral, X. Wang, L. W. Murray, W. B. Arendall, J. Snoeyink, J. S. Richardson, D. C. Richardson, *Nucleic Acids Res.* **2007**, *35*, W375–W383.

[15] S. McNicholas, E. Potterton, K. S. Wilson, M. E. M. Noble, *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2011**, *67*, 386–394.

[16] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J. D. Thompson, D. G. Higgins, *Mol. Syst. Biol.* **2014**, *7*, 539–539.

[17] S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, *Bioinformatics* **2009**, *25*, 1972–1973.

[18] A. M. Waterhouse, J. B. Procter, D. M. A. Martin, M. Clamp, G. J. Barton, *Bioinformatics* **2009**, *25*, 1189–1191.

[19] S. Guindon, F. Lethiec, P. Duroux, O. Gascuel, *Nucleic Acids Res.* **2005**, *33*, 557–559.

[20] I. Letunic, P. Bork, *Bioinformatics* **2016**, *23*, 127–128.

Appendices

Appendix 1. Chemical shifts of β-mannopyranoside in 13C NMR spectra of methyl β-D-mannopyranoside[1] and disaccharide 1.

	C-1	C-2	C-3	C-4	C-5	C-6
Me β-Man	101.89	71.10	73.93	67.89	77.10	62.02
β-Man-(1→3)-Glc	100.6/100.5	70.5	72.9/72.8	66.7	76.3	61.0

Appendix 2. Simulation of 1H and 13C NMR spectra of disaccharide 1 using CASPER programme[2]

Residue	H-1	H-2	H-3	H-4	H-5	H-6a	H-6b
→3)α-D-Glc^{i}	5.26	3.67	3.94	3.54	3.85	3.76	3.84
→3)β-D-Glc	4.67	3.38	3.72	3.54	3.47	3.72	3.90
β-D-Man(1→	4.89	4.13	3.65	3.61	3.40	3.74	3.92

Residue	c-1	C-2	C-3	C-4	C-5	C-6
→3)α-D-Glc^{i}	92.81	71.98	82.59	69.27	71.98	61.85
→3)β-D-Glc	96.66	74.71	85.57	69.27	76.37	61.85
β-D-Man(1→	101.30	71.40	73.88	67.72	77.21	61.94

Appendix 3. List of Euglena proteins identified from affinity proteomics and their predicted functions

Euglena sequences identified from Mascot search	MW (kDa)	Exclusive unique peptide count	BLAST hits	% identity	E-value
m.79870_dark	11	3	zinc-binding alcohol dehydrogenase family protein [Hyalangium minutum]	62	1.00E-24
m.21700_dark	13	3	Amoebiasin-1, putative [Entamoeba invadens IP1]	38	2.00E-11
m.11948_dark	13	3	stress protein [Microtetraspora glauca]	42	6.00E-12
m.7066_dark	14	3	vacuolar H+ ATPase F subunit [Dictyostelium discoideum AX4]	63	1.00E-44
m.18114_dark	15	5	hypothetical protein BI308_08655 [Roseofilum reptotaenium AO1-A]	51	4.00E-35
m.11320_light	15	7	PREDICTED: reactive Intermediate Deaminase A, chloroplastic-like [Brassica napus]	61	4.00E-42
m.10229_light	15	7	glutathione S-transferase [Rhizobium anhuiense]	40	6.00E-18
m.12772_dark	15	3	DNA/RNA-binding protein, putative [Plasmodium falciparum 3D7]	33	4.00E-07
m.13968_dark	16	3	PREDICTED: succinate dehydrogenase assembly factor 4, mitochondrial [Gossypium raimondii]	49	3.00E-09
m.10231_light	16	3	glutaredoxin [Novosphingobium capsulatum]	40	2.00E-18
m.15369_dark	16	5	hypothetical protein GUITHDRAFT_151036 [Guillardia theta CCMP2712]	29	4.00E-10
m.11225_dark	16	7	glutathione S-transferase [Methylobacterium radiotolerans]	50	2.00E-19
m.37482_light	18	3	PREDICTED: oxalate--CoA ligase [Erythranthe guttata]	41	4.00E-17
m.286_dark	18	4	eukaryotic translation initiation factor 5A [Alternaria alternata]	61	9.00E-63
m.9214_light	19	9	predicted protein [Phaeodactylum tricornutum CCAP 1055/1]	49	1.00E-43
m.15794_light	19	4	PREDICTED: dual specificity protein phosphatase 5 [Callorhinchus milii]	37	2.00E-23
m.12763_dark	20	3	hypothetical protein EMIHUDRAFT_239514 [Emiliania huxleyi CCMP1516]	68	5.00E-79
m.21012_dark	20	3	UV excision repair protein Rad23 [Fistulina hepatica ATCC 64428]	30	8.00E-18
m.33004_light	21	5	hypothetical protein EMIHUDRAFT_445645 [Emiliania huxleyi CCMP1516]	53	5.00E-50
m.3416_dark	21	3	NADP-specific glutamate dehydrogenase [Acidovorax sp. RAC01]	81	6.00E-110
m.67685_light	22	3	hypothetical protein M438DRAFT_357250 [Aureobasidium pullulans EXF-150]	40	3.00E-04
m.80031_light	22	3	no hit		
m.17258_light	23	7	glycine cleavage system protein H [Hoeflea sp. IMCC20628]	38	5.00E-22
m.172_dark	23	14	FMN-dependent NADH-azoreductase [Pseudomonas linyingensis]	61	3.00E-89
m.32921_dark	23	3	PREDICTED: iron-sulfur cluster co-chaperone protein HscB, mitochondrial-like isoform X2 [Sinocyclocheilus anshuiensis]	30	3.00E-21
m.18100_dark	24	6	hypothetical protein EMIHUDRAFT_444876 [Emiliania huxleyi CCMP1516]	62	1.00E-88
m.331_dark	24	5	superoxide dismutase [Sedimenticola thiotaurini]	57	3.00E-71
m.32739_light	24	5	hypothetical protein A2167_06680 [Planctomycetes bacterium RBG_13_46_10]	36	2.00E-29
m.50981_dark	25	3	zinc-binding alcohol dehydrogenase family protein [Hyalangium minutum]	56	1.00E-66
m.19245_dark	25	4	SHMT, serine hydroxymethyltransferase [Blastocystis sp. ATCC 50177/Nand II]	64	1.00E-99
m.3618_dark	25	4	vacuolar ATP synthase subunit e [Culex quinquefasciatus]	38	1.00E-37

Euglena sequences identified from Mascot search	MW (kDa)	Exclusive unique peptide count	BLAST hits	% identity	E-value
m.3628_dark	25	3	type 1 glutamine amidotransferase domain-containing protein [Rhizobium sp. 2MFCol3.1]	53	8.00E-70
m.32237_dark	26	3	Rab family, other [Strigomonas culicis]	49	2.00E-71
m.21646_dark	26	3	triosephosphate isomerase [Trachelomonas volvocina]	86	7.00E-107
m.3446_light	26	14	predicted protein [Phaeodactylum tricornutum CCAP 1055/1]	52	2.00E-72
m.50359_dark	26	4	unnamed protein product [Bodo saltans]	32	2.00E-10
m.17475_dark	27	7	4-carboxymuconolactone decarboxylase [Fragilariopsis cylindrus CCMP1102]	38	9.00E-32
m.46464_dark	27	8	hypothetical protein MNEG_4609 [Monoraphidium neglectum]	63	2.00E-98
m.15106_dark	27	3	conserved unknown protein [Ectocarpus siliculosus]	38	1.00E-42
m.10828_dark	27	5	Adaptin earbinding coat-associated protein 2 (NECAP-2) isoform 2, putative [Acanthamoeba castellanii str. Neff]	49	2.00E-57
m.53400_dark (+1)	28	4	no hit		
m.49487_light	29	12	PREDICTED: ES1 protein homolog, mitochondrial isoform X2 [Strongylocentrotus purpuratus	56	8.00E-98
m.16269_dark (+1)	29	3	glutathione-s-transferase (ISS) [Ostreococcus tauri]	48	2.00E-65
m.86177_dark	30	3	hypothetical protein CONCODRAFT_19301 [Conidiobolus coronatus NRRL 28638]	34	1.00E-41
m.12512_dark (+1)	30	7	calcineurin B-like protein [Euglena gracilis]	49	4.00E-72
m.81_dark	30	3	Tankyrase [Symbiodinium microadriaticum]	42	4.00E-65
m.76395_light (+1)	30	3	hypothetical protein AUJ96_28535 [Armatimonadetes bacterium CG2_30_66_41]	49	4.00E-73
m.56797_light	31	7	succinate-semialdehyde dehydrogenase [Salinivirga cyanobacteriivorans]	59	1.00E-115
m.16140_dark	31	7	caffeoyl-CoA O-methyltransferase [Crinalium epipsammum]	45	4.00E-55
m.68978_dark (+1)	31	4	Hypothetical protein, putative [Bodo saltans	48	9.00E-79
m.125_dark (+1)	32	3	ribulose-bisphosphate carboxylase [Euglena gracilis]	99	0.00E+00
m.10171_light	32	4	methyltransferase [Candidatus Nitrospira inopinata]	48	6.00E-59
m.51716_light (+1)	32	3	hypothetical protein LOTGIDRAFT_236265 [Lottia gigantea]	54	2.00E-91
m.60001_dark (+3)	33	3	Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase [Pycnoporus coccineus BRFM310]	60	2.00E-109
m.28123_light (+1)	33	4	Short chain dehydrogenase [Labilithrix luteola]	53	3.00E-95
m.35384_light	33	5	NAD(P)-dependent oxidoreductase [Noviherbaspirillum humi]	47	1.00E-74
m.39212_dark (+1)	33	4	S-formylglutathione hydrolase [Marinobacter sp. T13-3]	59	6.00E-110
m.3232_dark (+1)	33	3	pyridoxal biosynthesis lyase PdxS [Chloroflexi bacterium OLB13]	72	2.00E-148
m.53885_dark (+1)	34	7	CBS domain containing protein [Acanthamoeba castellanii str. Neff]	28	7.00E-24
m.34687_light (+1)	34	3	3-hydroxyisobutyrate dehydrogenase [Advenella kashmirensis]	51	4.00E-89
m.31657_dark (+1)	34	3	hypothetical protein CBC74_03460 [Crocinitomicaceae bacterium TMED114]	46	3.00E-49
m.138_dark	34	4	RecName: Full=Elongation factor 1-alpha; Short=EF-1-alpha	99	0.00E+00
m.29683_light	34	14	quinone oxidoreductase [Dechloromonas aromatica]	70	5.00E-155

Euglena sequences identified from Mascot search	MW (kDa)	Exclusive unique peptide count	BLAST hits	% identity	E-value
m.15568_light	34	11	Methylmalonate-semialdehyde dehydrogenase [Phytophthora megakarya]	60	2.00E-141
m.244_dark	34	7	RGG repeats nuclear RNA binding protein A [Arachis duranensis	29	0.086
m.4018_light (+1)	34	3	hypothetical protein BSLG_05151 [Batrachochytrium salamandrivorans]	27	5.40E-01
m.6701_dark	35	3	alkene reductase [Pedosphaera parvula]	61	2.00E-131
m.7023_light	35	17	hypothetical protein AK812_SmicGene31809 [Symbiodinium microadriaticum]	50	1.00E-95
m.30042_dark (+2)	35	3	hypothetical protein PBRA_006275 [Plasmodiophora brassicae]	43	6.00E-68
m.11257_light (+1)	35	4	phosphoglycerate mutase [Reticulomyxa filosa]	60	7.00E-111
m.11743_dark	35	6	hypothetical protein SDRG_00935 [Saprolegnia diclina VS20]	33	7.00E-40
m.3271_dark (+2)	35	12	ornithine transcarbamylase 2 [Euglena gracilis]	100	0.00E+00
m.42147_dark (+1)	35	6	Haloacid dehalogenase-like hydrolase [Ectocarpus siliculosus]	53	3.00E-100
m.39672_dark (+3)	36	3	no hit		
m.31673_dark (+1)	36	3	conserved hypothetical protein, fragment [Trypanosoma vivax Y486]	47	1.00E-92
m.6706_light (+1)	37	5	hypothetical protein CICLE_v10028769mg [Citrus clementina]	45	6.00E-77
m.10767_dark (+1)	37	11	Adenosine kinase 2 [Ananas comosus]	56	4.00E-123
m.6726_dark (+3)	37	11	putative aldo-keto reductase 1 [Symbiodinium microadriaticum]	53	6.00E-78
m.13900_light (+1)	37	8	hypothetical protein DB31_7462 [Hyalangium minutum]	58	6.00E-122
m.15850_light	38	15	transaldolase [Rattus norvegicus]	66	2.00E-146
m.37271_dark (+3)	38	8	PREDICTED: N-acetylneuraminate lyase-like [Crassostrea gigas]	33	3.00E-44
m.32429_dark (+3)	38	3	hypothetical protein GUITHDRAFT_82106 [Guillardia theta CCMP2712]	43	2.00E-80
m.7641_dark	39	5	heterogeneous nuclear ribonucleoprotein A0-like [Oryzias latipes]	28	6.00E-34
m.11846_light (+5)	39	7	cysteine proteinase [Fragilariopsis cylindrus CCMP1102]	56	7.00E-132
m.20849_dark (+1)	39	7	GDP-fucose synthetase [Exiguobacterium undae]	62	2.00E-150
m.27481_light (+2)	39	5	lactyolglutathione lyase [Phaeodactylum tricornutum CCAP 1055/1]	62	3.00E-135
m.15732_dark (+1)	39	4	L-iditol 2-dehydrogenase [Fistulifera solaris]	58	1.00E-139
m.11264_dark (+4)	39	3	hypothetical protein AK812_SmicGene9704 [Symbiodinium microadriaticum]	40	4.00E-59
m.49_dark (+1)	39	13	glyceraldehyde-3-phosphate dehydrogenase [Euglena gracilis]	100	0.00E+00
m.20090_light (+5)	39	13	hypothetical protein GPECTOR_161g124 [Gonium pectorale]	72	0.00E+00
m.15808_dark (+1)	40	5	hypothetical protein SELMODRAFT_411711 [Selaginella moellendorffii]	48	3.00E-105
m.26124_dark	40	9	oxidoreductase [Euglena gracilis]	99	4.00E-112
m.11777_dark (+18)	41	3	methyltransferase type 11 [Candidatus Marithrix sp. Canyon 246]	64	1.00E-165
m.3497_light (+1)	41	12	fructose-1,6-bisphosphate aldolase [Euglena gracilis]	95	0.00E+00
m.372_dark (+9)	41	6	alanine dehydrogenase [Shewanella sp. ANA-3]	73	0.00E+00
m.219_dark	41	3	oxygen-evolving enhancer protein 1 precursor [Euglena gracilis]	100	0.00E+00
m.4324_dark	41	12	no hit		
m.75_dark (+2)	42	18	NAD-dependent alcohol dehydrogenase [Euglena gracilis]	99	0.00E+00

Euglena sequences identified from Mascot search	MW (kDa)	Exclusive unique peptide count	BLAST hits	% identity	E-value
m.18570_dark (+1)	42	6	PREDICTED: chaperone protein ClpB4, mitochondrial [Citrus sinensis]	52	2.00E-88
m.442_dark	42	14	ketol-acid reductoisomerase [Proteobacteria bacterium TMED154]	69	0.00E+00
m.25814_light	42	7	aspartate-semialdehyde dehydrogenase [Neiella marina]	66	4.00E-180
m.3284_dark (+1)	43	17	phosphoserine aminotransferase [Galdieria sulphuraria]	65	0.00E+00
m.13040_dark (+1)	43	11	putative 3-ketoacyl-CoA thiolase [Euglena gracilis]	99	0.00E+00
m.13600_dark (+4)	43	12	aldo/keto reductase [Euglena gracilis]	99	ae-172
m.6874_dark (+1)	43	4	pyruvate dehydrogenase E1 alpha subunit [Euglena gracilis]	100	0.00E+00
m.1629_dark	43	3	actin [Euglena gracilis]	99	0.00E+00
m.14951_dark (+3)	43	6	NAD-dependent epimerase/dehydratase family protein [Phycisphaera mikurensis]	73	0.00E+00
m.30422_light (+1)	43	5	predicted protein [Nematostella vectensis]	46	4.00E-102
m.435_dark (+1)	45	4	DEAD-box ATP-dependent RNA helicase [Klebsormidium nitens]	82	0.00E+00
m.20595_dark (+1)	45	5	acyl-CoA dehydrogenase [Hahella chejuensis]	59	7.00E-166
m.77563_light	45	4	hypothetical protein SELMODRAFT_120297 [Selaginella moellendorffii]	45	8.00E-86
m.19242_dark (+1)	46	7	hypothetical protein GUITHDRAFT_91943 [Guillardia theta CCMP2712]	74	0.00E+00
m.44288_dark (+1)	46	7	hypothetical protein A2Y12_15210 [Planctomycetes bacterium GWF2_42_9]	62	0.00E+00
m.27401_dark	46	11	phosphoribosylamine--glycine ligase [Kordiimonas lipolytica]	61	3.00E-176
m.37348_dark (+1)	47	15	ATP sulfurylase [Coccomyxa subellipsoidea C-169]	70	0.00E+00
m.3987_dark (+3)	47	5	succinate-CoA ligase [Polysphondylium pallidum PN500]	51	9.00E-140
m.65974_dark (+1)	47	10	histidinol dehydrogenase [Chitinophaga rupis]	63	0.00E+00
m.405_dark	47	7	cytosolic phosphoglycerate kinase [Euglena gracilis]	98	0.00E+00
m.1600_dark (+1)	47	6	D-3-phosphoglycerate dehydrogenase [Galdieria sulphuraria]	62	0.00E+00
m.198_dark	48	8	enolase [Euglena gracilis]	100	0.00E+00
m.2778_dark (+2)	49	24	PREDICTED: ornithine aminotransferase, mitochondrial-like [Lingula anatina]	68	0.00E+00
m.65835_light (+5)	49	3	no hit		
m.50174_light	50	4	succinate semialdehyde dehydrogenase [Phaeodactylum tricornutum CCAP 1055/1]	52	2.00E-164
m.86_dark	50	19	glutamate dehydrogenase [Burkholderiales bacterium RIFCSPLOWO2_12_FULL_65_40]	81	0.00E+00
m.13194_light (+1)	50	19	putative RAB GDP dissociation inhibitor alpha [Trypanosoma congolense IL3000]	59	0.00E+00
m.8374_dark	50	7	DUF1015 domain-containing protein [Malonomonas rubra]	55	1.00E-170
m.39260_dark (+6)	50	3	PREDICTED: polyadenylate-binding protein 4-like isoform X1 [Ciona intestinalis]	47	3.00E-103
m.10179_light (+1)	51	9	hypothetical protein PHYSODRAFT_554034 [Phytophthora sojae]	60	0.00E+00
m.216_dark (+7)	51	3	plastid glyceraldehyde 3-phosphate dehydrogenase [Euglena gracilis]	100	0.00E+00
m.12750_light (+1)	51	7	saccharopine dehydrogenase [Candidatus Thorarchaeota archaeon SMTZ1-45]	54	6.00E-159
m.12542_dark (+3)	52	13	dihydrolipoyl dehydrogenase [Euglena gracilis]	99	0.00E+00
m.23377_dark (+1)	52	3	putative Aspartyl aminopeptidase [Leptomonas seymouri]	52	2.00E-167

Euglena sequences identified from Mascot search	MW (kDa)	Exclusive unique peptide count	BLAST hits	% identity	E-value
m.4323_dark	52	6	cytosolic fructose-1,6-bisphosphatase [Euglena gracilis]	99	0.00E+00
m.7243_dark	52	11	glycine hydroxymethyltransferase [Strigomonas galati]	63	0.00E+00
m.11360_light (+1)	52	6	no hit		
m.3928_light	53	22	methionine S-adenosyl transferase [Euglena gracilis]	100	0.00E+00
m.11230_dark (+3)	53	10	orotate phosphoribosyltransferase and orotidine-5'-monophosphate decarboxylase [Euglena gracilis]	100	0.00E+00
m.4000_dark (+2)	54	5	UDP-glucose dehydrogenase [Capsaspora owczarzaki ATCC 30864]	65	0.00E+00
m.17873_dark (+1)	54	6	6-phosphogluconate dehydrogenase [Euglena gracilis]	99	0.00E+00
m.34150_dark (+2)	54	8	PREDICTED: ubiquitin carboxyl-terminal hydrolase 6 [Rhinolophus sinicus]	43	1.00E-122
m.182_dark	54	14	fructose-1,6-bisphosphate aldolase [Euglena gracilis]	80	0.00E+00
m.20400_dark (+1)	55	3	hypothetical protein DICPUDRAFT_8590 [Dictyostelium purpureum]	30	6.00E-26
m.11265_light (+3)	55	18	succinate-semialdehyde dehydrogenase [Armatimonadetes bacterium CG2_30_66_41]	59	0.00E+00
m.7166_dark	55	18	vacuolar ATP synthase subunit B [Trypanosoma rangeli SC58]	76	0.00E+00
m.12514_dark	56	3	predicted protein [Ostreococcus lucimarinus CCE9901]	60	0.00E+00
m.10288_dark (+1)	56	3	unnamed protein product [Vitrella brassicaformis CCMP3155]	65	0
m.4766_dark	56	4	PREDICTED: serine/threonine-protein phosphatase 5 [Neodiprion lecontei]	58	0
m.52788_light	57	14	aldehyde dehydrogenase [bacterium SM23_31]	64	0.00E+00
m.7611_dark (+1)	57	4	cytosolic NADP malic enzyme [Coccomyxa subellipsoidea C-169]	58	0
m.21334_dark (+1)	57	3	predicted protein [Ostreococcus lucimarinus CCE9901]	60	0
m.10263_dark	58	4	no hit		
m.10548_dark (+1)	58	19	predicted protein [Physcomitrella patens]	54	0
m.12633_dark (+1)	59	4	T-complex protein 1 subunit alpha [Batrachochytrium dendrobatidis JEL423]	62	0
m.36204_dark (+1)	59	15	imidazole glycerol phosphate synthase hisHF [Salpingoeca rosetta]	66	0
m.10160_light (+5)	61	3	3-oxoacyl-synth [Coccomyxa subellipsoidea C-169]	67	0
m.17815_dark (+1)	61	3	DUF2252 domain-containing protein [Catalinimonas alkaloidigena]	32	5.00E-46
m.322_dark	62	9	heat shock protein 60 [Euglena gracilis]	100	0
m.25777_dark (+1)	62	8	hypothetical protein GUITHDRAFT_80753 [Guillardia theta CCMP2712]	52	0
m.36998_light (+2)	62	4	ATP-dependent chaperone ClpB [cyanobacterium PCC 7702]	57	0
m.1638_dark (+2)	62	3	heat shock protein 60 [Euglena gracilis]	63	0
m.37941_dark (+1)	62	5	NAD(P)/FAD-dependent oxidoreductase [Aquabacterium sp. NJ1]	37	1.00E-67
m.26_dark (+1)	63	5	hypothetical protein GUITHDRAFT_166346 [Guillardia theta CCMP2712]	36	7.00E-91
m.24430_dark (+3)	63	3	glycine--tRNA ligase [Opitutales bacterium TMED158]	61	0
m.12652_dark (+1)	64	7	dihydroxy-acid dehydratase [Sphaeroforma arctica JP610]	76	0
m.12627_dark (+2)	64	4	fumarate hydratase [Nannochloropsis gaditana]	77	0
m.3567_dark	64	7	hsp70-Hsp90 organizing protein-like [Ananas comosus]	48	1.00E-168

Euglena sequences identified from Mascot search	MW (kDa)	Exclusive unique peptide count	BLAST hits	% identity	E-value
m.52047_dark (+1)	64	4	regulator of chromosomal condensation, putative [Bodo saltans]	28	3.00E-44
m.7034_dark	64	14	phosphoglucomutase [Saprolegnia diclina VS20]	60	0
m.19787_dark (+1)	65	6	unnamed protein product [Coffea canephora]	59	0
m.3234_dark (+6)	66	18	NAD-dependent malic enzyme [Thermus amyloliquefaciens]	55	0
m.7761_dark	66	6	glucose-6-phosphate isomerase [Clarkia lewisii]	65	0
m.22013_dark (+2)	67	3	coagulation factor 5/8 type domain-containing protein [Reticulomyxa filosa]	48	3.00E-166
m.21806_light	67	8	phosphoribosylaminoimidazolecarboxamide formyltransferase/IMP cyclohydrolase [Hesseltinella vesiculosa]	64	0
m.1598_dark (+1)	67	25	hypothetical protein TRIADDRAFT_35480 [Trichoplax adhaerens]	55	7.00E-170
m.11884_light (+1)	68	15	V-type ATPase, A subunit [Trypanosoma rangeli SC58]	70	0
m.3556_dark (+5)	69	6	polyadenylate-binding protein 4-like [Anoplophora glabripennis]	47	3.00E-175
m.30934_light	69	4	cytosolic glutamyl tRNA synthetase [Euglena gracilis]	99	0
m.71638_light (+1)	69	3	leukotriene A-4 hydrolase [Oryzias latipes]	45	3.00E-162
m.16531_dark (+2)	70	3	WD40 repeat-containing protein, putative [Bodo saltans]	66	0
m.12312_light (+1)	70	16	molecular chaperone DnaK [Pleomorphobacterium xiamenense]	57	0.00E+00
m.11231_light (+1)	71	3	PREDICTED: vigilin-like [Sinocyclocheilus rhinocerous	25	2.00E-09
m.3566_dark	71	5	putative kinesin [Leishmania major strain Friedlin]	32	2.00E-63
m.2723_dark (+1)	71	29	heat shock protein mitochondrial precursor [Perkinsela sp. CCAP 1560/4]	66	0
m.3148_dark	72	4	polyadenylate binding protein [Neocallimastix californiae]	49	0
m.23187_dark (+1)	72	5	aspartyl-tRNA synthetase [Trypanosoma theileri]	55	0
m.10084_light (+1)	72	4	hypothetical protein STCU_05414 [Strigomonas culicis]	47	2.00E-09
m.21935_dark (+1)	73	9	unnamed protein product [Phytomonas sp. isolate EM1]	63	0
m.120_dark (+1)	73	24	hsp70 [Perkinsela sp. CCAP 1560/4]	78	0
m.12767_light (+13)	73	27	transketolase [Euglena gracilis]	99	0
m.11240_dark (+1)	73	3	cysteinyl-tRNA synthetase [Polysphondylium pallidum PN500]	51	0
m.199_dark	74	27	acetyl CoA synthetase [Chlamydomonas reinhardtii]	60	0
m.168_dark	75	4	hypothetical protein DICPUDRAFT_147099 [Dictyostelium purpureum]	33	2.00E-32
m.14720_light	75	4	plastid sedoheptulose-1,7-bisphosphatase [Euglena clara]	85	0
m.57688_dark (+1)	75	3	mitochondrial intermediate peptidase [Perkinsela sp. CCAP 1560/4]	33	9.00E-98
m.3346_light (+1)	76	4	hypothetical protein FisN_14Lh325 [Fistulifera solaris]	27	6.00E-04
m.11898_light (+1)	78	4	prolyl-tRNA synthetase, putative [Bodo saltans]	55	0.00E+00
m.11199_light (+1)	79	5	Isopropylmalate dehydratase [Guillardia theta CCMP2712]	69	0.00E+00
m.3154_dark (+3)	79	7	methylmalonyl-CoA mutase [Euglena gracilis]	99	0
m.11283_dark	80	4	asparagine--tRNA ligase [Microcystis aeruginosa]	52	5.00E-161
m.10910_dark (+1)	80	16	malate synthase G [Acidovorax delafieldii]	78	0
m.10568_dark (+1)	82	19	glycyl-tRNA synthetase [Emiliania huxleyi CCMP1516]	53	0

Euglena sequences identified from Mascot search	MW (kDa)	Exclusive unique peptide count	BLAST hits	% identity	E-value
m.28527_light	83	34	isocitrate dehydrogenase (NADP(+)) [Paraburkholderia xenovorans]	71	0
m.21623_dark (+1)	86	5	WD40 repeat domain-containing protein [Planktothrix prolifica]	41	1.00E-33
m.37470_dark (+7)	86	6	selenoprotein O [Micromonas commoda]	46	2.00E-173
m.3583_light (+2)	86	3	kinesin [Trypanosoma cruzi strain CL Brener]	46	3.00E-91
m.17398_dark (+1)	87	20	hypothetical protein GUITHDRAFT_159401 [Guillardia theta CCMP2712]	50	0
m.3102_dark	90	4	plastid transketolase [Euglena gracilis]	99	0
m.8430_light	90	6	Hemicentin-1 [Symbiodinium microadriaticum]	31	1.00E-10
m.14502_dark	92	5	membrane-associated protein, putative [Bodo saltans]	26	9.00E-52
m.24362_dark	92	8	PREDICTED: glycerol kinase-like isoform X4 [Lingula anatina]	56	2.00E-161
m.14920_light	95	3	major vault protein, putative [Bodo saltans]	64	0.00E+00
m.15620_light	97	14	Puromycin-sensitive aminopeptidase [Symbiodinium microadriaticum]	52	0.00E+00
m.468_dark	98	23	bifunctional aconitate hydratase 2/2-methylisocitrate dehydratase [Gammaproteobacteria bacterium RIFCSPLOWO2_02_FULL_57_10]	76	0
m.31840_dark	99	13	argonaute [Naegleria gruberi]	30	3.00E-95
m.20583_dark	100	4	aconitase, putative [Bodo saltans]	63	0
m.12398_light	100	3	PREDICTED: cell division cycle and apoptosis regulator protein 1-like [Trachymyrmex septentrionalis]	28	1.00E-05
m.77593_light	101	3	predicted protein [Naegleria gruberi]	26	3.00E-76
m.12663_dark	102	3	PREDICTED: multiple C2 and transmembrane domain-containing protein 1-like [Sinocyclocheilus grahami]	28	5.00E-27
m.17228_light	104	3	serine/threonine protein kinase [Salpingoeca rosetta]	27	6.00E-20
m.3538_light	104	7	pyruvate, phosphate dikinase [Omnitrophica bacterium GWA2_50_21]	63	0.00E+00
m.65276_dark	106	3	Lysosomal alpha-mannosidase [Symbiodinium microadriaticum]	52	0
m.12864_dark	107	17	hypothetical protein [Paramecium tetraurelia strain d4-2]	26	1.00E-21
m.6710_dark	107	13	glycine dehydrogenase [Spizellomyces punctatus DAOM BR117]	59	0
m.15695_dark	114	8	uncharacterized protein LOC105157340 [Sesamum indicum]	35	2.00E-45
m.12386_dark	123	4	PREDICTED: isoleucine--tRNA ligase, cytoplasmic [Equus caballus]	55	0
m.1626_dark	127	8	hybrid cluster protein, putative [Bodo saltans]	42	0.00E+00
m.14570_dark	127	36	laminaridextrin phosphorylase-like protein [Euglena gracilis]	53	0
m.13041_light	128	14	Pyrophosphate--fructose 6-phosphate 1-phosphotransferase subunit beta 2 [Symbiodinium microadriaticum	49	0.00E+00
m.11124_light	129	21	PREDICTED: pyruvate carboxylase, mitochondrial-like [Limulus polyphemus]	60	0
m.14_light	130	50	malate synthase-isocitrate lyase [Euglena gracilis]	99	0.00E+00
m.20564_dark	141	16	protein synthetase, putative [Acanthamoeba castellanii str. Neff]	42	0
m.13551_light	162	3	hypothetical protein NIES22_38050 [Calothrix brevissima NIES-22]	26	3.00E-80
m.45901_dark	162	8	hypothetical protein SARC_00815 [Sphaeroforma arctica JP610]	50	0

Euglena sequences identified from Mascot search	MW (kDa)	Exclusive unique peptide count	BLAST hits	% identity	E-value
m.10688_dark	164	6	argonaute [Naegleria gruberi]	33	5.00E-110
m.17301_dark	168	3	hypothetical protein THRCLA_00008 [Thraustotheca clavata]		
m.2768_dark	170	3	hypothetical protein C366_01032 [Cryptococcus neoformans var. grubii Tu401-1]	46	7.00E-73
m.186_dark	201	30	pyruvate: NADP+ oxidoreductase [Euglena gracilis]	99	0
m.55348_light	223	5	hypothetical protein AUK03_00510 [Anaerolineae bacterium CG2_30_64_16]	37	0.00E+00
m.462_dark	226	7	glutamate synthase, NADH/NADPH, small subunit, variant [Spizellomyces punctatus DAOM BR117]	59	0

Appendix 4. Sequence IDs for GH149 members

App. 4.1. GH149 sequences from NCBI database

RefSeq	Species	Phylum	Genome/ Nucleotide accession	Isolation source
OIO83857.1	Anaerolineae bacterium CG2_30_57_67	Chloroflexi	MNXJ01000074.1	groundwater filtered through a 3.0 um filter (groundwater metagenome)
OGS36884.1	Elusimicrobia bacterium RIFOXYB2_FULL_49_7	Elusimicrobia	MGVN01000056.1	Rifle well CD01 at 16ft depth; 0.2 micron filter at time point B (subsurface metagenome)
AEJ20297.1	Treponema caldarium DSM 7334	Spirochaetes	CP002868.1	Fresh water, Hot spring
EYR63563.1	Actinotalea ferrariae CF5-4	Actinobacteria	AXCW01000085.1	iron mine of Hongshan
WP_013115824.1	Cellulomonas flavigena DSM 20109	Actinobacteria	NC_014151.1	soil
WP_013769870.1	Cellulomonas fimi ATCC 484,	Actinobacteria	NC_015514.1	soil
WP_013882364.1	Cellulomonas gilvus ATCC 13127	Actinobacteria	NC_015671.1	bovine feces
WP_019134858.1	Cellulomonas massiliensis JC225	Actinobacteria	NZ_HE978588.1	human stool sample
WP_034634502.1	Cellulomonas cellasea DSM 20118	Actinobacteria	NZ_AXNT01000160.1	no data
WP_052022748.1	Actinotalea ferrariae CF5-4	Actinobacteria	NZ_AXCW01000085.1	iron mine of Hongshan
WP_055916706.1	Cellulomonas sp Root485	Actinobacteria	NZ_LMFC01000002.1	Arabidopsis root
WP_056027623.1	Cellulomonas sp. Leaf395	Actinobacteria	NZ_LMQI01000005.1	Arabidopsis leaf
WP_056085790.1	Cellulomonas sp Root137	Actinobacteria	NZ_LMFG01000001.1	Arabidopsis root
WP_056589364.1	Cellulomonas sp Leaf334	Actinobacteria	NZ_LMOO01000005.1	Arabidopsis leaf
WP_057208477.1	Cellulomonas sp Root930	Actinobacteria	NZ_LMJI01000001.1	Arabidopsis root
AFR34657.1	Riemerella anatipestifer RA-CH-1	Bacteroidetes	CP003787.1	sick duck (vaccine strain)
CDF79585.1	Formosa agariphila KMM 3901	Bacteroidetes	HG315671.1	an isolate from the green algae Acrosiphonia sonderi. F. agariphila is a facultative anaerobe. Potential algal degrader
EON75098.1	Lunatimonas lonarensis AK24	Bacteroidetes	AQHR01000110.1	Lonar Lake, a halo-alkaline lake in Buldana District
GAK91234.1	Nonlabens ulvanivorans JCM 19297	Bacteroidetes	BBMJ01000009.1	NA
GAK94769.1	Nonlabens ulvanivorans JCM 19298	Bacteroidetes	BBMK01000007.1	NA
GAP42007.1	Lentimicrobium saccharophilum TBC1	Bacteroidetes	DF968182.1	waste water treatment system
KPP94675.1	Bacteroidetes bacterium HLUCCA01	Bacteroidetes	LIHN01000021.1	microbial mat in Hot Lake
KRO66514.1	Cryomorphaceae bacterium BACL21 MAG-121220-bin10	Bacteroidetes	LICC01000048.1	Brackish water

RefSeq	Species	Phylum	Genome/ Nucleotide accession	Isolation source
KRO88543.1	Cryomorphaceae bacterium BACL29 MAG-121220-bin8	Bacteroidetes	LICI01000110.1	brackish water collected from the surface of the Baltic Sea
KRP08637.1	Sphingobacteriales bacterium BACL12 MAG-120802-bin5	Bacteroidetes	LICM01000125.1	brackish water collected from the surface of the Baltic Sea
KRP13118.1	Sphingobacteriales bacterium BACL12 MAG-120813-bin55	Bacteroidetes	LICN01000063.1	brackish water collected from the surface of the Baltic Sea
KRP29741.1	Cryomorphaceae bacterium BACL22 MAG-120619-bin32	Bacteroidetes	LIDG01000033.1	brackish water collected from the surface of the Baltic Sea
KUO68202.1	Lutibacter sp BRH_c52	Bacteroidetes	LOEY01000026.1	Opalinus Clay rock porewater BRC-3 borehole
OFX41757.1	Bacteroidetes bacterium GWA2_40_14	Bacteroidetes	MENH01000108.1	Rifle well CD01 at time point 1 / A; 5m depth; 0.2 filter , subsurface metagenome
OFX90003.1	Bacteroidetes bacterium GWE2_32_14	Bacteroidetes	MENV01000003.1	Rifle well CD01 at time point 5 / E; 5m depth; 0.2 filter
OGS61795.1	Flavobacteria bacterium GWF1_32_7	Bacteroidetes	MGWI01000014.1	Rifle well CD01 at time point 6 / F; 5m depth; 0.1 filter
OGS73573.1	Flavobacteria bacterium RIFCSPLOWO2_12_FULL_35_11	Bacteroidetes	MGWK01000012.1	Rifle well FP-101 under low O2 conditions; 1.2 μm filter;
OIP46152.1	Flavobacteriaceae bacteri□m CG2_30_31_66	Bacteroidetes	MNYM01000037.1	groundwater filtered through a 3.0 μm filter
OIP52753.1	Flavobacteriaceae bacterium CG2_30_34_30	Bacteroidetes	MNYN01000001.1	groundwater filtered through a 3.0 μm filter
OIP85247.1	Porphyromonadaceae bacterium CG2_30_38_12	Bacteroidetes	MNZK01000004.1	groundwater filtered through a 3.0 μm filter
OIQ30310.1	Bacteroidetes bacterium MedPE-SWsnd-G2	Bacteroidetes	MPDE01000002.1	50m depth Mediterranean seawater enriched with a specific particulate matter
OIQ41396.1	Bacteroidetes bacterium MedPE-SWsnd-G1	Bacteroidetes	MPDD01000002.1	50m depth Mediterranean seawater enriched with a specific particulate matter
OJJ21699.1	marine bacterium AO1-C	Bacteroidetes	MLAY01000004.1	central great barrier reef, assoicated with Pavona duerdeni
SCY20814.1	Nonlabens sp Hel1_33_55	Bacteroidetes	LT627735.1	NA
SDC80925.1	Algoriphagus faecimaris DSM 23095	Bacteroidetes	FNAC01000006.1	NA
SDG38233.1	Psychroflexus sediminis DSM 19803	Bacteroidetes	FNCW01000001.1	NA
SDI57228.1	Winogradskyella thalassocola DSM 15363	Bacteroidetes	FNCZ01000013.1	NA
SDM08669.1	Chryseobacterium taihuense CGMCC 110941	Bacteroidetes	FNHD01000012.1	plant associated
SDR87359.1	Polaribacter sp KT25b	Bacteroidetes	LT629752.1	NA
SDS23837.1	Formosa sp Hel1_31_208	Bacteroidetes	LT629733.1	NA
SDU10865.1	Polaribacter sp Hel1_33_78	Bacteroidetes	LT629794.1	NA
SDX94361.1	Lutibacter oricola DSM 24956	Bacteroidetes	FNNJ01000012.1	NA

RefSeq	Species	Phylum	Genome/ Nucleotide accession	Isolation source
SEC77108.1	Tenacibaculum sp MAR_2009_124	Bacteroidetes	FNSF01000007.1	NA
SEF71509.1	Chryseobacterium humi DSM 21580	Bacteroidetes	FNUS01000001.1	NA
SEP83958.1	Hyunsoonleella jejuensis DSM 21035	Bacteroidetes	FOFN01000001.1	NA
SEQ41340.1	Lewinella agarilytica DSM 24740	Bacteroidetes	FOFB01000009.1	NA
SFG38750.1	Algoriphagus hitonicola DSM 19315	Bacteroidetes	FOPC01000003.1	NA
SFI14923.1	Chryseobacterium frigidisoli DSM 26000	Bacteroidetes	FOQT01000002.1	NA
SFN60596.1	Algoriphagus ornithinivorans DSM 15282	Bacteroidetes	FOVW01000001.1	NA
SFO59267.1	Prevotella sp tf2-5	Bacteroidetes	FOWK01000003.1	no data
SFS48762.1	Lutibacter maritimus DSM 24450	Bacteroidetes	FOZP01000003.1	NA
SHN04610.1	Polaribacter sp KT 15	Bacteroidetes	LT670850.1	NA
SIT04910.1	Belliella pelovolcani DSM 46698	Bacteroidetes	FTOP01000013.1	NA
WP_003437557.1	Psychroflexus gondwanensis	Bacteroidetes	NZ_APLF01000004.1	organic, hypersaline lake
WP_004571105.1	Polaribacter irgensii	Bacteroidetes	NZ_CH724148.1	surface water from the Penola Strait, Antarctica
WP_007093907.1	Kordia algicida OT-1	Bacteroidetes	NZ_DS544873.1	type strain (algicidal)
WP_008253479.1	Flavobacteria bacterium BAL38	Bacteroidetes	NZ_AAXX01000001.1	marine
WP_008269464.1	Flavobacteriales bacterium ALC-1	Bacteroidetes	NZ_ABHI01000001.1	prevents the growth of a red tide dinoflagellate
WP_008629750.1	Mariniradius saccharolyticus AK6	Bacteroidetes	NZ_AMZY02000016.1	seawater of a fish culture pond
WP_009183469.1	Cecembia lonarensis LW9	Bacteroidetes	NZ_AMGM01000004.1	surface water (4.5 m depth)
WP_013306491.1	Maribacter sp HTCC2170	Bacteroidetes	NC_014472.1	collected from the coastal area of Newport, Oregon at a depth of 10 meters
WP_015025425.1	Psychroflexus torquis ATCC 700755	Bacteroidetes	NC_018721.1	solated from Prydz Bay, Antarctica from a sea-ice algal assemblage
WP_015362984.1	Nonlabens dokdonensis DSW-6	Bacteroidetes	NC_020156.1	marine
WP_015479867.1	Polaribacter sp MED152	Bacteroidetes	NC_020830.1	marine,
WP_019037051.1	Psychroflexus tropicus DSM 15496	Bacteroidetes	NZ_KB905342.1	Salinewater; Lake; Fresh water; Aquatic
WP_019668396.1	Eudoraea adriatica DSM 19308	Bacteroidetes	NZ_KB907546.1	Marine; Aquatic; Marine; Sea water; Surface water
WP_020534502.1	Lewinella cohaerens DSM 23179	Bacteroidetes	NZ_KB903532.1	Sand; Marine
WP_020570649.1	Lewinella persica DSM 23188	Bacteroidetes	NZ_KB890450.1	Brown mud, Galway, Ireland
WP_020895545.1	Winogradskyella psychrotolerans RS-3	Bacteroidetes	NZ_ATMR01000164.1	sediment sample from Arctic
WP_024282467.1	Algoriphagus marincola HL-49	Bacteroidetes	NZ_JAFX01000001.1	NA

RefSeq	Species	Phylum	Genome/ Nucleotide accession	Isolation source
WP_024980038.1	Flavobacterium succinicans LMG 10402	Bacteroidetes	NZ_JATV01000002.1	NA
WP_025007312.1	Marinilabilia salmonicolor JCM 21150	Bacteroidetes	NZ_AJKI01000016.1	Marine mud
WP_026716257.1	Flavobacterium gelidilacus	Bacteroidetes	NZ_AUGN01000004.1	Microbial mats in Antarctic lakes
WP_026755330.1	Sediminibacter sp Hel_I_10	Bacteroidetes	NZ_JHZX01000001.1	Seawater
WP_026951677.1	Algoriphagus mannitolivorans DSM15301	Bacteroidetes	NZ_AUBV01000007.1	Sediment sample of getbol, Korean tidal flat
WP_027136710.1	Gaetbulibacter saemankumensis DSM 17032	Bacteroidetes	NZ_AUDV01000005.1	Tidal flat sediment
WP_028873565.1	Psychroserpens burtonensis DSM 12212	Bacteroidetes	NZ_KE384347.1	Antarctic; Lake; Fresh water; Aquatic
WP_033957934.1	Psychroserpens jangbogonensis PAMC 27130	Bacteroidetes	NZ_JSWG01000006.1	Antarctic marine sediment
WP_034058179.1	Lacinutrix jangbogonensis PAMC 27137	Bacteroidetes	NZ_JSWF01000017.1	Antarctic marine sediment
WP_035128081.1	Flavobacterium aquatile LMG 4008	Bacteroidetes	NZ_JRHH01000005.1	aquatic
WP_035328191.1	Dokdonia donghaensis DSW-1	Bacteroidetes	NZ_CP015125.1	sea water
WP_035483300.1	Algoriphagus marincola DSM 16067	Bacteroidetes	NZ_AUBU01000003.1	Sea water; Marine; Aquatic
WP_035655455.1	Flavobacterium filum DSM 17961	Bacteroidetes	NZ_AUDM01000005.1	Wastewater treatment plant
WP_036786794.1	Polaribacter sp Hel1_33_49	Bacteroidetes	NZ_JPDI01000001.1	surface seawater, North Sea
WP_036821579.1	Polaribacter sp Hel1_85	Bacteroidetes	NZ_JPDS01000001.1	surface seawater, North Sea
WP_036841651.1	Polaribacter sp Hel_I_88	Bacteroidetes	NZ_JHZZ01000001.1	seawater
WP_038600314.1	Dokdonia sp MED134	Bacteroidetes	NZ_CP009301.1	Mediterranean sea collected at a depth of 1 meter
WP_038693341.1	Riemerella anatipestifer RA-CH-1	Bacteroidetes	NC_018609.1	sick duck
WP_041494898.1	Nonlabens marinus S1-08	Bacteroidetes	NZ_AP014548.1	solated from seawater from the Pacific Ocean
WP_042279461.1	Nonlabens sediminis JCM 19294	Bacteroidetes	NZ_BBML01000006.1	NA
WP_044225198.1	Phaeodactylibacter xiamenensis	Bacteroidetes	NZ_JPOS01000079.1	seawater
WP_044637609.1	Siansivirga zeaxanthinifaciens CC-SAMT-1	Bacteroidetes	NZ_CP007202.1	marine water
WP_045371590.1	Jejuia pallidilutea 11shioA1	Bacteroidetes	NZ_BBNY01000005.1	surface of seaweed
WP_045469600.1	Winogradskyella sp PG-2	Bacteroidetes	NZ_AP014583.1	marine
WP_046745370.1	Kordia zhangzhouensis MCCC 1A00726	Bacteroidetes	NZ_LBMH01000007.1	Surface sediment of the Jiulong River
WP_046758148.1	Kordia jejudonensis SSK3-3	Bacteroidetes	NZ_LBMG01000037.1	water area where the ocean and a freshwater spring meet
WP_047420121.1	Cellulophaga sp Hel_I_12	Bacteroidetes	NZ_JUHB01000001.1	Northsea
WP_050378900.1	Chryseobacterium sp Hurlbut01	Bacteroidetes	NZ_LGIP01000019.1	Indoor Water Fountain

RefSeq	Species	Phylum	Genome/ Nucleotide accession	Isolation source
WP_050977373.1	Nitritalea halalkaliphila LW7	Bacteroidetes	NZ_AJYA01000010.1	Lonar Lake, a halo-alkaline lake in Buldana District
WP_051113358.1	Lunatimonas lonarensis AK24	Bacteroidetes	NZ_AQHR01000110.1	Lonar Lake, a halo-alkaline lake in Buldana District
WP_051554804.1	Maribacter antarcticus DSM 21422	Bacteroidetes	NZ_KK211214.1	marine, associated with Green alga Pyramimonas gelidicola
WP_051774666.1	Formosa agariphila KMM 3901	Bacteroidetes	NZ_HG315671.1	isolate from the green algae Acrosiphonia sonderi
WP_052143767.1	Flaviramulus ichthyoenteri Th78	Bacteroidetes	NZ_KN525716.1	fish intestine
WP_052910403.1	Riemerella anatipestifer RA-JLLY	Bacteroidetes	NZ_LAVB01000002.1	isolated from the brain of an old duck
WP_053974405.1	Polaribacter dokdonensis DSW-5	Bacteroidetes	NZ_LGBR01000001.1	seawater
WP_054560158.1	Croceitalea dokdonensis DOKDO 023	Bacteroidetes	NZ_LDJX01000007.1	seawater
WP_054718612.1	Marinifilum fragile JCM 15579	Bacteroidetes	NZ_BAZX01000008.1	NA
WP_055412108.1	Nonlabens sp YIK11	Bacteroidetes	NZ_LBMJ01000001.1	marine sediment from tidal flat
WP_056016978.1	Chryseobacterium aquaticum KCTC 12483	Bacteroidetes	NZ_LLYZ01000020.1	fresh water resevoir
WP_057778920.1	Formosa algae KMM 3553	Bacteroidetes	NZ_LMAK01000007.1	marine, associated with the brown alga Fucus evanescens
WP_057953806.1	Salinivirga cyanobacteriivorans DSM:27204	Bacteroidetes	NZ_CP013118.1	Hypersaline microbial mat
WP_059137603.1	Chryseobacterium greenlandense UMB34	Bacteroidetes	NZ_LMAI01000010.1	ice core, tundra
WP_060524661.1	Nonlabens sp MIC269	Bacteroidetes	NZ_CP011373.1	marine sediment, tidal flat
WP_062037044.1	Lentimicrobium saccharophilum	Bacteroidetes	NZ_DF968182.1	anaerobic bioreactor
WP_062055367.1	Sediminicola sp YIK13	Bacteroidetes	NZ_CP010535.1	sediment
WP_064969793.1	Riemerella anatipestifer	Bacteroidetes	NZ_LUDR01000005.1	NA
WP_065318445.1	Polaribacter vadi LPB0003	Bacteroidetes	NZ_CP017477.1	associated with sea snail Thais luteostoma
WP_068356408.1	Polaribacter reichenbachii 6Alg 8T	Bacteroidetes	NZ_CP019419.1	marine, assoicated with Green Alga, Ulva fenestrata
WP_068448247.1	Polaribacter atrinae KACC 17473	Bacteroidetes	NZ_LVWE01000005.1	associated with comb pen shell
WP_069675026.1	Formosa sp Hel3_A1_48	Bacteroidetes	NZ_CP017259.1	Marine phytoplankton bloom
WP_069677645.1	Formosa sp Hel1_33_131	Bacteroidetes	NZ_CP017260.1	Marine phytoplankton bloom
WP_069829226.1	Algibacter sp SK-16	Bacteroidetes	NZ_MDJD01000014.1	surface water of a coastal marine environment, Aburatsubo Inlet, Misaki, Japan
WP_071669937.1	Flavobacteriaceae bacterium UJ101	Bacteroidetes	NZ_CP016269.1	associated with crab
WP_073000654.1	Mariniphaga anaerophila DSM 26910	Bacteroidetes	NZ_FQUM01000003.1	NA
WP_073082982.1	Winogradskyella jejuensis DSM 25330	Bacteroidetes	NZ_FQWS01000001.1	NA

RefSeq	Species	Phylum	Genome/ Nucleotide accession	Isolation source
WP_073176157.1	Spongiibacterium flavum DSM 22638	Bacteroidetes	NZ_FQWL01000001.1	NA
WP_073192712.1	Psychroflexus salarius DSM 25661	Bacteroidetes	NZ_FQTW01000003.1	NA
WP_073319268.1	Aquimarina spongiae DSM 22623	Bacteroidetes	NZ_FQYP01000008.1	NA
WP_073362871.1	Flavobacterium fontis DSM 25660	Bacteroidetes	NZ_FQVQ01000006.1	NA
WP_075340764.1	Tenacibaculum sp HZ1	Bacteroidetes	NZ_MSMP01000002.1	leaf of Porphyra yezoensis Ueda (cold water seaweed)
WP_075349687.1	Algoriphagus sp am2	Bacteroidetes	NZ_MSPQ01000002.1	marine sediment
WP_075591299.1	Labilibacter marinus Y11	Bacteroidetes	NZ_LXYE02000049.1	marine sediment from coastal area
WP_076392929.1	Chryseobacterium gambrini DSM 18014	Bacteroidetes	NZ_FTOV01000005.1	NA
WP_076502411.1	Belliella pelovolcani DSM 46698	Bacteroidetes	NZ_FTOP01000013.1	NA
WP_076621343.1	Winogradskyella sp J14-2	Bacteroidetes	NZ_CP019388.1	seawater
WP_066217685.1	Balneola sp EhC07	Balneolaeota	NZ_LXYG01000004.1	Pacific Ocean: South Pacific,associated with Emiliania huxleyi
WP_014854794.1	Melioribacter roseus P3M	Chlorobi	NC_018178.1	microbial mat, developing on the wooden surface of a chute, under the flow of hot water coming from an oil exploring well
BAJ63785.1	Anaerolinea thermophila UNI-1	Chloroflexi	AP012029.1	no data
GAP15770.1	Longilinea arvoryzae KOME-1	Chloroflexi	DF967972.1	paddy field
GAP18847.1	Levilinea saccharolytica KIBI-1	Chloroflexi	DF967974.1	anaerobic bioreactor
WP_062419172.1	Levilinea saccharolytica strain KIBI-1	Chloroflexi	NZ_LGCM01000060.1	sludge granules from sugar-processing plant
WP_070105148.1	Anaerolinea thermophila UNI-1	Chloroflexi	NC_014960.1	thermophilic sludge,reactor treating fried soybean-curd manufacturing waste water.
WP_075075299.1	Longilinea arvoryzae KOME-1	Chloroflexi	NZ_DF967972.1	paddy field soil
AOM82220.1	Bacillus beveridgei MLTeJB	Firmicutes	CP012502.1	lake sediment
OGS53170.1	Firmicutes bacterium GWF2_51_9	Firmicutes	MGWF01000062.1	Rifle well CD01 at time point 6 / F; 5m depth; 0.2 filter
OJV66797.1	Clostridiales bacterium 38-18	Firmicutes	MKTL01000001.1	Thiocyanate stock biobioreactor
ONI39082.1	Epulopiscium sp SCG-B11WGA-EpuloA1	Firmicutes	LJDB01000073.1	foregut of surgeonfish gut
OOB78065.1	Epulopiscium sp Nuni2H_MBin001	Firmicutes	MDJM01000060.1	hindgut of Naso unicorns (bluespine unicornfish
SEL97130.1	Alkalibacterium putridalgicola DSM 19182	Firmicutes	FOBL01000018.1	NA
SFC24100.1	Alkalibacterium subtropicum DSM 23664	Firmicutes	FOLT01000004.1	NA
SFO91185.1	Halolactibacillus alkaliphilus CGMCC 16843	Firmicutes	FOWN01000021.1	NA

RefSeq	Species	Phylum	Genome/ Nucleotide accession	Isolation source
SFP58978.1	Halolactibacillus halophilus DSM 17073	Firmicutes	FOXC01000031.1	NA
SFT05733.1	Halolactibacillus miurensis DSM 17074	Firmicutes	FPAI01000035.1	NA
WP_006038298.1	Paenibacillus curdlanolyticus	Firmicutes	NZ_AEDD01000005.1	Soil from Kobe City, Japan
WP_013172256.1	Bacillus selenitireducens MLS10	Firmicutes	NC_014219.1	isolated from alkaline, hypersaline, arsenic-rich mud from Mono Lake, California
WP_013487088.1	Bacillus cellulosilyticus DSM2522	Firmicutes	NC_014829.1	alkaliphilic strain from soil
WP_017472050.1	Amphibacillus jilinensis Y1	Firmicutes	NZ_JH976435.1	sediment from a soda lake
WP_034302189.1	Alkalibacterium sp AK22	Firmicutes	NZ_JANL01000036.1	alkaliphilic bacterium isolated from Lonar lake water sample
WP_041967361.1	Bacillus selenatarsenatis SF-1	Firmicutes	NZ_BASE01000091.1	soil (type strain
WP_069364333.1	Bacillus beveridgei MLTeJB	Firmicutes	CP012502.1	lake sediment
WP_071848865.1	Alkalibacterium sp 20	Firmicutes	NZ_LSRN01000056.1	antarctic canary pond (beta-galactosidase producing bacteria
WP_074912612.1	Proteiniclasticum ruminis ML2	Firmicutes	NZ_FOVK01000010.1	NA
EMI21452.1	Rhodopirellula maiorica SM1	Planctomycetes	ANOG01000246.1	sediment marine water
WP_008676918.1	Rhodopirellula sallentina SM41	Planctomycetes	NZ_ANOH01000136.1	sediment near San Cataldo, Italy
WP_009100872.1	Rhodopirellula sp SWK7	Planctomycetes	NZ_ANOQ01000091.1	isolated from sediment near Tjärnö, Sweden
WP_075085889.1	Planctomycetaceae bacterium FC18	Planctomycetes	NZ_LWSI01000041.1	associated with macroalge from Portugal: Porto, Foz
ABM02404.1	Psychromonas ingrahamii 37	Proteobacteria	CP000510.1	sea ice, off Point Barrow in northern Alaska
EEB78106.1	marine gamma proteobacterium HTCC2148	Proteobacteria	DS999230.1	10 m water depth from station NH15
EED28642.1	Vibrio parahaemolyticus 16	Proteobacteria	DS999325.1	no data
EJP72459.1	SAR86 cluster bacterium SAR86B	Proteobacteria	JH611193.1	isolated as part of a large dataset composed predominantly from surface water marine samples collected along a voyage from Eastern North American coast to the Eastern Pacific Ocean, including locations in the Sargasso Sea, Panama Canal, and the Galapagos Islands
ERB64611.1	Vibrio coralliilyticus OCN008	Proteobacteria	AVOO01000042.1	coral reef
GAC16427.1	Aliiglaciecola lipolytica E3	Proteobacteria	BAEN01000070.1	no data
GAC38158.1	Paraglaciecola psychrophila 170	Proteobacteria	BAES01000049.1	sea ice at high-latitude Arctic locations
GAD01867.1	Agarivorans albus MKT 106	Proteobacteria	BARX01000011.1	intestine of marine animals
KPK53854.1	Gammaproteobacteria bacterium isolate SG8 31	Proteobacteria	LJTI01000177.1	Sulfate-rich zone estuary sediments 8-12 cm

RefSeq	Species	Phylum	Genome/ Nucleotide accession	Isolation source
KPK55639.1	Gammaproteobacteria bacterium SG8_31	Proteobacteria	LJTI01000137.1	Sulfate-rich zone estuary sediments 8-12 cm
OGA97325.1	Burkholderiales bacterium RIFCSPHIGHO2_12_FULL_69_20	Proteobacteria	MESD01000215.1	subsurface metagenome, Rifle well FP-101 under high O2 conditions
OIQ00041.1	Zetaproteobacteria bacterium CG2 30 46 52	Proteobacteria	MNZV01000010.1	ground water
SEG54255.1	Vibrio hangzhouensis CGMCC 17062	Proteobacteria	FNVG01000018.1	no data
SET34366.1	Thalassotalea agarivorans DSM 19706	Proteobacteria	FOHK01000006.1	no data
WP_001913576.1	Vibrio cholerae CT 5369-93	Proteobacteria	NZ_ADAL01000121.1	sewage
WP_002541194.1	Grimontia indica AK16	Proteobacteria	NZ_ANFM02000039.1	no data
WP_005428312.1	Vibrio campbellii HY01	Proteobacteria	NZ_DS179490.1	marine
WP_006015369.1	Glaciecola pallidula DSM 14239 = ACAM 615	Proteobacteria	NZ_BAEQ01000066.1	marine
WP_006877539.1	Vibrio brasiliensis LMG 20546	Proteobacteria	NZ_AEVS01000008.1	no data
WP_007224925.1	marine gamma proteobacterium HTCC2143	Proteobacteria	NZ_AAVT01000002.1	coastal region of Newport, Oregon.
WP_007619979.1	Paraglaciecola arctica BSs20135	Proteobacteria	NZ_BAEO01000029.1	no data
WP_008041327.1	Reinekea blandensis MED297	Proteobacteria	NZ_CH724149.1	mediteranian sea
WP_009548992.1	Burkholderiales bacterium JOSHI_001	Proteobacteria	NZ_CM001438.1	no data
WP_009706764.1	Vibrio sp HENC-03	Proteobacteria	NZ_JH976124.1	no data
WP_010319050.1	Vibrio ordalii ATCC 33509	Proteobacteria	NZ_AEZC01000100.1	no data
WP_010453521.1	Vibrio rotiferianus DAT722	Proteobacteria	NZ_AFAJ01000064.1	mud crab larvae aquaculture tank
WP_011044475.1	Colwellia psychrerythraea 34H	Proteobacteria	NC_003910.7	Arctic marine sediment
WP_011082183.1	Vibrio vulnificus FORC_017	Proteobacteria	NZ_CP012740.1	no data
WP_012326956.1	Shewanella woodyi ATCC 51908	Proteobacteria	NC_010506.1	Seawater
WP_014110183.1	Glaciecola nitratireducens FR1064	Proteobacteria	NC_016041.1	seawater
WP_015430975.1	Paraglaciecola psychrophila 170	Proteobacteria	NC_020514.1	sea ice at high-latitude Arctic locations
WP_015727420.1	Vibrio vulnificus MO6-24/O	Proteobacteria	NC_014966.1	no data
WP_016961547.1	Enterovibrio norvegicus FF-454	Proteobacteria	NZ_AJWN02000035.1	seawater
WP_017003336.1	Enterovibrio norvegicus FF-33	Proteobacteria	NZ_AJYD02000038.1	Filtered seawater
WP_017006505.1	Enterovibrio norvegicus FF-162	Proteobacteria	NZ_AJYE02000109.1	Filtered seawater
WP_017010416.1	Enterovibrio calviensis FF-85	Proteobacteria	NZ_AJYF02000222.1	seawater
WP_017014771.1	Enterovibrio calviensis 1F-211	Proteobacteria	NZ_AJYG02000063.1	seawater
WP_017029323.1	Vibrio breoganii ZF-55	Proteobacteria	NZ_AJYL02000001.1	seawater

RefSeq	Species	Phylum	Genome/ Nucleotide accession	Isolation source
WP_017033027.1	Vibrio breoganii ZF-29	Proteobacteria	NZ_AJYM02000002.1	seawater
WP_017052035.1	Vibrio genomosp F6 FF-238	Proteobacteria	NZ_AJYW02000160.1	seawater
WP_017071890.1	Vibrio crassostreae ZF-91	Proteobacteria	NZ_AJZC02000073.1	seawater
WP_017242682.1	Vibrio breoganii 1C10	Proteobacteria	NZ_AKXW02000168.1	seawater
WP_017421774.1	Vibrio vulnificus NBRC 15645 = ATCC 27562	Proteobacteria	NZ_BCUB01000003.1	no data
WP_018606806.1	Uliginosibacterium gangwonense DSM 18521	Proteobacteria	NZ_KB892840.1	soil, wetland
WP_019026149.1	Colwellia piezophila ATCC BAA-637	Proteobacteria	NZ_KB905160.1	marine sediment
WP_019615046.1	Psychromonas ossibalaenae JAMM 0738	Proteobacteria	NZ_KB906989.1	Marine sediment beside whale carcasses
WP_019821075.1	Vibrio splendidus 12E03	Proteobacteria	NZ_AJZD02000183.1	seawater
WP_022774065.1	Candidatus Symbiobacter mobilis CR	Proteobacteria	NC_022576.1	prokaryotic symbiosis in the phototrophic consortium "Chlorochromatium aggregatum
WP_026376345.1	Aestuariibacter salexigens DSM15300	Proteobacteria	NZ_AUBH01000006.1	no data
WP_026958287.1	Aliagarivorans taiwanensis DSM 22990	Proteobacteria	NZ_AUBZ01000007.1	no data
WP_026971459.1	Aliagarivorans marinus DSM 23064	Proteobacteria	NZ_AUBY01000017.1	no data
WP_028022281.1	Enterovibrio calviensis DSM 14347	Proteobacteria	NZ_JHZA01000007.1	seawater
WP_028864419.1	Psychromonas aquimarina ATCC BAA-1526	Proteobacteria	NZ_AUAM01000036.1	Marine sediment beside whale carcasses
WP_029408591.1	Thiomicrospira sp Milos-T2	Proteobacteria	NZ_KL370892.1	marine shallow-water hydrothermal vent
WP_032550704.1	Vibrio fortis Dalian14	Proteobacteria	NZ_JFFR01000013.1	associated with Strongylocentrotus intermedius (sea urchins)
WP_038135364.1	Vibrio caribbeanicus T14	Proteobacteria	NZ_JRWR01000003.1	marine water
WP_038177488.1	Vibrio pacinii DSM 19139	Proteobacteria	NZ_KL543978.1	no data
WP_038188127.1	Vibrio sinaloensis T08	Proteobacteria	NZ_JRWP01000004.1	marine water
WP_038212124.1	Vibrio variabilis T01	Proteobacteria	NZ_JRWM01000003.1	seawater
WP_038887920.1	Vibrio rotiferianus CAIM 577 = LMG 21460	Proteobacteria	NZ_BAOI01000269.1	no data
WP_038888526.1	Vibrio campbellii 200612B	Proteobacteria	NZ_BANY01000003.1	no data
WP_038939200.1	Vibrio vulnificus JY1305	Proteobacteria	NZ_AFSW01000031.1	no data
WP_038968329.1	Vibrio vulnificus E64MW E NeaNN c6,	Proteobacteria	NZ_AFSX01000006.1	no data
WP_039442894.1	Vibrio navarrensis 08-2462	Proteobacteria	NZ_JMCI01000086.1	human blood
WP_039448752.1	Vibrio vulnificus ATCC 33147	Proteobacteria	NZ_JRQR01000014.1	fish pathogen

RefSeq	Species	Phylum	Genome/ Nucleotide accession	Isolation source
WP_039456055.1	Photobacterium gaetbulicola AD005a	Proteobacteria	NZ_JWLZ01000001.1	Coastal marine water
WP_039463204.1	Vibrio navarrensis 0053-83	Proteobacteria	NZ_JMCF01000065.1	human wound
WP_039467995.1	Vibrio vulnificus 99-578 DP-B1	Proteobacteria	NZ_JSVD01000004.1	no data
WP_039475600.1	Vibrio sinaloensis AD032	Proteobacteria	NZ_JWLU01000007.1	Coastal marine water
WP_039481470.1	Vibrio sinaloensis AD048	Proteobacteria	NZ_JWLV01000001.1	Coastal marine water
WP_039536454.1	Vibrio vulnificus 2322	Proteobacteria	NZ_JQDS01000004.1	fish pond water-fish farm
WP_039624851.1	Vibrio sinaloensis T47	Proteobacteria	NZ_JXBJ01000003.1	marine water
WP_039840769.1	Vibrio owensii 47666-1	Proteobacteria	NZ_JPRC01000005.1	pathogen of prawns
WP_040110433.1	Vibrio vulnificus 93U204	Proteobacteria	NZ_CP009262.1	no data
WP_040307136.1	749888648 Agarivorans albus MKT106	Proteobacteria	NZ_BARX01000011.1	no data
WP_042488649.1	Vibrio diazotrophicus NBRC 103148	Proteobacteria	NZ_BBJY01000019.1	Echinoderm GI tract
WP_043011791.1	Vibrio coralliilyticus RE98	Proteobacteria	NZ_CP009620.1	shellfish hatchery
WP_043886430.1	Vibrio parahaemolyticus 16	Proteobacteria	NZ_DS999325.1	no data
WP_044621379.1	Photobacterium gaetbulicola Gung47	Proteobacteria	NZ_CP005973.1	tidal flat on the west coast in Korea
WP_045371119.1	Vibrio campbellii 051011F	Proteobacteria	NZ_BBKV01000055.1	marine
WP_045383138.1	Vibrio campbellii 151112C	Proteobacteria	NZ_BBKW01000134.1	marine
WP_045392215.1	Vibrio rotiferianus Oz08	Proteobacteria	NZ_BBLC01000464.1	no data
WP_045569123.1	Vibrio sp S234-5	Proteobacteria	NZ_JZAM01000072.1	no data
WP_045591188.1	Vibrio vulnificus SC9613	Proteobacteria	NZ_JZER01000124.1	seafood
WP_045593734.1	Vibrio vulnificus SC9729	Proteobacteria	NZ_JZEQ01000021.1	no data
WP_045598648.1	Vibrio vulnificus NV1	Proteobacteria	NZ_JZET01000036.1	no data
WP_045609964.1	Vibrio vulnificus SC9740	Proteobacteria	NZ_JZEP01000105.1	no data
WP_045612706.1	Vibrio vulnificus NV22	Proteobacteria	NZ_JZES01000092.1	no data
WP_045623449.1	Vibrio vulnificus SC9794	Proteobacteria	NZ_JZEN01000041.1	tidal mudflat
WP_045628599.1	Vibrio vulnificus SC9761	Proteobacteria	NZ_JZEO01000016.1	seafood
WP_045977480.1	Vibrio neptunius S2394	Proteobacteria	NZ_JXXU01000068.1	Jellyfish
WP_046029587.1	Vibrio vulnificus SC9629	Proteobacteria	NZ_JYGG01000066.1	seafood
WP_046304591.1	Grimontia sp AD028 1	Proteobacteria	NZ_LASY01000054.1	Coastal marine water
WP_047111508.1	Vibrio vulnificus CladeA-yb158	Proteobacteria	NZ_LBNN01000017.1	ills, fins and scales pooled from 10 healthy Tilapia fish grown in aquaculture ponds
WP_047476905.1	Vibrio campbellii UMTGB204	Proteobacteria	NZ_JSFE01000001.1	marine associated with Didemnum molle (Green Barrel Tunicate)
WP_048613273.1	Vibrio sp J2-1	Proteobacteria	NZ_CCJZ01000013.1	associated with living oyster

RefSeq	Species	Phylum	Genome/ Nucleotide accession	Isolation source
WP_048659631.1	Vibrio crassostreae J5-19	Proteobacteria	NZ_CCJU01000055.1	associated with disease in oyster
WP_050634916.1	Vibrio sp. J2-15	Proteobacteria	NZ_LK934448.1	no data
WP_050651495.1	Vibrio sp. J2-31	Proteobacteria	NZ_LK933984.1	no data
WP_050909111.1	Vibrio campbellii 602L	Proteobacteria	NZ_JPTF01000041.1	marine surface water
WP_050913756.1	Vibrio harveyi KC13175	Proteobacteria	NZ_DF967931.1	associated with Litopenaeus vannamei (whiteleg shrimp)
WP_051117377.1	Vibrio campbellii CAIM1500	Proteobacteria	NZ_JPTH01000245.1	Lutjanus guttatus liver (Spotted Rose Snapper)
WP_053543277.1	Vibrio vulnificus V252	Proteobacteria	NZ_LIIO01000045.1	human wound
WP_054021206.1	Ideonella sakaiensis	Proteobacteria	NZ_BBYR01000044.1	sediment from PET degrading factory
WP_054547313.1	Vibrio splendidus UCD-SED7	Proteobacteria	NZ_LIZK01000004.1	seagrass sediment
WP_055050307.1	Vibrio metoecus 07-2435	Proteobacteria	NZ_LCUE01000001.1	human wound
WP_055319761.1	Vibrio crassostreae J5-5	Proteobacteria	NZ_CCJV01000093.1	associated with disease in oyster
WP_055731608.1	Agarivorans gilvus WH0801	Proteobacteria	NZ_CP013021.1	no data
WP_058645752.1	Vibrio vulnificus ATCC 29306	Proteobacteria	NZ_LORY01000002.1	vibrio infection in human
WP_058666427.1	Vibrio vulnificus ATCC 43382	Proteobacteria	NZ_LOSI01000001.1	vibrio infection in human
WP_060533144.1	Vibrio vulnificus FORC_016	Proteobacteria	NZ_CP011776.1	human blood
WP_060982384.1	Vibrio splendidus MOR2	Proteobacteria	NZ_LNRE01000004.1	shore seawater
WP_061057768.1	Vibrio vulnificus ATL 6-1306	Proteobacteria	NZ_CP014048.1	human infection
WP_061895861.1	Vibrio cidicii 2423-01	Proteobacteria	NZ_LOBQ01000100.1	human blood
WP_061898254.1	Vibrio cidicii 2538-88	Proteobacteria	NZ_LOBR01000123.1	human blood
WP_061900117.1	Vibrio cidicii 1048-83	Proteobacteria	NZ_LOBP01000148.1	human blood
WP_062664862.1	Grimontia celer CECT 9029	Proteobacteria	NZ_FIZX01000002.1	surface seawater
WP_062711646.1	Grimontia marina CECT 8713	Proteobacteria	NZ_FIZY01000029.1	surface seawater
WP_063606839.1	Vibrio barjaei 3062	Proteobacteria	NZ_LQXO02000023.1	associated with Ruditapes decussatus (Grooved carpet shell)
WP_064791263.1	Shewanella sp Alg231 23	Proteobacteria	NZ_FKJK01000001.1	Marine Sponge
WP_065209351.1	Vibrio breoganii FF50	Proteobacteria	NZ_CP016177.1	seawater
WP_065303737.1	Vibrio natriegens CCUG 16371	Proteobacteria	NZ_CP016348.1	isolate from salt marsh mud
WP_065598362.1	Vibrio breoganii 1A06	Proteobacteria	NZ_MAKG01000024.1	sea water
WP_067411848.1	Enterovibrio coralii CAIM912T1	Proteobacteria	NZ_LNTY01000006.1	Water extract of bleached Merulina ampliata
WP_068544274.1	Thalassotalea sp. LPB0090	Proteobacteria	NZ_CP017689.1	associated with Crassostrea gigas (Pacific oyster)
WP_070374120.1	Colwellia sp PAMC 20917	Proteobacteria	NZ_CP014944.1	no data
WP_071235803.1	Vibrio rotiferianus HM-10	Proteobacteria	NZ_KV861323.1	liver of juvenile Hippocampus erectus (lined seahorse) with tail-rotted disease
WP_072600469.1	Vibrio vulnificus DAL-79087	Proteobacteria	NZ_LMXZ01000025.1	vibrio infection in human
WP_072604925.1	Vibrio fluvialis NCTC 11327	Proteobacteria	NZ_LMTE01000007.1	human feces

RefSeq	Species	Phylum	Genome/ Nucleotide accession	Isolation source
WP_073581065.1	Vibrio quintilis CECT 7734	Proteobacteria	NZ_FRFG01000017.1	seawater
WP_073605828.1	Vibrio aerogenes CECT 7868	Proteobacteria	NZ_FQXZ01000046.1	seagrass bed sediment
WP_076651708.1	Vibrio sp. 10N26145E1 1	Proteobacteria	NZ_MKKP01000081.1	NA
WP_077202203.1	Vibrio campbellii LMB29	Proteobacteria	NZ_CP019294.1	45.2
WP_077262559.1	Vibrio campbellii 1114GL	Proteobacteria	NZ_CP019635.1	45.3
WP_002702259.1	Treponema saccharophilum DSM 2985	Spirochaetes	NZ_AGRW01000030.1	Bovine rumen
WP_009103739.1	Treponema sp. JC4	Spirochaetes	NZ_AJGU01000004.1	bovine rumen
WP_013759594.1	Treponema brennaborense DSM 12168	Spirochaetes	NC_015500.1	skin from ulcerative lesion of bovine foot infected with digital dermatitis
WP_022933544.1	Treponema bryantii NK4A124	Spirochaetes	NZ_ATWV01000024.1	no data
WP_024267116.1	Salinispira pacificaL21-RPul-D2	Spirochaetes	NC_023035.1	no data
WP_038073801.1	Treponema sp. C6A8	Spirochaetes	NZ_JHVB01000005.1	no data
WP_052296413.1	Treponema caldarium DSM7334	Spirochaetes	NC_015732.1	Fresh water hot spring
WP_069893415.1	Candidatus Marispirochaeta associata JC231	Spirochaetes	NZ_JWTM01000186.1	saline soil
WP_074645719.1	Treponema bryantii B25	Spirochaetes	NZ_FOFU01000018.1	no data
WP_074929665.1	Treponema bryantii XBD1002	Spirochaetes	NZ_FORI01000001.1	no data
GAK53004.1	Candidatus Moduliflexus flocculans UASB14	unclassified	DF820459.1	wastewater treatment system
OGF14517.1	Candidatus Eisenbacteria bacterium RBG 16 71 46	unclassified	MFFQ01000049.1	Rifle background sediment
WP_007412870.1	Pedosphaera parvula Ellin514	Verrucomicrobia	NZ_ABOX02000002.1	pasture soil
WP_008104697.1	Verrucomicrobiae bacterium DG1235	Verrucomicrobia	NZ_DS990592.1	marine
WP_052329088.1	Verrucomicrobia bacterium SCGC AAA164-O14	Verrucomicrobia	NZ_HF539891.1	surface water of the Gulf of Maine

App. 4.2. GH149 sequences from other databases.

Identification number	Database	Species	Domain	Phylum	Isolation sources
EgP1 (in this study)		*Euglena gracilis var. saccharophila*	Eukaryota	Euglenozoa	fresh water
EgP2		*Euglena gracilis var. saccharophila*	Eukaryota	Euglenozoa	fresh water
EgP3		*Euglena gracilis var. saccharophila*	Eukaryota	Euglenozoa	fresh water
EgP4		*Euglena gracilis var. saccharophila*	Eukaryota	Euglenozoa	fresh water
200380720	http://marinemicroeukaryotes.org/	*Eutreptiella gymnastica CCMP1594*	Eukaryota	Euglenozoa	marine
200416338	http://marinemicroeukaryotes.org/	*Eutreptiella gymnastica CCMP1594*	Eukaryota	Euglenozoa	marine

Identification number	Database	Species	Domain	Phylum	Isolation sources
200418968	http://marinemicroeukaryotes.org/	*Eutreptiella gymnastica CCMP1594*	Eukaryota	Euglenozoa	marine
750	http://euglenadb.org/	*Euglena gracilis Z*	Eukaryota	Euglenozoa	fresh water
1135	http://euglenadb.org/	*Euglena gracilis Z*	Eukaryota	Euglenozoa	fresh water
BAV19384.1	NCBI	*Euglena gracilis NIES-47*	Eukaryota	Euglenozoa	fresh water
Pro_7066	Prozomix metagenome database	NA	Bacteria	Bacteroidetes	NA

Appendix 5. BLAST analyses of GH149 amino acid sequences against EgP1

Sequence ID	Query coverage	E-value	% identity to EgP1
EgP2	100%	0	72%
EgP3	100%	0	73%
EgP4	100%	0	91%
Pro_7066	97%	0	45%
200380720	97%	0	62%
200416338	95%	0	50%
200418968	96%	0	51%
EG_750	97%	0	50%
EG_1135	96%	0	53%
BAV19384.1	93%	0	53%
WP_052329088.1	68%	0	40%
WP_017010416.1	71%	0	47%
KRO88543.1	74%	0	44%
WP_050977373.1	77%	0	41%
OOB78065.1	87%	0	40%
EMI21452.1	87%	0	46%
OGA97325.1	90%	0	43%
WP_054021206.1	91%	0	42%
SFO59267.1	91%	0	43%
EYR63563.1	91%	0	45%
WP_052022748.1	91%	0	45%
WP_009548992.1	92%	0	42%
GAK91234.1	92%	0	47%
WP_062419172.1	92%	0	47%
GAK94769.1	92%	0	48%
WP_009103739.1	93%	0	40%
WP_038073801.1	93%	0	40%
WP_074645719.1	93%	0	40%
SEF71509.1	93%	0	44%
SFI14923.1	93%	0	44%
KRP08637.1	93%	0	45%
KRP29741.1	93%	0	45%
EON75098.1	93%	0	46%
OIQ30310.1	93%	0	46%
SDS23837.1	93%	0	46%
WP_073176157.1	93%	0	47%
WP_076621343.1	93%	0	47%
WP_020534502.1	93%	0	48%

Sequence ID	Query coverage	E-value	% identity to EgP1
WP_044225198.1	93%	0	48%
ONI39082.1	94%	0	37%
WP_002702259.1	94%	0	37%
WP_013759594.1	94%	0	39%
SIT04910.1	94%	0	41%
WP_022933544.1	94%	0	41%
WP_074929665.1	94%	0	41%
WP_076502411.1	94%	0	41%
WP_006038298.1	94%	0	42%
OGS53170.1	94%	0	43%
OJV66797.1	94%	0	43%
WP_029408591.1	94%	0	43%
WP_071848865.1	94%	0	43%
AOM82220.1	94%	0	44%
OFX41757.1	94%	0	44%
OIP85247.1	94%	0	44%
SFC24100.1	94%	0	44%
WP_008041327.1	94%	0	44%
WP_014854794.1	94%	0	44%
WP_019821075.1	94%	0	44%
WP_069893415.1	94%	0	44%
WP_074912612.1	94%	0	44%
AFR34657.1	94%	0	45%
CDF79585.1	94%	0	45%
KRP13118.1	94%	0	45%
OGS61795.1	94%	0	45%
SEP83958.1	94%	0	45%
SFG38750.1	94%	0	45%
SFP58978.1	94%	0	45%
SFT05733.1	94%	0	45%
WP_009100872.1	94%	0	45%
WP_013172256.1	94%	0	45%
WP_027136710.1	94%	0	45%
WP_028873565.1	94%	0	45%
WP_032550704.1	94%	0	45%
WP_033957934.1	94%	0	45%
WP_034058179.1	94%	0	45%
WP_036821579.1	94%	0	45%
WP_038693341.1	94%	0	45%
WP_045371590.1	94%	0	45%
WP_051554804.1	94%	0	45%

Sequence ID	Query coverage	E-value	% identity to EgP1
WP_051774666.1	94%	0	45%
WP_052910403.1	94%	0	45%
WP_057778920.1	94%	0	45%
WP_064969793.1	94%	0	45%
WP_073082982.1	94%	0	45%
WP_073362871.1	94%	0	45%
WP_075085889.1	94%	0	45%
WP_075591299.1	94%	0	45%
GAP42007.1	94%	0	46%
KPP94675.1	94%	0	46%
KRO66514.1	94%	0	46%
KUO68202.1	94%	0	46%
OGS73573.1	94%	0	46%
OIP52753.1	94%	0	46%
SDI57228.1	94%	0	46%
SDM08669.1	94%	0	46%
SDX94361.1	94%	0	46%
SEC77108.1	94%	0	46%
SFN60596.1	94%	0	46%
SFS48762.1	94%	0	46%
WP_002541194.1	94%	0	46%
WP_004571105.1	94%	0	46%
WP_009183469.1	94%	0	46%
WP_013306491.1	94%	0	46%
WP_013769870.1	94%	0	46%
WP_015479867.1	94%	0	46%
WP_019037051.1	94%	0	46%
WP_019134858.1	94%	0	46%
WP_024282467.1	94%	0	46%
WP_025007312.1	94%	0	46%
WP_026716257.1	94%	0	46%
WP_035128081.1	94%	0	46%
WP_035483300.1	94%	0	46%
WP_042279461.1	94%	0	46%
WP_050378900.1	94%	0	46%
WP_052143767.1	94%	0	46%
WP_053974405.1	94%	0	46%
WP_054560158.1	94%	0	46%
WP_054718612.1	94%	0	46%
WP_056016978.1	94%	0	46%
WP_056085790.1	94%	0	46%

Sequence ID	Query coverage	E-value	% identity to EgP1
WP_057208477.1	94%	0	46%
WP_057953806.1	94%	0	46%
WP_059137603.1	94%	0	46%
WP_060524661.1	94%	0	46%
WP_062037044.1	94%	0	46%
WP_063606839.1	94%	0	46%
WP_066217685.1	94%	0	46%
WP_068448247.1	94%	0	46%
WP_068544274.1	94%	0	46%
WP_069675026.1	94%	0	46%
WP_070105148.1	94%	0	46%
WP_075340764.1	94%	0	46%
WP_076392929.1	94%	0	46%
GAP15770.1	94%	0	47%
OGS36884.1	94%	0	47%
OJJ21699.1	94%	0	47%
SCY20814.1	94%	0	47%
SDC80925.1	94%	0	47%
SDU10865.1	94%	0	47%
SET34366.1	94%	0	47%
WP_007093907.1	94%	0	47%
WP_007412870.1	94%	0	47%
WP_008253479.1	94%	0	47%
WP_015362984.1	94%	0	47%
WP_024980038.1	94%	0	47%
WP_026755330.1	94%	0	47%
WP_026951677.1	94%	0	47%
WP_036786794.1	94%	0	47%
WP_044637609.1	94%	0	47%
WP_055412108.1	94%	0	47%
WP_062055367.1	94%	0	47%
WP_075075299.1	94%	0	47%
GAK53004.1	94%	0	48%
WP_008629750.1	94%	0	48%
WP_041494898.1	94%	0	48%
WP_075349687.1	94%	0	48%
OGF14517.1	94%	0	49%
WP_020570649.1	94%	0	49%
ABM02404.1	95%	0	45%
AEJ20297.1	95%	0	43%
WP_008104697.1	95%	0	43%

Sequence ID	Query coverage	E-value	% identity to EgP1
WP_010453521.1	95%	0	43%
WP_013487088.1	95%	0	43%
WP_052296413.1	95%	0	43%
WP_071235803.1	95%	0	43%
WP_005428312.1	95%	0	44%
WP_009706764.1	95%	0	44%
WP_017052035.1	95%	0	44%
WP_017071890.1	95%	0	44%
WP_019668396.1	95%	0	44%
WP_038177488.1	95%	0	44%
WP_038887920.1	95%	0	44%
WP_038888526.1	95%	0	44%
WP_039840769.1	95%	0	44%
WP_045371119.1	95%	0	44%
WP_045383138.1	95%	0	44%
WP_045392215.1	95%	0	44%
WP_045469600.1	95%	0	44%
WP_047476905.1	95%	0	44%
WP_048659631.1	95%	0	44%
WP_050634916.1	95%	0	44%
WP_050651495.1	95%	0	44%
WP_050909111.1	95%	0	44%
WP_050913756.1	95%	0	44%
WP_051117377.1	95%	0	44%
WP_054547313.1	95%	0	44%
WP_055319761.1	95%	0	44%
WP_060982384.1	95%	0	44%
WP_065303737.1	95%	0	44%
WP_069364333.1	95%	0	44%
WP_073000654.1	95%	0	44%
WP_073581065.1	95%	0	44%
WP_076651708.1	95%	0	44%
WP_077202203.1	95%	0	44%
WP_077262559.1	95%	0	44%
EED28642.1	95%	0	45%
ERB64611.1	95%	0	45%
GAC38158.1	95%	0	45%
GAD01867.1	95%	0	45%
SEG54255.1	95%	0	45%
SEL97130.1	95%	0	45%
WP_001913576.1	95%	0	45%

Sequence ID	Query coverage	E-value	% identity to EgP1
WP_006015369.1	95%	0	45%
WP_006877539.1	95%	0	45%
WP_008676918.1	95%	0	45%
WP_010319050.1	95%	0	45%
WP_011082183.1	95%	0	45%
WP_013882364.1	95%	0	45%
WP_014110183.1	95%	0	45%
WP_015430975.1	95%	0	45%
WP_015727420.1	95%	0	45%
WP_017029323.1	95%	0	45%
WP_017033027.1	95%	0	45%
WP_017242682.1	95%	0	45%
WP_017421774.1	95%	0	45%
WP_017472050.1	95%	0	45%
WP_018606806.1	95%	0	45%
WP_020895545.1	95%	0	45%
WP_034302189.1	95%	0	45%
WP_035655455.1	95%	0	45%
WP_036841651.1	95%	0	45%
WP_038135364.1	95%	0	45%
WP_038188127.1	95%	0	45%
WP_038212124.1	95%	0	45%
WP_038939200.1	95%	0	45%
WP_038968329.1	95%	0	45%
WP_039442894.1	95%	0	45%
WP_039448752.1	95%	0	45%
WP_039463204.1	95%	0	45%
WP_039467995.1	95%	0	45%
WP_039475600.1	95%	0	45%
WP_039481470.1	95%	0	45%
WP_039536454.1	95%	0	45%
WP_039624851.1	95%	0	45%
WP_040110433.1	95%	0	45%
WP_040307136.1	95%	0	45%
WP_042488649.1	95%	0	45%
WP_043011791.1	95%	0	45%
WP_043886430.1	95%	0	45%
WP_045569123.1	95%	0	45%
WP_045591188.1	95%	0	45%
WP_045593734.1	95%	0	45%

Sequence ID	Query coverage	E-value	% identity to EgP1
WP_045598648.1	95%	0	45%
WP_045609964.1	95%	0	45%
WP_045612706.1	95%	0	45%
WP_045623449.1	95%	0	45%
WP_045628599.1	95%	0	45%
WP_045977480.1	95%	0	45%
WP_046029587.1	95%	0	45%
WP_047111508.1	95%	0	45%
WP_048613273.1	95%	0	45%
WP_053543277.1	95%	0	45%
WP_055050307.1	95%	0	45%
WP_055731608.1	95%	0	45%
WP_056589364.1	95%	0	45%
WP_058645752.1	95%	0	45%
WP_058666427.1	95%	0	45%
WP_060533144.1	95%	0	45%
WP_061057768.1	95%	0	45%
WP_061895861.1	95%	0	45%
WP_061898254.1	95%	0	45%
WP_061900117.1	95%	0	45%
WP_062711646.1	95%	0	45%
WP_065598362.1	95%	0	45%
WP_067411848.1	95%	0	45%
WP_069829226.1	95%	0	45%
WP_072600469.1	95%	0	45%
WP_072604925.1	95%	0	45%
WP_073605828.1	95%	0	45%
OIP46152.1	95%	0	46%
OIQ00041.1	95%	0	46%
SDR87359.1	95%	0	46%
SHN04610.1	95%	0	46%
WP_007619979.1	95%	0	46%
WP_008269464.1	95%	0	46%
WP_015025425.1	95%	0	46%
WP_017014771.1	95%	0	46%
WP_019026149.1	95%	0	46%
WP_019615046.1	95%	0	46%
WP_026376345.1	95%	0	46%
WP_026958287.1	95%	0	46%
WP_026971459.1	95%	0	46%
WP_028864419.1	95%	0	46%

Sequence ID	Query coverage	E-value	% identity to EgP1
WP_034634502.1	95%	0	46%
WP_035328191.1	95%	0	46%
WP_038600314.1	95%	0	46%
WP_046304591.1	95%	0	46%
WP_046745370.1	95%	0	46%
WP_051113358.1	95%	0	46%
WP_062664862.1	95%	0	46%
WP_065209351.1	95%	0	46%
WP_068356408.1	95%	0	46%
WP_069677645.1	95%	0	46%
WP_070374120.1	95%	0	46%
WP_071669937.1	95%	0	46%
WP_073192712.1	95%	0	46%
WP_073319268.1	95%	0	46%
GAC16427.1	95%	0	47%
GAP18847.1	95%	0	47%
KPK55639.1	95%	0	47%
OIO83857.1	95%	0	47%
SDG38233.1	95%	0	47%
WP_003437557.1	95%	0	47%
WP_007224925.1	95%	0	47%
WP_011044475.1	95%	0	47%
WP_046758148.1	95%	0	47%
WP_047420121.1	95%	0	47%
EEB78106.1	95%	0	48%
SEQ41340.1	95%	0	48%
WP_012326956.1	95%	0	48%
WP_039456055.1	95%	0	48%
WP_044621379.1	95%	0	48%
WP_064791263.1	95%	0	48%
KPK53854.1	95%	0	49%
WP_024267116.1	96%	0	42%
SFO91185.1	96%	0	43%
WP_041967361.1	96%	0	43%
EJP72459.1	96%	0	44%
OFX90003.1	96%	0	44%
WP_013115824.1	96%	0	45%
WP_016961547.1	96%	0	45%
WP_017006505.1	96%	0	45%
WP_028022281.1	96%	0	45%
WP_055916706.1	96%	0	45%

Sequence ID	Query coverage	E-value	% identity to EgP1
WP_056027623.1	96%	0	45%
WP_065318445.1	96%	0	45%
BAJ63785.1	96%	0	46%
OIQ41396.1	96%	0	46%
WP_017003336.1	96%	0	46%
WP_022774065.1	96%	0	46%

Appendix 6. GC content analysis of bacterial GH149 DNA sequences

GI number of the proteins and species	Genome/Nucleotide accession	GC content of genome	GC content of sequence
1043035842 Polaribacter vadi LPB0003	NZ_CP017477.1	29.6	30.1
1103706940 Flavobacteriaceae bacterium UJ101	NZ_CP016269.1	30.7	31.7
1085840169 Polaribacter sp KT25b	LT629752.1	30.8	30.6
505292765 Polaribacter sp MED152	NC_020830.1	30.8	29.6
780093728 Winogradskyella sp PG-2	NZ_AP014583.1	33.3	30.4
537741942 Formosa agariphila KMM 3901	HG315671.1	33.6	34.2
917167954 Formosa agariphila KMM 3901	NZ_HG315671.1	33.6	34.1
1140756745 Winogradskyella sp J14-2	NZ_CP019388.1	34.4	33
504838323 Psychroflexus torquis ATCC 700755	NC_018721.1	34.5	34.5
503071536 Maribacter sp HTCC2170	NC_014472.1	34.8	34.9
403311816 Riemerella anatipestifer RA-CH-1	CP003787.1	35	33.9
740908094 Riemerella anatipestifer RA-CH-1	NC_018609.1	35	33.7
505175882 Nonlabens dokdonensis DSW-6	NC_020156.1	35.3	34.2
1085761172 Formosa sp Hel1_31_208	LT629733.1	36.5	33.3
503252427 Bacillus cellulosilyticus DSM2522	NC_014829.1	36.5	36
740815031 Dokdonia sp MED134	NZ_CP009301.1	37.4	35.1
499355778 Colwellia psychrerythraea 34H	NC_003910.7	38	37.7
737345905 Dokdonia donghaensis DSW-1	NZ_CP015125.1	38.15	35.3
1081429352 Nonlabens sp Hel1_33_55	LT627735.1	38.2	38.4
983344310 Nonlabens sp MIC269	NZ_CP011373.1	38.2	34
1057156359 Thalassotalea sp LPB0090	NZ_CP017689.1	38.8	39.1
952995109 Salinivirga cyanobacteriivorans DSM:27204	NZ_CP013118.1	39	39.6
1084285548 Colwellia sp PAMC 20917	NZ_CP014944.1	39	40
753186985 Nonlabens marinus S1-08	NZ_AP014548.1	39.7	38.8
119862927 Psychromonas ingrahamii 37	CP000510.1	40.1	41.6
518444839 Psychromonas ossibalaenae JAMM 0738	NZ_KB906989.1	40.1	42
505243873 Paraglaciecola psychrophila 170	NC_020514.1	40.4	39.6
504667692 Melioribacter roseus P3M	NC_018178.1	41.3	39.2
503876189 Glaciecola nitratireducens FR1064	NC_016041.1	42.3	42.6
501283938 Shewanella woodyi ATCC 51908	NC_010506.1	43.7	43.6
1042741052 Vibrio natriegens CCUG 16371	NZ_CP016348.1	44.8	46.7
1147295716 Vibrio campbellii LMB29	NZ_CP019294.1	45.2	48.8
1148474972 Vibrio campbellii 1114GL	NZ_CP019635.1	45.3	48.6
1041569487 Vibrio breoganii FF50	NZ_CP016177.1	45.5	45.6

GI number of the proteins and species	Genome/Nucleotide accession	GC content of genome	GC content of sequence
944520030 Agarivorans gilvus WH0801	NZ_CP013021.1	46	45.4
1062696832 Bacillus beveridgei MLTeJB	CP012502.1	46.1	48.4
1063263829 Bacillus beveridgei MLTeJB	CP012502.1	46.1	48.4
505943926 Vibrio vulnificus MO6-24/O	NC_014966.1	46.7	51
499394716 Vibrio vulnificus FORC_017	NZ_CP012740.1	46.9	51
983353765 Vibrio vulnificus FORC_016	NZ_CP011776.1	47	51.2
749197755 Vibrio vulnificus 93U204	NZ_CP009262.1	47.1	51.3
992353306 Vibrio vulnificus ATL 6-1306	NZ_CP014048.1	47.3	50.9
765293668 Photobacterium gaetbulicola Gung47	NZ_CP005973.1	47.9	50.6
502937280 Bacillus selenitireducens MLS10	NC_014219.1	48.7	49.9
1079276783 Anaerolinea thermophila UNI-1	NC_014960.1	49.85	46.6
757794123 Vibrio coralliilyticus RE98	NZ_CP009620.1	50.2	47.2
609224323 Salinispira pacificaL21-RPul-D2	NC_023035.1	51.9	51
551029991 Candidatus Symbiobacter mobilis CR	NC_022576.1	59.1	62.5
497234730 Burkholderiales bacterium JOSHI_001	NZ_CM001438.1	68.8	70.9
502880848 Cellulomonas flavigena DSM 20109	NC_014151.1	70.4	73.4
503648288 Cellulomonas gilvus ATCC 13127	NC_015671.1	73.8	73.2
503535794 Cellulomonas fimi ATCC 484,	NC_015514.1	74.7	73.9

Appendix 7. Sequence ID of GHyyy members

App. 7.1. Bacterial GHyyy sequences from NCBI

RefSeq ID	Species and strain	Phylum	Nucleotide ID	Isolation source
ERJ13626.1	Haloplasma contractile SSD-17B	Haloplasmatales	AFNU02000001.1	sediment-brine interface of the Shaban Deep in the Red Sea
EKQ57686.1	Clostridium sp. Maddingley MBC34-26	Firmicutes	ALXI01000016.1	coal seam gas formation water, 100m depth water
EKQ50390.1	Clostridium sp. Maddingley MBC34-26 1	Firmicutes	ALXI01000175.1	coal seam gas formation water, 100m depth water
BAM00492.1	Caldilinea aerophila DSM 14535 = NBRC 104270	Chloroflexi	AP012337.1	no data
GAP62125.1	Ardenticatena maritima 110S	Chloroflexi	BBZA01000033.1	sediment, hydrothermal beach
ABP68120.1	Caldicellulosiruptor saccharolyticus DSM 8903	Firmicutes	CP000679.1	hot spring in New Zealand
AIQ33578.1	Paenibacillus sp. FSL R5-0345	Firmicutes	CP009281.1	pasteurized 2% milk
AQQ70274.1	Phycisphaerae bacterium SM-Chi-D1	Planctomycetes	CP019646.1	sediment of evaporation pond
CUN48701.1	Roseburia hominis 2789STDY5608834	Firmicutes	CYZJ01000001.1	huamn feces
CUX60029.1	Clostridium sp. C105KSO15 1	Firmicutes	FBWN01000177.1	no data
SCX05664.1	Lachnospiraceae bacterium YSD2013	Firmicutes	FMTX01000009.1	rumen microbiome
SDB19587.1	Pseudobutyrivibrio sp. YE44	Firmicutes	FMXT01000002.1	rumen microbiome
SDC24037.1	Pelagirhabdus alkalitolerans S5	Firmicutes	FMYI01000005.1	no data
SDF19476.1	Fontibacillus panacisegetis DSM 28129	Firmicutes	FNBG01000007.1	No data
SDH17459.1	Fervidobacterium changbaicum DSM 17883	Thermotogae	FNDL01000007.1	No data
SDI45081.1	Paenibacillus typhae CGMCC 1.11012	Firmicutes	FNDX01000005.1	type strain, soil and plant-assoicated
SDY04365.1	Bacillus caseinilyticus SP	Firmicutes	FNPI01000001.1	cultured
SEA82830.1	Pseudobutyrivibrio sp. ACV-2	Firmicutes	FNQZ01000012.1	no data
SEG11996.1	Caloramator fervidus DSM 5463	Firmicutes	FNUK01000032.1	New Zealand:Rotorua Park
SEF83514.1	Vibrio hangzhouensis CGMCC 1.7062	Proteobacteria	FNVG01000004.1	no data
SEJ83912.1	Paenibacillus sp. 181mfcol5.1	Firmicutes	FNYX01000005.1	plant associated
SEK23177.1	Paenibacillus sp. cl141a	Firmicutes	FNZO01000001.1	no data
CCV65814.1	Acholeplasma brassicae O502	Tenericutes	FO681348.1	plant associated
SEL48476.1	Alkalibacterium putridalgicola DSM 19182	Firmicutes	FOBL01000002.1	no data
SEO89603.1	Amphibacillus marinus CGMCC 1.10434	Firmicutes	FODJ01000016.1	cultured
SES15713.1	Salipaludibacillus aurantiacus S9	Firmicutes	FOGT01000009.1	no data

RefSeq ID	Species and strain	Phylum	Nucleotide ID	Isolation source
SER69947.1	Salisediminibacterium haloalkalitolerans 10nlg	Firmicutes	FOGV01000004.1	no data
SEU24299.1	Paenibacillus sp. NFR01	Firmicutes	FOIK01000005.1	root endophytes of switch grass
SFC11500.1	Alkalibacterium subtropicum DSM 23664	Firmicutes	FOLT01000003.1	no data
SFH69345.1	Pseudobutyrivibrio sp. OR37	Firmicutes	FOQF01000005.1	no data
SFN80781.1	Pseudobutyrivibrio sp. JW11	Firmicutes	FOWB01000001.1	no data
SFN98921.1	Anaerocolumna aminovalerica DSM 1283	Firmicutes	FOWD01000006.1	no data
SFO78737.1	Halolactibacillus alkaliphilus CGMCC 1.6843	Firmicutes	FOWN01000011.1	no data
SFP69713.1	Halolactibacillus halophilus DSM 17073	Firmicutes	FOXC01000044.1	Japan: Oura beach, Miura Peninsula, Kanagawa Prefecture
CBK73627.1	Butyrivibrio fibrisolvens 16/4	Firmicutes	FP929036.1	no data
SFS93792.1	Halolactibacillus miurensis DSM 17074	Firmicutes	FPAI01000019.1	Japan: Oura beach, Miura Peninsula, Kanagawa Prefecture
SFT60650.1	Lachnospiraceae bacterium XBD2001 1	Firmicutes	FPAN01000008.1	no data
CDA29483.1	Eubacterium sp. CAG:156	Firmicutes	FR878154.1	Human GI tract
CDB14106.1	Eubacterium sp. CAG:192 1	Firmicutes	FR883228.1	Human GI tract
CDF46130.1	Roseburia sp. CAG:100	Firmicutes	FR904080.1	human GI tract
CCX93064.1	Firmicutes bacterium CAG:110	Firmicutes	HF988006.1	Human GI tract
CCX92044.1	Firmicutes bacterium CAG:110 1	Firmicutes	HF988115.1	Human GI tract
CCY30928.1	Roseburia inulinivorans CAG:15	Firmicutes	HF989932.1	human GI tract
CCY69845.1	Eubacterium sp. CAG:161	Firmicutes	HF991851.1	human GI tract
CCZ09292.1	Clostridium sp. CAG:127 1	Firmicutes	HF993829.1	human GI tract
EGW21636.1	Methylobacter tundripaludum SV96 1	Proteobacteria	JH109152.1	arctic wetland soil
KLU72387.1	Robinsoniella sp. RHS	Firmicutes	JNGB01000015.1	mouse fecal sample colonised with soil-derived microbiota
KHD15380.1	Clostridium butyricum NOR 33234		JSEG01000009.1	human stool
BAU78235.1	Paenibacillus polymyxa NBRC 15309	Firmicutes	LC140880.1	no data
KUK24060.1	Thermotoga sp. 50_1627	Thermotogae	LGFH01000067.1	oil reservoir sample K2
KOS03742.1	Paenibacillus polymyxa YUPP-8	Firmicutes	LIYK01000254.1	associated with cotton plant
ONI39876.1	Epulopiscium sp. SCG-B11WGA-EpuloA1	Firmicutes	LJDB01000060.1	Surgeonfish GI tract
KPK04798.1	Gemmatimonas sp. SG8_28	Gemmatimonadetes	LJNQ01000117.1	Sulfate-rich zone estuary sediments 8-12 cm

RefSeq ID	Species and strain	Phylum	Nucleotide ID	Isolation source
KPK42408.1	Omnitrophica WOR_2 bacterium SM23_29	Candidatus Omnitrophica	LJUB01000015.1	Sulfate-methane transisition zone estuary sediments 16-26 cm
KPK41404.1	Omnitrophica WOR 2 bacterium SM23 29	Candidatus Omnitrophica	LJUB01000051.1	sediment metagenome, Sulfate-methane transisition zone estuary sediments 16-26 cm
KPK99010.1	Omnitrophica WOR 2 bacterium SM23 72	Candidatus Omnitrophica	LJUU01000004.1	sediment metagenome, Sulfate-methane transisition zone estuary sediments 16-26 cm
KPK97305.1	Omnitrophica WOR 2 bacterium SM23 72	Candidatus Omnitrophica	LJUU01000045.1	sediment metagenome, Sulfate-methane transisition zone estuary sediments 16-26 cm
KPK96707.1	Omnitrophica WOR 2 bacterium SM23 72	Candidatus Omnitrophica	LJUU01000084.1	sediment metagenome, Sulfate-methane transisition zone estuary sediments 16-26 cm
OOB78077.1	Epulopiscium sp. Nuni2H MBin001	Firmicutes	MDJM01000060.1	Hindgut of marine fish (Naso unicornis)
OFW62009.1	Actinobacteria bacterium RBG_19FT_COMBO_36_27	Actinobacteria	MELT01000140.1	Subsurface metagenome,Rifle background sediment; well D04 at 19ft depth
OGB89810.1	candidate division WOR-1 bacterium RIFCSPHIGHO2 01 FULL 53 15 1	unclassified	METM01000020.1	subsurface metagenome
OGB86825.1	candidate division WOR-1 bacterium RIFCSPHIGHO2_02_FULL_45_12	unclassified	METN01000072.1	subsurface metagenome, Rifle well FP-101 under high O2 conditions;
OGC07941.1	candidate division WOR-1 bacterium RIFCSPLOWO2 12 FULL 45 9 1	unclassified	METQ01000065.1	subsurface metagenome, Rifle well FP-101 under low O2 conditions; 1.2 micron filter; full assembly
OGC23581.1	candidate division WOR-1 bacterium RIFOXYA2_FULL_41_14	unclassified	METW01000020.1	subsurface metagenom, Rifle well CD01 at 16ft depth
OGC11849.1	candidate division WOR-1 bacterium RIFOXYA2_FULL_51_19	unclassified	METY01000001.1	subsurface metagenome, Rifle well CD01 at 16ft depth
OGC33960.1	candidate division WOR-1 bacterium RIFOXYB2 FULL 48 7 1	unclassified	MEUF01000052.1	Subsurface metagenome, Rifle well CD01 at 16ft depth; 0.2 micron filter at time point B
OGC42363.1	candidate division WOR-1 bacterium RIFOXYD2_FULL_41_8	unclassified	MEUL01000086.1	subsurface metagenome, Rifle well CD01 at 16ft depth

RefSeq ID	Species and strain	Phylum	Nucleotide ID	Isolation source
OGV37168.1	Lentisphaerae bacterium GWF2 49 21	Lentisphaerae	MHBH01000073.1	subsurface metagenome Rifle well CD01 at time point 6 / F; 5m depth; 0.2 filter
OGV52597.1	Lentisphaerae bacterium GWF2_50_93	Lentisphaerae	MHBI01000118.1	subsurface metagenome, Rifle well CD01 at time point 6 / F; 5m depth
OGV45644.1	Lentisphaerae bacterium GWF2 57 35]	Lentisphaerae	MHBK01000025.1	subsurface metagenome,Rifle well CD01 at time point 6 / F; 5m depth
OGW91974.1	Omnitrophica bacterium RIFCSPHIGHO2 02 FULL 63 14	Candidatus Omnitrophica	MHFJ01000017.1	subsurface metagenome, Rifle well FP-101 under high O2 conditions; 0.2 micron filter
OGX01506.1	Omnitrophica WOR 2 bacterium GWA2 44 7	Candidatus Omnitrophica	MHFW01000038.1	subsurface metagenome, Rifle well CD01 at time point 1 / A; 5m depth
OGX07258.1	Omnitrophica WOR 2 bacterium GWA2 47 8	Candidatus Omnitrophica	MHFY01000094.1	subsurface metagenom, Rifle well CD01 at time point 1 / A; 5m depth; 0.2 filter
OGX11645.1	Omnitrophica WOR 2 bacterium GWB2 45 9	Candidatus Omnitrophica	MHGB01000028.1	subsurface metagenome,Rifle well CD01 at time point 2 / B; 5m depth
OGX23872.1	Omnitrophica WOR_2 bacterium GWF2_38_59	Candidatus Omnitrophica	MHGE01000056.1	subsurface metagenome, Rifle well CD01 at time point 6 / F; 5m depth
OGX31245.1	Omnitrophica WOR 2 bacterium RIFCSPHIGHO2 01 FULL 48 9	Candidatus Omnitrophica	MHGK01000100.1	subsurface metagenome, Rifle well FP-101 under high O2 conditions
OGX25981.1	Omnitrophica WOR 2 bacterium RIFCSPHIGHO2 02 FULL 45 21	Candidatus Omnitrophica	MHGN01000023.1	subsurface metagenome, Rifle well FP-101 under high O2 conditions; 0.2 micron filter; full assembly
OGX24948.1	Omnitrophica WOR 2 bacterium RIFCSPHIGHO2 02 FULL 48 11	Candidatus Omnitrophica	MHGP01000073.1	subsurface metagenome, Rifle well FP-101 under high O2 conditions
OGX39498.1	Omnitrophica WOR_2 bacterium RIFCSPHIGHO2_02_FULL_50_17	Candidatus Omnitrophica	MHGQ01000014.1	subsurface metagenome, Rifle well FP-101 under high O2 conditions
OGX37135.1	Omnitrophica WOR_2 bacterium RIFCSPHIGHO2_02_FULL_52_10	Candidatus Omnitrophica	MHGR01000042.1	subsurface metagenome, Rifle well FP-101 under high O2 conditions
OGX34577.1	Omnitrophica WOR 2 bacterium RIFCSPLOWO2 02 FULL 50 19	Candidatus Omnitrophica	MHGY01000049.1	subsurface metagenome, Rifle well FP-101 under low O2 conditions

RefSeq ID	Species and strain	Phylum	Nucleotide ID	Isolation source
OGX43336.1	Omnitrophica WOR_2 bacterium RIFCSPLOWO2_12_FULL_51_24	Candidatus Omnitrophica	MHHC01000020.1	subsurface metagenome, Rifle well FP-101 under low O2 conditions
OGX46482.1	Omnitrophica WOR_2 bacterium RIFOXYA2_FULL_45_12	Candidatus Omnitrophica	MHHH01000060.1	subsurface metagenome, Rifle well CD01 at 16ft depth
OGX61342.1	Omnitrophica WOR 2 bacterium RIFOXYC2 FULL 45 15	Candidatus Omnitrophica	MHHM01000014.1	subsurface metagenome, Rifle well CD01 at 16ft depth
OHB53795.1	Planctomycetes bacterium GWF2_41_51	Planctomycetes	MHXZ01000005.1	subsurface metagenome, Rifle well CD01 at time point 6 / F; 5m depth
OHE28602.1	Tenericutes bacterium GWC2 34 14	Tenericutes	MICR01000012.1	subsurface metagenome, Rifle well CD01 at time point 3 / C; 5m depth
OHE30965.1	Tenericutes bacterium GWC2 39 45	Tenericutes	MICS01000018.1	subsurface metagenome, Rifle well CD01 at time point 3 / C; 5m depth
OJV66793.1	Clostridiales bacterium 38-18	Firmicutes	MKTL01000001.1	biofilm in bioreactor
OIO35236.1	Candidatus Omnitrophica bacterium CG1 02 44 16 1	unclassified	MNVR01000035.1	groundwater filtered through a a 3.0 um filter and then through a 0.2 um filter
OIO38639.1	Candidatus Omnitrophica bacterium CG1 02 49 16 1	unclassified	MNVU01000047.1	groundwater filtered through a a 3.0 um filter and then through a 0.2 um filter
OIP13455.1	Comamonadaceae bacterium CG2 30 57 122 1	Proteobacteria	MNXW01000309.1	groundwater filtered through a 3.0 um filter
OMC64708.1	Paenibacillus odorifer FSL F4-0077	Firmicutes	MPVO01000008.1	milk
OMG47019.1	Paenibacillus macerans FSL R5-0527	Firmicutes	MRTR01000036.1	milk
WP_062329571.1	Methylomonas	Proteobacteria	multispecies (remove)	
WP_041722924.1	Caldicellulosiruptor saccharolyticus DSM 8903	Firmicutes	NC_009437.1	hot spring, New Zealand
WP_011994121.1	Fervidobacterium nodosum Rt17-B1	Thermotogae	NC_009718.1	no data
WP_012198896.1	Lachnoclostridium phytofermentans ISDg	Firmicutes	NC_010001.1	no data
ABX81332.1	Acholeplasma laidlawii PG-8A	Tenericutes	NC_010163.1	no data
WP_012547574.1	Dictyoglomus thermophilum H-6-12	Dictyoglomi	NC_011297.1	no data
WP_004103012.1	Thermosipho africanus TCF52B 1	Thermotogae	NC_011653.1	Troll oil formation in the North Sea
ACK41445.1	Dictyoglomus turgidum DSM 6724	Dictyoglomi	CP001251.1	Type strain
WP_012635372.1	Halothermothrix orenii H168	Firmicutes	NC_011899.1	sediment of a salt lake in Tunisia
WP_015848606.1	Thermococcus sibiricus MM 739	Euryarchaeota	NC_012883.1	high-temperature Samotlor oil reservoir
WP_015736393.1	Paenibacillus sp. Y412MC10	Firmicutes	NC_013406.1	Obsidian Hot Spring in Yellowstone National Park

RefSeq ID	Species and strain	Phylum	Nucleotide ID	Isolation source
WP_012958999.1	Bacillus pseudofirmus OF4	Firmicutes	NC_013791.2	soil, New york
WP_013129904.1	Thermosphaera aggregans DSM 11486	Crenarchaeota	NC_014160.1	Hot spring from Obsidian Pool at Yellowstone Natl. Park
WP_013171295.1	Bacillus selenitireducens MLS10	Firmicutes	NC_014219.1	sediment
WP_013272305.1	Clostridium saccharolyticum WM1	Firmicutes	NC_014376.1	sewage sludge, canada
WP_013289460.1	Caldicellulosiruptor obsidiansis OB47	Firmicutes	NC_014392.1	Obsidian Pool, Yellowstone National Park, Wyoming, USA
WP_013311628.1	Paenibacillus polymyxa E681	Firmicutes	NC_014483.2	plant rhizosphere
WP_013429146.1	Caldicellulosiruptor kronotskyensis 2002	Firmicutes	NC_014720.1	terrestrial neutral geothermal spring
WP_013487305.1	Bacillus cellulosilyticus DSM 2522	Firmicutes	NC_014829.1	no data
WP_013658929.1	Cellulosilyticum lentocellum DSM 5427	Firmicutes	NC_015275.1	river sediment with paper mill waste
WP_014079648.1	Roseburia hominis A2-183	Firmicutes	NC_015977.1	human GI tract
BAM00492.1	Caldilinea aerophila DSM 14535 = NBRC 104270	Chloroflexi	AP012337.1	
WP_014451045.1	Fervidobacterium pennivorans DSM 9078	Thermotogae	NC_017095.1	sediment hotspring
WP_014454798.1	Spirochaeta africana DSM 8902	Spirochaetes	NC_017098.1	bacterial bloom in the brine under trona from alkaline equatorial lake
WP_014641703.1	Halobacillus halophilus NC_017668.1	Firmicutes	NC_017668.1	no data
WP_015325999.1	Halobacteroides halobius DSM 5150	Firmicutes	NC_019978.1	dead sea sediment
WP_008525786.1	Halorhabdus tiamatea SARL4B	Euryarchaeota	NC_021921.1	a deep-sea anoxic brine lake
WP_071386948.1	Acholeplasma brassicae	Tenericutes	NC_022549.1	no data (record removed)
WP_023989977.1	Paenibacillus polymyxa CR1	Firmicutes	NC_023037.2	maize rhizosphere soil
OGW98683.1	Omnitrophica WOR_2 bacterium GWA2_45_18	Candidatus Omnitrophica	no data	subsurface metagenome, Rifle well CD01 at time point 1 / A; 5m depth
OGX48084.1	Omnitrophica WOR_2 bacterium RIFCSPLOWO2_12_FULL_63_16	Candidatus Omnitrophica	no data	subsurface metagenome, Rifle well FP-101 under low O2 conditions
WP_016313215.1	Paenibacillus barengoltzii FSL H7-0744	Firmicutes	no data	no data
WP_042210248.1	Paenibacillus borealis	Firmicutes	no data	no data
WP_033191227.1	Fervidobacterium islandicum	Thermotogae	no data	no data
WP_051615820.1	Acholeplasma modicum	Tenericutes	none	no data
WP_009860980.1	Roseburia inulinivorans DSM 16841	Firmicutes	NZ_ACFY01000165.1	human GI tract
WP_008826269.1	Haloplasma contractile SSD-17B	Haloplasmatales	NZ_AFNU02000001.1	sediment-brine interface of the Shaban Deep in the Red Sea

RefSeq ID	Species and strain	Phylum	Nucleotide ID	Isolation source
WP_016818628.1	Paenibacillus polymyxa OSY-DF	Firmicutes	NZ_AIPP01000116.1	fermented vegetable (Kimchee) from a local store
WP_039834362.1	Paenibacillus sonchi X19-5	Firmicutes	NZ_AJTY01000375.1	rhizosphere soil of Sonchus oleraceus (sow thistle)
WP_019421873.1	Paenibacillus sp. OSY-SE	Firmicutes	NZ_ALKF01000184.1	soil
WP_004627615.1	Clostridium termitidis CT1112	Firmicutes	NZ_AORV01000046.1	termite hindgut
WP_003424021.1	Clostridium butyricum DKU-01	Firmicutes	NZ_APKZ01000007.1	infant feces
WP_020481102.1	Methylomonas sp. MK1	Proteobacteria	NZ_AQOV01000001.1	Lake Wahsington sediment
WP_006313788.1	Caldisalinibacter kiritimatiensis L21-TH-D2	Firmicutes	NZ_ARZA01000185.1	no data
WP_045143732.1	Clostridium butyricum CWBI1009	Firmicutes	NZ_ASPQ01000014.1	anaerobic sludge
WP_036702107.1	Paenibacillus sp. FSL R7-269	Firmicutes	NZ_ASPS01000144.1	milk
WP_063829815.1	Paenibacillus sp. FSL H8-237	Firmicutes	NZ_ASPV01000054.1	Pasteurised milk
WP_036663991.1	Paenibacillus sp. FSL H8-457	Firmicutes	NZ_ASPW01000045.1	pasteurised milk
WP_036732712.1	Paenibacillus sp. FSL R7-277	Firmicutes	NZ_ASPX01000119.1	milk
WP_025721489.1	Paenibacillus polymyxa TD94	Firmicutes	NZ_ASSA01000171.1	associated with scutellaria (mint family)
WP_025679076.1	Paenibacillus massiliensis T7	Firmicutes	NZ_ASSE01000098.1	isolated from Willow rhizosphere
WP_022766121.1	Butyrivibrio sp. XPD2006	Firmicutes	NZ_ATVT01000012.1	no data
WP_022787627.1	Clostridiales bacterium NK3B98	Firmicutes	NZ_ATWA01000053.1	rumen microbiome
WP_022767451.1	Butyrivibrio sp. NC2007	Firmicutes	NZ_ATWY01000002.1	Rumen microbiome
WP_024346902.1	Clostridium methoxybenzovorans SR3	Firmicutes	NZ_ATXD01000008.1	soil
WP_029228890.1	Caldicellulosiruptor acetigenus DSM 7040	Firmicutes	NZ_ATXO01000014.1	hotspring water
WP_035375542.1	Acholeplasma axanthum ATCC 25176	Tenericutes	NZ_AUAL01000004.1	no data
WP_028591106.1	Paenibacillus panacisoli DSM 21345	Firmicutes	NZ_AUFO01000019.1	Soil of a ginseng field
WP_035775095.1	Butyrivibrio sp. VCB2001	Firmicutes	NZ_AUIZ01000001.1	no data
WP_036951784.1	Pseudobutyrivibrio ruminis CF1b	Firmicutes	NZ_AUJB01000008.1	rumen microbiome
WP_027636584.1	Clostridium butyricum AGR2140	Firmicutes	NZ_AUJN01000016.1	rumen microbiome
WP_035779225.1	Butyrivibrio sp. MC2013	Firmicutes	NZ_AUKE01000011.1	no data
WP_036818050.1	Pontibacillus yanchengensis Y32	Firmicutes	NZ_AVBF01000015.1	salt field
WP_036781682.1	Pontibacillus chungwhensis BH030062	Firmicutes	NZ_AVBG01000003.1	solar saltern (salt field)
WP_027308906.1	Caloramator sp. ALD01	Firmicutes	NZ_AXAA01000020.1	Etoliko lagoon in Western Greece
WP_024291666.1	Clostridium indolis DSM 755	Firmicutes	NZ_AZUI01000001.1	soil

RefSeq ID	Species and strain	Phylum	Nucleotide ID	Isolation source
WP_034563905.1	Hungatella hathewayi VE202-04	Firmicutes	NZ_BAHR02000231.1	tarting from a healthy human faecal sample, we obtained gnotobiotic mice colonized with a human microbiota enriched in Treg-inducing strains. From these mice, we cultured and selected strains of bacteria, and performed genome sequencing.
WP_025532029.1	Hungatella hathewayi	Firmicutes	NZ_BAHY01000109.1	human GI tract
WP_041966187.1	Bacillus selenatarsenatis SF-1	Firmicutes	NZ_BASE01000056.1	no data
WP_067842620.1	Amphibacillus sediminis NBRC 103570	Firmicutes	NZ_BCQW01000021.1	cultured
WP_066420707.1	Bacillus cohnii NBRC 15565	Firmicutes	NZ_BCUW01000111.1	culture collection
WP_008909277.1	Caloramator australicus RC3	Firmicutes	NZ_CAKP01000100.1	no data
WP_005490316.1	Halanaerobium saccharolyticum type strain:DSM 6643	Firmicutes	NZ_CAUI01000023.1	type strain
WP_040952469.1	Gorillibacterium massiliense strain:G5	Firmicutes	NZ_CBQR020000153.1	no data
WP_040978143.1	Oceanobacillus jeddahense S5	Firmicutes	NZ_CCDM010000001.1	human GI tract
WP_054955257.1	Paenibacillus sp. FF9	Firmicutes	NZ_CDSE01000011.1	no data
WP_009168511.1	Clostridium sp. DL-VIII 1	Firmicutes	NZ_CM001240.1	anaerobic sludge
WP_005370277.1	Methylomicrobium album BG8	Proteobacteria	NZ_CM001475.1	muddy water
WP_062370841.1	Thermococcus guaymasensis DSM 11113	Euryarchaeota	NZ_CP007140.1	no data
WP_038585252.1	Paenibacillus sp. FSL H7-0357	Firmicutes	NZ_CP009241.1	Pasteurized 2% milk
WP_042184448.1	Paenibacillus sp. FSL H7-0737	Firmicutes	NZ_CP009279.1	pasteurized 2% milk
WP_042133118.1	Paenibacillus sp. FSL P4-0081	Firmicutes	NZ_CP009280.1	heat treated raw milk
WP_042131802.1	Paenibacillus sp. FSL R5-0345	Firmicutes	NZ_CP009281.1	Pasteurized 2% fat milk
WP_042232123.1	Paenibacillus sp. FSL R5-0912	Firmicutes	NZ_CP009282.1	Pasteurized 2% fat milk
WP_039869333.1	Paenibacillus sp. FSL R7-0273	Firmicutes	NZ_CP009283.1	Pasteurized 2% fat milk
WP_042172399.1	Paenibacillus sp. FSL R7-0331	Firmicutes	NZ_CP009284.1	Pasteurized 2% fat milk
WP_038698858.1	Paenibacillus stellifer DSM 14472	Firmicutes	NZ_CP009286.1	food-packaging paperboard
WP_025707431.1	Paenibacillus graminis DSM 15220	Firmicutes	NZ_CP009287.1	maize rhizospher soil
WP_038568482.1	Paenibacillus odorifer DSM 15391	Firmicutes	NZ_CP009428.1	soil from wheat roots
WP_052570529.1	Endomicrobium proavitum Rsa215	Elusimicrobia	NZ_CP009498.1	no data
WP_025365657.1	Paenibacillus polymyxa CF05	Firmicutes	NZ_CP009909.1	associated with Cryptomeria fortunei (Japanese red cedar)
WP_044877319.1	Paenibacillus sp. IHBB 10380	Firmicutes	NZ_CP010976.1	soil glacier

RefSeq ID	Species and strain	Phylum	Nucleotide ID	Isolation source
WP_064011647.1	Fervidobacterium pennivorans DYC	Thermotogae	NZ_CP011393.1	hotspring, New Zealand (capable of cellobiose fermentation)
WP_075154651.1	Paenibacillus polymyxa ATCC 15970	Firmicutes	NZ_CP011420.1	soil
WP_053325977.1	Paenibacillus peoriae HS311	Firmicutes	NZ_CP011512.1	soil
WP_069366220.1	Bacillus beveridgei MLTeJB	Firmicutes	NZ_CP012502.1	no data
WP_058146710.1	Clostridium butyricum KNU0-L09	Firmicutes	NZ_CP013252.1	no data
WP_064021676.1	Methylomonas sp. DH-1	Proteobacteria	NZ_CP014360.1	waste sludge of a brewery plant
WP_072580822.1	Bacillus weihaiensis Alg07	Firmicutes	NZ_CP016020.1	marine,china
WP_068660028.1	Paenibacillus sp. LPB0068	Firmicutes	NZ_CP017770.1	associated with Crassostrea gigas (pacific oyster)
WP_071558349.1	Paenibacillus polymyxa YC0573	Firmicutes	NZ_CP017968.2	rhizosphere of tobacco
WP_068782371.1	Paenibacillus sp. GM2	Firmicutes	NZ_CTEK01000044.1	no data
WP_055167725.1	Roseburia inulinivorans 2789STDY5608887	Firmicutes	NZ_CYXX01000003.1	human faeces
WP_005358765.1	Eubacterium ventriosum ATCC 27560	Firmicutes	NZ_DS264265.1	Human GI tract
WP_005363330.1	Eubacterium ventriosum ATCC 27560	Firmicutes	NZ_DS264285.1	Human GI tract
WP_068787086.1	Paenibacillus phoceensis MT24	Firmicutes	NZ_FCOQ01000036.1	human GI tract
WP_066716675.1	Clostridium sp. Marseille-P299 1	Firmicutes	NZ_FJVE01000007.1	human GI tract
WP_062525136.1	Paenibacillus jilunlii CGMCC 1.10239	Firmicutes	NZ_FNGM01000002.1	no data
WP_074755796.1	Butyrivibrio fibrisolvens AR40	Firmicutes	NZ_FOGJ01000010.1	cultured
WP_074912609.1	Proteiniclasticum ruminis strain ML2	Firmicutes	NZ_FOVK01000010.1	no data
WP_073298186.1	Atopostipes suicloacalis DSM 15692	Firmicutes	NZ_FQUF01000021.1	cultured
WP_073248831.1	Caloramator proteoclasticus DSM 10124	Firmicutes	NZ_FQVG01000027.1	no data
WP_073388888.1	Butyrivibrio fibrisolvens DSM 3071	Firmicutes	NZ_FQXK01000027.1	cultured
WP_073338818.1	Clostridium grantii DSM 8605	Firmicutes	NZ_FQXM01000014.1	no data
WP_072916440.1	Pseudobutyrivibrio xylanivorans DSM 14809	Firmicutes	NZ_FQYQ01000010.1	no data
WP_073275368.1	Anaerocolumna jejuensis DSM 15929	Firmicutes	NZ_FRAC01000010.1	no data
WP_073287569.1	Anaerosporobacter mobilis DSM 15930	Firmicutes	NZ_FRCP01000011.1	no data
WP_072757930.1	Fervidobacterium gondwanense DSM 13020	Thermotogae	NZ_FRDJ01000002.1	Australia:Great Artesian Basin of Australia
WP_073590012.1	Anaerocolumna xylanovorans DSM 12503	Firmicutes	NZ_FRFD01000010.1	cultured
WP_034546778.1	Carnobacterium alterfunditum 313	Firmicutes	NZ_FSRN01000001.1	Antarctica: Ace Lake
WP_076544684.1	Halanaerobium kushneri strain ATCC 700103	Firmicutes	NZ_FTNC01000008.1	no data

RefSeq ID	Species and strain	Phylum	Nucleotide ID	Isolation source
WP_068592622.1	Paenibacillus macquariensis ATCC 23464	Firmicutes	NZ_FTNK01000017.1	no data
WP_009226247.1	Paenibacillus sp. oral taxon 786	Firmicutes	NZ_GG695987.1	assoicated with human
WP_019688419.1	Paenibacillus polymyxa ATCC842	Firmicutes	NZ_GL905390.1	type strain
WP_018666383.1	Thermobrachium celere DSM 8682	Firmicutes	NZ_HF952039.1	hot spring sediment and water in New Zealand
WP_044477328.1	Paenibacillus antibioticophila GD11	Firmicutes	NZ_HG005139.1	human GI tract
WP_028539026.1	Paenibacillus sp. J14	Firmicutes	NZ_JADQ01000027.1	rhizosphere (lignin degrading bacteria)
WP_024834966.1	Clostridium sp. 12(A)	Firmicutes	NZ_JAGI01000001.1	no data
WP_024865374.1	Butyrivibrio sp. FCS014	Firmicutes	NZ_JAHA01000006.1	Rumen microbiome
WP_034301206.1	Alkalibacterium sp. AK22	Firmicutes	NZ_JANL01000029.1	Lonar Lake
WP_032550702.1	Vibrio fortis Dalian14	Proteobacteria	NZ_JFFR01000013.1	associated with Strongylocentrotus intermedius (sea urchins)
WP_068345671.1	Kosmotoga arenicorallina S304	Thermotogae	NZ_JFHK01000002.1	submarine hot spring shallow hydrothermal field
WP_071220018.1	Paenibacillus sp. LC231	Firmicutes	NZ_JFOM01000015.1	pristine cave enviroment (rock)
WP_040579857.1	Methylobacter tundripaludum SV96	Proteobacteria	NZ_JH109152.1	arctic wetland soil
WP_008676761.1	Clostridium sp. 7 2 43FAA	Firmicutes	NZ_JH815222.1	human GI tract
WP_026892135.1	Clostridium aerotolerans DSM 5434	Firmicutes	NZ_JHWJ01000010.1	sheep rumen
WP_027628574.1	Clostridium cellobioparum DSM 1351 = ATCC 15832	Firmicutes	NZ_JHYD01000024.1	cattle rumen
WP_026671722.1	Bacillus bogoriensis ATCC BAA-922	Firmicutes	NZ_JHYI01000006.1	no data
WP_027149823.1	Methylobacter tundripaludum 21/22	Proteobacteria	NZ_JMLA01000001.1	Lake Washington sediment
WP_028541566.1	Paenibacillus sp. UNCCL52	Firmicutes	NZ_JMLR01000008.1	plant associated
WP_031462091.1	Paenibacillus polymyxa CICC 10580	Firmicutes	NZ_JNCB01000003.1	Noni fruits Morinda citrifolia L from Paracel Islands
WP_031543443.1	Lachnospiraceae bacterium AC2014 1	Firmicutes	NZ_JNJE01000015.1	cow rumen
WP_031576719.1	Proteiniclasticum ruminis DSM 24773	Firmicutes	NZ_JNKC01000007.1	rumen microbiome
WP_033156181.1	Methylomonas sp. LW13 1	Proteobacteria	NZ_JNLB01000003.1	Lake Wahsington sediment
WP_033167163.1	Clostridium sp. KNHs205	Firmicutes	NZ_JNLE01000003.1	no data
WP_034443461.1	Butyrivibrio sp. AE2032	Firmicutes	NZ_JNLF01000003.1	cow rumen
WP_029501686.1	Lachnoclostridium phytofermentans KNHs2132	Firmicutes	NZ_JNLG01000001.1	no data
WP_052078517.1	Spirochaeta lutea JC230	Spirochaetes	NZ_JNUP01000048.1	soil
WP_038281003.1	Clostridium celerecrescens 152B	Firmicutes	NZ_JPME01000013.1	Sediment associated with methane hydrate from Krishna Godavari Basin
WP_031392740.1	Clostridium sp. KNHs209	Firmicutes	NZ_JPNB01000003.1	no data

RefSeq ID	Species and strain	Phylum	Nucleotide ID	Isolation source
WP_031438254.1	Methylobacter tundripaludum 31/32	Proteobacteria	NZ_JPOH01000001.1	no data
WP_036650681.1	Paenibacillus wynnii DSM 18334	Firmicutes	NZ_JQCR01000002.1	Antarctic soil
WP_044294592.1	Robinsoniella peoriensis WT	Firmicutes	NZ_JTGN01000004.1	Wood turtle fecaes
WP_047389514.1	Firmicutes bacterium ZOR0006	Firmicutes	NZ_JTLC01000064.1	no data
WP_039309715.1	Paenibacillus sp. IHB B 3415	Firmicutes	NZ_JUEI01000119.1	soil from cold environments
WP_058486419.1	Defluviitalea phaphyphila Alg1	Firmicutes	NZ_JWID01000011.1	marine sediment
WP_043852479.1	Clostridium butyricum HM-68	Firmicutes	NZ_JXBT01000001.1	healthy chicken intestine
WP_041902237.1	Bacillus thermoamylovorans B4166	Firmicutes	NZ_JXLT01000003.1	milk
WP_002603706.1	Hungatella hathewayi 12489931 1	Firmicutes	NZ_KB850951.1	human microbiome
WP_018751038.1	Paenibacillus sanguinis DSM 16941	Firmicutes	NZ_KB895378.1	human blood sample
WP_018885078.1	Paenibacillus massiliensis DSM 16942	Firmicutes	NZ_KB898189.1	human blood
WP_018247638.1	Orenia marismortui DSM 5156	Firmicutes	NZ_KB900617.1	salt-saturated mud, Israel
WP_019635978.1	Paenibacillus fonticola DSM 21315	Firmicutes	NZ_KB907247.1	freshwater, warm spring
WP_019913922.1	Paenibacillus sp. HW567	Firmicutes	NZ_KB910518.1	no data
WP_016169534.1	Lachnospiraceae bacterium 10-1	Firmicutes	NZ_KE159799.1	mouse ceca (beginning of large intestine)
WP_035771655.1	Butyrivibrio sp. FCS006	Firmicutes	NZ_KE384205.1	Rumen microbiome
WP_034488578.1	Butyrivibrio fibrisolvens MD2001	Firmicutes	NZ_KE384210.1	no data
WP_026691333.1	Bacillus aurantiacus DSM 18675	Firmicutes	NZ_KE386942.1	sediment, soda lake
WP_035212222.1	Bacillus chagannorensis DSM 18086	Firmicutes	NZ_KE386944.1	fresh water, soda lake
WP_026600659.1	Methylomonas sp. 11b	Proteobacteria	NZ_KI911557.1	no data
WP_027339873.1	Halonatronum saccharophilum DSM 13868	Firmicutes	NZ_KI912095.1	Coastal lagoon sediment; Lake Magad
WP_036623986.1	Paenibacillus macerans 8244	Firmicutes	NZ_KN125580.1	no data
WP_045173481.1	Caldicellulosiruptor sp. Wai35.B1	Firmicutes	NZ_LACM01000001.1	terrestrial hot springs at Yellowstone National Park
WP_045168667.1	Caldicellulosiruptor sp. Rt8.B8	Firmicutes	NZ_LACO01000001.1	terrestrial hot springs at Yellowstone National Park
WP_047001861.1	Clostridium sp. C8 1	Firmicutes	NZ_LBMX01000094.1	sediment sludge
WP_060626728.1	Paenibacillus etheri SH7	Firmicutes	NZ_LCZJ02000065.1	surface of hydrocarbon contaminated soil
WP_010174598.1	Bacillus coahuilensis m2-6	Firmicutes	NZ_LDYF01000006.1	water column (Churince system, cultured)
WP_059350157.1	Bacillus coahuilensis p1.1.43	Firmicutes	NZ_LDYG01000005.1	water column from Pozas Azules
WP_054492046.1	Ardenticatena maritima DSM23922	Chloroflexi	NZ_LGKN01000005.1	Coastal hydrothermal field in the Kirishima Volcanic Belt

RefSeq ID	Species and strain	Phylum	Nucleotide ID	Isolation source
WP_060862999.1	Paenibacillus riograndensis CAS34	Firmicutes	NZ_LIRB01000146.1	associated with rice (Oryza sativa)
WP_053489857.1	Bacillus sp. FJAT-18019	Firmicutes	NZ_LITP01000001.1	soil
WP_054400844.1	Paenibacillus sp. FJAT-22460	Firmicutes	NZ_LIUT01000001.1	soil
WP_045749352.1	Acholeplasma oculi 19L	Tenericutes	NZ_LK028559.1	no data
WP_050645896.1	Vibrio sp. J2-6	Proteobacteria	NZ_LK933653.1	no data
WP_045085406.1	Clostridium cellulosi DG5	Firmicutes	NZ_LM995447.1	Soil woodland
WP_054943147.1	Paenibacillus sp. GD6	Firmicutes	NZ_LN831201.1	no data
WP_020426902.1	Paenibacillus riograndensis SBR5	Firmicutes	NZ_LN831776.1	soil from wheat field
WP_053984600.1	Lachnospiraceae bacterium mt14	Firmicutes	NZ_LN875035.1	no data
WP_058257546.1	Herbinix luporum SD1D	Firmicutes	NZ_LN879430.1	no data
WP_058300329.1	Gorillibacterium sp. SN4	Firmicutes	NZ_LN881720.1	no data
WP_055105746.1	Paenibacillus ihumii AT5	Firmicutes	NZ_LN884297.1	human GI tract
WP_058296499.1	Clostridium neonatale	Firmicutes	NZ_LN890328.1	no data
WP_059051178.1	Paenibacillus sp. SIT18	Firmicutes	NZ_LN909045.1	no data
WP_059046741.1	Paenibacillus sp. MT18	Firmicutes	NZ_LN909066.1	no data
WP_071847807.1	Alkalibacterium sp. 20	Firmicutes	NZ_LSRN01000021.1	Antarctica, canary pond
WP_064008863.1	Methylomonas methanica R-45363	Proteobacteria	NZ_LUUG01000074.1	top layer of a denitrification tank of a waste water treatment plant
WP_064036102.1	Methylomonas methanica R-45371	Proteobacteria	NZ_LUUH01000037.1	top layer of a denitrification tank of a waste water treatment plant
WP_066976998.1	Methylomonas lenta R-45370	Proteobacteria	NZ_LUUI01000022.1	top layer of a denitrification tank of a waste water treatment plant
WP_064041461.1	Methylomonas koyamae R-45378	Proteobacteria	NZ_LUUJ01000100.1	Wetland
WP_064030858.1	Methylomonas koyamae R-45383	Proteobacteria	NZ_LUUK01000196.1	Wetland
WP_064029862.1	Methylomonas koyamae R-49807	Proteobacteria	NZ_LUUL01000127.1	facultative waste stabilization pond
WP_064212087.1	Acholeplasma laidlawii PG-8B	Tenericutes	NZ_LVCP01000008.1	plant associated
WP_068528722.1	Paenibacillus glacialis DSM 22343	Firmicutes	NZ_LVJH01000003.1	soil from altitude 3500m (Himalayas, Kafni glacier)
WP_068649783.1	Paenibacillus antarcticus CECT 5836 PBAT16	Firmicutes	NZ_LVJI01000016.1	sediment
WP_074101767.1	Paenibacillus sp. P3E	Firmicutes	NZ_LVWH01000160.1	rhizosphere of Helianthus annuus L. (sunflower)
WP_074107672.1	Paenibacillus sp. P26E	Firmicutes	NZ_LVWI01000038.1	rhizosphere of Helianthus annuus L. (sunflower)
WP_074086705.1	Paenibacillus sp. P32E	Firmicutes	NZ_LVYY01000024.1	rhizosphere of Helianthus annuus L (sunflower)
WP_074109980.1	Paenibacillus sp. P46E	Firmicutes	NZ_LVYZ01000002.1	rhizosphere of Helianthus annuus L. (sunflower)
WP_066173653.1	Bacillus sp. NC2-31 1	Firmicutes	NZ_LWBL02000040.1	marine sediment

RefSeq ID	Species and strain	Phylum	Nucleotide ID	Isolation source
WP_068717352.1	Orenia metallireducens Z6	Firmicutes	NZ_LWDV01000009.1	Deep subsurface groundwater
WP_063480048.1	Paenibacillus glucanolyticus SLM1	Firmicutes	NZ_LWMH01000002.1	black liquor from pulp mill stream
WP_069079905.1	Paenibacillus polymyxa CFSAN034342	Firmicutes	NZ_LYMY01000077.1	Agricultural soil
WP_066374772.1	Bacillus sp. FJAT-27264	Firmicutes	NZ_MAQW01000029.1	soil
WP_068500098.1	Paenibacillus kribbensis 6hRe76	Firmicutes	NZ_MCBP01000005.1	root of wild Cymbidium faberi (boat orchid)
WP_075679299.1	Roseburia sp. 831b	Firmicutes	NZ_MJHW01000002.1	domestic pig feces
WP_071159838.1	Methylomonas sp. LWB	Proteobacteria	NZ_MKMC01000045.1	Cave water and mat
WP_071308156.1	Anaerobacillus alkalilacustris DSM 18345	Firmicutes	NZ_MLQR01000001.1	bottom sediment, lake Khadyn, Russia
WP_076113870.1	Paenibacillus borealis	Firmicutes	NZ_MPTB01000052.1	milk
WP_076116623.1	Paenibacillus odorifer FSL H7-0710	Firmicutes	NZ_MPTC01000001.1	milk
WP_036682608.1	Paenibacillus odorifer FSL F4-0126	Firmicutes	NZ_MPTU01000007.1	milk
WP_076325048.1	Paenibacillus lautus FSL F4-0100	Firmicutes	NZ_MRTF01000009.1	milk
WP_076243564.1	Paenibacillus sp. FSL H8-0259	Firmicutes	NZ_MRTL01000004.1	milk
WP_076160827.1	Paenibacillus peoriae FSL H8-0551	Firmicutes	NZ_MRTM01000018.1	milk
WP_076150417.1	Paenibacillus glucanolyticus FSL R5-0817	Firmicutes	NZ_MRTT01000001.1	milk
WP_076083370.1	Paenibacillus sp. FSL R7-0333	Firmicutes	NZ_MRTW01000043.1	milk
WP_076160219.1	Paenibacillus sp. FSL R7-0337	Firmicutes	NZ_MRTX01000023.1	milk
WP_075682963.1	Bacillus pseudofirmus MRWL7	Firmicutes	NZ_MSLQ01000005.1	sediment
SDM08669.1	Chryseobacterium taihuense CGMCC 1.10941	Bacteroidetes	FNHD01000012.1	no data
WP_013402036.1	Caldicellulosiruptor hydrothermalis 108	Firmicutes	NC_014652.1	Terrestrial neutral geothermal spring; Russia, Kamchatka

App. 7.2. Heterokont GHyyy sequences from MMET

Sequence ID on MMET transcriptomes	Species and strain	Domain	Phylum
173148212	*Ochromonas sp.* Strain BG-1	Eukaryota	Ochrophyta
186488266	*Amphiprora sp*	Eukaryota	Bacillariophyta
186515412	*Amphiprora sp*	Eukaryota	Bacillariophyta
186521926	*Amphora coffeaeformis* CCMP127	Eukaryota	Bacillariophyta
186540750	*Amphora coffeaeformis* CCMP127	Eukaryota	Bacillariophyta
186651940	*Aureococcus anophagefferens* CCMP1850	Eukaryota	Bacillariophyta
186720192	*Aureoumbra lagunensis* CCMP1510	Eukaryota	Ochrophyta
187036080	*Chaetoceros affinis* CCMP159	Eukaryota	Bacillariophyta
187050904	*Chaetoceros curvisetus*	Eukaryota	Bacillariophyta
187175162	*Chattonella-subsalsa* CCMP2191	Eukaryota	Ochrophyta
187263766	*Dinobryon sp* UTEXLB2267	Eukaryota	Ochrophyta
187365798	*Ditylum brightwellii* GSO105	Eukaryota	Bacillariophyta
188154862	*Fragilariopsis kerguelensis* L26 C5	Eukaryota	Bacillariophyta
188556876	*Heterosigma akashiwo* CCMP2393	Eukaryota	Ochrophyta
188557734	*Heterosigma akashiwo* CCMP2393	Eukaryota	Ochrophyta
188604494	*Heterosigma akashiwo* CCMP3107	Eukaryota	Ochrophyta
188629668	*Heterosigma akashiwo* CCMP3107	Eukaryota	Ochrophyta
188639640	*Heterosigma akashiwo* CCMP452	Eukaryota	Ochrophyta
188654168	*Heterosigma-akashiwo* CCMP452	Eukaryota	Ochrophyta
190523074	*Pavlova sp CCMP459*	Eukaryota	Haptophyta
191215482	*Prymnesium parvum Texoma1*	Eukaryota	Haptophyta
191454482	*Pseudopedinella elastica* CCMP716	Eukaryota	Ochrophyta
192128108	*Skeletonema dohrnii SkelB*	Eukaryota	Bacillariophyta
192265400	*Skeletonema menzelii* CCMP793	Eukaryota	Bacillariophyta
192890408	*Thalassiosira oceanica* CCMP1005	Eukaryota	Bacillariophyta
192939680	*Thalassiosira oceanica* CCMP1005	Eukaryota	Bacillariophyta
192964706	*Thalassiosira rotula* CCMP3096	Eukaryota	Bacillariophyta
192982662	*Thalassiosira rotula* CCMP3096	Eukaryota	Bacillariophyta
192994062	*Thalassiosira rotula* GSO102	Eukaryota	Bacillariophyta
193040376	*Thalassiosira weissflogii* CCMP1010	Eukaryota	Bacillariophyta
193042706	*Thalassiosira weissflogii* CCMP1010	Eukaryota	Bacillariophyta
193072692	*Thalassiosira weissflogii* CCMP1336	Eukaryota	Bacillariophyta
193077254	*Thalassiosira weissflogii* CCMP1336	Eukaryota	Bacillariophyta
193079930	*Thalassiosira weissflogii* CCMP1336	Eukaryota	Bacillariophyta
193803118	*Pteridomonas danica PT*	Eukaryota	Ochrophyta
193954876	*Ditylum brightwellii GSO104*	Eukaryota	Bacillariophyta
199325532	*Nitzschia punctata CCMP561*	Eukaryota	Bacillariophyta
199411170	*Fragilariopsis kerguelensis L2_C3*	Eukaryota	Bacillariophyta
199646274	*Pseudo nitzschia-australis* 10249_10_AB	Eukaryota	Bacillariophyta

Sequence ID on MMET transcriptomes	Species and strain	Domain	Phylum
199673778	*Pseudo nitzschia australis* 10249_10_AB	Eukaryota	Bacillariophyta
199684218	*Pelagomonas calceolata* CCMP1756	Eukaryota	Ochrophyta
199818004	*Pseudo nitzschia fradulenta* WWA7	Eukaryota	Bacillariophyta
199870084	*Asterionellopsis-glacialis* CCMP134	Eukaryota	Bacillariophyta
199902980	*Asterionellopsis glacialis* CCMP134	Eukaryota	Bacillariophyta
199941396	*Durinskia baltica* CSIRO CS 38	Eukaryota	Miozoa
200128254	*Thalassiosira antarctica* CCMP982	Eukaryota	Bacillariophyta
200158720	*Proboscia alata* PI D3	Eukaryota	Bacillariophyta
200194834	*Thalassionema nitzschioides* L26 B	Eukaryota	Bacillariophyta
200218036	*Heterosigma akashiwo* NB	Eukaryota	Ochrophyta
200218328	*Heterosigma akashiwo* NB	Eukaryota	Ochrophyta
200377014	*Corethron pennatum* L29A3	Eukaryota	Bacillariophyta
200454796	*Extubocellulus spinifer* CCMP396	Eukaryota	Bacillariophyta
200467358	*Extubocellulus spinifer* CCMP396	Eukaryota	Bacillariophyta
200712054	*Thalassiosira gravida*GMp14c1	Eukaryota	Bacillariophyta
200715886	*Thalassiosira gravida* GMp14c1	Eukaryota	Bacillariophyta
200884264	*Chaetoceros debilis* MM31A 1	Eukaryota	Bacillariophyta
200963138	*Thalassiothrix antarctica* L6_D1	Eukaryota	Bacillariophyta
200988500	*Chaetoceros-neogracile* CCMP1317	Eukaryota	Bacillariophyta
201033068	*Thalassiosira miniscula* CCMP1093	Eukaryota	Bacillariophyta

Appendix 8. BLAST analysis of GHyyy members against OcP1

App. 8.1. Sequences from MMET

Accession	Query cover	E value	Identity
191215482	96%	0	41%
187263766	96%	0	59%
199941396	98%	0	53%
191215482	96%	0	41%
190523074	95%	0	43%
186720192	95%	0	42%
193803118	97%	0	43%
199684218	94%	0	41%
186651940	94%	0	40%
187036080	93%	0	41%
200988500	93%	0	40%
187175162	93%	0	41%
200194834	95%	0	38%
200128254	93%	0	39%
200884264	93%	0	39%
193040376	93%	0	40%
191454482	92%	0	38%
200454796	96%	0	38%
192265400	93%	0	38%
193072692	93%	0	40%
200467358	92%	0	39%
199870084	98%	0	38%
188556876	93%	0	40%
200218328	93%	0	40%
188639640	93%	0	40%
186488266	93%	0	40%
188629668	93%	0	40%
192964706	93%	0	39%
192994062	93%	0	39%
200963138	93%	0	38%
201033068	93%	0	39%
200712054	93%	0	39%
200218036	97%	0	40%
188557734	97%	0	40%
193954876	93%	0	39%
188654168	97%	0	40%
192939680	90%	0	39%

Accession	Query cover	E value	Identity
188604494	93%	0	38%
200377014	93%	0	38%
199325532	93%	0	37%
193077254	92%	0	39%
193042706	92%	0	39%
186540750	93%	0	37%
186521926	93%	0	37%
193079930	92%	0	38%
199902980	93%	0	39%
186515412	93%	0	38%
187365798	90%	0	38%
199646274	98%	0	36%
188154862	93%	0	37%
199411170	93%	0	37%
200158720	92%	0	37%
200715886	92%	0	38%
199818004	96%	0	37%
192982662	90%	0	38%
199673778	92%	0	36%
192890408	85%	0	39%
187050904	63%	0	40%
192128108	64%	0	39%
200418968	23%	1.00E-04	25%
200416338	24%	1.00E-04	26%
200380720	31%	1.00E-05	25%

App. 8.2. Sequences from NCBI

Accession	Query cover	E value	Identity
ABP68120.1	93%	0	32%
AIQ33578.1	92%	0	39%
AQQ70274.1	93%	0	32%
BAM00492.1	89%	0	38%
BAU78234.1	96%	0	58%
BAU78235.1	93%	0	37%
CBK73627.1	92%	0	36%
CCV65814.1	92%	0	33%
CCX92044.1	88%	0	33%
CCX93064.1	89%	0	36%
CCY30928.1	92%	0	35%
CCY69845.1	92%	0	35%
CCZ09292.1	92%	0	36%
CDA29483.1	92%	0	35%
CDB14106.1	71%	8.00E-176	37%
CDF46130.1	88%	0	37%
CUN48701.1	92%	0	36%
CUX60029.1	92%	0	37%
EGW21636.1	92%	3.00E-179	32%
EJK69967.1	93%	0	39%
EKQ50390.1	91%	0	35%
EKQ57686.1	92%	0	35%
ERJ13626.1	92%	0	35%
GAP62125.1	92%	0	38%
KHD15380.1	92%	0	34%
KLU72387.1	93%	0	36%
KOS03742.1	93%	0	36%
KPK04798.1	89%	0	40%
KPK41404.1	92%	0	33%
KPK42408.1	92%	0	32%
KPK96707.1	88%	0	33%
KPK97305.1	90%	0	34%
KPK99010.1	92%	0	34%
KUK24060.1	88%	0	35%
OEU10145.1	92%	0	38%
OEU15658.1	93%	0	37%
OFW62009.1	93%	9.00E-170	31%
OGB86825.1	92%	1.00E-178	32%

Accession	Query cover	E value	Identity
OGB89810.1	92%	0	33%
OGC07941.1	92%	1.00E-178	32%
OGC11849.1	92%	0	33%
OGC23581.1	92%	0	33%
OGC33960.1	64%	6.00E-146	35%
OGC42363.1	85%	6.00E-170	33%
OGV37168.1	92%	0	33%
OGV45644.1	92%	0	33%
OGV52597.1	92%	0	34%
OGW91974.1	93%	0	34%
OGW98683.1	93%	0	35%
OGX01506.1	92%	0	34%
OGX07258.1	93%	0	34%
OGX11645.1	92%	0	34%
OGX23872.1	93%	0	34%
OGX24948.1	93%	0	35%
OGX25981.1	92%	0	34%
OGX31245.1	93%	0	35%
OGX34577.1	92%	1.00E-176	32%
OGX37135.1	93%	0	35%
OGX39498.1	93%	0	35%
OGX43336.1	92%	2.00E-176	32%
OGX46482.1	92%	0	34%
OGX48084.1	92%	0	35%
OGX61342.1	92%	0	34%
OHB53795.1	93%	0	32%
OHE28602.1	92%	0	33%
OHE30965.1	92%	0	34%
OIO35236.1	92%	0	34%
OIO38639.1	92%	0	34%
OIP13455.1	88%	0	33%
OJV66793.1	89%	0	35%
OLP88051.1	93%	0	43%
OMC64708.1	92%	0	38%
OMG47019.1	92%	0	38%
ONI39876.1	92%	0	36%
OOB78077.1	88%	0	36%
SCX05664.1	92%	0	35%
SDB19587.1	92%	0	36%
SDC24037.1	92%	0	34%

Accession	Query cover	E value	Identity
SDF19476.1	92%	0	39%
SDH17459.1	93%	0	37%
SDI45081.1	92%	0	37%
SDY04365.1	92%	0	36%
SEA82830.1	93%	0	35%
SEF83514.1	92%	0	36%
SEG11996.1	93%	0	36%
SEJ83912.1	93%	0	37%
SEK23177.1	92%	0	37%
SEL48476.1	92%	0	34%
SEO89603.1	92%	0	35%
SER69947.1	88%	0	34%
SES15713.1	91%	0	35%
SEU24299.1	93%	0	37%
SFC11500.1	92%	0	33%
SFH69345.1	92%	0	35%
SFN80781.1	92%	0	35%
SFN98921.1	92%	0	37%
SFO78737.1	92%	0	34%
SFP69713.1	92%	0	34%
SFS93792.1	92%	0	34%
SFT60650.1	91%	0	36%
WP_002603706.1	92%	0	35%
WP_003424021.1	92%	0	34%
WP_004103012.1	92%	0	36%
WP_004627615.1	93%	0	35%
WP_005358765.1	92%	0	35%
WP_005363330.1	92%	0	32%
WP_005370277.1	92%	1.00E-161	32%
WP_005490316.1	93%	0	34%
WP_006313788.1	93%	0	39%
WP_008525786.1	92%	0	36%
WP_008676761.1	92%	0	34%
WP_008826269.1	92%	0	35%
WP_008909277.1	93%	0	37%
WP_009168511.1	89%	0	35%
WP_009226247.1	92%	0	38%
WP_009860980.1	90%	0	36%
WP_010174598.1	92%	0	37%
WP_011994121.1	93%	0	36%
WP_012198896.1	90%	0	37%

Accession	Query cover	E value	Identity
ABX81332.1	92%	2.00E-180	33%
WP_012547574.1	93%	0	33%
WP_012582531.1	88%	0	35%
WP_012635372.1	92%	0	33%
WP_012958999.1	92%	0	35%
WP_013129904.1	93%	0	38%
WP_013171295.1	93%	0	35%
WP_013272305.1	92%	0	37%
WP_013289460.1	92%	0	33%
WP_013311628.1	93%	0	37%
WP_013402036.1	92%	0	32%
WP_013429146.1	92%	0	33%
WP_013487305.1	91%	0	35%
WP_013658929.1	92%	0	38%
WP_014079648.1	92%	0	36%
WP_014451045.1	93%	0	37%
WP_014454798.1	92%	0	34%
WP_014641703.1	93%	0	37%
WP_015325999.1	92%	0	34%
WP_015736393.1	92%	0	38%
WP_015848606.1	93%	0	37%
WP_016169534.1	92%	0	35%
WP_016313215.1	93%	0	38%
WP_016818628.1	93%	0	37%
WP_018247638.1	93%	0	35%
WP_018666383.1	93%	0	37%
WP_018751038.1	92%	0	39%
WP_018885078.1	93%	0	36%
WP_019421873.1	92%	0	38%
WP_019635978.1	93%	0	39%
WP_019688419.1	93%	0	37%
WP_019913922.1	91%	0	38%
WP_020426902.1	91%	0	38%
WP_020481102.1	92%	4.00E-171	31%
WP_022766121.1	92%	0	35%
WP_022767451.1	92%	0	35%
WP_022787627.1	89%	0	36%
WP_023989977.1	93%	0	37%
WP_024291666.1	92%	0	37%
WP_024346902.1	92%	0	37%
WP_024834966.1	92%	0	36%

Accession	Query cover	E value	Identity
WP_024865374.1	92%	0	36%
WP_025365657.1	93%	0	37%
WP_025532029.1	92%	0	35%
WP_025679076.1	93%	0	36%
WP_025707431.1	91%	0	37%
WP_025721489.1	93%	0	37%
WP_026600659.1	92%	1.00E-173	32%
WP_026671722.1	92%	0	34%
WP_026691333.1	92%	0	36%
WP_026892135.1	92%	0	37%
WP_027149823.1	93%	0	32%
WP_027308906.1	93%	0	37%
WP_027339873.1	92%	0	34%
WP_027628574.1	93%	0	35%
WP_027636584.1	92%	0	34%
WP_028539026.1	93%	0	37%
WP_028541566.1	93%	0	37%
WP_028591106.1	93%	0	37%
WP_029228890.1	92%	0	33%
WP_029501686.1	92%	0	37%
WP_031392740.1	93%	0	37%
WP_031438254.1	93%	0	32%
WP_031462091.1	93%	0	37%
WP_031543443.1	91%	0	34%
WP_031576719.1	92%	0	35%
WP_032550702.1	92%	0	36%
WP_033156181.1	92%	7.00E-177	32%
WP_033167163.1	92%	0	36%
WP_033191227.1	93%	0	37%
WP_034301206.1	92%	0	35%
WP_034443461.1	92%	0	35%
WP_034488578.1	92%	0	33%
WP_034546778.1	93%	0	33%
WP_034563905.1	92%	0	35%
WP_035212222.1	91%	4.00E-176	33%
WP_035375542.1	92%	3.00E-180	32%
WP_035771655.1	92%	0	35%
WP_035775095.1	92%	0	35%
WP_035779225.1	92%	0	36%
WP_036623986.1	92%	0	38%

Accession	Query cover	E value	Identity
WP_036650681.1	93%	0	38%
WP_036663991.1	92%	0	38%
WP_036682608.1	93%	0	38%
WP_036702107.1	92%	0	37%
WP_036732712.1	91%	0	38%
WP_036781682.1	93%	0	38%
WP_036818050.1	92%	0	38%
WP_036951784.1	92%	0	36%
WP_038281003.1	92%	0	37%
WP_038568482.1	93%	0	38%
WP_038585252.1	92%	0	38%
WP_038698858.1	92%	0	37%
WP_039309715.1	92%	0	38%
WP_039834362.1	91%	0	38%
WP_039869333.1	92%	0	39%
WP_040579857.1	92%	9.00E-180	32%
WP_040952469.1	92%	0	38%
WP_040978143.1	92%	0	34%
WP_041722924.1	93%	0	32%
WP_041902237.1	92%	0	33%
WP_041966187.1	93%	0	37%
WP_042131802.1	92%	0	39%
WP_042133118.1	91%	0	38%
WP_042172399.1	92%	0	36%
WP_042184448.1	92%	0	39%
WP_042210248.1	92%	0	38%
WP_042232123.1	91%	0	38%
WP_043852479.1	92%	0	34%
WP_044294592.1	93%	0	36%
WP_044477328.1	91%	0	38%
WP_044877319.1	92%	0	38%
WP_045085406.1	92%	0	35%
WP_045143732.1	92%	0	34%
WP_045168667.1	93%	0	33%
WP_045173481.1	93%	0	33%
WP_045749352.1	92%	1.00E-172	32%
WP_047001861.1	92%	0	34%
WP_047389514.1	92%	0	37%
WP_050645896.1	92%	0	36%
WP_051615820.1	92%	1.00E-175	31%

Accession	Query cover	E value	Identity
WP_052078517.1	88%	0	35%
WP_052570529.1	92%	0	34%
WP_053325977.1	93%	0	37%
WP_053489857.1	92%	0	38%
WP_053984600.1	93%	0	37%
WP_054400844.1	92%	0	38%
WP_054492046.1	92%	0	38%
WP_054943147.1	92%	0	38%
WP_054955257.1	92%	0	38%
WP_055105746.1	93%	0	39%
WP_055167725.1	92%	0	35%
WP_058146710.1	92%	0	34%
WP_058257546.1	92%	0	36%
WP_058296499.1	92%	0	33%
WP_058300329.1	91%	0	38%
WP_058486419.1	93%	0	36%
WP_059046741.1	92%	0	38%
WP_059051178.1	92%	0	39%
WP_059350157.1	92%	0	37%
WP_060626728.1	92%	0	38%
WP_060862999.1	91%	0	38%
WP_062329571.1	92%	1.00E-171	32%
WP_062370841.1	93%	0	37%
WP_062525136.1	92%	0	37%
WP_063480048.1	92%	0	38%
WP_063829815.1	92%	0	38%
WP_064008863.1	92%	2.00E-175	33%
WP_064011647.1	93%	0	38%
WP_064021676.1	93%	3.00E-161	32%
WP_064029862.1	93%	1.00E-163	32%
WP_064030858.1	92%	6.00E-177	33%
WP_064036102.1	92%	2.00E-175	32%
WP_064041461.1	93%	4.00E-163	32%
WP_064212087.1	92%	2.00E-179	33%
WP_066173653.1	93%	0	37%
WP_066374772.1	92%	0	38%
WP_066420707.1	93%	0	34%
WP_066716675.1	92%	0	37%
WP_066976998.1	92%	4.00E-163	31%
WP_067842620.1	92%	0	37%

Accession	Query cover	E value	Identity
WP_068345671.1	93%	0	36%
WP_068500098.1	93%	0	38%
WP_068528722.1	92%	0	38%
WP_068592622.1	92%	0	38%
WP_068649783.1	92%	0	38%
WP_068660028.1	92%	0	37%
WP_068717352.1	92%	0	35%
WP_068782371.1	92%	0	39%
WP_068787086.1	92%	0	38%
WP_069079905.1	93%	0	37%
WP_069366220.1	92%	0	33%
WP_070097083.1	89%	0	38%
WP_071159838.1	92%	4.00E-173	33%
WP_071220018.1	92%	0	37%
WP_071308156.1	91%	0	35%
WP_071386948.1	92%	0	33%
WP_071558349.1	93%	0	37%
WP_071847807.1	92%	0	34%
WP_072580822.1	93%	0	37%
WP_072757930.1	92%	0	38%
WP_072916440.1	92%	0	36%
WP_073248831.1	93%	0	37%
WP_073275368.1	92%	0	36%
WP_073287569.1	93%	0	37%
WP_073298186.1	93%	0	33%
WP_073338818.1	93%	0	36%
WP_073388888.1	92%	0	33%
WP_073590012.1	92%	0	37%
WP_074086705.1	92%	0	38%
WP_074101767.1	92%	0	38%
WP_074107672.1	92%	0	38%
WP_074109980.1	92%	0	38%
WP_074755796.1	92%	0	34%
WP_074912609.1	92%	0	34%
WP_075154651.1	93%	0	37%
WP_075679299.1	92%	0	37%
WP_075682963.1	92%	0	35%
WP_076083370.1	92%	0	38%
WP_076113870.1	92%	0	38%
WP_076116623.1	92%	0	38%
WP_076150417.1	92%	0	38%

Accession	Query cover	E value	Identity
WP_076160219.1	92%	0	37%
WP_076160827.1	93%	0	37%
WP_076243564.1	91%	0	38%
WP_076325048.1	93%	0	37%
WP_076544684.1	92%	0	34%
XP_002177898.1	93%	0	38%
XP_002181175.1	93%	0	39%
XP_002290006.1	93%	0	40%
XP_002297139.1	85%	0	36%
XP_009032155.1	86%	0	41%

Appendix 9. GC content analysis of bacterial *GHyyy* genes and their associated genomes.

Species	RefSeq	Domain	Phylum	Nucleotide accession	%GC genome	%GC GHyyy
Thermosipho africanus TCF52B 1	WP_004103012.1	Bacteria	Thermotogae	NC_011653.1	30.7	29.1
Acholeplasma laidlawii PG-8A	WP_012242663.1	Bacteria	Tenericutes	NC_010163.1	31.9	30.3
Halobacteroides halobius DSM 5150	WP_015325999.1	Bacteria	Firmicutes	NC_019978.1	32.5	32.9
Dictyoglomus thermophilum H-6-12	WP_012547574.1	Bacteria	Dictyoglomi	NC_011297.1	33.7	30.9
Cellulosilyticum lentocellum DSM 5427	WP_013658929.1	Bacteria	Firmicutes	NC_015275.1	34.3	36.2
Fervidobacterium nodosum Rt17-B1	WP_011994121.1	Bacteria	Thermotogae	NC_009718.1	35	35.9
Caldicellulosiruptor kronotskyensis 2002	WP_013429146.1	Bacteria	Firmicutes	NC_014720.1	35.1	35.5
Caldicellulosiruptor obsidiansis OB47	WP_013289460.1	Bacteria	Firmicutes	NC_014392.1	35.2	35.5
Caldicellulosiruptor saccharolyticus DSM 8903	WP_041722924.1	Bacteria	Firmicutes	NC_009437.1	35.3	34.3
Lachnoclostridium phytofermentans ISDg	WP_012198896.1	Bacteria	Firmicutes	NC_010001.1	35.6	35.5
Bacillus cellulosilyticus DSM 2522	WP_013487305.1	Bacteria	Firmicutes	NC_014829.1	36.5	34.2
Halothermothrix orenii H168	WP_012635372.1	Bacteria	Firmicutes	NC_011899.1	37.9	38.4
Fervidobacterium pennivorans DSM 9078	WP_014451045.1	Bacteria	Thermotogae	NC_017095.1	38.9	37.8
Bacillus pseudofirmus OF4	WP_012958999.1	Bacteria	Firmicutes	NC_013791.2	39.9	42.4
Thermococcus sibiricus MM 739	WP_015848606.1	Archaea	Euryarchaeota	NC_012883.1	40.2	46.4
Halobacillus halophilus NC_017668.1	WP_014641703.1	Bacteria	Firmicutes	NC_017668.1	41.8	40.5
Paenibacillus polymyxa E681	WP_013311628.1	Bacteria	Firmicutes	NC_014483.2	45.5	47.4
Paenibacillus polymyxa CR1	WP_023989977.1	Bacteria	Firmicutes	NC_023037.2	45.5	47.3
Thermosphaera aggregans DSM 11486	WP_013129904.1	Archaea	Crenarchaeota	NC_014160.1	46.7	47
Paenibacillus sp. Y412MC10	WP_015736393.1	Bacteria	Firmicutes	NC_013406.1	47.8	53.9
Roseburia hominis A2-183	WP_014079648.1	Bacteria	Firmicutes	NC_015977.1	48.5	52.1
Bacillus selenitireducens MLS10	WP_013171295.1	Bacteria	Firmicutes	NC_014219.1	48.7	52.9
Clostridium saccharolyticum WM1	WP_013272305.1	Bacteria	Firmicutes	NC_014376.1	50.2	45.6
Spirochaeta africana DSM 8902	WP_014454798.1	Bacteria	Spirochaetes	NC_017098.1	57.8	58.9
Halorhabdus tiamatea SARL4B	WP_008525786.1	Archaea	Euryarchaeota	NC_021921.1	62.6	65.6
Phaeodactylum tricornutum CCAP 1055/1	XP_002177898.1	Eukaryota	Bacillariophyta	ABQD00000000.1	48.8	46.3
Phaeodactylum tricornutum CCAP 1055/1	XP_002181175.1	Eukaryota	Bacillariophyta	ABQD00000000.1	48.8	49.5
Thalassiosira pseudonana CCMP1335	XP_002290006.1	Eukaryota	Bacillariophyta	AAFD00000000.2	46.9	48.1
Thalassiosira pseudonana CCMP1335	XP_002297139.1	Eukaryota	Bacillariophyta	AAFD00000000.2	46.9	44.8
Thalassiosira oceanica CCMP1005	EJK69967.1	Eukaryota	Bacillariophyta	AGNL00000000.1	53.3	53

Species	RefSeq	Domain	Phylum	Nucleotide accession	%GC genome	%GC GHyyy
Aureococcus anophagefferens	XP_009032155.1	Eukary-ota	Bacillariophyta	NZ_ACJI000 00000.1	67.4	68.4
Symbiodinium microadriaticum	OLP88051.1	Eukary-ota	Miozoa	LSRX000000 00.1	46.6	54.9
Fragilariopsis cylindrus CCMP1102	OEU10145.1	Eukary-ota	Bacillariophyta	LFJG000000 00.1	38.8	41.4
Fragilariopsis cylindrus CCMP1102	OEU15658.1	Eukary-ota	Bacillariophyta	LFJG000000 00.1	38.8	41

Appendix 10. Synthetic genes

App. 10.1. EgP1

ATGTGGCATCGTGCAGTTAATGCAACCGCACGTGTTGCACAGAGTTTTCAGCGT
ACCATGTTTAGCCCGGCACCTGTGGCACCTCAGAGAGCACCTAGCAGTACCATG
GCAGCAGCCCCGGCATTAAGAGTTGGCGATGGCTGGTATAAAGCACCGGCAGG
TATTCCGGAAGGTGAATATGTGAGCCGTCTGGGCGAAGAATGGTATCGGATTGA
TAACTATGATCAGATGCCGCCGTTTTTTATGAGCATTGTTAGCCCGAGCAATTTTT
GGCTGTACTTGAGTACAACAGGTGGTCTTACGTGTGGTCGTGTGAGCGCCGAAC
ATGCACTGTTTCCGTATTATACCGTTGATAAAATTCGTGATAGCTCAACCAATACC
GGTCCGTGTACCCTGGTTCATGCAACCCGTGGTGGTGAAACCTTTTTATGGCAG
CCTTTTCTGCCGCAGCCTCGTCTCTATCAGACCACCCGTGCCATTTATAAAACCG
TTAGCGGCGATAAAGTTGAATTTGAAGAAGTGAATCATGATCTTAAACTGCGCTT
TAGCTATACCTGGTCCACCACCGAAGAGTTCGGTTTCGTTCGCAAATGTGTTCTG
ACCAATCTTAGTAATGATGACGTTGCAATCCGTATTTTAGATGGTATGCAGAATAT
CCTGCCGGCGGGCGTTAATACCACACTGCAAAATATTAAAAGCTGTCTGGTCGA
TGCATATAAACTGACCGAACTGGATCCGACCGTGCAGCTGGGTATCTTTTGTCTG
TATGCCCTGCCGTGGGATCGGGCAGAACCGAAAGAAAGCCTGAAGGCAAATAC
CGTTTTTCATCGTGGTCTGCCGGCACCGGTTGTTCTGCTGAGTACACGTCAGCTT
GATCGTTTTCGTGCAGGTCATCCGGCACAGGCCGAAGCACGTATTCGTGGTCAG
CGTGGTGCCTTTTTTGTTCAGAGCGAATTTAGCCTGAAAGGTCAGGCAGAACAG
CGTTGGGTTCTGGTTGCAGAAGTTGAACAGAATCATCGTAGCATTGCCGCCCTG
CGTCAGAAACTGACAGCAAGTACCGCAGACCAGGTTCTGGCCGCAATTACCGAA
GGCGAAAAAGCAGCAGGCGTTAGCCTGCGTAATTTTATTGCAGCAGCAGATGGT
CTTCAGTTAACAAAAAGCGTGGGTACCACCGCCCATCATTTTGCCAACGTTATGT
TTAATAGCATGCGTGGCGGTATTTTTGATAGCAATTATGTTGTTCACCGTGATGC
ATTTGGCGCATTTGTTCGTACCTGGAATAAACCGCTGGCAACCCAGCAGAAAGG
TTGGCTGAGCGCATTACCTGAAGCACTGCCGAGCAGCGAATTATTAGCACAGGC
AAAAGCAACCGGCGATCCTGATCTGGTGCGCCTGGCGCTGGAATATCTGCCGAT
TACCTTTAGCCGTCGTCATGGTGATCCGAGCCGTCCGTGGAATACATTTAGCATT
TGTCTGCGTGATGCCGCAGGTAATAAAGTTCTGGGTTATCAGGGCAATTGGCGT
GATATTTTTCAGAATTGGGAAGCACTTTGTATGAGCTACCCGAACTTTTTACCGA
GCATGATTGCCAAATTTGTCAATGCGACCACAGCCGATGGTTATAATCCTTATCG
TGTTAGCAATGATGGTATTGATTGGGAAACCGTTGAACCGCATGATCCGTGGAG
CTATATTGGTTATTGGGGTGATCATCAGATTGTGTATTTACTGCGTTTTCTGGAAA
GCTTACGTCGTTTTGATCCTAGCGCACTGGAAGAGTTACTGCATGAAGAAATCTT
CGCCTATGCAAACGTGCCGTATATCATTCGCCCGTATAATGAAATCGTGAAAAAT

CCGAAGGACACCATTGTTTTTGCCCATACCCGTCATGCAGATCTGATGGCACTG
GCCAAAACCATGGGTAGCGATGGTAAATTACTGCTGGATGGTCATGGTAAAGTG
CTGCGTGTTAATCTGGTGGAAAAACTGCTGGTGAGCCTGCTTGCAAAACTGAGT
AATTTTGTTCTTGATGGTGGTATTTGGCTGAATACACAGCGTCCGGAATGGAATG
ATGCCAATAATGCCCTGGTTGGTAATGGTATTTCAATGGTTACCACCTTTCATCT
GCGTCGTTGGTTAACCTTTGTTATTGCCGAACTGCAAGCAATTAAGGGCGAAACC
AAACTTAGCGCAGAAGTGAGCGTGTGGTTTGACGGTATTAAAAAAGTTTTTGCAG
AAAATGAAGGTATTCTGGGTGGTGCAGTGAGCGCGAGCCAGCGTCGTGCAATG
CTGGACGCACTGGGTGAAGCAGCAAGCGTTTATCGTGGTATCCTGTATGAAAAA
GGTCTGAGCGGCGCAAAAGTGGCAGTTCCGACCGCAAGCGTGGTCGAATTTTTA
CAGAGCGCCTTGAAATTTGTGGATCATACAATTCGTGCAAATAAAACCCCGGAAG
GCCTGTATCATTCATATAATCTGCTGGTTCTTGGTCCGGGCAGCGCAGATATCAA
ACATCTGTATCTGATGCTGGAAGGCCAGGTGTGTGCACTGAGCAGCGGTCTGAT
TACCGGTCAGGAAGCAGTTGAAATGCTGCGCCATCTGCGGGATAGTGCACTGTA
TCGTGCCGATCAGGACAGTTATACCCTGTATCCGGATCGTGAAGTTCCTCCGTTT
CTGGCACGTAATGTTATTCCTGCACCGCGTCTGGCACTGCCTGGCCTGAAATTC
GTTCTGGATCATGGTCTGCAGTCAATTGCCTATGTTGACGCCGAAGGTGTTGGT
CGCTTTGGCGATACCCTGAGCAATGCAGATGATCTGTTAGCAGCACTGGATAAA
CTGCAGGATAGTCATCCTGGTGTTGATGCACATCGCGCAGAACTGGCAGAAACA
TTTGAAGCCGTGTTTGATCATAAAGCATTTACCGGCCGTAGTGGTACAATGTTTA
GTTATGAAGGTCTGGGTTGTATCTATTGGCACATGTGTGCAAAGCTGGCATTAGC
AGTTAGCGAACTGTGTGCAGCCGCAGAAGGCACCGCAGATCATGTTGCACTGC
GTGACAAATATTATGAACTGCGTAAAGGTTTAGGTGGTTTTAATAAAAGCCCGGC
CGTTTATGGTGCATTTCCGGCAGATCCTTATAGTCATACCCCGGCACATGCGGG
TGCAAAACAGCCTGGTATGACCGGCCAGGTTAAAGAAGAAATTCTGACCCGTTTT
GCCGAATTGGGTGTGACCATTAGTGACGGTAAACTGGCATTTGCACCGACCATT
CTGCGCCAGAGTGAATTTCTGGCCCAGCCTGCAGAATTTGCATATGTGGATCTG
CGTGCACGTCAGTTGACCATTCCGCTGAAAGCAGGCGAACTGGCCTTTACCTAT
TGTCAGGTTCCGGTTGTGTATCGTATTGGTGACACCCCGGGTGTTACCCTGACC
AGCGCAAGCCAGGGTGAAGTCGCAGTTCCTGGAAATCAGCTGGATCGTCATTGG
AGCAGCCAGCTGTTTGGTCGTACCGGTGATGTTGTTAGCATTACCGTTGTTGTTC
CGAGTGCAGTTCTGTTT

App. 10.2. EgP2

ATGTGGCATCGCGCAGTGAATGCAACCGCACGTGTTGCACAGAGCTTTCAGAGA
ACCATGTTTAGCCCGGCACCGGTCGCACCTCAACGTGCACCGTCAAGCACAATG

GCTGCTGCCCCTGCACTGCGTGTTGGTGATGGTTGGTATAAAGCACCGGCAGGT
ATCCCTGAAGGTGAATACGTTAGTCGTCTGGGTGAAGAATGGTATCGTATTGATA
ATTATGATCAGATGCCGCCGTTTTTTATGAGCATTGTTAGCCCGAGCAACTTTTG
GCTGTATCTGAGCACAACAGGTGGCCTGACATGTGGTCGTGTTAGCGCCGAACA
TGCACTGTTTCCGTATTATACCGTTGATAAAATCCGTGATAGCAGCACCAATACA
GGTCCGTGTACCCTGGTTCATGCAACCCGTGGTGGTGAAACATTTCTGTGGCAG
CCGTTTCTGCCGCAGCCGCGTTTATATCAGACCACCCGTGCCATTTATAAAACCG
TTAGCGGAGATAAAGTGGAATTCGAAGAAGTTAATCATGATCTGAAACTGCGCTT
TAGCTACACCTGGAGCACCACCGAAGAATTCGGTTTTGTGCGTAAATGTGTGCT
GACCAACCTGAGCAATGATGATGTTGCAATTCGTATTCTGGATGGTATGCAGAAC
ATTCTTCCGGCAGGCGTTAATACCACCCTGCAGAATATTAAAAGCTGTCTGGTTG
ATGCCTATAAACTGACAGAATTAGATCCGACCGTTCAGTTAGGTATTTTCTGTCT
GTATGCACTGCCGTGGGATCGTGCAGAACCGAAAGAAAGCCTGAAAGCAAATAC
CGTTTTTCATCGTGGTCTGCCGGCCCCGGTTGTTCTTCTGAGTACCCGTCAGTTA
GATCGTTTTCGTGCAGGTCATCCAGCACAGGCCGAAGCCCGTATTCGTGGTCAG
CGCGGTGCATTTTTTGTTCAGAGCGAATTTAGCTTAAAAGGTCAGGCAGAACAGC
GTTGGGTTCTGGTGGCAGAAGTTGAACAGAATCATCGTAGCATTGCAGCATTAC
GTCAGAAACTGACCGCAAGCACCGCAGATCAGGTGTTAGCCGCAATTACCGAAG
GCGAAAAAGCCGCAGGAGTTAGCCTGCGTAATTTTATCGCCGCAGCAGATGGTT
TACAGTTGACCAAAAGCGTGGGTACCACCGCACATCATTTTGCAAATGTGATGTT
TAATAGCATGCGTGGCGGTATTTTTGATAGCAATTATGTTGTTCACCGCGATGCA
TTTGGTGCCTTTGTTCGTACCTGGAATAAACCGCTGGCAACCCAGCAGAAAGGT
TGGCTGTCAGCACTGCCTGAAGCATTACCGAGTAGCGAACTGCTGGCACAGGCA
AAAGCCACCGGTGATCCTGATTTAGTTCGTCTGGCACTGGAATATTTACCGATTA
CCTTTAGCCGTCGTCATGGTGATCCGAGCCGTCCGTGGAATACCTTTTCAATTTG
TCTGCGTGATGCCGCAGGCAATAAAGTTCTGGGTTACCAGGGCAATTGGCGCGA
TATTTTTCAGAACTGGGAAGCACTTACCATTAGCTTCCCTTATTATTTTGAAAGTA
TCATTGCCAAGTTTGTTAACGCCAGCACCGTGGACGGTTACAATCCGTACATTGT
TAATCGTACCGATGGCATTAATTGGGAATGTCCGGAACCGGATCATCCGCACGC
AAATATTGGTTATTGGGGCGATCATCAGATTATCTATCTGTTAAAACTGTTAGAAT
GGTGTGGTAAGTTTCAGCCGGGTAAGCTGGAGGCAAATTGTGATAAGGATTGGT
TTAGCTATGCCAACGTTCCGTACGAGATCCGCCCGTATGATAAGATTGTTGAGAA
TGCAAAAGATACGATCACCTTTAATTGGAATAAGCACAAACACATCGAAGCACTG
AAGAAAACCATGGGTGCCGATGCGAAACTGGTTCTGACCAAAGATGGTAAAGTT
GTTCATGTTAATCTGGCAGAAAAGTTATTAGTTCCGCTGATGGCCAAAGCAAGCA
ATTTTGTGATTGGTGGTGGTATCTGGCTGAATACCCAGCGTCCGGAATGGAATG
ATGCAAATAATGCGATTGTTGGTTTTGGTCTGAGTATGGTTACCGTGTATTACAT

GCGTCGTTATGTTAGCTTTCTGTTACAGCTGCTGAAAAGCCTGCCGACCCCGAC
CGTGAATCTGAGTGAAGAGGTTGCAGTTTGGCTGGCAGGTTTAACCAAAGTGTA
TGCCGATAATGTTGGCCTGATTGCCGGTGATAAAGATCTGAGTAGCCATGATCG
TCGCCAGTTACTGGACGCACTGGGTGTTGTTGCAAGCGAATATCGTTGGAAAAT
TTATGAGAATGGTTTTAGCGGTCAGAAAACCGCAGTTAAAGTTGCAGATGCGGTT
CATTTTCTCGATAGCCTGCTGCTGTTTATTGACTATAGCATCCGCAAGAATAAGA
AAACAGATGGTCTGTATCATGCATATAATCTGCTGAGCCTGCATCCGGGTAAAGC
AGATATTGGCTATCTGTATGTTATGCTGGAAGGCCAGGCAAGCGCACTGAGTTC
AGGTCTGATTACCAGCGATAGCGCCGCCGCACTGTTCAATCATATTTATAGCAGC
GACTTATATCGTCCGGATCAGCATAGCTTTATGCTGTATCCGGATCGCAAACTGA
AGGGTTTTATGGAACGTAATTGTATTCCGGAGTATCGTCGTGATGCAGCCGCAG
TTGAATTTTGTCTTACCAATAAACTGTCACGTATCCTGTATGAAGATGCAAACGGC
GATCTGCGTTTTGGTGCAACCCTGGCAAATGCCAATGATCTTAAAGCAGCAATTC
AGGCATGTGCAAATCCGAGCGTGCAGGATCATACCGCAGAATTATTAGATATTTA
TGAAGAAGTCTTTGTGCATCGTGCATTTACCGGTCGTAGCGGTACAATGTTTGGT
TTCGAAGGTCTGGGCTGTATTTATTGGCATATGGTTGCAAAAGTTTTACTGGCCG
CACAGGAAGTTACCCTGGATGCAATTGATCGTAATGACGGTTGTGTTGGTACACT
GCAGAAATATTATTATGAAGCACGTGCCGGTATTGGTTTTAATAAAAGCCCGGAA
GTTTATGGTGCGTTTCCGTGTGATCCGTATAGCCATACCCCGAAACAGGCAGGC
GCACAGCAACCGGGTATGACCGGCCAGGTTAAAGAGGAAGTTCTTACCCGTTTT
GGCGAATTAGGTGTTCGTATTGTTAAAGGTACCATCCTGCTGCAGCCGAAACTG
CTGCGTCGTCGCGAATTTTTACAGGCACCTCGTGTTTTTGAATATGTTAGTTATGA
TGGTAGTCTGGTTCAGCGTACCTTGCAGGCAGGTCAGCTGGGTTTTACCTATTGT
CAGTGTCCGTTTATTTACACCAAGGGTACCAAATTAAGCATTGTGGTTAAAACCT
CAGCAAAAGAAATTCCGGTGGAAGGTGATACCTTACCGGCAGATGTTGCCTTTA
AAATTTTTAGCCGGAGCGGTGAAGTTTTAGAAGTTAGCGTTGAAGTTCCGGATAG
TGTTCTGTTAGATTTT

App. 10.3. EgP3

ATGTGGCATCGGGCAGTTAATGCAACCGCACGTGTTGCACAGAGCTTTCAGCGT
ACAATGTTTAGTCCGGCACCTGTTGCCCCACAGCGTGCACCGAGCAGTACAATG
GCCGCTGCACCAGCACTTCGCGTTGGTGATGGTTGGTATAAAGCACCGGCTGGT
ATTCCGGAAGGCGAATATGTTAGTCGTCTGGGTGAAGAATGGTATCGCATTGATA
ATTATGATCAGATGCCGCCGTTTTTTATGAGCATTGTTAGCCCGAGCAATTTTTG
GCTGTATCTGAGCACCACAGGCGGCTTAACCTGTGGTCGTGTTAGCGCCGAACA
TGCCCTGTTTCCGTACTATACCGTTGATAAAATTCGTGATAGCAGTACCAATACC

GGTCCGTGTACCCTGGTGCATGCAACCCGCGGTGGCGAAACCTTTTTATGGCAG
CCGTTTCTGCCTCAGCCGCGTCTGTATCAGACCACCCGTGCAATTTATAAGACC
GTTAGCGGCGATAAAGTGGAATTTGAAGAAGTGAATCATGATCTTAAACTGCGGT
TTAGCTACACCTGGAGCACAACCGAAGAATTCGGCTTTGTGCGTAAATGTGTTCT
GACCAACTTAAGCAATGATGATGTTGCAATTCGCATTCTGGATGGTATGCAGAAC
ATTCTGCCGGCCGGTGTGAATACCACCCTGCAGAATATTAAAAGCTGTCTGGTT
GATGCATATAAACTTACAGAACTGGATCCGACCGTGCAGCTGGGCATCTTTTGTC
TGTATGCCCTGCCGTGGGATCGTGCAGAACCGAAAGAAAGCTTAAAAGCAAATA
CCGTTTTTCATCGTGGTTTACCGGCACCGGTTGTTTTATTAAGCACCCGTCAATT
AGATCGTTTTCGTGCAGGTCATCCTGCACAGGCAGAAGCACGTATTCGCGGTCA
GCGCGGCGCATTTTTTGTGCAGAGCGAATTTAGCCTGAAAGGACAGGCCGAACA
GCGTTGGGTTCTGGTTGCAGAAGTTGAACAGAATCATCGTAGCATTGCAGCATTA
CGTCAAAAATTAACCGCAAGTACAGCAGACCAGGTTCTGGCCGCAATTACCGAA
GGTGAAAAAGCCGCCGGTGTTAGCCTGCGTAATTTTATTGCCGCAGCAGATGGT
CTGCAACTGACCAAAAGCGTTGGCACCACCGCACATCATTTTGCAAATGTTATGT
TTAATAGCATGCGTGGTGGCATTTTTGATAGCAATTATGTTGTTCACCGTGATGC
ATTTGGCGCCTTTGTTCGCACCTGGAATAAACCTTTAGCCACCCAACAGAAAGGT
TGGCTGAGTGCACTGCCGGAAGCACTTCCGAGCAGCGAACTGCTGGCACAGGC
CAAAGCAACCGGCGATCCTGATCTGGTTCGTCTGGCATTAGAATATTTACCGATT
ACATTTAGCCGTCGTCATGGTGATCCGAGCCGTCCGTGGAATACCTTTAGTATTT
GTCTGCGTGATGCCGCAGGTAATAAAGTGCTTGGTTATCAGGGTAACTGGCGCG
ATATCTTTCAGAATTGGGAAGCCCTTACCATTTCATTCCCGTATTATTTTGAGTCG
ATCATTGCAAAGTTTGTGAATGCCAGCACCGTGGATGGTTACAATCCGTATATCG
TTAATCGTACCGATGGCATTAATTGGGAGTGTCCGGAACCGGATCATCCGCATG
CAAACATTGGTTATTGGGGTGATCATCAGATTATCTATCTGCTTAAGCTTCTGGAA
TGGTGTGGTAAGTTTCAGCCTGGTAAGCTGGAAGCAAATTGTGATAAAGACTGG
TTCAGCTACGCCAATGTTCCGTACGAAATTCGTCCGTATGACAAAATTGTGGAGA
ACGCCAAAGACACGATCACCTTCAATTGGAACAAGCATAAACACATTGAAGCCCT
GAAAAAAACCATGGGTGCAGATGCCAAACTGGTGCTGACCAAGGATGGTAAAGT
GGTTCATGTGAATCTGGCAGAGAAATTACTGGTTCCGCTGATGGCAAAAGCAAG
TAATTTTGTGATTGGTGGCGGAATTTGGCTGAATACACAGCGTCCGGAATGGAAT
GATGCAAATAATGCAATTGTGGGTTTTGGCCTTAGCATGGTTACCGTTTATTATAT
GCGTCGTTATGTTAGCTTTCTGCTGCAGCTGCTGAAAAGCCTGCCGACCCCGAC
CGTTAATCTGAGTGAGGAAGTTGCCGTTTGGCTGGCAGGTCTGACAAAAGTTTAT
GCAGATAATGTGGGTCTGATTGCAGGAGATAAAGATTTAAGCAGCCATGATCGT
CGTCAGTTACTGGATGCACTGGGTGTTGTTGCAAGCGAATATCGTTGGAAAATTT
ATGAGAATGGTTTTAGCGGTCAAAAAACCGCCGTTAAAGTGGCCGATGCCGTTC

ATTTTCTGGATAGCCTGCTGCTGTTTATTGATTATAGCATTCGCAAAAATAAGAAA
ACCGACGGCCTGTATCATGCCTATAATCTGCTGAGCCTGCATCCGGGCAAGGCA
GATATTGGCTATTTATATGTTATGCTGGAAGGTCAGGCAAGCGCACTGAGCAGC
GGTTTAATTACCAGCGATAGCGCGGCAGCACTGTTTAATCACATTTATAGCAGCG
ATCTGTATCGTCCGGATCAGCATAGCTTTATGCTGTATCCGGATCGTAAACTGAA
AGGCTTCATGGAACGTAATTGTATCCCGGAATATCGGCGCGATGCAGCCGCAGT
TGAATTTTGTTTAACCAATAAACTGAGCCGTATTCTGTATGAAGATGCAAACGGT
GATCTGCGCTTTGGTGCAACCCTGGCAAATGCAAATGATCTGAAAGCAGCAATT
CAGGCATGTGCAAATCCGAGCGTTCAGGATCATACCGCAGAGCTGCTGGATATT
TATGAAGAGGTTTTTGTTCATCGCGCATTTACCGGTCGTAGCGGTACCATGTTTG
GTTTTGAAGGTCTGGGTTGTATTTATTGGCATATGGTTAGCAAAGTTCTGTTAGC
AGCACAGGAACTGACCCTGGATGCCTTAGATCGCCGTGATGGTAGCATGGCCG
CACTGCGTGCAGCATATTATGAACTGCGGGGCGGCATTGGTTTTAATAAAGCGC
CGCAGGAATATGGTGCATTTCCGAGCGATCCTTATAGCCATACCCCGAAACATG
CCGGTGCACAGCAGCCGGGTATGACAGGTCAGGTTAAAGAAGAAATTCTGACCC
GTTTTGGTGAACTGGGTATTCGTATTGTTCATGGCACCATTCGTGTTGAACCGCG
TTTACTGCGCCCGAGTGAATTTCTGACCGCACCGAAGGATTTTGAACATGTGGC
ACTGGGCGGTGCACCTCAGCGTATGCTGCTGCAACCGGGTCAGCTGGGTTTTA
CCTATTGCCAGTGTCCGTTTGTGTATGCACGTGGAAAAGTTTTACGCATGGTGGT
TACCTGTCGCGCAAAAAAAGTGTATATTAATGGTCATGATTTACGCCGCGATCTG
GCAACCCATATTTTTAAACGCAGCGGTGAAGTTCTGCGTGTTGATGTTGAAATTC
CGGATAAAGAACTGTTTCATGTT

App. 10.4. EgP4

ATGTGGCATCGTGCCGTTAACGCAACAGCACGTGTTGCACAGAGCTTTCAGCGT
ACAATGTTTAGCCCTGCACCTGTTGCACCTCAGCGTGCACCGAGCAGCACAATG
GCAGCAGCACCGGCACTGAGAGTTGGTGATGGCTGGTATAAAGCACCAGCAGG
CATTCCGGAAGGCGAATATGTTAGCCGTCTGGGTGAGGAATGGTACCGTATTGA
TAATTATGATCAGATGCCTCCGTTTTTTATGAGCATTGTTAGCCCGAGTAATTTTT
GGCTGTATCTGAGCACAACCGGCGGTCTTACCTGTGGTCGTGTTAGTGCAGAAC
ATGCACTGTTTCCGTATTATACCGTTGATAAAATTCGCGATAGCAGCACGAATAC
CGGTCCGTGTACCCTGGTTCATGCAACCCGCGGTGGTGAAACCTTTCTTTGGCA
GCCGTTCCTGCCGCAGCCGCGTCTTTATCAGACCACCCGTGCAATTTATAAAAC
CGTTAGCGGCGATAAAGTGGAATTTGAAGAAGTGAATCATGATCTGAAACTGCG
CTTTAGCTATACCTGGAGCACCACCGAAGAATTCGGTTTTGTGCGTAAATGTGTT
CTGACCAATCTGAGTAATGATGATGTTGCCATTCGCATTCTGGATGGTATGCAGA

ATATCCTGCCGGCCGGTGTGAACACCACCTTACAGAACATTAAAAGCTGCCTGG
TTGATGCCTATAAACTGACCGAACTGGATCCGACCGTTCAGTTAGGTATTTTTTG
CTTATATGCCCTGCCGTGGGATCGCGCAGAACCGAAAGAAAGCTTAAAAGCAAA
TACCGTTTTTCATCGTGGTCTGCCGGCACCGGTGGTTTTACTGAGTACACGTCAG
TTAGACCGTTTTAGAGCAGGTCATCCGGCACAGGCCGAAGCACGTATTCGCGGT
CAGCGTGGTGCCTTTTTTGTTCAGAGTGAATTTAGCCTGAAAGGTCAGGCAGAA
CAGCGTTGGGTTCTGGTTGCAGAAGTTGAACAGAATCATCGTAGTATTGCAGCA
CTGCGTCAGAAACTGACAGCAAGTACAGCAGATCAGGTTCTGGCCGCAATTACC
GAAGGTGAAAAAGCAGCGGGTGTTAGTCTGCGTAATTTTATAGCAGCGGCAGAT
GGTCTTCAGCTGACAAAAAGCGTTGGTACAACCGCACATCATTTTGCCAACGTTA
TGTTTAATAGCATGCGTGGCGGCATTTTTGATAGCAATTACGTTGTTCACCGTGA
TGCCTTTGGTGCATTTGTTCGTACATGGAATAAACCGCTGGCCACCCAGCAGAA
AGGCTGGCTGAGTGCACTTCCGGAGGCACTGCCTAGCAGCGAATTATTAGCACA
GGCAAAAGCAACCGGTGATCCGGACCTGGTTCGCCTGGCCCTGGAATATCTTCC
TATCACCTTTAGCCGCCGTCATGGCGATCCGAGCCGTCCGTGGAATACCTTTAG
TATTTGTCTGCGTGATGCAGCCGGTAATAAAGTTCTGGGTTATCAGGGTAATTGG
CGTGATATTTTTCAGAATTGGGAAGCACTGTGTATGAGCTATCCGAATTTTCTGC
CGAGCATGATCGCAAAATTTGTGAATGCAACCACCGCAGATGGCTATAATCCGTA
TCGCGTTAGCAATGACGGTATTGATTGGGAAACCGTTGAACCGCATGATCCGTG
GAGCTATATTGGTTATTGGGGTGATCATCAGATTGTTTATCTGTTACGTTTTCTTG
AAAGCCTGCGTCGCTTTGATCCTAGCGCACTGGAAGAACTGCTGCATGAAGAAA
TTTTCGCCTACGCCAACGTGCCGTATATCATTCGTCCGTATAACGAAATTGTTAA
AAATCCGAAAGATACCATCGTGTTTGCCCATACCCGTCATGCAGATCTGATGGCA
CTGGCCAAAACCATGGGTAGCGATGGTAAATTACTGCTGGACGGTCATGGTAAA
GTGCTGCGTGTTAATCTGGTTGAGAAACTGCTGGTTTCACTGCTGGCAAAACTTA
GCAATTTTGTTTTAGATGGAGGCATTTGGCTGAATACCCAGCGTCCGGAGTGGA
ATGATGCAAATAATGCACTGGTTGGTAATGGTATTAGCATGGTTACCACCTTTCA
TCTGCGTCGTTGGCTGACATTCGTTATTGCAGAATTACAGGCAATCAAAGGCGAA
ACAAAACTGAGCGCCGAAGTTAGCGTTTGGTTTGATGGCATTAAAAAAGTGTTCG
CAGAAAACGAAGGTATCCTGGGTGGTGCAGTTAGCGCAAGTCAGCGTCGCGCA
ATGCTGGATGCCCTGGGTGAAGCAGCCAGCGTTTATCGTGGTATTCTGTATGAA
AAAGGTCTGAGCGGTGCAAAAGTTGCAGTTCCGACCGCAAGCGTTGTTGAATTT
TTACAGAGCGCACTTAAATTCGTGGATCATACCATTCGCGCCAATAAAACCCCGG
AGGGTCTGTATCATAGCTATAACCTGTTAGTTTTAGGTCCGGGTAGCGCAGATAT
CAAACATTTATATCTGATGCTGGAAGGCCAGGTTTGTGCACTGAGCAGCGGTCT
GATTACGGGTCAGGAAGCCGTTGAAATGCTGCGTCATTTACGTGATAGCGCATT
ATATCGCGCCGATCAGGATAGTTATACCCTGTATCCGGATCGTGAAGTTCCGCC

GTTTCTGGCACGTAATGTTATTCCGGCACCTCGTTTAGCACTGCCGGGTCTGAAA
TTTGTTCTGGATCATGGTCTGCAGAGCATTGCATATGTTGATGCAGAAGGTGTTG
GTCGTTTTGGTGATACCCTGAGCAATGCAGATGATTTACTGGCAGCCCTGGATAA
ACTGCAGGATTCACATCCGGGTGTGGATGCACATCGGGCAGAACTGGCAGAAA
CATTTGAAGCAGTTTTCGATCATAAAGCATTTACCGGTCGTAGCGGCACCATGTT
TAGTTATGAAGGTCTGGGTTGTATCTATTGGCATATGGTTAGCAAAGTTCTTCTG
GCAGCACAGGAACTGACCCTGGATGCATTAGATCGTCGTGATGGTAGCATGGCA
GCCTTACGTGCAGCATATTATGAATTACGTGGTGGTATTGGTTTTAATAAGGCAC
CGCAGGAATATGGCGCATTTCCGAGCGATCCGTATTCACATACCCCGAAACATG
CGGGTGCACAGCAGCCGGGTATGACAGGTCAGGTTAAAGAAGAAATCTTAACCC
GCTTTGGTGAACTGGGCATTCGTATTGTTCATGGCACCATTCGTGTGGAACCGC
GTCTGCTGCGCCCGAGCGAATTTCTGACCGCACCTAAAGATTTTGAACATGTTG
CATTAGGTGGTGCCCCGCAGCGTATGTTACTGCAACCGGGTCAGCTGGGTTTTA
CCTATTGTCAGTGTCCGTTTGTTTATGCACGTGGCAAAGTTTTACGCATGGTTGT
TACCTGCCGCGCCAAAAAAGTTTATATTAATGGTCACGATCTGCGCCGTGATCTG
GCAACCCATATTTTTAAACGTAGCGGTGAAGTGTTACGTGTTGATGTTGAAATTC
CGGATAAAGAACTGTTTCATGTT

App. 10.5. Pro_7066

ATGAGCCAAAGCCCAAATACATTAGCTAATGAAGAAACAACAAGTATCGACAAAT
CGATTACTATGGATATGGTTTCTATGAATGGAGAAATGTTTTATAAAATAGCTAAT
AATGATGCTATGCGTCCGTTTTTTATGACCATTGTTAGTGATTCAAATCATTGGAT
GTTTGTTTCAAGTAACGGAGGGTTAACAGCTGGGAGAAAAAATGCTGAATACGC
ACTTTTTCCTTATTATACGGATGACAAAATCACGGAATCTGCAGATATAACAGGTA
GTAAGTCTATTTTTCAAATCCAATACAATAACGAACTTATTGTTTGGGAGCCTTTT
TCAGAACGTTTTACAAATAAATTCAAGATTACCAGAAATTTGTACAAAAACTACTA
TGGTAATAAAATTATTTTTGAAGAAATAAATGAGGATTTAGGTCTGACATACCGTT
ATCAATGGTGCTCTAGCAATCAATTTGGATTCGTGCGAAAATCAGAATTATCTAAT
CATTCTAAAAACGTTTATGAAATTTCACTTTTAGATGGAATTCAAAATATTATGCCT
TACGGGGTGAGTTCAGATTTACAAAGCTCAACAAGTAATTTAGTAGACGCTTATA
AAAGAAGTGAATTACATCCAAAGAGTGGTTTAGGAATTTTTGCTTTAAGTGCTATT
ATTGTGGATAAAGCCGAGCCAAGTGAAGCCTTAAAAGCAAATATAGCCTGGTCTT
TAGGACTAAATAATCCAAAATATTTAGTATCATCGTTACAACTGAATCATTTCAGA
AATGGCAAAAGCATTAGCCCTGAAGATGATATAAAAGGAGAAAAAGGAGCTTATT
TTCTAAATACTGTAATGACTTTAGAGGCAAATACGCAAAAAGAATGGATGATTATT
GCGAATGTAAATCAAGATCATTCTGACATCATTGCAATTACTGAAACCATTCAAAA

TAATAAAAAGATTGCAGAAGATATAAACACCGATATCGAATTAGGAACCAAAAGA
TTAATCGAGTTAAATGCTTCGTCTGATGCTTTGCAACTCACAGCGGATAATTTAAG
AGATACAAGACATTTTTCCAATACACTTTTCAACATCATGCGTGGTGGAATTTTTG
ATAATAATTATCAAATTGAAAAAGGAGATTTTAGCAACTATATAAAAAAAGCTAATA
AATTAGTTTTTGATAAAATTGATCTGAATGCACTTGGTGAAATTTTTTCTTTAAATG
ATTTAAATGAATTTGCTTCTAAGCAAAAAGATGTCGATTTTGATCGTCTAGCTTTA
GAATATTTACCACTAAAATTTAGTAGACGACATGGTGACCCAAGTCGTCCTTGGA
ATAAATTTTCTATCAACACTCAAAGTGAAATTGATGGTTCAAAAGTTTTAGATTATG
AAGGAAATTGGAGAGATATTTTTCAGAATTGGGAAGCTTTAGCACATTCGTTTCC
TAACTTTATCGATAGTATGATTCATAAGTTCTTGAACGCTTCTACTTTTGATGGTTA
TAACCCTTACCGAGTCACTAAAGAAGGATTTGATTGGGAAACTATAGAGGAGGAT
AATCCATGGTCTTACATTGGATATTGGGGAGATCATCAAATTATTTATTTGTTGAA
GTTTTTAGAGTTTATTGAAAAGCATCAACCTGGTAAACTGCATTCGTATTTTGAAA
GTGAATGTTTTGTTTATGCAGCAGTTCCTTACACTATAAAACCTTATGAAGAGATT
TTAAATAATCCAAAGGATACAATAGGTTATAATCACGAGTGGGAAAAAGTAATCAA
CGAAAGAAAAAAATCAATAGGAGCAGATGGAGCCCTACTAAAAAGTAATGATAAA
TCCATTTATCACGTTAATTTTATTGAAAAAATATTAGCAACAGTTTTAGCTAAAATG
TCCAATTTCATTCCAGAAGCTGGAATTTGGTTAAACACGCAAAGACCAGAATGGA
ACGATGCCAACAACGCATTAGTTGGAAATGGAGTTTCAATGGTAACGCTCTATTA
CTTAAGAAGATTTCTAAAGTTCTTTGACCAACTTTTAGAAAATTCAACCTTAGAAA
ATATAAAAATTTCTAACGAGATGGTCGAATTTTACCATAAAGTCAGAGAAACTTTG
ATGGAAAATCAACACTTACTAGCTGGCAGTATTTCTGATACAGACCGAAAAGTAA
TTTTAGATAAATTAGGAAATGCAGCGGCAGATTATCGTTTTCAAATTTACAACAGT
GGATTTTGGGGTAAAAAACGTACGCATTCTATGCAAGGTTTGAAAAATTTCACCA
AAGTAAGTTTGCAATTTATAGACCATTCTATAAAAGCAAATCAAAGACCAGATAAA
TTATATCACGCGTACAATTTAATGTCTGTTGAAAAGAATAAAGAAATTGCAATTTC
TTATCTAAGTGAAATGTTGGAAGGTCAAGTGGCTGTTTTAAGTTCAGGTTTTTTAT
CCTCAAAAGAAAATTTAGCCGTTTTAGATGGCTTAAAAAACAGTGCCCTTTTTAGA
GAAGACCAATACAGTTATTTGTTGTATCCTAACAAGGAACTTCCAAAATTTTTAGA
CAAGAATACAATTTCTAAAGAAGCTGTTTCAAAATCAGAGCTACTTTCACTATTAG
TTTCTAAAAGCAACAAACAAGTTATTGAAAAAGACAGTATTGGTGAGTATCATTTT
AATGGAGAATTCAACAATGCCAGTAATTTAAAGCAGGCCTTAGAAGACTTGAGTC
AACAAAATGAGTACAAGGACTTAGTAGCGAAAGAATCCAAAACAGTAGAAGCTAT
TTTTGAAGACGTTTTCAATCATAAAGCATTTACAGGTCGTTCTGGGACTTTTTATG
GTTACGAAGGCTTGGGTTCTATTTATTGGCACATGGTTTCAAAATTGCAATTAGC
CGTTTTAGAATGTTGTTTAAAAGCTGTTGAAGAAAAAGAATCAGAAGAAGTTATAG
GAAGACTTTTAGAGCATTATTACGAAATAAATGAAGGCATTGGTGTGCACAAATC

GCCATCTTTATATGGAGCTTTTCCAACAGATGCATATTCACACACGCCAGCAGGA
AAAGGAGCACAGCAACCTGGAATGACAGGTCAAGTGAAAGAAGATATTTTAAGT
CGTTTTGGTGAACTAGGAATTTTTGTAAAAAATGGTTGTTTAGAACTTAACCCATG
TCTTTTACGAAAAGATGAGTTTTTAAAAGAAGCAAAAACGTTTGACTATGTAACGG
TTAATTTTCAACATCAATCGTTAGAATTAGTAGAAAAGTCATTAGCATTTACATATT
GTCAAATTCCAATTATATACAAAATAGCTAATCAGAAATGTATTGAAGTGTTTACT
AACGACGGAAAATCTGCAAAAGCAGCAAGTTTAATCTTAGATAAGCAAACTAGCC
AAGATGTTTTTGGACGAACTGGAATTATTAACAAAATCGAAGTATCTATTTTAGAA
AGTGATTTAAGATAA

App. 10.6. *Pap*P

ATGCAGCCGTATTATTTTGACAATCAAGGACGTTTTGTGATTGAAAATTTTGCGCA
TGCCAAGCCCTTTTCCAGTTTCTTGCCCGGCATTGCTGGTGTTCAGGGGATTCC
CCTGTGGGCGTATTATGTTAATCGTGGACAAGGAATTGCGAGCTTTGGTGTGGA
GGATAAAAATGGAGCGATTATGGAATTTTTTCCGGCTAACCGCAGCTATAGCCTC
GTACCGATCCATGGCTTCCGTACCTTTTTAAAGGTAAAAAGGCAGGATGAGTGG
AGCTATATGGAGCCGTTTTCGGCTGTGCGTCCTTTAGAGCATGAGCAACAAAAAA
TGTTAATATCCCGCCATATGCTGGAAATTAATAGCTCTCATCCAGATATAGGGCT
GGAGGTGCGTGTGCAATATTTTACGCTTCCTCATGAAAGCTACGCTGCTTTGGTG
CGCAAGGTTGAGATTGTCAATCATTCCGATCAAGAACTTGAGCTGGAGGTGCTG
GACGGCATGCCGGCTGTGCTGCCCTTCGGTATTGACAATACGGCTTACAAGGAG
CTTGGATATACACTGAAAAGCTGGATGGATGTGTATAACCTCAATAACAGCATAC
CTTACTATAAGGTACGTGGAAGTACAGCGGATACAAGTGAGGTCCATCAGATTG
AGGTGGGACATTTTATGTATAGCTTTGTGGAGCAGGATGGCAAGGCCAGACTAG
TACGTCCCATTGTCGATCCATCCATTATATTTGGTGAAAATACGTCCCTGACCTTC
CCTGACTGCTTTATCCATACACGTGTCGAAGACCTGATATCACAGCAGCCAGTGA
CGACGAATAAAGTGCCTTGTGGGTTTAGCGGTTATGCGGCACGACTGGCGGCA
GGTGCAGCTGTAAATTACTGTTCGGTCATCGGGCATGTGAATGATATTCAATTGA
TCAATGAACGGATTACAGAAATTGCAACGATGTCTTATATCGATCAAAAGGCGAA
GGAAGCTAGAACGTTAGTTGAAGAGTTGACGGATGGAATGGCCACGCAGACATC
CATTCCGTTGTTTGACGAGTACTGCCGTCAGAGCTATCTGGATAATCTGCTACGT
GGGGGTTATCCATTATTGCTGGACAATGGCACGGAGCAACCCTTTGTGTATCAT
GTATTTTCGCGCAAGCATGGAGATCTGGAACGAGACTATAATTTTTTCAAACTCC
AGGCGGGGTATTATTCGCAAGGGAATGGTAATTTCCGCGATGCGAACCAGAATC
GGAGAAATGATGTGTGGTTTAACCCGGGCGTCGGAGATTTTAATATTCGTTTGTT

TATGAGCCTGATTCAGCCGGACGGCTATAACCCGCTTGTTGTAAAAGGCTGTAG
CTTTCAAATACGAGAGCTCGAACCGTTGCTGGAACAGGTGGAAGTGGCGGCCCA
TTCGTCTTTGTGTACTTTTTTTCAATCATCATTTACCCCGGGCGAGCTGATTCAGC
ACATCATTCATGACAATATTGGACTAAAGCAAACGGTAGAGACATTTGTGACCGA
AGCCCTCAAGCATTCCGAGCAATGCTATGAGGCTGATTTTGGTGAAGGCTTCTG
GATCGACCATTGGACTTATAATATGGATCTAATCGAAAGCTATTTAAGTATCTTTC
CAGACAAACAAGAAGAACTCCTGCTCTCGGCCAGGGAGTACATGTATTACGACA
GTCCCGCATGGGTACAGCCGCGCCGAAACAAATACGTGCAGATGAAGGACGGA
CGTATTCGTCAGTACGGAGCGGTTGTAGAAAGCAAGGAGAAGAAAGCTCAACTT
CACATGCGGACCCGCGATGTTCACTGGGTTCGAACCGGACGAGGCAACGGAGA
GATTTACAAGTCCAATTTGTACGAGAAGCTGCTATCCTTGGCTGCCATCAAGCTG
TCTACATTAGATCCAGAAGGCTTGGGAATTGAGATGGAGGCGGGGAAACCAGGA
TGGAATGATTCCATGAACGGACTTCCGGGGCTGTTTGCTTCGGGCTTTAGTGAG
CTATGCGAGTTACAGCGACTCGTAAGATTTTTAATACATGCACAGATGAAGCTTT
CGGCAGACTTTGTGCTGAGGCTACCGCTAGAGGTAGCCAGGTTGATCCAAGAAT
TAGGGCTCGCAGTTGAAGAGTACTACCACGATTTTAATGATGCGGAGCCTGAGA
ATGGTAGTGACGATCATGGCTTCTGGTCACGGATTTCGGATGCTAGAGAAGCAT
ACCGTGAAACGGTTCGCTTTGGATTTGAAGGGACAGAAGCAGCATTATCCTTTGA
TGATGTACTGGAGTTTTTGCAGCGGGCTGAACGTAAGATTCAAGAGGGGATGGA
TAAGGCTCTGCGTATAGGAAAGGGGATATACCCAACCTTCTTTTATTATGAAGCC
AAAGAATTCGAGGTACAGCAGGATGAGAACGGGGGGTTGTTTGTAAATATACATT
CCTTTAAACGCAAGGATATACCGTTCTTCCTGGAAGGACCAACCCGGGCTATGA
AAGTAACTGGCGACATTGTAAATGGGAGGAAGCTGTATCGACTTATCAGACAGA
GCGAGCTTTACGACAAAAAGCTGGGTATGTATAAGGTGAATGAGAGTCTGGCGG
GACAGCCGATTGAACTGGGCCGTTCCAGAGCTTTCACGCCAGGGTGGCTGGAA
AATGAATCGATATTTATGCATATGGAATATAAATATTTACTTGAGTGTTTAAAGGT
GGGTCTATACACTGAATTTTTTGCAGACATCCAAAAAGCGCTACCCCCTTTTATG
GACGCTGCGATCTATGGAAGAAGTACGCTGGAAAACAGTTCCTTTATTGCTAGCA
GTGCTAATCCTGATGCAAGTTTGCATGGAACGGGCTATATTGCGAGGCTGAGTG
GATCAACGGCAGAATTTATTCATATGTGGCTATGGATGATGACAGGCGGGCAGC
CATTTGGCTATGATTCTAGGGGGCTTACCTTGAAGCTGGAGCCCAAGTTGCCAG
GCTGGCTGTTCACAGCTACAGGTGAGTTAACCTTCTGCTTTCTCGGAAGCTGCC
AGGTGACGTATTATAATACAAGTCGCATGGATACTTACGCGCCGGACTGCATAGT
GAAACAGTATATACTGCATTATAGCAATGAGGAACAGAAGACAATAAACCAGGCT
GTATTAGACCAAAAAGAGGCGCAGGATGTGAGAGACCGGCTGGTTTCTCGTATC
GAGGTGCATCTGGAGAGATCGCTTTAA

App. 10.7. *Ps*LBP

ATGGGTCAGAAAGGTTGGAAATTTCAGGGCGAACAGGGAGAATTTCGTTTAGAA
CAGCCGGAACATAATAGCTATCTGTATTTTCCGCTGGTTAATGAAGCAGGTATGA
TGAGTGCCGTTACCCCGAATCTGCATGGTGAAATTACCAGCGGTCATAATACCTT
TCTTATGGAGCCGGTTAGCGCCGAAAGCCTGCATAATAGTAAAGCAAGCCGCAA
TTTTTGGGTTTTTATTGAAGGCTATGGTGCATGGAGTGTTAGCGGTAACAGCGCA
CGTCAAAATGCAGCACGTTTTACCGGTGAAGAAGAACGTAGCGCCGTTGAAGCG
GGTTTTCTGTGGCATGCAGTTACCCGTGAAAATGAAAAAGCAGGTCTGAAAGCA
CGTACCGTTAGCTTTGTGCCGGTTACAGATGATAAAATTGAACTGATGCGTGTTA
CCCTGACCAATACAGGTAATGCACCGCTGAAACTGACCCCGACCGCAGCAATTC
CGCTGTATGGTCGTAGCGCAGATGATTTACGTGATCATCGTCATGTTACCTCACT
GCTGCATCGTATTTTTACCAGTGAATATGGTATTGAAGTGCAGCCGGCACTGAGC
TTTGATGAACGCGGTCATCGTGTTAATAAAGTTACCTATGGCGTTTTTGGTGCAG
AAGCCGGTGGTACCGCACCGGCAGGTTTCTTTCCGGTTACCGAAGATTTTATTG
GTGAAGGTGGTGCACTGGATTGGCCGGAAGCAGTCGTTGCAAATCGTGAACCG
GATGCACAGGCAGGTACCGCGGTTGAAGGATATGAAGCCGTTGGTGCCTTACGT
TTTGCACCGGTTGAACTGGCACCGGGTAAAAGCGTTAGTTATGTTGTTGCAATG
GTTATTAGCGGTGATCGTATTGATGTTGGTCGTTATGCAGCCGATTATCTGGCAG
CAGGTCGTTTTGATGCACTGCTGGAACAGAATCGTGCATATTGGCGCGATAAATT
AGATACCGTTCGTTTTAGCAGTGGTGATGGTGAACAGGATCTGTGGATGAAATG
GGTTACCTTACAACCGATTCTGCGTCGTCTTTATGGTAATAGCTTTCTGCCGTAT
CATGATTATGGCCGTGGTGGACGTGGTTGGCGTGATTTATGGCAGGATTGTCTG
GCACTGATGGTTATGGAACCGGCCGAAGTTCGTCATTTACTGCTTAATAATTATG
CCGGCGTTCGTATGGATGGTAGCAATGCCACCATTATTGGAGCAGGTCCGGGTG
AATTTGTTGCAGATCGTAATAATATTCCGCGTGTTTGGATGGATCATGGTGCCTG
GCCGCTGATGACCACCCTGCTGTATCTGCATCAGAGCGGCGATCTGGATTTACT
GTTTCAGCCGCAGAGCTATTTTCGTGATGTTTTTGTTAAACGTTGTCGTGAACGT
GATGCAAGCTGGACCCCGGAACAGGGTAATAAACTGCTGACCGCCGATGGCCA
GATTTATGAAGGTACCATTCTGGAACATATCCTGCTGCAGAATATCGTTCCGTTTT
TTAATGTGGGTGAGCATGGTAACATTAAACTGGAAGGCGCCGATTGGAATGATG
GTCTGGATCTGGCGCCGGAACGTGGCGAAAGCGTTGCATTTACCGCATTTTATG
CAAGCAATCTGATGGAACTGAGCGAACTGCTGTTAGAACTGCAGAAACGTACAG
GCAAAGATAGCCTTGATATTGCAGAAGAAATGGCACTGCTTCTGGATACCCTGG
GTAAACCGATTAGCTATGATAGCATTCAGGAAAAACGCAGCCTGTTAGATCGTTA
TTATGATGCAGTGACCCCGCGCGTTAGCGGCAAAAAACTTCTGCTGGATATTCG
TAAAGTTGCCGAAGATCTGAAACGTAAAGCAGATTGGGCAGTTGCACACTTACG

CGGTAGCGAATGGATTCAGAGCAAAGAAGGTTATGCCTGGTTTAATGGTTACTAT
AATAATGATGGCGAACGTGTTGAAGGCGATCATCCGGACGGTGTGCGTATGACC
TTAACCGGTCAGGTTTTTGCAATTATGGGTGGTGTTGCAACCGATGAACAGACC
GAAAAAATTAGCCAGGCAGTGAATCGTTATCTGAAAGATGAACGTATTGGCTATC
GCCTGAATAGCCGTTTTGGTGGTATTCAGCAGAATCTGGGTCGTGCCTTTGGTTT
TGCGTTTGGTCACAAAGAAAATGGTGCCATGTTCAGCCATATGACCGTTATGTAT
GCAAATGCACTGTATAAGCGTGGTTTTGTTCAGGAAGGCTTTGAAGTTCTGGATT
CCATCTATCGTCTGAGCGCGGATTTTGAAAATAGTCGCATCTATCCGGGTGTTCC
GGAATATATTAATGAACGTGGTCGTGGTATGTATACCTATCTGACCGGCAGCGCA
AGTTGGCTGTTACTGACCCAGCTGACCGAAGTTTATGGTGTTAAAGGTCGCTTTG
GTGATCTGCGTCTGGAACCGAAACTGGTTCAGGCACAGTTTGATGGTAGTGGTG
AAGCAGCAGTTGAAACCCTGTTTGCCGGTCGTATGCTGCGTGTTGTTTATCGCAA
TCCGCAGGCAGCAGAACATGGTCAGTATCGTGTTGATAGCGTTAGCCTGAATGG
TCAGAGCGTTGATTGTCAGAATGACGGTGCAGGTTGTTTAATTGGCCGTAGCCT
GATTGAAGCATTACCGGCAGATGGTGTTCATGAGCTTATCGTAACTCTGGGCCG
CAACATTTCA

App. 10.8. *Prym*P

ATGGACCGCGTATATCAACCATTTGGAGAAGAGGAATGTGCGCTGCGCGACGAG
GAACGTGTAGCTCGCGTGCAACCGATCTTCCGCGCATCAGCAGCTGGTGCAGC
GATTATGGGTGTATTAGCCTTATTTGGTATGCTGTCTTCAATGACACCGTCGGTT
CCACCCTCATCTCTGCACCTTGCTACGCCCTCTGTTCTGGCGGCCGAGGAAGAC
ACCCGCGTCTTTGATGAGTCCGGGCGCTACATTATGCGTGCCTTTGACCGTGTA
AAGCCTATGGCCAGCTTTCTTCCTGGAGTAGGAGGGTTATGGGGCGTTCCCATG
TGGACCTTTTACGTAAACCGTGGCCAAGGGCTGGCAACATTCGGTGTTAAATCC
AAGGACGGCGGCATCCTGCTGTTCCAGACCGCCGAGAAAGCATATCAGGTCACT
CCTTTCGTTGGGTTCCGTACTCTGTTAAAAGGGAAACGTGCTTCGGGCAAGACG
TTCGAGGCACAGCCCTTTTTCCCCCAATCTGATGCAGATGAGACCACTACCGCC
CGTCGTGATATGTTTATTGGGAATAATGAAATGGAAATTGAAGAGGCTGATCCAG
TAAGCGGCGTTCGTACTAACGTAGTCTACTTCACGGCGCCGAATGAGGCATTCC
CGGCATTAGTTCGCCGTGTGACTTATACGAACGAAGGAGATGATGCCGTCGAGT
TAGAAGTAGTTGACGGATTGGCGCGCTTGGAACCTGTTGGAACATCCCTTTTCAT
GTTGAATGCGATGGGTCGCACATTGGAGGGGTGGATGCACGTTTACAATTTTGA
AACAGACCACACTGCCCCATTCTTTCACCTTGTAACCGCTCCTGCCGACACCGC
CGATGTGACGATGATTAAGGATGGCTACTTTGCAGCGGCATTCTTAGACGATGA
AATGCTGCCGATGATCTGCGATCAGCAATTACTGTTTGGAACAGATACTACTTTA
GCAGTACCACGTCGTTTTTTCTCGTCTGCAAAACGCCCCGGCGGTGCCCCTCTT

GCGGAATTATTAAAGGCTCCTCAAAGCACTACTTCTCGTACCCCGTCGGCTTTTG
CCGCAGCAACCATCTCATTAAAGCCCGGAGAGTCCAAGACGATTTCGATTGTGT
ACGGTTATGCAAGTGATATTGATAGTTACCTTTCGACAGTCCTGCCGAAATTGAA
AGCGGACGAATATTTCAGCTTGAAGCGCGAAGAAGCCCAAAGTCTTGGTAAGAA
CTTGACCGAGCGTGTTGCTATGTCCAGTGGAGTGCCAGTCATGGACAATTATATT
AAGCAGAACTACCTTGATAATCTGCTGCGCGGCGGAATGCCTGTCCCCATTGGT
AATCCGGATCAGCCAAAGATCTACCACGCATACAGCCGTATCCATGGTGACTTG
GAGCGTGATTATAACAACTTCGTATTGGAGCCATCCTATTGGTCTCAGGGGCCG
GGGAACTTTCGTGATGTGAATCAGAATCGCCGTTGCGATGTATTGCAGTTACCAT
CGGTCAAGGATTTCAATGTGCGTCAGTTCTTCTCCTATATTCAATCCGATGGCTA
TAATCAACTGACAGTGGCAACGGCGTTTTTTAAGATTCACGATGTTTGGCGCGTT
AATGAGGTCGTTGAGCAATTGTTAGCACCCGGCGACATGCAGCAAAAATTAAAG
GAAAAGCTTTTGGCCCCTTTTCGCCCAGGACAACTTTTTACGGATTTGAATAACG
CCGGAGTTACGTTTAAGCTTCCACTGGATAAGGTTTTGGAGATCGTCACACGTTA
TGCGGAACAGGTTCCTGCGGGGCAATACTCTCAAAACGGGTTCTGGACTGACCA
TTTCACTTATCACTTGGATTTGGTTTGGAACTTCCTGGCAGTCTTCCCTGACCTGA
AAGAATGGTTACTGTACGACTCCAACACCCTGCCATTTTTTCTTAGCCCGGGCCG
TGTGGCAAATCGTACCGAGAAGAACATGCTGGTGGCTGAACCGAGCACTGTCCG
CCAGTATGACGCTGTACGCGACTCGGGAGCTAAACAGCAGCAGTTACAGATTAT
TTATGGTGACCCAAACTTCGTAGGAGATTGGGGAGGTGGCGGTACGTGGCAACT
TACGGCCTCTGGAACCACAATGAAAGTCTCTCCCTTGGCAAAACTTGTCGTTCTG
GCAGCAAACAAGTTCTGCATTTTAGACCCTTTGGGGATGGGGATCGAAATGGAA
GCAGGAAAACCGGGGTGGAACGATGCTATGAATGGATTGCCAGCGCTGTTTGGT
TCCGAAATGCCAAGTGCATATGAGTTGCACGAGATCGTGGATTTTGTCGGCGTA
GCTATTGACGAGGCTGCTCGCGCAGTATCCTTGCCGGAGGAGGTCGCGGCATT
GCTTGATGCGATCGACGCGCAATTACGCGCTGTTTTAGGAGGGTCAACCAGCGA
CTTCACCTATTGGGATAAAGTACATGATGCATTGGAAGCCTACCGTACGAAGACC
GAGGCCACTTTCACCGGAACTATGGTGAGTCGCTCTGCAAAGGAGATGGGAAAG
GCGACCGGGGTATTCGGTCGCATGTTGGCTCGTATGGACGTGGGGATTGCCCG
TGCGCTTTCCTACTCCACGGACCCGCATAAAAAGGTAAGTCCTACGTATTTTCGT
TTTAATGTCAAAAGTTTTAAGTTGGTGGGAGTAAGCGGGCGCGGACTGCCCACT
GTCCAGGTGACTGAGTTCGATGAGGGCCAACCTCTGCCGCTTTTCCTGGAAGGT
CCCACCCGCCACCTGAAAACGCTGAAACGCGCCTCACAGGCAGATAAGCAGGC
GGTCTACGACGCGGTAGCAGCTAGCACCCTTCACGATAACGAGTTAGGCATGTA
CAAAATCTCGGAGTCGCTTAAAGGCCAGCCATTGGAGATCGGTCGCATGATGGC
ATTTGACAGCGGATGGCTGGAAAACGAGTCGATCTGGCTGCATATGTCTTATAAA
TGGTACTTAGAACTTCTTCGCGCGGGTCTGTATGCGCAATTTTTTGAGGAGATTC

GTAAGGGGGTGGTCTGTTTTCTTGATACAAAAGTTTTCGGGCGCAGCCCCCTTG
AAGCGACATCATTTATCGTCTCCTCGGCCTTCCCAGATAAAGCCCTGCATGGGT
CAGGATTTTTGGCTCGTTTGTCGGGAACCACTGCGGAATTTTTATCAATGTGGAA
CCATATGATGGCAGGAGCAGCTCCTTTCAGTCTTAATGCAGACGGTAAGTTACAA
TTAGCCCTGCGTCCGATTGTTGCCGACTGGATGTGGCGTGACGACGGAACAATC
AACTTCAAATTTCTGGGAAATATCGACGTTACCTATGTCTTGCATAGTAAAAAAAA
CTCATGGGAGGCCACGGTCCAGTCCTACGTCCTGCACGCTGCTGACGGAGATC
TGAAAATTGACGGGGAAGTCATTCCGATGCCAGCAGCGGAACAAGTCCGCAACA
TGAAATACTCGGCCATCACTGTTTTTTTGTCATGA

App. 10.9. OcP1

AACCGTACCAGCAATAATGAATATCAGGCCATTCCGGGTGATGAAACCAATCTTG
AATTACCGAGCTGGAACAATGATATTAATAGCCGTGATAGTGCCTTAAGCAGTCA
GAACGTTAAAAATCCGAGCCATAGCAAACGCCTTAGCCTGATTGGCGTGAGTGT
TATTTTATTTGGTATTTTTGCATACGGTACATATCTGGCATTTAGCGCAAGCTTAC
AGCCGCCGTTACTGGCCGATACCAATAGCGAATTTGATGCACAGAATCGTTATAT
TCTGCGCGATTATGATCTGACCCGCCCGATGAGCAACTTTCTGAATGGACTGGC
GACCGTGTGGGGTATTCCGCTGTGGGCATTTTATGTTAATCGTGGTCAGGCAAT
TACCAGCTTTGGTAAACAGAATAAAGATGCAGCAATTGCGAAATTTGTGACCGCC
GAAAAAGCGTACTTTCAGACCCCTTTTACCGGTTTTCGTACATTTCTGAAAGGTA
AACGTAACGGTGATAGTTGGAATCATATGCCGTTTTTTCCGACAAGCGAAGAAAA
ACGTGCAAAACTGCAGCGCAATATGATGATTGGTCTGAATGAACTGGAAATTGAA
GAGGTTAGCAGCGAACATCAGCTGCAGACAAATATTCTGTATTTCGTTGCACCG
GGTCAAGATTTTCCGAGCCTGGTTCGCCGTACCACATTTACCAATCTGGATAGCT
ATAATAGTCTGGAATTAGAAGTTCTGGATGGCTTAGCGCAGCTGGTTCCTAATGG
TCTGGGTAACAGCGCCATTGATAATATGGGTCGTACAATGGAAGCATGGATGAA
TGTTTATAATGTTGGCTTTACCCAGGGTACAGACACCAAAATCACACAGCCGTTC
TTTCATATTTCTCAGGGTACCGCAGATAGCGCCCAGGTTCAGATTGTGCGTGATG
GCTTTTTTAGCTTAGCATATGTTGAAAAGACCAGCGATAAAGCAAGCACCAAAGC
CAATGCAATTGATAGCGAAAACCTGCATGATCTGTTACCGTTTGTTGTTGACCCG
AGCGTTGTGTTTAGTACCGATACAACCCTGACCAATCCGTCAGGTTTTTTCGATT
TTAGCGGTAGTGTTGAAGAACTGACCCAGAGCAGCCAAGGTACCACCAGCCGTA
CCCCGTGTAGCTTTGCGGGTGTTCGCGTTACCATTCCGCCGGGTGCAAATGTTA
CAGTTACCAGCGTTTATGGTTACGCAGAATCACTGGAGACATTAGTTGGCAAATA
TAGCCCGATTGTTCGCGAAGTTAAATATAGTCGTGATAAACGTCAGCTGGCGTAC
GATTTTGTGGCAGATATTACCAAACGCGTGGATACCAAAACCAGCTCCGATGTGT

TTGATGCCTATGTGAAACAGGATTTTTTAGATAATTTTCTGCGTGGTGGCCTGCC
GTTGCTGCTGGGTGGTAAAACCACCGCAGGTAATAGCGTGACCAGCCCGAGCA
GCAAAGTGTTTCATGTGTATAGCCGCATTCATGGCGATATCGAACGTGATTATAA
CTATTTTCAGATCGATACCACATATTTTTCACAGGGTCCGGGTAACTTTCGTGAT
GTTTGTCAGAACCGTCGTGTTGATGTGTCACATAGCCCGTTTGTGGGTGATTTTA
ATATTCGCCTGTTTCTGAGTTTTATTCAGAGCGATGCATTTAATCCGCTGACCATT
GCATCAACCGTGTTTAAAGTGCCGAATACCCAGCTGGAATTCGTTCTGGACTCAC
TGAAAATTCTGAATCCGGACGGTAGCTTAACCAATAATGGTGTTGGTAGTGGTGG
TCTGCATCGTGAAGCCGCACGCACCCTGCTGAGCAAAGCATTTCGTCCGGGCCA
GTTTTTCAAAGATGCCGCAACCGCCGGTGTTTCTTTTGCAATTAGTAAAGAAGAA
GTTGCAGACATTATCATTGGTTTTGCAGTTCAGGACTTTGCCGGTCAGTTTAGCC
AGAATGGTTATTGGAGCGATCATTGGACCTATATTTTAGATCTGGTGGATAACTA
TCTGACCGTTTTTCCTGATAAAGAAGCACAGCTGCTGTGGGATAGTGAACCGGTT
GCATTTTACGTTAGCCCGGCAGTTGTTAAACCGCGTCATAGCCGTTATGTTGCAA
TGGATAATCCGGCAAAACCGGGTAGCAGCGTGCTGCGTGTTTACAATGCCATCA
GTGTTTGGGGTGATAGCCAGTTTCCGGTTGAAAAAACAAATGCAATGAACGCAAT
TTTTCAGGATCCGAATTATCTGGTTGATGTTAATGGTGCAGGCTCAGTTTGGCAG
CGTAGCGTTAAAGACAATAGCGTTGTTCGTGTTAGCGTTATTGCCAAGCTGCTGC
TGTTAGGTATTGTTAAATTTAGCACCCTGGATCCGTATGGTTTAGGTGTTGAAAT
GGAAGGTGGTAAGCCGGGTTGGAATGATGCGATGAATGGTTTACCGGGTTTACT
GGGTAGCGGTATGCCGGAAACATATGAAATGTTGCGTATCCTGCGTTATACCCA
TAGCGCATTACTGAAATTTAGTCGTCCGGTTAGCTTTCCGAGTGAATTTGCAGAT
TTTCTGACACAGTTAGCCGCAGCGATTGATCGCTATAATTCTAGCCCGAAACAGC
TGGCAGATGAATTTGTTTATTGGGATAGCGCAAATACCGCACGTGAAACCTATCG
TGCAACCGTTGTTGCAACCTTCAAAGGTGAATTTAAAAGCCTGGCAGCAAGTGAT
ATTGTGGTTCTGCTGGAAAAATTCATTGCCAAAGTTGAGGGTGGCATTCAGCGC
GCACTGGCAGTGAATTCAAATAACGGTTTTCTTAGCCCTACCTATTTTTATTATGA
ATGTAGCCAGGGTAATTATGAGATCAGCACCGATGGCGTGCAAACCAGCATCAT
TGCAAAGAGCTTTGAACTGCGTACCCTGCCTCTGTTTTTAGAAGGTCCGACCCGT
CATCTGCGCGTTGTGCAGACCGTGGAAGAACGTCGTAGCATTTATGAAAAAGTT
AAAAGCAGCGCACTGTATGACAGCGCCCTGAAAATGTATACCCTGTGTGAATCAT
TAGCAGCCATGGGTCAGGAAGTTGGTCGTATGAAAGCCTTTAGCCCTGGTTGGT
TAGAAAATCAGAGCGTTTGGCTTCACATGAGCTATAAATTTTATCTGGAACTGCTT
CGTGGCGGTTTATATGAAGAGTTCTTTAGCGAAATGGCCACCGGTCTGGTGCCT
TTTATGGACAATAAAGTTTATGGCCGCAGCCCGCTGGAAGCAGCAAGCTTCATT
GTTTCAAGCGCATTTCCGGATAAAAAATTACATGGTGCAAGCTTTCTGGCACGCC
TGAGCGGTAGCACCGCAGAATTTCTGAGCATGTGGCTGCTGATTGTTAGCGGTC

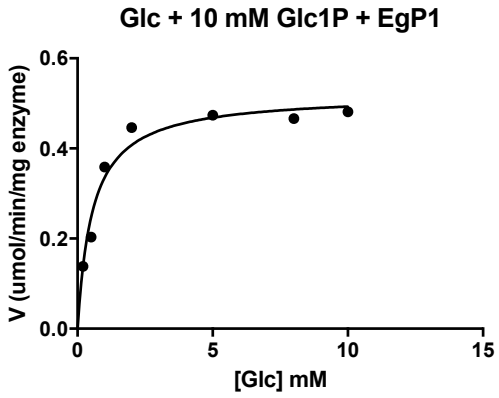
ATCAGCCGTTTAGCGTTGATCCGCAGACCAAAGAATTACTGCTGTCACTGCAGC
CGATTCTGCCGGGAAATTTTTTTGATGAGGATGATGAAGTTAGTTTTGTGTTTCTG
GGCAAAGTGGATGTGGTGTACCATAATCCGAGTCGCGAAGATACCTGGAAAATT
AGCGCCAAACGTGCCACCGTGACCAAACTGGATGGTAGCGTGGTTGAAGCAAG
CGATGCCGTTATTCGTGGTGATGTTGCACTGTTAGTTCGTAATCTGCAGGCAAAA
CGTGTGGATGTTTATTTTTAAACGAACCACCCTACGATTCCTCAGTACACCCCCC
ACCATACACCTCATCATACCCCGTTATCATTATCCGTCCACATTTTAATCTTCGAA
AGCGTATTCTGACGCTTATTTTTGTTGCTTAGCAATCGTACCATGGAGCGCCTTG
ACCCCCCTTGCCCATTGGTAGTAATGCTGTGTTACGGGTCTAGTCACTATTGACG
CTCGATCCTGTGACTTAAAAAAGATAATAACAATAATAACAAA

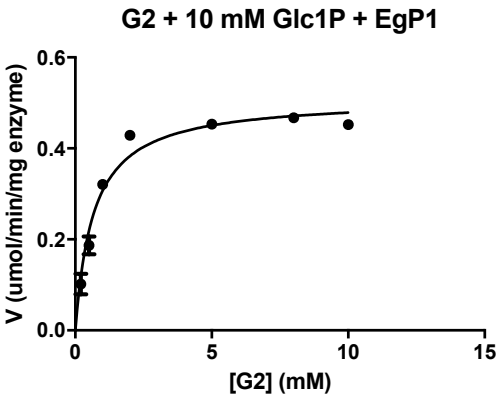
Appendix 11. Primers for PCR amplification and cloning into PopinF vector for gene expression in *E. coli*

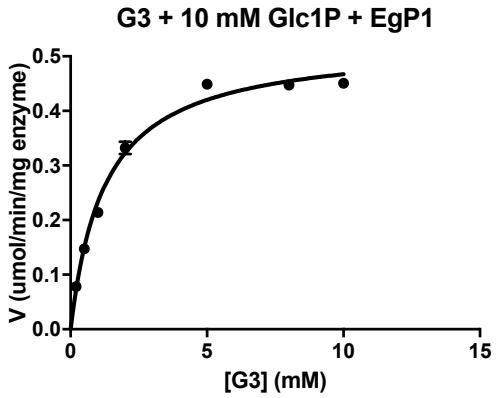
Primer direction	Primer sequence (5'→3')	Genes
Forward	AAGTTCTGTTTCAGGGCCCGATGGGTCAGAAAGGTTGG	*PsLBP*
Reverse	ATGGTCTAGAAAGCTTTATGAAATGTTGCGGCCC	*PsLBP*
Forward	AAGTTCTGTTTCAGGGCCCG ATGTGGCATC	*EgP2-4*
Reverse	ATGGTCTAGAAAGCTTTAAAAATCTAACAGAACACTATCCGG	*EgP2*
Reverse	ATGGTCTAGAAAGCTTTAAACATGAAACAGTTCTTTATCCGG	*EgP3-4*
Forward	AAGTTCTGTTTCAGGGCCCGAACCGTAC	*OcP1*
Reverse	ATGGTCTAGAAAGCTTTATTTGTTATTATTGTTATTATC	*OcP1*
Forward	AAGTTCTGTTTCAGGGCCCGCCTAGCAG	*EgP1* (truncated N-terminal signal peptide)
Forward	AAGTTCTGTTTCAGGGCCCGATGTGGCATCGTGCAGTTA	*EgP1* (full length)
Reverse	ATGGTCTAGAAAGCTTTAAAACAGAACTGCACTCGGAA	*EgP1*
Forward	AAGTTCTGTTTCAGGGCCCGATGCAGCCGTATTA	*PapP*
Reverse	ATGGTCTAGAAAGCTTTATTAAAGCGATCTCTCCAG	*PapP*
Forward	AAGTTCTGTTTCAGGGCCCGATGGACCGCGTAT	*PrymP*
Reverse	ATGGTCTAGAAAGCTTTATCATGACAAAAAAACAG	*PrymP*

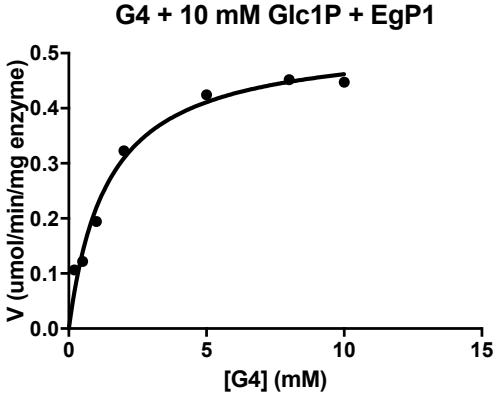
Appendix 12. Michaelis-Menten kinetic plots for the enzyme activity against different sugar acceptors in the presence of 10 mM Glc1P that were used to calculate the reported k_{cat} and K_M values reported in chapter 2, 3 and 4. The rates (v) were obtained from discontinuous assays which were performed in the presence of 10 mM of Glc1P and 0.2-10 mM of Glc-G6 as substrates (as indicated on each plot) in the presence of 25 μg/ml of the indicated enzymes. The reactions were incubated for 10 minutes with the assumption that the rate obtained at the end point is closed to initial rate of the reaction.

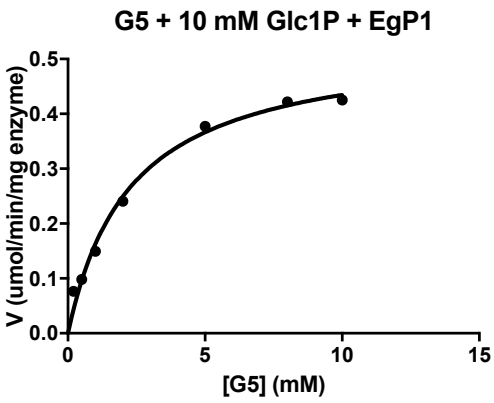
App. 12.1. EgP1

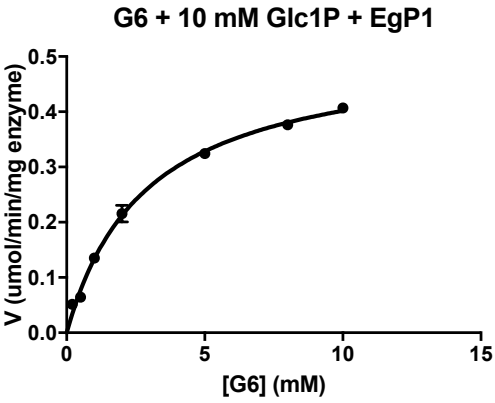




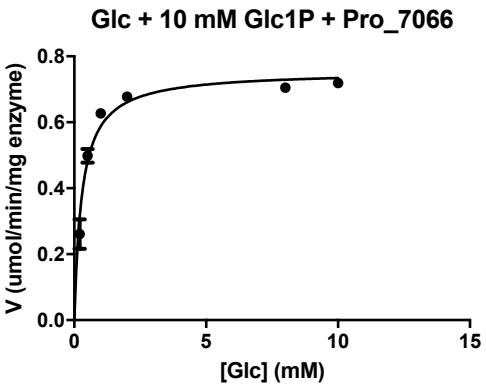


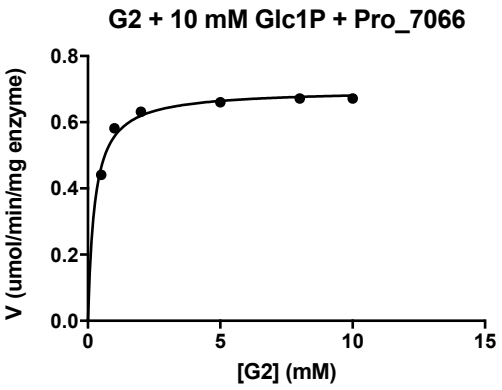


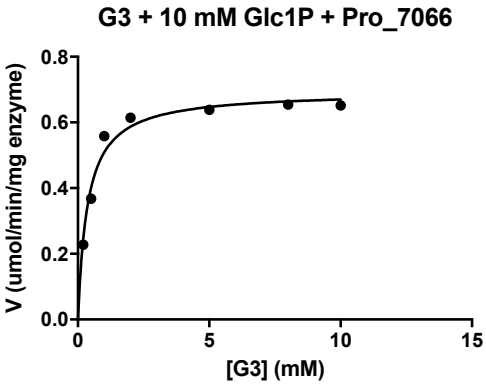

G5 + 10 mM Glc1P + EgP1
V (umol/min/mg enzyme)
0.0
0.1
0.2
0.3
0.4
0.5
0
5
10
15
[G5] (mM)

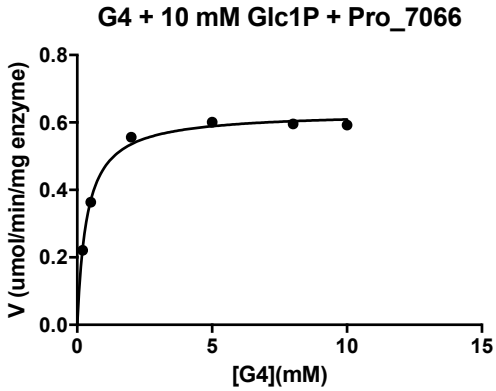

G6 + 10 mM Glc1P + EgP1
V (umol/min/mg enzyme)
0.0
0.1
0.2
0.3
0.4
0.5
0
5
10
15
[G6] (mM)

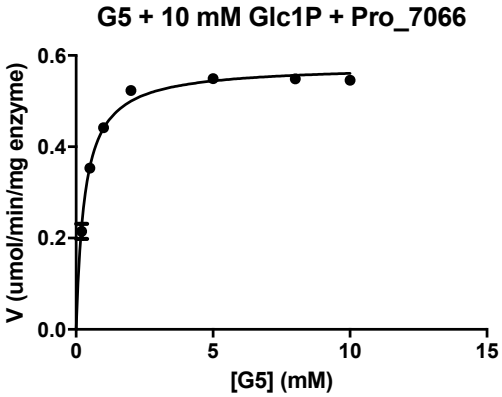
App. 12.2. Pro_7066

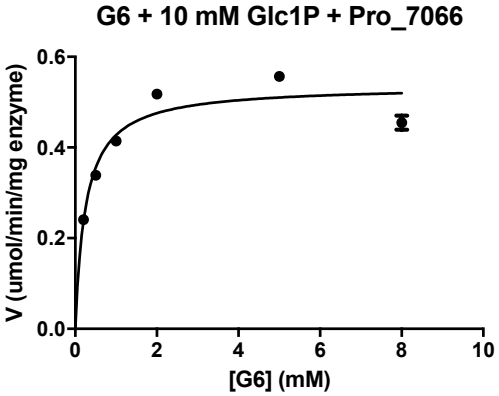




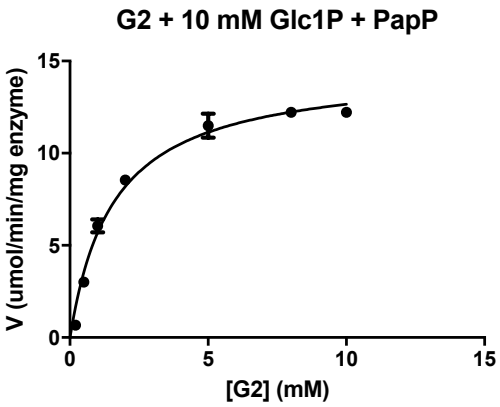


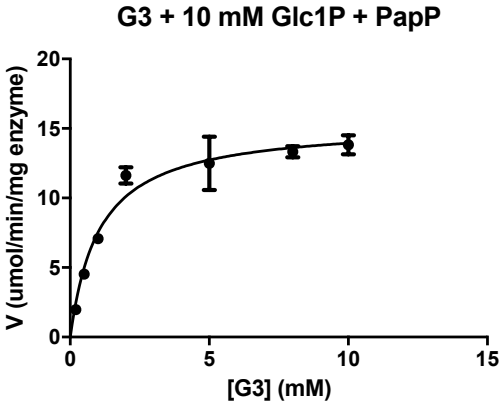


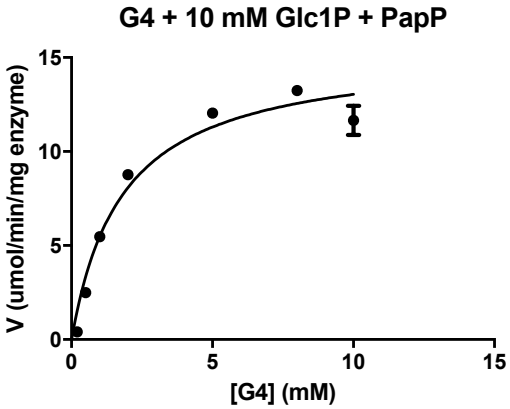


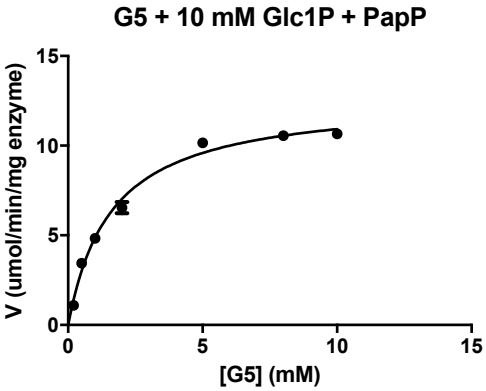


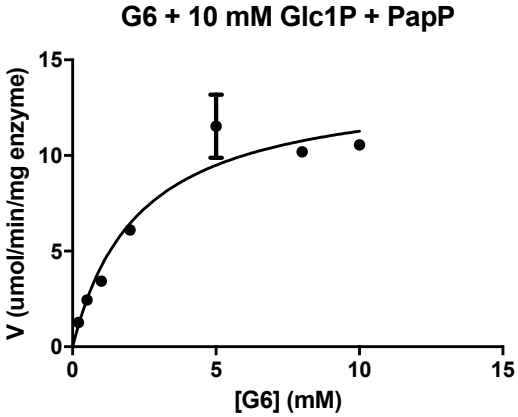
App. 12.3. PapP


G2 + 10 mM Glc1P + PapP
V (umol/min/mg enzyme)
[G2] (mM)
0
5
10
15


G3 + 10 mM Glc1P + PapP
V (umol/min/mg enzyme)
[G3] (mM)
0
5
10
15
20


G4 + 10 mM Glc1P + PapP
V (umol/min/mg enzyme)
[G4] (mM)
0
5
10
15


G5 + 10 mM Glc1P + PapP
V (umol/min/mg enzyme)
[G5] (mM)
0
5
10
15


G6 + 10 mM Glc1P + PapP
V (umol/min/mg enzyme)
[G6] (mM)
0
5
10
15

References

[1] P.-E. Jansson, J. Lindberg, G. Widmalm, *Acta Chem. Scand.* **1993**, *47*, 711–715.

[2] S. Roland, in *Bioinforma. Glycobiol. Glycomics*, Wiley-Blackwell, **2009**, pp. 311–320.