Discerning the Molecular Toolkit of Microalgae and Bacteria through *Omic* Approaches

A WRITTEN EXAM

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Natalia Calixto Mancipe

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BIOPRODUCTS AND BIOSYSTEMS SCIENCE, ENGINEERING AND MANAGEMENT

Brett Barney, Ph.D.

May 2022

Acknowledgements

I would like to thank my adviser Dr. Brett Barney and my lab mates and friends who made this work possible. Many thanks to Ben for keeping me company during late nights in the lab, to Alex for trivia nights and funzies in general, and to Carol for showing me the high level of detail and quality that she puts in her work. Even though we have shared less time together, I also want to thank Peter, Bahar and Luke for being excellent coworkers and friends. Thanks to Dr. Barney for sharing his thoughts on the various projects that interest him, which are the source of the work done in this dissertation, and for teaching me hands on a variety of technical skills.

Parts of this work would not have been possible without the help of experts in various fields. I specially would like to thank Dr. Michael Freeman and his PhD student Aman Imani for their assistance with mass spectroscopy.

I also want to thank the Department of Bioproducts and Biosystems Engineering and MnDRIVE for their financial support.

My family has had to endure my scarce free time, constrained vacations (compared to their standards) and limited international mobility for many years. I am sorry for not being able to spend more time together. Their support, love and example have made me the person I am, so the good and bad results from my deeds should be put on them as well.

I have known Aritra since my first day in graduate school. There is no space to write all he has done and continues doing for me. I want to thank him for his affection, enormous support and companionship throughout the ups and downs of these years. Thank you to Pablo and Julia, who have been friends-family for so many years. They are a connection to my younger and more carefree self that always brings joy and comfort to my life. And last, but not least, thanks to the latest addition to my family, Rummi, the sweetest cat.

Dedication

To my beloved family.

To my father. Here I give you work that I am proud of, I hope it reaches you and gives you a smile.

Abstract

Bacteria and microalgae present phenotypes of great advantage for a variety of biotechnological applications; however, the molecular tools they use to perform what we observe as beneficial traits are still unknown. This work aims to shed light on three specific processes: the production of waxes, the secretion of sugars and the degradation of plastics. Through the application of bioinformatic and evolutionary principles, we explore the targeted manipulation of the bacterial Wax Ester Synthase Ma1 to better understand its substrate preference, we also examine algal genomes and transcriptomes in the search of the genes behind the secretion of glucose and maltose, and we dig into the plastic degradation capabilities of Minnesotan bacteria. From the isolation of potential plastic degraders to the sequence analysis of algae with a specific phenotype and the targeted manipulation of Ma1, all projects aim to benefit our environment by discovering alternative sources of fuels and chemicals or by searching for better ways to manage our waste.

Table of Contents

List of tables

List of figures

Introduction

The unknowns

The work within this thesis is an application of bioinformatic and evolutionary principles to uncover the molecular tools that bacteria and microalgae use to perform what we observe as traits of interest. Since their molecular tools are encoded in their genomes, this thesis explores various approaches to dissect the meaning underneath their genomic data at the gene sequence, gene expression and protein expression levels.

Chapter 1 explores the effect of varying residues among the substrate binding regions of the enzyme wax ester synthase. This project was born from the multiple sequence alignment of ortholog wax ester synthases and the determination of conserved and variable residues. In chapter 2 the genomic sequence of a green microalga, *Scenedesmus glucoliberatum* PABB004, is defined and the coding regions within its genome are predicted as well as their possible function. Chapter 3 builds upon the results of chapter 2 and uses the genomic data of *S. glucoliberatum* along with other green microalgae to dissect the genomic tools that allow the release of simple sugars like glucose and maltose to the extracellular environment. Chapter 4 presents the preliminary work done towards understanding the bacterial degradation of paraffin, which ultimately aims to contribute to our knowledge of plastic biodegradation. The thesis ends with a short conclusion on the analysis of the -omic tools used throughout each chapter.

Omic tools

For this work our laboratory has used well stablished technologies for nucleic acid sequencing. We have sequenced and annotated one eukaryotic genome (*S. glucoliberatum* PABB004's) and 17 bacterial genomes from strains isolated from environmental samples or plastic waste. We have also sequenced, assembled, and annotated the transcriptomes of 6 green algae (*S. glucoliberatum, Coelastrum sp.* PABB002, *S. dimorphus, S. naegelii, C. sphaericum* UTEX 1354 and *C. proboscideum var. gracile* UTEX 184). All data produced in this work has been shared in public databases for the general use of the scientific community.

Evolution, homology and sequence similarity

What we see as evolution from an organism X to Y is widely thought to be the joint effect of random changes on X's genome that by chance either helped it thrive on its living conditions, or at least were not harmful. Harmful changes would be eliminated from the genetic pool by a competitive disadvantage (also known as "natural selection"). When X reproduces and passes its genome to its progeny the changing process continues until the genetic information is different enough to consider its holder a new organism (Y). In other words, *evolution can be considered as an ongoing addition of random and cumulative changes on a sequence that is biased towards beneficial traits*. This idea implies that there is a "path of changes" that connects Y's sequence to X's.

The concept of sequence similarity is used by many bioinformatic approaches and is based on computing the minimal path that connects two sequences (i.e., how many changes must be done to X's sequence to become Y's). In theory, the shorter the path, the more similar the sequences are. It is important to remark that "sequence similarity" is a metric calculated by comparing two sequences[1].

Homology, on the other hand, is an evolutionary concept (not a metric). *Two sequences are considered homologous if they share a common evolutionary ancestor*. *Homology is often inferred from sequence similarity*. If two sequences have more similarity than what would be expected by chance, they are presumed to have come from a common ancestor[2].

Gene function prediction: homologs, orthologs and paralogs

Gene annotation (i.e., the prediction of gene structure and function) is performed based on the search for homologous sequences. In similarity-based annotation approaches, homology is inferred from significantly similar sequences within a database. Some of the most common algorithms for sequence similarity searches are Blast[3,4] and HMMER[5] and highly useful databases are NCBI's Non-redundant (nr), SwissProt-Uniprot, Pfam, InterPro and the Kyoto Encyclopedia of Genes and Genomes (KEGG)[6,7].

The functional assignment relies on finding significant similarities between a protein (or gene) sequence with unknown function (the query) and a sequence with known function (the subject, part of the database). Highly similar sequences are assumed to be homologs (to have a common evolutionary ancestor) and thus are assumed to be connected through evolution. Since protein sequences determine the 3D protein structure, and the protein structure is required for its function, the idea behind a similarity-based function assignment is that highly similar sequences must have a similar structure and a similar function.

This approach, although commonly used, presents several challenges. In general, it makes sense when comparing "orthologs", which are homologs (sequences with a common ancestor) that diverged due to speciation. But one should be more cautious when comparing "paralogs", which are homolog sequences that appear in a species by duplication of parts of its own genome[2]. Although it might seem a technicality, the implications for gene function are important. If in a genome there are two copies of the same gene, random changes in one of the copies do not have such a pronounced effect on the organism because the other copy serves as a back-up system. Since mutations in one of the copies are better tolerated, paralog genes tend to have a higher diversification of functions even though they share a high sequence similarity. Furthermore, even for orthologous sequences the significance of the similarity depends on the size and content of the database used; the bigger the database, the less certainty in the inference but if the database is too small it is a problem too.

Additionally, genomic changes are not limited to simple sequence modifications or gene duplications. There are several other events like gene deletions, gene fusions, horizontal gene transfer and recombination, which make the inference of homology and function far more complex[8]. All these possibilities put together generate an organism's unique evolutionary history and with it the set of functionalities that determine its characteristics and behavior ("phenotype").

Summary

The evolutionary history of organisms is a complex process with multiple avenues for change. Understanding it helps us infer the functionalities of newly discovered sequences and determine the molecular basis of phenotypic traits.

The following key points are needed to frame all sequence comparisons throughout this thesis:

- The genomic sequence of an organism stores and encodes all required information to perform its living duties. In other words, the genotype encodes the phenotype.
- Bioinformatic approaches use sequence similarity to infer homology and functionality, but this approach only covers part of the possible genomic changes that occur during evolution.
- Homologs, orthologs and paralogs are evolutionary concepts. Homologs have common evolutionary ancestors, orthologs are homologs that diverged through speciation and paralogs are homologs that were generated through gene duplication.
- Homologous sequences can have similar functions (as many orthologs do) or have different functions (as some paralogs do). To go around this, function-oriented ortholog groups have been created[6], which only contain orthologs and paralogs with the same function. Functional orthologs do not necessarily share high sequence similarity.

Chapter 1: Canvasing the Substrate Binding Pockets of the Wax Ester Synthase

Natalia Calixto Mancipe1, Kalene M. Mulliner1, Mary H. Plunkett2 and Brett M. Barney1, 2, *

1 Department of Bioproducts and Biosystems Engineering, University of Minnesota, St. Paul, MN 55108, USA, 2 BioTechnology Institute, University of Minnesota, St. Paul, MN 55108, USA

ORCID: Brett Barney, 0000-0002-5976-5492; Natalia Calixto Mancipe 0000-0002-2558-1126;

This chapter is an almost verbatim reproduction of the manuscript published with the same title [9].

Synopsis

The biosynthesis of wax esters and triglycerides in bacteria is accomplished through the action of the wax ester synthase/acyl-coenzyme A:diacylglycerol acyltransferase (WS/DGAT) also known as wax ester synthase. A hallmark of these enzymes is the broad substrate profile that accepts alcohols, diglycerides and fatty acyl-CoAs of various carbon chain lengths and degrees of branching. These enzymes have a broad biotechnological potential due to their role in producing high-value lipids or simple fuels similar to biodiesel through biosynthetic routes. Recently, a crystal structure was solved for the wax ester synthase from *Marinobacter aquaeolei* VT8 (Maqu_0168), providing a much clearer picture of the architecture of this enzyme, and enabling a more precise analysis of the important structural features of the protein. In this work, we used the structure to canvas amino acids lining the proposed substrate binding pockets and tested the effects of exchanging specific residues on the substrate profiles. We also developed an approach to better probe the residues that alter fatty acyl-CoA selectivity, which has proven more

difficult to investigate. Our findings provide an improved blueprint for future efforts to understand how these enzymes position substrates for catalysis and to tailor or improve these enzymes in future biosynthetic schemes[10,11].

Introduction

The production of tailored wax esters through biosynthetic approaches presents a major opportunity to replace conventional petrochemicals with sustainable alternatives, particularly for fuels and high-value lubricants. Wax esters are a biologically produced class of chemicals generated when a fatty acid and fatty alcohol are combined to form an ester bond near the center of the new waxy compound (Figure 1). In nature, waxes are produced by wax ester synthase/acyl coenzyme A:diacylglycerol acyltransferases (WS/DGAT or wax ester synthase, E.C. 2.3.1.75)[12]. Understanding the precise locations where fatty alcohols and activated fatty acids (acyl-CoAs) bind to the WS/DGAT is an important first step toward engineering new biosynthetic pathways. Likewise, understanding the residues that are amenable to modifications in the protein sequence which alter substrate specificity without dramatically decreasing overall activity is required for more intricate studies to redesign or engineer these enzymes for more tailored applications. We thus chose to probe the wax ester synthase Ma1 from *Marinobacter aquaeolei* VT8 binding pockets in more detail to develop a clearer picture of the effects of various mutations.

This class of biological lipids have been recognized for more than two centuries. They are found broadly distributed through higher organisms as a component of the cuticle layer on the epidermis of plant tissues or the epidermis of skin in mammals and are also found in insects[13,14]. Wax esters can range in size based on the precursor fatty acids or fatty alcohols that serve as feedstocks. Many biological wax esters are derived from C14, C16 and C18 chain length fatty acids and alcohols, though this can vary for certain specific waxes[15].

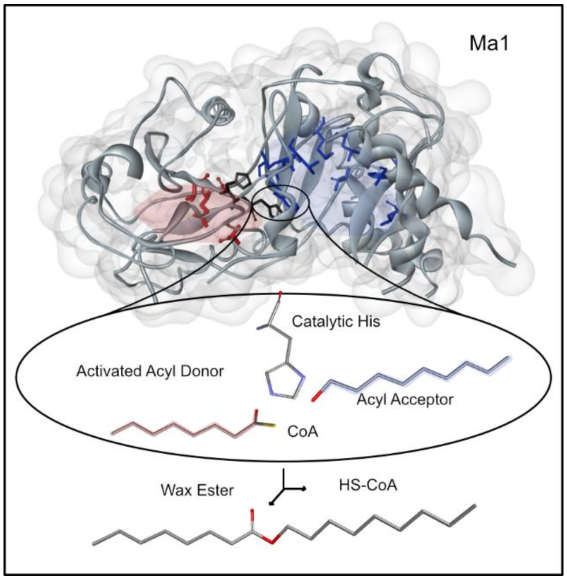


Figure 1. Scheme of wax ester synthesis by the WS/DGAT Ma1

Ma1 binds an activated fatty acid (Acyl-CoA) and a fatty alcohol or diglyceride and catalyzes the formation of an ester bond between the acid and the alcohol groups, freeing one mol of coenzyme-A per mol of wax ester produced. The figure shows Ma1's crystal structure with the catalytic residues (in black) at the center of the image and the putative binding pockets for the activated acid substrate (red) and fatty alcohol / diglyceride (blue).

Prior to the ban on commercial whaling, wax esters were isolated from spermaceti of sperm whales or from rendered blubber. Plant-based alternatives include feedstocks such as jojoba oil[16,17]. Wax esters are not as common in the microbial world, but are found in a small selection of bacteria, where they sometimes serve as an alternative carbon storage compound in place of polyhydroxyalkanoates, which are more common, and often referred to as bioplastics[18].

The bacterial WS/DGAT is a bifunctional enzyme that will add the fatty acid component of an activated fatty acid that is attached to coenzyme A (acyl-CoA) to either a fatty alcohol or a diacylglycerol molecule. In this manner, this enzyme catalyzes the final step in producing both wax esters and triglycerides. The ability to utilize such a diverse group of alcohols as this second substrate results in an enzyme with a very broad substrate range, leading to the characterization of these enzymes as being promiscuous in their substrate choice[19]. The range of organic alcohol compounds that are accepted by this enzyme is quite large, while the substrate profile for the fatty acid component is somewhat narrower.

The first bacterial wax ester synthase was identified from an eloquent study done with *Acinetobacter calcoaceticus* ADP1 in 2003 (later renamed *Acinetobacter baylyi* ADP1). Since that time, this enzyme has received a great deal of attention in relation to the breadth of substrates it will accept, and the potential for applications with this enzyme in biosynthetic schemes[12,18,20–23]. Biochemical characterization was performed, and additional enzymes from other bacteria were identified[24–27]. Since the discovery of this enzyme, there has been significant interest in obtaining an atomic model of the wax ester synthase to assist in potential bioengineering approaches, and several laboratories have attempted to obtain crystals of the enzyme to perform a structural analysis. In 2018, a structure (pdb 6CHJ) was obtained for a wax ester synthase (NCBI accession number WP_011783747.1) from *Marinobacter aquaeolei* VT8[28]. It is important to mention that this strain was later renamed *M. hydrocarbonoclasticus* VT8 and then more recently *M. nauticus* VT8[29,30]. Since the publication of the crystal structure, there have not been any reports on the structural base for the substrate specificity of this enzyme. In this study, we have performed a broad survey of various amino acids that line the putative substrate channels of the wax ester synthase in an effort to obtain more detailed evidence to support the notion that these binding pockets are indeed the locations where the two substrates of the enzyme bind during catalysis. Based on historical precedence and the predominant substrate preference of the enzyme, we will refer to this enzyme in this work as the wax ester synthase from *M. aquaeolei* VT8.

Results

Probing the residues lining the putative fatty alcohol binding pocket

The wax ester synthase catalyzes the formation of an ester bond arising from a fatty alcohol and an activated fatty acid. We previously developed an assay to probe the active site preference for a range of fatty alcohols of different chain lengths using a gas chromatography assay that includes equimolar amounts of various fatty alcohols, and a specific molar excess of the activated fatty acid palmitoyl-CoA. This assay is characterized by high reproducibility and, when run together with a wild-type control (Ma1), is able to differentiate subtle changes that modified amino acid residues have on substrate selectivity. The assay previously revealed that residue A360 plays an important role in determining the preference for long chain alcohols prior to any definitive structural information related to this enzyme. The recent structural determination of Ma1 by Petronikolou and Nair provided a more complete molecular view of this enzyme, confirming the location of the A360 lining a cavity near the active site H136 residue, and providing an opportunity to probe additional residues to better canvas the potential substrate binding pockets.

Figure 2 shows the putative fatty alcohol binding pocket of Ma1 identified from the crystal structure (pdb 6CHJ). We generated mutants for multiple amino acids that line the binding pocket and tested the impact these mutations had on fatty alcohol substrate selectivity. In each case, mutants were tested along a Ma1 control and performed in triplicate. Mutations that altered the substrate profiles for fatty alcohols are shown in Figure 2. Residues L14 and M21 are in close proximity to one another, lying near the bottom of the cavity shown in Figure 2, and resulted in similar changes in the substrate selectivity profile. Larger non-polar amino acid substitutions at A360 resulted in a substantial shift in preference for fatty alcohols around C9 and remains the most pronounced impact of the residues for improving selectivity for C9 or C8 alcohols, though a small shift toward C9 and C10 alcohols was found when L356 was decreased in size. Residues L363 and V377 both shifted preference toward slightly longer alcohols when mutations were made to increase the size of the side chains.

In addition to comparing the profiles with larger alcohols, we also tested a selection of smaller alcohols with unique features, including isoamyl alcohol, 2-phenyl ethanol and ethanol. Figure 3 shows residues of the putative fatty alcohol binding pocket that resulted in substantial changes to the substrate selectivity profile when tested with a mixture of these three alcohols and n-hexanol. In this assay, concentrations of these substrates were selected that resulted in roughly similar levels of product produced with Ma1. This assay had previously identified L356 and M405 as playing an important role in determining selectivity of these smaller alcohols. Additional residues from this putative binding pocket that resulted in changes to these substrate profiles included F11 when substituted with a tryptophan, I403 when substituted by alanine, leucine and phenylalanine. Residue L408 is also in close proximity to M405 and I403, and also altered the substrate profile when changed to phenylalanine. Residues V377 and F292, also altered the smaller alcohol profiles. Residue N411, which is highly conserved among wax ester synthases and lies in close proximity to I403 and L408, resulted in an inactive enzyme when substituted with a valine residue.

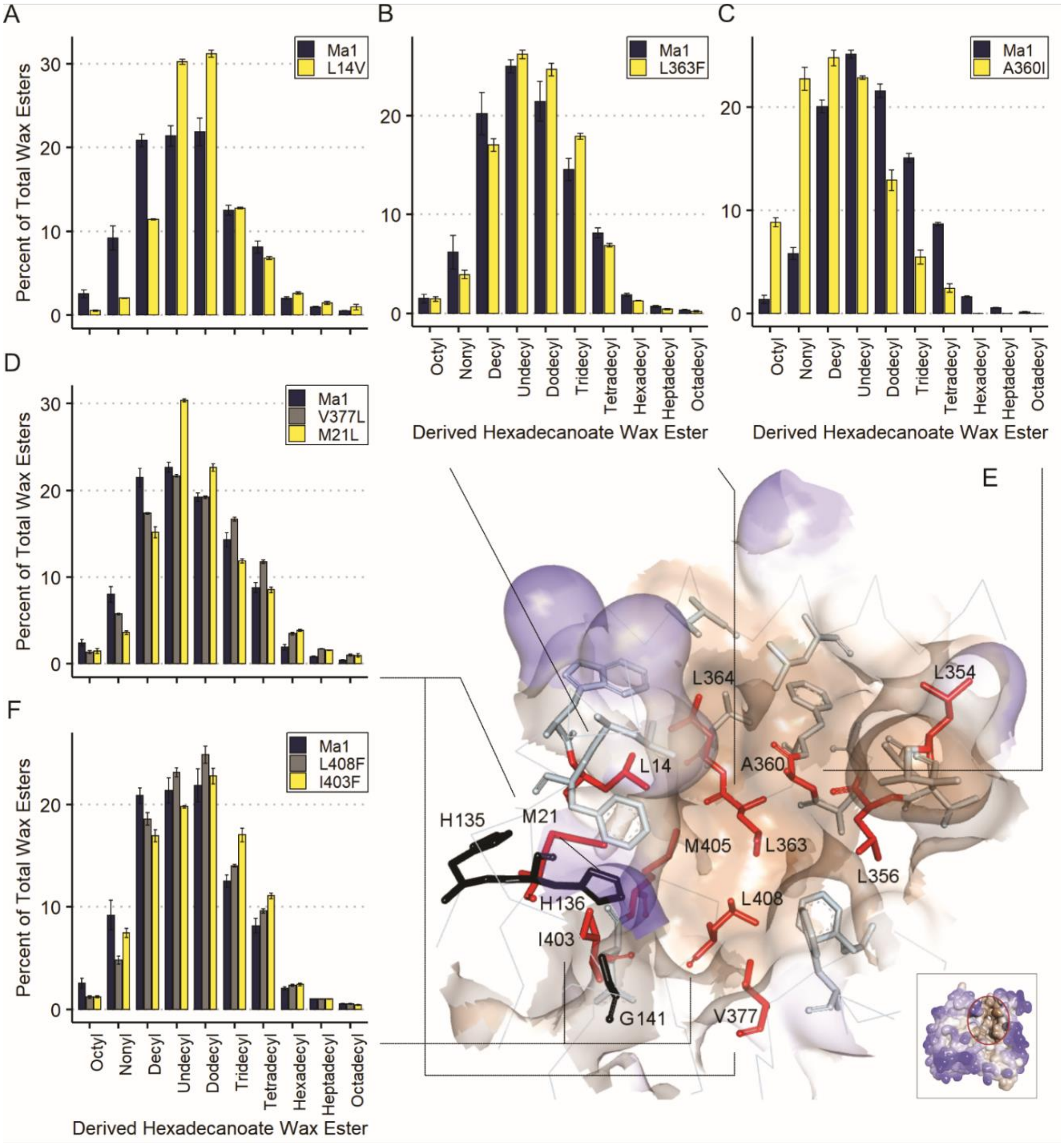


Figure 2. Substrate selectivity assays with various straight-chain fatty alcohols

Assays were performed by mixing a selection of fatty alcohols with palmitoyl-CoA and allowing the reaction to proceed until completion. Products were extracted and analyzed by gas chromatography. Error bars represent standard deviation (n=3). Each data set contains a wild-type (Ma1) control. Panels A-D, and F show product profiles of various mutations found to alter fatty alcohol selectivity and the location of these amino acid residues depicted from the crystal structure (pdb 6CHJ) shown in panel E.

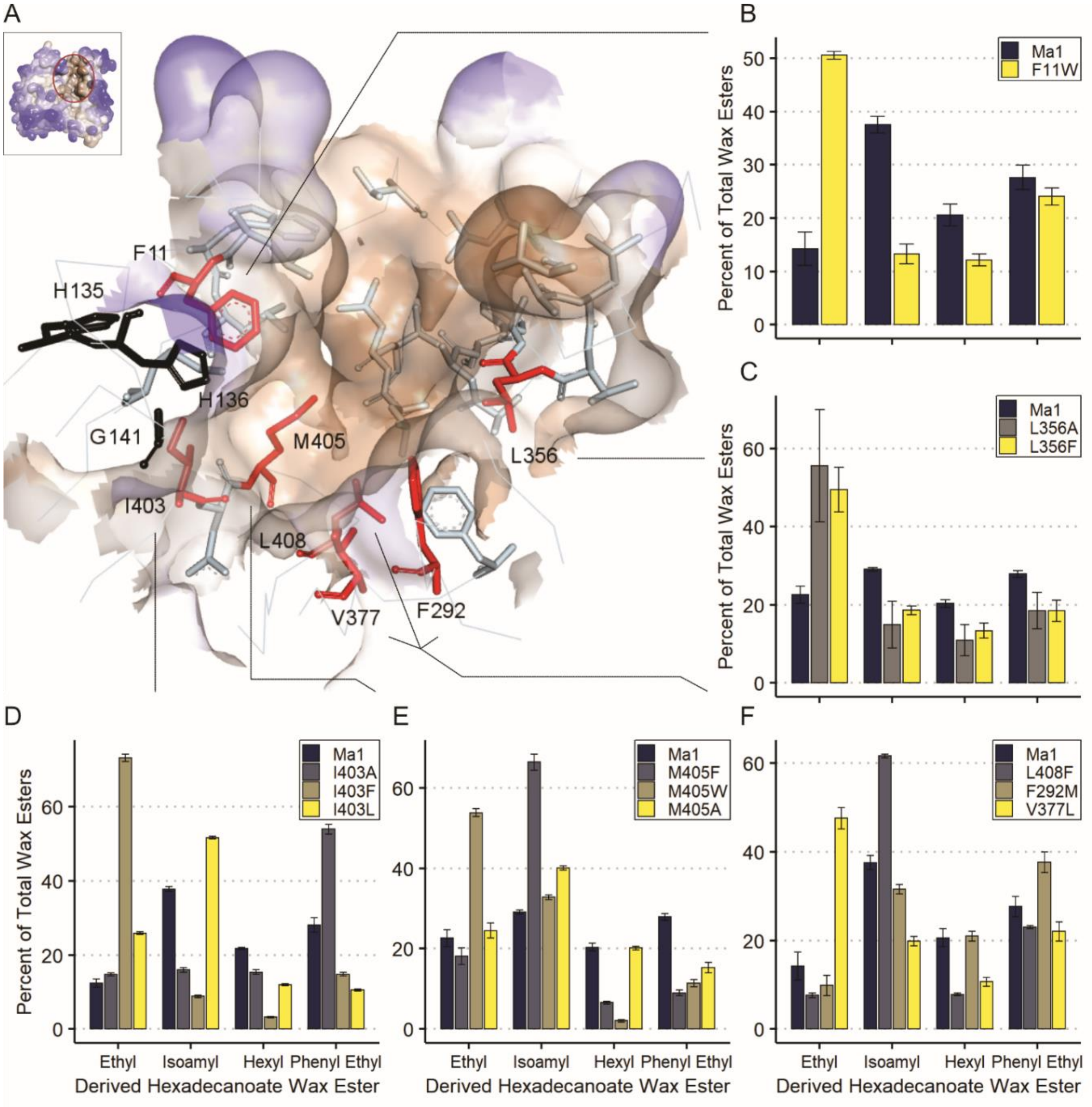


Figure 3. Substrate selectivity assays with various small and medium alcohols.

Assays were performed by mixing ethanol, isoamyl alcohol, n-hexanol and 2-phenyl ethanol with palmitoyl-CoA and allowing the reaction to proceed until completion. Products were extracted and analyzed by gas chromatography. Results are the mean and standard deviation (n=3). Each data set contains a wild-type (Ma1) control. Panel A shows the location of these amino acid residues depicted from the crystal structure (pdb 6CHJ), while panels B-F show product profiles of various mutations found to alter substrate selectivity.

Probing the residues lining the putative fatty acyl-CoA binding pocket

Testing the substrate preference of the fatty acyl-CoA binding pocket has proven to be more difficult than using the alcohol mixtures described above for the fatty alcohol binding pocket. To probe the selectivity of this putative pocket, we determined rates of specific activity with fatty acyl-CoAs of varying carbon length, including C16, C14, C12, C10 and C8, all fully saturated. Assays were performed in triplicate with controls to confirm that the enzymes had not lost significant activity during the duration of the assay period for each mutant. We generated a range of mutations near the putative fatty acyl-CoA binding pocket that was described previously. Residues A144 and G25 were tested previously, along with D8 and D140. Residues Q27, T132 and A148 are all in close proximity to one another. Modifications to Q27, T132 and A148 that replaced the sidechain with larger non-polar residues all resulted in similar changes to the specific activity profiles of these mutants (Figure 4). In each case, the activity for C10-CoA improved, while the activity for C16- and C14-CoAs decreased. Residues G25 and A144 lie closer to the H136 catalytic residue, and mutations for these residues resulted in further differences. Modification of A144 to a phenylalanine resulted in a more pronounced improvement toward C10-CoA, and a further decrease with C16-CoA, but did not alter C14-CoA activity. Modification of A144 to isoleucine resulted in a more subtle change in the profile, similar to the modifications to Q27, T132 and A148, and also decreased C14-CoA activity. Modification of G25 to valine resulted in a dramatic decrease in activity toward C16-, C14- and C12-CoAs, but was still able to catalyze reactions with C10-CoA at levels similar to the wild-type enzyme. For C8-CoA, only A148 to isoleucine and A144 to phenylalanine showed large differences, with all other modifications yielding similar activities to the wild-type enzyme.

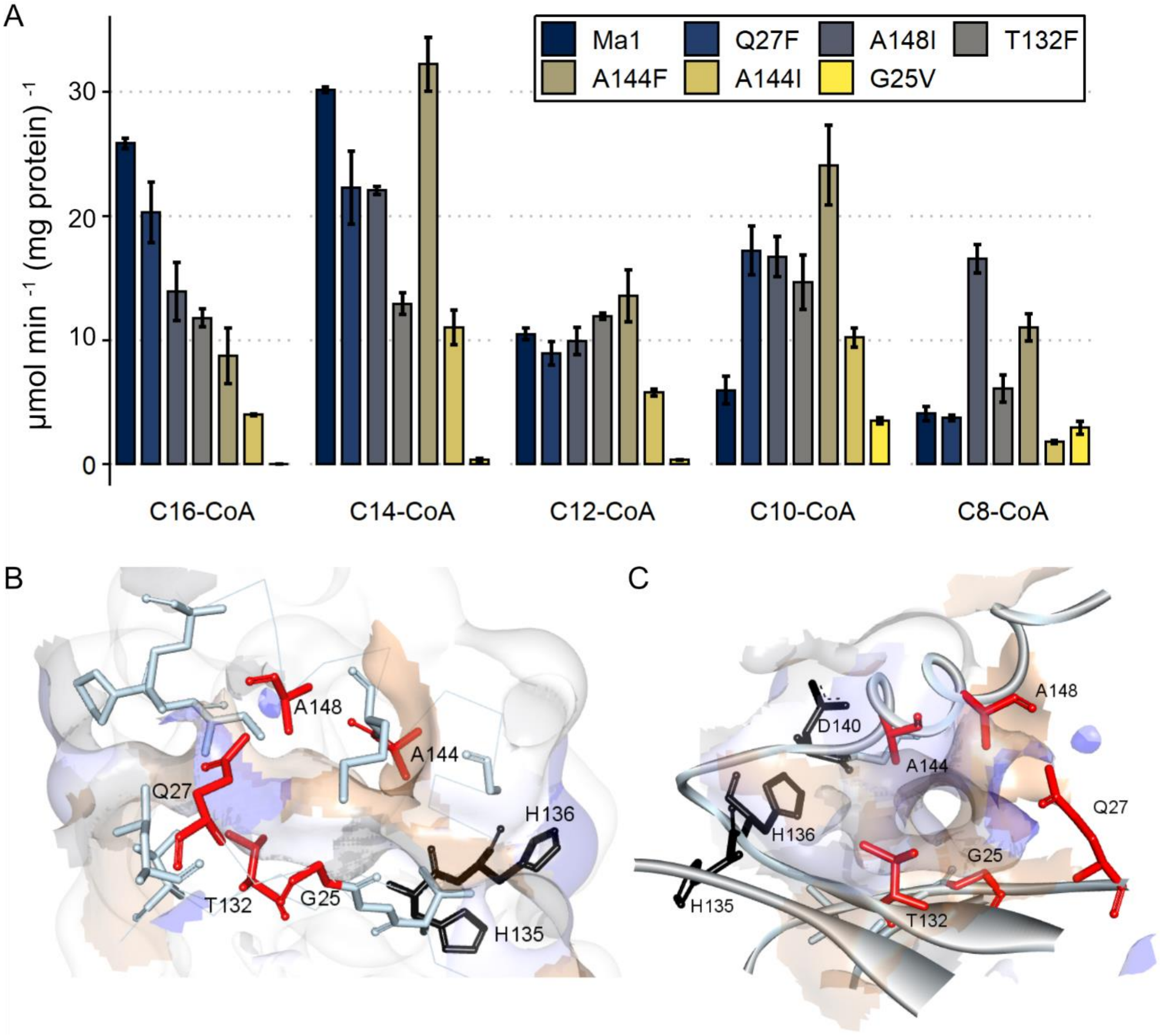


Figure 4. Specific activity assays with various fatty acyl-CoAs.

Panel A shows specific activity measurements performed by combining 1-dodecanol with the specific fatty acyl-CoA. Results represent the average for at least three replicates. Error bars represent standard deviation. Panels B and C show the location of these amino acid residues depicted from the crystal structure (pdb 6CHJ) from different perspectives.

Discussion

The wax ester synthase/acyl coenzyme A:diacylglycerol acyltransferase (WS/DGAT) was first identified from *Acinetobacter calcoaceticus* ADP1 by Kalscheuer et al in 2003 and further characterized by Stöveken et al in 2005. This enzyme catalyzes the final step in the production of wax esters from long-chain alcohols and activated fatty acids and has a biotechnological application as an alternative source of fuels and lubricants commonly obtained from petrochemical processes.

Structural studies provide important details of the atomic resolution for enzymes but are strengthened by biochemical and mutagenesis studies that further inform specific aspects of the protein that enable catalysis and substrate binding and recognition. Ma1 was selected by our laboratory for further biochemical studies following a comparison with four other wax ester synthases, as its heterologous expression in *E. coli* resulted in a product with high purity and activity.

Since the initial identification of the bacterial WS/DGAT (wax ester synthase) enzyme from *Acinetobacter calcoaceticus* by the Steinbuchel laboratory, many additional genes from a range of bacteria have been identified as a result of various sequencing projects. In a number of cases, genes have been cloned and their corresponding enzymes isolated and characterized to confirm activity[31–33]. While the catalytic residues and other regions of these enzymes share a high degree of sequence conservation, the residues that have been identified here and in previous studies that constitute the binding pockets show a much lower degree of conservation (Figure 5). This is particularly interesting, as many of these enzymes have been confirmed to yield wax esters or triglycerides in their native hosts. The level of variation in the primary sequence of these enzymes indicates an enzyme chassis that is highly amenable to manipulation. In some cases, these variations may come at a substantial cost in terms of lower specific activity. Indeed, in comparison studies, large differences in activity were found for substrates that are known to be the primary substrates in their native hosts.

```
    -10 -11 -14  -21 -25 -27  -132 -135 136 -140 141  -144 -148  -292  -356 -359 360 -363 364  -374 -377  -403 -405 -408 -411
Ma1 .LF..L. .M…G.Q. .T..HH…DG. .A..A. .F. .L..AA..LL. .F..V. .I.M..L..N.
Av2 .LF..L. .M…G.Q. .Y..HH…DG. .A..A. .M. .M..TG..LL. .F..V. .I.L..I..N.
Pc1 .LF..L. .M…G.F. .F..HH…DG. .A..V. .F. .Y..EG..LA. .F..I. .I.F..Q..N.
Ac1 .IF..L. .M…G.F. .F..HH…DG. .G..I. .M. .Y..AG..II. .F..V. .I.L..Q..N.
Ma2 .SW..V. .M…T.Q. .T..HH…DG. .G..M. .G. .M..YI..LM. .F..T. .L.A..G..N.
Av1 .GW..M. .M…G.A. .T..HH…DG. .G..M. .Q. .N..LS..QV. .F..V. .L.L..Y..N.
Rj1 .IF..G. .M…S.E. .T..HH…DG. .G..L. .M. .M..IA..PV. .F..I. .V.L..Q..N.
Go1 .SF..L. .M…G.F. .L..HH…DG. .G..L. .M. .P..TL..VM. .F..L. .P.M..L..N.
St1 .AF..I. .M…A.G. .F..HH…DG. .A..A. .L. .P..GH..GG. .F..L. .P.A..Q..A.
```

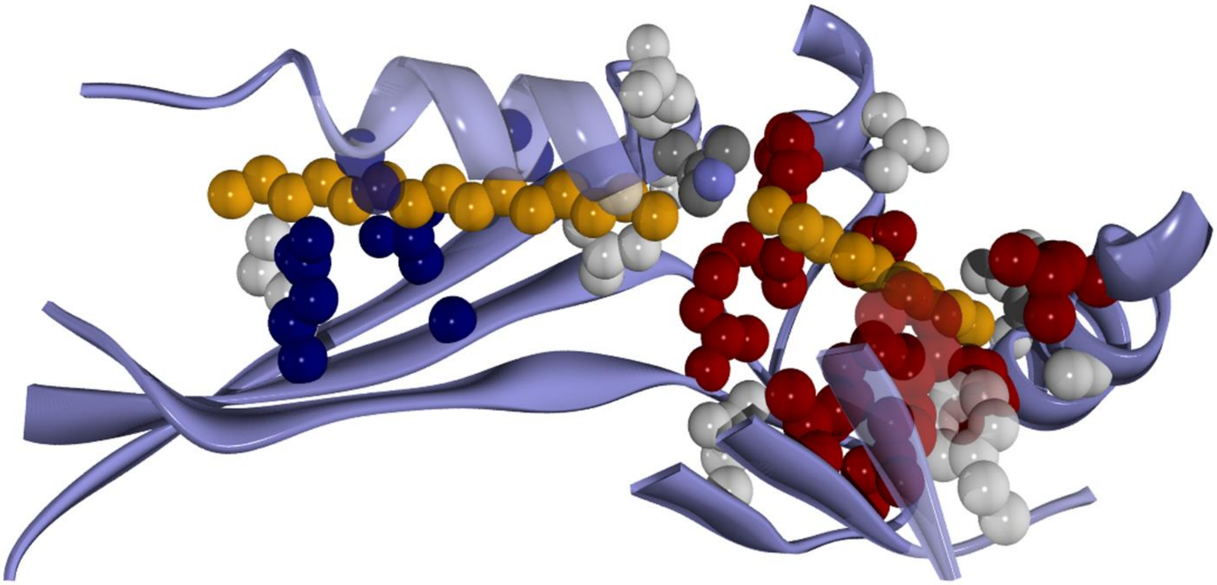


Figure 5. Alignment of protein primary sequences from enzymes confirmed to display wax ester synthase activity.

Shown above are alignments of the primary sequences of wax ester synthases from *M. aquaeolei* VT8 (Ma1 and Ma2, Maqu_0168 and Maqu_3067), *Alcanivorax borkumensis* (Av1 and Av2, ABO_2742 and ABO_1804), *Psychrobacter cryohalolentis* (Pc1, Pcryo_0247), *Acinetobacter baylyi (Ac1, ACIAD0832)*, *Rhodococcus jostii* (Rj1, RHA1_ro01601), *Gordonia* sp. KTR9 (Go1, KTR9_3844), *Streptomyces avermitilis* (St1, SAVERM_7256). Numbering is for Ma1. Alignments were performed using Multalin[34]. Accession codes are from the KEGG database. Selected sections represent the fatty alcohol binding pocket and the fatty acyl-CoA binding pocket, along with the conserved HHXXXDG motif representing the H136 catalytic residue, in conventional atom coloring. Residues found to alter specificity of fatty acyl-CoA are noted in blue (above and below) in the alignments and structural representation. Residues found to alter small and fatty alcohols are colored red. Aliphatic chains representing the acyl groups of the fatty acyl-CoA and the fatty alcohols have been modeled into the binding pockets. Image of active sites and binding pockets was prepared using POV-Ray.

Our research group recognized that, despite the very minimal primary sequence homology between the phthiocerol dimycocerosyl transferase PapA5[35] and the Ma1 wax ester synthase, the structure of PapA5 could be used as a guide to select several residues as potential targets for altering substrate selectivity. Based on this initial selection of residues, we hypothesized that the A360 residue would lie within one of the potential substrate binding pockets. Modifications of this alanine to larger non-polar side chains did indeed result in alterations to the substrate selectivity of fatty alcohols. Based on this initial assessment, we probed additional residues for activity with a range of alcohol substrates, identifying both the L356 and M405 residues as likely residing within the same binding pocket. While our own attempts to obtain crystals of Ma1 were unsuccessful, another laboratory successfully obtained crystals of the Ma1 enzyme from *M. aquaeolei* VT8, representing the first atomic-scale structural information related to a member of the WS/DGAT superfamily of enzymes.

While the structure of the Ma1 wax ester synthase provides the first glimpse of this enzyme at atomic resolution, there are still significant regions in the structure and the binding pockets that remain incomplete. Several sections of the wax ester synthase contained disordered regions with poor electron density, resulting in blind spots for significant segments of the structure. Importantly, one specific segment is near the entrance to the alcohol binding pocket, representing residues P168 to G191, leaving a void of twenty-two residues. Additionally, while crystals were grown in the presence of particular substrates, a substrate homolog was not found bound to the enzyme. This means that the precise location where the substrates bind to the enzyme remain uncertain, though prior studies by our group and additional mutations provided by Petronikolou and Nair do provide preliminary evidence for where these substrates might bind.

The fatty alcohol substrate binding pocket has been previously shown to accept a broad range of alcohol substrates, though the most common fatty alcohol found in the native wax esters of most bacteria are generally made of C14, C16 and C18 derived alcohols with the potential to contain unsaturated carbon bonds in the carbon backbone. The low specificity of these enzymes has resulted in them being referred to as promiscuous. Additionally, as has been pointed out previously, the residues found in the regions of these active sites have low conservation in comparison to the residues that account for catalysis[36].

A particularly interesting aspect of these enzymes is that they must contend with a unique problem, differentiating between two substrates that share significant similarity in the acyl constituent of the substrate. Indeed, the fatty alcohols produced by bacteria such as *Marinobacter aquaeolei* VT8 are derived from the fatty acid pool within the cell, and are reduced from fatty acyl-CoA or fatty acyl-ACPs by specific enzymes into the fatty alcohols[37–40]. As such, the chemical structure, surface area, degree of unsaturation, shape and polarity of most of the substrate, is nearly identical. Activating the fatty acid by attaching it to coenzyme A may help to properly align this substrate into the correct binding pocket, but the fatty alcohol component is likely to have as high an affinity for the fatty acid binding pocket as it does for the fatty alcohol binding pocket. In this manner, it is quite possible that the two similar substrates may serve as potential inhibitors for one another.

The binding pocket for the fatty alcohol has been defined based on the proximity to the H136 catalytic residue and based on evidence from prior studies that indicated L356, A360 and M405 should all lie within this binding pocket. As anticipated, these residues were all found within a large hydrophobic pocket in the Ma1 crystal structure. This pocket is characterized by a large internal surface that is populated with predominantly non-polar side chains. Additional residues lining this pocket include L10, F11, L14, M21, F292, A359, L363, L364, F374, V377, A403, L408 and N411. Our results demonstrate alterations in alcohol selectivity for mutations made for F11, L14, M21, F292, L356, A360, L363, V377, I403, M405, L408 (Figures 2 and 3), providing further evidence that this pocket is the primary site of alcohol binding and selectivity in the wax ester synthase. Substitutions to A357, A359 and L364 selected in this study did not significantly alter substrate selectivity toward fatty alcohols in comparison to other residues of the pocket (results not shown). Modification of N411 to a valine residue resulted in an inactive enzyme that could not be assayed. A large void at the top of the binding pocket may represent part of the residues near G191 that are missing from the structure and might be further probed in the future to provide evidence about their role in the absence of structural information related to these residues. It is also possible that additional structural studies may yield crystals with sufficient order in these regions, or the co-crystallization with other substrates might improve the order in this region of the crystal.

The fatty alcohol binding pocket of Ma1 is much larger than the average size of a fatty alcohol. It is important to recognize that this enzyme is also capable of catalyzing reaction to the free alcohol on a diglyceride in addition to fatty alcohols, yielding triglycerides instead of wax esters. This dual use is an aspect of the promiscuity that is attributed to these enzymes. There remains some debate as to whether all wax ester synthases are also able to catalyze the reaction with diglycerides. It has been reported that certain wax ester synthases do not catalyze this reaction with diglyceride, however, Ma1 (and the highly similar WS1 from *M. hydrocarbonoclasticus* DSM 8798, accession number EF219376) have been reported to catalyze the production of triglycerides and was found here to yield a specific activity with dilinoleoyl glycerol of 820 ± 60 nmol min^{-1} (mg of protein)$^{-1}$ in the wild-type Ma1 enzyme. It was pointed out that this fatty alcohol binding pocket of Ma1 is sufficiently large to accommodate a diglyceride. It has also been demonstrated that wax ester synthases can catalyze reactions with alcohols as small as ethanol, making it a strong target for the production of biodiesel style molecules in various biosynthetic approaches. Because wax ester synthase is able to catalyze such a broad range of chemical reactions, and because these enzymes are found to produce both wax esters and triglycerides even in their bacterial hosts, it remains uncertain whether the original substrate binding pocket accommodated diglycerides and evolved to also accept fatty alcohols, or accommodated fatty alcohols, and evolved to also accept diglycerides. The fatty alcohol binding pocket revealed from the crystal structure is indeed large enough to accommodate very large substrates. Similar enzymes are presumed to produce beeswax[41], which utilizes fatty alcohols of C24 to C34 lengths, indicating that substrates may likely penetrate through the alcohol binding pocket during catalysis.

The fatty acyl-CoA binding pocket was more difficult to probe using various substrates than the approach using fatty alcohols. Mixtures of fatty acyl-CoAs resulted in very low activities and incomplete reactions. We had previously uncovered minor differences between fatty acyl-CoA substrate selectivity in enzymes obtained from various organisms using a simple measure of specific activity with various substrates. In this work, we revisited this approach to analyze several mutations that were described in a prior report, and probed several additional residues that should lie further from the catalytic residues if this location were indeed the second binding pocket where fatty acyl-CoA binds in the wax ester synthase. Direct comparisons to the results reported by Petronikolou and Nair are difficult to make, as their assay used 1-hexanol as an alcohol and C16- and C6-CoAs for comparisons and was based on a gas chromatography analysis. While 1-hexanol is a substrate for wax ester synthase, it results in a pronounced decrease in activity versus 1-dodecanol (C12-OH), which is used as the primary alcohol in our real-time spectrophotometric acyl-CoA assays. We favor the use of real time assays to calculate specific activity. Real time assays can be performed over a shorter time frame, using substrates that are more similar to the natural or preferred substrates of the enzyme, and provide far more details about the time course of the reaction. This allows us to test many more samples with greater sensitivity so that a more stringent statistical analysis can be performed. Using these specific activity assays with C16-, C14-, C12-, C10- and C8-CoA substrates, we were able to probe two of the residues reported by Petronikolou and Nair, in addition to three additional residues lying a further distance away from the catalytic H136 residue (Figures 4 and 5). Modifications to Q27, T132 and A148 all resulted in improved activity with C10-CoAs but showed minimal changes with C12-CoAs. These modifications also resulted in lower specific activity for C16-CoA, to varying degrees. A144 also altered the specific activities with these various CoAs, without resulting in substantial losses in overall activity. The substitution of G25 with a valine, as described previously, resulted in an enzyme with very minimal activity toward C16-, C14- and C12-

CoAs, but retained most of its activity with C10- and C8-CoA. Taken together, these results provide additional and more precise evidence to support the notion that G25, Q27, T132, A144 and A148 constitute residues lining the binding pocket for fatty acyl-CoAs and should help direct future studies to fully probe additional aspects of this binding pocket.

Conclusions

The results presented here catalog a large number of residues lining the likely binding pocket of fatty and small alcohols for the wax ester synthase and provide further evidence to support this region of the enzyme as serving that role. The binding pocket for the fatty acyl-CoA required a different approach to provide sufficient evidence to begin to define this more elusive pocket. While the structural studies and selection of residues G25 and A144 mutants provided preliminary evidence of this binding pocket, the results presented here provide a more comprehensive analysis that supports this region and the residues A148, Q27 and T132 as also altering acyl-CoA substrate selectivity without resulting in the substantial losses in activity for the natural substrates of the wax ester synthase that were seen with the G25V modification. The approach taken here to characterize fatty acyl-CoAs, provides a detailed picture with statistical analysis that should have further application in exploring further details related to this second binding pocket. Combined, these results along with the structural picture of this enzyme, provide the first molecular scale blueprint for engineering this important class of enzymes, and should serve as a valuable tool to other laboratories in the future.

Experimental procedures

Strains and reagents

All chemicals were purchased through Fisher Scientific (Pittsburgh, PA) or Sigma Aldrich (St. Louis, MO) unless otherwise indicated. *Escherichia coli* JM109 and TB1 strains were obtained from New England Biolabs (Ipswich, MA).

Genetic constructs and mutations

All mutant enzymes described in this report were generated by performing site-specific mutagenesis using the QuikChange Site-Directed Mutagenesis system (Agilent Technologies, Santa Clara, CA). The parent plasmid used for these studies has been described previously[25] and contains the gene encoding a wax ester synthase from *Marinobacter aquaeolei* VT8 (Ma1, Maqu_0168 or NCBI reference sequence WP_011783747.1). This gene is fused to the gene for maltose binding protein on the N-terminus and also contains a polyhistidine tag on the C-terminus. Plasmid manipulations were first introduced into and isolated from *E. coli* JM109, then transferred to *E. coli* TB1 for expression. Once expressed in *E. coli* TB1, the resulting proteins were purified by successive steps of affinity chromatography. A list of plasmids used to construct mutants tested in this study is provided in Table 1. Primers used to construct each mutant are provided in Table 2.

Protein purification and quantification.

All proteins were expressed and purified as described before[25]. Briefly, transformed *E. coli* TB1 cells were grown in 10 mL of Lysogeny Broth (LB) media supplemented with 100 mg/L ampicillin. An overnight culture was used to inoculate 1 L of fresh LB media and grown on a shaker flask at 30°C. Protein expression was induced approximately 4 h after inoculation with 50 mg/L IPTG, when the culture reached an OD600 ~ 0.6. After 3 h of induction, cells were harvested by centrifugation at 7000 g and stored at -20°C until purification. Frozen cells were thawed, then resuspended in lysis buffer (50 mM potassium phosphate pH 7.2, 300 mM NaCl, 30 mg of CHAPS, 10 mg of lysozyme and 1 mg of DNase) and lysed by sonication. Cell debris was separated by centrifugation at 12000 g and the supernatant, containing the protein of interest, was loaded onto a metal affinity column charged with nickel (Chelating Sepharose Fast Flow; GE Healthcare, Piscataway, NJ) for purification using the polyhistidine tag attached to the C-terminus of the protein. The column was washed with buffer (50 mM potassium phosphate pH 7.2, 300 mM NaCl) supplemented with 50 mM imidazole and then eluted in 500 mM imidazole in the same buffer adjusted to pH 7.8. The eluted fractions were tested for protein using Coomassie Plus protein assay reagent (Thermo Scientific, Rockford, IL). As a second purification step, the protein-containing fractions eluted from the nickel column were immediately loaded onto an amylose resin column (New England Biolabs, Ipswich, MA) for purification using the maltose binding moiety fused to the N-terminus of Ma1. The column was washed with buffer (50 mM potassium phosphate pH 7.2, 300 mM NaCl) and the purified proteins were eluted with buffer supplemented with 10 mM maltose. Protein-containing fractions were pooled, and protein concentration was estimated using the absorbance at 280 nm and the extinction coefficient of the MBP-Ma1-polyhistidine given by the ExPASy ProtParam tool[42,43]. Purified protein was flash frozen as small pellets and stored in liquid nitrogen.

Fatty and small alcohol substrate preference assays.

Alcohol preference was surveyed using a competitive assay where a mixture of alcohol substrates and palmitoyl-CoA (the limiting reagent) were incubated with each Ma1 mutant until the palmitoyl-CoA was completely consumed. Fatty alcohol preference was tested with the following reagents, mixed in the order they appear: 1.5 mL of 50 mM potassium phosphate and 300 mM NaCl buffer at pH 7.2, 10 μL of DTNB (18 mg/mL in DMSO), 40 μL of alcohol mix (hexanol, octanol, nonanol, decanol, undecanol, dodecanol, tridecanol, tetradecanol, hexadecanol, heptadecanol, and octadecanol, 40 nmol each in DMSO), 50 μL of purified enzyme, and 200 μL of palmitoyl-CoA (200 nmol in H_2O). After the addition of each component the reaction was thoroughly mixed. Reaction completion was assessed by observing the color formation of TNB^{-2} (based on the assay described below). When finished, the mixture was flash-frozen in liquid nitrogen and freeze dried. The wax esters produced by the reaction were extracted in 1 mL of organic solvent (1:1 mixture of hexane and methylene chloride) and analyzed using GC/FID as described before[25,44]. Percentages of each wax ester product were quantified by determining the area of each peak and dividing each peak by the total area for all wax ester peaks. The small alcohols assay was performed as described above except for the addition of 100 μL of a different alcohol mixture (170 μmol ethanol, 3.8 μmol isoamyl alcohol, 160 nmol n-hexanol, and 4.6 μmol 2-phenyl ethanol in DMSO) and 100 μL of palmitoyl-CoA (100 nmol in H_2O). The relative quantities of small alcohols were selected so that the wild type Ma1 would produce similar amounts of esters with all substrates; therefore, a change in relative amounts of resulting waxes could be better appreciated[45].

Fatty acyl-CoA substrate preference assays.

Acyl-CoA preference was studied with a kinetic approach. The following reagents were mixed in order in a 1-cm pathlength quartz cuvette: 1 mL of 50 mM potassium phosphate, 300 mM NaCl buffer at pH 7.2; 5 μL of DTNB (18 mg/mL in DMSO); 50 μL of dodecanol (0.05 mg/mL in DMSO) and 50 μL of enzyme (varied concentration to optimize reaction in each case). The reaction mixture was thoroughly mixed using a pipette after the addition of each component. The absorbance at 412 nm (yellow color) was followed on a spectrophotometer (Model Cary 50 Bio; Varian, Palo Alto, CA). After a stable baseline was confirmed, the reaction was initiated by adding 20 μL of the pertinent acyl-CoA (1 mM in H_2O) and mixing thoroughly. The acyl-CoAs selected for this analysis included palmitoyl (C16), myristoyl (C14), lauroyl (C12), decanoyl (C10) and octanoyl (C8), all fully saturated. The reaction mixture contained approximately 13 nmol of dodecanol and 20 nmol of acyl-CoA at the start of the reaction. Data was obtained for 5 minutes, when the reaction had usually plateaued. Quantities of enzyme selected for these experiments were determined in advanced by confirming reaction completion with palmitoyl-CoA in approximately 3 minutes. Analysis with each acyl-CoA were performed in triplicate, assuring the protein was freshly thawed and diluted to the selected concentration right before testing each substrate. Initial reaction rates were calculated in Excel (Microsoft, Redmond, WA) based on the slope of the linear regression fitted to the initial change in absorbance and the extinction coefficient of the TNB^{-2} dianion at 412 nm of 14150 M^{-1} cm^{-1}. All rates of specific activity reported are the mean and standard deviation of at least three replicate assays.

Activity with diglyceride

The specific activity using 1,3-linoleoyl-rac-glycerol and palmitoyl-CoA as substrates was surveyed using a kinetic assay as described above. A 9 mM stock solution of 1,3-dilinoleoyl-rac-glycerol was prepared in DMSO. Assays substituted 50 μL of the stock diglyceride in place of the dodecanol solution. Concentrations greater than this resulted in significant precipitation that interfered with the assay. Reactions were initiated by adding 20 μL of palmitoyl-CoA (1 mM in H_2O)

Molecular visualization and image production.

Protein images were created using the protein data bank file for the Ma1 wild-type enzyme (accession number 6CHJ[28]) and the Biovia Discovery Studio 2017R2 software (Dassault Systèmes, Vélizy-Villacoublay, France). Assay plots and final images were created using RStudio (RStudio Team (2021), PBC, Boston, MA URL http://www.rstudio.com/) and Affinity Designer version 1.9/1.10 (Serif, Nottingham, UK).

Table 1. Plasmids used to construct mutant Ma1 wax ester synthases

Plasmid	DNA Primers used	Source or reference
pWEMBa1	See reference	
pWEMBa1-F11W	BBP3041 and BBP3042	This study
pWEMBa1-L14V	BBP2139 and BBP2140	This study
pWEMBa1-M21L	BBP3039 and BBP3040	This study
pWEMBa1-G25V	BBP3045 and BBP3046	This study
pWEMBa1-Q27F	BBP2978 and BBP2979	This study
pWEMBa1-T132F	BBP3055 and BBP3056	This study
pWEMBa1-A144I	BBP2972 and BBP2973	This study
pWEMBa1-A144F	BBP3047 and BBP3048	This study
pWEMBa1-F292M	BBP3049 and BBP3050	This study
pWEMBa1-L356A	See reference	
pWEMBa1-L356F	See reference	
pWEMBa1-L356V	See reference	
pWEMBa1-A357V	BBP1601 and BBP1602	This study
pWEMBa1-A359I	BBP1738 and BBP1739	This study
pWEMBa1-A360F	See reference	
pWEMBa1-A360I	See reference	
pWEMBa1-A360V	See reference	
pWEMBa1-L363F	BBP2995 and BBP2995	This study
pWEMBa1-L364F	BBP2997 and BBP2998	This study
pWEMBa1-V377L	BBP3053 and BBP3054	This study
pWEMBa1-I403A	BBP2232 and BBP2233	This study
pWEMBa1-I403F	BBP2154 and BBP2155	This study
pWEMBa1-I403L	BBP1891 and BBP1892	This study
pWEMBa1-M405F	See reference	
pWEMBa1-M405W	See reference	
pWEMBa1-L408F	BBP2993 and BBP2994	This study
pWEMBa1-N411V	BBP1595 and BBP1596	This study

Table 2. DNA primers used in this study.

Primer Designation	Primer DNA sequence
BBP1595	5’GACAGACTGGCCCTG**GTC**ATGACACTGACCAGC3’
BBP1596	5’GCTGGTCAGTGTCATGACCAGGGCCAGTCTGTC3’
BBP1601	5’CACGGCCCTGACCCTG**GTG**CCGGCCGCCTTCCAC3’
BBP1602	5’GTGGAAGGCGGCCGGCACCAGGGTCAGGGCCGTG3’
BBP1738	5’CCTGACCCTGGCGCCG**ATC**GCCTTCCACCTGCTG3’
BBP1739	5’CAGCAGGTGGAAGGCGATCGGCGCCAGGGTCAGG3’
BBP1891	5’ATGTATCCGGTGTCT**CTG**GATATGGACAGACTG3’
BBP1892	5’CAGTCTGTCCATATCCAGAGACACCGGATACAT3’
BBP2139	5’CAGCTCTTTCTCTGG**GTG**GAAAAACGCCAGCAGC3’
BBP2140	5’GCTGCTGGCGTTTTTCCACCCAGAGAAAGAGCTG3’
BBP2154	5’ATGTATCCGGTGTCT**TTC**GATATGGACAGACTG3’
BBP2155	5’CAGTCTGTCCATATCGAAAGACACCGGATACAT3’
BBP2232	5’GCATGTATCCGGTGTCT**GCC**GATATGGACAGACTGGC3’
BBP2233	5’GCCAGTCTGTCCATATCGGCAGACACCGGATACATGC3’
BBP2972	5’GGTGGACGGTGTCTCG **ATC** ATGCGCATGGCCACC3’
BBP2973	5’GGTGGCCATGCGCATGATCGAGACACCGTCCACC3’
BBP2978	5’CATGTGGGCGGCCTC**TTC**CTGTTTTCCTTCCC3’
BBP2979	5’GGGAAGGAAAACAGGAAGAGGCCGCCCACATG3’
BBP2993	5’CTATCGATATGGACAGA**TTC**GCCCTGAACATGACACTG3’
BBP2994	5’CAGTGTCATGTTCAGGGCGAATCTGTCCATATCGATAG3’
BBP2995	5’GCCGGCCGCCTTCCAC**TTC**CTGACCGGGCTGGCGCCC3’
BBP2996	5’GGGCGCCAGCCCGGTCAGGAAGTGGAAGGCGGCCGGC3’
BBP2997	5’GCCGGCCGCCTTCCACCTG**TTC**ACCGGGCTGGCGCCC3’
BBP2998	5’GGGCGCCAGCCCGGTGAACAGGTGGAAGGCGGCCGGC3’
BBP3039	5’GAAAAACGCCAGCAGCCC**CTG**CATGTGGGCGGCCTC3’
BBP3040	5’GAGGCCGCCCACATGCAGGGGCTGCTGGCGTTTTTC3’
BBP3041	5’CCCACTGACCAGCTC**TGG**CTCTGGCTGGAAAAACGC3’
BBP3042	5’GCGTTTTTCCAGCCAGAGCCAGAGCTGGTCAGTGGG3’
BBP3045	5’CCCATGCATGTGGGC**GTC**CTCCAGCTGTTTTCC3’
BBP3046	5’GGAAAACAGCTGGAGGACGCCCACATGCATGGG3’
BBP3047	5’GGTGGACGGTGTCTCG**TTC**ATGCGCATGGCCACC3’
BBP3048	5’GGTGGCCATGCGCATGAACGAGACACCGTCCACC3’
BBP3049	5’CCACTGGTGGCC**ATG**GTGCCGGTGTCACTACGC3’
BBP3050	5’GCGTAGTGACACCGGCACCATGGCCACCAGTGG3’
BBP3053	5’GGCAGACGTTCAATGTG**CTG**ATTTCCAATGTCCCC3’

BBP3054	5’GGGGACATTGGAAATCAGCACATTGAACGTCTGCC3’
BBP3055	5’GGCAGTTTGCGCTCTAC**TTC**AAGGTTCACCATTCCC3’
BBP3056	5’GGGAATGGTGAACCTTGAAGTAGAGCGCAAACTGCC3’

Accession Codes

The wax ester synthase used in these studies is Ma1 from *Marinobacter aquaeolei* VT8 (NCBI reference sequence WP_011783747.1 or KEGG gene ID Maqu_0168 or protein databank identifier 6CHJ).

Chapter 2: Genomic analysis of *Scenedesmus glucoliberatum* PABB004: an unconventional sugar-secreting green alga

Natalia Calixto Mancipe[1], Evelyn M. McLaughlin and Brett M. Barney[1, 2, *]

[1] Department of Bioproducts and Biosystems Engineering, University of Minnesota, St. Paul, MN 55108, USA

[2] BioTechnology Institute, University of Minnesota, St. Paul, MN 55108, USA

This chapter is an almost verbatim reproduction of the published manuscript with the same title [46].

Synopsis

In this work we present a new microalgal strain, *Scenedesmus glucoliberatum* PABB004, which is an interesting organism for fermentative processes due to its advantageous sugar-secretion phenotype and whose evolutionary history, inferred from its genomic traits, expands our current understanding of algal mutualistic relationships involving photosynthate exchanges. *S. glucoliberatum* PABB004 secreted ready-to-use fermentable sugars (glucose and maltose) directly to the extracellular media up to 2.7 g/L of free glucose and 1.2 g/L of maltose in batch cultures at pH 6.2. The range of pH values in which secretion takes place at elevated sugar concentrations ranged from 6.2 to 7.8 pH units. We sequenced, assembled and annotated the draft genome and transcriptome for the newly reported strain. The predicted proteome were compared with other green algae that show different sugar-secretion phenotypes aiming to help uncover their common features for maltose secretion and those unique to *S. glucoliberatum* PABB004.

Introduction

Biomass fermentation processes represent an alternative avenue to produce fuels and commodity chemicals with reduced environmental impacts. Currently, land crops (corn, sugarcane and sugar beets) are the main source of sugar-rich biomass for fermentations, but they require considerable cultivation time and post-harvest processing to extract and yield the monomeric sugars needed for downstream applications[23,47]. Thus, feedstock costs represent a large portion of the total process costs, imposing minimum molar yields for commercial viability. The use of lignocellulosic biomass, electro-biotechnology approaches and C1 feedstocks (CO, CO_2 and CH_4) have been proposed as potential solutions[48]. Microalgae stand out for their ability to use C1 feedstocks, having a faster growth rate than land plants and the availability of existing infrastructure for a variety of industrial applications[49]. Here we present a microalgal strain that secretes simple, ready to use sugars for fermentation processes into the extracellular environment.

Similar to what is found for land plants, sugar extraction from algae is generally based on biomass pretreatment[50] followed by saccharification of cellulose and starch, but algae have simpler cell walls that facilitate this process. The potential use of algae as a source of hydrocarbons for bioprocessing has been reviewed[51]. Additionally, cultivating natural and modified sucrose-accumulating algae and cyanobacteria to increase sugar productivity has also been reported[52]. Nonetheless, using natural sugar-secreting organisms as a source of readily recoverable sugars that do not require complex biomass processing has yet to be thoroughly explored.

Several species of sugar-secreting algae have been found forming symbiotic relationships with *Paramecium bursaria* and *Hydra viridis,* mainly from the *Chlorella* and *Micractinium* genera[53,54], or in cnidarian-dinoflagellates associations found in coral reefs[55]. When in the symbiotic state, algae cells are maintained within vacuoles inside the host species, where the release of simple sugars (predominantly maltose and lower amounts of glucose or glucose-6-phosphate) occurs. It has been suggested that such release is regulated by decreasing the pH in the vacuolar compartments, and this relationship between acidic pH and sugar secretion has been observed in vitro[56–58]. Particularly, *M. conductrix* has been reported to release a maximum of 5 g/L (14 mM) maltose at pH 5.7 in batch culture. Other studies describe the secretion of maltose from *Chlorella* spp. such as *C. sp*. 3N813A which at pH 5 reached concentrations in the low μM range in the media. Unfortunately, the molecular mechanisms that allow such secretion systems remain unknown, partly due to the minimal genomic data available for algal species[59] and a lack of transcriptomic profiles to correlate expression data with sugar-secreting phenotypes.

In this work, we describe a new sugar-secreting green microalga with differential characteristics compared to many previously reported sugar-accumulating *Chlorella* and *Micractinium* species. Linking phenotypic and genomic data, this work contributes new information filling the gap between the observation of desired phenotypes and the understanding of their molecular basis as well as increasing the general pool of algal strains with desired capabilities for biotechnological applications.

Results

Glucose and maltose secretion by *S. glucoliberatum* PABB004

To examine extracellular sugar production, *Scenedesmus glucoliberatum* PABB004 was cultured for 45 days at various pH levels ranging from 6.2 to 7.8 and the supernatant tested for the presence of sugars using an enzymatic approach. As shown in Figure 6, the concentration of free extracellular maltose reaches the mM level, which is significantly higher than what was previously described for *Chlorella* sp. 3N813A and in the same range as *M. conductrix*[56–58].

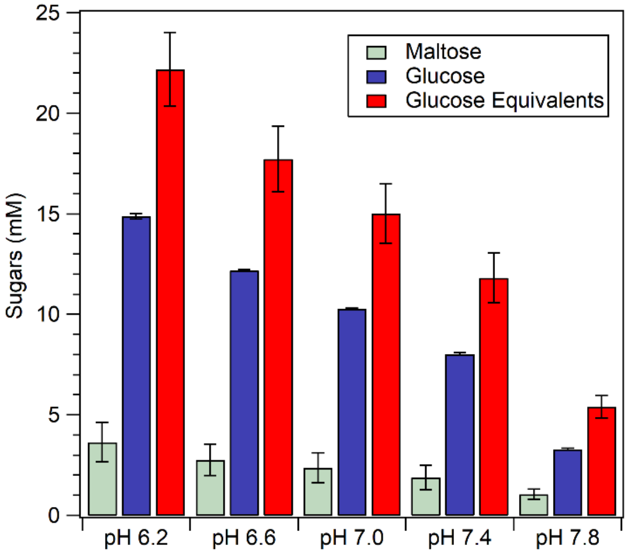


Figure 6. Effect of pH on the extracellular sugar accumulation by S. glucoliberatum PABB004

S. glucoliberatum PABB004 was cultured at various pH levels for 45 days under standard laboratory conditions.

Interestingly, extracellular glucose levels are consistently 3 times higher than maltose levels, reaching a maximum of 15 mM glucose versus 3.5 mM maltose at pH 6.2. As opposed to the other green algae, glucose is the main secreted photosynthate.

The concentrations of both glucose and maltose in the media depend on acidity, with higher sugar levels arising at lower pH. This relationship between acidity and sugar concentration appears to be more intense for the case of glucose, i.e., the reduction of extracellular glucose with increasing pH occurs faster than for the extracellular maltose at the same pH levels, suggesting a different response to pH changes in the mechanisms of glucose and maltose secretion.

S. glucoliberatum PABB004 cell morphology

Vegetative cells of *S. glucoliberatum* PABB004 (Figure 7) are ellipsoidal, approximately 4 µm long and 2 µm wide. Unlike other *Scenedesmus* spp., it does not present an elongated or sickle-like morphology nor form colonies in liquid culture. Instead, it lives predominantly as solitary cells. Outer structures such as flotation spines (common in *Scenedesmus* spp.) or flagella (common in *Chlamydomonas* spp.) were not observed.

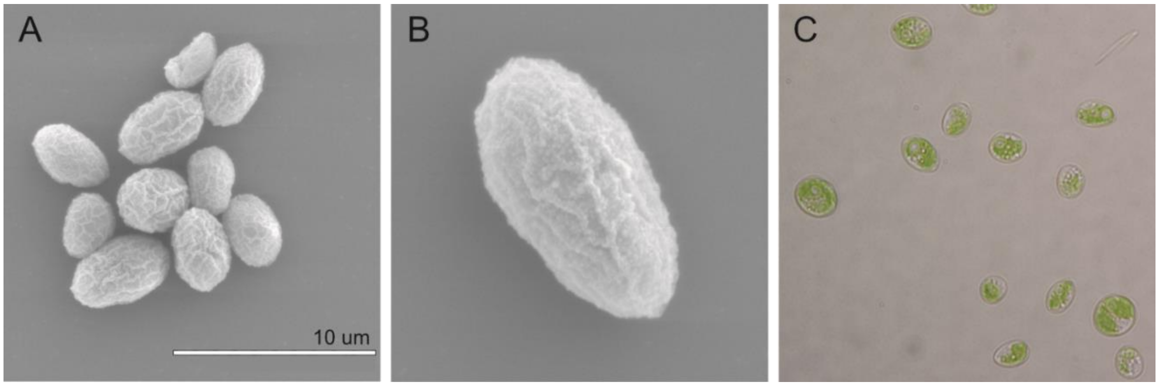


Figure 7. Scenedesmus glucoliberatum PABB004 cell morphology

Cell suspensions were imaged using SEM (A and B) and light microscopy (C). The predominant distribution showed solitary cells and lack of outer structures such as flotation spines or flagella.

Genome sequencing and assembly

S. glucoliberatum PABB004 *de novo* draft genome assembly was performed using 854,693 PacBio reads with an average length of 8.2 kb and an average sequence coverage of 183x. The complete draft genome (39.97 Mb) was assembled in 80 contigs. From those, 77 belong to nuclear DNA, two to the chloroplast chromosome and the complete mitochondrion chromosome (Table 3). Half of the assembled genome is contained in 10 contigs (L50), all of those longer than 1.3 Mb (N50). From the 1,519 Benchmarking Universal Single-Copy Orthologs (BUSCOs)[60] in the Chlorophyta dataset used to assess the assembly completeness, 93.7% were found as single copy, 1.3% were duplicated, 1.5% were fragmented and 3.5% missing. These percentages show that the assembly coverage and contiguity are of good quality and good representatives of the genetic background for the following comparisons.

Genomic features of *S. glucoliberatum* PABB004

The *S. glucoliberatum* PABB004 nuclear DNA contains two contigs with telomeric repeats on both ends, indicating full chromosomes (contigs 1 and 6) and 34 others with telomeric repeats on one end only. This suggests the existence of at least 19 nuclear chromosomes (Table 4). Assembly results did not show any evidence of polyploidy. Total nuclear DNA length is 39.8 Mb, considerably smaller than *Monoraphidium neglectum* (69.5 Mb, accession Uniprot: UP000054498, NCBI: GCA_000611645.1), which was the closest algae in taxonomic terms found with its plastid, mitochondrial and nuclear genomes annotated. *S. glucoliberatum* PABB004 has the third smallest genome and the highest GC content (78%, Table 4) within the Sphaeropleales genomes in the current NCBI database.

Table 3 S. glucoliberatum PABB004 genome assembly and annotation statistics

Assembly	
Total sequence length	39.97 Mb
Sequencing depth	183x
Number of contigs	80
Contig length range	[5.93 kb - 4,85 Mb]
Contig average length	499.66 kb
Contig median length	206.87 kb
N50	1.29 Mb
L50	10
PacBio reads used	835,035 (97.7%)
Illumina reads mapped	44,019,865 (98.4%)
BUSCO genome scores	C = 95.0% [S = 93.7%, D = 1.3%], F = 1.5%, M = 3.5%
GC content	78.38%
Annotation	
CDS count	6947
Models with homology support (SwissProt database)	3062 (44%)
Models with homology support (KEGG database)	3105 (45%)

The gene prediction approach resulted in the annotation of 6947 CDSs. Prediction quality was evaluated with BUSCO on protein mode and the chlorophyta_odb10 dataset, which contains 1519 orthologous genes. Results show a prediction outcome with 91.1% complete genes (88.2% found as single copy and 2.9% duplicated), 0.3% genes fragmented and 8.6% missing, which is comparable to the quality of the proteomes reported in the SwissProt database.

Table 4. Genomic features of S. glucoliberatum PABB004 and selected algae

	S. glucoliberatum	*M. neglectum †*	*C. reinhardtii ‡*	*C. variabilis §*
Nuclear DNA				
Total sequence length (Mb)	39.8	69.5	111.1	46.2
Chromosomes	>= 19	-	17	12
GC content (%)	78.6	64.2	61.9	65.5
CDS count	6855	16717	17700	9780
Plastid DNA				
Total sequence length (kb)	191.3	140	203.8	124.7
GC content (%)	32.1	32.4	34.5	33.9
CDS count	76	4	69	114
Mitochondrion chromosome				
Total sequence length (kb)	26.6	90	15.8	78.5
GC content (%)	36.8	45.5	45.2	28.3
CDS count	15	17	8	62

Genome assembly accessions † GCA_000611645.1, ‡ ABCN00000000.2 and §GCA_000147415.1

There were 6872 CDSs in the nuclear contigs, accounting for a gene density of 172.7 genes per Mb, which is slightly higher than *C. reinhardtii* (159 genes per Mb)[61] and lower than *M. neglectum* and *C. variabilis* (with 240 and 212 genes per Mb, respectively)[62]. It should be noted that this value for the *S. glucoliberatum* PABB004 genome only represents protein coding sequences, i.e., non-coding sequences, tRNA and rRNA genes were not included in the gene counts.

The chloroplast chromosome could not be completely closed with our assembly approach. Instead, its 191.3 kb are represented in two contigs of 154.4 kb and 36.9 kb. The shortest contig is highly similar to the longest (99.95% nucleotide identity on the reverse complementary sequence, 100% coverage) and it is likely the inverted repeat segment commonly found in plastid genomes[63,64]. The plastid GC content is similar to the reference green alga in Table 4 and its CDS density, 0.31 genes per kb, is similar to *C. reinhardtii* (0.34 genes per kb).

The mitochondrial genome was completely assembled in one circular contig with no gaps. The chromosome length was 26.6 kb with a GC content of 36.8%, 15 CDSs and a gene density of 0.56 genes / kb. These values are within the length and GC content range reported from other green algal genomes (Table 4). Moreover, the closely related *Tetradesmus obliquus* mitochondrion has 43 kb, 36.2% GC, 20 CDSs and 0.47 genes / kb (NCBI bioproject accession PRJNA11896) which shows that the *S. glucoliberatum* PABB004 mitochondrion is quite similar in GC content but slightly more compact, i.e. with only 62% of the length it contains 75% of the total CDSs in *T. obliquus*. Additionally, the relationship between these two species is supported by the use of the same mitochondrial genetic code, which differs from the standard code in the codons TCA being translated to stop instead of Ser and TAG to Leu instead of stop.

Taxonomic analysis of *S. glucoliberatum* PABB004

The taxonomic classification of *S. glucoliberatum* PABB004 was performed through a phylogenetic approach using the conserved SSU rRNA and variable ITS regions and a phylogenomic comparison based on the average nucleotide identity between the *S. glucoliberatum* PABB004 transcriptome and related green algae.

An initial BLASTn search of the *S. glucoliberatum* PABB004 genomic region containing the SSU rRNA and its ITS 1 and 2 revealed that it belongs to the Chlorophyceae clade within the green algae (Data S1). This result was unexpected since *P. bursaria* endophytes are usually found within the *Chlorella* and *Micractinium* genera, which belong to the Trebouxiophyceae clade[53,62]. Hence, a broader analysis including a wide range of Chlorophyceae organisms was performed. The resulting phylogenetic tree (Figure 8) supports the grouping of *S. glucoliberatum* PABB004 within the Chlorophyceae as part of the Sphaeropleales clade.

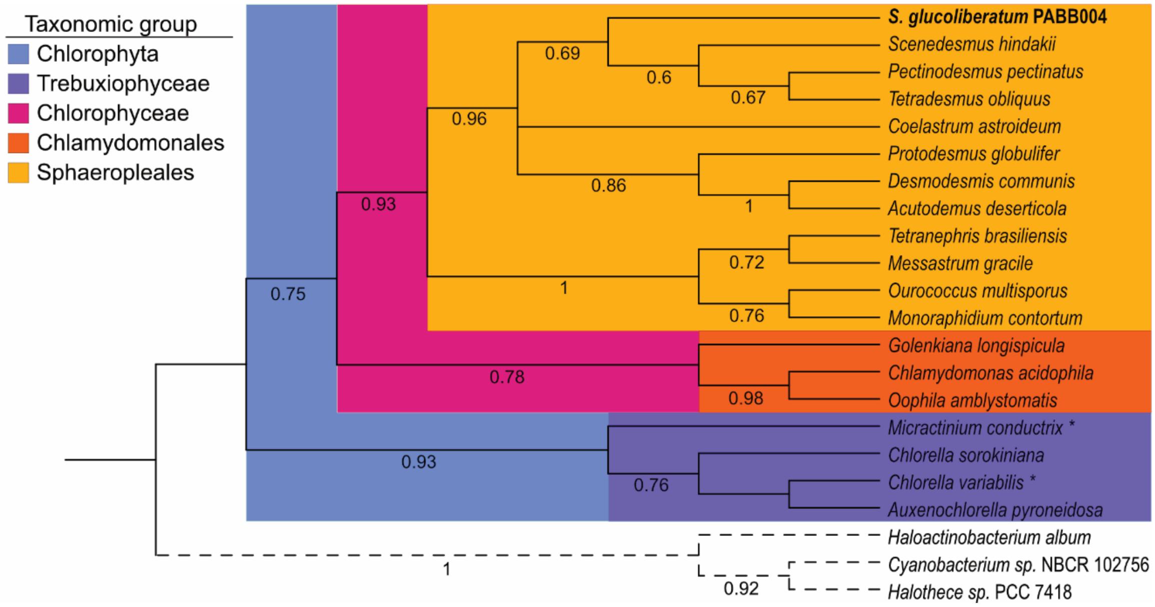


Figure 8. Scenedesmus glucoliberatum PABB004 phylogenetic classification

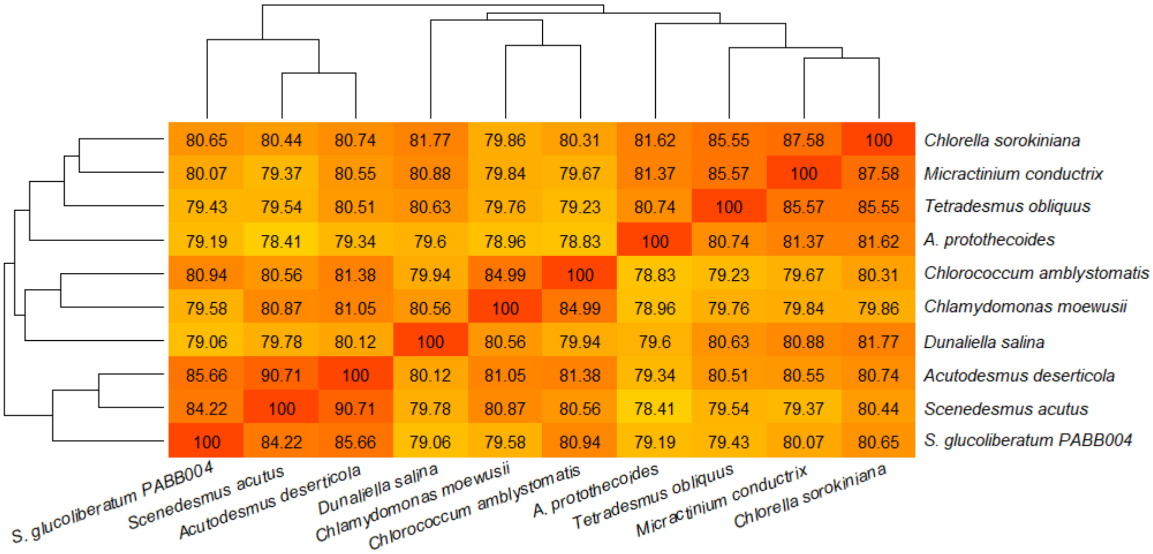
Evolutionary history of S. glucoliberatum PABB004 was inferred using the Maximum Likelihood method and Kimura 2-parameter model applied to calculate the evolutionary distances among 22 SSU rRNA sequences. The phylogenetic tree was generated with a bootstrap consensus from 100 replicates. The frequency of replicate trees in which the associated taxa clustered together in the bootstrap test are shown next to the branches. Taxonomic clades are color coded and known *P. bursaria* endosymbionts are marked with an asterisk. Dashed lines correspond to the outlier sequences branch.

Furthermore, the bootstrap consensus shows that in 96 out of 100 replicates *S. glucoliberatum* PABB004 is placed with the Scenedesmaceae family. Using this approach, the exact genus and species of the isolate could not be determined.

The *S. glucoliberatum* PABB004 transcriptome was compared with transcriptomes from other green algae. All transcripts were clustered based on sequence similarity and the average nucleotide identity of orthologous sequences was calculated between pairs of organisms.

As shown in Figure 9, *S. glucoliberatum* PABB004 is grouped with two other Scenedesmaceae algae (*S. acutus* and *Acutodesmus deserticola*) confirming the results of the previous phylogenetic analysis while the other Chlorophyceae and Trebouxiophyceae were grouped in different clades. The only exception to this classification was *T. obliquus*, which was placed among the Trebouxiophyceae, probably due to the small size of the transcriptome that was found on the TSA database, which might be more prone to error due to a low coverage.

Since both methods support the placement of the isolate within the Scenedesmaceae family and its closest organisms belong to the *Scenedesmus* genus based on the SSU rRNA data we chose to name our isolate *Scenedesmus glucoliberatum* PABB004.



	S. glucoliberatum PABB004	Scenedesmus acutus	Acutodesmus deserticola	Dunaliella salina	Chlamydomonas moewusii	Chlorococcum amblystomatis	A. protothecoides	Tetradesmus obliquus	Micractinium conductrix	Chlorella sorokiniana
Chlorella sorokiniana	80.65	80.44	80.74	81.77	79.86	80.31	81.62	85.55	87.58	100
Micractinium conductrix	80.07	79.37	80.55	80.88	79.84	79.67	81.37	85.57	100	87.58
Tetradesmus obliquus	79.43	79.54	80.51	80.63	79.76	79.23	80.74	100	85.57	85.55
A. protothecoides	79.19	78.41	79.34	79.6	78.96	78.83	100	80.74	81.37	81.62
Chlorococcum amblystomatis	80.94	80.56	81.38	79.94	84.99	100	78.83	79.23	79.67	80.31
Chlamydomonas moewusii	79.58	80.87	81.05	80.56	100	84.99	78.96	79.76	79.84	79.86
Dunaliella salina	79.06	79.78	80.12	100	80.56	79.94	79.6	80.63	80.88	81.77
Acutodesmus deserticola	85.66	90.71	100	80.12	81.05	81.38	79.34	80.51	80.55	80.74
Scenedesmus acutus	84.22	100	90.71	79.78	80.87	80.56	78.41	79.54	79.37	80.44
S. glucoliberatum PABB004	100	84.22	85.66	79.06	79.58	80.94	79.19	79.43	80.07	80.65

Figure 9. S. glucoliberatum PABB004 phylogenomic relationships

Average Nucleotide Identity (ANI) values were calculated based on BLAST identity scores generated by pairwise comparisons between organisms. Sequences were taken from 98 clusters of orthologous transcribed sequences among all taxa analyzed. Organisms were then classified based on their ANI scores. Note that S. glucoliberatum PABB004 clustered together with other Sphaeropleales (S. acutus and A. deserticola), other Chlorophyceae alga from different clades and Chlorellales were grouped in separate branches.

Proteomic comparisons among *S. glucoliberatum* PABB004 and selected green algae

Understanding the genetic mechanisms in green alga that allow the secretion of simple sugars to the extracellular environment could enable the manipulation of such pathways for their use in biotechnological applications. Therefore, the set of predicted CDSs and respective proteins contained in the *S. glucoliberatum* PABB004 genome was analyzed through a sequence similarity and clustering approach. The goal of this search was defining a subset of sequences that are present only in algae with sugar secreting phenotypes (*M. conductrix* and *S. glucoliberatum* PABB004) and absent from their non-sugar secreting controls. *C. reinhardtii* and *C. sorokiniana* were chosen as controls for *S. glucoliberatum*

PABB004 and *M. conductrix* due to their evolutionary relatedness and apparent lack of extracellular sugar secretion phenotypes.

The curated proteomes from *M. conductrix, C. sorokiniana* and *C. reinhardtii* were obtained from the Uniprot-Swissprot database, compared with *S. glucoliberatum* PABB004 and clustered based on sequence similarity (Table 5). From the 18829 proteins in the *C. reinhardtii* proteome, 12560 were classified in 7049 homologous groups ("clusters") and the remaining 6269 did not show homology relationships to other sequences ("singletons"). Similarly, the *C. sorokiniana* proteome with 9482 proteins was classified in 6710 clusters with 7527 proteins and 1955 singletons; the 9122 proteins of *M. conductrix* were organized in 6616 clusters with 7567 proteins and 1555 singletons and the 6947 proteins from *S. glucoliberatum* PABB004 were classified as 5354 proteins grouped in 4563 clusters and 1593 singletons.

As shown in Table 5 and Figure 10, most of the proteins analyzed belong to a core subset of 3098 clusters that are shared among all algae (labeled with the superscript "a" in Figure 10), which aligns with the expected results since all organisms belong to the Chlorophyta clade and hence must share a high percentage of their genomic resources. This core subset contains 68% of all *S. glucoliberatum* PABB004 clusters, 44% of *C. reinhardtii,* 46% of *C. sorokiniana* and 47% of *M. conductrix*. Proteins in this subset have basic functions that range from photosystem assembly to transcription and ribosomal activity, metabolism and transport processes.

Similarly, the overall level of homology among proteomes is expected to correlate with the evolutionary history of the algae analyzed. For example, the intersection between *M.*

conductrix and *C. sorokiniana* (*M. conductrix* ∩ *C. sorokiniana* in Table 5 and superscripts a, c, d and h in Figure 10) contains the highest number of shared clusters (6150), which represent 93% and 92% of each organism's total count. This result aligns with the expected outcome since both organisms are closely related and have similar sized proteomes.

Table 5. Proteome clustering results among selected green alga

Proteomes	**Total Proteins**	**Clustered**[†]	**Singletons**[‡]
C. reinhardtii (UP000006906)	18829	12560 (7049)	6269
C. sorokiniana (UP000239899)	9482	7527 (6710)	1955
M. conductrix (UP000239649)	9122	7567 (6616)	1555
S. glucoliberatum PABB004 (this study)	6947	5354 (4563)	1593
Selected intersection sets			
M. conductrix ∩ *C. sorokiniana*	-	(6150)	-
S. glucoliberatum PABB004 ∩ *C. reinhardtii*	-	(4258)	-
Core	14411	(3098)	-
S. glucoliberatum PABB004 ∩ *M. conductrix* only	52	(23)	-
S. glucoliberatum PABB004 only	269	(97)	-

[†] Values in parenthesis represent the number of protein clusters

[‡] Protein sequences not classified within any homology group

For *S. glucoliberatum* PABB004, 93% of its clusters are shared with *C. reindhardtii* (*S. glucoliberatum* PABB004 ∩ *C. reinhardtii* in Table 5 and superscripts a, e, f and g in Figure 10). Without including the core set shared by all algae, the overlap between *S. glucoliberatum* PABB004 and *C. reindhardtii* accounts for 26% of all its clusters, which is more than three times greater than its overlap with *C. sorokiniana* or *M. conductrix* (about 8% each). This clearly supports the classification of *S. glucoliberatum* PABB004 in the Chlorophyceae clade.

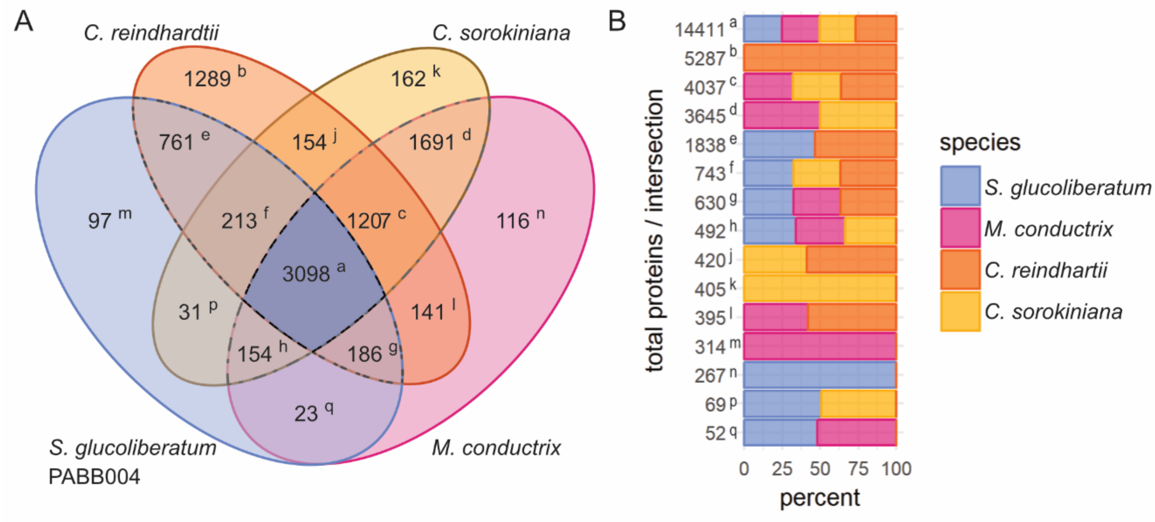


Figure 10. Proteomic comparisons among sugar-secreting and non-secreting algae

The Uniprot-Swissprot proteomes of *C. sorokiniana*, *M. conductrix* and *C. reindhardtii* were compared to *S. glucoliberatum* PABB004 predicted proteome using OrthoVenn2 with a blastp cutoff of 1e-10. The proteins were clustered in homologous groups based on sequence similarity. A: Venn diagram depicting the number of protein clusters shared by or unique to the four green algae. The number of clusters is shown instead of number of individual proteins due to the different copy number of some homologous sequences among species. B: Protein count in each intersection in A and the percent contribution of each species. As a reference, each intersection is labeled with a letter superscript in A and B.

Moreover, if the maltose-secreting phenotype is related to the existence of unique genes that enable the process, then such genes are likely contained in the subset common to *S. glucoliberatum* PABB004 and *M. conductrix* and absent from the other two algae ((*S. glucoliberatum* PABB004 ∩ *M. conductrix)* ⊄ (*C. sorokiniana* ∪ *C. reinhardtii*), superscript q). Since both sugar-secreting algae share a large percentage of their proteome with non-sugar-secretors (> 90%), a comparison between all proteomes, followed by the selection of clusters that exist exclusively in the sugar-secreting species results in a reduced subset of proteins. This subset (*S. glucoliberatum* PABB004 ∩ *M. conductrix* only) contains 23 clusters with 27 proteins from *M. conductrix* and 25 from *S. glucoliberatum*

PABB004. The 25 *S. glucoliberatum* PABB004 proteins in this subset were further analyzed through BLASTp and InterProScan searches, looking for evidence of a unique maltose transporter which would be a key component for the secretion pathway hypothesized in previous studies[58]. Our preliminary results fail to identify such a transporter, nonetheless, a deeper examination of the shared sequences between both algae and their transcription profiles would be necessary to further understand their potential role in the maltose secretion process. Following the same logic, we generated a pool of protein sequences unique to *S. glucoliberatum* PABB004 (superscript m in Figure 10) that might be involved in the unconventional sugar secretion characteristics of this alga, i.e. the secretion of glucose as the main photosynthate.

Metabolic traits of *S. glucoliberatum* PABB004 inferred from its genomic data

The metabolic characteristics of *S. glucoliberatum* PABB004 were examined through the Kyoto Encyclopedia of Genes and Genomes (KEGG) annotation and their Mapper tool. A total of 3105 proteins (45% of the proteome) were assigned a KEGG orthology accession, mapping to 380 pathways and were classified in 45 different BRITE functional hierarchies. It is worth mentioning that due to the role of some proteins in various pathways or cell processes, the total count of hits for all categories is much greater than the actual protein count.

KEGG found 92 orthologs for carbon metabolism pathways (ko01200) in *S. glucoliberatum* PABB004 including 26 for glycolysis/gluconeogenesis (ko00010), 21 for the TCA cycle (ko00020), 16 for the pentose phosphate pathway (ko00030), 22 in carbon fixation (ko00710), among others. All these basic pathways had a complete representation in *S. glucoliberatum* PABB004 predicted proteome. KEGG also found 23 orthologs for fatty acid metabolism (ko01212) which generate an uninterrupted pathway from Acetyl-CoA to Stearoyl-CoA, the pathway connecting longer-chain compounds has several missing genes. There were also 88 orthologs for biosynthesis of amino acids (ko01230), which completed an uninterrupted path to produce all amino acids except histidine, tryptophan, methionine, arginine and proline. This could be due to an incomplete annotation process whether at the structural or functional steps, which the authors acknowledge is a limiting factor for the full metabolic analysis of the organism.

Additionally, *C. reinhardtii* and *C. variabilis* were used as reference organisms to contrast the distribution of BRITE hierarchies. The most common hierarchies for *S. glucoliberatum* PABB004 are shown in Table 6. The three analyzed algae have a different number of proteins classified in each category, which is expected as each have different sized proteomes. Therefore, the comparison of ranks was favored over the raw protein count. Paired comparisons between the three algae were performed with a Wilcoxon-rank test (H_0: the pair of organisms compared have the same distribution of hits among BRITE categories, H_a: they have a different distribution of hits). Our results confirmed the alternative hypothesis for *S. glucoliberatum* PABB004 (p-values 2.9e-5 for *S. glucoliberatum* PABB004 vs. *C. reinhardtii* and 3.8e-5 for *S. glucoliberatum* PABB004 vs. *C. variabilis*), while there was no evidence of a difference between *C. reinhardtii* and *C. variabilis* (p-value 0.9).

This observation means that *S. glucoliberatum* PABB004 has a different relative content of proteins within its proteome when compared to the other two organisms. In detail, the hierarchies “Ribosome biogenesis” and Ubiquitin system” and “Messenger RNA biogenesis” seem to have a higher relative content in *S. glucoliberatum* PABB004 than in *C. variabilis*. Since the raw number of proteins in those categories among the three organisms is similar, we infer that *S. glucoliberatum* PABB004 seems to have a higher representation due to its reduced proteome size. This implies that proteins from other hierarchies have been lost through evolution or were never present in *S. glucoliberatum* PABB004.

Table 6. Selected BRITE functional hierarchies for S. glucoliberatum PABB004 proteins and other green algae†

Functional category	***S. glucoliberatum***	***C. reinhardtii***	***C. variabilis***
ko01000 Enzymes	1254 (I)	1903 (I)	1919 (I)
ko03036 Chromosome and associated proteins	260 (II)	628 (II)	407 (II)
ko04131 Membrane trafficking ‡	212 (III)	357 (IV)	344 (III)
ko03009 Ribosome biogenesis ‡ §	156 (IV)	189 (X)	202 (IX)
ko04121 Ubiquitin system ‡ §	139 (V)	177 (XII)	166 (XIII)
ko03041 Spliceosome	135 (VI)	285 (V)	299 (V)
ko03400 DNA repair and recombination proteins ‡	129 (VII)	198 (IX)	223 (VIII)
ko04147 Exosome	127 (VIII)	426 (III)	320 (IV)
ko03019 Messenger RNA biogenesis §	125 (IX)	202 (VIII)	186 (XII)
ko03029 Mitochondrial biogenesis ‡	121 (X)	179 (XI)	195 (X)
ko02000 Transporters	117 (XI)	247 (VI)	240 (VI)
ko03011 Ribosome	102 (XII)	243 (VII)	237 (VII)
ko01002 Peptidases and inhibitors ‡ §	100 (XIII)	129 (XIV)	128 (XIV or XV)
ko03016 Transfer RNA biogenesis	89 (XIV)	165 (XIII)	188 (XI)
ko03021 Transcription machinery ‡	74 (XV)	108 (XVII)	128 (XIV or XV)
ko03110 Chaperones and folding catalysts	72 (XVI)	114 (XV)	103 (XVI)
ko03032 DNA replication proteins ‡	66 (XVII)	78 (XIX)	84 (XVII or XVIII)
ko01009 Protein phosphatases and associated proteins	58 (XVIII)	97 (XVIII)	84 (XVII or XVIII)
ko04812 Cytoskeleton proteins §	56 (XIX)	109 (XVI)	54 (XX)
ko00194 Photosynthesis proteins ‡ §	48 (XX)	60 (XXI)	49 (XXI or XXII)
ko03051 Proteasome	35 (XXI or XXII)	49 (XXII)	49 (XXI or XXII)
ko01003 Glycosyltransferases	35 (XXI or XXII)	65 (XX)	76 (XIX)
ko02044 Secretion system	17 (XXIII)	27 (XXIII)	28 (XXIII)

† Number of hits per hierarchy, sequences may have hits in various categories. The ranked order of hierarchies for each alga is given in parenthesis.

‡ Hierarchy with higher rank in *S. glucoliberatum* PABB004 than in *C. reinhardtii*

§ Hierarchy with higher rank in *S. glucoliberatum* PABB004 than in *C. variabilis*

Discussion

In this work we present the microalga *Scenedesmus glucoliberatum* PABB004. The genome sequence of *S. glucoliberatum* PABB004 expands the limited number of green algae assembled genomes and offers a stark contrast to other sequenced strains of *Scenedesmus*. *S. glucoliberatum* PABB004 is of particular interest for fermentative processes due to its advantageous sugar-secretion phenotype. The evolutionary history, inferred from its genomic traits, will also expand our current understanding of algal mutualistic relationships involving photosynthate exchanges.

Sugar-secreting phenotype

S. glucoliberatum PABB004 was shown to secrete ready-to-use fermentable sugars (glucose and maltose) to the culture media, demonstrating an advantage over other feedstock sources that require biomass pretreatment and saccharification steps. The highest carbohydrate levels recorded in batch cultures were 2.7 g/L of free glucose and 1.2 g/L of maltose at pH 6.2 (Figure 6).

These glucose levels are higher than what has been reported previously for *Micractinium* and *Chlorella* spp. Furthermore, sugar concentration in the media also shows an inverse relationship with pH, but in the case of *S. glucoliberatum* PABB004, the pH range in which the secretion takes place at significant sugar concentrations is broader than that from species within these two reference genera. This isolate secreted considerable levels of sugars at all pH levels tested (pH 6.2, 6.6, 7.0, 7.4 and 7.8) while *M. conductrix* and *Chlorella* sp. 3N813A secreted negligible sugars at pH levels above 7.6 and 7.0, respectively. This trait represents a potential advantage in industrial applications over other sugar-secreting algae, since it offers a greater process flexibility and resilience. Considering these attributes, *S. glucoliberatum* PABB004 constitutes a promising strain to produce fermentable carbon sources for biotechnological applications and in co-culture schemes investigating nutrient exchange[65,66]. Based on its unique sugar secreting phenotype and phylogenetic history we propose to name this new strain as *Scenedesmus glucoliberatum*.

An unexpected phylogenetic lineage

The evolutionary history of *S. glucoliberatum* PABB004 was investigated using sequence similarity and likelihood-based inference approaches. First, we compared its SSU rRNA gene and ITS regions as phylogenetic markers with a broad range of green algae homologs (Figure 8). We found that the 5' segment of the SSU rRNA gene contains a highly variable region that does not align well with any sequence tested, but the rest of the sequence is highly similar to those from species from the Sphaeropleales and Scenedesmaceae clades. Moreover, the average nucleotide identity between expressed transcripts from various Chlorophyceae and Trebouxophiceae algae as well as our proteome comparisons support this strain designation (Figures 9 and 10).

Surprisingly, our results show that *S. glucoliberatum* PABB004 does not belong to the *Micractinium* or *Chlorella* genera, known for their close mutualistic relationships with *P. bursaria* involving maltose secretion and exchange. Instead, it is placed on the Chlorophyceae – Sphaeropleales – Scenedesmaceae branch of the tree of life. While the potential to maintain certain *Scenedesmus* strains in *P. bursaria* has been reported, to our knowledge, there are no reports of Scenedesmaceae algae isolated from *P. bursaria* and shown to release extracellular simple sugars. It is possible that prior reports have mistakenly classified prior *Scenedesmus* species as *Chlorella* based solely on visual inspections and morphology. Hence, *S. glucoliberatum* PABB004 presents an interesting opportunity to expand our understanding on these key aspects of algal physiology and plant evolution.

S. glucoliberatum PABB004 was isolated from ruptured cells of *P. bursaria* isolated from a small selection of natural environments surrounding Minneapolis. This approach was utilized to screen for additional wild strains of sugar-secreting algae, and yielded several strains of algae outside of the *Micractinium* and *Chlorella* genera with initial sugar-releasing phenotypes. This screen was by no means exhaustive and represents only a minimal effort directed at expanding the range and versatility of sugar release. It is thus surprising that a broader description of sugar-release has not been reported for this classical example of a mutually beneficial symbiotic endophyte relationship. While these simple screens were successful in uncovering new examples of the sugar-releasing phenotype, studies to determine if these strains can be successfully added back to algal-deficient *P. bursaria* to rescue the symbiotic relationship were beyond the scope of this study and generally outside of the current skillset of our laboratory. However, the success of this simple screen illustrates the potential of such studies to screen for new strains of sugar-releasing algae from natural environments, which could be much more successful than broad screens of random algal strains, as the *P. bursaria* host strain pre-selects for this phenotype, or simply digests that algal cell in the absence of sugar-release.

A general hallmark of the *Scenedesmus* genera is the multicellular colonial lifestyle where they often are found in coenobia of four or eight cells. While there are also examples of unicellular *Scenedesmus* strains, this morphological feature is often used to classify these algae following an initial visual inspection. Symbiotic sugar-releasing algae found in *P. bursaria* are generally housed in what is referred to as a perialgal vacuole, which may not be amenable to algae that employ colonial lifestyles.

S. glucoliberatum PABB004 genomic features and resources

Another interesting feature of this isolate is its compact genome with several differential characteristics compared to other green algae in its clade. *S. glucoliberatum* PABB004 has the third smallest genome and the highest GC content within the Sphaeropleales genomes currently published in the GenBank database. Its predicted proteome contains only about 7000 proteins, merely 1.5 times the protein count of many strains of *Escherichia coli*.

Such a small genome size and low protein count in the light of the proteome comparison performed here (Figure 10 and Table 6) suggests a possible reduction in the *S. glucoliberatum* PABB004 genome to contain mostly essential and highly conserved genes. In fact, 68% of its protein families are clustered within a core set shared by other green algae (*C. reinhardtii*, *C. sorokiniana* and *M. conductrix*), while this same set only contains on average 46% of the total clusters in the other species (Table 5 and Figure 10A). This trait is commonly found in symbiont organisms and is believed to have played a key role in the evolution of organelle structures, as a feature often attributed to endosymbiotic organisms is a reductive evolution related to prolonged growth in a stable and specific ecological niche[67,68]. In this context, it is not surprising that the genome of *S.*

glucoliberatum PABB004 has a smaller genome versus other reported strains of *Scenedesmus*. However, it is also uncertain the length of time that any specific alga has been in close association with strains of *P. bursaria*, which is especially true in this case. Many endosymbionts obtained from *P. bursaria* are capable of independent growth outside of the strain, so the question of how stable and prolonged the mutualistic symbiotic relationships are between green algae and *P. bursaria* is still very much outstanding.

Finally, by means of proteome comparisons, we provide two sequence sets that could be used to direct further research on the molecular mechanisms of sugar secretion by green algae. With these, and a further collection of additional strains and characterization in the near future, we hope to help uncover the common directives for maltose secretion and those unique to the *S. glucoliberatum* PABB004 glucose secretion phenotype.

Experimental procedures

Scenedesmus glucoliberatum PABB004 isolation and culture conditions

Natural isolates of *P. bursaria* were enriched by collecting 1 L of water from various lakes and ponds in the Minneapolis metro region and growing for several days under fluorescent lights. Initial medium was based on Bristol's recipe and contained 250 mg $NaNO_3$, 250 mg K_2HPO_4, 75 mg $MgSO_4{\cdot}7H_2O$, 25 mg $CaCl_2{\cdot}H_2O$, 25 mg NaCl and 15 mg ferric ammonium citrate, all per liter, adjusted to pH 7.8. Medium was supplemented with two grains of rice or a single wheat seed (per 250 mL) prior to autoclaving in a 500 mL Erlenmeyer flask and grown on a light table until *P. bursaria* were observed swimming in the medium. Aliquots of healthy *P. bursaria* were transferred several times to enrich until there were approximately 4 or 5 *P. bursaria* cells per 50 μL aliquot. Cultures were then spotted onto freshly prepared agar plates of freshwater SAG medium as drops containing 2-3 μL of liquid and drops containing *P. bursaria* were marked on the plate following visual inspection under a microscope. *P. bursaria* cells generally burst on the plate as the droplet evaporated or was absorbed into the plate, releasing the algae contained within. Algal cells derived from the *P. bursaria* that formed colonies were picked with sterile toothpicks, and carefully transferred to sterile media plates to isolate and purify the strains. Strains were passaged several times on solid agar plates until pure cultures were obtained. One strain, later designated *Scenedesmus glucoliberatum* PABB004 was selected for further studies.

Sugar secretion experiments

Sugar secretion experiments were conducted in 1.5 L tubular photobioreactors using the same modified freshwater SAG medium as described previously[58], and supplemented with 10 mM each of MES, MOPS and PIPES, then brought to the desired pH. Culture pH levels were adjusted throughout the experiment on a semi-daily basis to maintain it within 0.1 pH units of the target value. Levels of extracellular glucose and maltose were analyzed enzymatically as described previously.

Light and scanning electron microscopy

Cells of *S.* glucoliberatum PABB004 cultivated on agar plates were harvested and prepared for microscopy imaging. For light microscopy 5 μL of algal suspension were mounted on alcohol-cleaned glass slides with a coverslip and sealed with dental wax. Images were acquired in a Nikon 90i microscope equipped with a Plan Apo VC 100x Oil DIC N2 1.4 NA objective and a DS-Fi2-U3 camera. The microscope and camera were controlled by Nikon Elements software (5.02). After adjusting the condenser for Kohler illumination, software was used to adjust the white balance. Images (24-bit RGB) were acquired with Auto exposure time and pixel resolution of 2560x1920 (fine, final pixel size 30 nm).

For SEM, samples were suspended in 2% gluteraldehyde and 0.1 M phosphate buffer, kept at 4°C for at least 12 hours, and rinsed in 0.1 M phosphate buffer (3 times, 10 minutes each). Samples were then placed in 1% osmium tetroxide and 0.1 M phosphate buffer for 2 hours at room temperature. Specimens were rinsed in ultrapure water (NANOpure Infinity; Barnstead/Thermo Fisher Scientific; Waltham, Maryland) (3 times, 10 minutes each) and dehydrated in an ethanol series (50%, 75%, 95%, 100%; 5 minutes, 2 times each). After the samples were in 100% ethanol, they were put through two changes of hexamethyldisilazane (HMDS) for 5 min each. Drops of the suspension were placed on individual round glass cover slips cleaned with acetone, mounted on aluminum stubs, and allowed to air dry. The stubs were sputter-coated with gold-palladium (60-40) and observed in a Hitachi S3500N scanning electron microscope (Hitachi High Technologies America, Inc.; Schaumberg, Illinois) at an accelerating voltage of 10 kV.

Genomic DNA isolation and sequencing

Cells of *S. sp* PABB004 grown on agar plates were used for total DNA isolation using the ZR Fungal/Bacterial DNA Microprep kit (Zymo Research, Irvine, CA, USA) as directed by the manufacturer. Following isolation, DNA quantity and purity was measured with a NanoDrop 2000 spectrophotometer (Thermo Scientific, Waltham, MA, USA).

Illumina sequencing was performed at the University of Minnesota Genomics Center, following the standard protocol for pair-end reads Illumina HiSeq 2500 (Illumina, San Diego, CA, USA). Library preparation and long read sequencing was carried out by the Rochester Mayo Medical Genome Facility with PacBio SMRT-RS technology as previously described. Sequencing was done using 6 SMRT cells.

Genome assembly

PacBio long reads were assembled using Canu version 1.8[69]. Genome size was estimated through consecutive iterations of the assembly process with starting values ranging between 30 and 120 Mb. Contigs with negative covStat, as reported by Canu, were used as queries in BLASTn searches to look for and remove bacterial or PacBio contaminants. Mitochondrial DNA was manually circularized by aligning and clipping the contig overhangs. The resulting contig set was polished with Illumina short reads using Pilon version 1.22, with default parameters for 7 iterations, when there were no more changes reported by the pipeline.

Assembly quality was assessed with N50 and L50 statistics, the proportion of PacBio reads that were used by Canu and the percent Illumina reads that mapped back to the assembly. Completeness was evaluated with BUSCO quality scores (BUSCO version 4.0.5) using the genomic mode and the chlorophyta_odb10 dataset (creation date Nov/20/2019), which contains 16 species and 1519 BUSCOs.

Additionally, each contig end was examined to calculate GC% and find repetitive regions. Contig ends with GC content below 50%, and 100% repetitive were marked as telomers and confirmed by manual inspection. Contigs with telomers $(TTTAGGG)_n$ on both ends were considered full chromosomes.

Repetitive regions were soft-masked with RepeatMasker version 4.0.5 using the available *Scenedesmus* repeat library.

S. glucoliberatum PABB004 RNA isolation, sequencing and transcriptome assembly

S. glucoliberatum PABB004 cells were harvested at midday after 5 and 13 days of cultivation, frozen in liquid nitrogen and stored at -80°C until RNA extraction. Total RNA extraction was performed as described previously. Poly(A)+-tag based mRNA isolation and sequencing were conducted by the University of Minnesota Genomics Center. Poly(A)+ RNA was isolated with oligo (dT) magnetic beads and fragmented at elevated temperatures. Synthesis of cDNA used the fragmented mRNA as template and random primers. Additionally, a cDNA library was constructed by reverse PCR amplification of adapter-mRNA sequences. All cDNA was sequenced with HiSeq 2500 Illumina technology.

RNA reads were aligned to the genome using STAR version 6.06.017_01. The resulting alignment file (bam) was used to guide Trinity version 2.9.1 for a genome-guided de novo transcriptome assembly using the *S. glucoliberatum* PABB004 genome as reference.

Gene annotation and sequence analysis

Gene annotation was carried out separately for the nuclear and organellar contigs. Nuclear genes were predicted using AUGUSTUS version 3.2.3 and following Basic Protocol 11 and Alternative Protocol 13 as described by the pipeline authors[70]. The "chlamydomonas" species was used for the prediction parameters. RNA reads aligned to the genome as described above were used to generate an intron hints file for AUGUSTUS. Chloroplast contigs were annotated using the online GeSeq[71] prediction tool (https://chlorobox.mpimp-golm.mpg.de/geseq.html) with *Tetradesmus obliquus, Chlamydomonas reinhardtii* and *Chlorella sorokiniana* as reference sequences for BLAT searches. The resulting gff3 files were manually inspected to remove duplicate features. Reading frame convention (column 8 of the gff3 tables) was changed from [1-3] to [0-2] to be in accordance with the AUGUSTUS output. Mitochondrial genes were predicted using Prodigal version 2.50 and selected based on sequence similarity searches using BLASTp.

Nuclear and organellar annotation gff3 files were joined and further polished based on the output discrepancy and validation reports given by table2asn (https://ftp.ncbi.nih.gov/toolbox/ncbi_tools/converters/by_program/). Briefly, the “transcript”, “start_codon”, “stop_codon” and “transcription_end_site” features were removed, duplicated and contained features were filtered out and the mitochondrial genetic code was changed to *S. obliquus* Mitochondrial Code (transl_table=22).

Functional annotation was produced by comparing the translated *S. glucoliberatum* PABB004 predicted CDSs with the reviewed Swiss-Prot protein database (available from https://www.uniprot.org/downloads, accessed on 06/2020) through a BLASTp (version 2.8.1) search using an e-value of 1e-6 and selecting only the best-score hit for each protein. Protein names were given to the predicted CDSs based on these BLASTp results and manually curated following NCBI recommendations. Annotation quality was assessed with BUSCO version 4.0.5 using the protein mode and the chlorophyta_odb10 dataset.

The *S. glucoliberatum* PABB004 proteome was also examined through the KEGG annotation tool using default thresholds, BLASTp and BBH algorithms and the following GENES dataset: hsa, dme, cel, ath, sce, cho, eco, nme, hpy, rpr, bsu, lla, cac, mge, mtu, ctr, bbu, syn, bth, dra, aae, mja, ape, cre, mng, apro, olu, ota, mis, and mpp.

S. glucoliberatum PABB004 phylogenetic analysis

S. glucoliberatum PABB004 contigs were initially examined to find the SSU rRNA gene through a BLASTn search against green algae SSU rRNA genes. Then, a diverse dataset of SSU sequences from green algae was gathered to infer *S. glucoliberatum* PABB004 evolutionary history. This dataset contained representatives of the Chlorophyceae and Trebouxiophyceae clades[68,72] among the green algae lineage and two bacterial SSU rRNA gene sequences (outliers). Data was downloaded from the SILVA (https://www.arb-silva.de/) and NCBI nucleotide databases. Only SSU sequences derived from PCR amplifications (not genomic segments) were used to avoid alignment issues with introns and gene prediction artifacts.

SSU rRNA sequences were aligned using MUSCLE (UPGMA clustering, gap opening penalty -400 and no gap extension penalty) to remove *S. glucoliberatum* PABB004 overhangs as described by Hoshina et al. (2010)[53]. The final sequence set was aligned again using the same strategy and an evolutionary analysis was conducted in MEGA X[73]. A Maximum Likelihood analysis with the Kimura 2-parameter model was chosen to estimate evolutionary distances as suggested by MEGA X Best Model Tool. The bootstrap consensus phylogenetic tree was inferred from 100 replicates. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates were collapsed. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (2 categories (+G, parameter = 0.5045)) on a total of 3580 positions (gaps included) in the final dataset. Tree image was prepared with the web iTOL tool.

S. glucoliberatum PABB004 phylogenomic analysis

To have a better understanding of *S. glucoliberatum* PABB004's origin, an Average Nucleotide Identity (ANI) analysis was performed using its transcriptome and transcriptomic data from Chlorophyceae and Trebouxiophyceae algae. Transcriptomes were analyzed using GET_HOMOLOGUES_EST[74,75] version 25082020. Briefly, transcripts were clustered (OrthoMCL version 1.4) based on sequence similarity with sequences showing at least 70% nucleotide identity and 75% sequence coverage classified on the same cluster. Sequence similarity was calculated from BLASTn searches with an e-value threshold of 1e-5. Sequence clusters present in all taxa (i.e. that had at least one transcript from each species) were used to calculate the percent average nucleotide identity among pairs of organisms.

Proteomic comparisons

To explore the molecular basis of the sugar secretion phenotype in algae, the predicted proteomes of two sugar-secreting organisms (*S. glucoliberatum* PABB004 *and M. conductrix*) and two non-sugar secretors (*C. reinhardtii and C. sorokiniana*) were compared using OrthoVenn2[76]. In detail, the curated proteomes from *M. conductrix*, *C. sorokiniana* and *C. reinhardtii* were downloaded from the Swissprot-Uniprot database (accession numbers UP000239649, UP000239899 and UP000006906, respectively) and all protein sequences including *S. glucoliberatum* PABB004's were clustered using DIAMOND version 0.9.24 with an e-value threshold of 1e-10 and the OrthoMCL algorithm with an inflation value of 1.5. The protein set shared by *M. conductrix* and *S. glucoliberatum* PABB004 and absent from the other alga was analyzed for its potential relationship with sugar-secretion mechanisms. This potential sugar-related subset of proteins was further examined with BLASTp searches against the NCBI non-redundant and InterProScan v5.23-62.0 databases.

Data statement

All datasets generated in the production of this manuscript can be accessed at the DDBJ/EMBL/GenBank databases under accession numbers: BioProject PRJNA637367, BioSample SAMN15103613, genome assembly JABVCE000000000, locus tag HT031, genomic PacBio read library SRR12546905, genomic Illumina read library SRR12974924 and RNA-Seq Illumina read library SRR12973628.

Chapter 3: Genomic and Transcriptomic Analysis of Sugar-secreting and Non-sugar Secreting Algae through Orthologous Groups

Synopsis

Some microalgae from the *Chlorella, Micractinium* and *Scenedesmus* genera have been reported to secrete considerable amounts of simple sugars to the extracellular environment. Despite it being an attractive trait for downstream fermentative applications, the molecular toolkit that these microalgae use to secrete simple sugars remains unknown. This work compares the genomic and transcriptomic data of several phylogenetically related strains with different sugar secretion phenotypes using a different approach than previous studies. Here the focus is put on the functional orthologs present in strains from different phenotypic groups. The comparisons performed revealed sugar transporters and glucosidases that are likely to play an important role in the sugar secretion pathways of *M. conductrix* SAG 241.80 and *S. glucoliberatum* PABB004.

Introduction

Some species of green microalgae have been previously reported to secrete simple sugars to the extracellular media[46,56–58], which is an advantageous trait for the downstream use of such sugars in fermentative processes[51,77]. However, the molecular tools that enable sugar release are still unknown. This hinders algal strain engineering and their biotechnological use. A possible approach to uncover their sugar-releasing molecular toolkit are genomic and transcriptomic comparative studies, but the scarcity of genomic data and the lack of gene predictions in the sequenced genomes make this difficult[59]. This study explores the use of functional orthologous groups to offset the low abundance of genomic data related to algae with different sugar secretion phenotypes.

The prediction of genes and their function are heavily based on the concept of sequence similarity[2,8,76]. Highly similar genes are believed to have a common evolutionary ancestor and therefore share functional characteristics. Hence, gene functions are assigned to newly discovered sequences by similarity to other genes previously characterized. Although valid for functional orthologs with high sequence similarity, this procedure fails to correctly characterize the function of genes that have high sequence similarity and different functions (i.e., paralogs with different functions), and functional orthologs with low sequence similarity.

This limitation is intensified if the functional prediction relies on position-independent algorithms to quantify the distance between sequences. One such algorithm is the commonly used Blast[3]. An alternative is a position-aware algorithm such as HMMER[5], which calculates the probability of mutations by considering the different positions in the sequence using Hidden Markov Models (HMM). The relationship between structure and function in proteins is better characterized by a position-aware algorithm because it allows for a greater divergence in the sequences if those changes are not located in critical sites such as catalytic or other highly conserved residues.

Additionally, when comparing large sets of sequences like genomic or transcriptomic data, one is challenged by the correct choice of the subset of sequences that have a biological meaning pertaining to the research question. The clustering of sequences by similarity is rarely enough to dissect the genetic traits that underlie a phenotype. Therefore, the common alternative to uncover key genes is analyzing the differential gene expression between two or more conditions in which the phenotype of interest is and is not observed. However, this approach is not possible in the case of genes that are expressed constitutively or when there are no known conditions that result in a different phenotype to compare with. The glucose secretion by *Senedesmus glucoliberatum* PABB004 belongs to the later group[46].

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a “knowledgebase” in which the gene and genomic data from a variety of organisms are grouped into functional orthologous groups (KOs). KOs are also organized based on our current knowledge of biochemical pathways (PATHWAY database), types of proteins (BRITE hierarchies) and other manually curated categories. Mapping an organism’s genome or proteome to the orthologous groups in the KEGG database facilitates their analysis by giving a common designation of KEGG identifiers (KO). These can easily be compared among different proteomes. Additionally, they can be analyzed using existing KEGG infrastructure such as the PATHWAY, BRITE and MODULE tools[6,7].

In this work the genomic data of sugar-secreting and non-secreting algae are analyzed through the comparison of functional orthologous groups. Functional orthologs are inferred from HMM based sequence similarity searches and mapped to the KOs from KEGG. This comparison aims to uncover the unique functionalities that allow sugar-secretion in *S. glucoliberatum* PABB004 and *M. conductrix*. Additionally, the transcriptomes of two sugar-secreting algae and four non-sugar secretors are analyzed.

Results

Allocation of KEGG identifiers

The predicted proteome of *S. glucoliberatum* PABB004 contains 6947 protein-coding sequences (CDS). They were initially annotated through a sequence similarity approach using the SwissProt-Uniprot database as reference and the Blastp algorithm. With this approach, 3062 proteins (44%) found homology support and were described by a possible function based on a similar protein or by the domains they contain[46]. When the same proteome was annotated using two different sequence similarity search algorithms, the position-independent BlastKOALA and the HMM based KofamKOALA[77], 3105 (45%) and 3255 (47%) proteins were allocated a KEGG identifier (KO), respectively. The joint annotation resulted in 3772 proteins (54%) with a predicted function. Using both approaches, increased the fraction of proteins with a predicted function by approximately 10%. The proteomes of *C. sorokiniana* and *M. conductrix* were annotated using KofamKOALA and the already existing annotation files for *C. reinhardtii* were downloaded directly from KEGG. All annotation results are summarized in Table 7.

The complete dataset used for further comparisons contains 51990 genes, of which 15486 (29.8%) are annotated. There is a total of 4244 unique KEGG identifiers.

Genomic comparisons from a functional perspective

The proteins coded in the genomes of all organisms were analyzed through their KEGG orthologs. To find the unique properties of each individual, all KOs in each genome were compared against all other genomes. Then, to investigate the genetic basis of the sugar secretion phenotype of *M. conductrix* and *S. glucoliberatum*, their joint annotations

(“SRA” from sugar releasing algae) were compared to the joint annotations for *C. reinhardtii* and *C. sorokiniana,* two phylogenetically close algae that do not release sugar to the media.

Table 7. Allocation of KEGG identifiers to selected green algae proteomes

Individual strains	CDS count	CDS with KO	Non redundant KOs	Species-specific KOs
C. reinhardtii	17875	3679 (21 %)	3679	117
C. *sorokiniana*	9482	3882 (41 %)	4385	69
M. conductrix	9122	4153 (45 %)	4673	63
S. glucoliberatum PABB004	6947	3772 (54 %)	4760	245
Pooled genomes				
C. reinhardtii and *C. sorokiniana* (“control”)	27357	7561 (28 %)	3841	
M. conductrix and *S. glucoliberatum* (“SRA”)	26026	11595 (44 %)	3994	
Complete dataset	51990	15486 (29.8%)	4244	

The first approach compared all KOs in each organism’s proteome and found the species-specific set for each one (Table 7). Interestingly, *S. glucoliberatum* had the largest species-specific set, despite having the smallest genome of all green algae tested. This is possibly due to the evolutionary distance that separates it from the rest of the species.

To further understand what makes each species and each phenotypic group unique, the species-specific KOs were grouped by its BRITE category (i.e., the type of molecule). Without counting enzymes, all species had unique KOs in 11 BRITE categories (Figure 11), suggesting that enzymes and these 11 categories might be the most mutation-tolerant groups within these genomes, possibly due to a high gene copy number. Furthermore, unique KO from the categories ko03036 (“Chromosome and associated proteins”) and ko01001 (“Protein kinases”) seem to be more common in the SRA group than in the control group, indicating a possible relationship with their phenotype. The “transporters” and “membrane trafficking” functions did not show a greater abundance of unique KOs in any of the phenotypic groups.

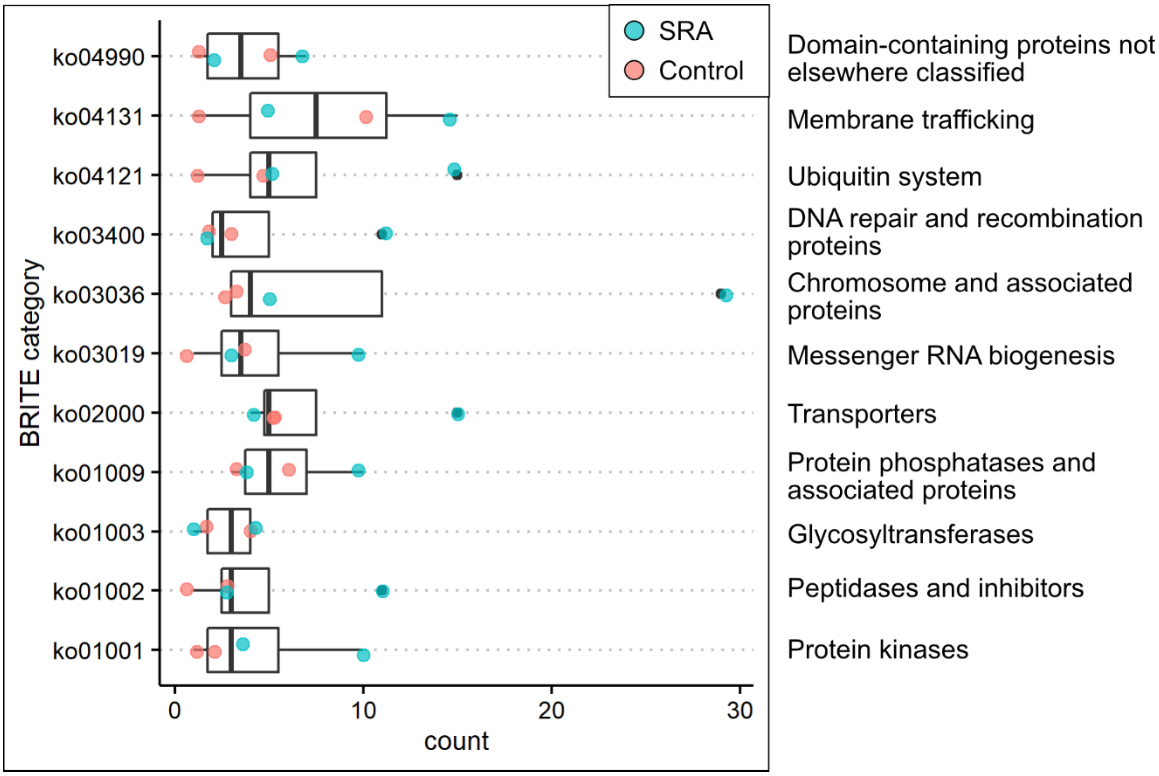


Figure 11. BRITE categories of species-specific KOs

The unique KEGG Orthology identifiers (KO) found in each species were clustered by their BRITE category. The 11 categories shared by all species besides “enzymes” are sown in the image.

The subset shared by all SRA species was also investigated. From its 3994 unique KOs, 402 were unique to the group (i.e., not found in either *C. reinhardtii* or *C. sorokiniana*). Similarly, from the 3841 unique KOs in the control set, 250 KOs were exclusive to the non-sugar secreting algae. Most of the unique KOs in both sets were species-specific

(357/402 for the SRA and 238/250 for the control group) but 45 KOs were shared by both SRA genomes and 12 KOs were found in both control genomes. The presence of these common functionalities in all organisms from a specific phenotypic group hints of their possible key role in their common lifestyle. However, they are not connected within a specific pathway or biochemical module, nor they are represented by a unique type of molecule; therefore, there are no straightforward commonalities to make sense of their conjoint biological role.

To offset this handicap, the pooled genomes from SRA and control groups were compared with a pathway completeness analysis. This aims to balance the deficiencies in the individual genome annotations (missing genes due to prediction errors) and provide a common frame to understand the overall picture of the proteins that define such different phenotypes.

Pathway completeness analysis

A Relative Pathway Completeness (RPC) analysis compares the total number of functions from a specific pathway in two groups. RPC gives a general idea of the relative importance of a specific pathway between the groups compared and provides a context for the inference of the biological meaning of their differences. In this case, the RPC was calculated using the unique 402 functions from the SRA group with respect to the full set (3841 KOs) of the control group. As expected, most of the pathways were overrepresented in the full annotation set, but 3 pathways were overrepresented in the SRA group (RPC > 1.5). These are pathway 531 (Glycosaminoglycan degradation), 604 (Glycosphingolipid biosynthesis) and 4923 (Regulation of lipolysis in adipocytes). Given the difference between the sizes of the two sets that were compared, the overrepresentation of these three pathways in the SRA group (the smaller set) is specially meaningful.

Functional analysis of overrepresented pathways in the SRA group

Glycosaminoglycan degradation

The SRA group has the functions K01135 (arylsulfatase B [EC:3.1.6.12]), a sulfur-ester hydrolase that acts on N-acetyl-D-galactosamine-4-sulfate or N-acetylglucosamine 4-sulfate. Function K01205 (alpha-N-acetylglucosaminidase [EC:3.2.1.50]), a hydrolase that performs the cleavage of terminal non-reducing N-acetyl-D-glucosamine residues in N-acetyl-alpha-D-glucosaminides. And the function K12309 (beta-galactosidase [EC:3.2.1.23]), a hydrolase that cleaves the non-reducing beta-D-galactose residues in beta-D-galactosides, some enzymes in this group also act on beta-D-fucosides and beta-D-glucosides.

Interestingly, functions K01205 and K01230 also have a role in the acid hydrolysis that occurs in lysosomes. Exocytosis is an alternative avenue to membrane transporters for the secretion of simple sugars. Therefore, the *genes coding for the acid hydrolases KAF8069596.1 (K01205) in S. glucoliberatum and A0A2P6V0S3_9CHLO (K01230) in M. conductrix* are interesting targets to further study their effects on sugar secretion by these algae.

Glycosphingolipid biosynthesis

The function K12309 appears again, suggesting multiple roles for the A0A2P6V0S3_9CHLO protein in *M. conductrix.* Additionally, the function K03372, an

MFS transporter also predicted to have a role in the acetylation of glycans appears in this pathway and has two representatives in each SRA species, these are: *proteins A0A2P6VJ16_9CHLO and KAF8059178.1*.

Regulation of lipolysis in adipocytes

There are molecules from different parts of this pathway in both groups; however, one of the main differences among the SRA and control species is that *S. glucoliberatum* has a membrane receptor (K12323) which is described as an atrial natriuretic peptide receptor A [EC:4.6.1.2]. The human ortholog of this protein is contains a 21 residue transmembrane domain and a 568 residue cytoplasmic domain with homology to the protein kinase family and to a subunit of the soluble guanylate cyclase and a protein kinase[78]. It has a role in the cGMP-PKG and cAMP signaling pathways.

In the presence of natriuretic peptide A (ANP), the receptor NPR-A makes guanosine 3',5'-cyclic phosphate (cGMP), which activates the cGMP-dependent protein kinase 1 (PKG). PKG has a lot of regulatory functions in the cell. PKG is part of 10 different pathways in the KEGG PATHWAY database including modulating the cell growth, differentiation and apoptosis and regulating Ca^{+2} levels.

In turn, when ANP binds its receptor, the cAMP signaling pathway is indirectly activated. cAMP is a central signal transduction metabolite in the cell. It is connected to the Ca^{+2} signaling pathway and activates the protein kinase A (PKA). PKA has even more functions than PKG. PKA regulates the exchange of ions (Na^+, K^+, Ca^{+2}, H^+, Cl^-), inhibits apoptosis, activates lipolysis and fatty acid degradation, and in summary has roles in 74 different pathways in the KEGG PATHWAY database. cAMP also activates the Rap1 signaling pathway which controls cell movements and proliferation. And finally, cAMP also regulates secretion, exocytosis, cytoskeleton reconstruction, cell proliferation and apoptosis through the metabolite 1,2-Diacyl-sn-glycerol 3-phosphate.

The importance of this receptor is evident from its various roles in the cell and the exclusive presence in *S. glucoliberarum*'s genome points towards an important role in its phenotype. *S. glucoliberatum's ortholog of this receptor is the protein KAF8068337.1.*

Sequence similarity and clustering of functional orthologs

To test whether this KO mediated approach provides different insights than a sequence similarity-based clustering, the sequence similarity of *M. conductrix* and *S. glucoliberatum* PABB004 proteins belonging to the same KO within the unique SRA functionalities was calculated with BlastP (Table 8). The KOs compared were specifically chosen from the unique pool of SRA genes that have only one ortholog per strain. Blastp comparisons were done using the *M. conductrix* protein as query and the *S. glucoliberatum* protein as subject.

As shown in Table 8, most (but not all) of *M. conductrix* and *S. glucoliberatum* functional orthologs predicted with KEGG annotation tools have also high similarity scores (they have an e-value in the Blastp searches small enough to be considered homologs). However, OrthoVenn, the sequence similarity clustering algorithm used before, did not cluster together the SRA functional orthologs in a unique group separated from similar proteins from the non-sugar secretor algae.

This demonstrates that sequence similarity approaches based only in a position-independent alignment algorithm like BLAST are insufficient to associate all functional orthologs. Moreover, even when associated in the same homology cluster due to a high sequence similarity, the approach used before was unable to exclude similar sequences from the control group. This suggests that another layer of information is required to successfully separate homologous sequences into functional orthologs. In the clustering approach used in this study the biochemical module completion requirement used by KEGG to form their orthologous groups accomplished a better separation of functional orthologs.

Table 8. Sequence similarity within functional groups in M. conductrix and S. glucoliberatum

KO	Proteins				Similarity based clustering		
	S. glucoliberatum		M. conductrix		Blastp		OrthoVenn
	ID	length (aa)	ID	length (aa)	E-value	Query coverage (%)	clusters
K01342	KAF8068246	1120	A0A2P6VKR1_9CHLO	1703	5e-13	47	103,182
K23012	KAF8057630	705	A0A2P6VRN2_9CHLO	720	2e-6	28	3619 both[α]
K03131	KAF8068263	715	A0A2P6VL26_9CHLO	366	5e-27	86	3275 both[α]
K12617	KAF8058443	1048	A0A2P6VBK2_9CHLO	885	2e-37	35	3734 both[α]
K13094	KAF8072973	1297	A0A2P6UZS2_9CHLO	2027	NA	NA	3176, 2283
K15601	KAF8059101	1426	A0A2P6V1E4_9CHLO	1735	3e-49	31	261 both[α]
K03372	KAF8059178	1103	A0A2P6VJ16_9CHLO	623	7e-73	75	538 both[α]
K03641	KAF8056023	2720	A0A2P6V648_9CHLO	367	6e-46	83	62, NA[β]
K00545	KAF8067201	533	A0A2P6VE74_9CHLO	907	6e-61	31	955 both[α]
K02414	KAF8063050	2731	A0A2P6V3M9_9CHLO	2985	NA	NA	177, 985
K14611	KAF8056380	698	A0A2P6V9T9_9CHLO	953	3e-157	52	3125 both[α]

[α] Clusters with proteins from *C. reinhardtii*, *M. conductrix*, *C. sorokiniana* and *S. glucoliberatum*
[β] Not included in any cluster, marked as a singleton

Unique transporters found in *S. glucoliberatum* PABB004

Given the advantageous and unique glucose secretion phenotype of *S. glucoliberatm* PABB004, its unique KOs from the transporter category were investigated (Table 9).

Table 9. Unique transporter KOs from S. glucoliberatum PABB004

KO	Description
K05673	ATP-binding cassette, subfamily C (CFTR/MRP), member 4
K08175	MFS transporter, FHS family, Na+ dependent glucose transporter 1
K08191	MFS transporter, ACS family, hexuronate transporter
K07552	MFS transporter, DHA1 family, multidrug resistance protein
K08153	MFS transporter, DHA1 family, multidrug resistance protein
K02440	glycerol uptake facilitator protein
K03322	manganese transport protein
K03325	arsenite transporter
K14445	solute carrier family 13 (sodium-dependent dicarboxylate transporter), member 2/3/5
K14708	solute carrier family 26 (sodium-independent sulfate anion transporter), member 11
K09869	aquaporin-8
K09884	aquaporin rerated protein, invertebrate
K09936	bacterial/archaeal transporter family-2 protein
K19791	iron transport multicopper oxidase
K21993	formate transporter

One glucose transporter is found in this set: the Na+ dependent glucose transporter 1 (K08175, protein accession KAF8055830.1). There were no unique sugar-transporters shared by *M. conductrix* and *S. glucoliberatum.*

Transcriptomes assembly and annotation

To test the unique KO sets found with the genomic comparisons, six phylogenetically related strains with diverse sugar secretion phenotypes were cultivated in equivalent conditions and their mRNA sequenced. The strains selected are *C. sphaericum* UTEX 1354, *C. proboscideum var. gracile* UTEX 184, *S. naegelii* and *S. dimorphus,* all non-sugar secretors, and two strains which secrete high amounts of glucose and maltose to the extracellular media: *S. glucoliberatum* PABB004 and *C.* sp. PABB002.

Illumina sequencing generated over 120 million reads with all expected barcodes detected. All libraries had a mean quality score above Q30 and the insert size is approximately 200 bp. All read libraries had between 19 and 31 million reads, except *C. sp.* PABB002, which had almost triple that number (58 million reads). All de novo assemblies were constructed using the same parameters for Trinity. Tables 10 and 11 summarize the transcriptome assembly and KEGG identification results for all strains.

Table 10. De novo transcriptome assemblies

Strain	Reads (M)	Transcripts	Complete	Internal	5’ partial	3’ partial	C/I[α]
S. glucoliberatum	22.1	10820	624	6488	2741	967	0.0962
S. dimorphus	31.0	52999	19937	12526	11134	9402	1.5916
S. naegelii	20.8	44841	17256	11014	9108	7463	1.5667
C. sp. PABB002	58.0	31137	4071	14827	8158	4081	0.3696
C. sphaericum UTEX 1354	21.8	23621	6089	6255	9306	1971	0.9735
C. proboscideum UTEX 184	19.7	24821	10924	4968	6536	2393	2.1988

[α] Ratio between complete and internal transcripts

Table 11. De novo transcriptomes' allocation of KEGG identifiers

Strain	Transcripts	Annotated	% Annotated	Unique KO
S. glucoliberatum	10820	2455	23	1532
S. dimorphus	52999	5562	10	2747
S. naegelii	44841	5087	11	2693
C. sp. PABB002	31137	4160	13	2193
C. sphaericum UTEX 1354	23621	3419	14	1916
C. proboscideum UTEX 184	24821	3924	16	2432

Interestingly, *S. glucoliberatum*, *C. sp.* PABB002 and *C. sphaericum* UTEX 1354 show a lower ratio of complete/internal transcripts compared to the other species, suggesting a more complex splicing process than their closely related algae.

KO comparison among transcriptomes from sugar-secreting and non-sugar secreting algae

The KOs from the de novo transcriptomes from *S. glucoliberatum* and *C. sp.* PABB002 were compared to the KOs from the de novo transcriptomes from *C. sphaericum* UTEX 1354, *C. proboscideum var. gracile* UTEX 184, *S. naegelii* and *S. dimorphus*. There are 55 unique functionalities in the transcriptomes of the sugar-secreting algae. From those, 9 KOs are shared by both strains, 35 are unique to *C. sp.* PABB002 and 11 are unique to *S. glucoliberatum* PABB004.

None of the 9 shared KOs between *S. glucoliberatum* and *C. sp.* PABB002 were present in the annotated proteome of *S. glucoliberatum* PABB004 (Table 12) and only 8 of the 11 KOs that were unique to its transcriptome were found in its predicted proteome (Table 13).

Table 12. Unique KOs shared by S. glucoliberatum PABB004 and C. sp. PABB002's transcriptomes

KO	Short name	Description
K04768	acuC	acetoin utilization protein AcuC
K24826	CRIPT	cysteine-rich PDZ-binding protein
K08506	SYP7	syntaxin of plants SYP7
K12813	DHX16	pre-mRNA-splicing factor ATP-dependent RNA helicase DHX16 [EC:3.6.4.13]
K03963	NDUFB7	NADH dehydrogenase (ubiquinone) 1 beta subcomplex subunit 7
K15746	crtZ	beta-carotene 3-hydroxylase [EC:1.14.15.24]
K11130	NOP10, NOLA3	H/ACA ribonucleoprotein complex subunit 3
K07342	SEC61G, SSS1, secE	protein transport protein SEC61 subunit gamma and related proteins
K02634	petA	apocytochrome f

Table 13. Unique KOs in S. glucoliberaum's de novo transcriptome

KO	Protein	Description
K00412	KAF8054287.1	ubiquinol-cytochrome c reductase cytochrome b subunit
K02923	KAF8071164.1	large subunit ribosomal protein L38e
K02979	KAF8060549.1	small subunit ribosomal protein S28e
K03358	KAF8066294.1	anaphase-promoting complex subunit 11
K03938	KAF8069460.1	NADH dehydrogenase (ubiquinone) Fe-S protein 5
K06228	KAF8056751.1	fused [EC:2.7.11.1]
K10845	KAF8061303.1	TFIIH basal transcription factor complex TTD-A subunit
K12625	KAF8062733.1	U6 snRNA-associated Sm-like protein LSm6
K18633	NA	mitotic-spindle organizing protein 1
K23408	NA	cell division cycle-associated protein 7
K24140	NA	glutaredoxin-dependent peroxiredoxin [EC:1.11.1.25]

Expression levels of selected KOs in *S. glucoliberatum* PABB004

S. glucoliberatum has 158 KO assignments within the transporter category in its genome. Due to its particular sugar-secretion phenotype, its unique sugar efflux transporters (K15382: KAF8054899.1, KAF8061939.1, KAF8068188.1) and the Na$^+$ dependent glucose transporter (K08175: KAF8055830.1) are of special interest. These and other unique transporters from Table 8 have been selected to verify their expression levels using *S. glucoliberatum* genome-guided transcriptome and the estimated FPKM values. Additionally, the genes of special interest identified in the sections above (the acid hydrolases KAF8069596.1 and KAF8059178.1 and the receptor KAF8068337.1) were also examined (Table 14).

Table 14. Expression level of selected genes from S. glucoliberatum PABB004

	Gene product	KO	Length (aa)	Short name	Locus tag	Transcript id (CUFF.)	FPKM
transporters	KAF8054899.1	K15382	759	APY5	HT031_006898	4952	184.246
	KAF8061939.1	K15382	248	SWEET6B	HT031_004199	-	0
	KAF8068188.1	K15382	871	hypothetical	HT031_001874	1495	449.567
	KAF8055830.1	K08175	1775	hypothetical	HT031_006605	4717	992.504
	KAF8071161.1	K08191	590	PHT4	HT031_001243	see note[α]	11.57
	KAF8067209.1	K02440	1606	GLPF	HT031_002256	1746	172.351
	KAF8065942.1	K14445	783	SPBC3B8.04c	HT031_003003	see note[α]	17.900
	KAF8060559.1	K14708	2514	SULTR2	HT031_004736	3368	18.002
	KAF8063729.1	K09936	667	MAN4	HT031_003585	see note[α]	0.405
	KAF8068170.1	K21993	5122	RH36	HT031_001856	1483	79.774
	KAF8069596.1	K01205	886	hypothetical	HT031_001713	1371	3.947
	KAF8059178.1	K03372	1103	hypothetical	HT031_005350	3807	12.019
	KAF8068337.1	K12323	625	Gyc88E	HT031_002025	1599	9.063

[α] Transcript id is in the form of “gene-<locus_tag>”

Expression for all the selected proteins was detected (FPKM > 0) except for the sugar efflux transporter KAF8061939.1, this includes the two acid hydrolases and the NPR-A membrane receptor. Among the transporters examined, the highest level of expression is by far that of the Na^+ dependent glucose transporter KAF8055830.1, followed by the sugar efflux transporter KAF8068188.1. More modest expression levels are found in the glycerol uptake facilitator protein KAF8067209.1 and the sugar efflux transporter KAF8054899.1. In contrast, the hexuronate transporter KAF8071161.1, the sodium-dependent dicarboxylate transporter KAF8065942.1, the sodium-independent sulfate anion transporter KAF8060559.1 or the bacterial/archaeal transporter family-2 protein KAF8063729.1 have considerably lower transcription levels.

Discussion

This work compares the genomes and transcriptomes of several strains of green algae with two distinct sugar secretion phenotypes. *Senedesmus glucoliberatum* PABB004, *Micractinium conductrix* SAG 241.80 and *Coelastrum sp.* PABB002 secrete glucose and maltose to the extracellular media which accumulates to concentrations at the mM level. *S. dimorphus, S. naegelii, C. sphaericus* UTEX 1354, *C. proboscideum var. gracile* UTEX 184 and *Chlorella sorokiniana* UTEX 1602 do not secrete either sugar as per our knowledge. The sugar-secretion phenotypes from these strains have been previously characterized by our laboratory using enzymatic assays[46,58].

Differences in the similarity and orthology-based comparative approaches

Datasets were compared by their KEGG orthology identifiers (KOs) calculated with the KofamKOALA[7] algorithm as an alternative means to the more commonly used Blastp homology comparisons. As shown in Table 8 both approaches agree in the classification of most homologous proteins, but the KO comparisons allowed a better separation of sequences into functional groups. With a more detailed identification of the functional characteristics in the proteomes of all species analyzed, it was possible to discern unique functionalities in the sugar-secreting algae that are likely related to that desirable trait.

The predicted proteomes of *Chlamydomonas reinhardtii, C. sorokiniana, M. conductrix* and *S. glucoliberatum* used in our previous studies[46] were re-analyzed. *C. reinhardtii* and *C. sorokiniana* were chosen as non-sugar secreting controls due to their phylogenetic proximity to *S. glucoliberatum* and *M. conductrix,* respectively. This relationship facilitates the elimination of most KOs that are present in both sugar-releasing algae, which are also present in their closely related controls but do not participate in the sugar secretion. The effectiveness of this control is shown by the small number of common KOs among phenotypic groups compared to the large number of common KOs shared by all species.

Gene products likely related to sugar secretion pathways in *M. conductrtix* and *S. glucoliberatum*

We did not observe any evident specialization in the sugar-secreting algae with respect to transporters or membrane trafficking molecules, in terms of increased abundance of unique functions compared to the control group (Figure 11). However, *S. glucoliberatum* PABB004 has a unique functionality not found in any other algae tested which correlates with its advantageous glucose secretion phenotype. This function is the K08175, a Na$^+$ dependent glucose transporter, represented by a single copy gene in *S. glucoliberatum* PABB004's genome (locus tag identifier HT031_006605). The lack of gene duplications strengthens the classification of this sequence in its orthologous group. The predicted gene product of HT031_006605 is the protein identified with the NCBI accession number KAF8055830.1. KAF8055830.1's function was predicted based on BlastP searches against the Uniprot-SwissProt and Pfam databases which found similarities with domains in the "Major Facilitator Superfamily", "Protein prenyltransferase alpha subunit repeat", "putative chemical substrate binding pocket" and "Fungalysin metallopeptidase (M36)" protein families but could not predict a unique function; therefore, KAF8055830.1 is annotated as a hypothetical protein. The high gene expression level of this gene compared to other sugar, ion and glycerol transporters, which was estimated with the genome-guided transcriptome assembly of *S. glucoliberatum,* supports KAF8055830.1's key role in the glucose secretion phenotype.

Additionally, the pathway completeness analysis performed comparing the unique KOs from the sugar-secreting algae (n = 402) and the pooled set of KOs from the control group (n = 3841) showed three pathways overrepresented in the much smaller set. This type of comparison highlights differences among phenotypes and gives a metabolic framework to infer their biological meaning. The three pathways overrepresented in the unique KOs from sugar-secreting algae are "Glycosaminoglycan degradation", "Glycosphingolipid biosynthesis" and "Regulation of lipolysis in adipocytes". The functions in these pathways that are present in the sugar-secreting algae but absent from the controls are of special interest.

One of such functions, K01230, represented by the protein A0A2P6V0S3_9CHLO in *M. conductrix,* describes glucosidases activated at low pH levels. These enzymes have shown to play a role in lysosome-mediated degradation. Since maltose and glucose secretion in *M. conductrix* increases at lower pH and the vesicle-mediated secretion is a possible alternative to membrane transporters for the secretion of sugars, this enzyme could be one of the missing pieces on the maltose and glucose secretion pathways of *M. conductrix*[58].

Transcriptomic analysis

A second part of this work compared the de novo assembled transcriptomes of two sugar secreting algae (*S. glucoliberatum* and *C. sp.* PABB002) and four non-sugar secreting strains. To make this comparison, the transcripts were translated to peptides. Peptides without homologous sequences in the Pfam or Uniprot-SwissProt databases were eliminated. The remaining sets were annotated with KO identifiers and compared.

Interestingly, *S. glucoliberatum*, *C. sp.* PABB002 and *C. sphaericum* UTEX 1354 de novo assembled transcriptomes showed a lower ratio of complete/internal transcripts compared to the other species, suggesting a more complex splicing.

The functional analysis of *S. glucoliberatum* transcripts evidenced new genes not predicted in the first genome annotation of the organism. None of the 9 KOs shared by the two sugar secreting algae and absent from their controls were present in the annotated proteome of *S. glucoliberatum* PABB004 (Table 12) and only 8 of the 11 KOs that were unique to its transcriptome were found in its predicted proteome (Table 13). The implications of this finding are certainly an interesting area of study to fine tune the gene prediction algorithms for these algae. Moreover, the transcriptomic evidence suggests that *S. glucoliberatum*'s small but highly compact genome has a bigger set of interesting molecular tools to be discovered which might shed more light on its phenotypic traits.

Experimental procedures

Microbial strains and genomic data sources

The genome and predicted proteome from *Scenedesmus glucoliberatum* PABB004 was downloaded from the NCBI Genome database (assembly accession GCA_014905635.1 UMN_S.PABB004_v1). Predicted proteomes from *C. sorokiniana* and *M. conductrix* were downloaded from the Swissprot- Uniprot database (accession numbers UP000239649 and UP000239899). Protein annotation data from *C. reinhardtii* was downloaded from the KEGG database.

Coelastrum sphaericum UTEX 1354, *Coelastrum proboscideum var. gracile* UTEX 184, *Scenedesmus naegelii* and *Scenedesmus dimorphus* were purchased from the UTEX Culture Collection. *Scenedesmus glucoliberatum* PABB004 and *Coelastrum* sp. PABB002 were isolated from *Paramesium bursaria* enriched cultures as described previously [46].

Annotation of proteomes with KEGG identifiers

The proteomes from *C. sorokiniana, M. conductrix* and *S. glucoliberatum* PABB004 were re-annotated with the BlastKOALA and/or KofamKOALA online tools from the KEGG platform. Annotation outputs with their respective KEGG identifiers (also known as KOs) were organized into biochemical pathways and protein hierarchies (BRITE hierarchies) using the Reconstruct Pathway tool.

For comparative analyses, the non-sugar releasing algae set ("control") was formed by joining the proteomes (and respective KOs) from *C. sorokiniana* and *C. reinhardtii*. The sugar releasing algae set (“SRA”) was built by joining the proteomes (and KOs) from *S. glucoliberatum* and *M. conductrix.*

KO set comparisons

All comparisons were based on the abundance or the presence/absence of selected KOs in the specific genomic or transcriptomic sets. Abundance quantification was performed by adding the number of KOs within a pathway or BRITE category. The presence/absence comparison was used to find the unique KOs for each set of interest. The quantification of unique KOs was favored over the quantification of the total number of genes annotated with a specific KO to control for the different genome sizes among the species.

Pathway completeness analysis

The relative pathway completeness (RPC) for a set X compared to the set Y and a pathway *i* was calculated as the number of unique KOs predicted for set X in pathway *i* divided by the number of unique KOs predicted for set Y in pathway *i.*

RPC values were calculated for all pathways in the SRA exclusive set with respect to the complete control set. Pathways with RPC greater than 1.5 were considered to be overrepresented with respect to the control set.

Sequence conservation analysis

The sequence similarity of proteins from *M. conductrix* and *S. glucoliberatum* was tested using a local alignment approach (BlastP) with default settings. These results were compared with previous clustering results of the complete proteomes of *M. conductrix, C.*

sorokiniana, C. reinhardtii and *S. glucoliberatum* PABB004 [46] and with the functional clustering obtained by the KEGG annotation tools.

RNA libraries preparation and sequencing

Algal biomass from all six *Scenedesmus* and *Coelastrum* strains was collected from growths in tubular bioreactors, flash frozen and kept at -80°C until RNA extraction. RNA extraction and sequencing was performed as described previously [46,58]. A total of 6 dual-indexed Illumina TruSeq stranded mRNA libraries were created, pooled, and sequenced on a NextSeq Mid-output 2x75-bp flow cell by the University of Minnesota Genomics Center.

Transcriptome assembly and peptide search

S. dimorphus, S. naegelii, S. glucoliberatum, C. esphaericum, C. proboscideum and *C. sp.* PABB002's transcriptomes were assembled de novo using Trinity version 2.9.1 with the -jaccard_clip option. All de novo assembled transcripts were translated into peptides using the 6 possible open reading frames with Transdecoder version 5.5.0. The longest peptides for each transcript were searched for homology support against the Uniprot-SwissProt database using BlastP and an e-value of 1e-5 and for predicted domains using HMMER 3.1 and the Pfam database. Peptides supported by any of these searches were included in the final peptide set for each organism. All peptides were annotated using KofamKOALA as described above.

To quantify the expression of genes in *S. glucoliberatum,* and map it to the predicted proteome, its transcriptome was assembled using TopHat version 2.0.13 and Cufflinks version 2.2.1, with the genomic assembly and the published annotation file as templates.

Software

Data analysis was carried out with RStudio (RStudio Team (2021), PBC, Boston, MA URL http://www.rstudio.com/) and image production was done with Affinity Designer version 1.9/1.10 (Serif, Nottingham, UK).

Chapter 4: Preliminary Work on the Genomic and Proteomic Analysis of Bacterial Plastic Degradation

Synopsis

Plastic materials are accumulating in the environment due to their massive use, inadequate disposal methods, and long durability. Conventional waste management strategies are recycling and incineration, but they have been insufficient due to their complex implementation and toxic byproducts. In contrast, biorremediation of environmental pollutants offers the advantages of few input requirements, high adaptability, and safer application. Therefore, biodegradation of plastic waste is an alternative avenue to tackle this environmental issue.

Few organisms have been recently reported to degrade common plastic materials, but their slow degradation rates hamstring their use. Additionally, information about the molecular machinery that enables this process is very limited. Hence, a better understanding of the plastic degradation pathways is required to enhance the natural microbial metabolic characteristics and fully harness their degradative potential.

This work aims to examine plastic biodegradation at the molecular level through the genomic characterization and proteomic analysis of plastic-dwelling organisms. Success of this work will open new avenues for future studies of particular proteins and further improvement of the microbial strains and their plastic degradation toolkit.

Introduction

Plastic materials are a group of organic polymers generally characterized by their water barrier properties, high stability, cheap production, easy molding and light weight. They can be produced by the polymerization of synthetic and relatively simple monomers derived from fossil hydrocarbons or from biological sources[79]. Although plastic materials were already used in the early 20th century, a widespread production begun only until 1930s - 1950s[80]. Since then, plastics have become fundamental in the packaging, textile, automotive, construction and electrical industries, only to name a few, where they have replaced natural materials such as wood, paper and glass. While the high stability of plastics is desired from an engineering perspective, it also facilitates their accumulation in the environment due to their negligible degradation rates.

The demand for plastics has increased with the economic growth of middle and high income countries and the greater use of products such as disposable containers, clothing and electronics[81]. Recent estimates report that only 24 - 30% of all plastics made between 1950 and 2015 are currently in use; from the remaining 70%, approximately 12% has been incinerated and 9% recycled, resulting in the waste and accumulation of approximately 60% of all plastics ever made[79]. The evidence gathered to date has increased awareness on plastic accumulation and pollution, catalyzing various social, political and research initiatives[82].

Biocatalytic conversion of plastics is an alternative pathway to the existing plastic waste management methods (i.e., recycling and incineration). Microorganisms adapted to degrade stable natural polymers such as cutin could degrade plastics with similar chemical structures like polyurethanes and polyesters. It has been shown that microbes involved in the catabolism of lignin, a complex polyaromatic component of wood, can degrade synthetic polymers with highly stable C-C bonds after oxidation pretreatments[83]. In recent years, numerous bioprospecting efforts have found and characterized new microorganisms with plastic degrading capabilities, mostly for polyethylene terephthalate (PET)[84] and polyurethane (PU)[85]. However, the reaction rates for biological degradation of plastics are generally low, being among the most favorable a 0.02% w/w loss per hour at 70°C. Thus, strategies to enhance the native degradative processes through protein engineering[86] and co-culture have been explored[87] with some success in the degradation of PET and PU. Other plastic materials, such as polyethylene (PE), polypropylene (PP) and polystyrene (PS) have been long considered highly resistant to biodegradation.

Material properties of interest

The performance and degradability of plastic materials are determined by the type of constituting monomers, the spatial arrangement of the substituent groups, the polymer chain length, among others[88]. For instance, the spatial distribution of substituent groups and level chain branching determine the crystallinity of the end material. A disordered orientation of substituent groups or highly branched molecules inhibit the formation of crystalline regions, thus forming amorphous materials. The loose molecular organization of amorphous plastics favors the access of microbes and enzymes, making it more prone to biodegradation. In contrast, highly crystalline plastics are recalcitrant to microbial degradation.

Another important parameter is the polymer chain length. There is a direct correlation between chain length and material properties. In general, shorter hydrocarbon chains have more flexibility and kinetic freedom, so at normal temperature and pressure conditions they form gases if they have less than 4 carbon atoms per molecule, and liquids if they have from 5 to 25 carbon atoms per molecule. Hydrocarbons with medium size chains (between 25 and 50 carbon atoms per molecule) form weak solids such as paraffin waxes and longer molecules form stronger materials such as polyethylene.

Degradation routes of conventional plastics

Plastic degradability is understood differently from the engineering, environmental and biological perspectives. From the material engineer or consumer perspective, “degradation” implies the decline of properties that are needed for adequate use. Therefore, a higher degradability causes a faster disposal of plastic products. From an environmental and biological perspective, plastic degradation also involves the breakdown of polymer chains to form smaller chemicals or its complete oxidation to CO_2[89]. In this work, I focus on the later definition of plastic degradation.

Plastics in the environment are exposed to UV radiation, extreme temperatures, oxidants, physical stress and biological attacks, which contribute to weathering and degradation. In practice, all such factors participate collaboratively in plastic degradation processes. For instance, exposure to UV light causes chemical changes on the polymers making them brittle and more susceptible to further fragmentation. In the environment, bigger fragments will progressively degrade to smaller fragments, eventually forming micro and nanodebris. Smaller fragments are more likely to pass through cell membranes, increasing the rate of intracellular catabolism[90]. To facilitate the study of these different degradation routes, it is convenient to classify their key factors into abiotic and biotic stressors.

Abiotic degradation

All conventional polymers can be degraded under mechanical stress or extreme temperature conditions. However, under regular environmental settings with moderate temperatures, solar radiation and oxidizing atmospheres, the most common abiotic degradation pathways are photodegradation and hydrolysis.

Plastics with C-C backbones (PE, PP, PS and PVC) are theoretically resistant to photo-initiated oxidative degradation, because the presence of unsaturated groups that absorb light energy is required. However, impurities and structural abnormalities enable radical formation when UV radiation breaks C-H bonds in the polymer. Degradation continues with a propagation phase. Newly formed radicals react with oxygen forming highly reactive hydroperoxides that attack other polymer chains, leading to autoxidation. Propagation results in chain scission and crosslinking. The process terminates when two radicals combine and form inert products. Resulting compounds are shorter or branched polymer chains, olefins, and increased number of oxygen-containing groups such as aldehydes and ketones. These chemical changes make the material brittle and more susceptible to fragmentation, which in turn increases the surface area available for abiotic and biotic degradation[89].

Plastics containing heteroatoms in the main chain such as PET and PU are more prone to degradation in environmental conditions. Like C-C backbone polymers, PET and PU are susceptible to photo and photo-oxidative degradation. Additionally, PET and PU undergo hydrolysis of their C-O and C-N bonds in the polymer backbone. In aqueous environments the hydrolysis of ester bonds in PET occurs slowly, nonetheless it is the main cause of its low temperature degradation. Hydrolysis of PET results in chain scission and formation of terminal carboxyl and alcohol groups. Similarly, nitrogen containing groups in the backbone of PUs are prone to hydrolysis, but at slower rates compared to the ester bonds[89].

Biological degradation

General mechanism of plastic biodegradation

Biodegradation is as a complex process that requires several intra and extracellular stages and in some cases the collaborative action of several organisms in a microbial community[87]. It involves the polymer break down by microorganisms, its conversion to oligomers and monomers, assimilation, and in some cases its complete oxidation (i.e. mineralization). The process can produce intermediate chemicals such as MHET and BHET[84] or completely oxidize the polymer the substrate producing CO_2 and H_2O in aerobic environments, or CO_2, H_2O and CH_4 in anaerobic conditions.

Current understanding of the general plastic degradation process involves four stages: adsorption, depolymerization, assimilation and catabolism/mineralization[91]. The first stage is the adsorption of organisms or extracellular enzymes to the debris surface. Hence, any process that facilitates such adsorption is required to initiate biodegradation. Biosurfactants such as glycolipids or amphiphilic proteins may play an important role on facilitating adsorption processes. Then, plastics' long polymeric molecules are broken down by enzymes into smaller fragments, this step is known as depolymerization. Shorter chain products are then internalized to the cytosol (assimilation) where they can enter cellular catabolic pathways and be utilized as carbon sources. Membrane bound proteins and transporters are key elements of the internalization step. Finally, a complete catabolism of plastics to CO_2, H_2O, salts and CH_4 (if anaerobic) terminates the biodegradation process, this step is termed mineralization.

Microbial Strains with Plastic Degrading Abilities

Several fungal and bacterial strains have been reported to degrade plastic substrates. Hydrolysable bonds in PET and PUs are particularly susceptible to microbial attack but the high stability of C-C bonds make homochain polymers recalcitrant to biodegradation. It has been observed that organisms and enzymes capable of degrading weathered C-C plastics (PE, PP and PS) are generally related to lignin degradation[92].

Important enzymes that participate in degradation processes are laccases (EC 1.10.3.2), manganese peroxidases (MnP) (EC 1.11.1.13) and lignin peroxidases (LiP) (EC 1.11.1.14). Additionally, alkane hydrolases (AH) (EC 1.14.15.3) and hydroquinone peroxidases[93] have been reported to degrade low molecular weight PE[83].

To date, most reports on enzymatic degradation of plastics describe hydrolases involved in the depolymerization of PU and PET. PU can be depolymerized by microbial proteases, ureases (EC 3.5.1.5), esterases (EC 3.1.1.1) and cutinases (EC 3.1.1.3). PET can be degraded by fungal lipases, esterases and cutinases as well as bacterial carboxylestearases[83]. Fungal and bacterial cutinases present an active site close to the enzyme surface which facilitates the recognition and interaction with polymeric substrates and produce a higher activity than lipases[94]. Nonetheless, the activity of all these native enzymes on synthetic polymers is low, as would be expected given their selection bias towards the utilization of natural polymers such as cutin and lignin.

Limitations of plastic biodegradation

In practice, conventional plastics (PE, PVC, PS, PP, PET) have extremely slow or non-observable biodegradation rates. This is due to the limitations imposed by material properties, the lack of capable microbes and conducive environmental conditions. Important factors that limit plastics' degradability from the material perspective are their high hydrophobicity, smooth surface topography, low surface/volume ratio, high degree of crystallinity, high molecular weight, and lack of reactive functional groups on their backbones.

In aqueous environments the adsorption of organisms and enzymes is hampered by plastics' hydrophobic surfaces. Similarly, smooth surface topographies hinder microbial accumulation and biofilm formation. Additionally, the low surface to volume ratio of bulk plastic products restricts microbial and enzymatic adsorption.

The specific polymer type and characteristics also influence other stages of the degradative process. It has been observed that depolymerization is more favorable in the amorphous regions of plastic materials compared to the more crystalline parts. This is explained by the restricted access of extracellular enzymes to the tightly organized crystalline areas[95]. Likewise, the lack of hydrolysable functional groups in the polymer backbone prohibits an effective interaction with degrading extracellular enzymes[83].

The high molecular weight of polymer molecules hampers their internalization into the cytosol for consumption. It has been shown that oligomers of plastic repeating units can be metabolized better than long polymeric chains of the same monomer for PS[96] and PE. Furthermore, the products of polymer metabolism might also inhibit enzymatic activity and decrease the rate of reaction. Altogether, these explain why plastic biodegradation usually requires a previous abiotic aging phase for initiation and overall is a slow metabolic process.

This work presents a set of organisms isolated from environmental soil, water and plastic waste samples and explores the methodology to characterize and uncover key molecules used for the biodegradation of paraffin as a substitute to PE.

Preliminary Results

Microbial diversity of plastic-dwelling organisms

An initial identification of 30 culturable bacterial strains isolated from a variety of plastic waste samples was performed using their SSU rRNA gene similarity to reference sequences from the SILVA database. As shown in Figure 12 all bacterial isolates belong to four distinct phyla. Actinobacteria with 18 isolates is the most abundant phylum, followed by Proteobacteria with 9 isolates, while Bacteroidetes and Firmicutes have only 2 and 1 representatives, respectively. The relative presence of each group is not a good measure of the true diversity of plastic-debris dwellers since these results are biased by our culture method and isolation approach. However, the presence of organisms from three different clades long ago diverged within the Bacteria kingdom, i.e., Terrabacteria group (Actinobacteria and Firmicutes), Proteobacteria and the FCB group (Bacteroidetes), depicts the wide diversity of organisms that can be found inhabiting on plastic waste surfaces.

Genome sequencing of selected strains

Strains from various phylogenetic clades were selected for genomic sequencing based on the existence of genomes from closely related strains in publicly available databases and on their apparent growth rate on PVP/PVA plates. A total of 17 genomes were sequenced, 6 from bacteria isolated previously from environmental samples and 11 from this isolation effort. All genomes were published in the NCBI database under the bioproject with accession number PRJNA610894. The strains and genome assembly accession numbers are presented in Table 15.

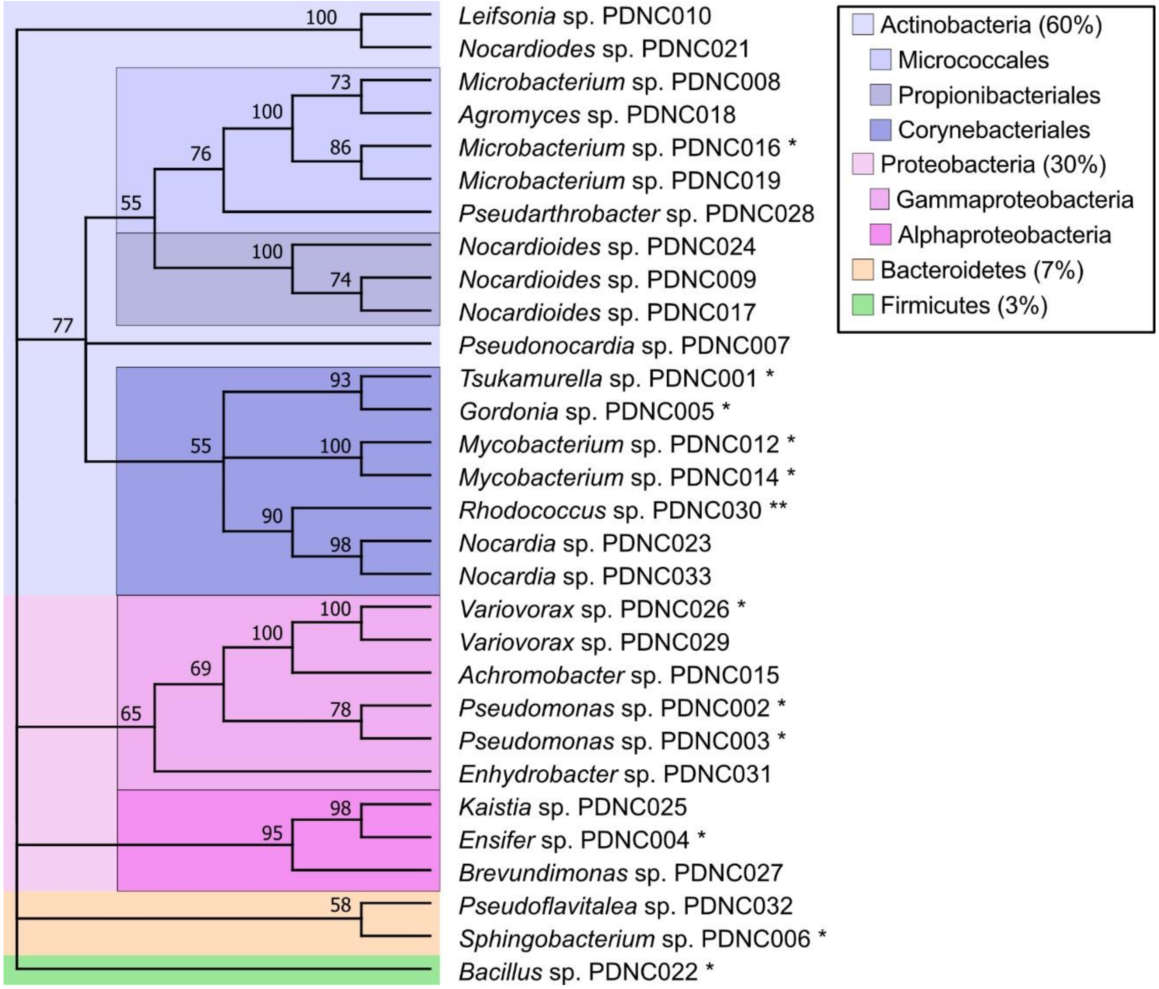


Figure 12. Phylogenetic tree of isolated plastic-dwelling bacteria

The evolutionary history of these isolates was inferred from their SSU rRNA sequences using the Maximum Likelihood method and Tamura 3-parameter model. The bootstrap consensus tree was inferred from 100 replicates. It is taken to represent the evolutionary history of the taxa analyzed. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates) are shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Tamura 3 parameter model, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (2 categories (+G, parameter = 0.3579)). This analysis involved 30 nucleotide sequences. There was a total of 838 positions in the final dataset. Evolutionary analyses were conducted in MEGA X[73]. (*) Organisms selected for genome sequencing are marked with an asterisk. (**) The genomes of three other *Rhodococcus* species previously isolated were sequenced.

Table 15. Bacterial genomes assembled in this study

Strain	Assembly accession
Bacillus sp. PDNC022	GCA_016919285.1
Bradyrhizobium sp. PSBB068	GCA_016839165.1
Ensifer sp. PDNC004	GCA_016919405.1
Gordonia sp. PDNC005	GCA_016919385.1
Microbacterium hominis PDNC016	GCA_016919305.1
Mycolicibacterium boenickei PDNC014	GCA_016919325.1
Mycolicibacterium septicum PDNC012	GCA_016919345.1
Mycolicibacterium septicum PSBB070	GCA_017052695.1
Pseudomonas aeruginosa PDNC003	GCA_016919425.1
Pseudomonas sp. PDNC002	GCA_016919445.1
Rhodococcus aetherivorans PSBB011	GCA_016839185.1
Rhodococcus sp. PSBB049	GCA_017068295.1
Rhodococcus sp. PSBB066	GCA_017068255.1
Shinella sp. PSBB067	GCA_016839145.1
Sphingobacterium siyangense PDNC006	GCA_016919365.1
Tsukamurella tyrosinosolvens PDNC001	GCA_016919465.1
Variovorax sp. PDNC026	GCA_016919265.1

Selection of plastic substitute substrates for bacterial growth

Due to the slow growth rate of bacterial strains on polyethylene, other substrates that resemble some of polyethylene's properties were tested for potential use as carbon sources to the bacterial isolates. PVA with average molecular weights of 70000 and 30000, PVP with average molecular weights of 40000 and 10000, and paraffin wax were tested.

None of the isolates grew to high densities in any of the soluble polymers (PVP or PVA) either in pure culture or in co-culture schemes. It was observed a general trend of increased pH in the media for all isolates except *Bacillus sp.* PDNC022 when cultivated on PVP. Adding extra buffering to the media controlled the pH increase but did not significantly improve strain growth. Moreover, the polymer size distribution observed with gel permeation chromatography did not show evident changes after being incubated with the organisms.

The possibility of using solid hydrophobic substrates such as paraffin and bee's wax as an alternative to polyethylene was assessed with the isolate *R. aetherivorans* PSBB011. As shown in Figure 13, *R. aetherivorans* PSBB011 grew faster and denser using bee's wax or paraffin compared to polyethylene. Therefore, paraffin was chosen as a substitute substrate to polyethylene for further experiments. The strain reached stationary phase after 3 days of growth, having duplicated its population 5 times (5 generations).

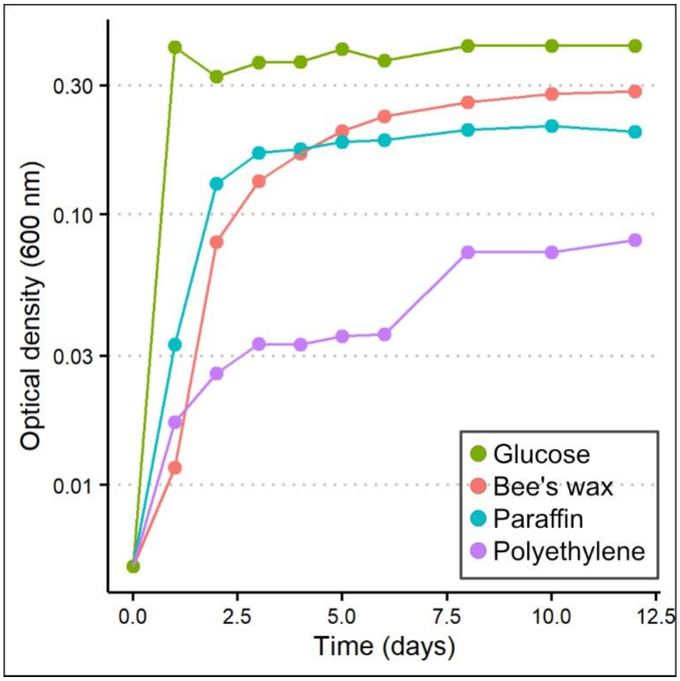


Figure 13. R. aetherivorans PSBB011's growth on solid hydrophobic carbon sources

R. aetherivorans PSBB011 was cultivated using different carbon sources. Paraffin wax, bee's wax and polyethylene were supplemented at 0.7 g/L. The growth on glucose (2 g/L) is used to verify the strain viability).

Exploratory growth of selected strains using paraffin as sole carbon source

Strains *T. tyrosinosolvens* PDNC001, *P. sp.* PDNC002 and *P. aeruginosa* PDNC003 were selected to characterize their growth using paraffin as sole carbon source. As shown in Figure 14 all selected strains increased their cell density over time. *T. tyrosinosolvens* PDNC001 doubled its population 5.5 times in 3 days while *P. aeruginosa* PDNC003 doubled its population 3.2 times over the course of the experiment (8 days) and *P. sp.* PDNC002 doubled its population 3 times over 8 days as well. This indicates that the fastest growing organism in the tested set is *T. tyrosinosolvens* PDNC001. *T. tyrosinosolvens* PDNC001's growth rate is comparable to that of *R. aetherivorans* PSBB011 in the conditions tested.

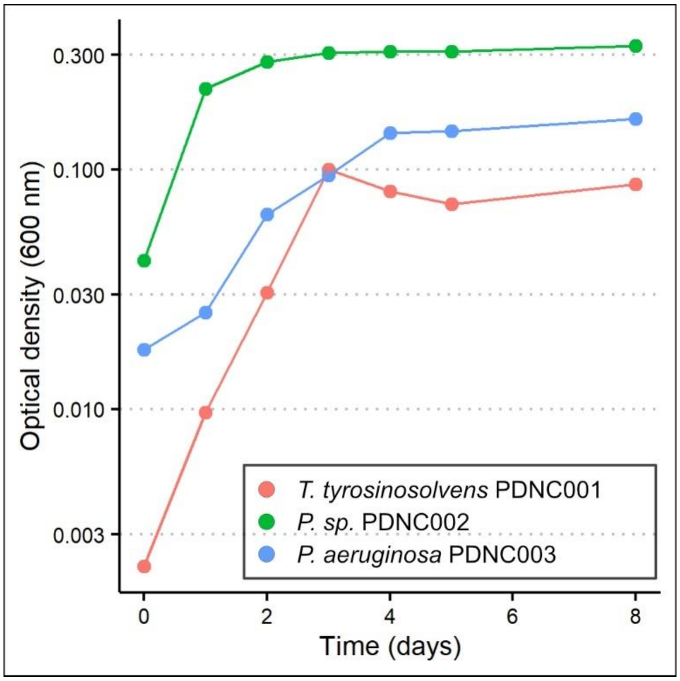


Figure 14. Growth of selected strains using paraffin as sole carbon source

Growth of *T. tyrosinosolvens* PDNC001 using paraffin as sole carbon source

T. tyrosinisolvens PDNC001 is an aerobic actinobacteria previously reported to degrade paraffin and oils[97,98]. Its genome is 4,6 Mbp long and was assembled in 1 contig. It contains 71.5% GC pairs and 4430 protein coding sequences. *T. tyrosinosolvens* PDNC001 has a hydrophobic surface and when grown in liquid media clumps at higher culture densities. When cultivated using paraffin as sole carbon source, a doubling time of 5.6 h during first 30 h (exponential growth) was calculated from four replicate growths (Figure 15).

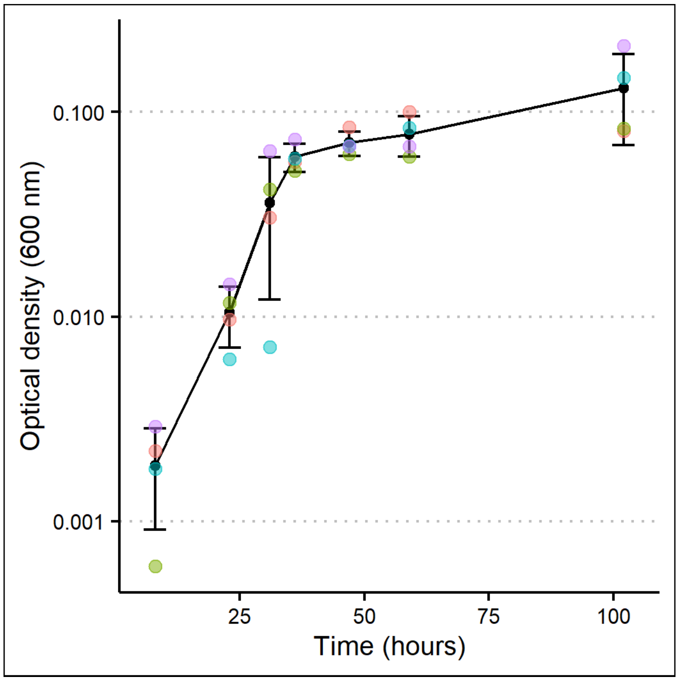


Figure 15. T. tyrosinosolvens PDNC001 growth using paraffin as sole carbon source

T. tyrosinosolvens PDNC001 was grown using paraffin as sole carbon source. Sterile, hot, liquid paraffin was dispersed on the growth flask surface to increase the area available to the bacteria. The data shown are the culture's optical densities at 600 nm wavelength for the first 100 hours of growth. The black points and error bars represent the mean and standard deviation calculated from four replicate experiments (colored points).

Characterization of substrates by gas chromatography

The gas chromatography profile of the paraffin substrate shows sharp peaks that elute between 15 and 23 minutes (Figure 16). The gas chromatography profile of the filtered growth media (after separating the organisms and paraffin particles) showed peaks that elute at shorter retention times than those from the paraffin peaks, this is, between 7.5 and 15 minutes. This suggests that molecules with lower molecular weight than those given with the paraffin substrate are present in the soluble fraction of the media. There are three main peaks with elution times of 7.5, 11 and 13 minutes in the filtered growth media. Furthermore, when the filtered growth media (without paraffin or organisms) was incubated with fresh paraffin in vitro a different chromatographic profile was observed (Figure 16). This profile did not show three prominent peaks but instead several small peaks that elute from 9 to 15 minutes. The prominent peaks found on the growth media were not evident in the in vitro reaction.

Proteomic analysis of *T. tyrosinosolvens* PDNC001 secretome

The existence and identification of proteins secreted to the media by *T. tyrosinosolvens* when growing with paraffin as sole carbon source was assessed with mass spectrometry. The culture media was filtered to separate the microorganisms and paraffin debris and lyophilized. The putative proteins in the sample were digested with trypsin and their MS/MS profiles compared to the predicted proteome from *T. tyrosinosolvens* PDNC001 using MaxQuant[99].

Proteins were identified using peptides that aligned perfectly to the protein sequence and that were identified by MS/MS with at least two evidence spectra. The total set of proteins identified in the sample after filtering out contaminants contains 80 proteins. Among them 51 proteins were predicted to have a signal peptide. This set of 51 proteins are assumed to be secreted by the organisms when grown using paraffin as sole carbon source. The 15 proteins with the highest relative abundance, after normalizing by the number of peptides identified per protein are shown in Table 16.

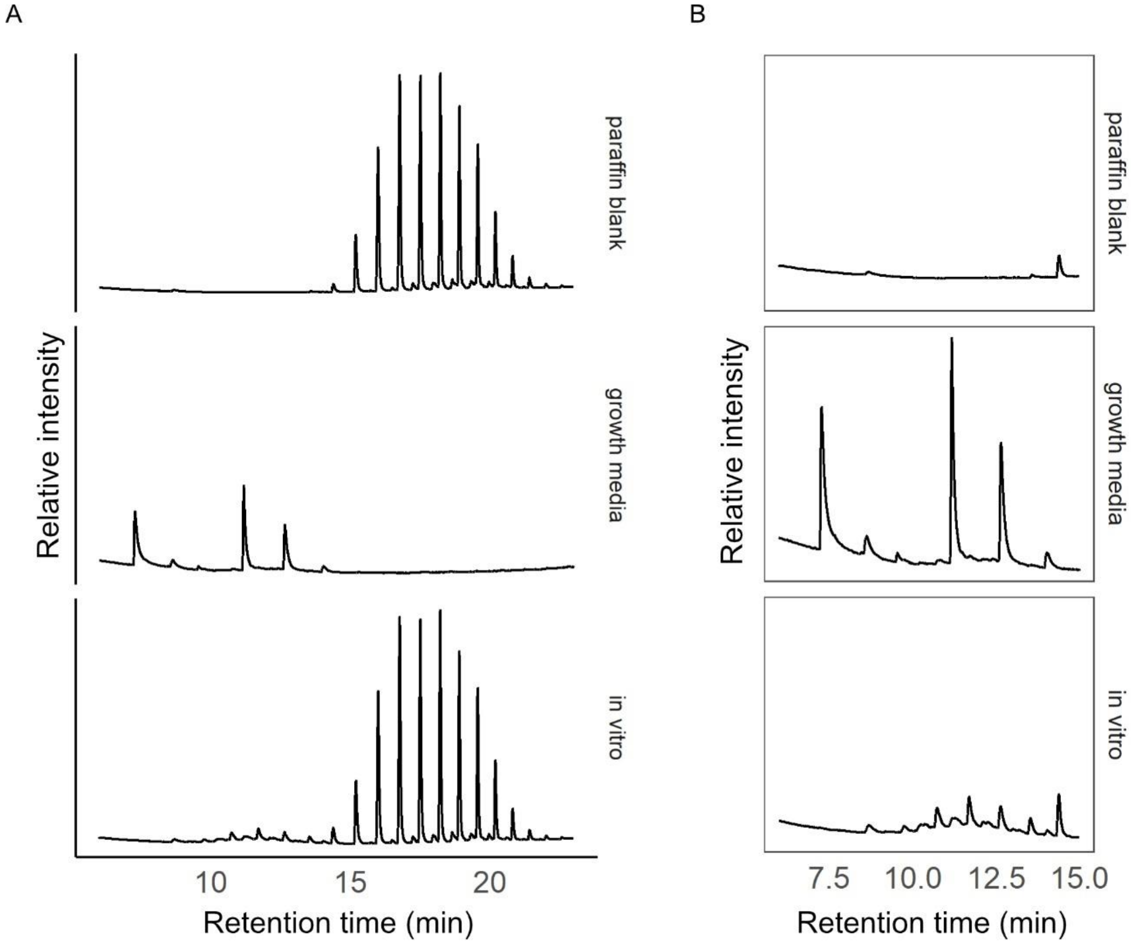


Figure 16. Gas chromatogram profiles of fractions from T. tyrosinosolvens growth using paraffin as sole carbon source

Panel A shows the full gas chromatography profile of the paraffin used as substrate (top), the filtered growth media (after separating the organisms and paraffin particles, middle) and the filtered growth media after in vitro incubation with fresh paraffin (bottom). Panel B shows in detail the elution peaks between 6 and 15 minutes from the chromatograms in A.

Table 16. Proteins identified from T. tyrosinosolvens growth on paraffin

ID	Description	Unique peptides	Intensity per peptide
QRY85577.1	esterase family protein	6	4.74E+08
QRY84254.1	N-acetylmuramoyl-L-alanine amidase	6	3.18E+08
QRY82583.1	hemophore-related protein	3	2.16E+08
QRY86593.1	lipoprotein LpqH	1	2.03E+08
QRY83621.1	hemophore-related protein	4	1.81E+08
QRY83611.1	hypothetical protein JVY00_17390	6	1.71E+08
QRY82898.1	DUF4822 domain-containing protein	3	1.48E+08
QRY84432.1	lysozyme	2	1.3E+08
QRY82630.1	prolyl oligopeptidase family serine peptidase	1	1.29E+08
QRY85555.1	SpoIID/LytB domain-containing protein	6	1.08E+08
QRY86105.1	transglycosylase family protein	1	1.06E+08
QRY83235.1	hypothetical protein JVY00_15275	1	89615000
QRY85875.1	zinc ABC transporter substrate-binding protein	6	83625000
QRY85970.1	hypothetical protein JVY00_07900	1	78874000
QRY83763.1	hypothetical protein JVY00_18245	3	75366667

Among the 15 proteins identified with highest relative intensity per peptide, there are 5 hydrolases, one lipoprotein, metal scavenging proteins and proteins with undefined function. Of special interest are the 5 enzymes with hydrolase activity and the lipoprotein which might have roles in the modification of the paraffin molecules and as surfactants, respectively. This preliminary results need further confirmation.

Experimental procedures

Chemical reagents

All chemicals were purchased through Fisher Scientific (Pittsburgh, PA) or Sigma Aldrich (St. Louis, MO) unless otherwise indicated.

Strains isolation and identification

Soil, water and weathered plastic debris samples were collected from lake and river shore environments in the greater Minneapolis – Saint Paul area during Fall 2019. Samples were incubated in 10 ml of screening media (10 mM potassium phosphate, 2 g/l ammonium sulfate, pH 7) supplemented with 0.5 g/l yeast extract and with a thin strip of solid plastic-like materials (30 % bee's wax, 30% paraffin, 30 % polyethylene average molecular weight 4000 and 10 % polystyrene, all mass percentages). The samples were maintained at 30 °C and 250 rpm shaking until a robust biofilm was observed on the surface of the plastic strip (approximately 3 weeks after inoculation). A small piece of the strip containing microbial biofilm was transferred to fresh screening media supplemented with 0.1 g/l yeast and a new strip with the same composition. Samples were maintained at 30 °C with shaking for 2 additional weeks. Aliquots of the second culture were transferred to agar plates containing poly(vinyl alcohol) (PVA) and poly(vinyl pyrrolidone) (PVP) as sole carbon source. Colonies with different morphologies were transferred onto fresh plates until pure cultures were achieved. Pure strains were preserved in lysogeny broth (LB) with 20% (v/v) glycerol or 1% (v/v) DMSO in water at −80◦C. For an initial strain identification, the SSU rRNA gene was amplified by PCR using the primers BBP3196 (5' – GAACGCTGGCGGCRKGCYTWAYACATGC – 3') and BBP3197 (5' –

TACGGYTACCTTGTTACGACTTM – 3’) and sequenced. Sequences were then compared to the SILVA database (https://www.arb-silva.de/aligner).

Phylogenetic analysis

SSU rRNA sequences were initially aligned using MUSCLE (UPGMA clustering, gap opening penalty -400 and no gap extension penalty). The evolutionary history was inferred using MEGA X [73]. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Tamura 3 parameter model, as suggested by MEGA X Best Model Tool, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (2 categories (+G, parameter = 0.3579)). The bootstrap consensus phylogenetic tree was inferred from 100 replicates. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates were collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates) are shown next to the branches. This analysis involved 30 nucleotide sequences and a total of 838 positions in the final dataset.

Genome sequencing, assembly, and annotation

Cells grown on LB or B-urea plates were used for total DNA isolation using the ZR Fungal/Bacterial DNA Microprep kit (Zymo Research, Irvine, CA, USA) as directed by the manufacturer. DNA quantity and purity was measured with a NanoDrop 2000 spectrophotometer (Thermo Scientific, Waltham, MA, USA). Sequencing was performed by the University of Minnesota Genomics Center from a barcoded multiplexed library using a 8 M 15 hour Sequel II SMRT cell. Genome assembly was carried out using Flye version 2.6 and annotation was done with the PGAP pipeline by the NCBI. All genomes were published in the NCBI genome database under the bioproject PRJNA610894.

Culture conditions with paraffin as sole carbon source

Selected strains were grown on 250 ml of modified Bristol media (150 mg/l potassium phosphate dibasic, 200 mg/l sodium sulfate, 25 mg/l sodium chloride, 10 mg/l ferric ammonium citrate, 250 mg/l sodium nitrate, 150 mg/l ammonium sulfate, 80 mg/l magnesium sulfate heptahydrate, 20 mg/l calcium chloride dihydrate and 0.1 % v/v trace metals solution as described in Villa et al. [100], pH 7) supplemented with 1 g/l of paraffin as a sole carbon source. The walls of the growth flasks were coated with a thin film of paraffin to increase the available surface in contact with the cells. Cultures were kept at 30 °C and 250 rpm shaking for the length of the growth experiment. Cell growth was assessed with the culture optical density at 600 nm. The liquid or biomass fractions were collected from the culture and lyophilized as required.

Gas chromatography

Hydrophobic compounds present in the lyophilized samples were extracted with hexane and analyzed with gas chromatography as described before[25].

Mass spectrometry

Lyophilized culture supernatant from the stationary growth phase (1 week after inoculation) were used for LC-MS/MS analysis. Samples were reconstituted in denaturation buffer (8 M urea, 0.5 M ammonium bicarbonate, pH 8.0 and 4 mM Tris(2-carboxyethyl)phosphine) and incubated at 37 °C for 45 min. To inhibit the reformation of disulfide bonds, thiol groups were alkylated by adding 1 volume of 20 mM iodoacetamide and incubating at room temperature and darkness for 30 min. Trypsin digestion was performed at 37 °C overnight. Samples were desalted and purified with C18 ZipTips (Millipore) and dried in vacuo. Peptides were reconstituted in 20% acetonitrile (ACN) and 0.1% fluoroacetic acid (FA) prior to analysis.

LC-MS/MS data were recorded on a Thermo Scientific Fusion mass spectrometer equipped with a Dionex Ultimate 3000 UHPLC system using a nLC column (200 mm × 75 µm) packed using Vydac 5-µm particles with a 300 Å pore size (Hichrom Limited). Elution was performed with a linear gradient using water with 0.1% FA (solvent A) and ACN with 0.1% FA (solvent B) at a flow rate of 0.3 µl/min. The column was equilibrated with 20% solvent B for 5 min, followed by a linear increase of solvent B to 90% over 32 min and a final elution step with 90% solvent B for 2 min. Mass spectra were acquired in positive-ion mode. Full MS was done at a resolution of 60,000 [automatic gain control (AGC) target, 4 × 105; maximum injection time (IT), 100 ms; range, 300 to 1800 m/z], and data-dependent MS/MS was performed at a resolution of 15,000 (AGC target, 5 × 105; maximum IT, 100 ms; isolation window, 2.2 m/z) using higher-energy collisional dissociation (HCD) energies of 19% with steps of ±4%. Data were processed using Thermo Fisher Xcalibur software and MaxQuant v1.6.10. Proteins were analyzed for the presence of signal peptides using SignalP version 5.0.

Conclusions

This thesis explored different approaches to define protein characteristics and the molecules related to specific phenotypes of interest through sequence comparisons.

The local alignment of orthologous proteins, such as the various bacterial wax ester synthases mentioned in Chapter 1, helped define highly conserved residues along Ma1 sequence that have important roles in its activity and substrate specificity. This sequence comparisons were performed using the BlastP[3] algorithm.

In chapters 2 and 3 much larger datasets were analyzed to find key molecules related to sugar secretion pathways in green algae. To this end, the protein and peptide sequences encoded in the genomes and transcriptomes from sugar-secreting and non-sugar secreting algae were compared. Similarity comparisons based on local alignments were less successful in clustering protein sets in functional orthology groups than the ortholog mediated HMM[7] prediction of gene function. The second approach allowed the selection of a few genes that likely participate in the sugar secretion pathways of *S. glucoliberatum* PABB004 and *M. conductrix* SAG 241.80. Finding these sequences had not been possible in previous attempts using similarity-based genomic comparisons and gene expression analysis.

The work described in chapters 2 and 3 highlights the importance of considering the limitations of homology inferences based on sequence similarity when aiming to predict the genetic basis of phenotypic traits. This must be considered when analyzing large sets of data that include paralog and orthologous genes.

Finally, chapter 4 is the initial exploration of plastic biodegradation. This chapter focuses on gathering microbial strains with potential biodegradative capabilities and defining their genomic sequences. The methodology to characterize their phenotype and molecular toolkit was tested with a promising isolate: *T. tyrosinosolvens* PDNC001. Proteins with predicted hydrolase and biosurfactant activity were found by comparing the peptides identified by LC-MS/MS and its predicted proteome.

All bioinformatic analyses performed in chapters 2, 3 and 4 aim to find genes that could be used in different industrial and environmental applications. The path from an observed phenotypic trait to its application is long and requires many steps. It starts with finding strains that possess interesting phenotypic traits. The characterization of such traits requires establishing adequate methodologies (growth conditions, biochemical assays, etc.). Then, engineering approaches can be used to improve such strains or to better understand the molecules involved.

This thesis contributes to several long-term projects in the Barney laboratory that are at different development stages. From the isolation of microorganisms with potential plastic degradation capabilities and the exploration of alternative methodologies for their characterization, to the analysis of genomic and transcriptomic data searching for molecules likely related to a specific phenotype and the targeted manipulation of Ma1 to better understand its characteristics, all projects aim to benefit our environment by discovering alternative sources of fuels and chemicals or by searching for better ways to manage our waste.

References

1. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).

2. Pearson, W. R. An introduction to sequence similarity ('homology') searching. *Curr. Protoc. Bioinforma.* (2013). doi:10.1002/0471250953.bi0301s42

3. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

4. McGinnis, S. & Madden, T. L. BLAST: At the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* **32**, (2004).

5. Eddy, S. HMMER user's guide: biological sequence analysis using prole hidden Markov models. (2020).

6. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27–30 (2000).

7. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).

8. Fang, G., Bhardwaj, N., Robilotto, R. & Gerstein, M. B. Getting started in gene orthology and functional analysis. *PLoS Comput. Biol.* **6**, (2010).

9. Calixto Mancipe, N., Mulliner, K. M., Plunkett, M. H. & Barney, B. M. Canvasing the Substrate-Binding Pockets of the Wax Ester Synthase. *Biochemistry* (2022). doi:10.1021/ACS.BIOCHEM.2C00076

10. Kalscheuer, R. *et al.* Neutral lipid biosynthesis in engineered Escherichia coli: Jojoba oil-like wax esters and fatty acid butyl esters. *Appl. Environ. Microbiol.* **72**, (2006).

11. Röttig, A., Wenning, L., Bröker, D. & Steinbüchel, A. Fatty acid alkyl esters: Perspectives for production of alternative biofuels. *Applied Microbiology and Biotechnology* **85**, (2010).

12. Kalscheuer, R. & Steinbüchel, A. A novel bifunctional wax ester synthase/acyl-CoA:Diacylglycerol acyltransferase mediates wax ester and triacylglycerol biosynthesis in Acinetobacter calcoaceticus ADP1. *J. Biol. Chem.* **278**, 8075–8082 (2003).

13. Busta, L. & Jetter, R. Moving beyond the ubiquitous: the diversity and biosynthesis of specialty compounds in plant cuticular waxes. *Phytochemistry Reviews* **17**, (2018).

14. Nelson, D. R., Fatland, C. L., Buckner, J. S. & Freeman, T. P. External lipids of adults of the giant whitefly, Aleurodicus dugesii. *Comp. Biochem. Physiol. - B Biochem. Mol. Biol.* **123**, (1999).

15. Teerawanichpan, P., Robertson, A. J. & Qiu, X. A fatty acyl-CoA reductase highly expressed in the head of honey bee (Apis mellifera) involves biosynthesis of a wide range of aliphatic fatty alcohols. *Insect Biochem. Mol. Biol.* **40**, (2010).

16. Greene, R. A. & Foster, E. O. The Liquid Wax of Seeds of Simmondsia californica. *Bot. Gaz.* **94**, (1933).

17. Miwa, T. K. Jojoba oil wax esters and derived fatty acids and alcohols: Gas chromatographic analyses. *J. Am. Oil Chem. Soc.* **48**, (1971).

18. Wältermann, M. *et al.* Mechanism of lipid-body formation in prokaryotes: How bacteria fatten up. *Mol. Microbiol.* **55**, (2005).

19. Wältermann, M., Stöveken, T. & Steinbüchel, A. Key enzymes for biosynthesis of neutral lipid storage compounds in prokaryotes: Properties, function and occurrence of wax ester synthases/acyl-CoA:diacylglycerol acyltransferases. *Biochimie* **89**, (2007).

20. Stöveken, T., Kalscheuer, R., Malkus, U., Reichelt, R. & Steinbüchel, A. The wax ester synthase/acyl coenzyme A:diacylglycerol acyltransferase from Acinetobacter sp. strain ADP1: Characterization of a novel type of acyltransferase. *J. Bacteriol.* **187**, (2005).

21. Kalscheuer, R., Stölting, T. & Steinbüchel, A. Microdiesel: Escherichia coli engineered for fuel production. *Microbiology* **152**, (2006).

22. Stöveken, T. & Steinbüchel, A. Bacterial acyltransferases as an alternative for lipase-catalyzed acylation for the production of oleochemicals and fuels. *Angewandte Chemie - International Edition* **47**, 3688–3694 (2008).

23. Steen, E. J. *et al.* Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. *Nature* **463**, (2010).

24. Alvarez, A. F., Alvarez, H. M., Kalscheuer, R., Wältermann, M. & Steinbüchel, A. Cloning and characterization of a gene involved in triacylglycerol biosynthesis and identification of additional homologous genes in the oleaginous bacterium Rhodococcus opacus PD630. *Microbiology* **154**, (2008).

25. Barney, B. M., Wahlen, B. D., Garner, E., Wei, J. & Seefeldt, L. C. Differences in Substrate Specificities of Five Bacterial Wax Ester Synthases. (2012). doi:10.1128/AEM.00534-12

26. Holtzapple, E. & Schmidt-Dannert, C. Biosynthesis of isoprenoid wax ester in Marinobacter hydrocarbonoclasticus DSM 8798: Identification and characterization of isoprenoid coenzyme a synthetase and wax ester syethases. *J. Bacteriol.* **189**, 3804–3812 (2007).

27. Manilla-Pérez, E., Lange, A. B., Luftmann, H., Robenek, H. & Steinbüchel, A. Neutral lipid production in Alcanivorax borkumensis SK2 and other marine hydrocarbonoclastic bacteria. *Eur. J. Lipid Sci. Technol.* **113**, (2011).

28. Petronikolou, N. & Nair, S. K. Structural and Biochemical Studies of a Biocatalyst for the Enzymatic Production of Wax Esters. *ACS Catal.* **8**, 1–12 (2018).

29. Márquez, M. C. & Ventosa, A. Marinobacter hydrocarbonoclasticus Gauthier et al. 1992 and Marinobacter aquaeolei Nguyen et al. 1999 are heterotypic synonyms. *International Journal of Systematic and Evolutionary Microbiology* **55**, 1349–1351 (2005).

30. Tindall, B. J. Marinobacter nauticus (Baumann et al. 1972) comb. nov. arising from instances of synonymy and the incorrect interpretation of the International Code of Nomenclature of Prokaryotes. *Arch. Microbiol.* **202**, (2020).

31. Indest, K. J., Eberly, J. O., Ringelberg, D. B. & Hancock, D. E. The effects of putative lipase and wax ester synthase/acyl-CoA:diacylglycerol acyltransferase gene knockouts on triacylglycerol accumulation in Gordonia sp. KTR9. *J. Ind. Microbiol. Biotechnol.* **42**, (2015).

32. Kaddor, C., Biermann, K., Kalscheuer, R. & Steinbüchel, A. Analysis of neutral lipid biosynthesis in Streptomyces avermitilis MA-4680 and characterization of an acyltransferase involved herein. *Appl. Microbiol. Biotechnol.* **84**, (2009).

33. Röttig, A., Wolf, S. & Steinbüchel, A. In vitro characterization of five bacterial WS/DGAT acyltransferases regarding the synthesis of biotechnologically relevant short-chain-length esters. *Eur. J. Lipid Sci. Technol.* **118**, 124–132 (2016).

34. Corpet, F. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* **16**, 10881–10890 (1988).

35. Buglino, J., Onwueme, K. C., Ferreras, J. A., Quadri, L. E. N. & Lima, C. D. Crystal structure of PapA5, a phthiocerol dimycocerosyl transferase from Mycobacterium tuberculosis. *J. Biol. Chem.* **279**, (2004).

36. Stöveken, T., Kalscheuer, R. & Steinbüchel, A. Both histidine residues of the conserved HHXXXDG motif are essential for wax ester synthase/acyl-CoA:Diacylglycerol acyltransferase catalysis. *Eur. J. Lipid Sci. Technol.* **111**, (2009).

37. Lenneman, E. M., Ohlert, J. M., Palani, N. P. & Barney, B. M. Fatty alcohols for wax esters in Marinobacter aquaeolei VT8: Two optional routes in the wax biosynthesis pathway. *Appl. Environ. Microbiol.* **79**, 7055–7062 (2013).

38. Reiser, S. & Somerville, C. Isolation of mutants of Acinetobacter calcoaceticus deficient in wax ester synthesis and complementation of one mutation with a gene encoding a fatty acyl coenzyme A reductase. *J. Bacteriol.* **179**, (1997).

39. Wahlen, B. D., Oswald, W. S., Seefeldt, L. C. & Barney, B. M. Purification, characterization, and potential bacterial Wax production role of an nadph-dependent fatty aldehyde reductase from Marinobacter aquaeolei VT8. *Appl. Environ. Microbiol.* **75**, (2009).

40. Willis, R. M., Wahlen, B. D., Seefeldt, L. C. & Barney, B. M. Characterization of a fatty acyl-CoA reductase from Marinobacter aquaeolei VT8: A bacterial enzyme catalyzing the reduction of fatty acyl-CoA to fatty alcohol. *Biochemistry* **50**, (2011).

41. Blomquist, G. J., Chu, A. J. & Remaley, S. Biosynthesis of wax in the honeybee, Apis mellifera L. *Insect Biochem.* **10**, (1980).

42. Wilkins, M. R. *et al.* Protein identification and analysis tools in the ExPASy server. *Methods Mol. Biol.* **112**, 531–52 (1999).

43. Gasteiger, E. *et al.* Protein Identification and Analysis Tools on the ExPASy Server. in *The Proteomics Protocols Handbook* 571–607 (Humana Press, 2005). doi:10.1385/1-59259-890-0:571

44. Barney, B. M., Mann, R. L. & Ohlert, J. M. Identification of a residue affecting fatty alcohol selectivity in wax ester synthase. *Appl. Environ. Microbiol.* **79**, 396–399 (2013).

45. Barney, B. M., Ohlert, J. M., Timler, J. G. & Lijewski, A. M. Altering small and medium alcohol selectivity in the wax ester synthase. *Appl. Microbiol. Biotechnol.* **99**, 9675–9684 (2015).

46. Calixto Mancipe, N., McLaughlin, E. M. & Barney, B. M. Genomic analysis and characterization of Scenedesmus glucoliberatum PABB004: An unconventional sugar-secreting green alga. *J. Appl. Microbiol.* **132**, 2004–2019 (2021).

47. Sheldon, R. A. Chemicals from renewable biomass: A renaissance in carbohydrate chemistry. *Curr. Opin. Green Sustain. Chem.* **14**, 89–95 (2018).

48. Zeng, A. P. New bioproduction systems for chemicals and fuels: Needs and new development. *Biotechnology Advances* **37**, 508–518 (2019).

49. Fabris, M. *et al.* Emerging Technologies in Algal Biotechnology: Toward the Establishment of a Sustainable, Algae-Based Bioeconomy. *Frontiers in Plant Science* **11**, 279 (2020).

50. Jeong, G. T., Kim, S. K. & Oh, B. R. Production of fermentable sugars from Chlorella sp. by solid-acid catalyst. *Algal Res.* **51**, 102044 (2020).

51. Ramachandra, T. V. & Hebbale, D. Bioethanol from macroalgae: Prospects and challenges. *Renewable and Sustainable Energy Reviews* **117**, 109479 (2020).

52. Sanz Smachetti, M. E., Coronel, C. D., Salerno, G. L. & Curatti, L. Sucrose-to-ethanol microalgae-based platform using seawater. *Algal Res.* **45**, 101733 (2020).

53. Hoshina, R., Iwataki, M. & Imamura, N. Chlorella variabilis and Micractinium reisseri sp. nov. (Chlorellaceae, Trebouxiophyceae): Redescription of the endosymbiotic green algae of Paramecium bursaria (Peniculia, Oligohymenophorea) in the 120th year. *Phycol. Res.* **58**, 188–201 (2010).

54. Yellowlees, D., Rees, T. A. V. & Leggat, W. Metabolic interactions between algal

symbionts and invertebrate hosts. *Plant, Cell and Environment* **31**, 679–694 (2008).

55. Davy, S. K., Allemand, D. & Weis, V. M. Cell Biology of Cnidarian-Dinoflagellate Symbiosis. *Microbiol. Mol. Biol. Rev.* **76**, 229–261 (2012).

56. Brechignac, F. & Schiller, P. Pilot CELSS based on a maltose-excreting Chlorella: concept and overview on the technological developments. *Adv. Sp. Res* **12**, 33 (1992).

57. Dorling, M., Mcauley, P. J. & Hodge, H. Effect of pH on growth and carbon metabolism of maltose-releasing Chlorella (Chlorophyta). *Eur. J. Phycol.* **32**, 19–24 (1997).

58. Arriola, M. B. *et al.* Genome sequences of Chlorella sorokiniana UTEX 1602 and Micractinium conductrix SAG 241.80: implications to maltose excretion by a green alga. *Plant J.* **93**, 566–586 (2018).

59. Blaby-Haas, C. E. & Merchant, S. S. Comparative and Functional Algal Genomics. *Annu. Rev. Plant Biol* **70**, 605–638 (2019).

60. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness. *Methods Mol. Biol.* **1962**, 227–245 (2019).

61. Merchant, S. S. *et al.* The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science (80-.).* **318**, 245–251 (2007).

62. Blanc, G. *et al.* The Chlorella variabilis NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. *Plant Cell* **22**, 2943–2955 (2010).

63. Turmel, M., Otis, C. & Lemieux, C. *The complete chloroplast DNA sequence of the green alga Nephroselmis olivacea: Insights into the architecture of ancestral chloroplast genomes*. **96**, (1999).

64. Lemieux, C., Vincent, A. T., Labarre, A., Otis, C. & Turmel, M. Chloroplast phylogenomic analysis of chlorophyte green algae identifies a novel lineage sister to the Sphaeropleales (Chlorophyceae) Phylogenetics and phylogeography. *BMC Evol. Biol.* **15**, (2015).

65. Romero-Rodríguez, A. *et al.* Transcriptomic analysis of a classical model of carbon catabolite regulation in Streptomyces coelicolor. *BMC Microbiol.* **16**, (2016).

66. Barney, B. M., Eberhart, L. J., Ohlert, J. M., Knutson, C. M. & Plunkett, M. H. Gene deletions resulting in increased nitrogen release by Azotobacter vinelandii: Application of a novel nitrogen biosensor. *Appl. Environ. Microbiol.* **81**, 4316–4328 (2015).

67. Moya, A. *et al.* Toward minimal bacterial cells: Evolution vs. design. *FEMS Microbiology Reviews* **33**, (2009).

68. Leliaert, F. *et al.* Phylogeny and Molecular Evolution of the Green Algae. *CRC.*

Crit. Rev. Plant Sci. **31**, 1–46 (2012).

69. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).

70. Hoff, K. J. & Stanke, M. Predicting Genes in Single Genomes with AUGUSTUS. *Curr. Protoc. Bioinforma.* (2018). doi:10.1002/cpbi.57

71. Tillich, M. *et al.* GeSeq-versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* **45**, (2017).

72. Burki, F., Roger, A. J., Brown, M. W. & Simpson, A. G. B. The New Tree of Eukaryotes. (2020). doi:10.1016/j.tree.2019.08.008

73. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. doi:10.1093/molbev/msy096

74. Contreras-Moreira, B. & Vinuesa, P. GET_HOMOLOGUES, a Versatile Software Package for Scalable and Robust Microbial Pangenome Analysis. *Appl. Environ. Microbiol.* **79**, 24 (2013).

75. Contreras-Moreira, B. *et al.* Analysis of plant pan-genomes and transcriptomes with GET_HOMOLOGUES-EST, a clustering solution for sequences of the same species. *Front. Plant Sci.* **8**, (2017).

76. Xu, L. *et al.* OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.* **47**, (2019).

77. Wijffels, R. H., Barbosa, M. J. & Eppink, M. H. M. Microalgae for the production of bulk chemicals and biofuels. *Biofuels, Bioproducts and Biorefining* **4**, 287–295 (2010).

78. Lowe, D. G. *et al.* Human atrial natriuretic peptide receptor defines a new paradigm for second messenger signal transduction. *EMBO J.* **8**, 1377–1384 (1989).

79. Geyer, R., Jambeck, J. R. & Law, K. L. Production, use, and fate of all plastics ever made. *Sci. Adv.* **3**, e1700782 (2017).

80. Brydson, J. The Historical Development of Plastics Materials. in *Plastics Materials* (Elsevier, 1999).

81. Jambeck, J. R. *et al.* Plastic waste inputs from land into the ocean. *Science (80-.).* **347**, 768–771 (2015).

82. Research highlights true impacts of plastics on our planet, ecosystems, people | UNEP - UN Environment Programme. Available at: https://www.unenvironment.org/news-and-stories/press-release/research-highlights-true-impacts-plastics-our-planet-ecosystems. (Accessed: 3rd October 2019)

83. Wei, R. & Zimmermann, W. Microbial enzymes for the recycling of recalcitrant petroleum-based plastics: how far are we? *Microb. Biotechnol.* **10**, 1308–1322 (2017).

84. Yoshida, S. *et al.* A bacterium that degrades and assimilates poly(ethylene terephthalate). *Science* **351**, 1196–9 (2016).

85. Schmidt, J. *et al.* Degradation of polyester polyurethane by bacterial polyester hydrolases. *Polymers (Basel).* **9**, 65 (2017).

86. Biundo, A., Ribitsch, D. & Guebitz, G. M. Surface engineering of polyester-degrading enzymes to improve efficiency and tune specificity. *Appl. Microbiol. Biotechnol.* **102**, 3551–3559 (2018).

87. Drzyzga, O. & Prieto, A. Plastic waste management, a matter for the 'community'. *Microbial Biotechnology* **12**, 66–68 (2019).

88. Mills, N. J. Molecular structures and manufacture of polymers. in *Plastics - Microstructure and Engineering Applications* 1–30 (Elsevier, 1993).

89. Gewert, B., Plassmann, M. M. & Macleod, M. Pathways for degradation of plastic polymers floating in the marine environment. *Environmental Sciences: Processes and Impacts* **17**, 1513–1521 (2015).

90. Gamerith, C. *et al.* Enzymatic recovery of polyester building blocks from polymer blends. *Process Biochem.* **59**, 58–64 (2017).

91. Shah, A. A., Hasan, F., Hameed, A. & Ahmed, S. Biological degradation of plastics: A comprehensive review. *Biotechnol. Adv.* **26**, 246–265 (2008).

92. Iiyoshi, Y., Tsutsumi, Y. & Nishida, T. Polyethylene degradation by lignin-degrading fungi and manganese peroxidase. *J. Wood Sci.* **44**, 222–229 (1998).

93. Nakamiya, K., Ooi, T. & Kinoshita, S. Degradation of synthetic water-soluble polymers by hydroquinone peroxidase. *J. Ferment. Bioeng.* **84**, 213–218 (1997).

94. Sulaiman, S., You, D. J., Kanaya, E., Koga, Y. & Kanaya, S. Crystal structure and thermodynamic and kinetic stability of metagenome-derived LC-cutinase. *Biochemistry* **53**, 1858–1869 (2014).

95. Austin, H. P. *et al.* Characterization and engineering of a plastic-degrading aromatic polyesterase. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E4350–E4357 (2018).

96. Nakamiya, K., Sakasita, G., Kinoshita, S., Ooi, T. & Kinoshita, S. Enzymatic degradation of polystyrene by hydroquinone peroxidase of Azotobacter beijerinckii HM121. *J. Ferment. Bioeng.* **84**, 480–482 (1997).

97. Romanova, V. A. *et al.* Draft Genome Sequence of a Medium- and Long-Chain n -Alkane-Degrading Bacterium, Tsukamurella tyrosinosolvens Strain PS2, with Two Genetic Systems for Alkane Degradation . *Microbiol. Resour. Announc.* **8**, (2019).

98. Chiciudean, I., Nie, Y., Tănase, A. M., Stoica, I. & Wu, X. L. Complete genome sequence of Tsukamurella sp. MH1: A wide-chain length alkane-degrading actinomycete. *J. Biotechnol.* **268**, 1–5 (2018).

99. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).

100. Villa, J. A., Ray, E. E. & Barney, B. M. Azotobacter vinelandii siderophore can provide nitrogen to support the culture of the green algae neochloris oleoabundans and scenedesmus sp. BA032. *FEMS Microbiology Letters* **351**, 70–77 (2014).