Characterization and diversity of the complete set of GH family 3 enzymes from *Rhodothermus marinus* DSM 4253

Kazi Zubaida Gulshan Ara1*, Anna Månberger1, Marek Gabriško2, Javier A. Linares-Pastén1, Andrius Jasilionis1, Ölafur H. Friðjónsson3, Guðmundur Ó. Hreggvíðsson3,4, Štefan Janeček2,5 & Eva Nordberg Karlsson1*

The genome of *Rhodothermus marinus* DSM 4253 encodes six glycoside hydrolases (GH) classified under GH family 3 (GH3): *RmBgl3A*, *RmBgl3B*, *RmBgl3C*, *RmXyl3A*, *RmXyl3B* and *RmNag3*. The biochemical function, modelled 3D-structure, gene cluster and evolutionary relationships of each of these enzymes were studied. The six enzymes were clustered into three major evolutionary lineages of GH3: \( \beta \)-N-acetyl-glucosaminidases, \( \beta \)-1,4-glucosidases/\( \beta \)-xylosidases and macrolide \( \beta \)-glucosidases. The *RmNag3* with additional \( \beta \)-lactamase domain clustered with the deepest rooted GH3-lineage of \( \beta \)-N-acetyl-glucosaminidases and was active on acetyl-chitooligosaccharides. *RmBgl3B* displayed \( \beta \)-1,4-glucosidase activity and was the only representative of the lineage clustered with macrolide \( \beta \)-glucosidases from Actinomycetes. The \( \beta \)-xylosidases, *RmXyl3A* and *RmXyl3B*, and the \( \beta \)-xylosidases *RmBgl3A* and *RmBgl3C* clustered within the major \( \beta \)-glucosidases/\( \beta \)-xylosidases evolutionary lineage. *RmXyl3A* and *RmXyl3B* showed \( \beta \)-xylosidase activity with different specificities for para-nitrophenyl (pNP)-linked substrates and xyloooligosaccharides. *RmBgl3A* displayed \( \beta \)-1,4-glucosidase/\( \beta \)-xylosidase activity while *RmBgl3C* was active on pNP-\( \beta \)-Glcp and \( \beta \)-1,3,1,4-linked glucosyl disaccharides. Putative polysaccharide utilization gene clusters were also investigated for both *R. marinus* DSM 4253 and DSM 4252T (homolog strain). The analysis showed that in the homolog strain DSM 4252T *Rmar_1080* (*RmXyl3A*) and *Rmar_1081* (*RmXyl3B*) are parts of a putative polysaccharide utilization locus (PUL) for xylan utilization.

Marine extremophilic biotopes, such as hot springs and hydrothermal vents, harbour diverse microbes hitherto underexploited and unexplored. Recent genomic studies show that many of the species, especially those found in coastal geothermal areas surrounded by profusion of carbohydrate rich biomass (seaweeds as well as terrestrial species), contain a wide array of novel glycoside hydrolases (GHs)1,2. Thermostable GHs have numerous applications in different fields, making marine thermophiles targets for prospecting of industrially interesting enzymes3,4. *Rhodothermus marinus* are Gram-negative marine thermophilic bacteria, previously classified under the phylum Bacteroidetes, but recently assigned to the new phylum Rhodothermaeota5. The type-species was isolated from a coastal hot spring on the North-West coast of Iceland and has an optimum temperature of 65 °C and is slightly halophilic6. *R. marinus* can utilize a variety of sugars as carbon sources and produces a wide range of GHs6–12. Sequence analysis shows that the *R. marinus* genome contains a large number of genes encoding GH enzymes, many of which are secreted extracellularly; yet, several of them appear to be attached to the cell surface13. These putative enzymes include six members of GH family 3 (GH3) (*RmBgl3A*, *RmBgl3B*, *RmBgl3C*, *RmXyl3A*, *RmXyl3B* and *RmNag3*), which have not yet been studied.

1Division of Biotechnology, Dept. of Chemistry, Lund University, P.O. Box 124, SE-221 00, Lund, Sweden. 2Laboratory of Protein Evolution, Institute of Molecular Biology, Slovak Academy of Sciences, Dúbravská cesta 21, SK-84551, Bratislava, Slovakia. 3Matís, Vínlandsleið 12, IS-113, Reykjavík, Iceland. 4Faculty of Life and Environmental Sciences, University of Iceland, Askja, IS-101, Reykjavík, Iceland. 5Department of Biology, Faculty of Natural Sciences, University of SS Cyril and Methodius, Nám. J. Herdu 2, SK-91701, Trnava, Slovakia. *email: zubaida.gulshan_kazi@biotek.lu.se; eva.nordberg_karlsson@biotek.lu.se
Glycoside hydrolase family 3 (GH3) is intriguing due to possible roles in cellulosic biomass degradation, bacterial and plant cell wall remodelling, recycling and in pathogen defence\textsuperscript{14–17}. GH3 members are retaining enzymes, capable of hydrolysing the terminal glycosidic bond in the non-reducing end of a number of glycosides and glyco-conjugates\textsuperscript{18}. According to the Carbohydrate-Active enZyme (CAZy) database\textsuperscript{19} (http://www.cazy.org), this family is widely distributed in bacteria, fungi and plants. It contains more than 27,000 genes (December 2019) encoding putative enzymes and most are of bacterial origin. Members of the family are known for diverse activities, such as β-N-glucosidases, β-p-xyllosidases, α-l-arabinofuranosidases and β-N-acetyl-β-glucosaminidases. This activity spectrum makes the family interesting for biotechnological applications such as degradation of renewable resources for biofuel production\textsuperscript{20,21}. As sequenced-based families group together GHs of different specificity\textsuperscript{22}, without biochemical investigation it is difficult to resolve substrate specificities for individual enzymes. Looking at the CAZy database the total number of biochemically characterised enzyme is significantly less compare to total gene numbers\textsuperscript{23}.

In this study, we report biochemical characterisation of the complete set of GH3 enzymes from \textit{R. marinus} DSM 4253. The hydrolysis data show that the six enzymes encoded in the genome have non-redundant substrate specificities which will help in future characterisation of other enzymes from this family. This study gave us significant insights into novel structural features of the enzymes as well as on corresponding loci in the \textit{R. marinus} genome.

**Results**

**Gene identification and cluster analysis.** Six genes encoding glycoside hydrolase (GH) family 3 (GH3) enzymes were identified in the \textit{Rhodothermus marinus} DSM 4253 genome by BLAST analysis, and designated according to the locus tags of their homologs in \textit{Rhodothermus marinus} DSM 4252\textsuperscript{7} (Supplementary Table 2). The positions were in accordance with those of the type strain, and as indicated by the gene number, the four genes \textit{Rmar}_0536 (encoding \textit{RmBgl3A}), \textit{Rmar}_0925 (encoding \textit{RmNag3}), \textit{Rmar}_2069 (encoding \textit{RmBgl3B}) and \textit{Rmar}_2616 (encoding \textit{RmBgl3C}) were found at different positions in the genome, while \textit{Rmar}_1080 (encoding \textit{RmXyl3A}) and \textit{Rmar}_1081 (encoding \textit{RmXyl3B}) were located adjacent to each other (Fig. 1). Clustering with other GH-encoding, regulatory orpor-secretory genes, was investigated by analysing neighbouring genes in both strains.

For \textit{RmBgl3A}, \textit{RmNag3} and \textit{RmBgl3B} no adjacent glycoside hydrolase encoding genes were found (Fig. 1A,B,D). \textit{RmNag3} is predicted to be co-transcribed with four intracellular proteins including a 23s rRNA methyltransferase, a putative ATPase and a transcriptional regulator (Fig. 1B). A gene encoding an ABC transporter was located in the vicinity of \textit{RmBgl3B} that may be of importance for substrate uptake (Fig. 1D). Signal peptides were predicted for \textit{RmNag3}, \textit{RmBgl3B} and \textit{RmBgl3C} but not for \textit{RmBgl3A}.

\textit{RmXyl3A} and \textit{RmXyl3B} were located in a putative polysaccharide utilization locus (PUL)\textsuperscript{24,25} for xylan utilization (Fig. 1B). This putative xylan utilization locus is also present in the type strain, consisting of six GHs, including the previously cloned and characterized \textit{GH10} endo-xylanase \textit{RmXyn10A}\textsuperscript{26–29} (encoded by \textit{Rmar}_1069) isolated from the type strain. Table 1 describes the cluster which consists of \textit{Rmar}_1067, annotated as a transcriptional regulator of the Lacl family, and the 14 following genes (\textit{Rmar}_1067 – \textit{Rmar}_1081). Several versions and settings of SignalP were used to predict presence of signal peptides. No signal peptide was found for \textit{Rmar}_1075, \textit{Rmar}_1076 and \textit{Rmar}_1081 (encoded \textit{RmXyl3B}) which were concluded to code for intracellular proteins, while \textit{Rmar}_1071, \textit{Rmar}_1073 and \textit{Rmar}_1080 (encoding \textit{RmXyl3A}) code for proteins containing a signal peptide. The other genes predicted signal peptide patterns with lower score. A putative por-secretion system C-terminal sorting domain (CDD: cl22550 in the Conserved Domain Database (CDD)) was encoded in \textit{Rmar}_1071 and \textit{Rmar}_1078. Three genes, two encoding GHs, and a homologue to the RfaB glycosyltransferase (a family GT4 glycosyl transferase) are located downstream \textit{Bgl3C} but not for \textit{Bgl3A}.

\textit{Rmar}_1074 was predicted downstream \textit{Bgl3C} (Fig. 1E). The two GHs, of which one contains two domains, are classified into GH43 by CDD, but into GH130 by the Carbohydrate-Active enZyme (CAZy) database. A signal peptide was predicted for \textit{RmBgl3C}, while surrounding genes were predicted to be intracellular (lacking signal peptides).

**Enzyme purification.** The activity of all six GH3 candidates was confirmed using the respective cell extract after expression in \textit{E. coli}. Purification by nickel affinity chromatography resulted in enzyme purities of 70–80% for all six enzymes (as judged by SDS-PAGE, Supplementary Fig. 1), which also showed that the molecular mass of each enzyme was in accordance with the theoretical molecular mass. Dependent on the domain composition (see below), the molecular mass of the GH3 enzymes ranged from 66 to 107 kDa (Supplementary Table 2).

**Activity screening.** Specific activity was first determined using α- and β-pNP-linked sugars at 60 °C (Supplementary Table 3), a temperature where all enzymes were shown to be stable. The results showed that \textit{RmNag3} was only active on pNP-β-GlcNAc, which makes it a putative β-N-acetyl-β-glucosaminidase. \textit{RmBgl3A} showed highest hydrolytic activity towards pNP-β-Glc and pNP-β-cellobiose (pNP-β-Cel) with some activity on pNP-β-Xyl. \textit{RmBgl3B} was most active on pNP-β-Glc, but with significant activity on pNP-β-Cel and pNP-β-Xyl, and minor activity on pNP-α-L-Ara. In contrast, \textit{RmBgl3C} only hydrolysed pNP-β-Glc (with low
A. *Rmar*._0536 (RmBgl3A) gene cluster spanning 15 kb

B. *Rmar*._0925 (RmNag3) gene cluster spanning 16 kb

C. *Rmar*._1080 and *Rmar*._1081 (RmXyl3A and B) gene cluster spanning 15 kb

D. *Rmar*._2069 (RmBgl3B) gene cluster spanning 14.4 kb

E. *Rmar*._2616 (RmBgl3C) gene cluster spanning 13.7 kb

**Figure 1.** Gene cluster analysis in *Rhodothermus marinus* DSM 4253. The gene cluster corresponding to GH3 loci in strains DSM 4252 and DSM 4253 is identical. (A) Downstream of *Rmar*._0536 are genes encoding: membrane proteins (*yedZ* and *yedy*); a membrane transporter protein (*mfs*); an arylesterase precursor (*aryleste*); an ABC transporter ATP-binding protein (*yvcR*); and a hypothetical protein (*hyp*). Upstream of *Rmar*._0536 are genes encoding: a DNA-binding protein (*hiran*); two hypothetical proteins (*hyp*); a potential methyltransferase (*duf43*). (B) Downstream of *Rmar*._0925 are genes encoding: a methylmalonyl-CoA mutase (*mm_coa_mut*); a 1,4-dihydroxy-2-naphthoyl-CoA synthase (*sfp*); a 4′-phosphopantetheinyl transferase (*′*); a 4-s RNA methyltransferase (*rnmV*); an O-acetylhomoserine aminocarboxypropyltransferase (*acp*); a homoserine O-acetyltransferase (*esterase*). (C) Downstream of *Rmar*._1080 and *Rmar*._1081 are genes encoding: two hypothetical proteins (*hyp*); a periplasmic ABC transporter substrate-binding protein (*abc*), a potential covalin binding protein (*yvcR*); and a cobalamin biosynthesis protein (*cbiX*); a hemi-binding protein (*hmuY*). Upstream of *Rmar*._1080 and *Rmar*._1081 genes encoding: an α-Glucuronidase (α-glucuronidase) and a pair of *susC* (a TonB-dependent receptor) and *susD* homologues. (D) Downstream of *Rmar*._2069 are genes encoding: a RNA-Gln-CTG,tRNA (gray arrow); a hypothetical protein similar to ABC transporter (*hyp. abc*); a hypothetical protein (*hyp*); a riboflavin synthase (*ribD*); a phosphohydrolase (*hd*); a riboflavin. Upstream of *Rmar*._2069 are genes encoding: an asparagine synthase (*asn_synth*); two hypothetical proteins (*hyp.*); a Fe-S oxidoreductase (*redox*); a hypothetical protein (*hyp.*). (E) Downstream of *Rmar*._2616 are genes encoding: a glycose hidrolase of the GH43/DUF377 family (*gh43_62_32_68*); a predicted glycosyl hydrolase of the GH43/DUF377 family (*gh43*); a protein similar to glycosyltransferase involved in cell wall biosynthesis (*rfB*); a short-chain dehydrogenase (*sdr*); a saccharopine dehydrogenase (*sacc*). Upstream of *Rmar*._2616 are genes encoding: a RNA polymerase ECF-type sigma factor gene (*sig70*); a RNA (cytosine-5-)- methyltransferase (*met_trn*); a Fmu domain protein (*fmu*); a flavin containing amine oxidoreductase (*amine oxida*); a bacterioferritin (*ferr*); an exodeoxyribonuclease III (*exoIII*).

**specific activity**. Both *RmXyl3A* and *RmXyl3B* showed highest specific activity towards *pNP-β-Xyl*, but displayed side activity on *pNP-α-l-Ara* and *pNP-β-Glc*.

**Temperature and pH optima.** The effect of temperature and pH on activity was studied using the *pNP*-substrate with highest specific activity for the respective enzyme (Supplementary Table 4). Highest optimal temperature was observed for *RmNag3* and *RmBgl3B* at 90 °C, while the remaining enzymes displayed activity optima between 60–80 °C. The pH optima for all six enzymes were between pH 5.0–5.6 except *RmBgl3C*, which was pH 7.0.

**Kinetic analysis.** Kinetic parameters on the same substrates (Table 2), showed that *RmBgl3A* had a relatively low *Kₘ* and higher catalytic efficiency (kcat/Kₘ) for *pNP-β-Glc* compared with *pNP-β-Xyl* (Table 2). *RmBgl3A* also had the highest catalytic efficiency for *pNP-β-Glc* of the three enzymes with β-glucosidase activity (*RmBgl3A, B* and *C*). The *Kₘ* of *RmBgl3C* was an order of magnitude higher than the *Kₘ* of the two other enzymes, showing lower affinity for this substrate. The turnover number for *RmBgl3B* (kcat) was in the same range for *pNP-β-Glc*, *pNP-β-Xyl* and *pNP-α-l-Ara*. The higher kcat/Kₘ of 511 s⁻¹·M⁻¹ for *pNP-β-Glc* (due to lower *Kₘ*) suggests that this is the most preferred synthetic substrate. All three enzymes, *RmBgl3A, RmBgl3B* and *RmBgl3C*, showed substrate inhibition while hydrolysing *pNP-β-Glc*, *RmXyl3A* and *RmXyl3B* were active against *pNP-β-Glc, pNP-β-Xyl* and *pNP-α-l-Ara*, and comparison of kinetic parameters revealed a definite preference for *pNP-β-Xyl* with a low *Kₘ* of *RmXyl3A* and *RmXyl3B* showed a similar substrate preference, but *RmXyl3B* had higher turnover number and catalytic efficiency.

The kinetic parameters for *RmNag3* were determined using *pNP-β-GlcNAc* and the *Kₘ* value was 0.1 mM (Table 2), in a reaction run in the presence of 20 mM sodium phosphate buffer. Additional kinetic analysis was then made using different concentrations of sodium phosphate in 50 mM HEPES buffer (Table 3), to investigate...
any potential phosphorolytic function for \(Rm\) Nag3. The \(K_M\) value increased between 0 and 50 mM of phosphate added to HEPES buffer from 0.16 to 0.30 mM, with no further increase at higher phosphate concentration (100 and 200 mM). For \(k_{cat}\), a similar trend was observed (Table 3). However, there was no clear trend in \(k_{cat}/K_M\) values to suggest any significant effect of phosphate on enzyme activity (Table 3).

Table 1. Putative locus for degradation and uptake of xylan in \(Rhodothermus marinus\) DSM 4252\(^\text{T}\). aEC - extracellular, CA - cell attached, OM - outer membrane, PS - periplasmic space, IC – intracellular, IM – inner membrane. The cellular location is predicted based on prediction of signal peptides and putative por secretion system C-terminal sorting domain. bCo-transcription is based on predicted operons, Rho-independent translational termination sites and promoters.

| Gene loci   | Annotation                        | Potential cellular location | Potential co-transcription |
|-------------|-----------------------------------|----------------------------|---------------------------|
| \(Rmar\)\_1067 | LacI family transcriptional regulator | IC                         | 1                         |
| \(Rmar\)\_1068 | GH43, CBM6                        | EC and CA                  | 2                         |
| \(Rmar\)\_1069 | CBM4, CBM4, GH10- RmXyn10A        | EC and CA                  | 2                         |
| \(Rmar\)\_1070 | TonB-dependent receptor            | OM                        | 2                         |
| \(Rmar\)\_1071 | Hypothetical protein               | EC and CA                  | 2                         |
| \(Rmar\)\_1072 | Hypothetical protein               | PS or EC                   | 3                         |
| \(Rmar\)\_1073 | Hypothetical protein               | EC and CA                  | 3                         |
| \(Rmar\)\_1074 | GH10                              | PS or EC                   | 4                         |
| \(Rmar\)\_1075 | RNA polymerase \(\sigma^{34}\)-factor | IC                      | 4                         |
| \(Rmar\)\_1076 | anti-FecI \(\sigma\)-factor, FecR | IM                        | 4                         |
| \(Rmar\)\_1077 | TonB-dependent receptor, SusC/RagA domain protein | OM                         | 4                         |
| \(Rmar\)\_1078 | SusD/RagB domain protein          | EC and CA                  | 4                         |
| \(Rmar\)\_1079 | \(\alpha\)-glucuronidase, GH67    | PS or EC                   | 5                         |
| \(Rmar\)\_1080 | GH3 - RmXyl3A                     | PS or EC                   | 5                         |
| \(Rmar\)\_1081 | GH3 - RmXyl3B                     | IC                        | 5                         |

Table 2. Kinetic parameters on aryl substrates. Data presented as means ± standard error from three independent experiments. *Substrate inhibition was observed for \(pNP-\beta\)-Glc at concentration above 2 mM.

| Enzyme | Substrate | \(K_M\) (mM) | \(k_{cat}\) (s\(^{-1}\)) | \(k_{cat}/K_M\) (s\(^{-1}\)mM\(^{-1}\)) |
|--------|-----------|--------------|---------------------------|------------------------------------------|
| \(RmBgl3A\) | \(pNP-\beta\)-Glc\(^c\) | 0.1 ± 0.0 | 79.9 ± 3.4 | 754.9 ± 28.1 |
| \(RmBgl3A\) | \(pNP-\beta\)-Xyl | 11.8 ± 1.0 | 140.8 ± 0.6 | 12.0 ± 0.0 |
| \(RmBgl3B\) | \(pNP-\beta\)-Glc\(^c\) | 0.1 ± 0.0 | 49.5 ± 4.0 | 511.4 ± 66.1 |
| \(RmBgl3B\) | \(pNP-\beta\)-Xyl | 1.2 ± 0.1 | 65.8 ± 0.4 | 56.2 ± 0.7 |
| \(RmBgl3C\) | \(pNP-\alpha\)-l-Ara | 1.9 ± 0.1 | 47.7 ± 0.8 | 25.0 ± 1.2 |
| \(RmXyl3A\) | \(pNP-\beta\)-Glc | 1.6 ± 0.9 | 67.2 ± 0.5 | 42.9 ± 2.2 |
| \(RmXyl3A\) | \(pNP-\beta\)-Xyl | 0.4 ± 0.0 | 50.1 ± 1.9 | 140.17 ± 15.7 |
| \(RmXyl3B\) | \(pNP-\beta\)-Glc | 6.5 ± 0.8 | 29.2 ± 0.5 | 4.5 ± 0.3 |
| \(RmXyl3B\) | \(pNP-\alpha\)-l-Ara | 8.8 ± 1.0 | 94.2 ± 11.2 | 10.7 ± 1.0 |
| \(RmNag3\) | \(pNP-\beta\)-Glc\(^c\) | 0.3 ± 0.0 | 160.1 ± 0.6 | 481.9 ± 21.5 |
| \(RmNag3\) | \(pNP-\beta\)-Glc\(^c\) | 1.6 ± 0.4 | 57.5 ± 0.6 | 36.1 ± 0.9 |
| \(RmNag3\) | \(pNP-\alpha\)-l-Ara | 2.0 ± 0.2 | 556.8 ± 8.6 | 280.9 ± 2.4 |
| \(RmNag3\) | \(pNP-\beta\)-Glc\(^c\) | 0.1 ± 0.0 | 102.3 ± 3.0 | 1314.4 ± 50.3 |

Table 3. Kinetic parameters for aryl substrate hydrolysis by \(RmNag3\). All the reactions were performed in 50 mM HEPES buffer pH 6.0 in absence or presence of sodium phosphate. Data presented as means ± standard error from three independent experiments.

| Phosphate [mM] | Substrate | \(K_M\) (mM) | \(k_{cat}\) (s\(^{-1}\)) | \(k_{cat}/K_M\) (s\(^{-1}\)mM\(^{-1}\)) |
|----------------|-----------|--------------|---------------------------|------------------------------------------|
| 0              | \(pNP-\beta\)-Glc\(^c\) | 0.2 ± 0.1 | 144 ± 4 | 920 ± 50 |
| 50             | \(pNP-\beta\)-Glc\(^c\) | 0.3 ± 0.1 | 223 ± 4 | 740 ± 20 |
| 100            | \(pNP-\beta\)-Glc\(^c\) | 0.3 ± 0.1 | 189 ± 3 | 630 ± 10 |
| 200            | \(pNP-\beta\)-Glc\(^c\) | 0.3 ± 0.1 | 232 ± 1.1 | 860 ± 30 |

\(^a\)Potential cellular location is predicted based on prediction of signal peptides and putative por secretion system C-terminal sorting domain. \(^b\)Co-transcription is based on predicted operons, Rho-independent translational termination sites and promoters.
Additional but instead of the typical single or two-domain architecture, \( Rm \) other one is missing the C-terminal domain 2. \( Rm \) domains 1, 2 and FnIII, it comprises the PA14 domain (Fig. 3F) inserted between biotics, activated by the removal of glucosyl moieties from their non-active glycosylated forms. In addition to -glucosidases involved in activation of secreted anti-group together with bacterial (mostly actinobacterial) -l-arabinofuranosidase (EC 3.2.1.55) and -glucosidase (EC 3.2.1.21) activities (Fig. 2). This group is

| Substrate | Specific activity (μmol min\(^{-1}\) mg\(^{-1}\)) |
|-----------|-----------------|
| Cellobiose (β-1,4) | \( 333 \pm 1 \) |
| Xylobiose (β-1,4) | \( 67 \pm 0 \) |
| Laminaribiose (β-1,3) | — |
| Cellohexaose (β-1,4) | \( 197 \pm 0 \) |
| Xylohexaose (β-1,4) | — |
| Xylan (β-1,4) | — |
| Chitobiose (β-1,4) | — |
| Chitopentose (β-1,4) | — |

Table 4. Hydrolysis of oligosaccharides. Type of linkage for each oligosaccharide is indicated in parentheses.

Activity on natural substrates. Hydrolysis of natural substrates was monitored using a set of oligosaccharides (Table 4). \( Rm \) Nag3 was able to release glucosamine from chitobiose and chitopentaose, consistent with the \( pNp,\beta\text{-GlcNAc} \) hydrolysis, confirming it as a \( \beta\text{-N-acetyl-glucosaminidase} \).

\( Rm \) Bgl3A produced glucose from both cellubiose and cellohexaose, which confirmed exo-glucanase activity. In addition, the enzyme maintained its bifunctional activity by hydrolysing xylobiose to xylose. \( Rm \) Bgl3B hydrolysed cellubiose, but was not active on cellohexaose, limiting the activity to short substrates. \( Rm \) Bgl3C hydrolysed all glucoligosaccharides tested, but showed clear preference for laminaribiose over cellubiose, and thus a clear preference for hydrolysis of \( \beta\text{-1,3-} \) over \( \beta\text{-1,4-linkages} \).

Both \( Rm \) Xyl3A and \( Rm \) Xyl3B were hydrolysing xylobiose, but only \( Rm \) Xyl3B showed exo-glucanase activity and released xylose from both xellohexaose and xylan.

Phylogenetic analysis of the \( R. \) marinus GH3 enzymes. A phylogenetic analysis was performed on the amino acid sequences of 100 characterized GH3 proteins, including the six enzymes from \( R. \) marinus. Multiple sequence alignment revealed strictly conserved residues within the family and highlights the lack of the otherwise conserved catalytic acid/base among the \( \beta\text{-N-acetyl-glucosaminidases} \) (Supplementary Fig. 2). Based on the maximum likelihood tree (Fig. 2), it can be seen that the \( \beta\text{-N-acetyl-glucosaminidase} \) cluster is the deepest rooted group within GH3, separated from others. Characterized enzymes clustering in this lineage have two different domain architectures, single-domain or two-domain enzymes. One subgroup consists of both domain 1 and 2, the other one is missing the C-terminal domain 2. \( Rm \) Nag3 clusters with the lineage of \( \beta\text{-N-acetyl-glucosaminidases} \), but instead of the typical single or two-domain architecture, \( Rm \) Nag3 displays three domains, containing an additional \( \beta\text{-lactamase} \) domain at its C-terminus. BLAST search using the BALSTp tool\(^{30} \), revealed a number of gene sequences encoding proteins with similar domain architecture (i.e. domain 1, domain 2 and the \( \beta\text{-lactamase} \) domain) e.g. genes encoding putative GH3 candidates from \( Salinibacter altiplanensis \) (RefSeq: WP_103019701) and \( Rhodobacter halophilus \) (RefSeq: WP_083750206), indicating a common role. To date, \( Rm \) Nag3 is however the only characterized enzyme with this type of domain architecture.

The other five \( Rm \) GH3 enzymes cluster with different subgroups in two main evolutionary lineages. The first lineage consists of the three subgroups: thermostable multifunctional \( \beta\text{-xylosidases} \), closely related \( \beta\text{-glucosidases} \) and exo-\( \beta\text{-1,3,1,4-glucanases} \) from plants and bacteria (Fig. 2). Four of the \( R. \) marinus GH3 enzymes cluster with subgroups within this first lineage (\( RmBgl3A \), \( RmBgl3C \), \( RmXyl3A \) and \( RmXyl3B \)). \( RmBgl3A \) is part of a weakly defined cluster formed by three bacterial \( \beta\text{-glucosidases} \) (Fig. 2), all clustering within the \( \beta\text{-glucosidase} \) subgroup. The enzyme consists of domains 1, 2 and the fibronectin type III (FnIII) domain. The protein \( RmBgl3C \) clusters together with exo-\( \beta\text{-1,3,1,4-glucanase} \) from plants and bacteria and is a two-domain enzyme consisting of domain 1 and 2. At its C-terminus it has a C-terminal extension, like exo-glucanase (ExoI) from \( Hordeum vulgare \) (UniP: Q9XE13), instead of the domain 3 found in e.g. exo-glucanase (ExoP) from \( Pseudomanares \) sp. BB1 (UniP: Q0QJA3). \( RmXyl3A \) and \( RmXyl3B \) both consist of domain 1, 2 and the FnIII domain and cluster together with the group of thermostable enzymes with predominant \( \beta\text{-xylosidase} \) activity (EC 3.2.1.37) activity with minor \( \alpha\text{-l-arabinofuranosidase} \) (EC 3.2.1.55) and \( \beta\text{-glucosidase} \) (EC 3.2.1.21) activities (Fig. 2). This group is separated from, and therefore not directly phylogenetically related to, other \( \beta\text{-xylosidase} \) clusters originating mainly from plants and fungi.

A second lineage is formed by two large subgroups. The first subgroup comprises fungal and bacterial \( \beta\text{-glucosidases} \), including a group of macrolide \( \beta\text{-glucosidases} \), whereas the second subgroup is formed by \( \beta\text{-xylosidases} \) from plants, fungi and bacteria (Fig. 2). \( RmBgl3B \) clusters with the macrolide \( \beta\text{-glucosidase} \) subgroup together with bacterial (mostly actinobacterial) \( \beta\text{-glucosidases} \) involved in activation of secreted antibiotics, activated by the removal of glucosyl moieties from their non-active glycosylated forms. In addition to domains 1, 2 and FnIII, it comprises the PA14 domain (Fig. 3F) inserted between \( \beta\text{-strand K} \) and \( \alpha\text{-helix K} \).

Structural analysis of the \( R. \) marinus GH3 enzymes. Domain organization, tertiary structure and active site arrangement for the six GH3 enzymes were investigated by homology modelling and compared to closely related structures. Sequence similarities with template structures and model validation are presented in Table 5. Homology model validation show an overall good quality of the six models however indicating some parts of low quality. The modelled structure of \( RmBgl3B \) have low quality in parts of the PA14 and FnIII domains and in loop I in domain 2 close to the active site. For \( RmXyl3A \) and \( RmXyl3B \), several long loops in the modelled
structures were not present in the templates, making loop predictions in these regions poor. The RmNag3 model had very good quality, despite lower sequence identity with its template.

The overall structures of the most common domains in GH3 enzymes: $(\alpha/\beta)_8$ barrel of domain 1, $(\alpha/\beta)_6$ sandwich of domain 2, and $\beta$-sandwich of FnIII were conserved, when present in the modelled structures (Fig. 3)\textsuperscript{31}. The FnIII domain is found in RmBgl3A, RmBgl3B, RmXyl3A and RmXyl3B, and is situated on the side of the domain 1 barrel, in contact with domain 2. The PA14 domain of RmBgl3B is situated on top of the barrel in contact with domain 2. RmBgl3A, RmXyl3A and RmXyl3B were modelled as dimers, with the monomers positioned with the surface created by the FnIII domain, domain 2 and domain 1 next to the opposite chain, and with the linker between domain 2 and FnIII positioned on top of the barrel of the opposite chain.

The active sites are situated on top of the barrel (Supplementary Figs. 3 and 4), are pocket shaped and are made up of both domain 1 and 2, the exception being RmNag3, which will be discussed separately below. RmBgl3B has a very deep and narrow active site pocket, and the PA14 domain as well as a long loop g covers the surface over the active site. The PA14 domain of RmBgl3B differs considerably from the PA14 domain present in closely related GH3 enzyme KmBglI from Kluyveromyces marxianus which have been shown to be involved in substrate interaction\textsuperscript{32}. For RmBgl3A, the linker between domain 2 and the FnIII domain on the opposite chain is involved in shaping the active site. This linker makes the active site deeper on one side, compared to the active site of RmBgl3C and the $\beta$-xylosidases. However, RmBgl3A displays a shorter loop b which opens up the active site on the opposite side. RmBgl3C has a different conformation of loop j on domain 2, making the active site narrower in this direction.
The homology model of \( RmNag3 \), the \( \beta \)-N-acetyl-glucosaminidase, displays a well-conserved tertiary structure as well as active site arrangement (Fig. 3F, Supplementary Figs. 3F, 4F and Table 6). The enzyme consists of domain 1 and 2 but only domain 1 builds up the active site (Supplementary Fig. 3F). Conserved residues important for catalysis were found in positions corresponding to those in structure determined GH3 \( \beta \)-N-acetyl-glucosaminidases: catalytic nucleophile Asp118 on \( \beta \)-strand \( g \), the Asp227-His229 dyad on loop \( e \), and Arg186 and Phe188 on loop \( d \) (Supplementary Fig. 4F). Similar to other GH3 enzymes, Lys216 and His217 on \( \beta \)-strand \( e \) and Asp118 on \( \beta \)-strand \( c \) was found hydrogen bonding to GlcNAc in subsite \(-1\). Superimposition with NagZ from \( Pseudomonas aeruginosa \) in complex with GlcNAc and l-Ala-1,6-anhydroMurNac (PDB: 5G3R), showed that Arg60 on \( \beta \)-strand \( a \), Arg126 on loop \( c \) and Glu306 on loop \( g \) were potentially hydrogen bonding to a sugar unit in subsite \(+1\) (Supplementary Figs. 3F and 4F). All residues important for substrate binding were conserved in the closely related NagA from \( Streptomyces thermoviolaceus \) (Table 6).

The modelled structures of the \( \beta \)-glucosidases and \( \beta \)-xylosidases presented \(-1\) subsite-architectures typical of GH3, with small but important differences (Fig. 4A, Supplementary Figs. 3, 4 and Table 6). The strictly
Table 6. Comparison of potential substrate interacting residues with closely related characterized GH3 enzymes based on sequence alignment and structure. The identities between the enzymes are indicated in parenthesis. Strictly conserved residues presented in Supplementary Fig. 2 are excluded from the table. The table is broken before the β-N-acetyl-glucosidases and the columns are not comparable above and under this line.

*Gap in the sequence alignment. The residue is not in proximity for substrate interaction based on structures.

| Enzyme                              | Subsite – 1 | Subsite +1 |
|-------------------------------------|-------------|------------|
| RnbBlg3A                            | D110 G218*  | A428* Y287 Y614 G431 ←3 R696 F254 |
| Elizabethkingia menengoseptica (46%) AAB66561.1 | D71 G179 S387 Y248 E587 S590 ← R572 F215 |
| RnbBlg3B                            | D71 R178 S380 W246 Y703 V383 ← S682* Y213 |
| Streptomyces venezuelae – DesR (44%) PDB: 4L3G | D98 R206 S410 W274 Y691 V413* ← D670* Y241 |
| RnbBlg3C                            | D121 D234* T450* Y311 ← W453 ← Y278 |
| Pseudalteromonas sp. BB1 – ExoP (45%) PDB: 3F93 | D136 G252 S460 W321 ← W463 ← F288 |
| Hordeum vulgare – ExoI (48%) PDB: 1EX1 | D95 G217 T431 W286 ← W434 ← Y253 |
| RnXyl3A                             | E123 G237* S446* W307 ← P449* Y545 E648* Y274 |
| RnXyl3B                             | E125 G239* S448* W309 ← P451* W547 E650* Y276 |
| Caldanaerobius polysaccharolyticus – Xyl3A (42, 43%) AFM44469.1 | E108 G218 A429* Y288 W631 C432 C524 G627 Y255 |

Table 6. Comparison of potential substrate interacting residues with closely related characterized GH3 enzymes based on sequence alignment and structure. The identities between the enzymes are indicated in parenthesis. Strictly conserved residues presented in Supplementary Fig. 2 are excluded from the table. The table is broken before the β-N-acetyl-glucosidases and the columns are not comparable above and under this line.

*Gap in the sequence alignment. The residue is not in proximity for substrate interaction based on structures.

| Enzyme                              | Subsite – 1 | Subsite +1 |
|-------------------------------------|-------------|------------|
| Bgl3A                               | D118 R186 D227 H229 R60 R126 R306 |
| Streptomyces thermoviolaceus – NagA (39%) BAA32403.1 | D164 R232 D274 H276 R80 R172 G352 |

Figure 4. Comparison of active sites of the β-glucosidase and β-xylanases showing residues involved in (A) subsite –1 and (B) subsite +1. Superimposition of homology models of RnbBlg3A (pink), RnbBlg3B (blue), RnbBlg3C, RnXyl3A (yellow) and RnXyl3B (orange). Cellobiose (gray) and laminaribiose (light gray) in subsite –1 and +1 from PDB entries 1IEX and 1J8V respectively, both solved in complex with ExoI from barley. RnXyl3A and RnXyl3B have similar –1 and +1 subsites and RnbBlg3C have a similar –1 subsite as RnbBlg3A in terms of possible interacting residues.

Conserved catalytic nucleophile on β-strand e28, the catalytic acid/base on loop l in domain 229, a tandem Lys and His on β-strand e, Arg on loop d and Met on β-strand f were in positions corresponding to those in other GH3-structures29. Subsite –1 of RnXyl3A and RnXyl3B were very similar. This was also the case for subsite +1 of RnbBlg3A and RnbBlg3C, while RnbBlg3B revealed two non-conserved residues: Arg178 on loop e and Ser380 on loop j potentially hydrogen bonding with the glucose in subsite –1 (Fig. 4A). A difference between the β-glucosidases and β-glucosidasises was found on β-strand c, the two β-xylanases possess a Glu and the three β-glucosidasises an Asp, potentially hydrogen bonding with the respective sugar in subsite –1 (Fig. 4A). These features were conserved in the closest relative of the respective enzyme (Table 6).

Subsite +1 was studied by superimposition of structure determined enzyme-ligand complexes (PDB: 1IEX, 1J8V, 4Z09 and 5AE6) (Fig. 4B, Supplementary Figs. 3, 4 and Table 6). Both β-glucosidasises and β-xylanases displayed an aromatic residue on top of subsite +1 on loop d. In RnbBlg3A, RnbBlg3C and the two β-xylanases an aromatic residue below subsite +1, was potentially sandwiching the sugar unit. This residue was positioned in different loops in the respective enzyme (Fig. 4B). In RnbBlg3A, Tyr614 is located in the linker region between domain 2 and FnIII of the opposite chain. In RnbBlg3C, Trp545 is located in loop j in domain 2. In RnXyl3A and RnXyl3B, Tyr545 and Trp547, respectively, are located two positions after the acid/base catalyst on loop l in domain 2. Due to low sequence identity to the model templates, the prediction of loop l in domain 2 is,
however uncertain, including the potential involvement of Tyr545 and Trp547 in subsite +1. In *RmBgl3A*, an additional residue, Arg606 in the linker between domain 2 and FnIII, is potentially interacting with subsite +1. In *RmBgl3B*, *RmBgl3C*, *RmXyl3A* and *RmXyl3B*, a Tyr on loop f creates hydrogen bonding possibilities. *RmBgl3A* has Phe254 in the corresponding position. The interacting residue in subsite +1 were not completely conserved in the closest related structure of the respective enzyme (Table 6). Tyr614 in *RmBgl3A* corresponded to Glu587 in the 3-glucosidase from *Elizabethkingia meningoseptica* and the aromatic residue on the bottom of subsite +1 in *RmXyl3A* and *RmXyl3B* was replaced by Cys524 in *Xyl3A* from *Caldanaeobius polysaccharolyticus*.

**Discussion**

*Rhodothermus marinus* has served as a source of various thermostable enzymes. The Carbohydrate Active enZymes (CAZy) database ([http://www.cazy.org](http://www.cazy.org)) reports 55 glycoside hydrolase (GH) encoding genes in 33 different GH families and two non-classified genes in the type strain (DSM 42527). Six genes encodes GH family 3 (GH3) candidates and five of them are annotated as 3-glucosidases in Kyoto Encyclopedia of Genes and Genomes (KEGG). However, the GH3 members of strain DSM 4253 have different substrate specificities, consistent with differences in their modelled structures. It is very important to mention that the GH3 candidates from both strains of *R. marinus* DSM 4252 and 4253 are homologous. The most common domain architecture in GH3 is either a two domain type: domain 1 and 2, with the catalytic nucleophile in domain 1 and acid/base in domain 2 or a three domain type, with an FnIII-domain combined with domain 1 and 2.10. Besides these domain architectures, there are two clusters of proteins in the phylogenetic tree containing an additional domain type, PA14, inserted between the two strands, there are two clusters of proteins in the phylogenetic tree containing an additional domain type, PA14, inserted between the two strands, there are two clusters of proteins in the phylogenetic tree containing an additional domain type, PA14, inserted between the two strands, there are two clusters of proteins in the phylogenetic tree containing an additional domain type, PA14, inserted between the two strands, there are two clusters of proteins in the phylogenetic tree containing an additional domain type, PA14, inserted between the two strands, there are two clusters of proteins in the phylogenetic tree containing an additional domain type, PA14, inserted between the two strands. These two clusters are phylogenetically separated (Fig. 2) and differ in both type and orientation of the PA14 domain, it is likely that they are result of two evolutionary independent insertions.

*RmNag3* has a well-conserved -1 subsite and several possible hydrogen bonding residues in subsite +1 (Fig. 4F). Among the interacting residues in subsite +1 Arg126 was found conserved in almost all structure-determined GH3 3-N-acetyl-glucosaminidases. *RmNag3* possesses a signal peptide, a 3-lactamase domain and showed optimal activity at high temperature. It was earlier proposed that all 3-N-acetyl-glucosaminidases were phosphorylases. However, this seems less likely as the presence of phosphate in the kinetic reaction showed little effect on hydrolytic activity of *RmNag3*. An increase in KM was observed at 50 mM phosphate, but KM remained unaffected with increasing phosphate concentration. A similar trend was also observed for the 3-N-acetyl-glucosaminidase from *Herbaspirillum seropedicae* which does not have phosphorylase activity.42. For the GH3 phosphorylase from *Cellulomonas fimii* the presence of phosphate had significant effect on the hydrolytic activity42 compared to *RmNag3*. It is also interesting to see that *RmNag3* had no activity on pNP-3-Glc while both enzymes from *H. seropedicae* and *C. fimii* were active on this substrate.42,43. This narrow substrate specificity could be due to the presence of the 3-lactamase domain in *RmNag3*. Two main biological roles for 3-N-acetyl-glucosaminidases are degradation of chitin and peptidoglycan turnover. *R. marinus* possesses an extracellular GH18 chitinase and an intracellular GH20 3-N-acetyl-hexosaminidase, proposing chitin utilization in *R. marinus*. However, the 3-lactamase domain in *RmNag3* suggests association with peptidoglycan turnover as the degradation products have been shown to function as inducers for 3-lactamase.43, 3-N-acetyl-glucosaminidases involved in peptidoglycan turnover are often but not exclusively intracellular in Gram-negative bacteria.44. Comparison with peptidoglycan turnover in *Escherichia coli* strain K12 shows that the *R. marinus* type strain genome contains orthologous genes for degrading peptidoglycan and release of GlnNac, including anhMurNAc and tetratetrapeptide in the periplasmic space (endolytic murein transglycosylase (MltG)), membrane-bound lytic murein transglycosylase (MltD), N-acetylmuramoyl-L-alanine amidase (AmiA) and NagZ (ortholog to *RmNag3*). No intracellular orthologs to NagZ, the intracellular N-acetylmuramoyl-L-alanine amidase (AmpD) or the permease AmpG for transportation of GlnNac-1,6-anhydro-MurNAc-peptides across the inner membrane are found in *R. marinus*. The lack of AmpG makes it likely that *RmNag3* is present in the periplasm. This set-up is found in the Gram-positive *Bacillus subtilis* which lacks AmpG and AmpD, and has a 3-N-acetyl-glucosaminidase (NagZ) and an N-acetylamuramoyl-2,6-anamine amidase (AmiE) in the periplasm.47. In *B. subtilis*, transportation of GlnNac and anhMurNAc is believed to be done by the phosphotransferase system enzymes MurP (anhMurNAc) and NagE (GlnNac). No orthologs for these proteins were found in *R. marinus*, however, AmnK (anhydro-N-acetylmuramic acid kinase) and NagK (N-acetyl-D-glucosamine kinase) were present as intracellular proteins.

Both *RmBgl3A* and *RmBgl3B* showed specificity for cellobiose but only *RmBgl3A* could hydrolyse cellobiose. *RmBgl3A* is expected to be intracellular while *RmBgl3B*, which has a higher temperature optimum and a predicted a signal peptide, is likely exported. *RmBgl3A* is a dimer with a conserved subsite −1 and a well-defined subsite +1 including two aromatic residues sandwiching the sugar and Arg606 on the linker positioned between domain 2 and the FnIII domain on the opposite chain. These residues are also conserved in the closest related structure-determined enzymes (Supplementary Table 1). The corresponding residue to Arg606 have been hypoth-

However, in the closest related structure of the respective enzyme (Table 6). Tyr614 in *RmBgl3A* corresponded to Glu587 in the 3-glucosidase from *Elizabethkingia meningoseptica* and the aromatic residue on the bottom of subsite +1 in *RmXyl3A* and *RmXyl3B* was replaced by Cys524 in *Xyl3A* from *Caldanaeobius polysaccharolyticus*. However, the

However, in the closest related structure of the respective enzyme (Table 6). Tyr614 in *RmBgl3A* corresponded to Glu587 in the 3-glucosidase from *Elizabethkingia meningoseptica* and the aromatic residue on the bottom of subsite +1 in *RmXyl3A* and *RmXyl3B* was replaced by Cys524 in *Xyl3A* from *Caldanaeobius polysaccharolyticus*. However, the
DesR from *Streptomyces venezuelae*, which is placed in the same cluster (Fig. 2) and active on glucosylated macrolides56. The PA14 domain in other carbohydrate-active proteins, including the closely related *Rm*Bgl1 from *Kluyveromyces marxianus*, have been associated with oligosaccharide binding57,58. However, no substrate interaction possibility for the PA14 domain was found, neither in the model of *Rm*Bgl3A nor in the structure of DesR.

*Rm*Bgl3C has linkage preference towards 1,3,1,3-linked glucose which is found in laminarin, a polysaccharide present in brown algae53, in a bacterial polysaccharide called curdlan54 and in various mixed-linkage glucans. It clusters under the lineage that consists of exo-glucanases from plants and bacteria. Despite C-terminal extension differences, the exo-glucanase ExoP from *Pseudalteromonas* sp. BB1 and *Rm*Bgl3C display similar activity on laminarioligosaccharides55. *Rm*Bgl3C has a clear signal peptide and higher pH optimum than the other GH3 enzymes in *R. marinus*. Therefore, it is likely that the enzyme is exported out of the outer membrane.

Corresponding pH optima were also observed in *Rm*Xyn10A and *Rm*Cel12 which are exported13,58. *Rm*Bgl3C displays a subsite −1 similar to *Rm*Bgl3A and a subsite +1 with two aromatic residues sandwiching the sugar, the conserved Tyr311 on loop g and Trp453 on loop j, and Tyr278 with a hydrogen bonding possibility. Trp453 is located in a similar position as Tyr614 of *Rm*Bgl3A but the two residues are places on different loops (Fig. 4B). *Rm*Trp453 is only found in a few structures including the closest structure-determined enzymes (Supplementary Table 1). The two characterised Exol from barley (*Hordeum vulgare*) and ExoP, are in similarity with *Rm*Bgl3C 1,3,1,3,1,4-glucanases. The structural basis for this specificity has been investigated but not clearly understood55. Interestingly, loop e which is involved in shaping subsite +1 is longer in *Rm*Bgl3C than in any of the closest structure-determined enzymes. An accurate conformation of this loop can be important in understanding the specificity of *Rm*Bgl3C, as Arg228 on loop e in ExoP was found to hydrogen bond to the glucose in subsite +1, when laminaribiose was modelled into the structure58.

*Rm*Xyl3A and *Rm*Xyl3B are both bifunctional β-xylosidases/β-1,4-gluco-sideases but show different activity patterns. They are 94% identical, positioned next to each other on the chromosome, and thus, likely a result of a relatively recent gene duplication. There are two less similar regions identifiable in the sequences. The first region is located in domain 1 in between β-strand g and α-helix H2, whereas the second one is found in domain 2 on an unusually long loop between β-strand k and α-helix K. *Rm*Xyl3B has endo-β-xylose activity while *Rm*Xyl3A only hydrolysed xylobose and, interestingly, cellobiohexose. Based on kinetic analysis *Rm*Xyl3B also showed specificity for pNP-α-1-Ara. A similar type of bifunctional β-xylosidase/β-1,4-gluco-sidease has been characterised from *Caldanaeroebius polysaccharolyticus*59, in the same phylogenetic cluster as *Rm*Xyl3A and *Rm*Xyl3B (Fig. 2). Subsite −1 of *Rm*Xyl3A and *Rm*Xyl3B is similar to that of *Rm*Bgl3A and *Rm*Bgl3C except for one residue potentially for the difference in specificity. The two β-xylosidases/β-1,4-gluco-sideases display a Glu instead of an Asp on β-strand c in subsite −1 (Fig. 4A). This difference could affect the preference for glucose or xylose in subsite −1 as the Asp is making no steric hindrance for the additional CH2OH group of glucose. The same pattern was found in other GH3 enzymes (Table 6), including Exol from barley and Xyl3B from *Prevotella bryantii* where substitution of Glu115 to an Asp increased catalytic activity of Xyl3B on pNP-β-Glc60. Another interesting feature is the aromatic residue on loop j in domain 2, two positions away from the potential acid/base catalyst in *Rm*Xyl3A and *Rm*Xyl3B (Fig. 4B), with the possibility to be in a position analogous to Tyr614 in *Rm*Bgl3A. This was not found in the *Caldanaeroebius polysaccharolyticus* β-xylosidase/β-1,4-gluco-sidease, but comparison with structures of other GH3 enzymes, reveals an aromatic residue two positions after the acid/base catalyst and a loop j of similar length in several fungal β-glucosidases with very low sequence identity to the *R. marinus* β-xylosidases/β-1,4-gluco-sideases (see PDB: 4IVF, 5F1L, 5F1J, SNBS, 4DOJ and 3ZVZ), and stacking with a glucose in subsite +1 has been shown for AaBGL1 from *Aspergillus aculeatus*56.

The genomic context of the GH3 enzymes was informative to some extent concerning their potential role in utilization of carbohydrate substrates. The genome comparison between *R. marinus* DSM 42527 and 4253 showed that all the corresponding GH3 gene loci are identical. In *R. marinus* DSM 42527 and DSM 4253, the putative PUL consists of six GHs: the endo-1,4-β-xylose *Rm*Xyn10A19 with putative location in the extracellular space, an uncharacterized GH10 (a family containing mainly endo-1,4-β-xylosanases) and the two GH3 β-xylosidases characterized in the current study, suggesting that the cluster is involved in utilization of 1,4-β-linked xylans. The potential extracellular GH43 belongs to subfamily 15 with no characterized proteins so far. The uncharacterized GH67 is annotated as an α-glucuronidase in NCBi, suggesting that xylan substituted with glucuronic acid or methyl glucuronic acid could be utilised by the PUL. Besides glycan degrading enzymes, a canonical PUL involves a pair of susC and susD homologues and a regulator61, components which are all found in the described cluster. SusC and SusD, first described as vital components of a starch utilisation system (SUS) in *Bacteroides thetaiotaomicron*, are co-regulated genes involved in coordinated binding, transport and degradation of carbohydrates from the extracellular space via the outer membrane into the periplasmic space55. SusC is a type of TonB-dependent receptor, an outer membrane transporter known to transport ferric chelates, but also known to transport carbohydrates. TonB-dependent transporters require energy and three inner membrane proteins in a complex, TonB-ExbB-ExbD, for the transportation62. SusD is a lipoprotein attached to the outer cell membrane bound to carbohydrate. In addition, the putative regulation system identified in the cluster, called trans-envelope signalling, involving an extracytoplasmic function (ECF) σ-factor/anti-σ-factor system is found in PULs from *Bacteroidetes*63. Biochemical and structural characterization of the hydrolytic enzymes involved in the cluster, transcriptome analysis during growth on different substrates as well as knock-out of the genes are necessary to fully understand the mechanisms involved in the PUL described in this study.

In summary, the biochemical and structural characterisation of the GH3 enzymes from *R. marinus* DSM 4253 (and 42527), shows that the six GH3 enzymes encoded in the genome have non-redundant substrate specificities which are involved in extracellular laminarin, potential macrolide degradation, as well as intracellular cellobiose to glucose conversion, the conversion of xylans, and recycling of peptidoglycans, giving significant insights into structural features important for the specificity of these enzymes as well as the organization of corresponding loci in the *R. marinus* genome.
Materials and Methods

Chemicals. Laminariobiose, cello-, xylo- and acetyl-chitoooligosaccharides were obtained from Megazyme (Wicklow, Ireland). Sodium acetate, laminarin from Laminaria digitata, xylan from birch wood, para-nitrophenol and all para-nitrophenyl-β-D-glycosides were purchased from Sigma-Aldrich (St. Louis, Mo). All other chemicals were of molecular biology or analytical grades and purchased from VWR International (Stockholm, Sweden).

Bacterial strains, genome sequencing and gene cluster analysis. *Rhodothermus marinus* strain DSM 4253 was isolated from an intertidal hot spring in Iceland, at a location close to that of the *Rhodothermus marinus* DSM 4252 (Type strain). The genome was sequenced using TrueSeq chemistry for library construction and MiSeq sequencing platform (unpublished data). Sequencing data was assembled using GS De Novo Assembler software (Roche) and annotated using the RAST annotation server at rast.nmpdr.org. Genes, encoding enzymes of glycoside hydrolase (GH) family 3 (GH3), were identified by BlastX using GH3 amino acid sequences from the *R. marinus* type strain retrieved from the Carbohydrate-Active Enzyme (CAZY) database (http://www.cazy.org) as query sequences. Sequence regions harbouring the GH3 genes along with flanking sequences, spanning approximately 15 kb, were extracted from the genome sequence using the Geneious molecular biology tool. The structure of corresponding loci of the GH3 genes was resolved in a genome viewer and the neighbouring genes were analysed by BLAST.

Gene clusters were analysed in the genome of the type strain by annotations presented in the National Center for Biotechnology Information (NCBI) Nucleotide database, Reference Sequence (RefSeq); NC_013501.1, if genes in proximity to the six genes 
\\texttt{(Rm\textsubscript{arunus}_0536, Rm\textsubscript{arunus}_0925, Rm\textsubscript{arunus}_1080, Rm\textsubscript{arunus}_1081, Rm\textsubscript{arunus}_2069, and Rm\textsubscript{arunus}_2616)}
encoding the putative GH3 enzymes: *RmBgl3A*, *RmNag3*, *RmXyl3A*, *RmXyl3B*, *RmBgl3B*, and *RmBgl3C*. Additional annotations were done by comparing the information from searches in the Conserved Domain Database (CDD) from NCBI and in the CAZY database in the case of annotated GHs. Operons were predicted by DOOR 2.055,66 and Genome2D.55 Putative promoters were predicted by PePPER and Rho-independent translational terminator sites were predicted by ARNold and DOOR 2.0. Potential co-transcription was based on predicted operons, promoters and transcription terminators. Predictions of signal peptides were done by SignalP versions 3.0 and 4.7. BLASTp was used to investigate poorly annotated genes and search for domains. Potential cellular location was based on prediction of signal peptides and the putative por-secretion system C-terminal sorting domain.

Sequence alignment and evolutionary relationships. From 301 characterized protein members classified in GH3 of the CAZY database, 100 with known activity were selected (Supplementary dataset file). Amino acid sequences of these proteins were retrieved mainly from the UniProt database and some were obtained from GenBank and RefSeq databases. All sequences of characterized GH3 proteins together with those of the six proteins from *R. marinus* were aligned using the programme Clustal-X and then the alignment was further manually fine-tuned. Maximum likelihood phylogenetic tree was calculated using the PhyML algorithm available through the T-REX server (http://www.trex.uqam.ca). Gamma shape parameter and proportion of invariant sites were estimated by the PhyML itself. Number of relative substitution rate categories was set to four and WAG substitution model was used. The starting tree was calculated using BIONJ. NNI was used for a tree improvement. Tree topology and branch lengths were optimized. Reliability of tree topologies was evaluated using the bootstrap test with 100 replications. For the phylogenetic analysis, only amino acid sequences of domain 1, which is universally present in all proteins from GH3, was used.

Structure homology modelling. Homology modelling of the six GH3 of *R. marinus* DSM 4253 was carried out using the YASARA program with default settings except templates that were manually inserted (Supplementary Table 1). For each enzyme, a BLASTp search in the Protein Data Bank (PDB) was done and the five enzymes with the highest score were chosen. If several structures were available, the structure without mutations, with a high resolution and relevant ligand was chosen. If the sequence identity was more than 10% units lower than the hit with the highest sequence identity, this template was not used. Only relevant ligands were kept in the structure and in some cases ligands were manually modified: in *RmBgl3A* and *RmBgl3B* into a xylose and in the case of *RmNag3* into a β-N-acetyl-glucosamine before entering the homology modelling. Only ions conserved among the templates were kept in the structures. Only dimerization conserved among the templates were kept and in other cases only chain A was kept. The β-lactamase domain of *RmNag3* was removed before homology modelling since no hit was found for the entire protein and no hit with a sequence identity above 30% was found for the domain alone. YASARA homology modelling generates five alignments for each template and builds a model for each alignment. A hybrid model is generated by combining the best part of the models. Each hybrid model was manually checked by superimposition and comparison of templates and other GH3 structures in Chimera. Side chains in the active site and ligand positioning were manually changed based on conserved features in the closest related structures. The peptide bonds linking the conserved residues Lys and His on β-strand and the following two amino acids were manually modified into cis-conformation, which are conserved within GH3. Modified models were energy minimised and then refined in YASARA. Refinement was done with default setting, only changing temperature and pH to the optima for each enzyme (Supplementary Table 4), with a 500 ps simulation of molecular dynamics with the YASARA2 force field. During the simulation, 20 trajectories were saved, energy minimised and analysed by checking the energy of the system as well as dihedral angles, packaging and packing3D. The trajectory with best overall quality was further evaluated. Evaluation of each refined structure was done manually and by several online validation tools. Superimposition in Chimera verified that the active site arrangement was kept and stabilized by the simulation. If not, modifications of the hybrid model were revised and a new version was refined. The quality of each refined structure was evaluated by average 3D-1D score.
Cloning and expression of GH3 genes. *R. marinus* DSM 4253 was grown in 500 mL of DIFCO™ Marine Broth (BD, NJ, USA) at 65 °C for 12 hours. The cell pellet was harvested by centrifugation and washed twice with 200 mM sodium phosphate buffer pH 7. Genomic DNA was extracted using ZymoBead™ Genomic DNA Kit (Zymo Research, CA, U.S.A). The GH3 encoding genes, without signal peptide, were amplified using the primers listed (Supplementary Table 2). A C-terminal His-tag was introduced in the primer-designs, except for *Rmar_0926*. Genes digested with *Nde*I and *Bgl*II were ligated into pOE307592, a non-commercial vector and transformed to *Escherichia coli* BL21 C43. *Rmar_0926* was digested by *Nde*I and *Hind*III, propagated in vector pUC19 followed by sub-cloning into pET-21b(+) (Novagen, Madison, WI) in frame with the C-terminal His-tag and transformed to *E. coli* BL21 C43. Expression was performed at 0.5 L cultivation scale in Erlenmeyer flasks at 37 °C in LB containing 100 µg/mL ampicillin. After reaching an optical density at 620 nm of 0.5, gene expression was induced by 1 mM IPTG for the pET-construct and 0.2% (w/v) L-rhamnose for pOE3075-constructs92. Production was continued at 30°C over-night. Cells were harvested by centrifugation at 5000 × g for 15 min at 4°C and washed twice with 20 mM sodium phosphate buffer pH 7.0.

Hydrolysis of para-nitrophenyl glycosides. Enzyme catalysed hydrolysis of para-nitrophenyl (pNP)-linked substrates was assayed spectrophotometrically at 405 nm using a UV-1650PC spectrophotometer (Shimadzu, Kyoto, Japan) connected to a JulaboMB (Labortechnik GMBH, Teltow, Germany). For *Rmar_0925* encoding the RmNg3, the binding buffer contained 20 mM Tris–HCl instead of sodium phosphate buffer Cell debris was removed by centrifugation (4000 × g, 20 min, 4 °C), prior to purification by nickel affinity chromatography using an AKTA™ start system (GE Healthcare) with a HisTrap FF crude 1 mL column (GE Healthcare). Bound protein was eluted using gradient of imidazole 20–500 mM and fractions of 1 mL were collected. All purified proteins were stored at 4 °C.

Hydrolysis of oligosaccharides. Oligosaccharides (10 mg/mL) with a degree of polymerization (DP) of 2 (cellobiose, laminaribiose and xylobiose) and DP 6 (cellotetraose and xylosexose) in 500 µL, 20 mM citrate phosphate buffer, pH 5.6 were incubated with 5–10 µg/mL enzyme in duplicate reactions in a ThermoMixer (HLC Biotech, Bovenden, Germany) at 60 °C, 600 rpm. Samples (30 µL) were withdrawn at 20 min intervals for 120 min, and diluted with 970 µL 0.5 mM NaOH before analysis on high-performance anion exchange chromatography with pulsed amperometric detection (HPAEC-PAD) (Thermo Fisher Scientific, Waltham, USA), using a CarboPac PA200 column (250 mm × 3 mm, 5.5 µm) and a guard column (50 mm × 3 mm) (Thermo Fisher Scientific, Waltham, USA) of the same material. A mobile phase of 100 mM NaOH at 0.5 mL/min and a linear gradient of sodium acetate (0–120 mM) was used for 25 min. Standards included glucose, cellobiose, laminaribiose and xylose (DP 2 to 6). An identical method was applied to analyse hydrolysis of acetylated chito-oligosaccharides, with N-acetylated chito-oligosaccharides (DP 2 to 5) and glucosamine as standards.

References
1. Sana, B. *In Marine Microbiology* 491–508 (Wiley-VCH Verlag GmbH & Co. KGaA, 2013).
2. Giordanò, A., Andreotti, G., Tramice, A. & Trincone, A. Marine glycosyl hydrolases in the hydrolysis and synthesis of oligosaccharides. *Biotechnology Journal* 1, 511–530, https://doi.org/10.1002/biot.200500036 (2006).
3. Zamost, B., Nielsen, H. & Starnes, R. Thermostable enzymes for industrial applications. *Journal of Industrial Microbiology* 8, 71–81, https://doi.org/10.1002/biot.157785757 (1991).
4. Linares-Pasten, J. A., Andersson, M. & Nordberg Karlsson, E. Thermostable Glycoside Hydrolases in Biorefinery Technologies. *Current Biotechnology* 3, 26–44 (2014).
5. Munoz, R., Rosselló-Móra, R. & Amann, R. Revised phylogeny of *Bacteroidetes* and proposal of sixteen new taxa and two new genera including *Rhodothermaceota* phyI. *Current and Applied Microbiology* 39, 281–296, https://doi.org/10.1016/j.camb.2016.04.004 (2016).
35. Acebrón, I.
36. Bacik, J.-P.
37. Faure, D.
38. Li, Y.-K., Chir, J., Tanaka, S. & Chen, F.-Y.
39. Dodd, D., Kiyonari, S., Mackie, R. I. & Cann, I. K. O.
40. Varghese, J. N., Hrmova, M. & Fincher, G. B.

33. Vocadlo, D. J., Mayer, C., He, S. & Withers, S. G.
34. Martens, E. C., Koropatkin, N. M., Smith, T. J. & Gordon, J. I.
35. Nordberg Karlsson, E., Bartonek-Roxå, E. & Holst, O.
36. Nordberg Karlsson, E., Dahlberg, L., Torto, N., Gorton, L. & Holst, O.
37. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B.
38. Lee, R. C., Hrmova, M., Burton, R. A., Lahnstein, J. & Fincher, G. B.
39. Grondin, J. M., Tamura, K., Déjean, G., Abbott, D. W. & Brumer, H.
40. Varghese, J. N., Hrmova, M. & Fincher, G. B.

6. Alfredsson, G. A., Kristjansson, J. K., Hörnlöfsdottir, S. & Stetter, K. O.
7. Dahlberg, L., Holst, O. & Kristjansson, J. K.
8. Gomes, J. & Steiner, W.
9. Gomes, J. et al.
10. Blücher, A., Nordberg Karlsson, E. & Holst, O.
11. Cantarel, B. L. et al.
12. Nolan, M. et al.
13. Nordberg Karlsson, E. et al.
14. Johnson, J. W., Fisher, J. F. & Mobashery, S.
15. Bhatai, Y., Mishra, S. & Bisaria, V. S.
16. Lee, R. C., Hrmova, M., Burton, R. A., Lahnstein, J. & Fincher, G. B.
17. Faure, D.
18. Davies, G. & Henrissat, B.
19. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B.
20. Martens, E. C., Koropatkin, N. M., Smith, T. J. & Gordon, J. I.
21. Nordberg Karlsson, E.
22. Abou Hachem, M.
23. Garron, M. L. & Henrissat, B.
24. Nordberg Karlsson, E., Dahlberg, L., Torto, N., Gorton, L. & Holst, O.
25. Nordberg Karlsson, E.
26. Abou Hachem, M.
27. Nordberg Karlsson, E., Dahlberg, L., Torto, N., Gorton, L. & Holst, O.
28. Altschul, S. F.
29. Aronsson, A.
30. Altschul, S. F. et al.
31. Johnson, J. W., Fisher, J. F. & Mobashery, S.
32. Martens, E. C., Koropatkin, N. M., Smith, T. J. & Gordon, J. I.
33. Nordberg Karlsson, E., Bartonek-Roxå, E. & Holst, O.
34. Martens, E. C., Koropatkin, N. M., Smith, T. J. & Gordon, J. I.
35. Nordberg Karlsson, E., Bartonek-Roxå, E. & Holst, O.
36. Nordberg Karlsson, E., Dahlberg, L., Torto, N., Gorton, L. & Holst, O.
37. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B.
38. Lee, R. C., Hrmova, M., Burton, R. A., Lahnstein, J. & Fincher, G. B.
39. Grondin, J. M., Tamura, K., Déjean, G., Abbott, D. W. & Brumer, H.
40. Varghese, J. N., Hrmova, M. & Fincher, G. B.

6. Alfredsson, G. A., Kristjansson, J. K., Hörnlöfsdottir, S. & Stetter, K. O.
7. Dahlberg, L., Holst, O. & Kristjansson, J. K.
8. Gomes, J. & Steiner, W.
9. Gomes, J. et al.
10. Blücher, A., Nordberg Karlsson, E. & Holst, O.
11. Cantarel, B. L. et al.
12. Nolan, M. et al.
13. Nordberg Karlsson, E. et al.
14. Johnson, J. W., Fisher, J. F. & Mobashery, S.
15. Bhatai, Y., Mishra, S. & Bisaria, V. S.
16. Lee, R. C., Hrmova, M., Burton, R. A., Lahnstein, J. & Fincher, G. B.
17. Faure, D.
18. Davies, G. & Henrissat, B.
19. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B.
20. Martens, E. C., Koropatkin, N. M., Smith, T. J. & Gordon, J. I.
21. Nordberg Karlsson, E.
22. Abou Hachem, M.
23. Garron, M. L. & Henrissat, B.
24. Nordberg Karlsson, E., Dahlberg, L., Torto, N., Gorton, L. & Holst, O.
25. Nordberg Karlsson, E.
26. Abou Hachem, M.
27. Nordberg Karlsson, E., Dahlberg, L., Torto, N., Gorton, L. & Holst, O.
28. Altschul, S. F. et al.
29. Aronsson, A. et al.
30. Altschul, S. F. et al.
31. Johnson, J. W., Fisher, J. F. & Mobashery, S.
32. Martens, E. C., Koropatkin, N. M., Smith, T. J. & Gordon, J. I.
33. Nordberg Karlsson, E., Bartonek-Roxå, E. & Holst, O.
34. Martens, E. C., Koropatkin, N. M., Smith, T. J. & Gordon, J. I.
35. Nordberg Karlsson, E., Bartonek-Roxå, E. & Holst, O.
36. Nordberg Karlsson, E., Dahlberg, L., Torto, N., Gorton, L. & Holst, O.
37. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B.
38. Lee, R. C., Hrmova, M., Burton, R. A., Lahnstein, J. & Fincher, G. B.
39. Grondin, J. M., Tamura, K., Déjean, G., Abbott, D. W. & Brumer, H.
40. Varghese, J. N., Hrmova, M. & Fincher, G. B.
78. Guindon, S. & Gascuel, O. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Systematic Biology 52, 696–704, https://doi.org/10.1080/10635150390235520 (2003).
79. Boc, A., Diało, A. B. & Makarenkov, V. T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. Nucleic Acids Research 40, W573–W579, https://doi.org/10.1093/nar/gks485 (2012).
80. Whelan, S. & Goldman, N. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. Molecular Biology and Evolution 18, 691–699 (2001).
81. Felsenstein, J. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution 39, 783–791, https://doi.org/10.2307/2408678 (1985).
82. Krieger, E. & Vriend, G. YASARA View—molecular graphics for all devices—from smartphones to workstations. Bioinformatics 30, 2981–2982, https://doi.org/10.1093/bioinformatics/btu426 (2014).
83. Krieger, E. et al. Improving physical realism, stereochimetry, and side-chain accuracy in homology modeling: four approaches that performed well in CASP8. Proteins: Structure, Function, and Bioinformatics 77, 114–122 (2009).
84. Pettersen, E. F. et al. UCSF Chimera—A visualization system for exploratory research and analysis. Journal of Computational Chemistry 25, 1605–1612, https://doi.org/10.1002/jcc.20084 (2004).
85. Krieger, E., Darden, T., Nabaurs, S. B., Finkelstein, A. & Vriend, G. Making optimal use of empirical energy functions: Force-field parameterization in crystal space. Proteins: Structure, Function, and Bioinformatics 57, 678–683, https://doi.org/10.1002/prot.20251 (2004).
86. Bowie, J., Luthy, R. & Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 253, 164–170, https://doi.org/10.1126/science.1853201 (1991).
87. Lüthy, R., Bowie, J. U. & Eisenberg, D. Assessment of protein models with three-dimensional profiles. Nature 356, 83–85 (1992).
88. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. Journal of Applied Crystallography 26, 283–291, https://doi.org/10.1107/S0021889992099944 (1993).
89. Sippl, M. J. Recognition of errors in three-dimensional structures of proteins. Proteins: Structure, Function, and Genetics 17, 355–362 (1993).
90. Wiederstein, M. & Sippl, M. J. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Research 35, W407–W410, https://doi.org/10.1093/nar/gkm290 (2007).
91. Colovos, C. & Yeates, T. O. Verification of protein structures: Patterns of nonbonded atomic interactions. Protein Science 2, 1511–1519, https://doi.org/10.1002/pro.5560020916 (1993).
92. Wegener, A., Sun, T. & Altenbuchner, J. Optimization of an E. coli L-rhamnose-inducible expression vector: test of various genetic module combinations. BMC Biotechnology 8, 2 (2008).

Acknowledgements
Support from the Swedish research council Formas (Grant 2015-769), the BBI, Horizon 2020 project "Macro Cascade" and the Era-nets MBT "Thermofactories" and SusFood2 "ProSeaFood" as well as the NovoNordisk Grant (NNF18OC0034792) is greatly appreciated. Open access funding provided by Lund University.

Author contributions
K.Z.G.A.: Conceptualization, expression and characterization of all six enzymes. Data analysis and M.S. writing. A.M.: Building homology models for all the enzymes, gene cluster analysis for the strain DSM 4252T, conceptualization, homology model validation and M.S. writing. M.G.: Multiple sequence alignment, conserved sequence analysis and M.S. writing. J.L.P.: Gene cloning and MS writing. A.J.: Production of Rm Nag3 protein and Nag3 protein and M.S. (revision) writing. O.H.F.: Cloning of the genes, gene cluster analysis for strain DSM 4253 and M.S. writing. G.O.H.: Genome sequencing of the strain DSM4253 and M.S. writing. S.J.: Phylogenetic analysis of the G.H. family 3, M.S. writing. E.N.K.: Conceptualization, funding acquisition and M.S. writing.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-58015-5.

Correspondence and requests for materials should be addressed to K.Z.G.A. or E.N.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s) 2020