Engineering *Parageobacillus thermoglucosidans* as a robust platform for bioethanol production

Leonardo Joaquim van Zyl

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of philosophy (PhD)

Institute for Microbial Biotechnology and Metagenomics, Department of Biotechnology, Faculty of Natural Science

University of the Western Cape,

Bellville, Cape Town, South Africa

Supervisor: Professor Marla Trindade

Co-supervisor: Professor Donald Cowan

Administrative supervisor: Professor Ndiko Ludidi

March 2018

Declaration

I, Leonardo Joaquim van Zyl, hereby declare that "Engineering *Parageobacillus thermoglucosidans* as a robust platform for bioethanol production" is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Date: ……16/03/2018…….

Signed: 

Abstract

Parageobacillus thermoglucosidans is a promising "platform" organism to use in the production of a range of useful metabolites with demonstrated ability to produce ethanol, isobutanol and polylactic acid for bio-degradable plastics. Extensive work has been done in engineering the organism for enhanced ethanol production. However, an often used and highly effective alternative pathway (pyruvate decarboxylase mediated) for ethanol production has not yet been demonstrated in *P. thermoglucosidans*. We first characterize two novel bacterial pyruvate decarboxylase enzymes (PDC's) then attempt to express the more thermostable of these enzymes from *Gluconobacter oxydans* in *P. thermoglucosidans* to improve ethanol yields. Initial expression was unsuccessful. Analysis of the codon usage pattern for the gene revealed that the codon usage was suboptimal in the heterologous host *P. thermoglucosidans*. After codon harmonization, we could demonstrate successful expression of the enzyme at 45°C, however not at the bacterium's optimum growth temperature of 60°C. This was concomitant with enhanced ethanol production close to the theoretical yield possible (0.5g/l).

It is well known that commercial bacterial fermentations are prone to bacteriophage (virus) attack and as such any strains used either should be engineered or selected for resistance against the phage. As *P. thermoglucosidans* is to be used to ferment a range of feedstocks from around the world, it will from time to time, be exposed to phages from a variety of environments which may lead to failed or stuck fermentations and associated financial loss. This also means that it will be an on-going problem for those wishing to employ this organism for large scale metabolite production. Although several *Parageobacillus* species phages have been described (GVE1, GVE2, GBSV1, GBK2, DE6 and ϕOH2), sequenced and one, GVE2, well studied, none have been found that infect *P. thermoglucosidans.* We describe a novel bacteriophage (GVE3) that infects *P. thermoglucosidans* and we develop strains resistant against the phage in two ways. The one mechanism is the overexpression of an immunity protein encoded by the phage in *P. thermoglucosidans* and the second is mutation of the polysaccharide pyruval transferase (*csaB*), likely involved in phage infection. Both mechanisms appear to give complete phage resistance however the exact mechanism, DNA exclusion or interference during phage binding, remains to be elucidated. The combination of improvement in ethanol yield and phage resistance should make *P. thermoglucosidans* a robust platform to produce a wide range of metabolites while taking advantage of the thermophilic nature of the microorganism.

Keywords: Thermophilic, phage, pyruvate decarboxylase, GVE3, codon harmonization

Acknowledgements

I would like to express my gratitude to my parents, Leon and Elsa van Zyl without who's support, none of the work I have performed included here or outside of this thesis would have ever been possible. Apart from this obvious statement, they laid the foundations of my ability to ask questions and persist in finding solutions, whether in life or biological research. I give them credit for this small aspect here, however they have instilled in me much more than this and I am grateful to them for their unwavering love and support for all these years.

I want to thank my love, partner, lab mate, confidant, collaborator and supervisor for much of the work performed here, Prof. Marla Trindade. Although not officially being given credit for my supervision, due to the conflict of interest that presents, she was nonetheless the functional supervisor throughout most of this work. She's given a tremendous amount of input and numerous discussions over red wine led to the work presented here. Without her continued support, personally and professionally, the work would simply lose its flavour. Life has a strange way of working out. There is a lot of water that has passed under our bridge and we have had many challenges, but we've weathered these together and will continue to do so for many years to come. I look forward to sharing a life with you and working by your side to uncover all that (micro)biology has to offer. We really need to focus our projects a little!

I would like to thank Prof. Don Cowan who in 2008 (at Marla's suggestion) invited me to join his burgeoning group at the University of the Western Cape. Don is a fantastic enabler of scientific research and he and IMBM gave me (and many others) the opportunity to explore my creativity in doing research, parts of which are on display in this work and for this I will always be thankful. For me, Don took the limits off of what can be achieved in a project, and showed that one can, and should, dream big. Thank you for your infectious enthusiasm and the many lessons both Marla and I will carry forward with us. Apart from this, his inclusion of those prepared to put up their hand to do work in all sorts of (scientific) ventures saw me be everything form the IT guy at the 2008 Extremophiles conference to four years later exploring viruses in the Namib Desert, in my native homeland. This whirlwind of field trips, lab work and conferences injected with his exuberance for exploration made for a fun place to work.

I would like to extend my gratitude to the University of the Western Cape who have employed me and therefore provided me a means to exercise my love of research for the past nine years.

The work presented here would not have been possible had I not received, what in my opinion, was excellent training up to the point of starting this degree. I am therefore heavily indebted to Dr. Shelly Deane and Prof. Douglas Rawlings, from the University of Stellenbosch, under who's tutelage my skills as a researcher were developed.

Table of Contents

This thesis takes the form of four published manuscripts bookended by an introductory chapter and general discussion chapter.

Table of Figures

List of Tables

Abbreviations

℃	Celsius
µg	Microgram
µg/ml	Microgram per millilitre
µl	Microlitre
µm	Micrometre
µmol	Micromoles
µM	Micromolar
Å	Angstrom
A	Absorbance
ATP	Adenosine triphosphate
BLAST	Basic local alignment tool
BLASTn	Nucleotide BLAST
BLASTp	Protein BLAST
BLASTx	Translated nucleotide query BLAST against a protein database
bp	Base pair
$CaCl_2$	Calcium chloride
C:I	Chloroform : Isoamyl alcohol
$CoSO_4$	Cobalt sulphate
$CuSO_4$	Copper sulphate
Contig	Contiguous sequence
DCO	Double crossover
DNA	Deoxyribonucleic acid
dNTPs	Deoxyribonucleic acid triphosphates
EDTA	Ethylenediaminetetraacetic acid
$FeSO_4$	Ferrous sulphate
G	G-force or relative centrifugal force
g/g	Gram ethanol produced per gram glucose consumed

g/l	Grams per litre
h	Hours
HCD	High energy collision dissociation
HPLC	High performance liquid chromatography
H_2SO_4	Sulphuric acid
K	Kelvin
k_{cat}	Turnover number - the number of substrate molecules each enzyme active site converts to product per unit time
kDa	Kilo Dalton
K_M	Michaelis constant
k_{cat}/K_M	Catalytic efficiency
kJ/mol	Kilo Joule per mole
K_2HPO_4	Potassium phosphate dibasic
K_2SO_4	Potassium sulphate
kV	Kilo volt
LB	Lysogeny broth
MeOH	Methanol
Min	Minutes
$MgSO_4$	Magnesium sulphate
$MgCl_2$	Magnesium chloride
ml	Millilitre
ml/min	Millilitre per minute
mM	Millimolar
mm	Millimetre
MMTS	Methane methylthiosulfonate
MnSO4	Manganese sulphate
MOI	Multiplicity of infection
$M^{-1}.s^{-1}$	Molar per second
MW	Molecular weight

m/z	Mass to charge ratio
ng	Nanogram
nm	Nanometres
NaCl	Sodium chloride
NAD^+/NADH	Nicotinamide dinucleotide / reduced
NaOH	Sodium hydroxide
Na_2MoO_4	Sodium molybdate
NaH_2PO_4	Sodium dihydrogen phosphate
$NiSO_4$	Nickel sulphate
OD	Optical density
ORF	Open reading frame
PCR	Polymerase chain reaction
P:C:I	Phenol : Chloroform : Isoamyl alcohol
PEG	Polyethylene Glycol
pH_{opt}	pH optimum
ppm	Parts per million
RNA	Ribonucleic acid
rpm	Revolutions per minute
rRNA	Ribosomal ribonucleic acid
s	Seconds
SCO	Single crossover
SM	Sodium chloride / Magnesium sulphate buffer
$T_{1/2}$	Enzyme half-life at the given temperature
TCEP	Tris-carboxyethyl phosphine
TE	Tris-EDTA buffer
TEAB	Triethylammonium bicarbonate
TFA	Trifluoroacetic acid
TGP	Tryptone glucose pyruvate medium
ThDP	Thiamine diphosphate

TPP	Thiamine pyrophosphate
T_{opt}	Temperature optimum
U	Unit
U/mg	units per milligram
U/ml	Units per millilitre
USM	Urea sulphates medium
UV/Vis	Ultraviolet / Visible spectrum
v/v	Volume per volume
WT	Wild type
w/v	Weight per volume
wt/vol	Weight per volume
$ZnSO_4$	Zinc sulphate

Chapter 1

LITERATURE REVIEW

It has recently been proposed that the genus *Geobacillus* be spilt into two genera: *Geobacillus* and *Parageobacillus*. The species name of *Geobacillus thermoglucosidasius* has also changed in recent years to *thermoglucosidans*. The manuscripts presented here were written at a time before these name changes took effect, and in this thesis, we consider *Geobacillus thermoglucosidasius* and *Parageobacillus thermoglucosidans* as synonymous, as are different combinations of the genus and species descriptors.

1.1 Fermentation

The process of fermentation is one that has been employed by humans, initially probably inadvertently, and later knowingly for thousands of years for food preservation, production of intoxicating beverages or acidic dairy products (McGovern *et al*., 1986; Oberman and Libudzisz 1998; Zhong *et al*., 2016). The benefits of fermentation on food, is that it makes it more digestible, introduces compounds such as vitamins (folic acid, niacin), produces a product with better organoleptic qualities and improves shelf life through the introduction of antimicrobial compounds and reduction in pH (small molecule antibiotics or proteins) (Pederson 1971). The microorganisms involved are usually fungi (*Aspergillus*, *Rhizopus*, *Mucor*, *Actinomucor*, and *Neurospora*) or bacteria (*Bacillus*, *Clostridium*, *Lactococcus*, *Lactobacillus*, *Streptococcus* and *Pediococcus*). The single-celled fungus, *Saccharomyces cerevisiae*, is arguably the most important microorganism used in commercial fermentation processes to produce bread, beverages (beer and wine) as well as grain alcohol, while lactic acid bacteria (LABs) are most often used in dairy products.

Fermentation can be defined as a metabolic process that generates energy from organic molecules, which does not require oxygen or an electron transport chain (ETC) and uses an organic molecule as

final electron acceptor. The fermentation pathway is considered ancient compared to oxidative phosphorylation as the concentration of oxygen in the atmosphere only started increasing ±2 billion years ago allowing organisms to take advantage of this process (Tadege *et al*., 1999; Sessions *et al*., 2009). In general, when oxygen is present to serve as final electron acceptor, glucose will be processed through the glycolysis pathway to produce 2 molecules of acetyl-coA (**Figure 1.1.1**) where it "joins" the tricarboxylic acid cycle (TCA). Through glycolysis, nicotinamide adenine dinucleotide (NADH x2) and adenosine triphosphate (ATP x2) are generated through substrate level phosphorylation. Through the TCA cycle more NADH/FADH is produced. These reduced molecules (NADH/FADH) transfer electrons to a series of electron acceptors and donors (ubiquinone, cytochrome bc1 complex, cytochrome c, cytochrome c oxidase) in the ETC embedded in the bacterial inner plasma membrane. This transfer powers the translocation of protons (H^+) across the plasma membrane, from inside the cytoplasm to the periplasmic space. The electrons are eventually handed to molecular oxygen as final acceptor and coupling of oxygen to protons leads to the formation of metabolic water. This process also serves to regenerate NAD^+/FAD^+ to take part in more rounds of glycolysis and the TCA cycle. The local buildup of positive charge on the outside of the membrane and resulting negative charge on the inside, creates a proton gradient, referred to as the proton motif force (pmf). As these protons flow "down" the gradient back through the membrane *via* a transmembrane protein complex (F1F0 ATP synthase), they drive the rotation of the synthase which couple's inorganic phosphate (P_i) to adenosine diphosphate (ADP) to produce adenosine triphosphate (ATP), the cells' energy unit. Theoretically, prokaryotes can generate 38 molecules of ATP from one glucose molecule, however in reality there are losses due to leakiness of the membrane and the pmf being used for purposes other than energy generation. As stated above, two molecules of ATP are generated from glycolysis, while each NADH generates 2.5 ATP units (1.5 / FADH) in ETC. Thus, from one glucose molecule a total of 14 molecules of ATP are generated if converted to acetyl-CoA in glycolysis. An additional 22 molecules of ATP are generated from the ATP and NADH/FADH produced in the TCA cycle. This makes the total yield of ATP under aerobic conditions 38 (+2 ATP in prokaryotes that don't have to move NADH across the mitochondrial membrane) compared to just 2 under anaerobic conditions (Teusink and Molenaar 2017).

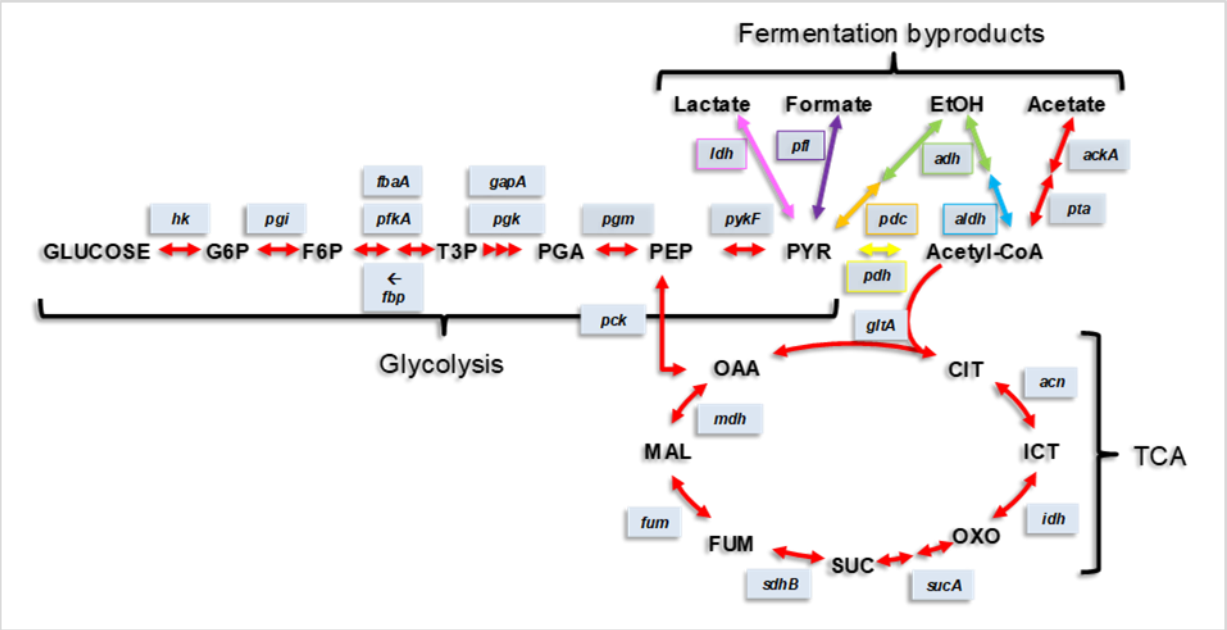


Figure 1.1.1 Broad overview of glycolysis, TCA cycle, fermentation end products and the enzymes involved. 6PG, 6-phosphogluconate; acetyl-CoA, acetyl-coenzyme A; CIT, citrate; E4P, erythrose-4-phosphate; EtOH, ethanol; F6P, fructose-6-phosphate; FUM, fumarate; G3P, glyceraldehyde-3- phosphate; G6P, glucose-6-phosphate; ICT, isocitrate; MAL, malate; OAA, oxaloacetate; OXO, 2-oxoglutarate; PEP, phosphoenolpyruvate; PGA, 3-phosphoglycerate; PYR, pyruvate; R5P, ribulose-5-phosphate; S7P, sedoheptulose-7-phosphate; SUC, succinate; T3P, triose-3-phosphate; X5P, xylose-5-phosphate; *hk*, hexokinase; *pgi*, phosphoglucose isomerase; *pdc*, pyruvate decarboxylase; *adh*, alcohol dehydrogenase; *aldh*, aldehyde dehydrogenase; *ackA*, acetate kinase; *pdh*, pyruvate dehydrogenase; *ldh*, lactate dehydrogenase; *pfl*, pyruvate formate lyase; *sdhB*, succinate dehydrogenase iron-sulfur subunit; *sucA*, succinate dehydrogenase catalytic subunit; *pgk*, phosphoglycerate kinase; *gapA*. glyceraldehyde 3-phosphate dehydrogenase; *pfkA*, phosphofructokinase; *pykF*, pyruvate kinase; *fbaA*, fructose-1,6-bisphosphate aldolase; *fbp*, fructose bisphosphatase; *pck*, phosphoenolpyruvate carboxykinase; *pta*, phosphotransacetylase; *pgm*, phosphoglycerate mutase; *acn*, aconitase; *idh*, isocitrate dehydrogenase; *fum*, fumarase; *mdh*, malate dehydrogenase; *gltA*, citrate synthase.

In the absence of oxygen, the cell relies mostly on this substrate level phosphorylation for its ATP production as the ETC is rendered inactive due to the lack of oxygen to accept electrons. As mentioned above, during fermentation, the end products are the result of organic molecules acting as final electron acceptor and they can either be a single product (homofermentative) or a mix of products (heterofermentative). Depending on the organism involved, two main pathways are followed: homofermentative with lactic acid as main end product, or heterofermentative with a combination of organic acids and alcohols (heterofermentative; ethanol / acetate / formate / propionate / butyrate / succinate / butanol / acetone). The lactic acid pathway is present in many bacteria including *Parageobacillus thermoglocosidans* (see section 1.2.4; Cripps *et al*., 2009). In this pathway, pyruvate serves as final electron acceptor with NADH as donor, and the reaction is catalyzed by the enzyme

lactate dehydrogenase (Tadege *et al*., 1999). This regenerates NAD^+ to again take part in glycolysis. During alcoholic fermentation, which occurs in a wide range of organisms including plants, bacteria, yeast and some animals, the enzyme pyruvate decarboxylase catalyzes the cleavage of pyruvate to yield acetaldehyde which is then converted by alcohol dehydrogenase to alcohol using NADH as cofactor (van Waarde 1991). This again recycles NAD^+. A second pathway is *via* acetyl-CoA where it is reduced to acetaldehyde by aldehyde dehydrogenase with conversion of acetaldehyde to ethanol. Another important product of fermentation is acetate from acetyl-CoA. This benefits the organism capable of acetate production due to the generation of an additional ATP molecule by acetate kinase when converting acetyl-phosphate to acetate.

1.2 Introduction to biofuels

It is now accepted that fossil fuel reserves, the main source for liquid petroleum, will eventually be depleted. It is also established that the use of fossil fuels has a negative impact on the environment, contributing to global warming, through re-introduction of trapped carbon, a greenhouse gas, into the atmosphere (Naik *et al*., 2010, Rulli *et al*., 2016). Thus, an alternative source for liquid fuels needs to be found and one of the proposed alternatives is biofuels. Biofuels refers to technologies that use renewable feedstocks (plant material, autotrophic photosynthetic organisms) and often employ living organisms, mostly yeast, algae or bacteria, to convert biomass to liquid fuels. Apart from the environmental benefits which come with biofuels, they may also contribute to the enhancement of energy security in countries which don't have access to fossil fuel deposits, and offer a more profitable use of crops other than as a food source. The biomass used may be sugars from food crops such as maize (corn) or can be lignocellulosic (non-edible parts of the plant) in nature. While ethanol has been the major focus as a fuel produced from biomass, more recently researchers have shifted their focus to longer chain-length molecules such as biodiesel and isobutanol, due to their higher energy density and compatibility with existing motor vehicle engines and infrastructure for refining, housing and dispensing of these liquid fuels (Taylor *et al*., 2009; Lin *et al*., 2014; http://tinyurl.com/po6a52q). Most biofuels processes can be classified into four different categories: First -, second -, third - and fourth

generation biofuels which will be discussed in more detail below. There are several factors to consider when comparing biofuels processes. There are the technical differences in the processes, energetic inputs and outputs, financial competitiveness and environmental impact costs (Hill *et al*., 2006). **Figure 1.2.1** shows a summary of the benefits and drawbacks of the various processes. Arguably the most prominent of the biofuels processes to date are the first-generation processes (1G), which, until recently was also the only truly commercially viable biofuels process. Other processes (2G and 3G) are just coming on-line (INEOS Bio, Abengoa, POET, India Glycols) however will have to prove their worth in years to come. As this study only addressed the technical challenges associated with these processes, the financial, energetic, and environmental impacts will not be discussed in detail.

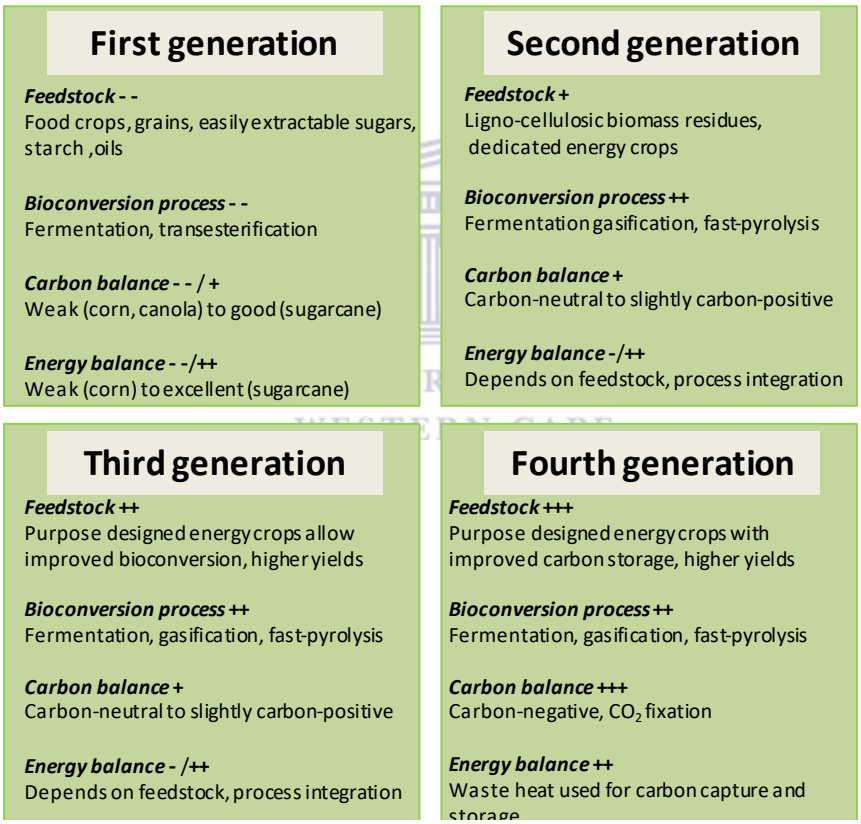


Figure 1.2.1 Summary of the four generations of biofuels processes and their main advantages and disadvantages. Adapted from http://tinyurl.com/jqrc8f8

1.2.1 Types of biofuel process

1.2.1.1 First generation biofuels

The main products of 1G processes are biodiesel, bioethanol, or biogas, produced primarily through transesterification of vegetable oil, yeast fermentation and anaerobic digestion respectively (Naik *et al*., 2010). First generation processes refer to those that, at least for biodiesel and bioethanol, use the edible parts of plants, or rather sugars derived from them, as feedstock and the methods employed in 1G processes are well established as these have been used and refined over many years. Although commercially successful, production of biofuels from edible plant sugars is not favoured as this leads to competition for the resource to either be used as food or fuel and this has become known as the "food vs fuel debate" (Mohr and Raman 2013). Additionally, these processes place a high demand on available arable land, and fresh water resources making them less attractive (Rulli *et al*., 2016).

The production of first generation biodiesel, also known as fatty acid methyl esters (FAME), happens through transesterification of vegetable oil, used cooking and frying oil with an alcohol, in the presence of a catalyst (KOH / NaOH / CH_3ONa / CH_3OK) (**Figure 1.2.1.1.1**). The various catalysts have different efficiencies in converting the oil/alcohol mixture to methyl esters with CH_3ONa (>98 wt%) being most efficient. Two more factors impede the efficient conversion of the oil to methyl esters. These are the water content of the oil feedstock and the presence of free fatty acids. The water, in the presence of alkali, leads to hydrolysis of the triglyceride ester bond and leads to the release of free fatty acid. These free fatty acids can react with the catalysts to produce soaps thereby using up catalyst in the reaction. A solution to this is the use of a heterogeneous solid base/acid catalyst such as alkali doped materials (Na_2SiO_3/Li_4SiO_4), alkaline earth oxides ([CaO/Al_2O_3] / [$Cs_2Mg(CO_3)_2(H_2O)_4$]), hydrotalcites ($[M(II)_{1-x}M(III)_x(OH)_2]x+(A^{n-}_{x/n})\cdot mH_2O$), sulfonated multi-walled carbon nanotubes (s-MWCNTs) or sulfated zirconium-alumina (SZA) to prevent saponification (Semwal *et al*., 2011; Lee *et al*., 2014). Biodiesel can also be produced by other methods such as direct/blends, microemulsion or pyrolysis (Atadashi *et al*., 2013).

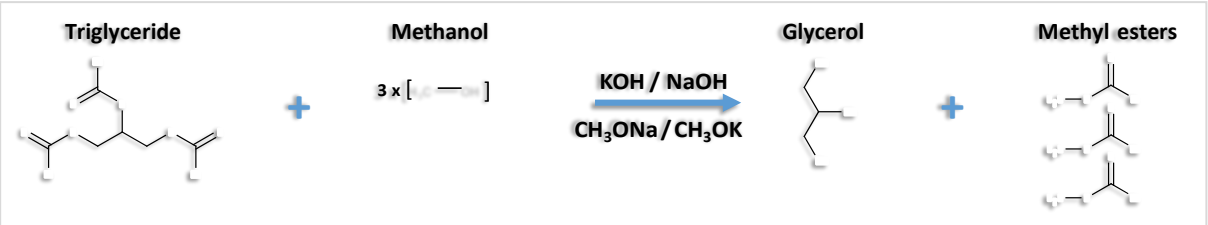


Figure 1.2.1.1.1 General formula for the production of first generation biodiesel.

When using sugar from food crops such as that used in 1G processes, the product is referred to as grain ethanol. The crops used to produce grain ethanol include sugar cane, wheat, beet root, palm juice, wheat, barely, rice, sweet sorghum, corn, potato and cassava. The first step in the process is to remove the easily recoverable sugars (starch) from the plant material. Starch is composed of a chain of glucose residues and depending on the plant species, is composed of 20-30% linear glucose chains (amylose) or 70-80% branched glucose (amylopectin). Starch removal is done by grinding the plant material, mixing with water and heating to generate a mash containing up to 20% starch (Naik *et al*., 2010). Next, three enzymes (amylase, pullulanase and glucoamylase) are added sequentially to hydrolyse the released starch to glucose, maltose (two glucose subunits α-1,4 linked) and isomaltose (two glucose subunits α-1,6 linked) which can be fermented by the yeast. Pullulanase and glucoamylase have important debranching activity to hydrolyse α-1,6 linkages in amylopectin (**Figure 1.2.1.1.2**).

Bioethanol is produced mainly by fermentation using the yeast *Saccharmoyces cerevisiae* although processes utilizing *Kluyveromyces marxianus* and bacteria *E. coli* and *Zymomonas mobilis* have also been developed (section 1.2.4). Although *S. cerevisiae* is currently preferred due to its higher ethanol tolerance and production capability, there are benefits to a bacterial process. In *S. cerevisiae*, the organism uses the Embden-Meyerhof-Parnas glycolysis pathway, whereas *Z. mobilis* uses the Entner-Doudoroff pathway that produces 50% less ATP resulting in higher ethanol yields (Yang *et al*., 2016). Other benefits are faster glucose consumption rates due to larger cell surface area and the ability to utilize pentose sugars (Altintas *et al*., 2006). The theoretical maximum yield from 1 gram of glucose is 0.51 grams of ethanol, as much of the carbon is released as carbon dioxide and some used for biomass

production (**Figure 1.2.1.1.3**). The percent conversion achieved in practice is between 40 and 48% (Naik *et al*., 2010).

The 1G biofuels processes, and in particular the grain ethanol processes, are however likely to be discontinued as they do not appear to provide any reduction in greenhouse gas emissions and its ability to sustain modern energy demands with an energy return on investment of 2.2:1 is woefully inadequate compared with current technologies (Gallagher *et al*., 2015; Aro 2016). As this study relates to liquid fuel production, 1G biogas production will not be discussed.

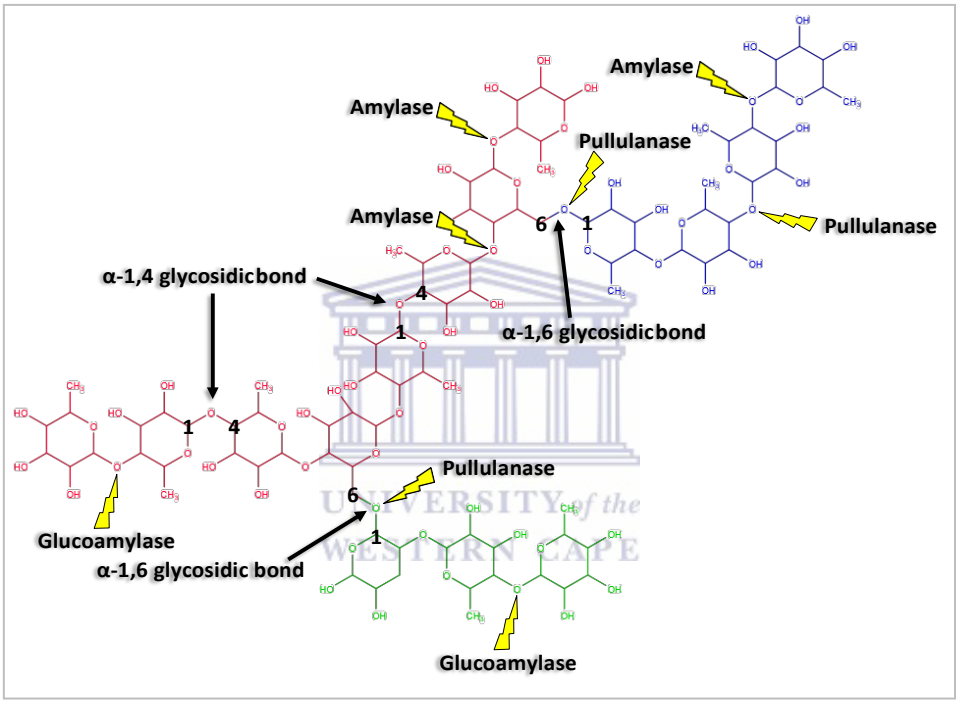


Figure 1.2.1.1.2. Structure of amylopectin and the enzyme activities used to degrade the molecule to its dimer and monomer subunits

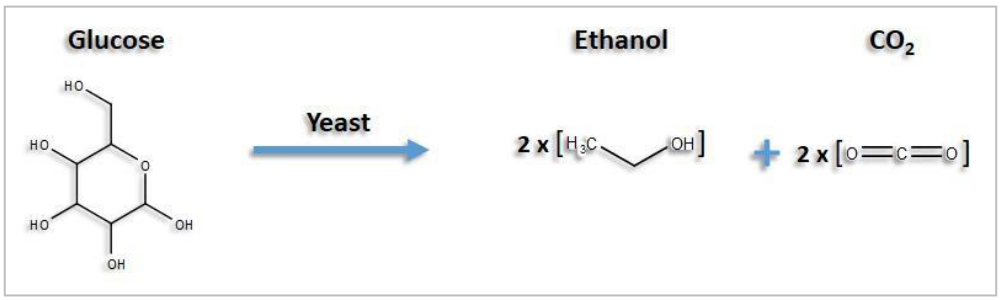


Figure 1.2.1.1.3. General formula for the production of grain alcohol.

1.2.1.2 Second generation biofuels

Second generation biofuels processes (2G) aim to improve on 1G by using the non-edible parts of plants (lignocellulose) as feedstock for biofuels production (Naik *et al*., 2010). 2G is considered carbon neutral as the CO_2 released from their combustion should be equal to the CO_2 sequestered by the biomass during its growth, resulting in no net CO_2 release. This picture is naïve in that it supposes that enough energy is produced from the plant material to overcome the energy requirements of collecting plant material, run production facilities and produce and deliver liquid fuels. Usually the process still requires a fossil fuel energy input, meaning that although not carbon neutral, the process can help mitigate the damage from using fossil fuels alone to produce liquid transport fuels. Although research and development of 2G processes has been on-going for 30+ years, the first commercial 2G plant was only commissioned in July 2013 by INEOS Bio (UNCTAD/DITC/TED/2015/8, http://tinyurl.com/zx8ouv4), but operated only intermittently between 2013 and 2015 and INEOS sold its cellulosic ethanol business in September 2016. As the organisms and techniques for fermentation of the various sugars released from lignocellulose are well developed, the lag in establishing these technologies demonstrates the challenges faced when attempting to liberate fermentable sugars, in an economically competitive manner, from lignocellulose (Mir *et al*., 2014). Thus, the main difference between the 1G and 2G bioprocess is the feedstock employed.

Lignocellulose is composed of various fractions including carbohydrates (cellulose and hemicellulose), lignin, pectin and protein (**Figure 1.2.1.2.1**). Some of the plant materials being targeted as feedstocks for 2G processes include: switch grass, sugarcane bagasse, corn stover, miscanthus and distillers dried grain

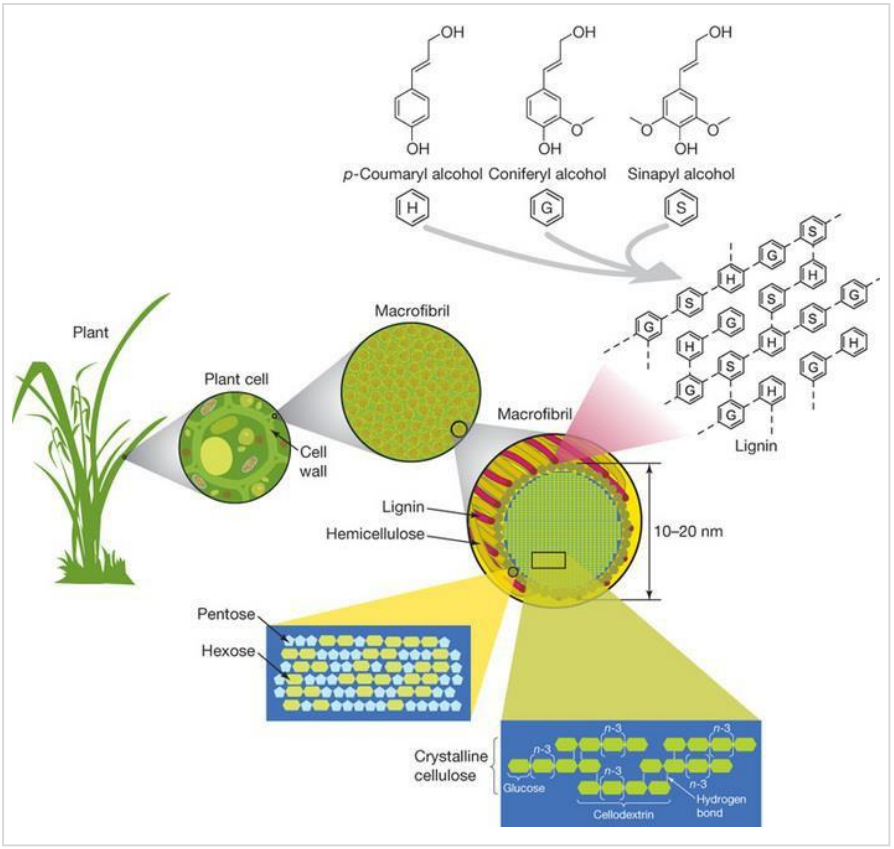


Figure 1.2.1.2.1. Composition of lignocellulose (From Rubin 2008)

(left over from 1G process) (Akhtar *et al*., 2015; Kumar and Sharma 2017). However, any cellulose containing substrate can be employed such as municipal solid waste (Rocha–Meneses *et al*., 2017). The cellulose and hemicellulose fractions are relatively easy to break down to constituent sugars for fermentation, whereas the lignin component is recalcitrant to enzymatic degradation and is often just burned as energy source, rather than forming part of the fermentation process (Rabelo *et al*., 2011). Cellulose is a polymer of β-1,4 linked D-glucose subunits, is a major component of the plant cell wall (Updegraff 1969; Thomas *et al*., 2013) and the amount of cellulose in plants can vary from 35%-45% (**Figure 1.2.1.2.1**; Ververis *et al*., 2004). The polymer can be hundreds or thousands of glucose units in length and, unlike starch (α-1,4 linkage) does not have any branching or coiling and forms straight chains. In plants, these link with neighbouring polymers through hydrogen bonding to produce

microfibrils (crystalline) or can be amorphous. These microfibrils are laid on the surface of the plant wall by cellulose synthase catalytic subunits. These extrude the individual growing polymers from the cell near one another allowing hydrogen bonding to take place (Jarvis 2013; Thomas *et al*., 2013).

The second most abundant natural polymeric carbohydrate, next to cellulose, is hemicellulose. This is a heterogeneous polymer which comprises 15-35% of the plants biomass and is usually composed of a mix of pentose (xylose) and hexose (mannose) sugars as well as uronic acids (**Table 2.1, Figure 1.2.1.2.1 and 1.2.1.2.2**; Gírio *et al*., 2010). The structure consists of a main-sugar backbone (mix of repeating xylose, mannose or glucose subunits) substituted with other sugars and uronic acids. Glucuronoxylan and xyloglucans are predominantly found in hardwoods, while galactoglucomannans are the main hemicelluloses found in softwoods. Arabinoglucuronoxylans and arabinoxylans are the main hemicelluloses in agricultural crops (grasses). This is an important fraction to take advantage of in the process as, without doing so, bioethanol would be too expensive to produce and cannot compete with existing products (Wyman 1999).

Table 1.2.1.2.1. Comparison of cellulose and hemicellulose characteristics

Charcteristic	Cellulose	Hemicellulose
Monomer	Pure glucose	Mixed sugars
Polymer chain length	Long (5µm)	Short
Mol. Weight	High (10000 units)	Low (hundred units)
Polymer topology	Linear	Branched
Side groups substitution	No substitution	On C2, C3, and C6
Polymer morphology	Crystalline + Amorphous	Amorphous
Solubility	Low	High
Reactivity	Less reactive	More reactive
Hydrolysis	Partial	Readily (susceptible)

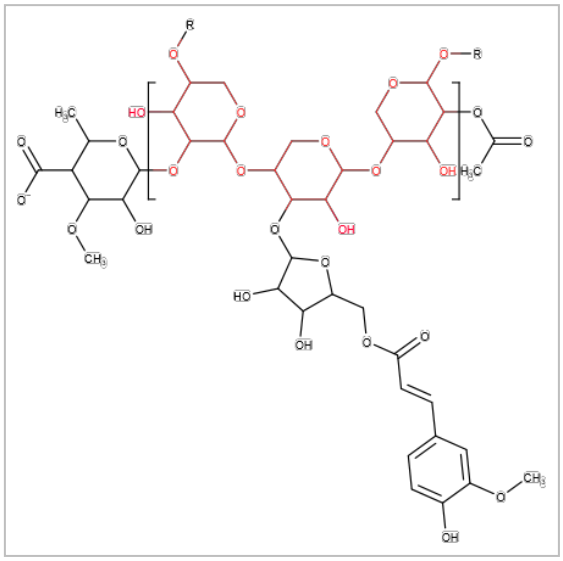


Figure 1.2.1.2.2. Example of polymers in hemicellulose. Structure of xylan with repeated xylose units coloured red.

Processing this biomass into constituent sugars, termed pre-treatment, is normally achieved through a chemical/physico-chemical treatment alone, or in combination with enzymatic degradation (Bhatia *et al*., 2017). Various techniques to break down the plant material to constituent sugars have been developed and each of these have their advantages and drawbacks (**Table 1.2.1.2.2**). Enzymatic hydrolysis, as opposed to physico-chemical hydrolysis of lignocellulose, results in fewer fermentation inhibitors released from the material, especially from the hemicellulose and lignin fractions. Several compounds including ferulic acid, 5-hydroxymethylfurfural, furfural, formic acid, acetic acid, levulinic acid and others are toxic to microbes used to ferment released sugars, and techniques have been developed to cope with these inhibitors (**Table 1.2.1.2.3**; Jönsson *et al*., 2013). The cellulosic to ethanol plant previously owned by INEOS, mentioned earlier, was bought by Alliance, and their process uses a proprietary dry powder catalyst to break down plant material to constituent sugars in minutes without requiring high temperature/pressure, enzymes or acid (https://tinyurl.com/y7cn8fxe). The addition of enzymes to break down the material to constituent sugars represents a significant portion of the bioethanol production cost, due to the enzymes' high production costs (Klein-Marcuschamer *et al*., 2012; Balan 2014). Any technology that can reduce the reliance on enzymes for pre-treatment would be preferred and may make the process more economically feasible.

The work described in subsequent Chapters was done in collaboration with a UK-based biofuel company (TMO) looking to develop a 2G process. Their process relied on steam explosion to open up the highly compact physical structure of the biomass. Following physical grinding of the material, the biomass is subjected to high pressure/temperature steam, often under acidic conditions using sulphuric or phosphoric acid (Schell *et al*., 1998). The steam condenses under high pressure penetrating the material (wetting). The pressure is released rapidly which means that water molecules in the plant material instantly turn to steam and the expanding gas breaks the material apart, thereby increasing the surface area accessible to enzymes. This treatment results in high hemicellulose solubility (release of xylo- and mannooligosaccharides) as well as solubilizing a small fraction of the lignin present (Gírio *et al*., 2010). The next step is enzymatic hydrolysis. This can either be performed by addition of commercial crude enzyme preparations (Novozymes) or by allowing microorganisms, mostly brown and white rot fungi, to grow on the material (Akhtar *et al*., 2015; Kumar and Sharma 2017). A drawback of allowing microorganisms to grow on the material is that some of the material is assimilated into fungal biomass as opposed to being turned into ethanol and hydrolysis times are long. Following steam explosion, the material, is warm, and needs cooling down to allow enzymes that are not thermostable to work on the material. Cooling large volumes comes at considerable financial cost and this is one reason why a high temperature process is beneficial, as the material would not have to be cooled if thermostable enzymes could be added at this point.

The enzymes required for hydrolysis of cellulose include: exoglucanases (EC 3.2.1.4), endoglucanases (EC 3.2.1.74) and β-glucosidases (EC 3.2.1.21). The exoglucanases act in a processive manner on the reducing *and* non-reducing ends of the cellulose polysaccharide chains. This liberates either cellobiose or glucose as end products. Endoglucanases randomly cleave β-1,4 bonds inside cellulose chains generating new ends and β-glucosidases specifically cut the cellobiose dimers to yield glucose monomers. Xylanases (EC 3.2.1.8/37) and mannanases (EC 3.2.1.25/78) are the two classes of enzyme mainly required to break down hemicelluloses. These attack the main chain backbone, however a suite of accessory enzymes is needed to cleave off side chain residues as well as reduce disaccharides produced by these enzymes to monomers. Removal of side chain residues is important to allow the main enzyme classes access to the backbone. These include β-xylosidase, arabinofuranosidase,

acetylxylanesterases, β-mannosidase, 4-O-glucuronoyl methylesterases, α-galactosidases and α-xylosidases.

Table 1.2.1.2.2. Chemical and physico-chemical pretreatment of different lignocellulosic biomass. Adapted from Akhtar *et al*., 2015.

Pretreatment method	Advantage (s)	Disadvantage (s)
Alkaline hydrolysis		
Pre-treatment can be performed at low temperature for a long time with high concentration of alkali.	-Effective in breaking the ester bonds between lignin, hemicellulose and cellulose without fragmentation of hemicellulose.	-High cost of alkaline catalyst -Alteration of lignin structure
Oxidizing agent		
An oxidative pre-treatment employs oxygen or air as catalyst, allows reactor operation at relatively low temperatures and short reactor times.	-Less toxic compound generation -Efficient removal of lignin -Minimizes the energy demand	-High cost of oxidizing agents used.
Organosolv process		
Organic liquid and water is heated to dissolve the lignin and part of the hemicellulose, leaving reactive cellulose in the solid phase.	-Relatively pure lignin recovery -Effective for both hardwood and softwood.	-High capital investment -Formation of toxic inhibitors -Need of solvent recycling
Green Solvents/Ionic liquid		
The chemistry of the anion and cation can be tuned to generate a wide variety of liquids which can dissolve a number of biomass types.	- Enhanced recovery rate - No formation of toxic products	-Deactivate the enzymes -Requires longer pre-treatment time. -Can cause explosion effects.
Microwave-chemical pre-treatment		
Process utilizes thermal and non-thermal effects generated by microwaves in aqueous environments as it breakdown lignin-hemicellulose complex and expose more accessible surface area of cellulose.	-Accelerate reactions during the pre-treatment process -Improve sugar content	-Not feasible for large scale operations. -High cost and slow process
Ultrasound pre-treatment		
Ultrasound greatly enhance the transport of enzyme macro-molecules toward the substrate surface by cavitation and break down cell wall to make the surface accessible for enzymes	-Higher enzymatic hydrolysis	-Not feasible on large scale.
Acid hydrolysis		
Operated either under a very high temperature with dilute acid or under a low temperature with concentrated acid.	-Simple technique -Does not require thermal energy -Effectively hydrolyse hemicellulose with high sugar yield	-High cost of acid recycling -Generate toxic inhibitors
Ammonia fibre explosion (AFEX)		
Biomass is exposed to liquid ammonia at relatively high temperature for a period of few minutes followed by immediate reduction of pressure.	-Increases accessible surface area -Low formation of inhibitors -Effectiveness for herbaceous material and low lignin content biomass	-High cost of ammonia and its recycling -Less hemicellulose solubilisation -Alters lignin structure
CO_2 explosion		
CO_2 can penetrate the minute pores of lignocelluloses by high pressure resulting into disruption of cellulose and hemicellulose structure and makes the surface more accessible to enzymatic attack.	-Availability at relatively low cost -Non-toxic and non-flammable -Easy recovery after extraction -Environmental acceptability.	-Less effect on lignin -Very high-pressure requirement -Too expensive for industrial application
Explosion or autohydrolysis		

Thermo-mechanical force results into disruption of biomass by saturation of biomass with sudden release of pressure generated by steam.	-Cost effective -High yield of glucose and hemicellulose in two-step process	-Incomplete destruction of lignin-carbohydrate matrix -Toxic compound generation
Super critical fluid pre-treatment		
Super critical fluid explosion opened up minute pores of the biomass which increased accessible surface area for subsequent enzymatic hydrolysis.	-Low degradation of sugars -Cost effective -Increases cellulose accessible area	-High pressure requirement -Less effect on lignin and hemicellulose

Table 1.2.1.2.3. Techniques for detoxification of lignocellulose hydrolysates and slurries (Taken from Jönsson *et al*., 2013)

Technique	Procedure
Chemical additives	Alkali [such as $Ca(OH)_2$, NaOH, NH_4OH] Reducing agents [such as dithionite, dithiothreitol, sulfite]
Enzymatic treatment	Laccase Peroxidase
Heating and vaporization	Evaporation Heat treatment
Liquid-liquid extraction	Ethyl acetate Supercritical fluid extraction [such as supercritical CO_2] Trialkylamine
Liquid–solid extraction	Activated carbon Ion exchange Lignin
Microbial treatment	*Coniochaeta ligniaria* *Trichoderma reesei* *Ureibacillus thermosphaericus*

An alternative to complete enzymatic hydrolysis of the substrate prior to fermentation, is to employ microorganisms that naturally (*P. thermoglucosidans*; β-xylosidase, α-N-arabinofuranosidase, endo/exoglucanase activity), or following genetic engineering, are capable of producing glycoside hydrolase enzymes which can aid in the breakdown of the material (section 1.2.4; Bartosiak-Jentys *et al*., 2013; Studholme 2014). This can help to reduce the cost of having to purchase enzymes for pre-treatment, however the trade-off is having some of the energy and carbon from the plant material put into enzyme production rather than producing the desired end product. Once the sugar solution has been prepared it is fermented and distilled in a manner similar to that of 1G processes.

1.2.1.3 Third and fourth generation biofuels

In 3G processes, the feedstocks are engineered crop plants, which allow for easy degradation in the 2G process. Thus, 3G processes sought to improve on 2G through focusing on the type of feedstock used, whereas 2G processes mainly focused on improving the bioconversion of non-engineered plant material. Examples of these are low-lignin content *Eucalyptus* hybrids and sorghum (Pedersen *et al*., 2006; Sykes *et al*., 2015), higher biomass crops, drought tolerant crops as well as engineering plants to produce the enzymes required for their eventual breakdown. These enzymes are mostly from (hyper)thermophiles which allow for their activation during pre-treatment (Mir *et al*., 2014; Mir *et al*., 2017). Fourth generation (4G) processes look to replace lignocellulose derived biofuels with fuels produced by photosynthetic autotrophic (CO_2 fixing) microorganisms. The input would therefore be sunlight, water and carbon dioxide to yield liquid fuels and are thought to be truly carbon negative (Aro 2016). An example of this is the engineering of cyanobacteria to produce farnesene, by introducing a farnesene synthase from Norway spruce into an *Anabaena* sp. (Halfmann *et al*., 2014).

1.2.1.4 Thermophilic ethanologenesis

A high temperature process to produce biofuels would be beneficial for several reasons. Firstly, process kinetics are often faster, not because thermophiles have a higher inherent metabolic rate, but because substrates (particularly polymeric substrates) are typically more accessible and digestible at elevated temperatures. Secondly, thermophilic fermentations are known to generally be less susceptible to contamination with competing microorganisms (*Lactobacillus* sp.), that plague mesophilic fermentations, due to the higher temperature and reduced oxygen availability in the bulk solution. Thirdly, where target products have relatively high volatility (such as ethanol or butanol), high temperature fermentations offer greatly simplified product recovery through continuous 'stripping' of ethanol under low vacuum from the fermentation vessel headspace (Taylor *et al*., 2009; Eram *et al*., 2013; van Dyk and Pletschke 2012). As mentioned earlier this process would also reduce cooling costs, needed to enable the addition of mesophilic enzymes for substrate pre-treatment, as well as having to heat the fermentation broth post-fermentation for distillation (Wang and Wang 2014). High temperature

organisms are mostly dominated by bacteria and archaea, with a few fungal representatives and some animals (Tardigrades). The moderately thermophilic yeast *K. marxianus* is the organism that has received most attention as candidate for establishing a high temperature bioethanol process (Wang *et al*., 2013; section 1.2.4). However, thermophilic bacteria offer several benefits over *K. marxianus* and several potential hosts have been identified that could fulfil this role: *P. thermoglucosidans*, *Clostridium thermocellulum*, *Thermoanerobacter mathranii*, *Thermoanaerobacter ethanolicus* and *Thermoanerobacterium saccarolyticum* (Taylor *et al*., 2009; Jiang *et al*., 2017). Many of these microorganisms, unlike many yeasts and some bacteria, have the ability to ferment polymeric precursors or complex polycarbohydrates (cellulose) in addition to hexose and pentose sugars released from pre-treated material (Sommer *et al*., 2004). A "holy grail" for biofuels production is the so-called consolidated bioprocessing (CBP). This concept would combine the four biological events required for plant biomass conversion to biofuels (production of saccharolytic enzymes, hydrolysis of the polysaccharides present in pre-treated biomass, fermentation of hexose sugars, and fermentation of pentose sugars) in one reactor. No one microorganism found in nature, thus far, can degrade lignocellulose efficiently *as well as* produce high ethanol yields from it. Thus, an efficient lignocellulose degrader has to be engineered to produce ethanol, and *vice versa* (van Zyl *et al*., 2007). Due to their higher rates of native lignocellulose hydrolysis, thermophilic microorganisms are seen as good candidates for a CBP process if they can be engineered to produce and survive in ethanol (Jiang *et al*., 2017).

Of relevance to the research presented in this thesis is the engineering of a microorganism towards ethanol production. Various yeast and bacterial strains have been modified, and this will be discussed in more detail in section 1.2.4.

1.2.2 Pyruvate decarboxylase

The enzyme pyruvate decarboxylase (Pdc) is responsible for the non-oxidative conversion of pyruvate to acetaldehyde, which in turn is used by alcohol dehydrogenase as substrate to convert to ethanol using NADH as reducing agent (König 1998; **Figure 1.2.2.1**). It is thus the key enzyme in homofermentative metabolism, where ethanol is the main fermentation product. Its catalytic mechanism relies on two cofactors, namely thiamine diphosphate (ThDP) and Mg^{2+}, and it thus belongs to a very large family of ThDP dependent enzymes which include pyruvate oxidase, pyruvate dehydrogenase, acetohydroxyacid synthase, benzaldehyde lyase, and several 2-ketoacid decarboxylases (benzoylformate- and phenylpyruvate decarboxylase). The enzyme is often found in plants and yeasts and related enzymes (2-keto-decarboxylases) are often identified in a wide range of bacteria (KdcA – *Lactococcus lactis*; BFD – *Pseudomonas putida*; iPDC – *Enterobacter cloacae*). Although PDC activity has been described in two other bacteria (*Erwinia amylovora* and *Clostridium botulinum*) neither has been confirmed as "true" PDCs and these seem exceedingly rare in prokaryotes with only six having been characterized (Haq 1983; König 1998; Eram and Ma 2013; Buddrus *et al*., 2016).

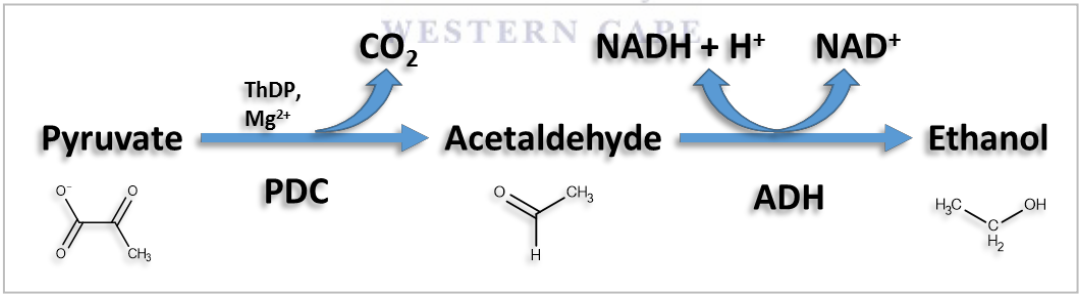


Figure 1.2.2.1. Pdc mediated conversion of pyruvate to ethanol *via* alcohol dehydrogenase (Adh).

Not only do these enzymes have application for metabolic engineering to produce biofuels, they may also be useful in other biotechnology applications such as catalysing the initial step of pseudoephedrine synthesis (Demir *et al*., 2007; Andrews *et al*., 2016) by performing a carboligation reaction to produce (R)-phenylacetylcarbinol (**Figure 1.2.2.2**).

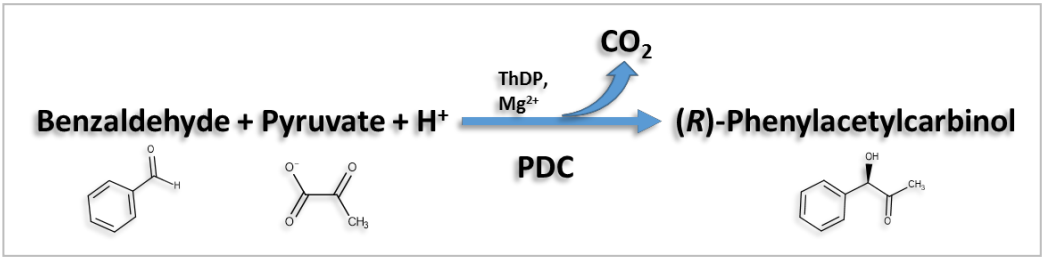


Figure 1.2.2.2. Formation of (R)-phenylacetylcarbionol (pseudoephidrine precursor) through ligation of pyruvate and benzaldehyde by Pdc.

The strategy for improvement in ethanol yield in *P. thermoglucosidans*, described in this thesis, involves the expression a bacterial Pdc and the structure of another bacterial enzyme was solved in this work. An introduction to the enzyme is therefore needed, prior to discussing its use for improving ethanol production and discussing aspects related to the physical structure and biochemical properties of the enzymes described in the Chapters below.

As stated previously, only a handful of bacterial Pdcs have been described. Previous to the work presented in this thesis, the four Pdcs from bacteria (*Zymomonas mobilis* (ZmPDC), *Zymobacter palmae* (ZpPDC), *Acinitobacter pasteurianus* (ApPDC) and *Sarcina ventriculi* (SvPDC) as well as the enzyme from *S. cerevisiae* (ScPDC) had been thoroughly characterized biochemically and three crystal structures had been solved (**Table 1.2.2.1;** Buddurus *et al*., 2016). The work presented in Chapters 2 and 3 describe the characterisation of two new PDCs from *Gluconacetobacter diazotrophicus* and *Gluconobacter oxydans* and therefore they will not be discussed in detail here. All four bacterial Pdc crystal structures described thus far show that the biologically active molecule is a homotetramer which can best be described as a dimer of dimers, as shown in **Figure 1.2.2.3**. Each monomer consists of three domains: R-domain or regulatory domain, PYR-domain or pyrimidine ring binding domain and the PP- or pyro(di)phosphate binding domain (Muller *et al*., 1993). Their overall structures are highly similar with r.m.s.d values no greater than 0.7Å when compared with one another (**Table 1.2.2.1**). The active sites are formed at the interface of the monomers. This means that each monomer contains "half" of an active site. The ThDP molecule and Mg^{2+} which are essential cofactors, and which play pivotal roles in catalysis are part and parcel of the active site environment. The ThDP molecule is coordinated by

residues in both monomers which come together to make up an active dimer. The role of Mg^{2+} is mainly to interact with ThDP *via* the pyrophosphate moiety, thereby positioning the cofactor correctly in the active site (**Figure 1.2.2.4 and 1.2.2.5**). A feature that seems to be common to all ThDP dependent enzymes is some form of chemical communication that exists between the active sites in one dimer (Frank *et al*., 2004). As is the case for many ThDP dependent enzymes, this mechanism of communication is thought to be a network of water molecules which connects the two active sites and acts as a proton relay (Buddrus *et al*., 2016). This water network has been observed in all bacterial Pdc structures solved to date and it was demonstrated that only one of the two active sites in a dimer can be involved with catalysis at any one-time, with this proton relay thought responsible for the cycling (Schrö der-Tittmann *et al*., 2013).

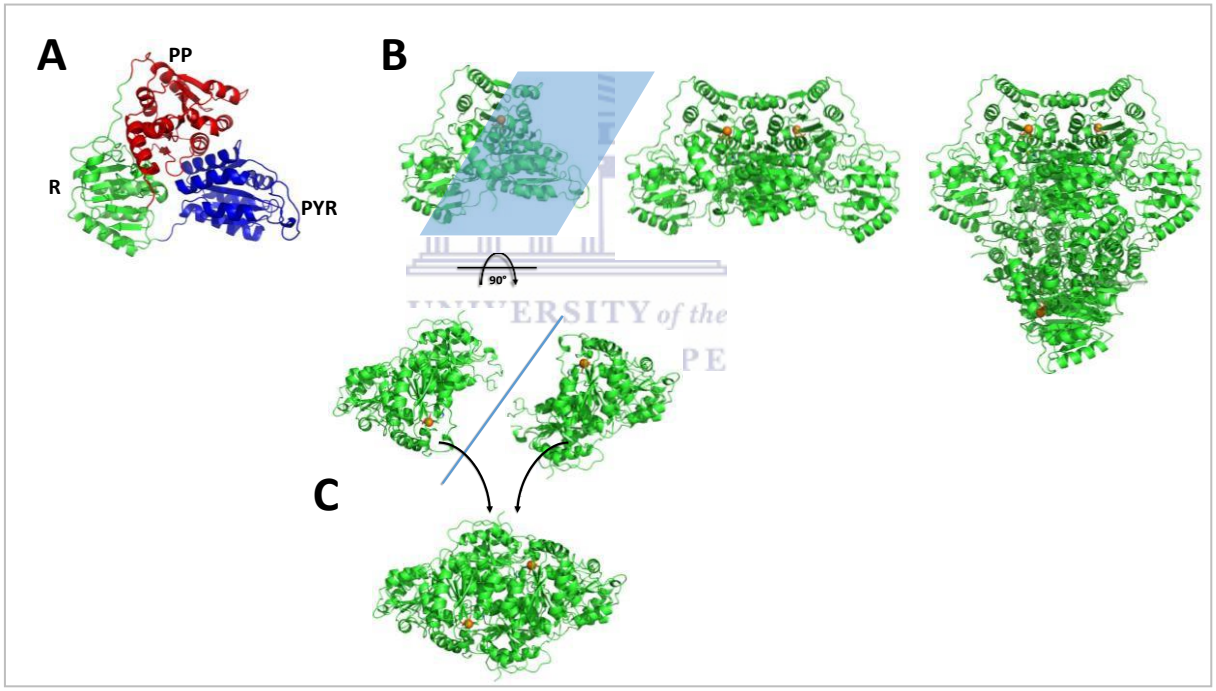


Figure 1.2.2.3. Tertiary and quaternary structure of *Z. palmae* Pdc (5EUJ). A) The three domains of the Pdc monomer are coloured in blue (Pyr), red (PP) and green (R) respectively. B) Shows the monomer, dimer and tetramer respectively from left to right. C) Shows how two monomers come together to form the dimer (Top view relative to side view in A). The blue plane indicates the interaction surface of the two monomers. All three-dimensional representations were created with PyMol 2.0.7

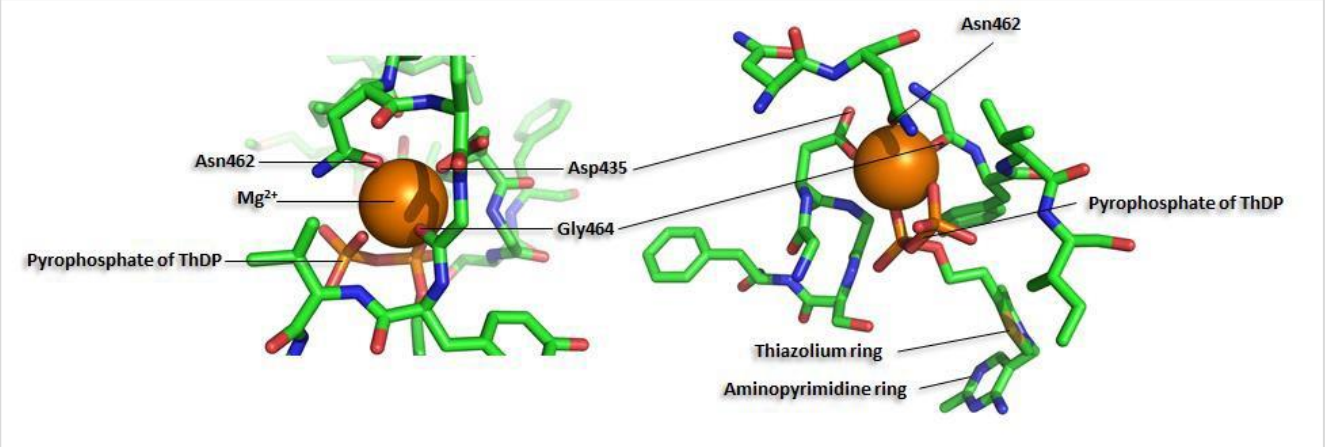


Figure 1.2.2.4. Coordination of Mg^{2+} by residues from Pdc and its placement relative to the ThDP cofactor in structure 5EUJ

Mutagenesis studies have shown that the catalytic mechanism of Pdc and related enzymes, such as benzoylformate decarboxylase (Bfd), are conserved with only a few residues in the active site selecting the substrate of choice (Siegert *et al*., 2005). Whereas Pdc from *Z. mobilis* has two isoleucine (Ile) residues (Ile472 and Ile 476) that help coordinate pyruvate in the active site, *Pseudomonas putida* Bfd has an alanine (Ala460) and phenylalanine (Phe464) in the corresponding positions. Exchange of the Ile472 residue in ZmPDC with Ala resulted in an enzyme capable of decarboxylation of benzoylformate, where it could not previously. Conversely, substitution of Ala460 with tyrosine resulted in a Bfd variant with decarboxylase activity on pyruvate, whereas the wild type enzyme has none (Yep and McLeish 2009; Andrews *et al*., 2016).

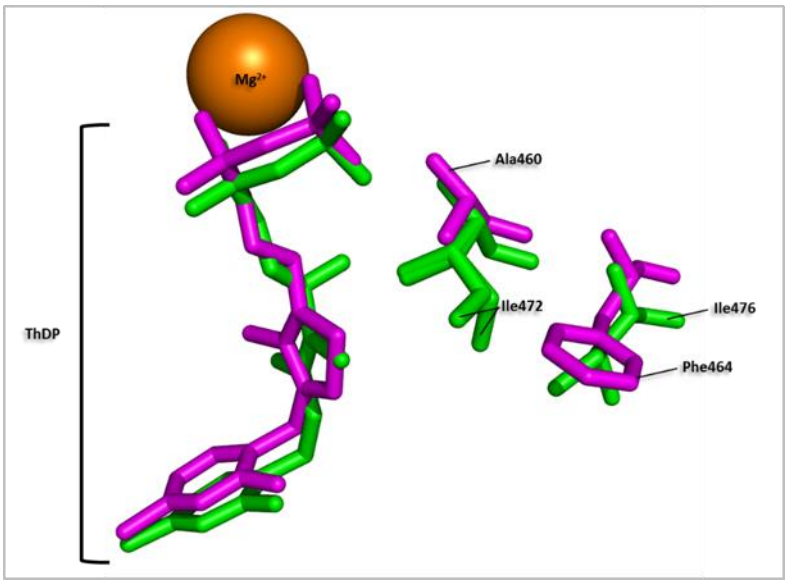


Figure 1.2.2.5. Overlay of Pdc (green; 1ZPD) and Bfd (magenta; 1BFD) showing substrate binding residues. The close superposition of the ThDP cofactors demonstrate the similarities of the active site. Adapted from Siegert *et al*., 2005

One of the features which differentiates the bacterial enzymes from their yeast counterparts is that the tetramer is "locked" in the "closed" or active conformation, whereas the *S. cerevisiae* Pdc appears to alternate between an "open" conformation and the catalytically active "closed" conformation (**Figure 1.2.2.6**). The ability to switch between the two conformations is brought about, by the way in which the dimers interface to form the tetramer. In the bacterial enzymes, the dimers are rotated relative to one another by ~78° when compared with the *S. cerevisiae* enzyme and that of *Kluyveromyces lactis* (**Figure 1.2.2.6**). This gives the bacterial enzyme a much greater area of interaction between the two dimers (greater number of hydrogen bonds and salt bridges), reducing its ability to undergo large conformational changes (Dobritzsch *et al*., 1998) and likely adding to its stability.

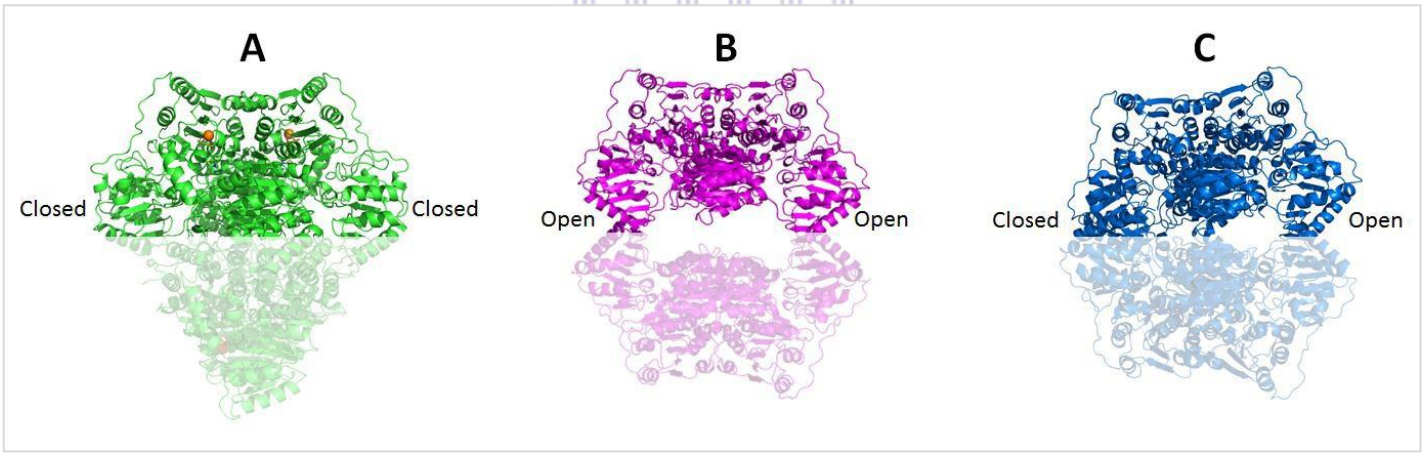


Figure 1.2.2.6. Quaternary structures of A) *Z. palmae* PDC (5EUJ) compared with that of B) *S. cerevisiae* (1PVD) and the activated conformation of C) *K. lactis* (2VK4). The bottom dimers are shaded to help indicate the orientation of the dimer sets relative to one another as well as the conformational change upon substrate analogue binding

When looking at the kinetic properties of the bacterial enzymes, we see that they all display Michaelis-Menten kinetics except for the enzyme from the Gram-positive organism *Sarcina ventriculi* (**Table 1.2.2.1**). What is striking about these enzymes, is that for the most part their pH optima, affinity for the substrate and catalytic efficiencies are much the same under similar assay conditions (Buddrus *et al*.,

2016). Their thermostability profiles do differ with the *G. diazotrophicus* enzyme being the least stable, and the *Zymobacter palmae* PDC the most stable.

Table 1.2.2.1. Comparison of all bacterial PDC biochemical characteristics. Adapted from Buddrus *et al*., 2016

	ZpPDC	ApPDC	ZmPDC	GdPDC	GoPDC	SvPDC
Gram status	Negative	Negative	Negative	Negative	Negative	Positive
Amino-acid identity (%)	Reference	73	63	71	67	31
Temperature optimum (℃)	65	65	60	45–50	53	N/A
Temperature dependence of activity retention	60 ℃, 100% 65 ℃, 80% 70 ℃, 0%	50 ℃, 100% 60 ℃, 65% 65 ℃, 45% 70 ℃, 5%	45 ℃, 85% 60 ℃, 65% 65 ℃, 45% 70 ℃, 0%	N/A (half-life at 60 ℃, 0.3 h)	55 ℃, 98% 60 ℃, 70% 65 ℃, 40%	45 ℃, 95% 50 ℃, 0%
Kinetics	Michaelis–Menten	Michaelis–Menten	Michaelis–Menten	Michaelis–Menten	Michaelis–Menten	Sigmoidal
V_{max} (U mg ^{1})	165 (pH 6.5) 116 (pH 6.5) 130 (pH 6)} 140 (pH 7)}	110 (pH 6.5) 97 (pH 5)} 79 (pH 7)}	121 (pH 6.5) 100 (pH 6) 78 (pH 7) 120 181	20 (pH 5) 39 (pH 6) 43 (pH 7)	57 (pH 5) 47 (pH 6) 125 (pH 7)	103 45 (pH 6.5) 35 (pH 7)
K_m ($S_{0.5}$) (mM)	0.67 (pH 6.5) 2.5 (pH 6.5) 0.24 (pH 6) 0.71 (pH 7)	2.8 (pH 6.5) 0.39 (pH 5) 5.1 (pH 7)	1.3 (pH 6.5) 0.43 (pH 6) 0.94 (pH 7) 0.31 (pH 6) 1.1 0.4 (pH 6)	0.06 (pH 5) 0.6 (pH 6) 1.2 (pH 7)	0.12 (pH 5) 1.2 (pH 6) 2.8 (pH 7)	13 5.7 (pH 6.5) 4.0 (pH 7)
PDB entry	5euj	2vbi	1zpd	4cok	NA	NA
GenBank gene	AF474145	AF368435.1	M15393.2	KJ746104.1	KF650839.1	AAL18557.1
GenBank protein	AAM49566.1	AAM21208.1	AAA27696.2	AIG13066.1	AHB37781.1	AF354297.1
R.m.s.d.	Reference	0.70	0.70	0.62	N/A	N/A
Q scores	Reference	0.94	0.90	0.94	N/A	N/A

The catalytic mechanism of PDCs, and ThDP dependent enzymes in general, has been extensively studied (Kern *et al*., 1997, Zhnag *et al*., 2004, Lie *et al*., 2005, Baykal *et al*., 2006, Brandt *et al*., 2009, Nemeria *et al*., 2009, Chakraborty *et al*., 2009, Meyer *et al*., 2010) even so, the exact mechanism has not been solved.

The latest work on the reaction mechanism of ThDP-dependent enzymes involved the study of transketolase, the enzyme responsible for the interconversion of sugars D-xylulose-5-phosphate and D-erythrose-4-phosphate to D-fructose-6-phosphate and D-glyceraldehyde-3-phosphate (Lüdtke *et al*., 2013; Nauton *et al*., 2016). These studies have demonstrated that ThDP-dependent enzymes hold the substrate-cofactor intermediate in a destabilizing position, straining the substrate bond to be broken. This was indicated by way of the substrate carbonyl carbon, covalently attached to the C2 of the

thiazolium ring of ThDP, being out of plane with this ring by ±22°. The ThDP cofactor usually assumes a "V"-conformation, once bound to the enzyme, with the methylene joining the aminopyrimidine (AP) and thiazolium rings at the bottom of the "V" (Andrews *et al*., 2013). At least for transketolase this conformation puts the N4′ atom of the AP ring in close proximity to a OH group on the substrate which is a highly energetic and unfavourable contact. This interaction likely further helps to push the substrate and sessile bond out of plane.

In all ThDP-dependent enzymes the AP ring in ThDP exists as a mesomeric structure with the iminopyrimidine form (IP) (**Figure 1.2.2.7**; Paulikat *et al*., 2017). The catalytic cycle starts with the abstraction of a proton from the N4´ of the of the AP ring of ThDP. The enzyme residue responsible for "starting" this process has been debated over, but not yet unequivocally established (Meyer *et al*., 2010; Paulikat *et al*., 2017). In ZmPDC two glutamate residues are positioned within hydrogen bond distance from either the N1´ of the AP ring (Glu50) or from the pre-decarboxylation intermediate LThDP (Glu473). Early data indicated that Glu473 does regulate the tautomeric equilibrium (AP-IP) of the cofactor as substitution of Glu473 with glutamine resulted in a perturbed equilibrium (AP-IP) and the enzyme failed to bind the substrate analogue acetylphosphinate (Meyer *et al*., 2010). Glu473 has however been shown to play multiple roles during catalysis, particularly in substrate binding, decarboxylation and product release (Meyer *et al*., 2010). The Glu50 residue, highly conserved in ThDP-dependent enzymes, is however currently considered as the cofactor activating residue. Its protonation is proposed to lead to protonation of the N1´ of the AP ring. This results in a knock-on effect which leads to deprotonation of N4´ and subsequently C2 of the thiazolium ring, thereby generating the ylide. The ylide executes nucleophilic attack on the carbonyl carbon of the substrate, pyruvate, in this case. This generates the first covalent cofactor-substrate intermediate, 2-α-lactylthiamin diphosphate (LThDP). Decarboxylation, driven by Glu473 through a suspected, but as yet unexplained, stereoelectronic effect results in formation of the second major intermediate, 2-hydroxyethyl-ThDP (HEThDP). Protonation of this intermediate by Glu473, together with Asp27 and one of the highly conserved active site residues His113, releases the product (acetaldehyde) and regenerates the ylide (Meyer *et al*., 2010). This allows another round of catalysis to be initiated.

An interesting observation was made by Meyer and co-workers where they gave structural and spectroscopic evidence to show that the C2 position of the thiazolium ring of ThDP in pyruvate oxidase from *Lactobacillus plantarum* had a pair of unpaired electrons as an intermediate in the reaction scheme (Meyer *et al*., 2013). This arrangement is referred to as a carbene (neutral carbon atom with a valence of two and two unshared valence electrons). These are normally highly reactive species which, in the case of ThDP, appears to be stabilized in that form by the active site environment. It may, and probably does, exist in equilibrium with the mesomeric carbanion zwitterion form of C2 (**Figure 2.15**), either of which can initiate nucleophilic attack on the 2-keto substrate. The investigators could see this pre-reaction state due to their use of phosphate as a non-reactive analogue for pyruvate. In the actual catalytic cycle this intermediate would exist for a fraction of a second prior to covalent binding of C2 to the alpha carbon of pyruvate. This was the first demonstration of the existence of such an arrangement in nature and it may point to this mechanism being in operation in all ThDP-dependent enzymes.

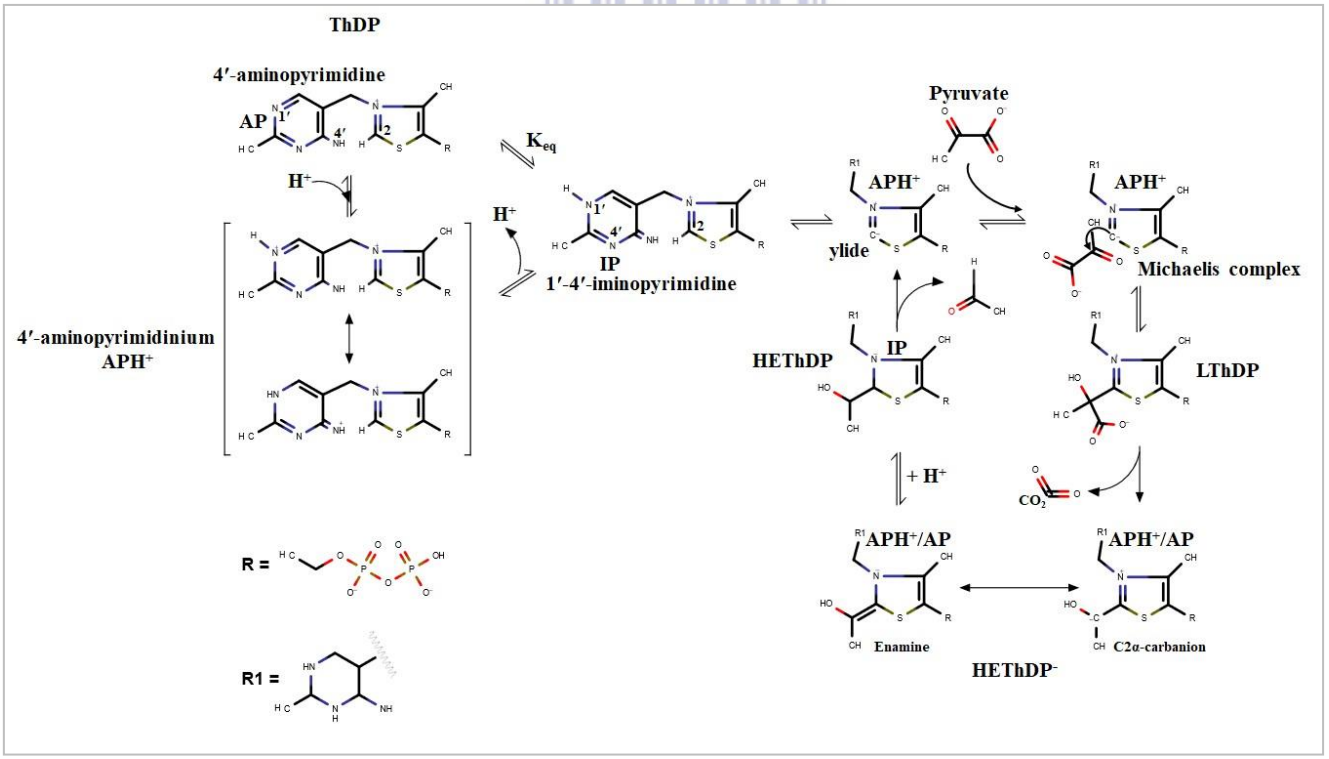


Figure 1.2.2.7. Reaction scheme of ThDP-catalyzed non-oxidative decarboxylation of pyruvate. Adapted from Meyer 2010.

The conservation of structural features and biochemical characteristics seems to suggest that, at least for the enzymes from Gram negative hosts, the selective pressures on them are much the same, indicative of the bacterial hosts' lifestyles (*Zymomonas mobilis* - isolated from plant sap and spoiled

beer; *G. diazotrophicus* - plant endophyte; *G. oxydans* - found in flowers, fruits, garden soil, alcoholic beverages, cider; *Acetobacter pasteurianus* - sugar-rich substrates such as such as fruits, flowers, and vegetables, *Zymobacter palmae* – palm sap).

The benefit of having enzyme kinetic data as well as structural information, is that these can shed light on the catalytic mechanism and substrate range. This is useful from a fundamental biology perspective, but also to know how the enzyme may best be utilized or how to modify it to suit an application. In the case of Pdc, both substrate range and thermostability are areas where investigators would like to make changes or improvements to the enzyme. The bacterial enzymes which, apart from the *S. ventriculi* enzyme (Lowe *et al*., 1992), are not affected by substrate activation and which have higher thermostabilities and activities compared with their yeast and plant counterparts are particularly attractive for engineering purposes (Buddrus *et al*., 2016). Dual function pyruvate ferrodoxin oxidoreductase / pyruvate decarboxylase enzymes from several thermophilic archaea have been described, opening the possibility of using these for thermophilic ethanologenesis. Some of their biochemical characteristics however (low PDC activity, high pH optima and oxygen sensitivity), make them unsuitable for engineering of certain ethanologenic strains that operate under microaerobic conditions or low temperature (*S. cerevisiae*) (Eram *et al*., 2014).

1.2.3 Heterologous protein expression

Heterologous expression of proteins has been used to produce proteins and peptides such as insulin to harvest from the cell and use in treatment of diabetes, as well as for strain improvement such as expression of glycoside hydrolases for improved lignocellulose degradation by ethanol producing strains or enzymes that take part in metabolic pathways for the production of desired metabolites (Lambertz *et al*., 2014; Baeshen *et al*., 2014). Anfinsen proposed in 1973 that a polypeptide contains all the necessary information for folding into its final functional form and this has remained unchallenged (Anfinsen 1973). The successful expression of bacterial proteins, is however a highly regulated event and influenced by several factors. Evolution has selected the gene content (codons),

regulatory network, RNA polymerase, chaperones, ribosomes, tRNA pool ratio, temperature, secretion signals, genomic location etc. to result in optimal expression of proteins under the range of physiological conditions the host finds itself (Ardell and Kirsebom 2005; Couturier and Rocha 2006; Yona *et al*., 2013; Englaender *et al*., 2017). The factors affecting protein expression and their strength has therefore evolved to increase the fitness of the cell, rather than to achieve the highest gene expression. The heterologous expression of non-native proteins in a particular host often fails due to one or several of these factors being sub-optimal for its expression. To overcome the difficulties with heterologous expression several aspects of protein expression have been investigated and modifications made, either to the promoter sequence, the gene sequence or the expression host to enable proper expression. Various aspects that play a role in protein expression will be discussed below as well as the modifications made to ensure successful expression.

Promoter sequences

A gene promoter is a nucleotide sequence element, usually found directly upstream of the gene. This sequence serves to act as signal to recruit the enzyme complex RNA polymerase and associated sigma factors, from which to start transcription (making a messenger RNA copy of the gene) of the gene. In bacteria, these can roughly be divided into four motifs: -10 region, -35 region, UP element and discriminatory element (**Figure 1.2.3.1**; Haugen *et al*., 2008; Browning and Busby 2016). They can further be classified as constitutive (expressed continually) or inducible (only express when a given signal is present). Several inducible promoter systems have been developed for *E. coli*. One, is a strong phage T7 promoter employed in the pET expression system, which is a good example of an inducible promoter. In this system, engineered *E. coli* hosts express the T7 RNA polymerase from a *lacUV5* (IPTG inducible) promoter integrated on its chromosome. The genes of interest cloned behind a T7 promoter on plasmid vectors (pET) are then transcribed from this promoter by the highly promoter specific T7 RNA polymerase once IPTG inducer is introduced into the growth medium (Marschall *et al*., 2017). If the *E. coli* host has an intact lactose operon, adding inducer will result in an all or nothing response, whereas in certain host strains (BL21(DE3) Tuner™; *lacZY*$^{-}$), the expression can be tuned depending on the amount of inducer added (Hartinger *et al*., 2010). Several other *E. coli* systems that

rely on induction with sugars (arabinose, rhamnose) and osmotic shock have also been developed (Marschall *et al*., 2017).

It was demonstrated that the transcription level is likely more important for correct protein expression than codon usage (discussed below), at least for membrane proteins in *E. coli* (Claassens *et al*., 2017). The authors speculate that reduced transcription leads to reduced translation of mRNA to polypeptide, and this reduces the load on chaperones which have to help proteins fold thereby allowing enough time for correct protein folding. In this respect promoter modifications should be considered when attempting to optimize heterologous protein expression. The importance of choosing a promoter that expresses the protein at the "correct" level so as not to overload the cellular machinery was also shown for proteins expressed in the yeast *Yarrowia lipolytica* demonstrating that this likely should be considered a general concept in heterologous protein expression (Dulermo *et al*., 2017).

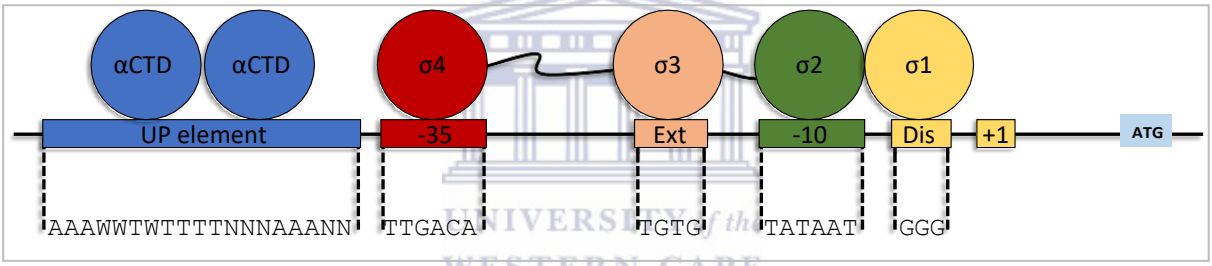


Figure 1.2.3.1. The principal DNA elements recognized by RNA polymerase in bacterial promoters include the UP element (positions –37 to –58, if the transcriptional start site is denoted +1), the –35 element (positions –35 to –30), the extended –10 element (Ext; positions –17 to –14), the –10 element (positions –12 to –7) and the discriminator element (Dis; –6 to –4). The exact positions of each element can vary according to the particular promoter. The regions of the sigma factor (domains 1-4) and of the carboxy-terminal domain of the α-subunit of RNA polymerase (αCTD) that contact these promoter elements are shown. ATG indicates the translational start. Adapted from Browning and Busby 2016.

Recently, a suite of semi-synthetic promoters both inducible (P*xylA*) and constitutive, based on the promoter of *Geobacillus* sp. GHH01 *groESL* operon, for controlled gene expression in *Geobacillus* and *Parageobacillus* species, was developed (Pogrebnyakov *et al*., 2017). Promoters of various strengths have also been developed for other microbial hosts such as *E. coli*, *Lactobacillus* sp. and *S. cerevisiae* (Jensen and Hammer 1998; Braatsch *et al*., 2005; Nevoigt *et al*., 2006; Tauer *et al*., 2014). In the case of the constitutive promoters developed for *Parageobacillus*, several modifications were made to achieve a range of promoter strengths. The CIRCE regulatory sequence was deleted and the sequences between and around its -35 and -10 regions were randomized using a degenerate oligonucleotide

sequence. Using their suite of promoters, they showed promoter dependent expression of super folder green fluorescent protein in *P. thermoglucosidans* which allowed a 76-fold increase in expression. For metabolic engineering, constitutive promoters are often preferred, however depending on the process conditions and pathways present, as well as the flux through these in the cell, an inducible promoter may be preferred. The application of a particular host and promoter combination is usually determined empirically (Jensen and Hammer 1998). These engineered promoters should allow researchers to match promoter strength to metabolic flux for optimal strain performance in this platform organism.

Codon usage

The genetic code is degenerate in that several nucleotide triplets (codons) can code for the same amino acid to be incorporated into a protein sequence. During translation of mRNA to protein, transfer RNA, to which is bound a particular amino acid (aminoacyl-tRNA), enters the ribosome and binds to its cognate anticodon and the attached amino acid is transferred to the growing polypeptide chain. It has been demonstrated that there is a relationship between the frequency with which particular codons occur in gene sequences and the pool of cognate aminoacyl-tRNAs in a particular organism (Ikemura 1985; Spencer and Barral 2012). Most organisms show a bias towards certain codons in their gene sequences with a corresponding aminoacyl-tRNA ratio. This results in the rate at which mRNA translation proceeds (speed at which the ribosome moves along the mRNA) at rare or infrequently used codons versus frequently used codons. Rare codons refer to those that occur infrequently in gene sequences from a particular organism and for which there is a matching low aminoacyl-tRNA pool inside the cell. The reason for the slowdown is the time it takes the cognate aminoacyl-tRNA to find its way into the A-site of the ribosome. If its concentration is low inside the cell, it will take longer to find its way into the A-site. It has been demonstrated that clusters or single rare/infrequently used codons do play a role in protein folding by slowing the rate of translation allowing the immature polypeptide produced up to that point, to adopt its correct secondary/tertiary structure, or the correct intermediate (Quax *et al*., 2015). The rate at which ribosomes initiate translation of mRNA is also controlled by rare codons at the 5-prime end the mRNA molecule, which further modulates the density of ribosomes on a particular mRNA molecule (Plotkin and Kudla 2011). The role that codon usage plays in protein expression is

such that if selected correctly, expression of a protein can be increased by 1000-fold, just by adjusting the codon usage profile (Gustafsson *et al*., 2004).

The *E. coli*-based pET expression system has also been modified to take codon bias into account. The pRARE vector, from which several tRNAs that recognize codons which are often rare in *E. coli* (AGG, AGA, AUA, CUA, CCC, GGA) are expressed, was designed to change the ratio of these aminoacyl-tRNAs in *E. coli* and facilitate improved protein folding. This approach works for those proteins where these codons are not rare in their native host, however if rare codons and translational pausing is required for correct folding, these are not accommodated.

A modern approach is to chemically synthesize genes that have their codon usage modified (codon optimized) such that it still encodes the same polypeptide, but the codon usage frequency is altered relative to the expression host being targeted (Claassens *et al*., 2017). To improve protein expression several algorithms have been developed: Codon Adaptation Index (CAI), the tRNA Adaptation Index (tAI), Competition Adaptation Index (CompAI) and the Effective Number of Codons (Nc) (Dilucca *et al*., 2015); to indicate the probability of the proteins' expression in a particular system. All of these attempt to condense a range of factors (codon frequencies; reference gene sets; deviation from a postulated distribution; information theory; interactions among tRNAs) to generate this probability, despite several other factors also being involved in protein expression such as genomic location, expression frequency and gene length (Dilucca *et al*., 2015). Commercial gene synthesis providers have developed their own codon optimization algorithms and one of the central tenets is the replacement of rare/infrequently used codons with frequently used ones in the heterologous host used. This again does not account for the presence of rare codons in specific regions of the sequence (in its native host) to allow for translation slowdown to assist protein folding, and the application of these algorithms often result in a mixed bag of success. Angov and coworkers proposed a new approach to codon optimization for heterologous expression (Angov *et al*., 2008). Their approach was to attempt to copy the codon usage frequency of the gene from its native host in the heterologous host to mimic as closely as possible the translation speed of the protein in the native host, in the new host (**Figure 1.2.3.2**). They demonstrated the usefulness of their approach by improving the expression of a *Plasmodium falciparum*

protein, for vaccine development, in *E. coli* showing up to 1000-fold increase in soluble protein expression (Angov *et al*., 2008). The method has now successfully been applied to expression of membrane- and fluorescent proteins in *E. coli* (Tian *et al*., 2017; Claassens *et al*., 2017).

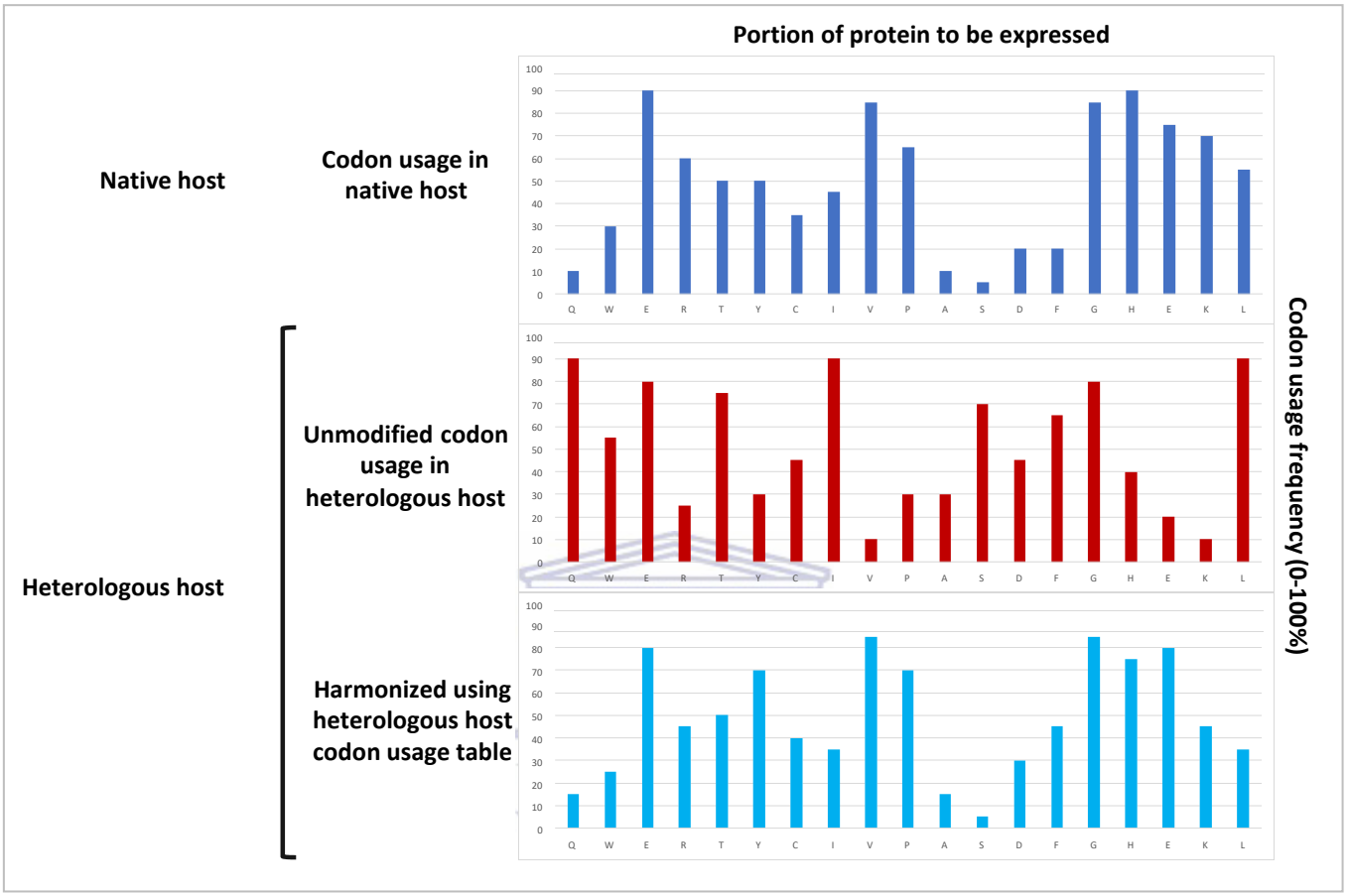


Figure 1.2.3.2. Comparison of codon usage frequencies (in percent) over a segment of a fictitious protein in both the native and heterologous host prior to adjustment (Top two graphs). The bottom graph shows the codon usage frequency adjustment using the heterologous host codon usage table, which is more similar to what it was in the native host (Top graph). Adapted from Angov *et al*., 2008.

Chaperones

Many cellular proteins are incapable of folding on their own and require the assistance of chaperones to do so (Lin and Rye 2004). Two ATP-dependent protein-folding machines, heat shock protein (Hsp70; DnaK) and the chaperonins (CPN) are present in the majority of cells (**Figure 1.2.3.3**). The CPNs are ubiquitous, hollow nanomachines with two cavities that open and close to encapsulate and actively fold non-native proteins, and three distinct groups of these chaperones have now been described (An *et al*., 2017). The difference between the classes is that group I requires a separate GroES co-chaperone,

whereas this function (lid) is built into the GroEL of group II (An *et al*., 2017). Group III chaperones are similar to the proteins from group II in that the lid structure is built in. The differences between group II and group III are mostly mechanistic, dictated by structural differences. These also make them fall into a monophyletic group, separate to the group I and II chaperones. A second difference is that group III chaperones, with one exception, are encoded in the same operon as DnaK, suggesting a closer association between DnaK and these chaperones than between group I/II and DNaK (An *et al*., 2017).

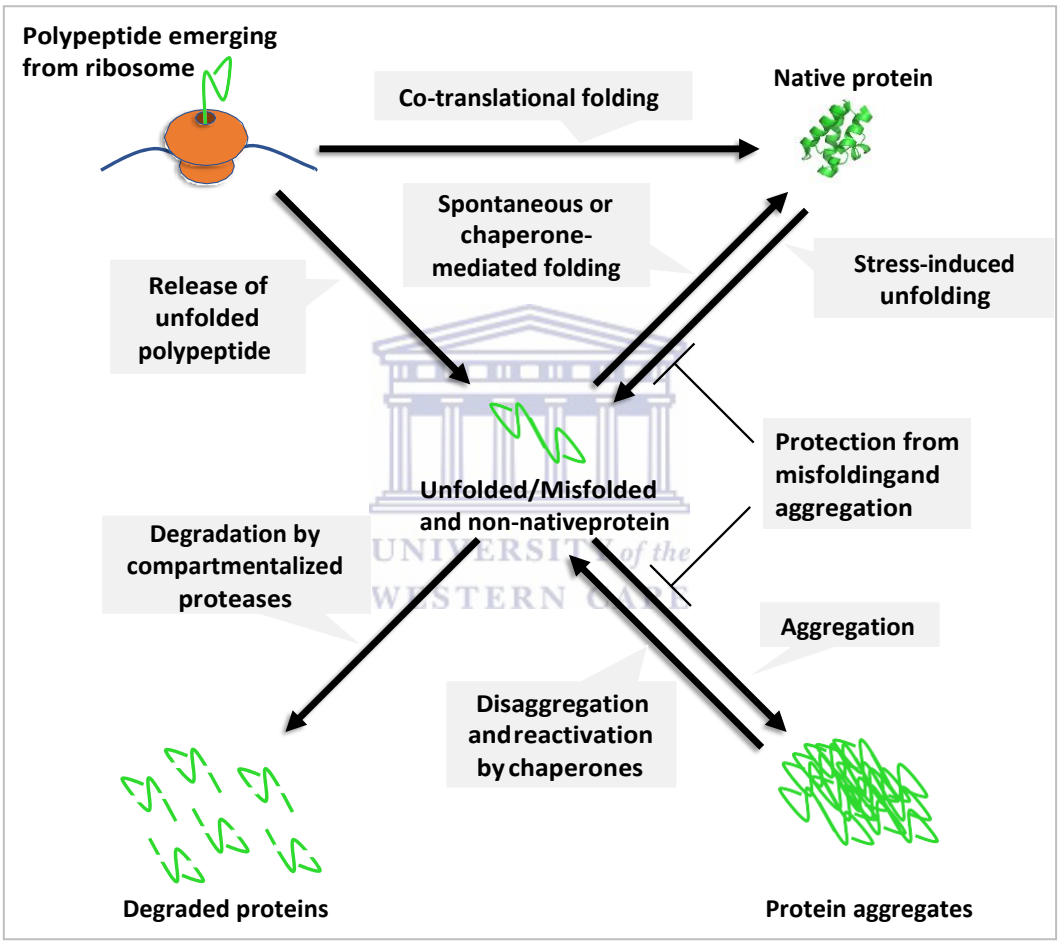


Figure 1.2.3.3. Folding pathways for proteins in the bacterial cytosol. The protein quality control network combines the actions of many chaperones to facilitate the folding of proteins into their native conformations, to prevent misfolding and to 'reactivate' misfolded and aggregated proteins. Some molecular chaperones slow or prevent protein misfolding and aggregation. If the chaperone network becomes overwhelmed during stress, non-native proteins may form large, amorphous aggregates. However, other chaperones can extract and unfold polypeptides from aggregates, thus providing another chance for proper folding. Adapted from Doyle *et al*., 2013.

The best studied is the GroEL/ES from *E. coli* (group I) and this protein complex resembles a "pot" and "lid", with GroEL (large subunit) being the "pot" and GroES (small subunit) forming the "lid" (**Figure**

1.2.3.4). Each subunit of GroEL binds an ATP molecule leading to conformational changes that allows the binding of GroES (Clare *et al*., 2012). These structural changes prevent ATP binding by the opposing ring of GroEL, thus each chamber works in turn to assist protein folding. For group I CPNs which consist of seven subunits each protein folding cycle consumes at least 7 ATP molecules, meaning that chaperone assisted protein folding is a highly energy expensive process in the cell. ATP is not required for folding inside the chamber, but rather advances the protein complex through its cycle of opening and closing. The exact mechanism by which proteins are assisted in folding once inside GroEL is not yet understood. However, it is thought that the inside cavity of GroEL and bottom of the GroES lid, makes a highly hydrophilic environment which forces the hydrophobic proteins of the nascent protein (buried inside of the correctly folded protein) to self-associate thereby promoting protein folding. Although it has been demonstrated that overproduction of chaperonins GroEL and GroES in *E. coli* promote the folding and assembly of a functional El subunit of the mammalian branched-chain α-keto acid dehydrogenase complex (Wynn *et al*., 1992), as well as for many other proteins, its involvement in the folding of bacterial Pdcs has never been investigated. Given their monomer size, it is expected that these proteins are likely to require chaperones to fold correctly.

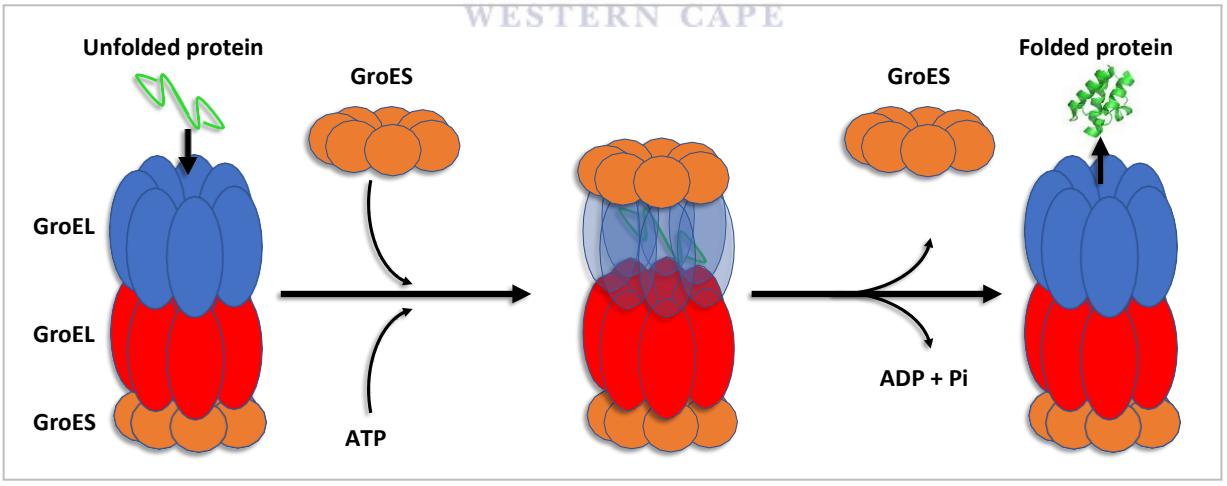


Figure 1.2.3.4. Simplified model of GroEL/ES operation in assisting folding of nascent or stress-denatured polypeptides**.**

Another important chaperone is DnaK (Doyle *et al*., 2013). This protein helps nascent proteins to fold as well as with remodelling of protein complexes. It is composed of an N-terminal protein (substrate biding domain) and a C-terminal nucleotide (ATP) binding domain. The substrate binding domain

usually recognizes a seven-amino acid stretch in the polypeptide consisting of 4 to 5 hydrophobic residues flanked by basic residues (Baneyx and Palumbo 2003). The substrate binding domain can be further subdivided into two regions: i) a "lid" and ii) "substrate binding domain". The catalytic cycle starts by the C-terminal domain being loaded with an ATP molecule. The "lid" domain undergoes large conformational changes upon substrate binding, closing around the unfolded protein. Once this happens, the bound ATP molecule is hydrolysed, and ADP released, and a new ATP molecule bound. This leads to the reversal of lid closure and release of the native protein. Each catalytic cycle therefore consumes one ATP molecule. DnaK often works with two co-chaperones (DnaJ and GrpE) which helps DnaK with ATP hydrolysis and directing unfolded proteins to DnaK. The crystal structure of *Geobacillus kaustophilus* HTA426 DnaK in complex with GrpE has been determined (Wu *et al*., 2012). This showed an altered stoichiometry to that of the *E. coli* protein with two molecules of DnaK bound to two of GrpE as opposed to a 2:1 stoichiometry for the *E. coli* protein. This suggests a slightly altered mechanism of action and may have consequences for heterologous proteins expressed in this organism.

The formation and rearrangement of disulphide bonds, which play a critical role in protein folding and stability, are predominantly catalysed in the periplasm by a suite of Dsb proteins (DsbA, DsbB, DsbC and DsbD; Baneyx and Palumbo 2003). In the *E. coli* pET expression system, two expression hosts have been developed to promote disulphide bridge formation. SHuffle T7 expresses DsbC protein to improve the rate of cytoplasmic disulphide bridge formation, whereas Origami / BL21trxB has mutations in the *trxB* thioredoxin and glutathione reductase (*gor*) genes. These proteins reduce the disulphide bonds when formed, destroying them, and this mutation allows these bonds to more easily be formed and maintained in the cytoplasm.

Temperature

Bacteria can control their intracellular environment composition even against a range of steep chemical gradients. The intracellular pH of obligate acidophiles and alkaliphiles is circumneutral while the [Na^+] in the cytoplasm of halophiles is far below that of their surrounding medium (Oren 2003; Baker-Austin and Dopson 2007). However, one of the few parameters they don't have this level of control over, is the temperature at which they find themselves. Whereas the membrane allows physical separation of protons or other solutes, temperature penetrates the whole cell and all cell-bound proteins have therefore had to adapt to perform optimally in a particular temperature range. Although many studies have been conducted to determine the source of protein thermostability, no one factor has been found to allow proteins to maintain stability at elevated temperatures. Some general features have however been observed, such as increased hydrogen bonding and disulphide bridges, increased isoleucine, proline and valine content, particular amino acid couplings, smaller protein volume (compactness) and larger polar surface area (Panja *et al*., 2015; Modarres *et al*., 2016). In one exceptional circumstance the presence of a particular salt concentration was the sole reason for a thermophilic protein being stable at high temperature (Zhang *et al*., 2012). Irrespective, there are two features that a protein has to possess to operate at high temperature *in vivo*: i) it has to be able to fold into its final tertiary/quaternary structure at high temperature and ii) its final tertiary/quaternary structure has to then be stable at this temperature. An example of this is the phage P22 tailspike protein where the native trimer has a thermal denaturation midpoint temperature of 88°C, but nascent polypeptide chain mis-folds at temperatures above 40°C (Pope *et al*., 2004). Protein folding is a highly complex process, that includes the actions of chaperones, co-translational folding, and the evolved amino acid sequence of the protein exploring the folding energy landscape (Mallamace *et al*., 2016). Many small proteins appear to have an intrinsic folding pathway, and do not require assistance to fold (Pfanner 1999; Mayor *et al*., 2003; Turoverov *et al*., 2010; Tzul *et al*., 2017). Larger proteins (20-65kDa) however, usually have a more complex folding pathway and require the assistance of GroEL/ES chaperones to fold correctly (Houry *et al*., 1999; Turoverov *et al*., 2010). If an intermediate does not fold correctly (gets trapped in an unproductive energy low), the non-functional protein may be shunted to inclusion bodies inside the cell or be turned

over by proteases (DegP, S and Q in *E. coli*) to recycle the amino acids (Clausen *et al*., 2002; Turoverov *et al*., 2010; Pope *et al*., 2014).

Mesophilic proteins with desirable properties for a particular biotechnological application are likely best expressed in a mesophilic host. Thermophilic proteins display desirable properties such as ease of purification when expressed in a mesophilic host, reduced susceptibility to organic solvents and improved stability at elevated temperatures and extremes of pH (Sarmiento *et al*., 2015). However, numerous thermophilic proteins express poorly in *E. coli* (mesophilic host), likely due to the expression temperature being far below what they have evolved at *and* the major differences in mesophilic and thermophilic organisms' codon usage patterns (Schultes and Jaenicke 1991; Siddiqui *et al*., 1998; Diruggiero and Robb 1995; Singer and Hickey 2003; Hidalgo *et al*., 2004). Likewise, the expression of mesophilic proteins in thermophilic hosts very often fails because, apart from all the other factors that affect heterologous protein expression, they have not evolved folding intermediates or tertiary/quaternary structure which are stable at increased temperatures. For metabolic pathway engineering of thermophiles, the unavailability of a suitable thermophilic version of a particular protein (such as Pdc), may require the expression of a mesophilic variant. A strategy to overcome the inability of the mesophilic protein to work at the increased temperature is directed evolution of the enzyme to improve thermostability or improved folding at these temperatures (Frappier and Najmanovich 2014). This entails either the generation of a random mutant library which is then screened for activity or improved activity at higher temperatures. If the crystal structure is available for the protein, a rational design strategy could be employed to make selected amino acid changes to bring about thermostability (Modarres *et al*., 2016). Alternatively, domain swapping or gene shuffling, using a mix of pyruvate ferrodoxin oxidoreductase/pyruvate decarboxylase-like enzymes identified in several thermophilic archaea together with the mesophilic bacterial variants may be used to engineer a suitable enzyme (Eram et al., 2014).

1.2.4 Metabolic engineering of microorganisms for bioethanol production

The ability to genetically manipulate organisms has started to unlock their full potential. Engineering of microorganisms can be broadly divided into two categories. Directed evolution, where random changes brought about by physical (transposon-mediated) or chemical mutagenesis to the hosts genetic material, may result in improved phenotypic traits which are then selected for; or rational engineering which requires knowledge of the genes and regulatory elements present in the host organism.

Metabolic or pathway engineering in microorganisms is an often-used technique to redirect the flow of carbon inside the cell towards a desired product (small molecule / metabolite or protein) and most often takes a rational engineering approach (Kumar and Prasad 2011). The ability to do so requires several elements to be in-place: 1) a method for introducing DNA into the microorganism (electroporation, chemical competence, natural competence, conjugation, biolistic injection, transfection), 2) a suitable vector which can carry the DNA payload (cloning and/or expression vectors, shuttle vectors, suicide/integration vector), 3) a method to select for the incoming DNA (antibiotic or auxotrophic markers), and 4) knowledge of regulatory networks or elements (promoter sequences, repressors proteins and genes involved) inside the host cell. The general idea is to i) introduce a new gene(s) to generate new phenotypic features, ii) to delete genes/pathways which can siphon carbon from the desired pathway, iii) modify the regulation of genes and gene clusters to change the flux through existing pathways. Here I will have a look at some examples of engineering of microorganisms to produce biofuels to gain a better understanding of the targets selected for manipulation, the reason for their selection and the mechanisms by which changes were introduced. **Table 1.2.4.1** shows the starting characteristics of some model organisms which have been engineered for enhanced ethanol production. It therefore also shows which features, that might be desirable to have in an ethanol producing strain, are absent in these organisms.

Table 1.2.4.1. Characteristics of the most relevant microorganisms considered for ethanol production. (Adapted from Gírio *et al*., 2010)

Characteristic	Microorganism					
	E. coli	***Z. mobilis***	***S. cerevisiae***	***P. stipitis***	***K. marxianus***	***P. thermoglucosidans***
D-glucose fermentation	+	+	+	+	+	+
Other hexose utilization (D-galactose and D-mannose)	+	-	+	+	+	+
Pentose utilization (D-xylose and L-arabinose)	+	-	-	+	+	+
Direct hemicellulose utilization	-	-	-	W	+	+
Anaerobic fermentation	+	+	+	-	+	-
Mixed-product formation	+	W	W	W	W	+
High ethanol productivity	-	+	+	W	W	-
Ethanol tolerance	W	W	+	W	-	-
Tolerance to lignocellulose derived inhibitors	W	W	+	W	-	W
Osmotolerance	-	-	+	W	V	-
Acidic pH range	-	-	+	W	+	-
GRAS microorganism	-	+	+	+	+	+

+ Positive; - Negative; W weak; V variable.

A general feature of microorganisms investigated for bioethanol production is that they are either good natural producers of ethanol, and have to be engineered for lignocellulose breakdown and *vice versa*. Recently though, the discovery of a yeast (*Scheffersomyces shehatae*) capable of direct fermentation of starch to ethanol opens the possibility that an organism capable of breaching this divide has been identified and puts emphasis on continued biodiscovery efforts (Tanimura *et al*., 2015).

The yeast *K. marxianus* is the sexual stage of *Candida kefyr* and is most often isolated from dairy products or human / animal lesions and airway tissues. This thermotolerant yeast is capable of fermentation in the temperature range of 38℃-45℃ and can survive up to a temperature of 52℃, making it a good candidate for a high temperature bioethanol process. It has been shown capable of fermenting several economically relevant industrial and experimental substrates such as whey, starch, Jerusalem artichoke tubers, orange peel waste, paper sludge, agave and carrot pomace (Wilkins *et al*., 2007; Yuan *et al*., 2008; Yu *et al*., 2013; Villegas-Silva *et al*., 2014; Wang *et al*., 2014; Gabardo *et al*., 2015). Apart from its ability to naturally ferment substrates to produce high ethanol yields, it also produces a range of biotechnologically interesting enzymes (Fonseca *et al*., 2008), including some that assist in the breakdown of varied substrates for ethanol production (inulin in the Jerusalem artichoke). One of the main problems with using *K. marxianus* for ethanol production is that, compared with *S. cerevisiae*, it has a lower ethanol tolerance (Costa *et al*., 2014). Whether the organism can compete with *S. cerevisiae*

through engineering or selection to be more ethanol tolerant, or the tolerance offset through early product removal aided by the increased fermentation temperature, is yet to be determined (Hack and Marchant 1998; Pang *et al*., 2010). A genetic system has been established for *K. marxianus* (Fonseca *et al*., 2008), however the transformation efficiencies for most techniques are relatively low (hundreds to thousands of transformants per µg of DNA). New techniques have been developed to improve on both the transformation efficiency and allow for targeted gene disruption (Abdel-Banat *et al*., 2010). As mentioned, the organism is naturally capable of fermenting a range of feedstocks to ethanol, but it has also been engineered or selected for in several ways to improve this ability.

To ferment xylose, the sugar must first be converted to xylulose. In bacteria, capable of xylose fermentation, xylose is usually converted to xylulose by xylose isomerase, in a redox-neutral conversion (Jeffries 1983). The product is then phosphorylated by xylulokinase and the phosphorylated product metabolized by the pentose phosphate pathway. A second pathway is through conversion of xylose to xylitol, by xylose reductase and subsequent conversion to xylulose by xylitol dehydrogenase. Both these steps require cofactors which, when these enzymes are over-expressed or expressed in a non-native host, can lead to redox imbalance and poor growth and fermentation performance, especially under anaerobic conditions (**Figure 1.2.4.1**).

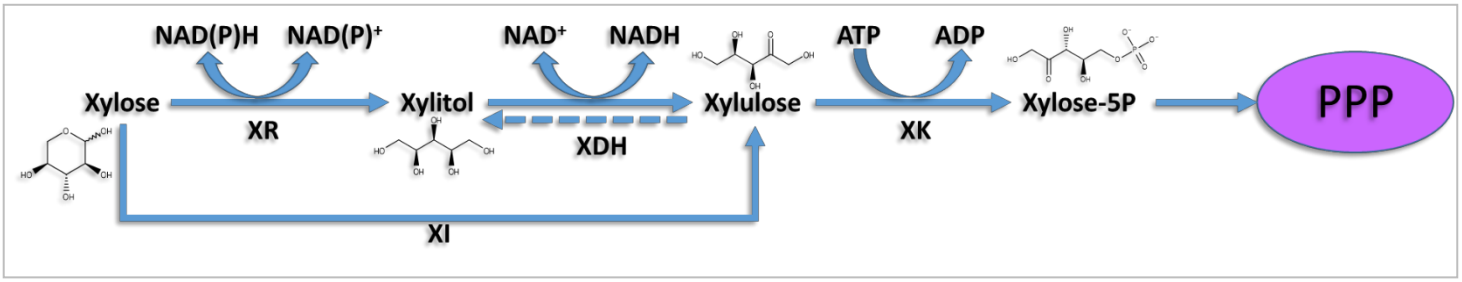


Figure 1.2.4.1. D-xylose conversion in the PPP. XI – xylose isomerase, XR – xylose reductase, XK – xylulokinase, XDH – xylitol dehydrogenase, PPP – pentose phosphate pathway

To improve *K. marxianus*' ability to ferment this pentose sugar to ethanol, Wang and co-workers engineered it by interrupting the native XR and XDH and introducing the XI from an *Orpinomyces* sp. (Wang *et al*., 2013). The strain was capable of utilizing xylose as a sole carbon source and xylose fermentation was demonstrated by fermentation of corn cob hydrolysate, rich in xylose, producing the same result as fermentation of pure xylose (6.83 g/l ethanol produced from 17.474 g/l xylose consumed

in 7 days at 2% wt/vol xylose) with the production of 8.25 g/l ethanol from 20.04 g/l xylose (23.61g/l present in the hydrolysate).

Pang and co-workers used a random mutagenesis strategy to improve both ethanol production and tolerance in *K. marxianus* (Pang *et al*., 2010). They screened 25200 mutated colonies and demonstrated both improved ethanol production and tolerance through a mixture of UV radiation and exposure to the mutagen nitrosoguanidine with selection *via* tetrazolium test (Pang *et al*., 2010). For their best performing mutant, they saw a 25% increase in maximum ethanol concentration produced and the strain could tolerate ethanol concentrations up to 11% (v/v) as opposed to the wild type strain at 8% (v/v). This demonstrates that although unpredictable, with the correct selection, random mutagenesis strategies can be successfully employed to deliver improved strain characteristics.

S. cerevisiae is the undisputed microorganism-of-choice when it comes to high concentration ethanol fermentations, and is the main workhorse for first- and second-generation bioethanol production (Thomas and Ingledew 1992). One of the reasons it is capable of such high ethanol production is its tolerance to ethanol, allowing it to continue metabolizing glucose to ethanol despite rising ethanol concentrations in its surroundings. The main drawbacks of this organism are its inability to ferment some C5 sugars and the fact that it does not natively produce hydrolytic enzymes capable of degrading lignocellulose or lignocellulose breakdown products (den Haan *et al*., 2013). It is also susceptible to lignocellulose hydrolysis products (section 1.2.1.2). Thus, the main focus of engineering of this organism has been to improve its substrate degradation capabilities and improve on the range of sugars it can metabolize. To improve its ability to degrade amorphous and crystalline cellulose, several yeast and fungal enzymes (*Saccharomycopsis fibuligera* β-glucosidase; *Trichoderma reesei* endoglucanase 2 and cellobiohydrolase 2; *Aspergillus aculeatus* β-glucosidase; *T. reesei* endoglucanase 1 (*cel7B*); *S. fibuligera* β-glucosidase (*cel3A*)) have been expressed in *S. cerevisiae* (Van Rooyen *et al*., 2005; den Haan *et al*., 2013). The ability of various yeast and fungal sugar transport systems to enable uptake of these substrates in *S. cerevisiae* have also been investigated. The cellodextrin transport system of *Neurospora crassa* was introduced into *S. cerevisiae*. Expression of this system, together with a β-glucosidase allowed the strain to grow on cellotetraose, while the *Kluyveromyces lactis* lactose uptake

system enabled uptake of cellobiose, when co-expressed with a *Clostridium stercorarium* cellobiose phosphorylase (*cepA*). To take advantage of the hemicellulose fraction (mainly xylose and arabinose), endoxylanase, β-xylosidase, α-arabinofuranosidase, α-glucuronidase, acetylxylan esterase and ferulic acid esterase activities are required for complete hydrolysis of the backbone sugars (section 1.2.1.2). β-xylosidase genes from *Aspergillus niger, Aspergillus oryzae* and xylanase II from *T. reesei* have been expressed enabling the breakdown of birchwood xylan to short-chain xylo-oligomers.

As the organism cannot metabolize D-xylose, XR, XDH or XI encoding genes have had to be expressed in *S. cerevisiae* to enable xylose utilization. XR and XDH both use NAD(P)H as cofactor, thus XI is the preferred enzyme activity as this does not interfere with the intracellular redox balance, particularly under anaerobic conditions (Zhou *et al*., 2012). Expression of these enzyme activities has resulted in strains capable of xylose utilization, however the expression of XI from a *Piromyces* species, together with engineering a pentose phosphate pathway into *S. cerevisiae* through expression of *Pichia stipites* XK, gave the best result thus far, with the strain able to produce 41 grams of ethanol per gram of xylose consumed (Zhou *et al*., 2012). Together, these transport and degradation enzymes have enabled *S. cerevisiae* to degrade these substrates and convert them to ethanol with varying degrees of success, demonstrating that, in principle, the organism could be engineered towards a CBP platform organism.

One big difference between 1G and 2G processes is the release of pentose sugars in 2G processes, which represents a valuable waste stream. As seen above, many yeasts are not naturally capable of (taking up) fermenting certain pentose sugars and have to be engineered to do so. One benefit of utilizing bacteria for 2G fermentations, is that some such as *E. coli*, can naturally take advantage of these sugars.

E. coli's natural fermentation ability results in mixed acid fermentation products (lactate, formate, acetate, ethanol and succinate), and compared with any of the engineered microorganism mentioned above, has unmatched ability to ferment pentose sugars (Ingram *et al*., 1999). The amount of ethanol produced relative to the organic acids made it unfeasible to use for commercial ethanol production, thus it had to be engineered to redirect the carbon from organic acid production to ethanol. Early efforts to engineer a homoethanolic pathway in *E. coli* made use of the Pdc and Adh discovered in *Z. mobilis* to do the two-step conversion of pyruvate to acetaldehyde and ethanol (Ingram *et al*., 1999). It was found

that the high-level expression of Pdc from plasmid vectors effectively redirected the flow of carbon to acetyl-CoA and ethanol, without the need to inactivate the native Ldh. Once the gene cassette was integrated on the chromosome (into *pfl* gene), to avoid loss of the genes, it was found that expression levels were too low and the strain had to be evolved for improved gene expression. The inactivation of fumarate reductase (*frd*) reduced the amount of succinic acid produced and this strain, KO11, has been the basis for much research on *E. coli*-based bioethanol production (Ohta *et al*., 1991a; Gonzalez *et al*., 2003; Huerta-Beristain *et al*., 2016). This approach (Pdc and Adh expression) was duplicated in *Klebsiella oxytoca* with similar results (Ohta *et al*., 1991b). More recently, the difficulty in using genetically modified organisms for commercial processes has forced researchers to adopt a different approach to engineering *E. coli* for ethanol production. Kim and co-workers engineered the organism by inactivation of the *ldh* and *pfl* genes (Kim *et al*., 2007). The PflB is part of the native ethanol pathway in *E. coli*, so inactivation of this gene, should not have allowed the bacterium to produce ethanol. It was hypothesized that its deletion, which prevents formation of formate and acetyl-CoA from pyruvate, has allowed the activity of a Pdc-like enzyme to be seen, although no Pdc activity could be detected in cell extracts. Another alternative is that the Pdh complex is responsible, converting pyruvate to acetyl-CoA which can then, through the action of acetaldehyde/aldehyde dehydrogenase, be converted to acetaldehyde and ethanol. The second pathway was shown to be the one responsible for ethanol production in this deletion strain by knocking out of *aceF*, an integral part of the Pdh complex. This was rather unusual as the Pdh complex genes are normally only expressed under aerobic conditions. Mutations that allow increased Pdh expression under microaerobic conditions are thought to have enabled this pathway in *E. coli*.

Even though many lignocellulose degrading enzymes that are functionally expressed in *E. coli*, have been identified, the main drawback and possibly its Achilles heel as a CBP organism, is its ability to efficiently secrete proteins. Despite this, Bokinsky and co-workers engineered *E. coli* to produce and secrete an endocellulase from *Bacillus* sp. D04 and a *Clostridium stercorarium* endoxylanase (*xyn10B*), both fused to OsmY, for secretion; as well as a β-glucosidase (*cel3A*) and a xylobiosidase (*gly43F*) from *Cellvibrio japonicus* (Bokinsky *et al*., 2011). The expression of these were coupled to promoter sequences that drive transcription of genes when *E. coli* is starved of carbon (biomass-consumption

pathways). This did away with the need to add a chemical inducer and provided substrate induced induction allowing rapid growth on both cellobiose and enzymatically hydrolyzed xylan at rates nearly equalling growth on their constituent monomer sugars. These strains were capable of growing on substrates (EZ-Switchgrass) pre-treated with ionic liquids to produce either pinene, butanol or fatty acid ethyl esters (FAEE; biodiesel), without the need to add exogenous cellulases.

Bacteria belonging to the genus *Parageobacillus* are Gram positive, thermophilic, spore-forming and capable of growth at 45-75℃ (De Maaayer *et al*., 2014). They have been isolated from a variety of environments including soil, oil fields, compost heaps, deep sea sediment and hot springs (Nazina *et al*., 2001; Schmidt *et al*., 2011; Ziegler 2013). Formerly belonging to group 5 of the genus *Bacillus*, many *Parageobacillus* species were lumped together and described as one species, *Bacillus stearothermophilus* (Studholme 2014). Since description of the genus *Geobacillus* in 2001, fifteen new species have been validly described (Nazina *et al*., 2001, De Maaayer *et al*., 2014, Aliyu *et al*., 2016). More recently however, a phylogenomic approach demonstrated that the genus possibly comprises two genera leading to the proposal of a new genus: *Parageobacillus* (Aliyu *et al*., 2016; Burgess *et al*., 2017). Comparison of several *Parageobacillus* genomes has shown that they harbour a shared genomic island encoding a range of thermostable hemicellulose degrading enzymes, transporters and genes for the utilization of the monomer sugars generated (Ziegler 2013; De Maayer *et al*., 2014). Most of these are dedicated to the breakdown of xylooligosacharides, methylglucuronate side chains and L-arabinan arabinosaccharides. Despite the presence of these genes, some hydrolytic activities are not natively encoded by *Geobacillus*/*Parageobacillus* species. The successful heterologous expression of β-galactosidase (*G. stearothermophilus*), α-amylase (*G. stearothermophilus*), cellulase (*Pyrococcus horikoshii*) and esterase (*Pyrobaculum calidifontis*) in *Geobacillus* HTA426 means that these activities could be used to augment the already impressive suite of enzymes found natively in several *Geobacillus*/*Parageobacillus* species (Suzuki *et al*., 2013).

These organisms naturally produce large amounts of lactic acid from pyruvate under microaerobic conditions, and cannot grow under fully anaerobic conditions. To redirect its metabolism from lactate towards ethanol production several metabolic changes were made and these closely mimic what was

done in *E. coli*. This was achieved by knocking out (interrupting) *ldh* and *pfl*, as well as up regulation of *pdh*. These changes meant that most, if not all, pyruvate was directed towards ethanol production (Cripps *et al*., 2009).

As can be seen, many studies have described the engineering of metabolic pathways in microorganisms for the purposes of producing ethanol and in many of these microorganisms the strategy for conversion of sugars to ethanol relies on the Pdc enzyme for conversion of pyruvate to acetaldehyde (Eram *et al*., 2013). This approach, expression of *pdc* and *adh*, in hosts with a broad catabolic phenotype, essentially creates an Entner–Doudoroff-type metabolism (Taylor *et al*., 2009). As seen in the *E. coli* example discussed above, depending on the metabolic flux through various pathways, the Pdc mediated path may prove advantageous over ones that have been engineered in *P. thermoglucosidans* thus far (Cripps *et al*., 2009). Several bacterial pyruvate decarboxylase enzymes have been expressed in *P. thermoglucosidans*, but none of these resulted in improved ethanol yields, even though evidence existed for expression of the genes in the organism, such as increased pyruvate consumption and positive Western blot results (Thompson *et al*., 2008). This suggested that even though the proteins were produced they were non-functional or minimally functional in *P. thermoglucosidans*. Thus, there is an opportunity to identify novel pyruvate decarboxylases for functional expression in this organism which may lead to improved fermentation performance such as higher ethanol yields (g/g) or reduced residence times.

1.3 Introduction to bacteriophages

A virus is a small infectious agent that replicates only inside the living cells of other organisms and infects all types of life forms including animals, plants and microorganisms. Bacteriophages are viruses that specifically infect bacteria. They are composed of nucleic acid encapsidated in protein (with or without a lipid layer envelope; **Figure 1.3.1**), where the nucleic acid can be dsDNA, ssDNA, dsRNA or ssRNA (Hyman and Abedon 2012). Recent estimates suggest that these are the most abundant biological entities on the planet, with an estimated 10 virus particles to every bacterial cell and a total estimate of 10^{31} virus particles (Breitbart and Rohwer 2005). Although probably recorded as early as 1896 to 1898 by Ernest Hankin and Gamaleya respectively, these entities are considered to have been co-discovered by Frederick Twort and Felix d'Herelle in 1915 and 1917 (Samsygina & Boni 1984). Since then, the study of phages has contributed enormously to our understanding of the basic principles of biology and led to the birth of modern molecular biology (Salmond and Fineran 2015). Some of the highlights include: 1) The discovery that DNA is the molecule coding for genes, not proteins, 2) Discovery of the triplet nature of the genetic code 3) Gene regulation 4) Discovery of restriction endonucleases, and 5) The recent discovery of the bacterial CRISPR adaptive immunity system and its adoption as a Eukaryotic genetic modification tool.

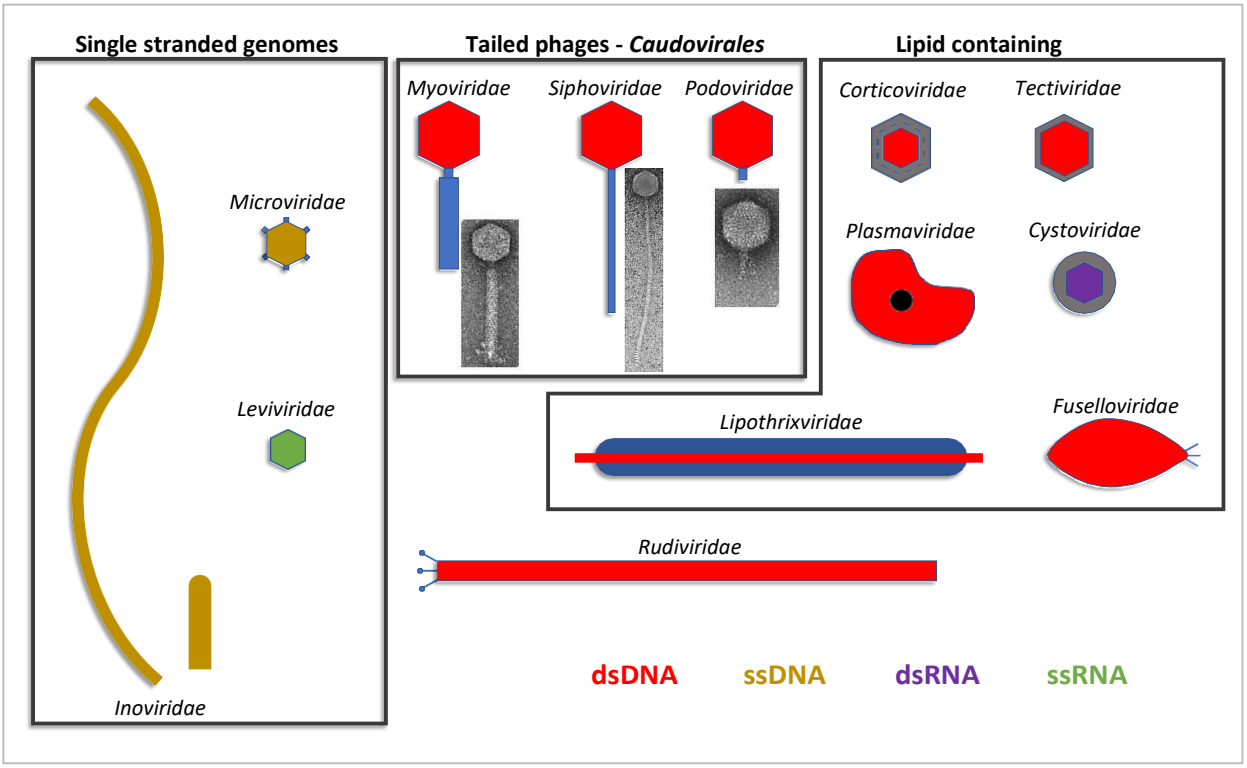


Figure 1.3.1. Morphologies and nucleic acid content of bacteriophages. Adapted from Ackermann 2003.

Phages appear to be ubiquitous in nature, having been found in all temperate and many extreme environments, including soda lakes, terrestrial hot springs, deep sea hydrothermal vents, hot/cold deserts and hypersaline springs (Le Romancer 2007; Moser *et al*., 2012; Adriaenssens *et al*., 2014). Here, they have the demonstrated ability to influence biogeochemical cycling through their lytic and lysogenic effects on their host and in some cases can lead to the collapse of trophic structures as they interact with their host through predator-prey relationships (Fuhrman 1999; Clokie *et al*., 2011; Donavaro *et al*., 2011; Peduzzi *et al*., 2014). These small biological entities therefore have an enormous, and direct ecological impact. They are also intimately associated with human beings and their surrounds (Hannigan *et al*., 2015, Manrique *et al*., 2016).

With the advent of next generation sequence technology and the development of a plethora of tools for virome analysis (MetaVir, MG-RAST, Virsorter, VIP, ViromeScan, PHACCS, VIROME, CAMERA, vConTACT, WIsH; Roux et al., 2014; Roux et al., 2015; Meyer et al., 2008; Yang et al., 2016; Rampelli et al., 2016; Angly et al., 2005; Wommack et al., 2012; Seshadri et al., 2007; Bolduc et al., 2017; Galiez et al., 2017), which allows for the study of whole virus populations, it has become apparent that there is a very wide diversity of bacteriophages (Breitbart *et al*., 2002). When sampling most environments, it has been found that phages dominate the viral population, and of these, the tailed phages (Order *Caudovirales*) are most dominant (Monier *et al*., 2008; Broecker *et al*., 2016; Adriaenssens *et al*., 2017). Myoviruses and podoviruses can be the dominant tailed phages (rather than siphoviruses), especially in marine environments (Ackermann 2007; Williamson *et al*., 2008; Huang *et al*., 2010; Brum *et al*., 2016). Because of the genomic diversity observed in these studies, they are also thought of as a vast untapped reservoir of unique genomic sequence which may harbour biotechnologically useful enzymes.

Phages are also being considered a serious alternative/augmentation to antibiotic treatments. In the mid-20th century, the use of phages to treat a variety of infections was actively investigated in the former Soviet Union due to lack of access to Western antibiotics, however had fallen out of favour due to the efficacy of the new antibiotics against a wide range of pathogens (Salmond and Fineran 2015). Recently, Western researchers have revived the idea of phages as treatment for bacterial infections due to the increase in antibiotic resistance among bacteria (Elbreki *et al*., 2014), and there are several documented

successes in treating infections, including a severe case of sepsis (https://tinyurl.com/kaxxhuz). With the phase I/II PhagoBurn clinical trial (http://www.phagoburn.eu/) and commercialization of products like Staphefekt® (https://www.staphefekt.com/en/; an endolysin cocktail specifically targeting *Staphylococcus aureus*), the PhageGuard range of products (https://www.phageguard.com/) and numerous phage-based companies (http://www.companies.phage.org/) established, the age of using phages to combat bacterial infections appears to have arrived.

In the pre-genomic era, phage classification was based mainly on their morphology. Later, a combination of the type of nucleic acid (DNA or RNA), capsid structure and whether the virus was enveloped was used (Salmond and Fineran 2015). The International Committee on the Taxonomy of Viruses (ICTV) was established in 1966 and is the body currently responsible for establishing the rules governing viral classification (Adams *et al*., 2017). With the exponential increase in phage genomic sequences being deposited in the GenBank database (Merrill *et al*., 2016), it is evident that there is enormous diversity viral and bacteriophage genomes (Mizuno *et al*., 2013; Zablocki *et al*., 2014). This necessitates a different approach to phage classification (Simmonds *et al*., 2017; **Figure 1.3.2**).

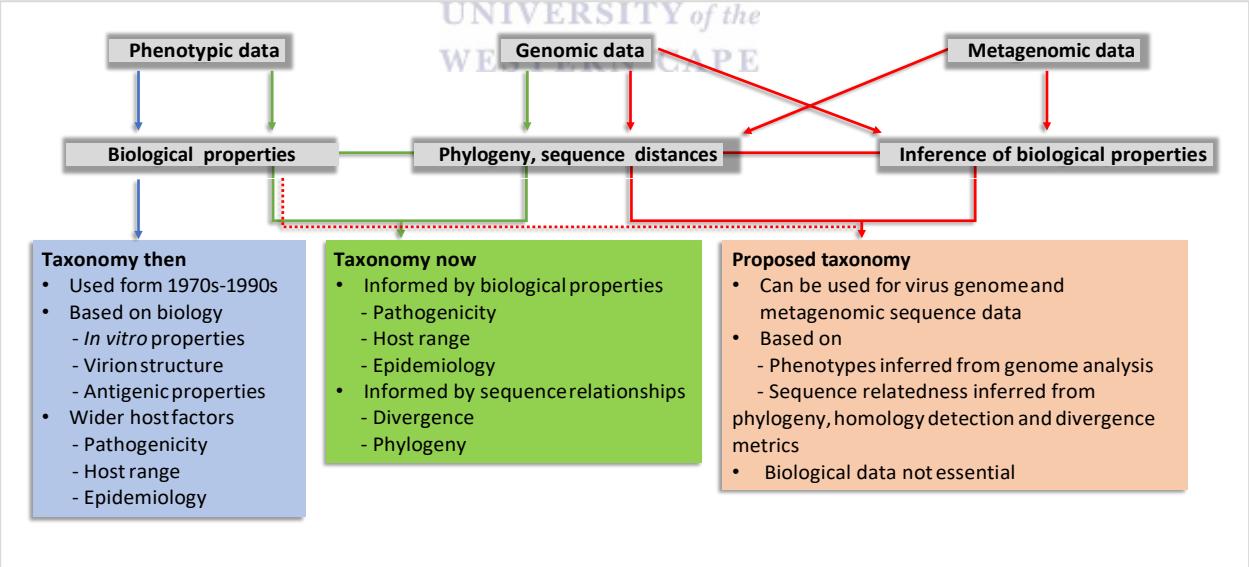


Figure 1.3.2. Summary of the proposed phage classification pipeline. The proposed classification pipeline (red arrows) enables both metagenomic sequence data and conventionally derived virus sequences to be classified. Inferred biological properties that are obtained by bioinformatic analysis of virus sequences together with information on sequence relatedness and gene content, and, optionally, any observed biological properties (dotted line), may all be used as defining criteria for species and higher rank taxonomic assignment in the International Committee on Taxonomy of Viruses (ICTV) taxonomy. This procedure differs from current (green arrows) and previous practice (blue arrows), in which biological data and/or host information and

sequence data (current), or biological data alone (1970s–1990s), were required for classification. (Adapted from Simmonds *et al*., 2017)

Phages are currently classified into Order, Family, Sub-family, Genus and Species, with a species being the lowest taxon (group) in a branching hierarchy of viral taxa. The main species demarcation criterion for bacterial and archaeal viruses is currently set at a genome sequence identity of 95%. The modern approach is to use a consensus group of properties to classify phages. These properties include overall DNA and protein sequence conservation between two phages, phylogenetic analysis of conserved proteins (terminase large subunit), the host range, and the phage morphology. The Bacterial and Archaeal Viruses Subcommittee (BAVS) of the ICTV are currently working on a new system that would abolish the old family classification based solely on morphology (*Siphoviridae*, *Myoviridae*, *Podoviridae*) in favor of genome/proteome-based family description. Currently, all viruses of bacteria and archaea are classified into 22 families, 14 subfamilies and 204 genera with 873 described species (Adriaenssens and Brister 2017).

1.3.1 Bacteriophage lifestyles

Broadly speaking, phages can have two "lifecycles": lytic or lysogenic. During the lytic cycle, the phage infects its host, produces more phage progeny and lyses the host to release the progeny. During the lysogenic cycle, the phage infects its host, integrates its genome into that of the host and persists in the host as part of its genomic content until conditions become unfavorable. At this point the phage can be induced to the lytic cycle. In reality, different phages can exist in a number of variations, or grayscales, of both these "lifecycles". These include states where viral particles are continuously produced without killing the host (chronic infection; e.g. M13; Siringan *et al*., 2014) and pseudolysogeny, where the virus (e.g. T4) neither replicates nor is integrated into the host genome and is in suspended animation until conditions for the host improve (Łoś and Węgrzyn 2012). Some phages are strictly lytic (cannot lysogenize the host) while so-called temperate phages can switch between the lytic and lysogenic states.

The reasons why phages have evolved these two distinct "lifecycles" is thought to revolve around their own survivability *as well as* the survival of their hosts (Paul 2008; Knowles *et al*., 2016). Several models

have been proposed to describe phage-host dynamics at the level of lysogeny and lysis in natural environments. The "kill-the-winner" hypothesis, where virulent phages reduce the most dominant host populations to a lower steady state level (followed by a switch to lysogeny), independent of the host's growth rate was first proposed by Thingstad in 1997 (Maslov and Sneppen 2017). This model took into account exclusion of viral predators by heterotrophic protists, the rise or transfer of resistance to viral infection, greater species-level host diversity and increasing viral decay to explain observed data. The model has since been amended to the "piggyback-the-winner" hypothesis, based on the observation, in several disparate environments, that viral numbers relative to host numbers decrease beyond a certain host density (Knowles *et al.*, 2016). This means that once a certain threshold host-density is reached, the phages "prefer" to enter the lysogenic cycle and the numbers are further reduced by a concomitant increase in superinfection immunity in the hosts. These two factors alone (the switch to lysogeny and superinfection immunity) appear to explain the observed data and are in agreement with single isolate studies (see below). The mechanism (signaling molecules, epigenetic effects) by which phages may exercise this "preference" has not yet been elucidated.

In general, when the host is plentiful, relative to the phage particle numbers (low MOI; multiplicity of infection - numbers of phage that infect one cell), and rapidly growing, the phage will infect at low MOI and likely enter a lytic phase. This should produce many phage progeny which will go on to infect more of the host. If the host is starved and/or the number of phage relative to the host is high (high MOI), the phage will enter a lysogenic phase to prevent killing all hosts. Thus, if conditions are favorable for the host, this should allow for rapid proliferation of the phage, and as conditions deteriorate for the host the phage becomes dormant by integrating into the host genome and replicating with it.

1.3.1.1 Lytic cycle – *Escherichia* virus T4

The *E. coli*-infecting myovirus T4 is the best studied example of a phage with a strictly lytic "lifecycle". The lytic development of a phage inside the host is usually tracked using soft agar overlays to observe plaques (clear zones) in a bacterial lawn and is used to determine the phage titre. In a natural setting, phages continually infect hosts at different time intervals meaning that one cell may be releasing progeny, while another has just been infected. This asynchrony does not allow one to determine the parameters associated with lytic growth and development. In studies to elucidate the nature of phage infection and lytic development, the infection of the culture by virus is synchronized such that the whole culture behaves as one organism. If the number of plaques is then plotted against time, a curve is observed as shown in **Figure 1.3.1.1.1**. The curve, known as a one-step growth curve, is meant to capture the information around the first lysis event and shows the various stages during phage development.

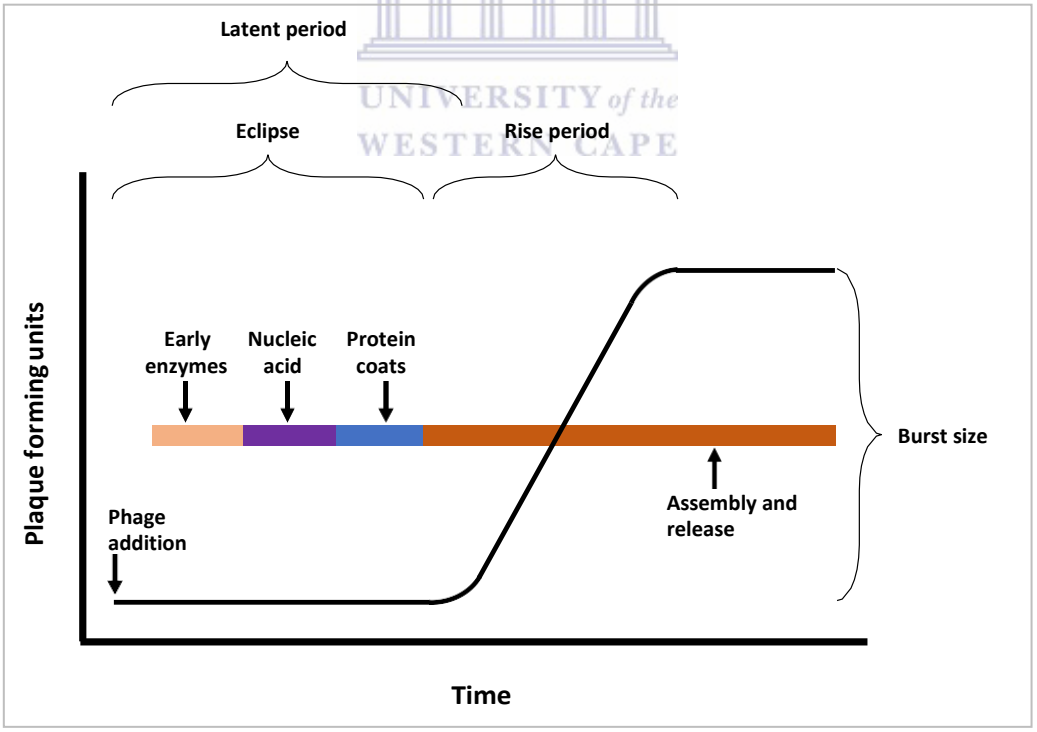


Figure 1.3.1.1.1. Phage one-step growth curve. During the eclipse period the host does not contain any complete, infective virions. During the remainder of the latent period, infective virions are assembled, but not released. Virions are released through host lysis during the rise period. The burst size indicates the number of phage progeny released per cell.

The obvious first step in phage infection is attachment to its host and injection of its genetic material into the host. Phages have evolved to exploit most cell surface features for attachment to the host including proteins, polysaccharide and lipopolysaccharides (LPS) (Samson *et al*., 2013). The current view is that most phages are highly specific for a particular host, due to the recognition of particular features on the bacterial cell surface (Rakhumba *et al*., 2010). For phage T4, the surface feature recognized by its tail fiber protein (Gp37) is the *E. coli* LPS. For the other model system described below, phage lambda, the attachment site is the LamB protein responsible for maltose uptake (Werts *et al*., 1994). Phage binding happens in two phases: they first bind reversibly to their target, followed by solidification of binding which is irreversible. In T4, this first stage is mediated by the Gp37 protein (distal protein on the main tail fibers; **Figure 1.3.1.1.2**), and allows T4 to "walk" the cell surface seeking out its cognate receptor (Rothenberg *et al*., 2011; Storms and Sauvageau 2015). The interaction between Gp37 and LPS is strong enough to allow the phage to stay in close association with the cell surface, but still weak enough to allow the phage to detach. Irreversible binding is driven by a conformational change in the baseplate that releases the short tail fibers which bind strongly to the keto-deoxyoctulosonate core subunit of LPS (Hu *et al*., 2015). This change in baseplate structure, is also the trigger for contraction of the tail sheath (outer protein layer of the tail). This drives the tail tube, the inner tail structure and conduit for DNA translocation from the head to the bacterial cytoplasm, through the outer cell membrane. The tip of the tube is composed of four proteins, three of which form a trimer $(gp27\text{-}gp5^*\text{-}gp5C)_3$-gp5.4, with gp5.4 at the very tail tip. gp5* has lysozyme activity, and it is thought that gp5C and gp5.4 dissociate during injection into the cell to reveal gp5* and allow it to degrade the cell wall peptidoglycan. The tube does not penetrate the full cell wall, but rather only extends into the periplasmic space. In the late stages of DNA injection, the inner membrane distorts and is seemingly fused to the tail tube tip. This is thought to allow the tube access to the cytoplasm for introduction of the phage genome. The mechanism for this fusion is not yet understood, and although an energized membrane is necessary for DNA translocation, membrane potential is not thought to play a role in puncturing the inner membrane as the fusion is still observed in the presence of uncouplers. While these steps describe the stages of membrane penetration for a myovirus, this process is not directly applicable

to other tailed and enveloped phages. For a review of strategies used by other phages see reviews by Molineux and Panja 2013 as well as Xu and Xiang 2017.

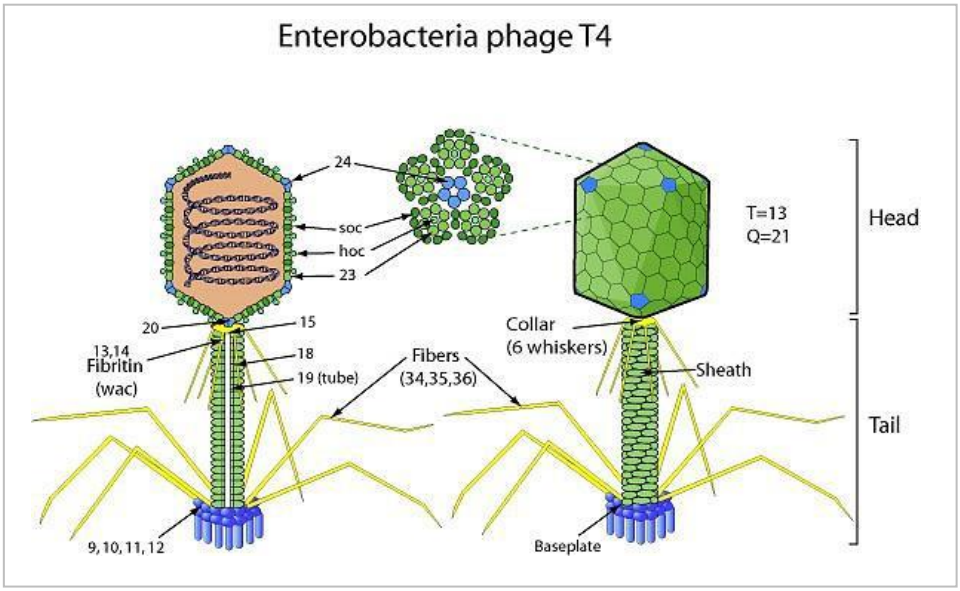


Figure 1.3.1.1.2. Physical structure of phage T4. http://viralzone.expasy.org/504?outline=all_by_species

There are currently large discrepancies between *in vitro* and *in vivo* measurements of DNA ejection for a range of phages studied (Molineux and Panja 2013). For some phages DNA ejection is a two-step process (T5), suggesting involvement of cellular machinery for internalization of the DNA, whereas for others it occurs as a single step. From *in vitro* measurements, the DNA translocation process is thought to be driven by one of two passive processes: a) The pressure from DNA packaging in the head (continuum mechanism model) or b) The inflow of water molecules through the capsid (hydrodynamic model) (**Figure 1.3.1.1.3**; Molineux and Panja 2013). The packaging of DNA into the head, towards the end of phage particle maturation, is energy dependent, and results in increased pressure within the head (>60 atm). The pressure comes both from the exclusion of water and counter ions (the capsid is permeable to water and other small solutes) and the increase in electrostatic interactions within the DNA molecule itself, and between the DNA and capsid proteins (Sun *et al*., 2008). One reason that the DNA is thought to be so tightly packed is that exclusion of water, reduces the ability of DNA to take part in biochemical reactions thereby reducing the chance of degradation and improving phage stability. Currently, there are two scenarios which are thought to dictate DNA ejection. If the channel through

which DNA is ejected is larger than the B-form of DNA, then the hydrodynamic model seems to explain most observations, as the water is allowed to move through the phage capsid and along with/next to the in-rushing DNA molecule into the cytoplasm. When the channel is just wide enough for the DNA to pass through (but not water or other ions), the initial driving force for ejection may come from the stored pressure in the head, followed by active internalization of the DNA by cellular machinery, such as the RNA polymerase transcribing early genes, by pushing against the inner membrane. As the head empties, water molecules may enter providing an additional driving force to help eject the last portion of the genome.

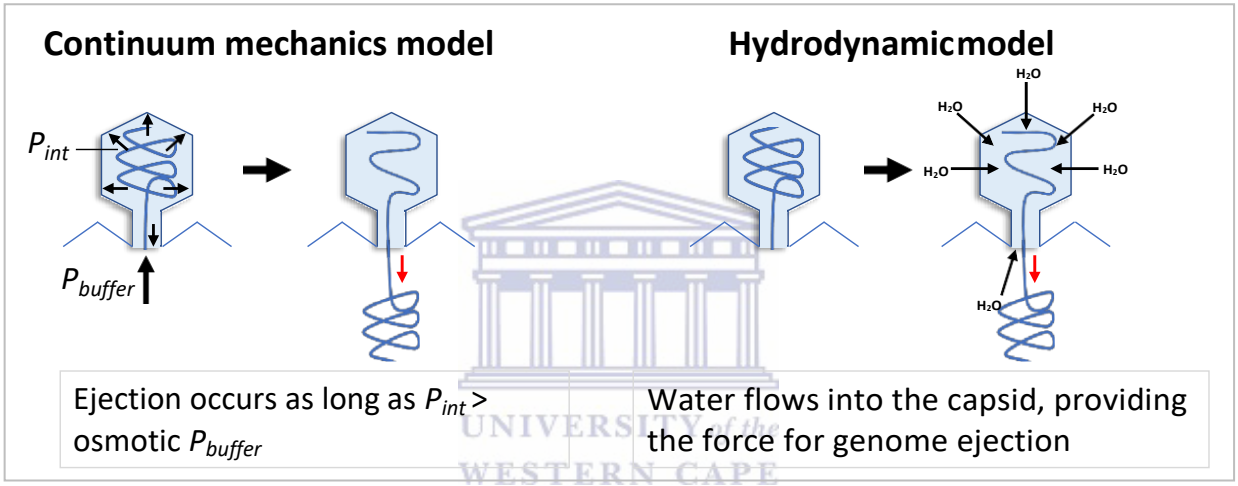


Figure 1.3.1.1.3. Models of phage DNA ejection. Red arrows indicate the movement of DNA out of the tail tube, and grey arrows indicate the direction of water flow for the hydrodynamic model. Ejection in the continuum mechanics model is facilitated by the internal ejection pressure (P_{int}), and ejection continues as long as P_{int} is higher than the osmotic pressure of the buffer (P_{buffer}). In the hydrodynamic model, water flows into the virion up the osmotic gradient along any path that it can find, and pushes the DNA out of the capsid. Adapted from Molineux and Panja 2013.

The main feature of T4 lytic growth is the complete hijacking of host cell metabolism to produce phage progeny. This is achieved by repressing host gene synthesis and degrading host DNA to generate nucleotides for replication of its own genome. The T4 lytic cycle can be divided into three stages (early, middle and late) with overlapped timing of gene expression to coordinate cessation of host replication and metabolism, phage replication, structural proteins synthesis, assembly and release (Luke *et al*., 2002; **Figure 1.3.1.1.4**). The entire cycle is completed in 30 minutes, with late gene transcription initiated as early as 7 minutes post-infection at 37°C. These three developmental stages can be described according to the promoter class (Pe, Pm, Pl) that drives gene transcription at the various stages.

Directly following infection, a set of 39 phage promoters are activated along with several host-like promoters. The Pe promoters are stronger than most host promoters. One of the proteins (gpAlt) that enters the cell along with the phage genomic DNA modifies the host RNA polymerase by adding ADP-ribosyl units to various subunits of the polymerase. These changes enhance the interaction between T4 early promoters (only one α subunit ADP-ribosylated) and the RNA polymerase to such an extent that phage early promoters are preferentially transcribed. Two more ADP-ribosyltransferases are produced (ModA and B). The modifications they make to RNA polymerase, including ADP-ribosylating the second α subunit, leads to inhibition of transcription from promoters with UP elements upstream of the promoter. The UP element consists of poly A tracts and induces DNA bending which enhances transcription from those promoters. Many of the host promoters as well as T4 early promoters contain these UP elements. This effectively switches off host gene transcription as well as early gene transcription. Transcriptional regulation is the predominant factor in T4 gene expression, however mRNA degradation also plays a role in determining the abundance of transcript and plays a pivotal role in gene expression regulation. RegB, a T4 encoded endoribonuclease produced from a Pe, specifically targets early gene transcripts for degradation helping the phage efficiently switch from early to late transcription. Another early gene (*61.5* or *dmd*) was shown to act as an antitoxin protein which neutralizes several toxin proteins in *E. coli*, the role of which is mRNA degradation (Otsuka and Yonesaki 2012). This action selectively stabilizes late transcripts and destabilizes middle transcripts.

Next, the 30 Pm's are activated. These are characterized by dependence on the transcriptional activator MotA (presence of a Mot box in the promoter sequence), and are weaker than the Pe promoters. AsiA, another phage encoded protein, is also required for Pm transcription. It binds to σ^{70} and stimulates middle-mode transcription. This association also further represses transcription from host promoters and Pe. The early and middle genes are mostly concerned with arrest of host metabolism, host DNA degradation (phage encoded endonuclease) and phage DNA replication (DNA polymerase).

T4 replication is recombination dependent, as opposed to rolling circle or theta type, and five possible T4 replication pathways have been identified (Mosig 1998). Any one or a combination of these pathways could be in operation during phage replication, however some are dependent on gene

expression at the various stages. The two main pathways are “join-copy” and “join-cut-copy”, and these pathways lead to the formation of a highly branched “concatenated” DNA molecule ready for packaging. The “join-copy” pathway only requires early and middle gene products, whereas the “join-cut-copy” pathway requires two late proteins (endonuclease VII and terminase), making this a pathway which only takes place later during the lytic cycle. As soon as a replication origin-initiated replication fork reaches one end of the phage genome, the join-copy pathway is initiated. Because of headful packaging, the terminal ends of the phage genome are complementary. This allows the single stranded DNA (ssDNA) of the unreplicated phage genome end to perform one of two functions. It can either invade the complementary end, at the other end of the phage genome, or invade another phage genome. In the first instance, this would serve as primer for replication of the middle region of the genome, while the antisense strand is made double stranded through discontinuous replication using Okazaki fragments (**Figure 1.3.1.1.5A**). In the second instance, this will generate a replication fork in the second genome which will lead to replication of a portion of this genome and generate more ssDNA which can invade another genome and so on. These processes rely on several phage-encoded enzymes such as UvsX, UvsY, gp32 and T4 topoisomerase. UvsY is a hexameric recombination mediator that facilitates loading of UvsX onto ssDNA where gp32 has bound. It has been shown that the T4 encoded DNA primase, responsible for laying down primers from which Okazaki fragments are generated, is dispensable for late stage DNA replication (Mosig 2001). The reason for this is that ssDNA generated at this stage can be replicated by the join-cut-copy mechanism (**Figure 1.3.1.1.5B**).

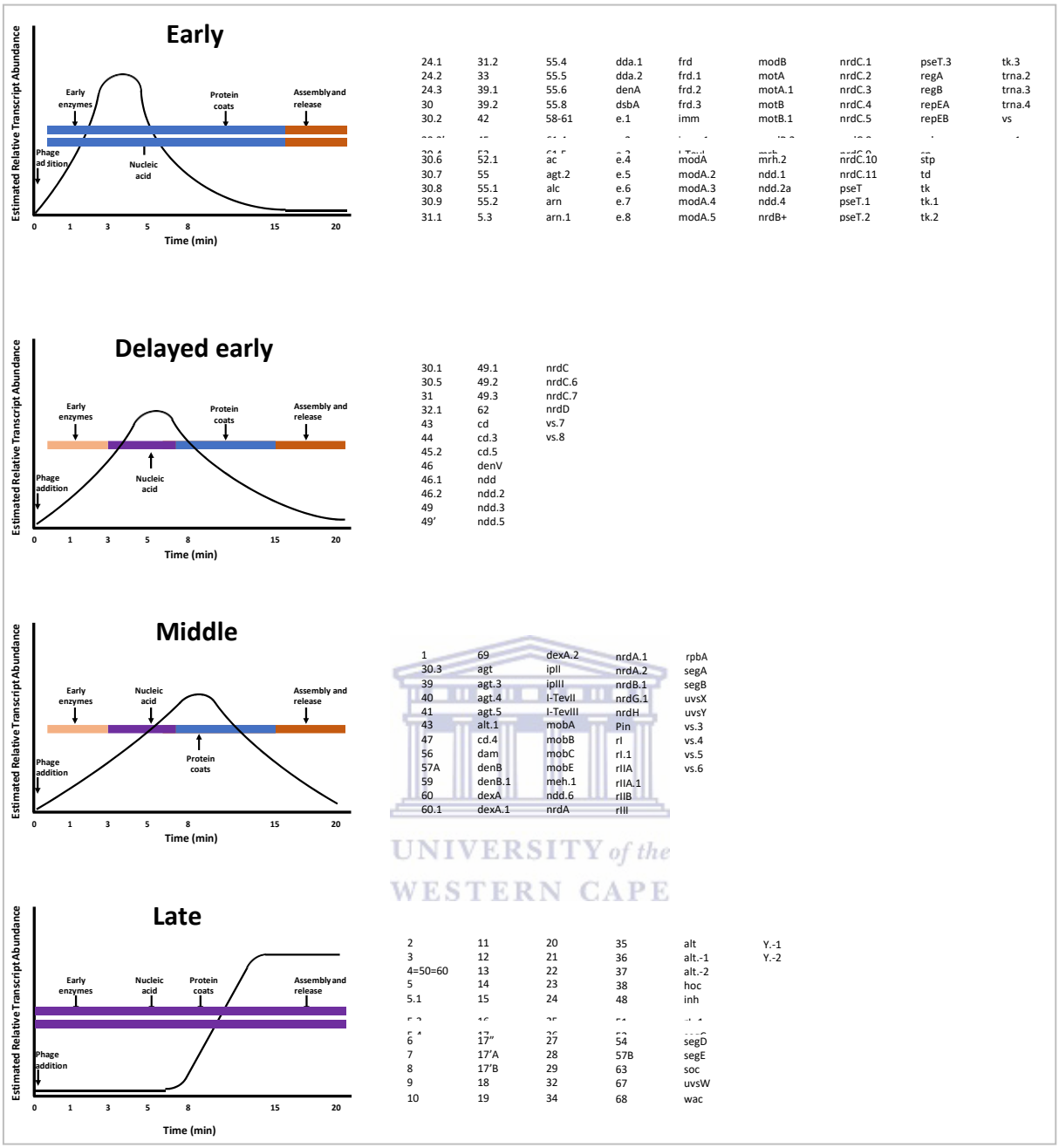


Figure 1.3.1.1.4. Averaged expression profiles of individual representative genes from each temporal class of T4 are shown in the figures at the left of the table. Next to each figure is a list of individual genes displaying similar expression profiles. Genes that do not fit any of the temporal classes such as e, alt.-3, t, and the six genes of the *mobD* cluster are not included here. Adapted from Luke et al., 2002.

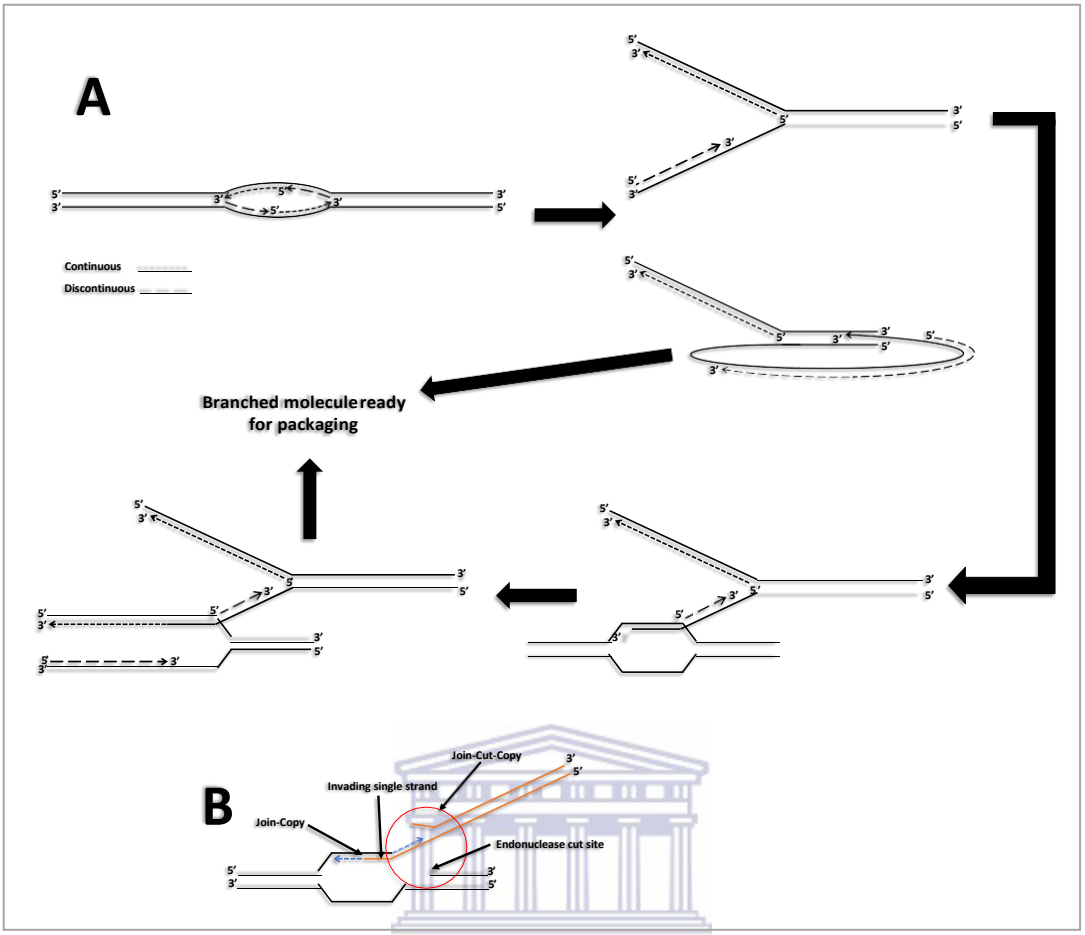


Figure 1.3.1.1.5. Phage T4 DNA replication. A) The join-copy pathway that is initiated as soon as the first replisomes starting at replication origins have reached a chromosome end. This pathway can be considered as a relative of copy-choice recombination. Recombining DNA strands guide replisomes to switch templates. B) The join-cut-copy pathway, where a single stranded nick is introduced to allow replication of invading ssDNA. Adapted from Mosig 1998.

The final stage of lytic development is characterized by the expression of genes for the physical structure of the phage, assembly chaperones and late stage recombination/replication genes from 50 Pl's. The promoter structure of the Pl's is very different to that of Pe and Pm with only a -10 region and no discernable -35 or MotA binding site. The phage also encodes a special sigma factor (σ^{55}) to aide host RNA polymerase in specifically recognizing the late promoters. The dual functional protein gp45 serves as both processivity enhancer of DNA polymerase, but is also an absolute requirement for late gene transcription together with gp33 which facilitates binding between σ^{55} and gp45. gp45 enhances the opening of late promoters, and these activated promoters outcompete Pm's. T4 also produces what are thought to be sigma factor decoys ($\sigma^{32, 38, 70}$) to further reduce expression from host promoters.

Packaging the phage genome into the head is an energy dependent process driven by three enzymes: portal (gp20) protein, small (gp16) and large terminase (gp17) subunits. The portal protein, as the name suggests, is the protein situated at the base of the phage head through which the phage genomic DNA is fed by the terminase. It is therefore also the site of attachment of the packaging motor to the head (Sun *et al*., 2008). The large subunit terminase is the main enzyme responsible for DNA packaging and five subunits come together to make up the T4 packaging motor. On binding ATP, the C-terminal of gp17 moves from an open conformation to a closed conformation bringing it in contact with the DNA. Hydrolysis of ATP causes a twisting motion in one portion of the N-terminal domain of gp17 pulling the C-terminal domain, attached to the DNA, towards the N-terminal (attached to the portal protein). This movement therefore advances the DNA by 6.8 Å (2 base pairs). Each of the five subunits cycles through these motions in turn and in so doing forces the DNA into the head. The small terminase subunit is not essential for packaging, but enhances gp17 activity 100-fold at low gp17 concentrations and its absence results in lower packaging efficiency *in vitro*. This enhancement happens through stimulation of gp17's ATPase activity. Different phages package their genomes in different ways. Some, such as T4, perform "headful" packaging, instead of cleaving the genome at a specific site (*cos*). This indicates that the phage packages a little more than one full length of its genome into the head. The reason it can do this is because the genome is concatenated end-to-end during replication. The main substrate for these enzymes is the concatenated phage genome produced during replication, as this is packaged at 100-fold greater efficiency than externally added mature DNA. Gp17 also possesses endonucleolytic activity and is further responsible for cutting the genome to length when the head is full. The mechanism by which gp17 senses when the correct amount of DNA has been packaged is not known, but may be mediated through changes in the portal protein (Smith *et al*., 2001). Finally, the terminase disconnects from the filled head, and the head is sealed with head completion or adapter proteins to avoid the phage genome from leaking out. The terminase, still connected to the concatenated DNA, can attach to a new prohead to repeat the process up to 12 times (Orlova 2012).

The dynamics of T4 development under various growth conditions have also been accurately modelled (Rabinovitch *et al*., 1999), which posited that the burst size is dependent on the "size" of the protein-

synthesizing system when the phage infects. Thus, the smaller the size of the cell on infection, the smaller the final burst size.

The lytic cycle is most often associated with exponentially growing host cells, however it was recently established that phage T4 can still infect cells during stationary phase (Bryan *et al*., 2016). Two behaviors were observed: The phage was seen to go into a "hibernation" mode where the phage does not replicate but starts protein production and stops full phage particle development half-way. Once fresh nutrients become available to the host, phage particle formation resumed. The phage also managed to scavenge, *via* the host metabolism, all nutrients available to it, including those of lysed bacteria in the surrounding medium to produce a small number of progeny. This shows that although under laboratory conditions idealized models are developed, the picture may be rather different in nature. The purely lytic nature of T4 has to be reconsidered in light of these new modes of existence.

Release of phage progeny is best studied in the phage Lambda system. The reasons for this are that it is virtually impossible to establish and maintain lysis deficient strains of virulent phages as well as the difficulty in syncing infection of a culture to capture lysis physiology and timing, something that's easily done with temperature sensitive mutants of lysogens, such as phage Lambda. For these reasons, this aspect of the lytic cycle will be described in the below section.

1.3.1.2 Lysogenic cycle – *Escherichia* virus Lambda

The siphovirus phage Lambda, which also infects *E. coli*, has been intensely investigated and has been the main workhorse behind the discovery of gene regulation principles (Dodd *et al*., 2005). The switch between lysis and lysogeny is stochastic and controlled by an intricate molecular regulatory network. Numerous attempts have been made to mathematically model this system to better understand it, however the decision seems primarily dictated by the MOI and host cell volume (Joh and Weitz 2011). The regulatory network has evolved such that each "lifecycle" is given an opportunity to take hold post- infection, meaning it is impossible to proceed down one developmental pathway without going at least part way down the other as well (Wulff and Rosenberg 1983). It has been demonstrated that a low MOI

(1) usually results in induction of the lytic cycle and *vice versa* (Kourilsky 1973; Kobiler *et al*., 2005). Also, the smaller the cell volume on infection, the greater the chances of the lysogenic state being induced and *vice versa*, pointing to the concentration of phage DNA (copy number) inside the cell as a major determining factor (St Pierre 2008; Zeng 2010).

The lysogenic cycle in phage Lambda is controlled by the expression of two genes, CII and CI (**Figure 1.3.1.2.1**, **1.3.1.2.2** and **Table 1.3.1.2.1**). Following infection, and depending on the MOI and cell metabolism, CII is produced from P_R, accumulates and dimerizes. The dimer binds to P_{RE} and induces transcription of CI. CI monomers dimerize and bind to the DNA at two operator sites (O_L and O_R). If a sufficiently high concentration of CI is reached, the further polymerization of this protein leads to a complex which induces a DNA bend over the region between the two operator sites blocking transcription from P_R and P_L shutting down transcription of the early-lytic genes, and with it, the activation of the lytic pathway (Dodd *et al*., 2005). Eventually, binding of CI dimers at O_R^3 represses CI transcription as it obscures the RNA polymerase binding sites. Once lysogeny is established following phage genome integration (see below), small bursts of transcription from P_{RM} (producing CI) are enough to then maintain the lysogenic state (Zong *et al*., 2010).

Table 1.3.1.2.1. Main regulatory elements involved in the switch between lysis and lysogeny in phage lambda

Regulatory element	Function
Proteins	
CI	At low concentrations, a repressor of P_R and P_L and an activator of P_{RM}; at high concentrations also represses P_{RM}
CII	An activator of P_{RE} and P_{int}
CIII	Controls the stability of the CII protein by inhibiting the action of the host encoded HflB(FtsH)-HflC-HflK protease complex
Cro	At low concentrations, a repressor of P_{RM}; at high concentrations also represses P_L and P_R
N	An antiterminator at tL1, tR1, tR2, and other terminators
Q	An antiterminator for late gene transcription
Promoters	
P_R	Major rightward transcription
P_L	Major leftward transcription
P_{RM}	Transcription for repressor maintenance
P_{RE}	Transcription for repressor establishment
P_{int}	Transcription of genes for integration and excision

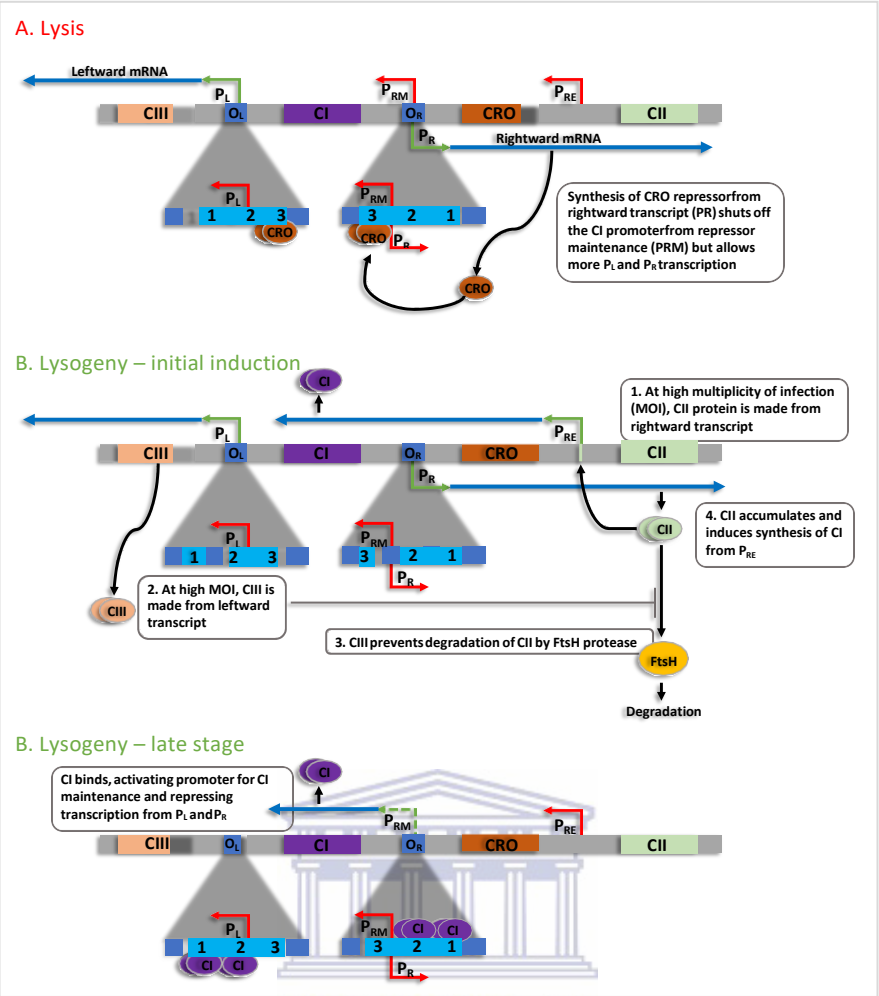


Figure 1.3.1.2.1. The phage lambda lysis/lysogeny decision. A. Lysis pathway. The right and left operators, O_L and O_R, possess three binding sites each for the CI and CRO repressors. Synthesis of CRO repressor shuts off the CI promoter for repressor maintenance. The phage chooses lysis. B. At high multiplicity of infection, CII protein is made and triggers the CI promoter for repressor establishment. CI is made. Once made, CI binds to sites 1 and 2 in the operator regions, turning on the promoter for CI maintenance while repressing the leftward and rightward promoters needed to make more phage particles. The phage chooses lysogeny. Adapted from https://tinyurl.com/y8a8lq24

Although the ratio of CI to Cro was thought to be the tipping point for the switch between lysis and lysogeny (bistable switch), the level of Q protein in relation to CI has recently been shown to be more important in this decision-making process (Kobiler *et al.*, 2005; Joh and Weitz 2011; Svenningsen and Semsey 2014; Casjens and Hendrix 2015) and the current view is that CII is the linchpin on which the lysis-lysogeny decision hangs (Casjens and Hendrix 2015).

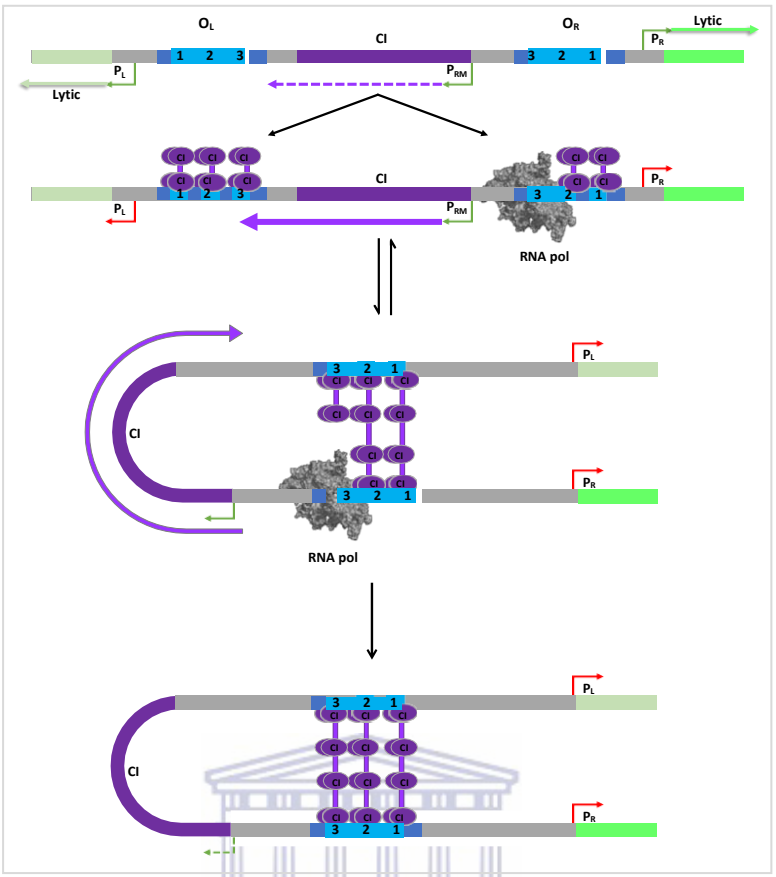


Figure 1.3.1.2.2. CI-mediated regulation of the P_R, P_L and P_{RM} promoters involves multiple levels of cooperativity. The DNA is shown as a bar with the CI gene in purple and the left and right early lytic operons coloured light and bright green respectively. RNA transcripts are shown as coloured arrows and promoter start points as bent arrows. CI monomers (purple ovals) bind as dimers to operators (numbered 1-3) at O_L and O_R, and adjacent dimers interact further to form tetramers through C-terminal domain contacts. CI binding blocks RNA polymerase access to P_R and P_L. However, the N-terminal domain of a monomer at O_{R2} contacts the s4 region of the s subunit of RNA polymerase while s4 is bound to the –35 region of P_{RM}, thereby activating transcription from that promoter. Further interaction of CI C-terminal domains drives its octamerization by long-range DNA looping, aligning O_{R3} and O_{L3} for cooperative CI binding to repress P_{RM}. Adapted from Dodd et al., 2005

As often happens, the development of new tools to probe the questions about regulation of this switch at a single cell level has led to a paradigm shift in what was thought to be a well understood system. Recently, Shao and coworkers demonstrated that instead of just choosing lysis or lysogeny, phage Lambda can exist in a "lyso-lysis" state within its host where phage DNA is integrated into host genomic DNA even when the phage enters the lytic cycle (Shao *et al*., 2016). It was observed that the first phage genomes to replicate would physically separate from one another inside the cell, possibly allowing each phage to make an independent decision whether to enter the lytic cycle, or not. The separation and localization of the genomes may be linked to protein Q as this was shown to work in *cis* and its localization restricted

(Echols *et al*., 1976). This seemed to support their earlier work which suggested that each phage which infects a host makes an independent decision and the fate of the cell is dependent on the consensus "vote" of all phages that infect. They then went on to show that the decision to enter either cycle is made at the level of each individual phage genome and that phages which enter the lytic cycle compete with one another for (replication) resources within one cell, whereas those which enter the lysogenic state act non-competitively, largely due to the far lower number of genomes present (Trinh *et al*., 2017).

Although some phages that enter the lysogenic cycle are maintained as plasmids inside the host (e.g. P1 and N15), and some phages integrate at random sites on host gDNA, as in the case of phage lambda, the phage is programmed to integrate its genome into that of the host at a specific site. To achieve this, the phage has to identify a particular site on the host genome called an attachment site (*att*). The site of recombination on the bacterial genome is referred to as *attB* while the site on the phage is called *attP*. The scanning of the genome for this specific sequence is thought to be achieved by making use of the natural replication of the host cell to pull the chromosome past the phage genome and associated proteins (Int), which appears to become localized at the point of cell entry (Abbani *et al*., 2014). These results do not appear entirely consistent with those of Shao and coworkers which does show colocalization of *attB* and lambda DNA post-infection over similar time ranges, however does not show a fixed position for lambda nor the migration of *attB* towards the phage genome. The Shao study further shows that during lytic development, the lambda DNA and *attB* are strongly colocalized and clearly migrate to the cell poles (Shao *et al*., 2016).

Proteins from both the phage (integrase; Int) and the host (integration host factor; IHF) are needed for integration to occur. Int is responsible for cleavage, strand exchange, and resealing of the DNA attachment site (**Figure 1.3.1.2.3**; Ross and Landy 1983). The protein has two DNA binding domains which interact with "arm-type" and "core-type" *att* sites. One domain, located in the N-terminal, interacts strongly with "arm" regions of *attP* which are not present in *attB*. A second domain interacts weakly with the "core" region of *attP* and *attB*. A second phage-encoded protein called exicionase (Xis), is inhibitory to the integration process, and is the main regulator of integration and excision. Xis interacts with three DNA sites (X1, X1.5 and X2) located adjacent to *attR* (**Figure 1.3.1.2.3**) as well as

with Int. Here, it induces a DNA bend which favors formation of the excisive intasome (nucleo-protein complex) but antagonizes the formation of the integrative intasome (Abbani *et al*., 2014). Its expression is under control of P_L, which is turned on if cellular DNA damage is detected. Expression from P_L is positively regulated by lambda protein N and negatively by CI. Once DNA damage is sensed through recombinase A (RecA), which induces CI autocatalytic cleavage, the expression of Xis leads to excision of the phage genome as part of lytic induction. A second host encoded protein, the factor for inversion stimulation (FIS) enhances excision (Thompson *et al*., 1987; Numrych *et al*., 1990; Ball and Johnson, 1991).

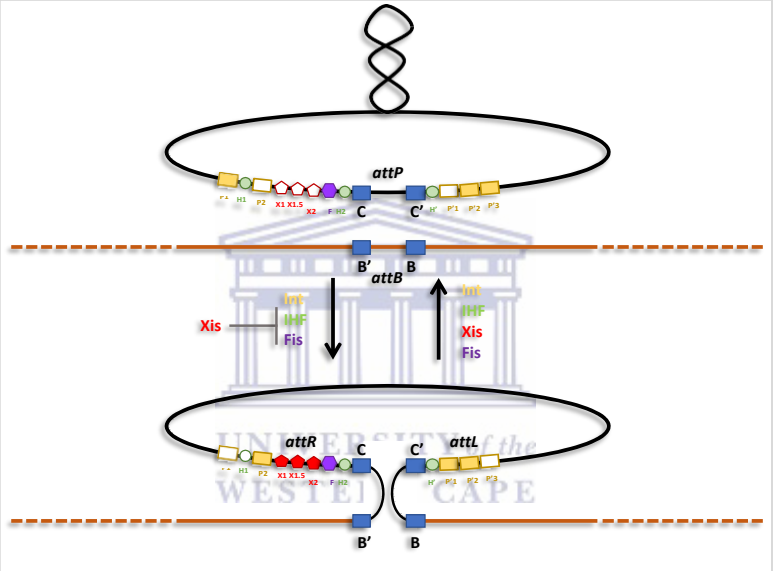


Figure 1.3.1.2.3. Integrative and excisive recombination of phage phage Lambda. Schematic representation of the phage phage Lambda site- specific recombination reactions. The supercoiled phage genome inserts into the *E. coli* chromosome by recombination between the *attP* and *attB* sites to generate the *attL* and *attR* sites that are substrates for excisive recombination. Filled symbols, binding sites used; open symbols, binding sites not used during the integration or excision reactions. Adapted from Abbani et al., 2006.

As said at the end of section 1.3.1.1 I will look at host lysis and phage release here. Expression of late genes begins ±8 minutes after the lytic cycle is induced and protein Q activates expression from P_R. Release of progeny relies on the action of four proteins: holin, endolysin and spanins. Two types of holin have been identified. The first are called canonical holins (phage Lambda, T4, P2), and these assemble at the cell membrane until a threshold concentration is reached. Once this concentration is reached (White *et al*., 2011), they self-assemble to create 1-3 micron-scale (average 340nm up to 1µm) pores in the inner

membrane through which endolysin (R), a peptidoglycan hydrolase, can escape into the periplasmic space. These enzymes attack the (1->4)-beta-glycosidic linkage between N-acetylmuramic acid (MurNAc) and N-acetylglucosamine (GlcNAc) residues in the bacterial cell wall peptidoglycan (Oliveira *et al*., 2013). White and co-workers described the formation of holin "rafts" on holin triggering. These experiments were performed using holin-GFP fusions and the formation of the "rafts" were only seen at times beyond 90 minutes which is well beyond the time normally required for phage induction and host lysis (±50 minutes). Based on their observations they proposed a revision of the "death raft" model where the timing of lysis by different mutants of S105 (the lambda holin) would purely be a function of the critical concentration at which they are triggered. The previous model suggested that the timing was due to the altered effects the mutants had on ion leakage and membrane potential. These two ideas are likely interconnected and difficult to deconvolute.

Membrane potential is expected to play a role in hole formation, as it has been demonstrated that premature triggering of the lambda holin can be effected by treatments (uncouplers such as dinitrophenol) that reduce the proton motive force (Gründling *et al*., 2001). The phage Lambda holin has several charged residues on either side of the inner membrane. The net overall charge on the cytoplasmic side is positive, while the net charge on the outside is negative. These are in opposition to normal proton motive force across the membrane (negative inside, positive outside). The reason uncouplers may lead to premature triggering is by producing a greater local positive charge, thereby allowing protons to flow into the cytoplasm, on the cytoplasmic side of the inner membrane where the concentration of holin is high enough. When relying solely on the holin to perform this action it may require sufficient buildup of holin molecules, and therefore charge, on either side of the membrane to cause an inversion which is why it is delayed until a certain concentration is reached. The role that charge may play has already been demonstrated for the T protein from phage T4 (Moussa *et al*., 2014). The second type of holin, designated pinholin (S68 from phage 21), does not form pores to effect release of endolysin, but rather causes membrane depolarization while the endolysins of these phages are transported across the membrane by the *sec* system (Sao-Jose *et al*., 2000; Oliveira *et al*., 2013).

Holins are further classified as either class I, II or III depending on the number of transmembrane domains present. For all holins studied thus far, these transmembrane domains have been shown to be critical for their function, as opposed to just being an anchoring or locating feature (Pang *et al*., 2013; Moussa *et al*., 2014; To and Young 2014). Most of these domains have a predominantly hydrophobic and hydrophilic side(s), and these rearrange when the critical concentration is reached such that the lumen of the hole that is produced is aligned with the hydrophilic faces of the transmembrane domains (To and Young 2014).

One aspect of the lysis pathway not yet discussed are spanins (i-spanin and o-spanin). These correspond to the products of genes Rz and Rz1 (transcribed from within Rz). In lambda, these are encoded next to the endolysin and holin genes. Where the holin and endolysin are responsible for breaking through the inner membrane and peptidoglycan layer, the outer membrane too has to be breached for effective release of progeny. The spanins connect the inner and outer membranes to each other (Rajaure *et al*., 2015). Endolysin-mediated destruction of the peptidoglycan liberates the spanins to oligomerize and undergo conformational changes which brings the opposing membranes into contact and stimulates IM–OM fusion (**Figure 1.3.1.2.4**).

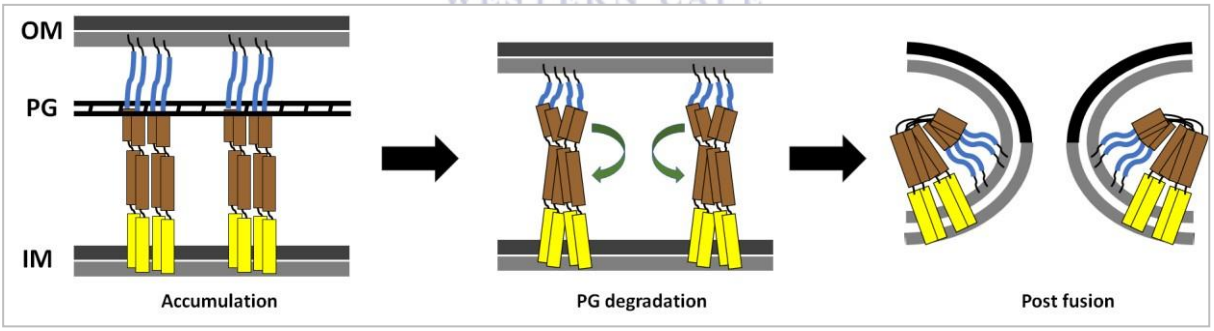


Figure 1.3.1.2.4. Membrane fusion model for Rz–Rz1 (Spanin) lytic function. During the infection cycle, spanin complexes connecting the inner membrane (IM) and outer membrane (OM) accumulate within the lacunae formed by peptidoglycan (PG). Adapted from Rajaure *et al*., 2015.

The concerted actions of these four proteins create a hole in the cell wall which results in "local blowout" of the cell cytoplasm contents, and along with it, phage progeny.

1.3.2 Natural phage resistance mechanisms

The most common mutations leading to bacteriophage-insensitive mutant might be expected to be in the proteins or pathways responsible for phage attachment to the host, however there are a range of natural defenses bacteria have developed against phage infection (Moineau and Lévesque 2005). Over fifty of these natural phage resistance mechanisms have been discovered and can be exploited, as discussed below (3.4.1), to engineer phage resistant strains.

Inhibition of phage adsorption

As the phage discussed in this thesis infects a Gram-positive bacterium, a short review of the Gram-positive cell wall structure and phage attachment targets may be relevant at this juncture. The general cell wall of Gram-positive bacteria is composed of thick, multilayered peptidoglycan (sacculus) that surrounds the inner membrane which encloses the cell cytoplasm. Threaded through the outer peptidoglycan layers are anionic polymers composed of glycerol phosphate, ribitol phosphate and glucosyl phosphate repeats. These are referred to as teichoic acid (Silhavy *et al*., 2010). The peptidoglycan layer is composed of disaccharide-peptide repeats coupled through a glycosidic bond to form linear glycan strands. Many of the surface proteins associate with peptidoglycan and teichoic acid *via* noncovalent ionic interactions, whereas others are covalently bound to the peptidoglycan through stem peptides (Silhavy *et al*., 2010).

In addition to having the peptidoglycan layer and cell membrane, many Gram-positive bacteria also have an S-layer. This is usually a monomolecular layer composed of a single (glyco)protein, that surrounds the peptidoglycan layer (**Figure 1.3.2.1**, **Table 1.3.2.1**). These proteins normally self-assemble into a crystal lattice-like structure which produces 2-8nm pores covering up to 70% of the cell surface (Sára and Sleytr 2000). The proteins that compose the S-layer in many Firmicutes have been shown to contain S-layer homology (SLH) domains which serve to bind secondary cell wall polymers (SCWP) and not peptidoglycan or teichoic acid (Ries *et al*., 1997). It has been determined for *B. anthracis* that these SLH motifs attach to pyruvyl moieties present on the SCWP (Kern *et al*., 2010). The enzyme shown to be responsible for attachment of these pyruvyl moieties is pyruvyl polysaccharide transferase (CsaB). Although the S-layer in *Bacillus anthracis* is primarily composed of two major

proteins (EA1 and Sap), more than 22 other SLH-domain containing proteins have been identified on its genome (Fagan and Fairweather 2014). These are thought to become part of the S-layer at different growth stages. The S-layer proteins of *Lactobacillus* species do not possess SLH domains however also bind to SCWP such as teichoic acids, lipoteichoic acids and neutral polysaccharides (Åvall-Jääskeläinen and Palva 2005).

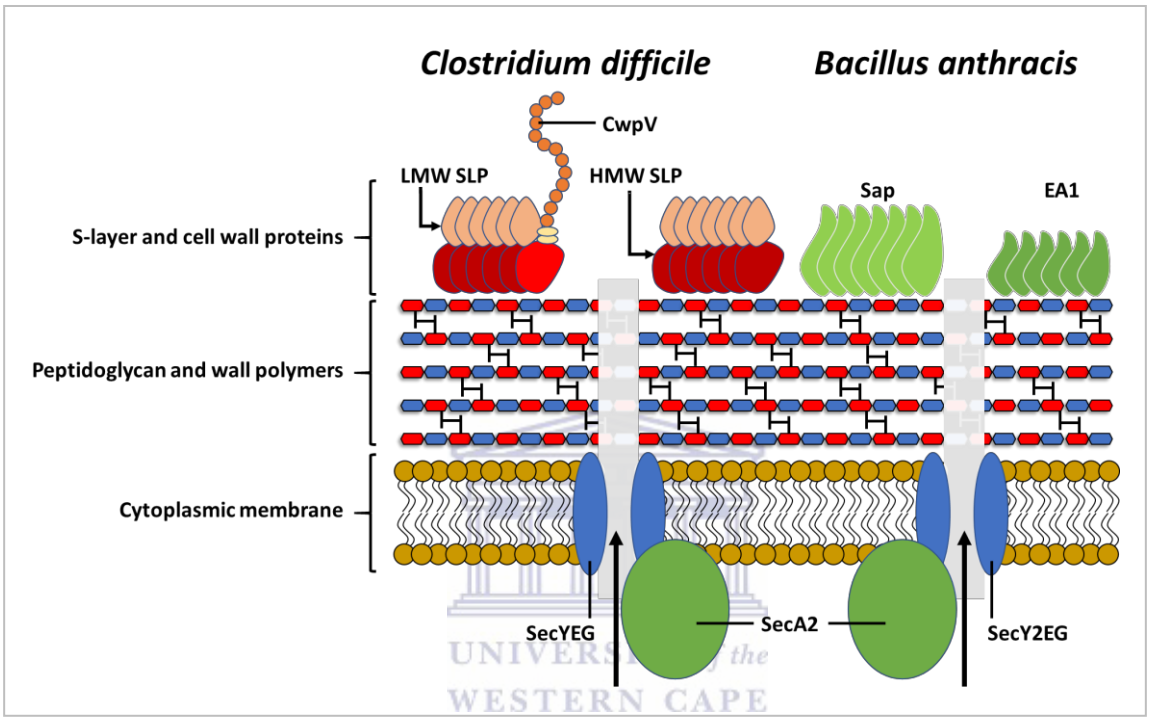


Figure 1.3.2.1. Cross section of the Gram-positive cell wall depicting secretion of bacterial S-layer proteins. In the Gram-positive bacteria *Clostridium difficile* and *Bacillus anthracis*, secretion of the S-layer precursors are mediated by the accessory Sec secretion system. The proteins contain an amino-terminal signal peptide which directs the nascent polypeptide to the secretion apparatus and is cleaved upon membrane translocation (indicated by the black arrow). In both these bacteria, translocation requires the accessory ATPase, SecA2. Following recognition by SecA2, the nascent polypeptide is translocated across the membrane through a pore that consists of SecY, SecE and SecG (in *C. difficile*) or SecY2, SecE and SecG (in *B. anthracis*). Adapted from Fagan and Fairweather 2014.

Table 1.3.2.1: Summary of SLPs (Taken from Fagan and Fairweather 2014)

Organism	SLPs	Features
Campylobacter fetus	SapA SapB	High-frequency antigenic variation of the S-layer by recombinational switching of *sap* homologues; secreted by a specific type I secretion system
Clostridium difficile	SlpA and the CWP family	SlpA is essential for cell growth; S-layer functionalized by decoration with up to 28 additional CWPs; secreted by the accessory Sec system; mediates interactions with epithelial cells; activates dendritic cells
Bacillus anthracis	Sap, EA1 and the BSL family	Sap and EA1 are alternate SLPs; S-layer functionalized by decoration with BSLs; secreted by the accessory Sec system; anchored *via* interaction with pyruvylated SCWP
Caulobacter crescentus	RsaA	Secreted by a specific type I secretion system; anchored *via* interaction with LPS
Aeromonas salmonicida	VapA	Secreted by a dedicated type II secretion system
Geobacillus stearothermophilus	SbsA SbsB SbsC SbsD SgsE	Anchored *via* interaction with pyruvylated SCWP (for SbsB) or *N*-acetylmannosaminuronic acid (for SbsA, SbsC, SbsD and SgsE); glycosylated
Tannerella forsythia	TfsA TfsB	Glycosylated; S-layer includes both SLPs; glycosylation required for biofilm formation; S-layer essential for virulence
Lactobacillus crispatus	CsbA SlpA SlpC	CsbA mediates binding to types I and IV collagen
Deinococcus radiodurans	SlpA Hpi	S-layer includes both SLPs; has a role in maintenance of envelope integrity
Synechococcus spp.	SwmA	Glycosylated; required for swimming motility

Few phage receptors have been identified in Gram-positive bacteria. Those that have been identified include the inner membrane protein PIP (phage infection protein), shown to be necessary for infection of *L. lactis* by phage c2 (Geller *et al*., 1993). YueB, encoded by a putative type VII secretion system gene cluster in *B. subtilis*, was identified as receptor for model phage SPP1 (Jakutytė *et al*., 2011). For *L. delbrueckii* phage LL-H, glucose substituted lipoteichoic acid serves for reversible binding and non- substituted glycerol subunits for irreversible binding (Munsch-Alatossava and Alatossava 2013).

Blocking the ability of the phage to attach to the host cell in the first place is one method by which host cells can avoid phage infection. The receptors on the cell surface have to be in the correct conformation/chemical state and available for binding of the phage. Lytic phages put great selective pressure on their hosts because of the predator-prey relationship and the presence of phages therefore rapidly selects for hosts that adapt to prevent phage infection. The hosts do this through mutation of proteins to which the phages attach. If the protein is non-essential, the mutation may also be in the promoter sequence from which the gene is transcribed. The host can produce decoys such as outer membrane vesicles to which the phage can bind nonproductively, titrating phage away from the host

cell (Manning and Kuehn 2011). To block access to receptor features the cell may produce more of an extracellular polymer or capsule such as to block access to the attachment site (Scholl *et al*., 2005). A good example of blocking of the phage receptor, is the ability of *Staphylococcus* species to prevent lytic infection by certain podoviruses. It was demonstrated that *Staphylococcus* species carrying a functional *tarM*, an accessory wall teichoic acid glycosyltransferase, protects cells from lytic podovirus infection through α-O-GlcNAcylation of wall teichoic acid (Li *et al*., 2015). Interestingly the pressure put on hosts to avoid phage infection may be responsible for phase variation of some opportunistic human pathogens and their ability to avoid host defenses (Seed 2015).

Restriction-Modification (R-M) systems

Various classes of nucleic acid degrading enzymes have been recognized (exonuclease, endonuclease, ribonuclease, Apurinic/apyrimidinic endonuclease, homing endonucleases, Cas enzymes). Bacteria possess restriction endonucleases capable of cleaving the phosphodiester bond within a polynucleotide (dsDNA) sequence generating 5′-, 3′ overhangs or blunt ends. To prevent self-cleavage of host DNA, the host also encodes a cognate methylase (MTase) for each restriction endonuclease, that transfers a methyl group from S-adenosyl methionine to the C-5 carbon or the N^4 amino group of cytosine or to the N^6 amino group of adenine (Vasu and Nagaraja 2013). These R-M systems are ubiquitous in bacteria having been identified in >90% of bacterial genomes sequenced, with >80% having multiple R-M systems. Phages in-turn have responded to the evolution of these R-M defense mechanisms in several ways, one of which is by modifying their nucleotides to avoid digestion through substitution of cytosine for hydroxymethylcytosine, thymine for uracil or 5-hydroxymethyluracil (Snyder *et al*., 1976; Neubort and Marmur 1973). Phage DNA produced inside a host can be modified by a host encoded MTase or sometimes phage encoded methylases, to prevent degradation on infection of an identical host. However, should the phage infect a closely related bacterium with endonucleases which digest at sites other than where the DNA has been protected or can cope with the methylation modification, the phage genome will be degraded preventing phage spread. This was first observed in *E. coli* using the model phage lambda, when it was found that phage prepped on *E. coli* B grew poorly on *E. coli* K12 (Luria and Human 1952) and subsequently observed in other phage-host systems (Sharp *et al*., 1986). Phages

further encode proteins which block digestion by restriction endonucleases such as the OCR (overcome classical restriction) or Ard (alleviation of restriction of DNA) proteins. These mimic small stretches of DNA to which the restriction enzymes bind, rather than phage genomic DNA, thereby titrating enzyme away from the genome (Vasu and Nagaraja 2013). R-M systems also put phages under selective pressure to reduce the number of restriction sites on their genome. It was demonstrated that lactococcal R-M systems were more effective against so called "young" or newly infecting P335 species which harbored many restriction sites, versus the 936 phages which are often encountered and represent an older lineage (Moineau *et al*., 1993). The 936 phages have vastly reduced numbers of a series of restriction sites on their genome. The R-M systems, together with CRISPR, could therefore be seen as a first line of defense against newly infecting phages.

Abortive infection (Abi)

Abortive infection mechanisms are post-infection systems that trigger programmed cell death by the infected bacterium. As this kills the infected host, the phage can't use the host to produce more phage progeny, thus limiting the spread of phage. The *rexAB* system was the first Abi described, and in *E. coli*, limits the spread of phage lambda. RexB, a transmembrane protein, is activated by RexA which in turn is activated by a protein-DNA complex that is formed as part of phage recombination and replication (Dy *et al*., 2014a). Once activated, the four transmembrane helices of RexB form a pore in the membrane through which ions can leak from the cell, thereby depolarizing the membrane (Parma *et al*., 1992). Lit and PrrC both act as translation inhibitors blocking translation of phage and host proteins on infection with phage. The T4 major capsid protein Gol (GP23) binds to the Lit protein, encoded by a defective prophage *e14* on the *E. coli* chromosome, activating the protease activity of Lit. This cleaves the Tu elongation factor responsible for catalyzing the binding of aminoacyl-tRNA to the ribosome. Thus, without this factor, translation stalls and both phage and host proteins can no longer be produced.

L. lactis harbors over 23 described Abi systems which are mostly plasmid encoded (Dy *et al*., 2014a). Three of these systems are carried on one plasmid, pTR2030: AbiA, AbiZ and LlaI. These three systems block phage infection and reproduction at three levels namely, degradation of incoming phage DNA

(LlaI – R-M system), inhibition of phage DNA replication (AbiA) and accelerating the timing of lysis (AbiZ). The killing activity of AbiZ is mediated through interaction with the phage-encoded holin (Durmaz and Klaenhammer 2007). This interaction leads to early lysis of the bacterial host, not giving enough time for phage progeny to form, although the exact mechanism is not fully understood. AbiA appears to be related to reverse transcriptases, and catalyzes long, random, nontemplated nucleotide polymerization. This system provides immunity against a broad range of phages (936, c2, and P335), however the mechanism by which it does so is unknown.

The AbiD1 system is one of the best studied. The mRNA from which AbiD1 is translated is a very short-lived molecule and translation from this RNA is made inefficient by the presence of a stem loop structure in the middle of the transcript meaning that AbiD1 is silent under normal conditions. A phage encoded protein, ORF 1 from phage biL66 (936 group), stabilizes the mRNA and enables translation and therefore production of AbiD1. The protein inhibits the product of ORF3, part of middle protein expression in phage groups 936 and c2, responsible for resolving branched DNA structures during phage DNA replication (Bidnenko *et al*., 1998). This halts phage DNA replication, resulting in abortion of phage multiplication.

A recently discovered Abi system is that of ToxIN, which is a classic toxin-antitoxin (TA) system, and shares homology with the AbiQ system from *L. lactis* (Dy *et al*., 2014b). TA systems are thought to play a role in many cellular processes, although they are often encountered on plasmids where they serve as stable inheritance systems. The antitoxin protein is a short-lived species while the toxin protein is not degraded as quickly. When the plasmid is present in the cell, production of the antitoxin from the plasmid neutralizes the toxin by binding to it. If the plasmid is lost from the host, the short-lived antitoxin breaks down, leaving the toxin to kill the cell. In this way, the population only ever consists of cells harboring plasmid. Class III TA systems are characterized by a protein serving as toxin and an RNA molecule which interacts with it, as inhibitor. The ToxIN system is the prototypical Class III TA system (Fineran *et al*., 2009). The ToxN protein cleaves free (as opposed to ribosome-bound) mRNA, thereby preventing translation. The ToxI (ToxN inhibitor) blocks the activity of ToxN through directly binding to it. How phage infection brings about the activation of ToxIN is currently unknown.

Clustered Regularly Interspaced Short Palindromic Repeats (**CRISPR)**

CRISPR-Cas can be described as an adaptive bacterial immune system, that protects the bacterium against invading mobile DNA elements such as plasmids and phages (Rath *et al*., 2015). In general, the genomic locus for this system consists of a suite of genes (*cas*, *csn, csm, cmr* or *csh*), located adjacent to a series of sequence repeats interspersed with spacer sequences. These spacer sequences correspond, in their nucleotide sequence, to segments of the invading DNA against which the CRISPR system has had to defend against in the past. The proteins produced from the genes associated with this DNA repeat-spacer array are responsible for incorporating new spacers from DNA invading for the first time, into the spacer array, as well as using this array to act against invading mobile DNA for which spacers are already in the array. The CRISPR-Cas system consists of a multiprotein complex called Cascade responsible for adaptation (spacer acquisition) and interference (targeting foreign DNA for degradation). Thus, the spacers in the array serve as a way for the bacterium to "memorize" which plasmids or phages it has encountered before (**Figure 1.3.2.2**). Not only do these spacers serve as a way to neutralize phages previously encountered, they can take part in new spacer acquisition from these DNA elements (primed spacer acquisition). Several CRISPR-Cas systems have been identified (Type I, II and III) as well as various subtypes of these (A-F). The various systems encode different numbers of *cas* genes and have different mechanisms of adaptation and interference. The way cells incorporate spacers from new invading DNA is not fully understood. For Type I-E systems, two proteins, Cas1 and Cas2 (endonucleases) are required for naïve (first time infection) spacer acquisition. The Cas1-2 complex recognizes a particular sequence on the invading DNA known as the protospacer adjacent motif (PAM). The PAM sequence together with the spacer to eventually be incorporated, together, are known as the proto-spacer. The complex formed by Cas1 and Cas2 is also responsible for incorporating the proto-spacer into the array. Whether the proto-spacer is copied or cut from the target molecule is not entirely clear. The repeats in the array form a cruciform secondary DNA structure that is, together with the leader sequence, recognized by the Cas1-2-proto-spacer complex. Integration is achieved through a mechanism similar to that used by transposases. The need to have the leader sequence, only present at the 5' end of the array, recognized means that new spacers are only incorporated at one end of the array. The spacers closest to the leader therefore represent the most recent infections. Host

encoded factors such as polymerases, ligases and recombination proteins participate in adaptation, by perform their generic roles during the process.

The mechanism of interference works on the basis of RNA directed targeting of the molecule to be degraded. An RNA molecule is made from the spacer, complementary to the incoming molecule, *via* host encoded RNA polymerase (Gorski *et al*., 2017). In Type I-E systems five proteins (CasA-E) make up the Cascade complex. The Cas6-type endoRNase (CasE) is central to this process of pre-crRNA maturation. The pre-crRNA consists of the spacer region flanked by repeats on either side, which, when single stranded forms hairpins. These hairpins are recognized by CasE and cleaved 8bp upstream of the spacer sequence and at a variable position at the 3' end repeat. The small bits of repeat sequence still attached to the spacer act as handles for two subunits (CasE-5' and CasD-3') of the Cascade complex to hold onto it. Once the mature crRNA is bound to Cascade, and the crRNA has hybridized to its target, Cas3 (nuclease/helicase) is recruited to the site. Cas3 then nicks the DNA strand and proceeds to degrade the whole DNA molecule thereby limiting phage replication (Brouns *et al*., 2008). It is important to note that many variations of this mechanism exist.

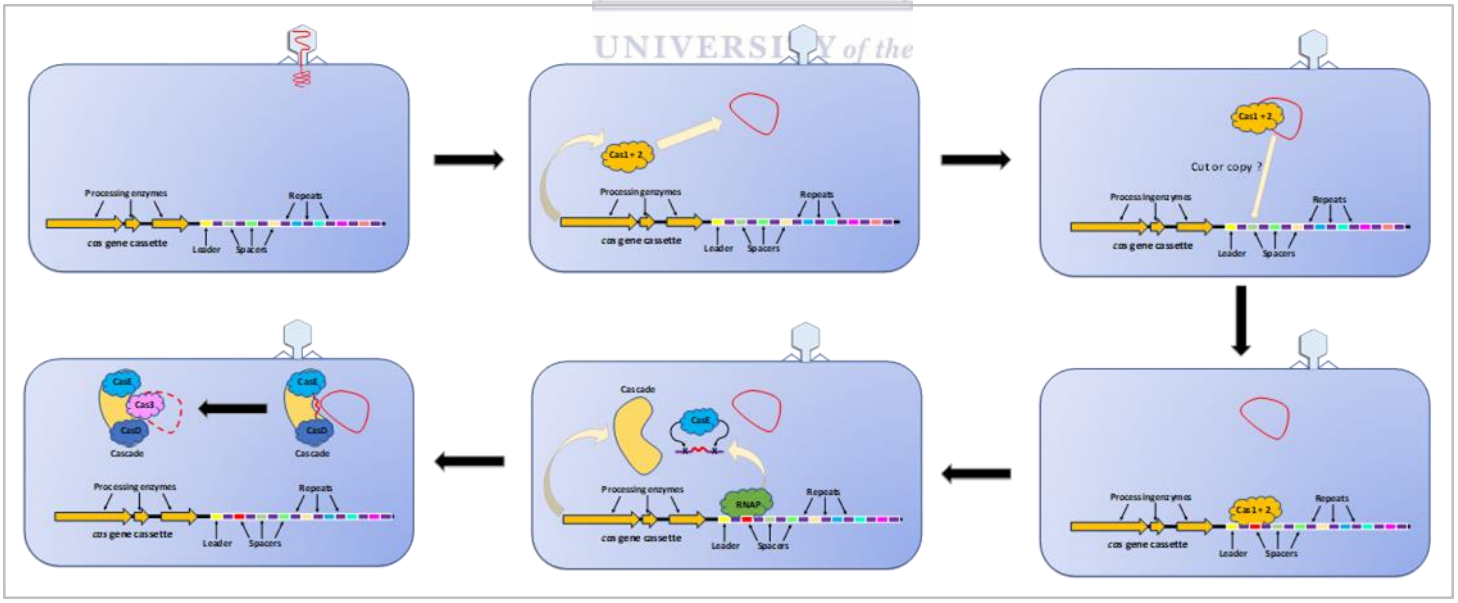


Figure 1.3.2.2. Schematic representation of the steps involved in CRISPR-Cas Type I-E adaptation and interference. See text for more detail.

Phages have also evolved systems to circumvent CRISPR-Cas degradation. An example is a group of Mu-like phages that infect *Pseudomonas aeruginosa* which are insensitive to the Type I-F CRISPR-

Cas system (Bondy-Denomy *et al*., 2013). These phages encode a series of small ORFs between a Mu G homolog and a scaffold protease which confers resistance. These genes were not present in all related genomes studied, and those without this suit of genes were found to be sensitive to the CRISPR-Cas system. It was found that the system could inactivate the Type I-F systems in other *Pseudomonas* sp., however not the closely related Type I-E system in *E. coli*. They showed that for two of these phage-encoded proteins (ORF 30 from D3112 and ORF 35 from JBD 30) they bind to the Cascade complex inhibiting its ability to bind to PAM sequences, however they do so through binding at different sites. A third phage encoded protein (ORF 35 from JBD5) binds to the Cas3 equivalent in this system preventing it complexing with Cascade and enabling its DNA degradation capability (Bondy-Denomy *et al*., 2015).

Superinfection immunity (SI) / Superinfection exclusion (SE)

Lysogenic conversion was first described by Hatano *et al*., 1959. This term describes the occurrence of phenotypic changes that accompany infection by temperate phages and expression of their genes. One of these changes is protection of the host from further infection by the same phage. Two molecular mechanisms are responsible for this phenomenon: SI and SE. In SI, the host is protected from lytic phage development through expression of the main prophage repressor (CI in lambda; section 1.3.1.2) which may also act on closely related phages. SE refers to the ability of a phage encoded protein to prevent DNA injection of related phages that bind to the cell after infection by the first phage. Most of the proteins that mediate exclusion have one or more transmembrane domains, perhaps except for the Glo protein of *Vibrio cholerae*, signifying where they localize to in the cell and the site at which they serve to block the injection (Ali *et al*., 2014). For phage TP-J34, that infects *S. thermophilus*, it was demonstrated that its Ltp protein was responsible for SE and it likely targets the tape measure protein (TMP) at the tail tip (Bebeacua *et al*., 2013). The TMP is an integral part of the tail of sipho-and myoviruses. The length of the TMP physically determines the length of the tail, and is usually the longest ORF on a phage genome. Although the use of lysogens as industrial workhorses is not preferred, lysogens of starter culture bacteria could be protected from infection by virulent versions of prophages they harbor, through SI, making a case for their use in industry (Marcó *et al*., 2012).

1.3.3 Viruses of thermophiles including those of *Parageobacillus* / *Geobacillus* sp.

Phages that infect thermophilic microorganisms were being described not too long after their initial discovery and thus the ability of these entities to adapt and infect at high temperature has been known for a long time (Koser 1926). Early on most of these phages were isolated on high temperature *Bacillus* species (White *et al*., 1955; Thompson and Shafia, 1962; Welker and Campbell 1965; Egbert and Mitchell 1967; Sharp *et al*., 1986). More recently phages infecting thermophilic bacterial genera *Meiothermus*, *Thermus*, *Geobacillus* and *Parageobacillus*, have been isolated (Nagayoshi *et al*., 2016). The other viruses known to infect high temperature microorganisms are those that infect archaea, and these were first described in 1974 (Torsvik and Dundas 1974). Although some archaeal viruses are also tailed, they mostly display other, highly unique morphologies compared with those that infect eubacteria. Viruses that infect thermophiles have been isolated from environments ranging in temperature between 50℃ and 92℃. The filamentous bacteriophage ϕOH3 has been shown to be the most thermostable bacteriophage known, stabile up to 70℃ which is similar to that of ϕYS40, both of which infect *Thermus thermophilus* (Nagayoshi *et al*., 2016). The study of thermophilic phage morphologies has led to the introduction of several new phage families, including *Lipothrixviridae*, *Rudiviridae* and *Fuselloviridae*, and their promoter structures have, in some cases (ϕYS40), been shown to be different to that of mesophilic phage (van Zyl *et al*., 2015). Thus, the study of thermophilic phage might shed light on phage-host interaction strategies that are perhaps radically different from those of mesophilic phage.

Although several phages (**Table 1.3.3.1**), isolated from deep-sea vents, hot spring sediment and compost heaps, have been discovered recently that infect species closely related to *P. thermoglucosidans*, none have been identified that specifically infect this organism (Liu *et al*., 2006; Liu *et al*., 2008; Liu *et al*., 2009; Wang *et al*., 2010; Doi *et al*., 2013; Song *et al*., 2011; Marks and Hamilton 2014). They have been described in terms of their morphology and genetic content; however, none except for GVE2, have been interrogated further to establish if they share mechanisms of infection, replication, assembly and release with known phages. All these phage genomes are smaller than 50kb and except for GVE2 are lytic on the respective hosts they were isolated on. There are many podoviruses known to infect Gram-positive bacteria including Firmicutes, thus there is no obvious reason why only sipho- and myoviruses have been isolated thus far, apart from sample size (Grose *et al*., 2014; Li *et al*., 2015).

Table 1.3.3.1. Summary of *Parageobacillus* and *Geobacillus* infecting phages isolated thus far

Phage	Type	Host	Genome size in (bp)	G+C %
GVE1	Siphovirus	*Parageobacillus* sp. E26323	± 41 000	N/A
GBSV1	Myovirus	*Parageobacillus* sp. 6k51	34 683	44.4
GVE2	Siphovirus	*Geobacillus* sp. E263	40 863	44.8
GBK2	Siphovirus	*Geobacillus kaustophilus* ATCC 8005	39 078	43
ϕOH2	Myovirus	*Geobacillus kaustophilus* ATCC 8005	38 099	44.7
D6E	Myovirus	*Geobacillus* sp. E263	49 335	46

First described in 2008, GVE2 was isolated on a *Geobacillus* species (E263) from a deep sea hydrothermal vent and transmission electron microscopy together with its genome sequence revealed it to be a lambdoid siphovirus (Liu and Zhang 2008). Initial studies to look at phage replication and gene transcription showed that the first transcripts were only detected an hour or more post infection (PI) increasing until six hours PI. This suggests a much slower rate of phage gene transcription and replication than other well-studied systems. This could be due to several factors. It may be that this phage-host combination does not face strong competition in its natural environment and there has not been strong selective pressure to select for phages that replicate quickly. Another alternative is that GVE2 is a phage that has only recently begun to infect this host and has not had time to optimise its transcription in the host. Phage metabolism (rate of transcription, replication, assembly and release) are

all dependent on host metabolism. If the host is particularly slow growing, the phage would follow suit. The information currently available on GVE2 and its host's metabolism does not make it clear which of these options is most likely.

The role of several proteins from GVE2 has since been studied. The HNH endonuclease encoded by GVE2 showed the same three-dimensional structure and DNA nicking activity as many other characterized HNH endonucleases (Zhang *et al*., 2017). The endolysin from this phage was also characterized and its primary amino acid sequence suggested that it belongs to the *N*-acetylmuramoyl-L-alanine amidase family and that there were few differences between this thermophilic and related mesophilic proteins (60% identity; 73% similarity to Bacillus virus 1; Ye and Zhang 2008). The only biochemical difference was that the enzyme could operate at high temperature. This indicates that the structure and mechanism of action by these enzymes are the same as those from mesophiles and that the modifications to the amino acid sequence have come about to ensure that the enzyme can fold and be thermostable under high temperature. The role of the endolysin during lysis was probed further, and it was shown to associate with an ABC transporter, apart from its association with the GVE2 holin, by bacterial two hybrid system (Jin *et al*., 2013). The association of an endolysin with host encoded proteins has never been reported for mesophilic phage systems, possibly indicating a unique lysis mechanism.

GVE2 lysogenizes its host and the regulation of the switch between lysis and lysogeny was investigated. A CI homolog identified on the GVE2 genome, divergently transcribed from a CRO homolog, was shown to bind to the promoter region located between the two genes, and its expression was responsible for lysogeny of the host (Song *et al*., 2011). This is similar to many other lambdoid phages such as phage Lambda (*E. coli*; section 1.3.1.2), TP901-1 (*L. lactis*), and P22 (*Salmonella typhimurium*). Given the many genetic, morphological and regulatory similarities to lambda it would appear that this thermophilic phage, and perhaps others, share evolutionary strategies, if not a common ancestor, with mesophilic phages.

Recently, the family *Sphaerolipoviridae*, that infect archaea and contain icosahedral, tailless haloarchaeal viruses with an internal lipid membrane, was expanded to include some thermophilic

bacterial viruses (infecting *Thermus thermophilus*) which share structural and genomic features, suggesting common origins (Pawlowski *et al*., 2014). It could be that other extremophile viruses including the tailed archaeal and bacterial viruses share common ancestry (Krupovic *et al*., 2011).

1.3.4 Phages in industrial fermentations

It has been known, since shortly after their discovery, that phages affect commercial fermentations and it was first noticed in 1923 in fermentations utilizing *Clostridium acetobutylicum* for production of acetone and butanol (AB) (Jones *et al*., 2000). The infection resulted in halving the production for a year. The AB process developed in the United States of America was eventually performed all over the world and was continually plagued by phage infection with factories in Japan being hardest hit or perhaps just best reported on throughout the 1930s up until the 1970s. Phage infections in commercial bacterial processes, are not often reported in scientific literature due to the ramifications it may have for that industry (Moineau and Lévesque 2005). This makes it difficult to assess the extent of the problem. A good example of this is the South African case, where infections in the AB process at the National Chemical Products company (NCP) were never reported publicly but are clearly outlined in company records and an M.Sc. study conducted at Rhodes University, which shows that the company suffered four confirmed phage infections during 1943 and 1980. Two phages, one a podovirus (CA1) and the other a siphovirus with a peculiarly short tail, were identified by electron microscopy (Barber, 1977). Characteristic to most of these infections is that the fermentations produce lower yields, change the ratio of acetone to butanol and take longer than normal to complete. One solution was the isolation of bacterial strains resistant to infection (immunized) through the use of mutagens or culturing in media that produced very high phage titers. However, even these newly developed strains would eventually succumb to infection by different phages (Jones and Woods 1986). If the problem persists today, it is not reported outside the facilities employing these bacteria. Although the AB fermentations were the first documented cases of phage infection of commercial fermentations, there have been many other cases and it appears that phage infection is a ubiquitous problem in bacterial fermentations (Lu *et al*., 2012; **Table 1.3.4.1**). Not many industrial processes make use of thermophilic microorganisms, and as mentioned earlier, high temperature fermentations are expected to be less prone to bacterial, and perhaps

bacteriophage related failure. However, bacteriophages that infect thermophiles are known as described in section 1.3.3, and are expected to eventually become a problem for industries making use of such organisms, especially given the history of phage infections of commercial fermentations.

Table 1.3.4.1. Examples of documented fermentations affected by phage infection (Updated from Moineau and Lévesque 2005)

Bacterial species	Product	Bacterial species	Product
Acetobacter sp.	Vinegar	*Lactobacillus plantarum*	Cucumber fermentation
Bacillus colistinus	Colistin	*Lactobacillus plantarum*	Silage, sauerkraut
Bacillus polymyxa	Polymycin	*Lactococcus lactis*	Buttermilk, cheese, sour cream
Bacillus subtilis	Amylase, protease	*Leuconostoc mesenteroides*	Sauerkraut, buttermilk, sour cream
Bacillus subtilis var. *natto*	Fermented soy beans	*Leuconostoc fallax*	Sauerkraut
Bacillus thuringiensis	Insecticide (BT)	*Oenococcus oeni*	Malolactic fermentation
Brevibacterium lactofermentum	L-glutamic acid	*Pediococcus* sp.	Cucumber fermentation
Clostridium sp.	Acetone, butanol	*Propionibacterium freudenreichii*	Cheese
Corynebacterium sp.	L-glutamic acid	*Pseudomonas aeruginosa*	2-Ketogluconic acid
Escherichia coli	Various biotechnology products	*Streptococcus thermophilus*	Cheese, yogurt
Gluconobacter sp.	Gluconic acid	*Streptomyces aureofaciens*	Tetracycline
Lactobacillus acidophilus	Fermented milk	*Streptomyces endus*	Endomycin
Lactobacillus brevis	Lactic acid	*Streptomyces griseus*	Streptomycin
Lactobacillus casei	Fermented milk	*Streptomyces kanamyceticus*	Kanamycin
Lactobacillus delbrueckii subsp. *bulgaricus*	Yoghurt	*Streptomyces venezuelae*	Chloramphenicol
Lactobacillus delbrueckii subsp. *lactis*	Cheese	*Tetragenococcus halophila*	Soy sauce
Lactobacillus fermentum	Sourdough bread	*Xanthomonas campestris*	Xantham
Lactobacillus helveticus	Cheese		

The industry that has arguably been most affected is the dairy industry, with an estimated 10% of fermentations negatively impacted, and much of what has been learned as to how best to manage infections, based on the work done there. Below I will look at some examples of how this industry has modified its practices and the bacteria they use to combat phage infection.

1.3.5 Methods of phage resistance engineering: Lessons from the dairy industry

Lactic acid bacteria (LAB) are nonsporulating, Gram-positive bacteria used in the manufacture of yoghurt and various cheeses among other fermented products (Park *et al*., 2011). The fermentations to produce these products are usually seeded with one or more LAB, referred to as a starter culture, which comprises of the following genera and species: *Lactococcus lactis*, *Lactobacillus* sp., *Leuconostoc* sp., and *Streptococcus thermophilus*. These bacteria, as already eluded to, are sensitive to phage infection and various techniques and phage resistant strains have been developed to limit the impact phage infection could have on these fermentations. The problem persists due to the re-

introduction of "wild phage" with the incoming raw milk, as well as factory practices such as re-use of whey protein concentrates which could act as a reservoir for phage, movement of personnel and raw materials which causes aerosolizing of phage (Marcó *et al*., 2012). Phage biodiversity, driven by rapid host growth rates, large burst sizes and genomic plasticity further contributes to this persistence. It appears that the bulk of viruses that infect starter cultures belong to the family *Siphoviridae*, with myoviruses being second most abundant, and in particular the lytic 936 group of phages are a problem (Marcó *et al*., 2012; Mahony *et al*., 2015; Muhammed *et al*., 2017). Murphy and coworkers studied thirty-eight 936 variants, and demonstrated that small gene acquisitions/deletions result in phages that are physically distinct from one another, potentially circumventing host mutations to abrogate phage binding (Murphy *et al*., 2016). The diversity of phage in fermented foods is often less than that found in other environmental settings, likely due to enrichment of a few host species during the fermentation process (Park *et al*., 2011; Muhammed *et al*., 2017). In the dairy setting, phage infection of the starter culture is characterized by late onset lactic acid production which results in the inoculated milk not coagulating and the organoleptic or physical properties of the product being altered which has, in some cases, led to products being downgraded (Marcó *et al*., 2012). There is some robustness to these fermentations, and phage are only likely to stop the fermentation altogether if the titre is at, or above, 10^5-10^6 plaque forming units/ml (Moineau and Lévesque 2005).

1.3.5.1 Modified media and strain rotation

One of the early strategies to reduce phage related failure in dairy fermentations was the development of phage inhibitory media (PIM) (Whitehead and Hunter 1945; Gulstrom *et al*., 1979). Calcium, as well as other divalent cations, have been shown to play either a critical role in phage infections or enhance phage infection (Bonhivers and Letellier 1995; Chhibber *et al*., 2013; Mahony *et al*., 2015). Milk, being rich in calcium, therefore represents an ideal medium to facilitate phage infection. At the start of the cheese making process, a starter culture of LAB is prepared at 1% the volume of milk to be inoculated and is cultured in PIM. The concept behind PIM is to remove calcium, and other divalent cations,

through the addition of phosphate and citrate salts to chelate the cations and make them bio-unavailable which limits the ability of phage to infect and proliferate (Gulstrom *et al*., 1979; Suárez *et al*., 2007).

Several BIM strains of LAB have been isolated, either through directed evolution, or through selection of natural mutants that spontaneously arise after prolonged exposure to lytic phage (Moineau and Lévesque 2005). As continued use of one of these strains would eventually lead to infection by phages specific to that strain, knowing the phage sensitivity profile of the starter strains allows for the development of a rotation strategy to avoid the buildup of resident phage populations in the factory. The type of products produced can also be alternated to switch between a moderately thermophilic starter culture (*S. thermophilus*), such as that used for yoghurt, and mesophilic starters used for cheese.

A third technique is to avoid the generation of a starter culture altogether. Here, the vat containing milk to be fermented, is directly inoculated with a very high concentration of starter bacteria, in so-called direct vat inoculation.

1.3.5.2 Examples of phage resistance engineering in industrial bacterial strains

As reviewed earlier (2.3) the phage life cycle provides many points for interference and these have been effectively exploited by researchers in engineering BIM (**Figure 1.3.5.2.1**). Several of the systems described above can be used in combination to generate BIM that are more robust. A drawback to any of the systems described below is that they target infection by individual phages and, apart from perhaps the CRISPR system, no general mechanisms have been identified to defend against a wide range of phages (Moineau and Lévesque 2005).

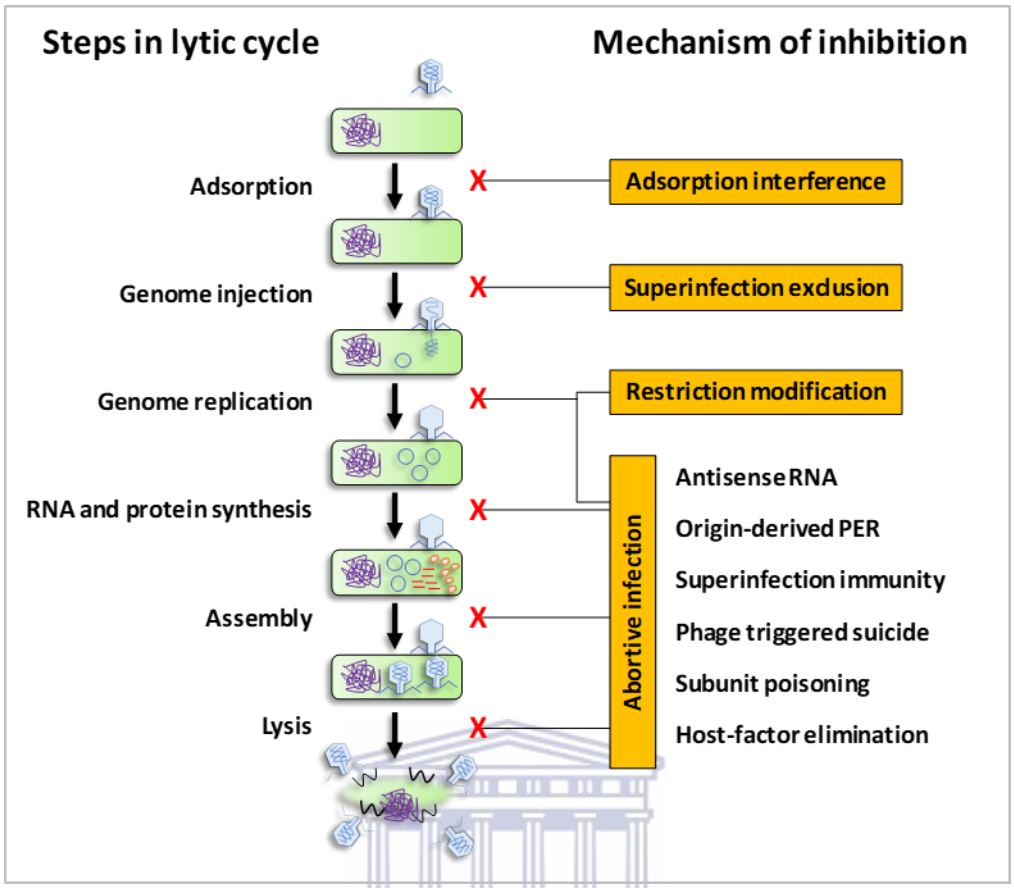


Figure 1.3.5.2.1. Points of inhibition during a generalized phage lytic cycle by the various phage resistance mechanisms used by researchers to develop phage resistant strains (Adapted from Sturino and Klaenhammer 2006).

The PER (phage encoded resistance) system was developed to target phage genome replication to prevent phage proliferation. The system consists of the origin of replication of a particular phage cloned into a high copy number plasmid, which is introduced into the starter culture bacterium. On phage infection, the proteins responsible for phage replication are titrated away from the phage genome to the high copy vector containing the origin. This effectively stops the phage genome from being replicated and new particles being generated. The system has been employed for phage phi50 and phi48 which target *L. lactis* and phage Sfi21 that infects *S. thermophilus* (Sturino and Klaenhammer 2006).

Gene silencing through the use of antisense RNA has been used to target a range of genes, key to the successful replication of phage genomes (DNA polymerase subunits, single stranded binding protein, helicase), formation of structural components (major tail protein, major capsid protein) and genome packaging (terminase) (Walker and Klaenhammer 2000; Sturino and Klaenhammer 2004). This

technique had limited success in terms of reducing the efficiency of plaquing, however it reduced the number of progeny produced substantially. No phage mutants that overcame the newly engineered resistance mechanism could be isolated, suggesting that the system would not easily select for phages capable of infecting the host again. The reason for this is thought to be the length of the antisense RNA used (1.5kb) which means that even if there were mutations at several positions along the RNA, there is still plenty of homologous sequence to allow hybridization between the RNA and the target DNA.

Expression of a restriction endonuclease, without its concomitant methylase, cloned onto a high copy number vector and driven from a phage induced promoter has also proved effective against phage infection (Djordjevic *et al*., 1997). This simple mechanism relies on the enzyme degrading both phage and host DNA that is unmethylated. Through the course of their study, several phage mutants capable of infecting the engineered host were isolated, and the mutations mapped to the Tac31 positive regulator of the promoter used to transcribe the endonuclease from. This indicates how rapidly phage mutants can be selected for once selection pressure is applied.

The CRISPR adaptive immune system, has also been employed in commercial starter culture strains to provide phage resistance. The Du Pont (Danisco) company sell the CHOOZIT SWIFT range of *Streptococcus thermophilus* strains, which have been "immunized" against some known *S. thermophilus* phages through the incorporation of sections of the phage genomes into the spacer array (https://tinyurl.com/oxtkvbj; Barrangou *et al*., 2007). A severe drawback to the CRISPR system, is that it has been shown that even a single base change to the region of the phage genome which was incorporated into the spacer array, could nullify the anti-phage activity of the array (Vale and Little 2010).

As mentioned in section 1.3.1, phages can either be mainly lytic or lysogenic. The regulation of the switch between the two lifecycles can be targeted to prevent the lytic cycle from taking hold when hosts are infected by temperate phages. In the case of phage lambda, expression of the CI protein blocks transcription of genes which lead to lytic development. If a CI-like regulator can be identified on a phage genome, overexpression of this CI-like protein from a high-copy number plasmid inside host cells can prevent the induction of the lytic cycle and thereby the spread of phage. This system has

been effectively employed to prevent proliferation of phage Sfi21 (Bruttin *et al*., 1997). As the protein may be highly specific for its cognate promoter sequence(s) it may not be effective at stopping even closely related phages from entering the lytic cycle. This was the case for Sfi21, which effectively stopped lytic development of the phage it was directed against but failed to protect against 30 other phages tested.

Blocking phage DNA injection is another way to prevent the spread of phage. If the cell surface feature responsible for phage attachment has been identified, this can be targeted for disruption if a genetic system is available for the host, or through chemical mutagenesis and rounds of selection by challenging the mutated host with the phage. As described in 1.3.2, phages also protect their host against further infection by related phages through superinfection exclusion by expressing proteins that block phage DNA injection. Orf203 of *S. thermophilus* phage Sfi21 is thought to encode such a protein, and its overexpression in the host, reduced the efficiency of plaquing of 12 of 21 virulent phages (Bruttin *et al*., 1997). Oddly, the expression of just this protein did not protect against further infection by Sfi21 itself, as its lysogen does, but against closely related phages.

The above systems act in a manner similar to abortive infection mechanisms, by allowing the phage to infect, but blocking the progression of its life cycle and production of progeny.

Some phages, including T4 and lambda, produce an antiholin protein. In the case of phage lambda, this protein is produced from the same open reading frame as the holin (S105), except that it has two more amino acid residues at the N-terminal (S107). The expression of this longer version of the holin, inhibits cell lysis by interacting with S105 (Bläsi and Young 1996). Overexpression of these proteins could be explored as a possible phage resistance mechanism.

1.4 Aims and Objectives

As there currently exists a need for the development of systems capable of the conversion of (ligno)cellulosic biomass, or other waste streams to biofuels, and given the advantages outlined for thermophilic systems, the development of *P. thermoglucosidans* as platform for this conversion seems a suitable solution. Prior to this study, the organism had already been engineered to produce ethanol from glucose at near theoretical yields. Here we wanted to demonstrate that an alternative, and historically successful method for homo-ethanol engineering in microorganisms (PDC pathway), can operate in *P. thermoglucosidans* to yield equivalent results, perhaps with improved growth kinetics. This would require the expression of either a known, characterized PDC in *P. thermoglucosidans*, or the discovery, characterization, and expression of novel PDCs in the organism. This would allow us to determine if the pathway offers any metabolic advantage over those already in operation in the existing ethanologenic strains.

The recent discovery of a bacteriophage capable of infecting the organism, especially during large scale fermentation, highlighted the need to develop strains resistant to the virus to enable its successful commercial application. The description of the new virus, together with resistance engineering would afford us the opportunity to gain insight into a unique thermophilic phage and its interaction with its host and compare it to those that are currently known.

The main aim was therefore to express PDC in *P. thermoglucosidans* to determine if the pathway offers a better solution to those that already exist and to engineer the organism to be resistant to the only virus currently known to infect it thereby creating a robust bioethanol producing platform organism.

The specific aims were:

1. Identify and characterize novel bacterial pyruvate decarboxylase enzymes for expression in *P. thermoglucosidans*

2. Express pyruvate decarboxylase enzymes in *P. thermoglucosidans* and characterize ethanol production

3. Characterize a novel *P. thermoglucosidans* infecting phage

4. Develop phage resistant strains of *P. thermoglucosidans*

Chapter 2

Author contributions

Don Cowan, Marla Tuffin, Kirsten Eley and Mark Taylor conceived the study of *Gluconobacter oxydans* pyruvate decarboxylase and its expression in *Parageobacillus thermoglucosidans*. Leonardo Joaquim van Zyl conceived the codon harmonization part of the study. Leonardo Joaquim van Zyl performed all experiments and analysis described in the manuscript. Leonardo Joaquim van Zyl wrote the bulk of the manuscript. All authors read and approved the final manuscript.

Appl Microbiol Biotechnol (2014) 98:1247–1259
DOI 10.1007/s00253-013-5380-1

Engineering pyruvate decarboxylase-mediated ethanol production in the thermophilic host *Geobacillus thermoglucosidasius*

L. J. Van Zyl & M. P. Taylor & K. Eley & M. Tuffin & D. A. Cowan

Received: 12 September 2013 / Revised: 30 October 2013 / Accepted: 2 November 2013 / Published online: 26 November 2013

Abstract This study reports the expression, purification, and kinetic characterization of a pyruvate decarboxylase (PDC) from *Gluconobacter oxydans* . Kinetic analyses showed the enzyme to have high affinity for pyruvate (120 μM at pH 5), high catalytic efficiency ($4.75\times10^5\ M^{-1}\ s^{-1}$ at pH 5), a pH_{opt} of approximately 4.5 and an in vitro temperature optimum at approximately 55 °C. Due to in vitro thermostablity (approximately 40 % enzyme activity retained after 30 min at 65 °C), this PDC was considered to be a suitable candidate for heterologous expression in the thermophile *Geobacillus thermoglucosidasius* for ethanol production. Initial studies using a variety of methods failed to detect activity at any growth temperature (45–55 °C). However, the application of codon harmonization (i.e., mimicry of the heterogeneous host's transcription and translational rhythm) yielded a protein that was fully functional in the thermophilic strain at 45 °C (as determined by enzyme activity, Western blot, mRNA detection, and ethanol productivity). Here, we describe the first successful expression of PDC in a true thermophile. Yields as high as 0.35±0.04 g/g ethanol per gram of glucose consumed were detected, highly competitive to those reported in ethanologenic thermophilic mutants. Although activities could not be detected at temperatures approaching the growth optimum for the strain, this study highlights the possibility that previously unsuccessful expression of *pdc* s in *Geobacillus* spp. may be the result of ineffective transcription/translation coupling.

Keywords Pyruvate decarboxylase · Bioethanol · *Gluconobacter* spp. · Thermophilic expression

L. J. Van Zyl · M. P. Taylor · M. Tuffin · D. A. Cowan
Institute for Microbial Biotechnology and Metagenomics (IMBM), University of the Western Cape, Modderdam Road, Bellville, Cape Town, South Africa

M. P. Taylor · K. Eley
TMO Renewables Limited, 40 Alan Turing Road, The Surrey Research Park, Guildford, Surrey GU2 7YF, UK

D. A. Cowan
Centre for Microbial Ecology and Genomics, Department of Genetics, University of Pretoria, Pretoria 0028, South Africa

D. A. Cowan (✉)
Institutional Research Theme in Genomics, University of Pretoria, Pretoria 0028, South Africa
e-mail: don.cowan@up.ac.za

Introduction

Pyruvate decarboxylase (PDC, EC 4.1.1.1) is responsible for the non-oxidative decarboxylation of pyruvate to acetaldehyde and carbon dioxide. PDCs are common in the plant and fungal kingdoms and at least in the latter, together with alcohol dehydrogenase (ADH, EC 1.1.1.1) form part of an ethanol fermentation pathway (Konig et al. 1998). Several plant and yeast PDCs have been isolated and characterized, but as yet only four of bacterial origin has been described—from *Zymomonas mobilis*, *Zymobacter palmae*, *Acetobacter pasteurianus*, and *Sarcina ventriculi* (Raj et al. 2002). Bacterial PDCs participate in ethanol production via the Entner–Doudoroff pathway and not through glycolysis for pyruvate production.

There has been an increased interest in the use of thermophiles, such as *Geobacillus thermoglucosidasius* , for ethanol production, primarily because of their catabolic promiscuity, an important benefit for a second-generation bioprocess design (Cripps et al. 2009; Taylor et al. 2009). Other advantages include improved product removal, reduced incidence of contamination, and high ethanol yields in selectively mutated strains (Taylor et al. 2009). Ethanol production in *G. thermoglucosidasius* and mutants with enhanced ethanologenic phenotypes relies on endogenous metabolic pathways,

generating acetyl CoA via pyruvate dehydrogenase and its subsequent conversion to acetaldehyde and ethanol by aldehyde dehydrogenase and alcohol dehydrogenase, respectively (Cripps et al. 2009). An alternative to further develop *G. thermoglucosidasius* as an ethanologenic strain is to engineer the expression of the PDC pathway. The in vitro high specificity and thermostability (half-life of 30 min at 60 °C) of the *Z. mobilis* PDC (ZmoPDC) has made it the main candidate for such engineering (Pohl et al. 1995). However, both the ZmoPDC and *Z. palmae* PDC (ZpaPDC) have been expressed in *G. thermoglucosidasius* but do not function at temperatures exceeding 55 °C, despite good in vitro thermostability at these temperatures. The reasons for the low levels of activity are not fully understood (Taylor et al. 2008; Thompson et al. 2008), but protein misfolding resulting in inactive protein has been proposed (Thompson et al. 2008). Attempts to express these proteins in mesophilic Gram-positive hosts (notably lactic acid bacteria and *Bacillus megatarium*) have also had limited success (Gold et al. 1996; Bongers et al. 2005; Kaczowka et al. 2005; Talarico et al. 2005; Liu et al. 2005, 2006, 2007; Orencio-Trejo et al. 2008; Bi et al. 2009).

The role that codon usage plays in heterologous protein expression has been recognized, but is not well understood (Gustafsson et al. 2004). It has been demonstrated that the position and usage frequency of codons, together, play a role in correct protein folding and that "codon harmonization" could be used to overcome poor expression, at least in *Escherichia coli* (Angov et al. 2008; Rosano et al. 2009). Incompatibilities in codon usage and their effect on expression of PDCs have been reported (Talarico et al. 2001, 2005; Lowe et al. 1992). Two examples of the effect of codon usage on PDC production include the five- to tenfold increase in soluble *S. ventriculi* PDC (SvePDC) when expressed in an *E. coli* strain with or without accessory tRNAs (specifically those which are rarely used in *E. coli*), as well as the superior production of this PDC relative to those from *A. pasteurianus* and *Z. mobilis* in *B. megatarium* (Raj et al. 2002; Talarico et al. 2001, 2005).

Despite the rarity of prokaryotic PDCs, we have identified a *pdc*-like gene sequence in the genome sequence of a Gram-negative acetic acid bacterium, *Gluconobacter oxydans* . A PDC enzyme from *G. oxydans* has previously been characterized (King et al. 1954). *G. oxydans* is often associated with sugar-rich environments such as ripe fruit, honey, and cider as well as in a variety of soil types and is used industrially to produce L-sorbose from D-sorbitol, D-gluconic acid and 5-keto- and 2-ketogluconic acids from D-glucose, and dihydroxyacetone from glycerol (Gupta et al. 2001). This organism uses the PDC as part of the well-characterized lactate oxidation and acetate excretion pathways (Raj et al. 2001; Peters et al. 2012).

The aim of this study was to evaluate the ethanologenic potential of the *G. oxydans* PDC (GoxPDC) expression in *G. thermoglucosidasius*. We report the cloning, expression, and characterization of GoxPDC as well as its "codon harmonization" for enhanced expression in this Gram-positive thermophilic host.

Materials and methods

Media, bacterial strains, and plasmids

Bacterial strains and plasmids used in this study are shown in Table 1. *E. coli* strains were grown in Luria–Bertani (LB) broth (Sambrook et al. 1989) with 200 μg/ml ampicillin or 50 μg/ml kanamycin added as required. *G. thermoglucosidasius* strains were cultured either in LB, 2TY, TGP media, or modified urea sulphates medium (USM). In general, *E. coli* DH5α was used for plasmid construction.

One liter of TGP broth contained 17 g tryptone, 3 g soy peptone, 2.5 g K_2HPO_4, and 5 g NaCl. The pH was adjusted to 7.3 before autoclaving, after which 4 g Na-pyruvate and 4 mL glycerol were added in the form of filter-sterilized 10× concentrates. For solid media, 15 g/L agar was added before autoclaving. LB and TGP were used during genetic manipulation and general maintenance of cultures.

Per liter, 2TY medium contained 10 g yeast extract, 5 g NaCl, 20 g tryptone, and 15 g agar (where applicable), with a final pH of 7.0. USM supplemented with yeast extract (USMYE) contained 10 g glucose, 0.42 g citric acid, 0.31 g $MgSO_4$, 3.1 g NaH_2PO_4, 3.5 g K_2SO_4, 3 g urea, 2.2 mg $CaCl_2$, 0.4 mg Na_2MoO_4, 1 g yeast extract, 1 g typtone, 8.36 g Bis-Tris, 12.08 g PIPES, 10.4 g HEPES, 1 ml silicone antifoam, and 5 ml trace elements solution per liter. The trace element solution contained (per liter) 1.44 g $ZnSO_4.7H_2O$, 0.56 g $CoSO_4.6H_2O$, 0.25 g $CuSO_4.5H_2O$, 5.56 g $FeSO_4.6H_2O$, 0.89 g $NiSO_4.6H_2O$; 1.69 g $MnSO_4$, and 5.0 ml 12 M H_2SO_4. The trace elements solution and glucose (50 ml of a 20 % *w* /*v* solution) were added aseptically after autoclaving. The pH of USM was adjusted to pH of 7 using 10 M NaOH. Cultures were incubated at 45, 52, or 60 °C as required with vigorous aeration.

G. oxydans was cultured in medium containing, per liter: 8 g yeast extract, 15 g peptone, 10 g glucose, 0.5 % (*w* /*v*) ethanol, and 0.3 % (*w* /*v*) acetic acid. The final pH was between 3.5 and 4. Ethanol, acetic acid, and glucose were added after autoclaving. Cultures were incubated at 25 °C.

DNA manipulations and sequencing

Restriction endonuclease digestion, gel electrophoresis, and ligation performed using standard methods or following the manufacturers' recommendations (Sambrook et al. 1989). Ultrapure plasmid DNA was obtained using the Wizard Plus SV miniprep DNA purification system (Promega™). Total

Table 1 Bacterial strains, plasmids, and primers used in this study. Underlined sections in primer sequences indicate restriction endonuclease sites

Strain or plasmid	Genotype or description	Source or reference
Strains		
G. thermoglucosidasius TM89	*ldh*A⁻ variant of *G. thermoglucosidasius* NCIMB 11955	TMO renewables
G. oxydans (DSMZ7145)	Isolated from beer	DSMZ
E. coli DH5α	F′/*endA1 hsdR17* ($r_K^- m_K^+$) *supE44 thi -1 reacA1 gyrA* (Nal^s) *relA1* Δ(*lacZYA -argF*)*U169* (ϕ*80dlac*Δ(*lacZ*)*M15*)	Promega Corp.
E. coli BL21-DE3	*E. coli* B F⁻ *dcm ompT hsd*S($r_B^- m_B^-$) *gal* phage Lambda(DE3)	Invitrogen Corp.
E. coli JM109	F′ *traD36 proA*$^+$*B*$^+$ *lacIq* Δ(*lacZ*)*M15*/Δ(*lac-proAB*) *glnV44 e14*⁻ *gyrA96 recA1 relA1 endA1 thi hsdR17*	New England Biolabs, Beverley, MA, USA
Plasmids		
pET28a	Kan^r; ColE1 replicon, HIS-tag expression vector	Novagen Corp.
pGO	Kan^r; ColE1 replicon; *G. oxydans pdc* gene cloned into pET28a	This study
pldhGO	Kan^r; ColE1 replicon; lactate dehydrogenase (P*ldh*) gene promoter region (±170 bp *Nco* I-*Nde*I) from *G. thermoglucosidasius* NCA1503 cloned upstream of the *G. oxydans pdc* gene in pET28a	This study
pTM049	Derivative of pUB190 containing the *ldh* promoter from *G. stearothermophilus* NCA1503.	TMO Renewables (Cripps 2009)
pTMO111	Amp^r, Kan^r (in *G. thermoglucosidasius*) ColE1 replicon, pUB110 IncA replicon, *E. coli - G. thermoglucosidasius* shuttle-suicide (>55 °C) vector containing a truncated *pflB* gene	TMO Renewables (Cripps 2009)
pGO111 ($GoxPDC_{WT}$)	Amp^r, Kan^r (in *G. thermoglucosidasius*), 3603 bp *Dra*III-*Eco*RV fragment, blunted at the *Dra*III end, from pldhGO cloned into the unique *Swa*I site of pTMO111	This study
pGOF111 ($GoxPDC_{OPT}$)	Amp^r, Kan^r (in *G. thermoglucosidasius*), 1,887 bp *Not*I-*Not*I fragment, containing the fully codon optimized *G. oxydans* PDC with the P*ldh* upstream, cloned into the unique *Not*I site of pTMO111	This study
Primers		
LDHF	5′-TATACCATGGGCGGGACGGGGAGCTGAGTGCTC-3′	Cripps (2009)
LDHR	5′-GCCGCATATGATTCATCCTCCCTCAATAT-3′	Cripps (2009)
GOPDCpETF	5′-GGAATTCCATATGACTTATACTGTCGG-3′	This study
GOPDCpETR	5′-CCGCTCGAGTCAGACGCTCTGCGG-3′	This study

DNA from all bacterial strains was prepared as described (Kotze et al. 2006). The QIAGEN plasmid midi kit was used for large-scale plasmid preparations. DNA was sequenced using an ABI Prism 377 automated DNA sequencer and sequences were analyzed with DNAMAN (version 4.1, LynnonBioSoft). Codon usage in *G. thermoglucosidasius* NCIMB 11955 (with particular reference to the PDC genes from *G. oxydans*) were analyzed using the web servers http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=1426&aa=1&style=N and http://gcua.schoedl.de/.

Polymerase chain reaction

Polymerase chain reaction (PCR) was performed using Phusion DNA polymerase (New England Biolabs™). Generally, 50 ng DNA were used in a 50 μl reaction volume containing 2 mM $MgCl_2$, 0.125 μM of each primer, 0.2 mM of each deoxynucleoside triphosphate, and 1 U DNA polymerase. Reactions were carried out in a Hybaid Sprint thermocycler, with an initial denaturation at 94 °C for 60 s, followed by 30 cycles of denaturation (30 s at 94 °C), annealing (30 s), and variable elongation (72 °C), where annealing temperatures and elongation times were adjusted as required. Primers are listed in Table 1.

Cloning of the *G. oxydans pdc* gene

The *pdc* gene was amplified using genomic DNA isolated from *G. oxydans* DSMZ7145 using primers GoxPDCpETF and GoxPDCpETR (Table 1). The gene encoding the *G. oxydans* pyruvate decarboxylase was cloned into the pET28a expression vector in two parts. The 5′ 913 bp fragment was cloned by digesting both PCR product and pET28a with *Nde*I and *Xho* I. The 779 bp 3′-fragment was first cloned into pBluescriptSK by digesting the GoxPDC PCR product with *Xho* I, and then excised by *Xho* I digestion and cloned into *Xho* I-digested pET28–GoxPDC-5′. Clones with the correct orientation of this 779 bp fragment were identified by restriction enzyme digest of the final construct (pGO) using *Sph*I, and confirmed by sequencing. The nucleotide sequences of the wild type and codon-optimized *pdc* genes have been submitted to the EMBL-GenBank database and are available

under accession numbers KF650838 and KF650839, respectively.

Purification of PDC protein

An overnight culture of pGO (Table 1) in *E. coli* BL21-DE3 with kanamycin was used to inoculate fresh LB broth (1/100 ml) and then incubated with aeration (120 rpm) overnight at room temperature to express the protein without IPTG induction. The cells were collected by centrifugation (3,214×*g* for 10 min). BugBuster™ was used to lyse cells (3 ml/g of wet cells) and the suspension incubated at room temperature for 20 min with shaking. After centrifugation to remove cell debris (12,857×*g* for 20 min), DNaseI and RNaseA (Fermentas) were added (10 U/ml) to the lysate to reduce the viscosity and incubated at room temperature with shaking for 30 min. The HisBind™ resin and buffer kit (Novagen) was used to purify the protein. After elution with 9 ml imidazole buffer (100 mM), the protein was dialysed against 200 volumes of buffer (50 mM MES pH 6.4) containing 1 mM TPP and 1 mM $MgCl_2$. Purity was estimated by reducing sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) gel (12 %) and the protein concentrations determined using Bradford reagent (Bio-Rad) with bovine serum albumin as the standard (Laemmli et al. 1970).

Steady state kinetics and substrate specificity

GoxPDC activity was measured using a coupled assay with baker's yeast ADH (Sigma-Aldrich) as described (Conway et al. 1987), and represents a different assay to that used in the initial *G. oxydans* PDC characterization in the King and Cheldelin study (King et al. 1954). The enzyme was purified independently four times and kinetic measurements performed with each preparation to give the average ($n \geq 3$) of the results shown. The reaction mixture (1 ml final volume) contained 0.25 mM NADH, 5 mM $MgCl_2$, 0.1 mM TPP, 5 mM pyruvate (unless stated otherwise), and 10 U of ADH in 50 mM MES or 200 mM Na citrate pH 6.4 or 6.0, respectively. For substrate range determination, ADH was replaced with 1 U/ml baker's yeast aldehyde dehydrogenase (ALDH, Sigma-Aldrich). β-mercaptoethanol was added to a final concentration of 3 mM and NADH replaced with NAD^+. Assays were performed in 100 mM citric acid/K_2HPO_4 buffer pH7.0 (Vuralhan et al. 2005). Activities were recorded at 25 °C unless otherwise indicated using a Cary 50 temperature controlled spectrophotometer (Varian). In the case where aldehydes produced by the decarboxylation of certain substrates by PDC were not recognized as a substrate for ALDH, high-performance liquid chromatography (HPLC) analysis was used to determine activity on that substrate. Reactions were run on a Rezex RHM monosaccharide column (Phenomonex), using 5 mM H_2SO_4 as mobile phase under isocratic elution (0.6 ml/min, 40 °C) on a Dionex Ultimate 3000 machine. Samples (20 μl) were injected by autosampler and the components detected using either a refractive index detector or a UV/Vis photodiode array at 215 nm. For kinetic data, initial rates were measured over the substrate range of 0.1 to 30 mM for pyruvate or 0.1 to 50 mM for other 2-keto acids. Kinetic parameters were determined by nonlinear fitting of data to hyperbolic curves according to Michaelis–Menten (GraphPad Prism v. 4.00, GraphPad Software, San Diego, CA, USA). One unit of enzyme activity corresponds to the amount of enzyme that generates 1 μmol of acetaldehyde per minute. k_{cat} values were calculated based on the MW of the monomer with one active site.

Construction of the *pdc* expression for transformation in *G. thermoglucosidasius*

For expression of PDC in *G. thermoglucosidasius*, the 170-bp promoter region of the lactate dehydrogenase gene P*ldh* from *G. thermoglucosidasius* NCA1503 was cloned upstream of the Gox*pdc* gene. This promoter has been shown to be induced under microaerobic conditions in *G. thermoglucosidasius* (Cripps et al. 2009). The promoter region was amplified from pTMO49 using the LDHF and LDHR primers and cloned into pGEM-T Easy. Sequencing confirmed that no DNA base changes had occurred. The p*ldh* was cloned into pGO (Table 1) using the *Nco* I and *Nde* I sites such that the promoter was functional for *pdc* expression. This construct was digested with *Dra* III and the ends filled in using T4 DNA polymerase (Fermentas, ThermoFisher). Digestion with *Eco* RV yields a 3,603 bp *Dra* III (blunt)–*Eco* RV fragment. The plasmid TMO111 was digested with *Swa* I and treated with rAPid™ alkaline phosphatase (Roche) to prevent self-ligation. The *Dra* III (blunt)–*Eco* RV fragment and *Swa* I digested pTMO111, were ligated using T4 DNA ligase (Fermentas). Insertion at the *Swa* I site leaves 809 and 436 bp of the *pflB* gene on either side of the *ldh -pdc* for recombination with its chromosomal counterpart.

The constructs were passaged through *E. coli* JM109 for DNA methylation prior to transformation to prevent endonuclease degradation in *G. thermoglucosidasius*. *G. thermoglucosidasius* competent cells were prepared and transformed (Cripps et al. 2009).

Fermentative product profile quantification

Cultures expressing *pdc* gene were grown overnight at 37 °C for 16 h and 200 rpm in LB media (*E. coli*) and 2TY media (*G. thermoglucosidasius*) at 60 °C. A volume of 0.5 ml of this culture was transferred to 10 ml of USMYE media, contained separately, in 50 and 15 ml screw-cap universal tubes. This effectively generated 40 ml and 5 ml headspaces, respectively, mimicking aerobic and microaerobic or fermentative culture

conditions. These cultures were grown overnight at 37 (*E. coli*), 45, or 52 °C (*G. thermoglucosidasius*) for 16 h and 200 rpm and the supernatant removed by centrifugation (2,057×*g* for 10 min). Metabolite concentrations in culture supernatants were determined by HPLC (see above) and products were compared to suitable standards of known concentration and against the media in which the cultures were grown. Experiments were carried out in triplicate for *E. coli* and at least in duplicate for *G. thermoglucosidasius*.

Western blotting

Rabbit anti-GoxPDC polyclonal antibodies were made by Antibodies Incorporated (Davis, CA, USA) using His-tag purified GoxPDC protein. Cells were harvested directly after fermentation by centrifugation at 3,214×*g* for 10 min. The cell pellet was resuspended in MES buffer pH6.5 and sonicated using five pulses of 30 s each. Cell debris was removed by centrifugation (15,682×*g* for 20 min), and the supernatant decanted. Protein concentrations were determined by Bradford assay. Forty micrograms of total protein was loaded for each sample and run on a 12 % SDS-PAGE gel. Protein was transferred from the gel to Biotrace™PVDF membrane by semidry blotting. For signal detection, the anti-rabbit Super Signal West Femto Chemiluminescent Substrate (Pierce) kit was used and the signal visualized using a chemiluminescent camera.

Results

Amino acid sequence considerations in the *G. oxydans* PDC

The cloned and sequenced GoxPDC gene differed from the published genome sequence (NC_006677.1) (Prust et al. 2005) at 22 positions, resulting in 5 amino acid changes (Y163F, S207N, A209T, I469M, and D517E) but no frame shifts or deleterious events. The sequence alignment (Fig. 1) indicated that none of the affected residues have been shown to be directly involved in catalysis, substrate, or co-factor binding. Most changes are conservative (F163, N207, and E517) and/or are located distantly from the active site (F163, N207, T209, and E517). Met469 is, however, located in a region of the enzyme which may be sensitive to changes. It is positioned adjacent to Glu468, important in catalytic activity; Ile467, involved in substrate recognition; and Ile471, crucial for substrate positioning (Prust et al. 2005; Pohl et al. 1998; Siegert et al. 2005). However, equivalent residues in other PDCs have not been associated with catalysis or substrate recognition (Fig. 1). Phe163 (Fig. 1) is an interesting change, as it is occupied by a tyrosine in all the PDCs (including the *G. oxydans* NC_006677.1 PDC, except for SvePDC). GoxPDC, unlike ZpaPDC and ApaPDC, also contains an extra six amino acid loop from position 498-503 (EESGKY), which is similar to positions 503-508 in ZmoPDC (DSGAGK), but the residues are not conserved. As the encoding gene was amplified using a polymerase with 3′-5′ exonuclease activity and the differences were consistent in several independent clones, we infer that the changes represent natural variations in the GoxPDC and are not artifacts of the cloning procedure.

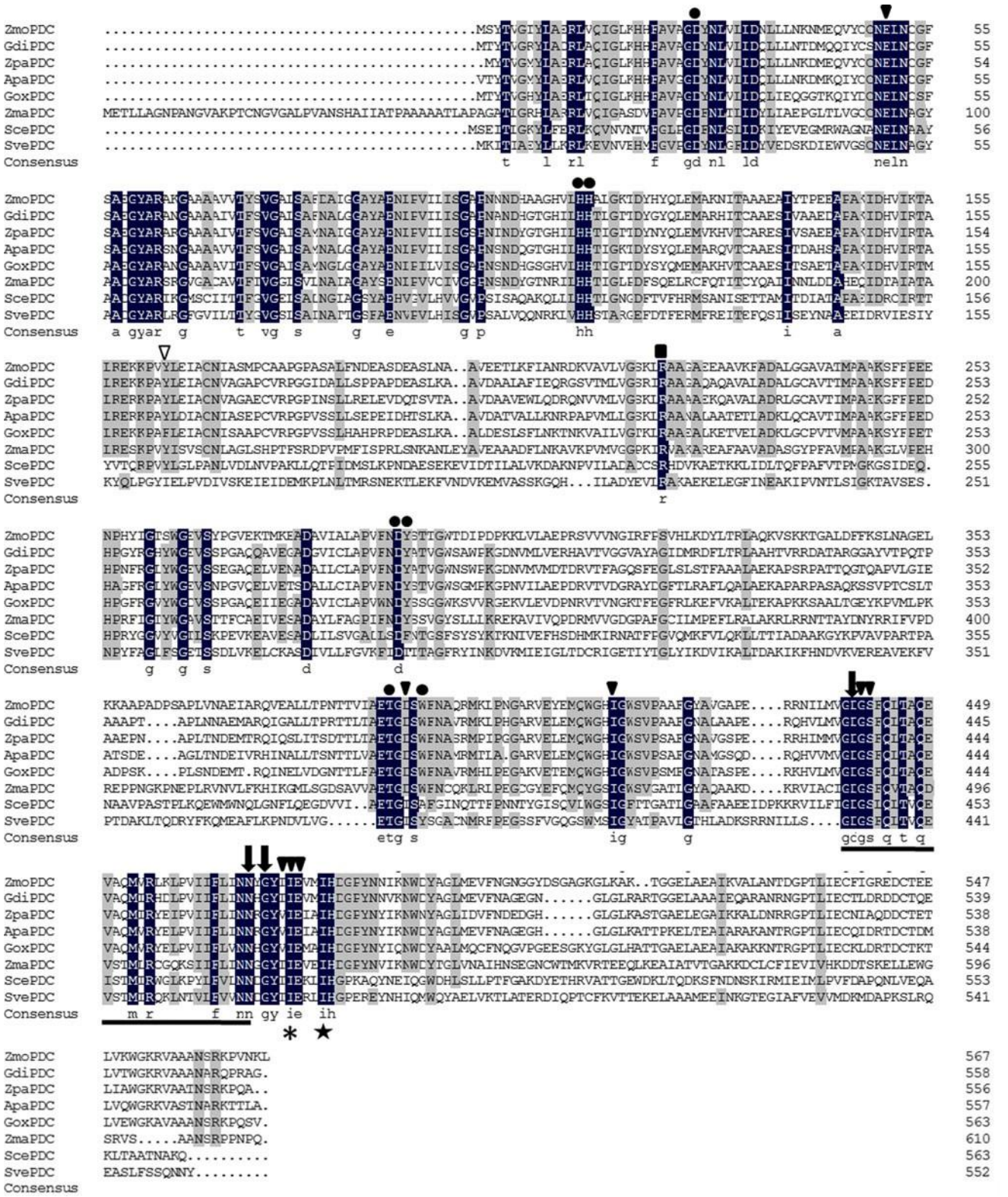
The protein sequence demonstrates all the features typical of ThDP-binding enzymes including the conserved ThDP binding motif GDGS-XXX-NN as well as several conserved residues required for substrate binding and catalysis (indicated in Fig. 1). It does, however, lack a cysteine residue equivalent to Cys221 in ScePDC, as shown for all the bacterial PDCs, thought to be involved in allosteric substrate activation (Lu et al. 2000).

Kinetic characterization of the GoxPDC enzyme

GoxPDC was purified to homogeneity by affinity chromatography, and the estimated molecular weight of the protein, at ±60 kDa (Fig. 2), corresponds well to the theoretical molecular mass of 60.8 kDa. The predicted pI value is 6.0.

Conventional enzyme characterization was performed using pyruvate as a substrate (kinetic data for GoxPDC are summarized in Table 2). The K_M value for pyruvate was found to be in the range of those determined for other PDCs from Gram-negative bacteria assayed under similar conditions. The enzyme also displayed a ±20-fold decrease in the K_M for pyruvate with a decrease in pH from 7 to 5, without an equivalent change in the catalytic rate (k_{cat} showed an approximate twofold decrease). This is in line with previous observations in related enzymes (Raj et al. 2002; Siegert et al. 2005) and supports the interpretation that PDCs require a protonated residue for efficient binding of the substrate, in which the ionizable group is thought to be the aminopyrimidine ring of the ThDP coenzyme (Meyer et al. 2010). The GoxPDC enzyme displayed normal Michaelis–Menten kinetics with pyruvate as the substrate and was not subject to allosteric substrate activation as has been reported for PDCs from plants, yeasts, and the SvePDC (Konig et al. 2009). The GoxPDC has a lower catalytic rate than ApaPDC but the catalytic efficiencies were similar to those reported for ZmoPDC and SvePDC.

The pH optimum of GoxPDC was determined to lie between 4.5 and 5.0, similar to ApaPDC and slightly lower than for other PDCs from Gram-negative bacteria (Fig. 3; Gocke et al. 2009). The temperature optimum of GoxPDC was between 50 and 55 °C. Thermal inactivation studies demonstrated that the enzyme was stable at the experimental T_{opt}, with no loss of activity after an hour of incubation. However, at temperatures ≥60 °C, moderate to rapid loss of activity was recorded, retaining 30 % of the initial activity at 65 °C after an hour of incubation (Fig. 4). These data demonstrate that GoxPDC thermostability is equivalent to the Gram-negative homologs.



The following substrates were shown by HPLC analysis to be decarboxylated by GoxPDC; 2-ketopropanoate (pyruvate), 2-ketobutanoate, 2-ketopentanoate, 2-keto-4-methylpentanoate, and 4-hydroxy-phenyl-pyruvate. Calculated specific activities were 32 U/mg for 2-keto-butanoate, 1.2 U/mg for2-keto-pentanoate, and 0.2 U/mg for 2-keto-4-methylpentanoate. No activity was recorded on 3-phenyl-2-oxopropanoate, benzoyl formate, 4-hydroxy-phenyl pyruvate, and indole-3-pyruvate.

Fig. 1 Multiple sequence alignment of the protein sequences from selected PDC proteins: *GdiPDC* , *G. diazotrophicus* (YP_001600462.1); *GoxPDC*, *G. oxydans*; *ZpaPDC*, *Z. palmae* (AF474145_1); ApaPDC, *A. pasteurianus* (AF368435_1); *ZmoPDC*, *Z. mobilis* (YP_163095.1); *ZmaPDC*, *Z. mays* (P28516.1); *ScePDC*, *S. cerevisiae* (EGA85775.1); *SvePDC*, *S. ventriculi* (AF354297_1). The alignment was generated using the "full alignment" feature in DNAman with default setting. Residues shaded in *black* are 100 % conserved while those in *grey* are 75 % conserved. The *underlined region* shows the conserved ThDP-binding motif and *triangles* indicate those residues which bind ThDP. *Arrows* indicate Mg^{2+} binding residues. *Circles* indicate residues which line the catalytic pocket and are thought to play a role in catalysis. *Asterisk* indicates the residue involved in substrate specificity. The *star* shows a residue thought to be needed for positioning of the substrate for catalysis. The *square* indicates the arginine residue involved in substrate activation of ScePDC and SvePDC. The *open triangle* shows the position of the unique phenylalanine residue in GoxPDC

Native *pdc* ($GoxPDC_{WT}$) expression in *E. coli* and *G. thermoglucosidasius*

The expression of $GoxPDC_{WT}$ in *E. coli* produced 0.5± 0.005 g/g ethanol per gram of glucose consumed under fermentative conditions, substantially higher than the control (DH5α-pTMO111) which produced only 0.1±0.01 g/g. Cell densities were of the same order of magnitude, demonstrating that the higher ethanol concentrations were not simply the result of higher biomass levels in GoxPDC-positive strains. It is noted that these elevated ethanol yields were achieved in the absence of a recombinant *adh* II, which has previously been shown to be essential for enhanced ethanol production in *E. coli* strains expressing ZmoPDC (Lawford et al. 1991). This constitutes the first report demonstrating ethanol production as a direct result of the functional expression of only a *pdc* gene in E. *coli* (Liu et al. 2005; Talarico et al. 2005; Lowe et al. 1992).

Expression in *G. thermoglucosidasius* under microaerobic conditions did not result in ethanol production more than the control strain (Fig. 5). RT-PCR confirmed that the gene was transcribed (data not shown); however, no soluble protein could be detected by Western blotting for cultures grown at 45 °C (Fig. 6) and no PDC activity was detected using cell free extracts from the same cultures. Together, these data suggest a failure at the level of translation, possibly with the generation of misfolded protein which would be targeted for intracellular proteolysis.

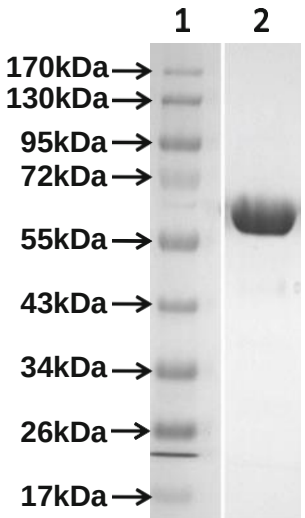


Fig. 2 A denaturing SDS-PAGE gel showing purified GoxPDC protein. *1* Molecular weight marker (#SM0671), *2* HIS-tag purified GoxPDC protein. The GoxDC protein was approximately to 59 kDa in size

Assessing codon usage and predicting a gene sequence for harmonization of GoxPDC expression in *G. thermoglucosidasius*

A relatively new concept, termed codon harmonization (Angov et al. 2008, 2011a, b), was proposed as a technique to assist protein folding during heterologous expression. Codon harmonization, as opposed to other codon optimization strategies, involves mimicking the translation rates of the native host in a heterologous strain. The translation rate is predominantly determined by codon usage frequencies, where the presence of infrequently used codons forces a reduction of the translation rate, allowing the protein being translated to fold in phases. Both the frequency and positioning of infrequently used codons is critical for protein folding (Clarke et al. 2008).

An analysis of the codon usage frequencies in GoxPDC for expression in *G. thermoglucosidasius* , *E. coli* , and *G. oxydans* was conducted (Table 3, Fig. 7). There are five codons which are rarely used in *G. oxydans*, the native host, one of which (CGA) i s also rarely used in *G. thermoglucosidasius*. However, in *G. thermoglucosidasius*, 20 codons in the GoxPDC gene (CCC and CTC) are recognized as rare, which are frequently used (>20 %) in the native host. It is thought that the inclusion of so many rare codons when expressing the gene in *G. thermoglucosidasius*, is detrimental to correct folding of the protein (Kane et al. 1995; Kim et al. 2006; Rosano et al. 2009). Similarly, in *E. coli*, the CTC and CGA codons are also rare. However, unlike in *G. thermoglucosidasius* , the CCC codon is infrequently used (Table 3). Based on the rare codon analysis, *E. coli* is expected to be a more suitable expression host for the WT GoxPDC relative to *G. thermoglucosidasius*. This was the case experimentally; however, the temperature had to be reduced (25 °C) to enable soluble expression of the GoxPDC protein in *E. coli*.

In order to codon harmonize the GoxPDC gene for *G. thermoglucosidasius* expression, 416 bp changes (348 codon substitutions) were made to the wild-type gene sequence so as to match the usage frequencies found in the native host for every codon position while maintaining the amino acid composition of the wild-type protein. Figure 7 shows a 20 amino acid section of the protein to demonstrate how the harmonization was performed.

This harmonization strategy was chosen to demonstrate whether the naturally evolved translation frequency alone would enable correct folding of the protein in *G. thermoglucosidasius*,

Table 2 Steady state kinetic constants for GoxPDC compared with those from other Gram negative bacteria. Values given in brackets indicate the pH at which measurements were made

PDC	K_M (mM)	Specific activity in (U/mg)	k_{cat} (s^{-1})	k_{cat}/K_M ($M^{-1}\,s^{-1}$)	T_{opt} (°C)	pH_{opt}	$T_{½}$ at °C
$GoxPDC_R$	0.12 (5.0)±0.005 1.2 (6.5)±0.2 2.8 (7.0)±0.4	57 (5.0) 47 (6.5) 125 (7.0)	57 (5.0) 47 (6.5) 125 (7.0)	4.75×10^5 (5.0) 3.6×10^4 (6.5) 4.2×10^4 (7.0)	53	4.5–5.0	10 min at 65 °C
$GoxPDC_N$	0.74 (6.0)	4.4 (6.0)[g]	N/A	N/A	N/A	6.0	[f]
ApaPDC	2.8 (6.5)[a]/0.39 (5.0)[e]	110 (6.5)[a]/97(5.0)[e]	341–508[e]	N/A	65[a]	3.5–6.5[a]	24 min at 70 °C[a]
ZpaPDC	2.5 (6.5)[a]/0.24 (6.0)[e]	116 (6.5)[a]/130 (6.0)[e]	341–508[e]	N/A	55[a]	7.0[a]	24 min at 60 °C[a]
ZmoPDC	1.3 (6.5)[a]/0.31 (6.0)[b]/1.1[c]	120 (6.5)[a]/120[c]	150 (6.0)[b]/486 (6.5)[c]	4.8×10^5 (6.0)[b]/4.4×10^5 (6.5)[c]	60[a]	6.0–6.5[a]	30 min at 60°C[a]
SvePDC	13[d]	103[d]	412[d]	3.2×10^{4d}/0.87×10^4 (7.0)	N/A	6.3–6.7[d]	30 min at 50 °C

For all experiments done on $GoxPDC_R$, $n \geq 3$. Values in brackets indicate pH

$_N$ native, $_R$ recombinant, *f* stable for 5 min at 55 °C but completely inactivated by incubating for 3 min at 80 °C, *g* crude extract

[a] Gocke et al. (2009)

[b] Meyer et al. (2010)

[c] Siegert et al. (2005)

[d] Lowe et al. (1992)

[e] Raj et al. (2002)

without the need to calculate link/end segments (Angov et al. 2008; Thanaraj et al. 1996).

Expression of $GoxPDC_{OPT}$ in *G. thermoglucosidasius*

Comparative expression of the wild-type ($GoxPDC_{WT}$) and codon harmonized ($GoxPDC_{OPT}$) GoxPDC in *G. thermoglucosidasius* TM89 was evaluated at 45 and 52 °C (Fig. 5). Expression of $GoxPDC_{OPT}$ at 45 °C produced 0.35±0.04 ethanol per gram of glucose consumed, compared to 0.26±0.04 g/g for TM89 alone or 0.24±0.02 g/g for $GoxPDC_{WT}$ (Fig. 5). This result clearly demonstrates that translational discord was, at least in part, a significant limitation to the functional expression of GoxPDC in *G. thermoglucosidasius*. However, at 52 °C strains expressing either $GoxPDC_{WT}$ (0.26±0.01 g/g) or $GoxPDC_{OPT}$ (0.25±0.04 g/g) produced lower ethanol yields than the control TM89 (0.32±0.05 g/g; Fig. 5). We speculate that the reduced performance of TM89 when

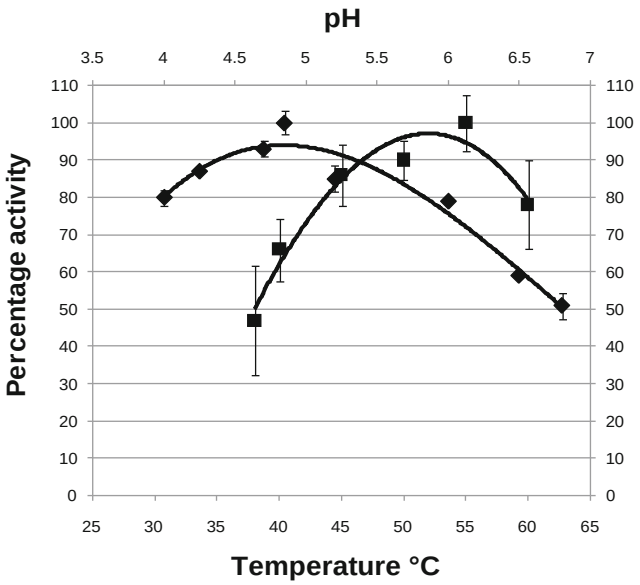


Fig. 3 Effect of pH and temperature (in degree Celsius) on the activity of GoxPDC (*black diamond*) when using pyruvate as substrate. For all data points, $n \geq 3$. The assay buffer used was 100 mM Na_2HPO_4/citrate buffer. The 100 % activity was analogous to a specific activity of 72 U/mg for T_{opt} and 160 U/mg for pH_{opt}

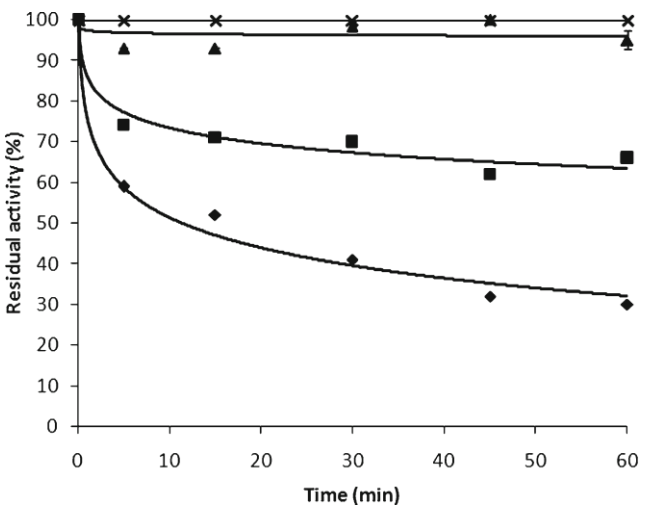


Fig. 4 Thermal inactivation profile of GoxPDC at 25 °C (*multiplication symbol*), 55 °C (*black triangle*), 60 °C (*black square*), and 65 °C (*in black diamond*). Activity is expressed as a percentage of that at time zero in the standard assay at 25 °C. The activity at 100 % correlates to a specific activity of 26 U/mg. Assays were performed in 50 mM MES buffer at pH6.5

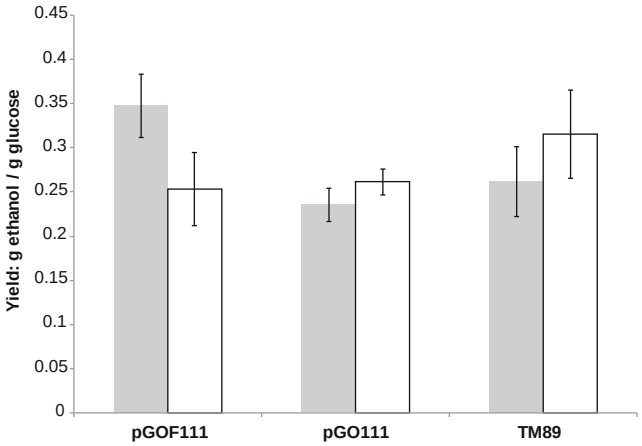


Fig. 5 Comparison of the yield of ethanol produced per gram of glucose consumed during 10/15 model fermentations in *G. thermoglucosidasius*. *Grey bars* represent fermentations performed at 45 °C and *white bars* 52 °C. The data represents an average of three independent fermentations (batches of medium) with three repeats of each culture in each fermentations after 48H at the given temperature

expressing either the unmodified or codon-harmonized GoxPDC at 52 °C may be due to the metabolic load imposed by the presence of the shuttle vector and/or the metabolic load imposed by misfolded PDC protein (proteotoxic stress).

Intracellular PDC activity was assayed from GoxPDC$_{OPT}$–*G. thermoglucosidasius* cultures grown at 45 °C and 52 °C. A specific activity of 0.22 U/mg was determined for GoxPDC$_{OPT}$ cultures grown at 45 °C, consistent with the observed phenotype. No detectable activity could be demonstrated for the control and GoxPDC$_{WT}$ cultures under similar conditions. No PDC activity was detectable in GoxPDC$_{OPT}$–*G. thermoglucosidasius* cultures grown at 52 °C. Furthermore, through Western blot analysis, a GoxPDC signal was only detectable in soluble protein extracts from *G. thermoglucosidasius* cultures expressing GoxPDC$_{OPT}$ at 45 °C (Fig. 6). Extracts from cells grown at 52 °C (GoxPDC$_{WT}$, GoxPDC$_{OPT}$, and control cultures) were reproducibly negative. Given that in vitro thermostability does not necessarily equate to in vivo stability, the lack of improved ethanol production at 52 °C is not unexpected. These results, taken together, indicate that although codon harmonization enhanced the expression of GoxPDC in *G. thermoglucosidasius* TM89 at 45 °C, a second limitation, probably thermally related, prevents functional enzyme accumulation at higher temperatures. The lack of detectable protein by Western blotting either indicates that a second limitation is present at the transcription–translation interface or that the protein produced is extremely unstable and the resultant aggregates or proteolysis products are not suitable for antibody binding.

Nevertheless, the increased ethanol yield at 45 °C represents a significant improvement in comparison to other engineered thermophiles (Cripps et al. 2009; Shaw et al. 2008). It has been demonstrated that effective partitioning of carbon into biosynthesis and fermentation is critical in achieving optimal production of ethanol under fermentative conditions (Underwood et al. 2002b). Typically, *G. thermoglucosidasius* TM89 fermentations are characterized by the production of formate (average of 40.5±7.2 mM after 48 h at 45 °C) and acetate with ethanol (Cripps et al. 2009). However, it was noted that for TM89-GoxPDC$_{OPT}$ fermentations at 45 °C, no formate was produced but low level (±185 μM) fumarate accumulation was detected in culture supernatants. This may be due to a reduced metabolic flux through *pfl* where the active GoxPDC$_{OPT}$ may outcompete the *pfl* enzyme for pyruvate (Orencio-Trejo et al. 2008; Tolan et al. 1987; Feldmann et al. 1989; Underwood et al. 2002a).

Discussion

The pyruvate decarboxylase from the acetic acid bacterium *G. oxydans* has been described in this study. The substrate recognition and decarboxylation range of the enzyme is similar to that of the other four Gram-negative PDCs identified to date, showing a preference for short-chain aliphatic 2-keto acids (Gocke et al. 2009). The value of k_{cat}/K_M for pyruvate compared to those for 2-ketobutanoate and 2-ketopentanoate, the nearest analogs, and the retention of Ile468, thought to be involved in substrate specificity, suggests that this enzyme favors pyruvate as its physiological substrate (Pohl et al. 1998; Gocke et al. 2009). The GoxPDC kinetics are also similar to the other Gram-negative bacterial PDCs, displaying

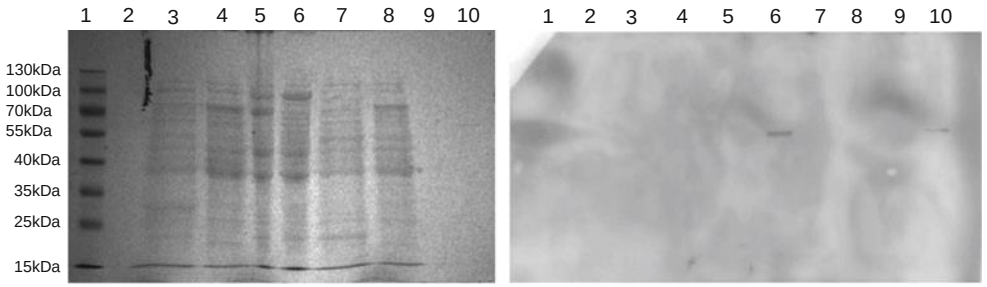


Fig. 6 SDS-PAGE and Western blots of cell extracts from *G. thermoglucosidasius* TM89, containing either empty vector pTMO111 (*5* and *8*) or pGO111 (GoxPDC$_{WT}$; *4* and *7*), or pGOF111 (GoxPDC$_{OPT}$; *3* and *6*). Cultures were grown at either 45 °C (*6*, *7*, and *8*) or 52 °C (*3*, *4*, and *5*), respectively. Purified GoxPDC protein served as positive control (*10*). Lanes *2* and *9* are empty

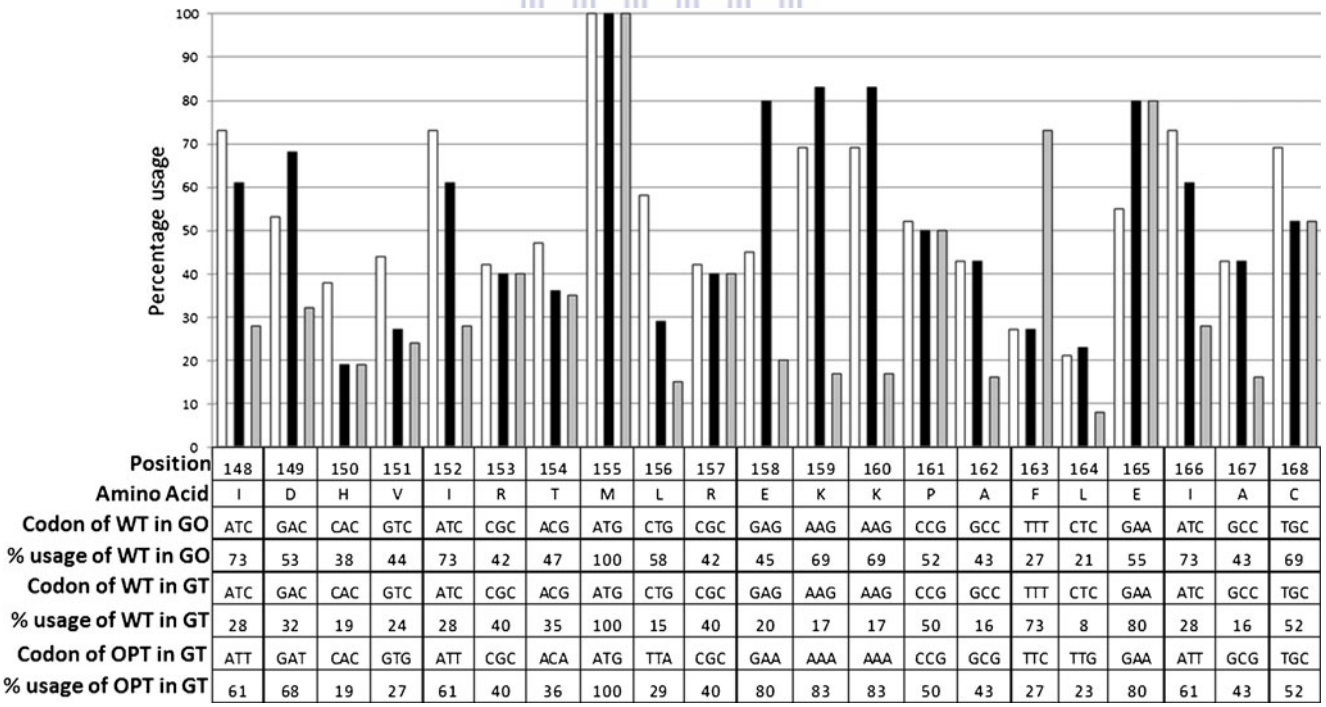
Table 3 The correspondence of rare codons (<10 % usage) for PDCs which have been expressed in *G. thermoglucosidasius* and in *E. coli*, between their native host and *G. thermoglucosidasius*

	All codons with <10 % usage in the native host (number of codons in respective *pdc* genes)	All codons with <10 % usage in *G. thermoglucosidasius* (amino acid position)	All codons with <10 % usage in *E. coli* (amino acid position)	% codon usage for selected codons in their native host
ZmoPDC	CTC (11)	CTC (2, 18, 30, 56, 164, 174, 215, 236, 306, 320, 348, 362, 400, 436, 567, 568), CCC (174, 320), AGT (2, 56, 362)	CGG (12), CTC (18, 30, 164, 215, 236, 306, 348, 400, 436, 567, 568, 569)	CCC 10–20 %, AGT 10–20 %
ZpaPDC	CTA (4), TTG (3), ACA (3), CTC (3), GTG (2), CTT (2), GCG (2), TCA (2), CCC (1), AGT (1), GGA (1), TCG (1), AAG (1), CCA (1)	CTC (35, 187, 517), CCC (356), CTA (12, 155, 231, 509), AGT (193)	CTC (35, 187, 517), CTA (12, 155, 231, 509),	N/A
GoxPDC	ACT (2), CGA (1), TCT (1), TCA (1)	CGA (12), CTC (30, 33, 84, 95, 164, 206, 226, 305, 431, 505, 531, 536, 545), CCC (174, 239, 251, 255, 309, 348, 396)	CGA (12), CTC (30, 33, 84, 95, 164, 206, 226, 305, 431, 505, 531, 536, 545)	CTC >20 %, CCC >20 %

N/A not applicable

the same pH dependent increase in k_{cat}/K_M while catalytic efficiency (k_{cat}) remaining largely unchanged.

Despite being an elusive enzyme in the bacterial kingdom, PDCs are sought after enzymes for the generation of ethanologens, and the engineering of a PDC-expressing pathway in fermentative bacteria is now a well-established procedure to achieve increased ethanol yields from mesophilic organisms (Taylor et al. 2008, Thompson et al. 2008, Talarico et al. 2005, Bi et al. 2009, Tolan et al. 1987, Ingram et al. 1987, Correa et al. 2011). In this study, expression of GoxPDC in *E. coli* resulted in a fivefold ethanol production increase, with a final yield of 0.5 g/g, similar to the best results reported for other recombinant *E. coli* strains, and moreover, without the co-expression of a heterologous *adh* (Ohta et al.



Position	148	149	150	151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168
Amino Acid	I	D	H	V	I	R	T	M	L	R	E	K	K	P	A	F	L	E	I	A	C
Codon of WT in GO	ATC	GAC	CAC	GTC	ATC	CGC	ACG	ATG	CTG	CGC	GAG	AAG	AAG	CCG	GCC	TTT	CTC	GAA	ATC	GCC	TGC
% usage of WT in GO	73	53	38	44	73	42	47	100	58	42	45	69	69	52	43	27	21	55	73	43	69
Codon of WT in GT	ATC	GAC	CAC	GTC	ATC	CGC	ACG	ATG	CTG	CGC	GAG	AAG	AAG	CCG	GCC	TTT	CTC	GAA	ATC	GCC	TGC
% usage of WT in GT	28	32	19	24	28	40	35	100	15	40	20	17	17	50	16	73	8	80	28	16	52
Codon of OPT in GT	ATT	GAT	CAC	GTG	ATT	CGC	ACA	ATG	TTA	CGC	GAA	AAA	AAA	CCG	GCG	TTC	TTG	GAA	ATT	GCG	TGC
% usage of OPT in GT	61	68	19	27	61	40	36	100	29	40	80	83	83	50	43	27	23	80	61	43	52

Fig. 7 A comparison of codon usage frequencies of GoxPDC in the native host (*G. oxydans*—*white bars*), the heterologous host (*G. thermoglucosidasius*—*grey bars*), and after harmonization (*black bars*), for the amino acid residues 148–168. Disparities exist between the percent usage of codons in *G. oxydans* for the wild-type PDC and those used in *G. thermoglucosidasius* (large differences between the *green* and *red bars*). The rarely used codons in *G. thermoglucosidasius* (<10 %), such as amino acid 164, would represent likely loci where a hiatus would be reached in translation of the gene in this strain. This does not correlate with a locus of low codon usage in *G. oxydans*. Post harmonization, it can be seen that the *black* and *white bars* are closer to each other in terms of percentage codon usage

1991). In *G. thermoglucosidasius* expression of the codon, harmonized GoxPDC resulted in a significant improvement in ethanol production. This represents the first account of in vivo PDC-mediated ethanol production in a Gram positive organism, and in a thermophile. While expression of several PDCs has been demonstrated in *Geobacillus* and *Bacillus megaterium* , none of these studies could demonstrate increased cellular ethanol production (Taylor et al. 2008, Thompson et al. 2008, Talarico et al. 2001). Considering that *G. thermoglucosidasius* is currently employed in a commercially viable bioethanol technology, this finding could represent a significant advancement in the engineering of a thermophilic ethanologen. The upregulation of pyruvate dehydrogenase and subsequent conversion of acetyl-CoA to acetaldehyde is stoichiometrically equivalent to a PDC intervention, and is the engineering strategy employed for the bioethanol process strain *G. thermoglucosidasius* TM242 (Cripps et al. 2009). It would be of value therefore for future studies to compare the performance of PDC- vs PDH-engineered *G. thermoglucosidasius* in a *pfl* -negative background, in order to maximize the flux of pyruvate to ethanol. However, despite the significance of these findings, PDC-mediated ethanol production in this study was only possible at a maximum temperature of 48 °C, and therefore the lack of PDC thermostability continues to limit the viability of the PDC route for a thermophilic commercial process.

It has long been known that codon usage differences during heterologous protein expression can result in low expression or formation of insoluble aggregates. The tendency has been to replace the rare codons in the protein of interest for codons used more frequently by the expression host. However, only recently has the suggestion been made that the frequency and positioning of infrequently used codons is critical for protein folding, and that the standard codon optimization approach is flawed. Instead, the principal of codon harmonization (Angov et al. 2008) involving the substitution of synonymous codons from the heterologous host such that the codon usage frequency, positioning, and therefore rhythm of translation follows that of the native host, was recently proposed. Codon harmonization has been exclusively applied to expression in *E. coli*, therefore this study provides further evidence that codon harmonization may provide a general strategy for improving the expression of soluble, functional proteins in a wide range of bacterial hosts. Precedence for this already exists if one analyses the expression of two other Gram-negative PDCs (ZmoPDC and ZpaPDC) in *G. thermoglucosidasius* . Production of soluble ZmoPDC in cell free extracts of *G. thermoglucosidasius* grown at 52, 54, 56, and 58 °C was observed to decrease with increased temperature, and PDC activity was undetectable above 52 °C (Thompson et al. 2008). For ZpaPDC, activity was absent at growth temperatures above 45 °C (Taylor et al. 2008). An analysis of the codon usage pattern of these PDCs expressed in *G. thermoglucosidasius* revealed that they have a higher coincidence of both frequency-of-usage and positioning of rare codons for expression in *G. thermoglucosidasius* (Table 3), which does not represent a large deviation from the frequency in the native host. Therefore, this provides further support for the codon harmonization concept, and this correlation may be responsible for the reported variations in PDC-expression efficiency in *G. thermoglucosidasius* : ZmoPDC> ZpaPDC> GoxPDC.

Despite the improvement in GoxPDC expression as a result of codon harmonization, there are still other factors which continue to play a major role in the functionality of the enzyme at higher growth temperatures. When correctly folded the GoxPDC protein displays relatively high thermostability when assayed in vitro. However, this does not necessarily translate to the ability to fold correctly at elevated temperatures, offering a possible explanation for the apparent failure of functional expression at 52 °C, unlike ZmoPDC when expressed in the same host (Thompson et al. 2008). We suggest that although codon harmonization contributes to the correct folding of a nascent protein during translation of the mRNA, it cannot necessarily compensate for the kinetics involved in protein folding in temperature ranges outside those for which the protein had been selected for or evolved under. The further stabilization of the enzyme therefore represents an area of future improvement for the use of PDC in engineering superior homo-ethanolic pathways in *G. thermoglucosidasius* .

The significant role that codon harmonization played in the correct processing of GoxPDC protein when expressed in *G. thermoglucosidasius* serves to reiterate the importance of codon usage in heterologous protein expression. This study represents the first account of improved expression of a protein of mesophilic origin in a thermophilic host using this technique, and demonstrates the potential benefits for microbial biotechnology. Due to its metabolic versatility, *Geobacillus* is a suitable platform organism for the synthesis of additional industrial products. In this light, codon harmonization should play a pivotal role in enabling and improving its development and in expanding its biotechnological repertoire.

Acknowledgments The *G. oxydans* strain was a kind gift from the Institute for Wine Biotechnology (IWBT) at the University of Stellenbosch. This work was supported by the National Research Foundation of South Africa.

References

Angov E (2011) Codon usage: nature's roadmap to expression and folding of proteins. Biotechnol J 66:650–659

Angov E, Hillier CJ, Kincaid RL, Lyon JA (2008) Heterologous protein expression is enhanced by harmonizing the codon usage frequencies of the target gene with those of the expression host. PLoS ONE 3: e2189

Angov E, Legler PM, Mease RM (2011) Adjustment of codon usage frequencies by codon harmonization improves protein expression and folding. Methods Mol Biol 705:1–13

Bi C, Zhang X, Ingram LO, Preston JF (2009) Genetic engineering of *Enterobacter asburiae* strain JDR-1 for efficient production of ethanol from hemicellulose hydrolysates. Appl Environ Microbiol 75:5743–5749

Bongers RS, Hoefnagel MH, Kleerebezem M (2005) High-level acetaldehyde production in *Lactococcus lactis* by metabolic engineering. Appl Environ Microbiol 71:1109–1113

Clarke TF IV, Clark PL (2008) Rare codons cluster. PLoS ONE 3:e3412

Conway T, Osman YA, Konnan JI, Hoffmann EM, Ingram LO (1987) Promoter and nucleotide sequences of the *Zymomonas mobilis* pyruvate decarboxylase. J Bacteriol 169:949–954

Correa A, Oppezzo P (2011) Tuning different expression parameters to achieve soluble recombinant proteins in *E. coli*: advantages of high-throughput screening. Biotechnol J 66:715–730

Cripps RE, Eley K, Leak DJ, Rudd B, Taylor M, Todd M, Boakes S, Martin S, Atkinson T (2009) Metabolic engineering of *Geobacillus thermoglucosidasius* for high yield ethanol production. Metab Eng 11:398–408

Feldmann S, Sprenger GA, Sham H (1989) Ethanol production from xylose with a pyruvate-formate-lyase mutant of *Klebsiella planticola* carrying a pyruvate-decarboxylase gene from *Zymomonas mobilis*. Appl Microbiol Biotechnol 31:152–157

Gocke D, Graf T, Brosi H, Frindi-Wosch I, Walter L (2009) Comparative characterisation of thiamin diphosphate-dependent decarboxylases. J Mol Catal B Enzym 61:30–35

Gold RS, Meagher MM, Tong S, Hutkins RW, Conway T (1996) Cloning and expression of the *Zymomonas mobilis* "production of ethanol" genes in *Lactobacillus casei*. Curr Microbiol 33:256–260

Gupta A, Singh VK, Qazi GN, Kumar A (2001) *Gluconobacter oxydans*: its biotechnological applications. J Mol Microbiol Biotechnol 33: 445–456

Gustafsson C, Govindarajan S, Minshull J (2004) Codon bias and heterologous protein expression. Trends Biotechnol 22:346–353

Ingram LO, Conway T, Clark DP, Sewell GW, Preston JF (1987) Genetic engineering of ethanol production in *Escherichia coli*. Appl Environ Microbiol 53:2420–2425

Kaczowka SJ, Reuter CJ, Talarico LA, Maupin-Furlow JA (2005) Recombinant production of *Zymomonas mobilis* pyruvate decarboxylase in the haloarchaeon *Haloferax volcanii* . Archaea 15: 327–334

Kane JF (1995) Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. Curr Opin Biotechnol 65: 494–500

Kim S, Lee SB (2006) Rare codon clusters at 5′-end influence heterologous expression of archaeal gene in *Escherichia coli*. Protein Expr Purif 50:49–57

King TE, Cheldelin VH (1954) Pyruvic carboxylase of *Acetobacter suboxydans* . J Biol Chem 208:821–831

Konig S (1998) Subunit structure, function and organisation of pyruvate decarboxylases from various organisms. Biochim Biophys Acta 1385:271–286

Konig S, Spinka M, Kutter S (2009) Allosteric activation of pyruvate decarboxylases. A never-ending story? J Mol Catal B Enzym 61: 100–110

Kotze AA, Tuffin IM, Deane SM, Rawlings DE (2006) Cloning and characterization of the chromosomal arsenic resistance genes from *Acidithiobacillus caldus* and enhanced arsenic resistance on conjugal transfer of *ars* genes located on transposon TnAtcArs. Microbiology 152:3551–3560

Laemmli UK (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 227:680–685

Lawford HG, Rousseau JD (1991) Ethanol production by recombinant *Escherichia coli* carrying genes from *Zymomonas mobilis*. Appl Biochem Biotechnol 28–29:221–236

Liu S, Dien BS, Cotta MA (2005) Functional expression of bacterial *Zymobacter palmae* pyruvate decarboxylase gene in *Lactococcus lactis*. Curr Microbiol 50:324–328

Liu S, Nichols NN, Dien BS, Cotta MA (2006) Metabolic engineering of a *Lactobacillus plantarum* double ldh knockout strain for enhanced ethanol production. J Ind Microbiol Biotechnol 33:1–7

Liu S, Dien BS, Nichols NN, Bischoff KM, Hughes SR, Cotta MA (2007) Coexpression of pyruvate decarboxylase and alcohol dehydrogenase genes in *Lactobacillus brevis*. FEMS Microbiol Lett 274: 291–297

Lowe SE, Zeikus JG (1992) Purification and characterization of pyruvate decarboxylase from *Sarcina ventriculi*. J Gen Microbiol 138:803–807

Lu G, Dobritzsch D, Baumann S, Schneider G, Konig S (2000) The structural basis of substrate activation in yeast pyruvate decarboxylase. A crystallographic and kinetic study. Eur J Biochem 267:861–868

Meyer D, Neumann P, Parthier C, Friedemann R, Nemeria N, Jordan F, Tittmann K (2010) Double duty for a conserved glutamate in pyruvate decarboxylase: evidence of the participation in stereoelectronically controlled decarboxylation and in protonation of the nascent carbanion/enamine intermediate. Biochemistry 49: 8197–8212

Ohta K, Beall DS, Mejia JP, Shanmugam KT, Ingram LO (1991) Genetic improvement of *Escherichia coli* for ethanol production: chromosomal integration of *Zymomonas mobilis* genes encoding pyruvate decarboxylase and alcohol dehydrogenase II. Appl Environ Microbiol 57:893–900

Orencio-Trejo M, Flores N, Escalante A, Hernandez-Chavez G, Bolivar F, Gosset G, Martinez A (2008) Metabolic regulation analysis of an ethanologenic *Escherichia coli* strain based on RT-PCR and enzymatic activities. Biotechnol Biofuels 1:8

Peters B, Junker A, Brauer K, Mühlthaler B, Kostner D, Mientus M, Liebl W, Ehrenreich A (2012) Deletion of pyruvate decarboxylase by a new method for efficient markerless gene deletions in *Gluconobacter oxydans*. Appl Microbiol Biotechnol 97:2521–2530

Pohl M, Mesch K, Rodenbrock A, Kula MR (1995) Stability investigations on the pyruvate decarboxylase from *Zymomonas mobilis* . Biotechnol Appl Biochem 22:95–105

Pohl M, Siegert P, Mesch K, Bruhn H, Grotzinger J (1998) Active site mutants of pyruvate decarboxylase from *Zymomonas mobilis*—a site-directed mutagenesis study of L112, I472, I476, E473, and N482. Eur J Biochem 257:538–546

Prust C, Hoffmeister M, Liesegang H, Wiezer A, Fricke WF, Ehrenreich A, Gottschalk G, Deppenmeier U (2005) Complete genome sequence of the acetic acid bacterium *Gluconobacter oxydans*. Nat Biotechnol 23:195–200

Raj KC, Ingram LO, Maupin-Furlow JA (2001) Pyruvate decarboxylase: a key enzyme for the oxidative metabolism of lactic acid by *Acetobacter pasteurianus*. Arch Microbiol 176:443–451

Raj KC, Talarico LA, Ingram LO, Maupin-Furlow JA (2002) Cloning and characterization of the *Zymobacter palmae* pyruvate decarboxylase gene (pdc) and comparison to bacterial homologues. Appl Environ Microbiol 68:2869–2876

Rosano GL, Ceccarelli EA (2009) Rare codon content affects the solubility of recombinant proteins in a codon bias-adjusted *Escherichia coli* strain. Microb Cell Fact 8:41

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: a laboratory manual. Cold Spring Harbour Press, Cold Spring Harbor, NY

Shaw AJ, Podkaminer KK, Desai SG, Bardsley JS, Rogers SR, Thorne PG, Hogsett DA, Lynd LR (2008) Metabolic engineering of a thermophilic bacterium to produce ethanol at high yield. Proc Natl Acad Sci 105:13769–13774

Siegert P, McLeish MJ, Baumann M, Iding H, Kneen MM, Kenyon GL, Pohl M (2005) Exchanging the substrate specificities of pyruvate decarboxylase from *Zymomonas mobilis* and benzoylformate decarboxylase from *Pseudomonas putida*. Protein Eng Des Sel 18:345–357

Talarico LA, Ingram LO, Maupin-Furlow JA (2001) Production of the Gram-positive *Sarcina ventriculi* pyruvate decarboxylase in *Escherichia coli*. Microbiology 147:2425–2435

Talarico LA, Gil MA, Yomano LP, Ingram LO, Maupin-Furlow JA (2005) Construction and expression of an ethanol production operon in Gram-positive bacteria. Microbiology 151:4023–4031

Taylor MP, Esteban CD, Leak DJ (2008) Development of a versatile shuttle vector for gene expression in *Geobacillus* spp. Plasmid 60: 45–52

Taylor MP, Eley KL, Martin S, Tuffin MI, Burton SG, Cowan DA (2009) Thermophilic ethanologenesis: future prospects for second-generation bioethanol production. Trends Biotechnol 27:398–405

Thanaraj TA, Argos P (1996) Protein secondary structural types are differentially coded on messenger RNA. Protein Sci 5:1973–1983

Thompson AH, Studholme DJ, Green EM, Leak DJ (2008) Heterologous expression of pyruvate decarboxylase in *Geobacillus thermoglucosidasius*. Biotechnol Lett 30:1359–1365

Tolan JS, Finn RK (1987) Fermentation of D-xylose to ethanol by genetically modified *Klebsiella planticola* . Appl Environ Microbiol 53:2039–2044

Underwood SA, Buszko ML, Shanmugam KT, Ingram LO (2002a) Flux through citrate synthase limits the growth of ethanologenic *Escherichia coli* KO11 during xylose fermentation. Appl Environ Microbiol 68:1071–1081

Underwood SA, Zhou S, Causey TB, Yomano LP, Shanmugam KT, Ingram LO (2002b) Genetic changes to optimize carbon partitioning between ethanol and biosynthesis in ethanologenic *Escherichia coli*. Appl Environ Microbiol 68:6263–6272

Vuralhan Z, Luttik MA, Tai SL, Boer VM, Morais MA, Schipper D, Almering MJ, Kotter P, Dickinson JR, Daran JM, Pronk JT (2005) Physiological characterization of the ARO10-dependent, broad-substrate-specificity 2-oxo acid decarboxylase activity of *Saccharomyces cerevisiae*. Appl Environ Microbiol 71:3276–3284

The final publication is available at Springer via http://dx.doi.org/10.1007/s00253-013-5380-1

Permission to reproduce the article here:

Excerpt From Springer Copyright Transfer Contract:

"Author retains the right to use his/her article for his/her further scientific career by including the final published journal article in other publications such as dissertations and postdoctoral qualifications provided acknowledgement is given to the original source of publication."

License no: 4171350839014

Correspondence with the journal:

"Dear Leonardo van Zyl,

We have just received the submission entitled: "Engineering pyruvate decarboxylase-mediated ethanol production in the thermophilic host Geobacillus thermoglucosidasius" for possible publication in Applied Microbiology and Biotechnology, and you are listed as one of the co-authors.

Could you please verify that you are affiliated with this submission?

Warning: This is an automated email, please DO NOT respond to this sender but through the links below:

If you are affiliated, please click this link:

http://amab.edmgr.com/l.asp?i=95347&l=S2EEL011 If

you are NOT affiliated, please click this link:

http://amab.edmgr.com/l.asp?i=95348&l=C3YKVRQ2

Thank you very much for your kind attention and cooperation.

Best regards,

Springer Journals Editorial Office

Applied Microbiology and Biotechnology"

From: "Alexander Steinbuchel" steinbu@uni-muenster.de

Date: 02 September 2013 0:02:52 SAST

To: "Don Cowan" <Don.Cowan@up.ac.za>

Subject: AMAB: Your manuscript entitled Engineering pyruvate decarboxylase-mediated ethanol production in the thermophilic host Geobacillus thermoglucosidasius

Ref.: Ms. No. AMAB-D-13-01789

Engineering pyruvate decarboxylase-mediated ethanol production in the thermophilic host Geobacillus thermoglucosidasius

Applied Microbiology and Biotechnology

Dear Professor Cowan,

I write you in regards to manuscript AMAB-D-13-01789 entitled "Engineering pyruvate decarboxylase-mediated ethanol production in the thermophilic host Geobacillus thermoglucosidasius" which you submitted to the Applied Microbiology and Biotechnology.

I am sorry to inform you that your manuscript does not meet the requirements as defined in the guidelines for authors (please see http://www.springer.com/chemistry/biotech/journal/253 for online instructions):

1.) It is the journal's strict policy that the major (wild-type) strains used in a study must be deposited in a publicly accessible culture collection belonging to the WDCM like DSMZ, ATCC, CGMCC etc. (see http://www.wfcc.info/index.php/collections/display/ for a list of the WDCM culture collections), and the collection numbers must be provided.

Please provide such a collection number for the Gluconobacter oxydans isolate used. If the strain has not been deposited yet, this must be done before your manuscript can be considered. Local deposit (at the IWBT, University of Stellenbosch lab collection) is not sufficient.

2.) It is also required that relevant nucleotide sequences must be submitted to a database, and the accession numbers must be provided. This also applies to codon-harmonized genes.

Please provide accession numbers for the wild-type (GoxPDCWT) and the codon harmonized (GoxPDCOPT) GoxPDC genes.

Your manuscript has therefore been denied publication in the Applied Microbiology and Biotechnology.

If you desire to resubmit your manuscript, please submit the revised version online as a "new manuscript", and indicate the ms-ID of the rejected version in this case.

This is your login information:

Your username is: DonCowan

Your password is:

In addition to addressing the comments above, please make sure that your resubmitted manuscript complies with the following journal requirements:

- Have all in the study described strains been deposited in a public strain collection? Has the collection number been mentioned in the manuscript?

- Does the manuscript meet the guidelines for authors of AMB?

- Have "Results" and "Discussion" sections been separated? - (For Original Paper articles only)

- Are the conclusions integrated in the discussion? (There is no separate "Conclusion" section allowed!) - (For Original Paper articles only)

- Have the references been prepared according to the format requested in the guideline to authors?

- Have all taxa names (species names, genus names, and names of higher categories) been italicized?

- It is not allowed to submit the text of your revised version as pdf-file. Please upload a word- or rtf-document.

Thank you for considering the Applied Microbiology and Biotechnology for the publication of your research.

Yours sincerely,

Dorothea Kessler

Dr. Dorothea Kessler

Managing Editor

on behalf of

Prof. Dr. Alexander Steinbuchel

Editor-in-Chief

Applied Microbiology and Biotechnology

From: "AMB Editorial Office" <amboffice@gmx.de>

Date: 12 September 2013 10:50:54 GMT+02:00

To: "Don Cowan" <Don.Cowan@up.ac.za>

Subject: AMAB: PDF Engineering pyruvate decarboxylase-mediated ethanol production in the thermophilic host Geobacillus thermoglucosidasius has been built and requires approval

Dear Prof. Cowan,

The PDF for your submission, "Engineering pyruvate decarboxylase-mediated ethanol production in the thermophilic host Geobacillus thermoglucosidasius" is ready for viewing.

Please go to http://amab.edmgr.com/ to approve your submission.

Your username is: DonCowan

Your password is:

Your submission must be approved in order to complete the submission process and send the manuscript to the Applied Microbiology and Biotechnology editorial office.

Please view the submission before approving it to be certain that your submission remains free of any errors.

Thank you for your time and patience.

Editorial Office

Applied Microbiology and Biotechnology

Dear Prof. Cowan,

Your submission entitled "Engineering pyruvate decarboxylase-mediated ethanol production in the thermophilic host Geobacillus thermoglucosidasius" has been assigned the following manuscript number: AMAB-D-13-01934.

You will be able to check on the progress of your paper by logging on to Editorial Manager as an author.

The URL is http://amab.edmgr.com/.

Thank you for submitting your work to this journal.

Kind regards,

Editorial Office

Applied Microbiology and Biotechnology

>>> "Alexander Steinbuchel" <steinbu@uni-muenster.de> 2013/10/17 12:42 AM >>>

Ref.: Ms. No. AMAB-D-13-01934

Engineering pyruvate decarboxylase-mediated ethanol production in the thermophilic host Geobacillus thermoglucosidasius

Applied Microbiology and Biotechnology

Dear Prof. Cowan,

your manuscript AMAB-D-13-01934 entitled Engineering pyruvate decarboxylase-mediated ethanol production in the thermophilic host Geobacillus thermoglucosidasius which you submitted to "Applied Microbiology and Biotechnology", has been reviewed.

The reviewers' comments can be found at the end of this email or can be accessed by following the provided link.

The reviewer recommended publication, but also suggest some minor revisions to your manuscript. Therefore, I invite you to respond to the reviewers' comments and revise your manuscript.

This is your login information:

Your username is: DonCowan

Your password is:

When revising your work, please submit a list of changes or a rebuttal against each point which is being raised when you submit the revised manuscript.

Please upload the revised version within the next three weeks.

To submit a revision, go to http://amab.edmgr.com/ and log in as an Author. You will see a menu item called 'Submissions Needing Revision'. You will find your submission record there.

Beside the comments of the reviewers, please make sure to address the points listed in the following checklist before you submit your revised version:

- Have all in the study described strains been deposited in a public strain collection? Has the collection number been mentioned in the manuscript?

- Does the manuscript meet the guidelines for authors of AMB?

- Have "Results" and "Discussion" sections been separated? - (For Original Paper articles only)

- Are the conclusions integrated in the discussion? (There is no separate "Conclusion" section allowed!) - (For Original Paper articles only)

- Have the references been prepared according to the format requested in the guideline to authors?

- Have all taxa names (species names, genus names, and names of higher categories) been italicized?

- It is not allowed to submit the text of your revised version as pdf-file. Please upload a word- or rtf-document.

Once again, thank you for submitting your manuscript to "Applied Microbiology and Biotechnology" and I look forward to receiving your revision.

Yours sincerely

Alexander Steinbuchel

Editor-in-Chief

Applied Microbiology and Biotechnology

Reviewers' comments:

Reviewer #1: Please enter your comments to the Author below

A well written document about ethanol production by Geobacillus at elevated temperatures by introduction of a codon harmonized PDC gene.

Only a few small remarks:

Page 3, line 48-49. This sentence suggests that yeasts also use the Entner Doudoroff pathway, please rephrase

Page 5, line 87. I would opt for honey in stead of honeybees....

Page 18, line 391-401. The absence of formate suggests that all ethanol was synthesized through PDC. That's interesting to note.

>>> "AMB Editorial Office" <amboffice@gmx.de> 2013/10/30 06:51 AM >>>

Ref.: Ms. No. AMAB-D-13-01934R1

Engineering pyruvate decarboxylase-mediated ethanol production in the thermophilic host Geobacillus thermoglucosidasius

Applied Microbiology and Biotechnology

Dear Prof. Cowan,

We are expecting the revision of AMAB-D-13-01934R1 by 06 Nov 2013.

If you require more time, please contact the journal office. If you are ready to submit your revision, then please go to http://AMAB.edmgr.com/ and submit the revision.

Your username is: DonCowan

Your password is: Kind

regards,

@ Reminder Service

Reminder Setup

Applied Microbiology and Biotechnology

>>> "AMB Editorial Office" <amboffice@gmx.de> 2013/10/30 08:11 AM >>>

Dear Prof. Cowan,

Re: Engineering pyruvate decarboxylase-mediated ethanol production in the thermophilic host Geobacillus thermoglucosidasius

Thank you for approving the changes that the Editor made to your submission or updating your submission according to the requested changes.

You will be able to check on the progress of your paper by logging on to Editorial Manager as an author. The URL is http://amab.edmgr.com/.

Thank you for submitting your work to this journal.

Kind regards,

Editorial Office

Applied Microbiology and Biotechnology

Dear Dr. van Zyl,

thank you for contacting us.

I just resent the letter requesting confirmation of authorship to KEley@tmo-group.com

Please note that sometimes these system emails are mis-identified as spam. It would therefore be good if you coauthor could also check her spam folder for the missing letter.

Best regards,

Dorothea Kessler

Dr. Dorothea Kessler

Managing Editor

Applied Microbiology and Biotechnology

Am 30.10.13 12:24, schrieb lonnie van zyl:

Dear Loida,

We recently submitted and had a paper accepted for publication by AMB titled "Engineering pyruvate decarboxylase-mediated ethanol production in the thermophilic host Geobacillus thermoglucosidasius", however our submission can't go any further without all authors replying to your office that they are indeed affiliated with the paper. One author is still outstanding on this manuscript: Dr. Kirstin Eley from TMO. I gave her e-mail address as keley@tmo-group.com, but the correct one may be KEley@tmo-group.com. Could you please see to it that she gets sent the e-mail to confirm her affiliation. She says that she has not received any notification from AMB as of yet.

Kind regards

Lonnie van Zyl

>>> "AMB Editorial Office" <amboffice@gmx.de> 2013/10/30 08:08 AM >>>

Dear Prof. Cowan,

Before your manuscript can proceed to the next process, you are

requested to accomplish the following:

The other author(s) listed below have yet to confirm that they are

affiliated with the submission. Therefore, a letter requesting

authorship verification has been sent to them in a separate email. Also,

please check if the contact address(es) entered in the system is/are

correct. To view the verification status, please click 'Author

Status' in the Action links.

Mark Paul Taylor

Kirstin Eley

*Note: Please do not resubmit unless all co-authors click the link to

approve the co-authorship verification letter. This is to formally

accept the affiliation online.

If you have any questions, please do not hesitate to contact me.

Kind regards,

Loida Escueta

JEO Assistant

Applied Microbiology and Biotechnology

From: "Alexander Steinbuchel" <steinbu@uni-muenster.de>

Date: 02 November 2013 10:14:50 CAT

To: "Don Cowan" <Don.Cowan@up.ac.za>

Subject: AMAB: Your manuscript entitled Engineering pyruvate decarboxylase-mediated ethanol production in the thermophilic host Geobacillus thermoglucosidasius

Ref.: Ms. No. AMAB-D-13-01934R1

Engineering pyruvate decarboxylase-mediated ethanol production in the thermophilic host Geobacillus thermoglucosidasius

Applied Microbiology and Biotechnology

Dear Prof. Cowan,

It is a pleasure to accept your manuscript in its current form for publication in Applied Microbiology and Biotechnology.

The article proofs will be sent to you about 3-4 weeks after receipt of this email.

You may check the progress of the publication process at My Springer under Article Tracking

(http://www.springer.com/generic/registration?SGWID=0-40105-6-1004221-0).

Thank you for your fine contribution. On behalf of the Editors of the "Applied Microbiology and Biotechnology", we look forward to your continued contributions to the Journal.

Best regards,

Alexander Steinbuchel

Editor-in-Chief

Applied Microbiology and Biotechnology

Chapter 3

Author contributions

Don Cowan and Marla Tuffin conceived the study and participated in its design and coordination. Leonardo Joaquim van Zyl performed cloning, purification of the protein, crystallization, sequence alignment, phylogenetic tree construction. Wolf-Dieter Schubert collected X-ray data. Wolf-Dieter Schubert and Leonardo Joaquim van Zyl solved and refined the crystal structure. Leonardo Joaquim van Zyl wrote the bulk of the manuscript and performed analysis of results. All authors read and approved the final manuscript.

BMC Structural Biology

RESEARCH ARTICLE Open Access

Structure and functional characterization of pyruvate decarboxylase from *Gluconacetobacter diazotrophicus*

Leonardo J van Zyl[1*], Wolf-Dieter Schubert[2], Marla I Tuffin[1] and Don A Cowan[3]

Abstract

Background: Bacterial pyruvate decarboxylases (PDC) are rare. Their role in ethanol production and in bacterially mediated ethanologenic processes has, however, ensured a continued and growing interest. PDCs from *Zymomonas mobilis* (ZmPDC), *Zymobacter palmae* (ZpPDC) and *Sarcina ventriculi* (SvPDC) have been characterized and ZmPDC has been produced successfully in a range of heterologous hosts. PDCs from the *Acetobacteraceae* and their role in metabolism have not been characterized to the same extent. Examples include *Gluconobacter oxydans* (GoPDC), *G. diazotrophicus* (GdPDC) and *Acetobacter pasteutrianus* (ApPDC). All of these organisms are of commercial importance.

Results: This study reports the kinetic characterization and the crystal structure of a PDC from *Gluconacetobacter diazotrophicus* (GdPDC). Enzyme kinetic analysis indicates a high affinity for pyruvate (K_M 0.06 mM at pH 5), high catalytic efficiencies ($1.3 \cdot 10^6\ M^{-1} \cdot s^{-1}$ at pH 5), pH_{opt} of 5.5 and T_{opt} at 45°C. The enzyme is not thermostable ($T_{½}$ of 18 minutes at 60°C) and the calculated number of bonds between monomers and dimers do not give clear indications for the relatively lower thermostability compared to other PDCs. The structure is highly similar to those described for *Z. mobilis* (ZmPDC) and *A. pasteurianus* PDC (ApPDC) with a rmsd value of 0.57 Å for Cα when comparing GdPDC to that of ApPDC. Indole-3-pyruvate does not serve as a substrate for the enzyme. Structural differences occur in two loci, involving the regions Thr341 to Thr352 and Asn499 to Asp503.

Conclusions: This is the first study of the PDC from *G. diazotrophicus* (PAL5) and lays the groundwork for future research into its role in this endosymbiont. The crystal structure of GdPDC indicates the enzyme to be evolutionarily closely related to homologues from *Z. mobilis* and *A. pasteurianus* and suggests strong selective pressure to keep the enzyme characteristics in a narrow range. The pH optimum together with reduced thermostability likely reflect the host organisms niche and conditions under which these properties have been naturally selected for. The lack of activity on indole-3-pyruvate excludes this decarboxylase as the enzyme responsible for indole acetic acid production in *G. diazotrophicus*.

Background

Pyruvate decarboxylase (PDC, EC 4.1.1.1) is the enzyme responsible for the non-oxidative decarboxylation of pyruvate to acetaldehyde and carbon dioxide. All characterized PDCs are dependent on the cofactors thiamine diphosphate (ThDP) and Mg^{2+}. A recent study proposed a PDC capable of co-factor independent decarboxylation of pyruvate [1], however this discovery has been refuted [2]. Although widespread in the plant kingdom and amongst ascomycetous yeasts and fungi, PDCs are comparatively rare in prokaryotes. Several of the plant and yeast PDCs have been isolated and characterized, however; by contrast only five bacterial PDCs have been described, namely those from *Zymomonas mobilis* (ZmPDC), *Zymobacter palmae* (ZpPDC), *Sarcina ventriculi* (SvPDC), *Acetobacter pasteurianus* (ApPDC) and *Gluconobacter oxydans* (GoPDC) [3-8].

In higher organisms and most prokaryotes (*Z. mobilis*, Z. *palmae* and *S. ventriculi*), the PDC forms part of the fermentative pathway leading to ethanol production. Therefore, bacterial PDCs and their hosts have been the focus of extensive characterization and engineering

* Correspondence: vanzyllj@gmail.com
[1]Institute for Microbial Biotechnology and Metagenomics (IMBM), University of the Western Cape, Robert Sobukwe Road, Bellville, Cape Town, South Africa
Full list of author information is available at the end of the article

efforts to develop ethanologenic strains [7,9-15]. In the *Acetobacteraceae* (*A. pasteurianus*, and *G. oxydans*) however, PDC links oxidative lactate assimilation (lactate dehydrogenase; pyruvate forming) and ethanol consumption (alcohol dehydrogenase; pyruvate forming) to the production of acetate, and therefore forms part of oxidative metabolism [4,16]. In *G. oxydans*, which only has a partial TCA cycle, all L-lactate, fructose and mannitol is converted to acetate via the PDC showing its metabolic importance in this organism [16].

Although the exact mechanism of ThDP dependent decarboxylation has not yet been fully described, it centrally involves the deprotonation of atom C2 of the thiazolium ring to yield a corresponding carbanion or ylide [17]. The latter nucleophillically attacks the carbonyl group of pyruvate substrate to yield a C2-α-lactylthiamin diphosphate intermediate [18,19]. The enzymes bind ThDP in a conformation that places the N4' atom of the aminopyrimidine ring near atom C2. N4' is a strong base in the imino tautomeric state of the aminopyrimidine ring allowing it to deprotonate C2 and activate the cofactor. Glu50, within hydrogen bonding distance of N1 and deprotonated under physiological conditions, was previously thought to induce the amino to imino tautomerization of the aminopyrimidine ring [20]. More recent studies of the pre-reaction state of ZmPDC, however, suggest that Glu469 instead directly abstracts a proton from N4' [21,22]. Decarboxylation of the lactyl cofactor adduct yields an enamine/carbanion mesomeric intermediate with concomitant CO_2 release. The carbanion/enamine intermediate becomes protonated to give hydroxyethyl ThDP and release of the acetaldehyde product regenerates the ylide [20,23-26]. Crystal structures for PDCs from *Z. mobilis* (ZmPDC) and *A. pasteurianus* (ApPDC) are published [27,28].

Gluconacetobacter diazotrophicus, a member of the family *Acetobacteraceae,* is a Gram negative, obligate aerobic bacterium. This organism is also nitrogen fixing and endophytic, setting it apart from other acetic acid bacteria. It is often found in association with sugar cane where it stimulates plant growth through the secretion of auxin-like compounds, notably indole acetic acid (IAA) [29,30]. No indolepyruvate decarboxylases could be identified on the *G. diazotrophicus* PAL5 genome sequence, however several decarboxylases were identified, one of which is possibly responsible for production of IAA from indole-3-pyruvate [31]. Of these, one showed significant sequence similarity to other true bacterial PDCs and although the role of PDC has been investigated in two other members of this family (see above), its role in this unique bacterium is not known.

As described, the enzyme fulfills multiple roles in key metabolic pathways and has potential for use in engineering of ethanologenic strains. In order to confirm the annotated sequence as a true PDC and to further elucidate the role of the enzymes in these plant-associated organisms, we kinetically characterized the PDC from *G. diazotrophicus* (GdPDC) and solved the GdPDC crystal structure at 1.7 Å, adding to our knowledge of these rare enzymes.

Results

Functional characterization of the *G. diazotrophicus* PDC

A search against the non-redundant NCBI database using the GdPDC protein sequence as query identified only 27 bacterial proteins (E-value = 0), despite the wealth of sequence data available, including metagenomic sequences. PDCs with identity to the bacterial enzymes which have been studied and which are not of *Acetobacteraceae* origin are few (Figure 1). All bacterial proteins related to GdPDC that are annotated as PDCs are shown in Figure 1, and included are the indole-3-pyruvate decarboxylase from *Enterobacter cloacae* and the benzoyl-formate decarboxylase from *Pseudomonas putida* for reference, as well as the best BLAST hit against the non-redundant NCBI environmental metagenomic proteins database. The same sequences are identified when using any of the five Gram negative PDCs as search query. The proteins related to the Gram negative PDCs from bacteria other than the *Acetobacteraceae* include putative enzymes from the family or order: *Chroococcales, Oscillatoriales* (2), *Alteromonadaceae, Legionellaceae* (2), *Chloroflexi, Acidobacteriaceae*, and *Beijerinckiaceae*.

G. diazotrophicus pdc was amplified, cloned and sequenced. PCR amplification introduced one amino acid change, P554Q, four residues from the end of the chain. As C-terminal deletions after this position do not affect activity for ZmPDC, this substitution is not expected to affect enzyme activity substantially [34]. Of the characterized PDC's, the amino acid sequence of GdPDC is most closely related to that of *Z. palmae* PDC sharing amino acid identity of 71%, followed by 70% to PDC from *A. pasteurianus*. The protein shares the typical ThDP binding motif GDGS-XXX-NN and retains conserved residues for substrate binding and catalysis (Additional file 1: Figure S1).

GdPDC was purified to homogeneity by affinity chromatography as judged by reducing SDS-PAGE analysis (Additional file 2: Figure S2). The MW of ±60 kDa corresponds well to the theoretical molecular mass of 59.2 kDa. The predicted pI is 5.8. The kinetic parameters of the enzyme are summarized in Table 1. The K_M value for pyruvate decreased ~20-fold on decreasing the pH from 7 to 5 and at pH 5 this value is twofold lower than the lowest K_M reported for any PDC at this pH [7]. The catalytic rate (k_{cat}) remains unaffected similar to related enzymes Table 1 [26] supporting the idea that PDC requires the de-protonation of the ThDP aminopyrimidine ring for catalysis [26]. The enzyme displays Michaelis

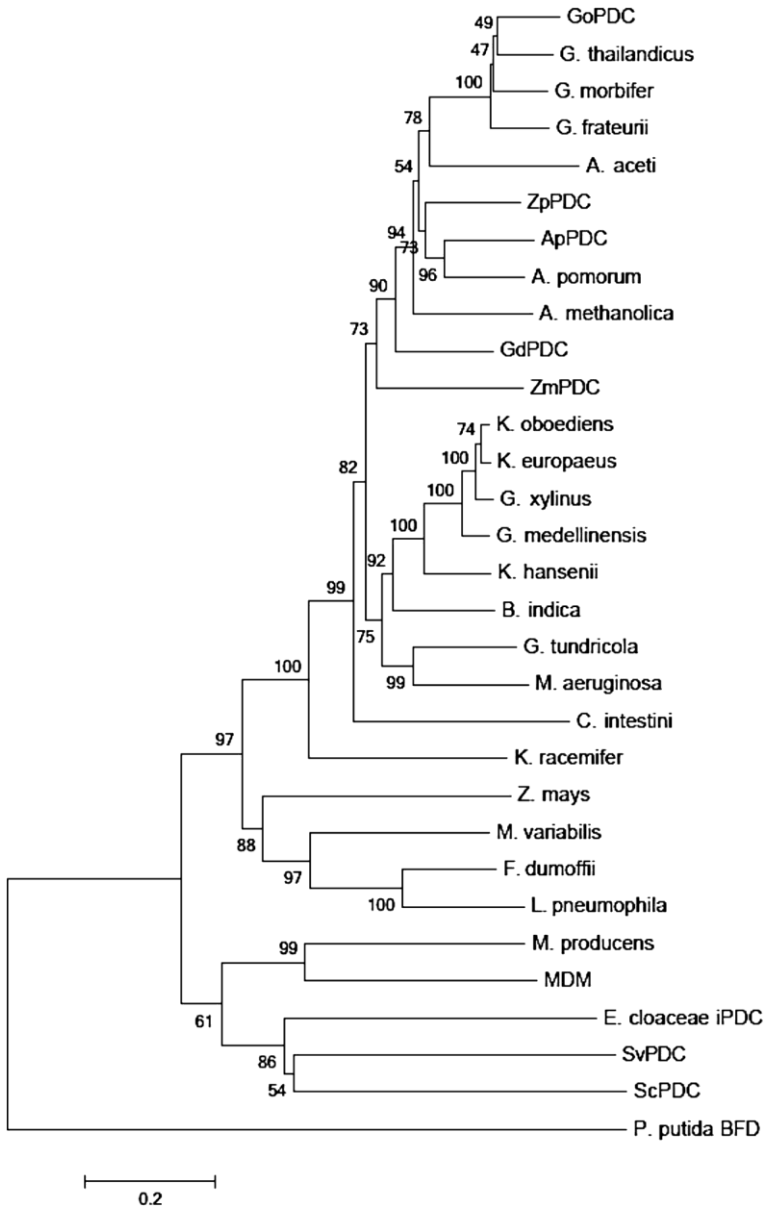


Figure 1 Neighbor-joining tree comparing full length amino acid sequences of PDC-related proteins. The optimal tree with the sum of branch length = 8.50849307 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) is shown next to the branches [32]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method [33] and are in the units of the number of amino acid substitutions per site (scale bar). The analysis involved 31 amino acid sequences. All ambiguous positions were removed for each sequence pair. There were a total of 633 positions in the final dataset. GdPDC - *G. diazotrophicus* (KJ746104); GoPDC - *G. oxydans* (KF650839); ApPDC *Acetobacter pasteurianus* (AF368435.1); ZpPDC - *Z. palmae* (AF474145); ZmPDC - *Z. mobilis* (AB359063); ZmPDC - *Z. mays* (X17555); ScPDC - *S. cerevisiae* (X04675); SvPDC - *S. ventriculi* (AF354297); *Lyngbya aestuarii* (WP023067698); *Acidomonas methanolica* (GAJ29946); *Acetobacter pomorum* (WP006115789); *Acetobacter aceti* (WP010667855); *Microcystis aeruginosa* (WP_0027648); *Moorea producens* (WP008180762); *Microbulbifer variabilis* (WP020414286); *Legionella pneumophila* (YP006505162); MDM (CBI10829); *Ktedonobacter racemifer* (WP007922190); *Komagataeibacter oboediens* (WP010515737); *Komagataeibacter hansenii* (WP003622049); *Komagataeibacter europaeus* (WP010509054); *Granulicella tundricola* (YP004210504); *Gluconobacter thailandicus* (WP007283613); *Gluconobacter morbifer* (WP008852112); *Gluconobacter frateurii* (WP023941876); *Gluconacetobacter xylinus* (AHI26557); *Gluconacetobacter medellinensis* (YP004868149); *Fluoribacter dumoffii* (WP010654974); *Enterobacter cloacae* iPDC (P23234); *Commensalibacter intestini* (WP008853550); *Beijerinckia indica* (YP001834435); *Pseudomonas putida* BFD (YP008115845) ; MDM- Mine Drainage Metagenome (CBI10829.1).

Table 1 Characterization data (Steady state kinetic constants, T_{opt} and pH_{opt}) for GdPDC using pyruvate and compared with those from other Gram negative bacteria (The values represent the average of at least two individual rounds of protein purification and assay)

PDC	K_M (mM)	Specific activity in (U/mg)	k_{cat}/K_M ($M^{-1}.s^{-1}$)	T_{opt} (°C)	$T_{½}$ at °C	pH_{opt}
GdPDC	0.06 (5.0)*	20 (5.0)	1.3×10^6 (5.0)	45-50	18 min at 60°C	5.0-5.5
	0.60 (6.0)	39 (6.0)	2.6×10^5 (6.0)			
	1.2 (7.0)	43 (7.0)	1.4×10^5 (7.0)			
GoPDC	0.12 (5.0) [7]	57 (5.0) [7]	1.9×10^6 (5.0) [7]	53 [7]	10 min at 65°C [7]	4.5-5.0 [7]
	1.2 (6.5) [7]	47 (6.5) [7]	1.6×10^5 (6.5) [7]			
	2.8 (7.0) [7]	125 (7.0) [7]	1.8×10^5 (7.0) [7]			
ApPDC	2.8 (6.5) [36]/0.39 (5.0) [5]	110 (6.5) [36]/97 (5.0) [5]	1.3×10^6 (5.0)# [5]	65 [36]	24 min at 70°C [36]	3.5 - 6.5 [36]
ZpPDC	2.5 (6.5) [36]/0.24 (6.0) [5]	116 (6.5) [36]/130 (6.0) [5]	1.4×10^6(6.0)# [5]	55 [36]	24 min at 60°C [36]	7.0 [36]
ZmPDC	1.3 (6.5) [36]/0.31 (6.0) [26]/1.1 [37]/0.4 (6.0) [38]	120 (6.5) [36]/120 [37]/181 [38]	1.9×10^6 (6.0) [26] / 4.4×10^5 (6.5) [37]/ 1.79×10^6 (6.0) [38]	60 [36]	30 min at 60°C [36]	6.0-6.5 [36]
SvPDC	13 [6]	103 [6]	3.2×10^4 [6] / 0.87×10^4 (7.0)	N/A	30 min at 50°C	6.3 - 6.7 [6]

Numbers after values are the references from which the numbers were obtained.
*Values in brackets indicate assay pH.
#Calculated based on values given in reference [5].

Menten kinetics with pyruvate as substrate and is not subject to allosteric substrate activation, as for the PDCs from plants, fungi, and the bacterium *S. ventriculi* [35]. Catalytic efficiencies were also similar to those reported for SvPDC, the only known representative from a Gram positive bacterium, and ZmPDC which is the best studied enzyme.

Its temperature optimum is between 45°C and 50°C (Figure 2A), one of the lowest for bacterial PDCs. GdPDC is less thermostable than PDCs from other Gram negative bacterial enzymes, retaining 15% activity after 30 min at 60°C (half-life of 18 min, Figure 2B) and no residual activity after 1 h at 60°C. The activation energy of GdPDC on pyruvate was determined in the linear range from 25°C to 45°C to be 46 kJ/mol, which is in agreement with values reported for other bacterial PDCs (44). The alanine, cysteine and phenylalanine content of PDCs was previously proposed to correlate with its thermostability [5]. Alanines constitute 17% of the residues in GdPDC (Cys 1.6%, Phe 2.5%) but 12% in GoPDC (2%, 3%), 15% in ZmPDC (1.2%, 3.1%), 13% in ZpPDC (1.8%, 2.7%), 13% in ApPDC (2%, 2.5%) and 6.9% in SvPDC (0.9%, 4.7%). Despite having the highest alanine content of all the bacterial PDCs, GdPDC is not the most thermostable, contradicting amino acid-based predictions [5]. Other factors might contribute to the lower *in vitro* thermostability of GdPDC observed

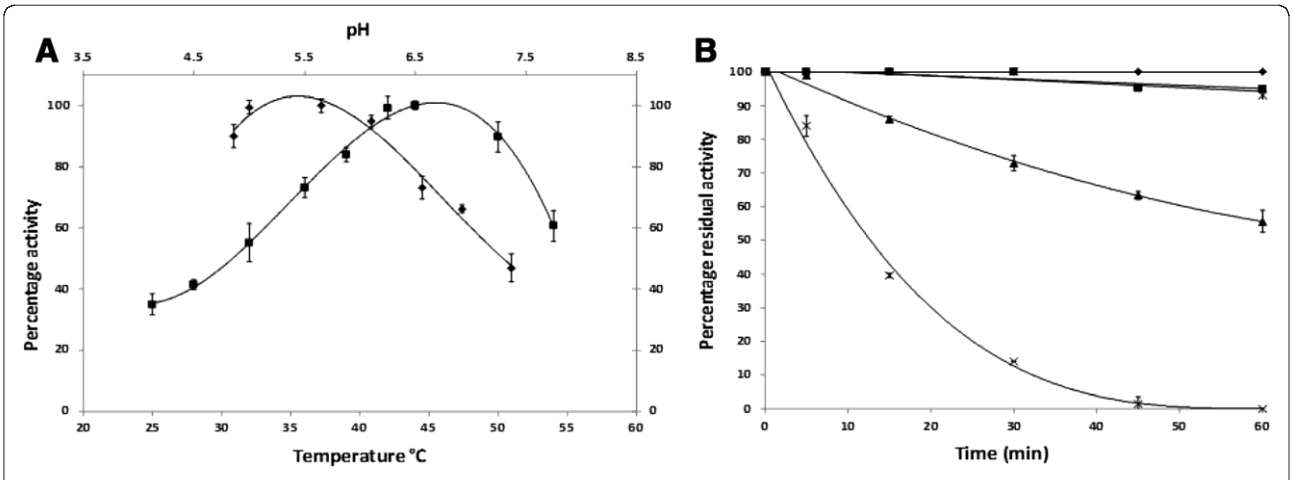


Figure 2 GdPDC characterization data. A) The temperature (■) and pH (♦) profiles of GdPDC using pyruvate as substrate. The assay is described in the methods section. 100% activity is analogous to a specific activity of 60 U/mg for T_{opt} and 36 U/mg for pH_{opt}. B) Thermal inactivation GdPDC at 25°C (♦), 40°C (■), 55°C (▲) and 60°C (×). Enzyme activity at zero time using the standard assay at 25°C is set to 100%, analogous to a specific activity of 10 U/mg. Assays were performed in 200 mM citrate buffer at pH 6.0 and pyruvate as substrate. The data represents the average of at least three individual rounds of protein purification and assay.

here, as has been summarized in a comparative study conducted by Pohl and coworkers [39]. For example, the use of $MgCl_2$ instead of $MgSO_4$ to provide the Mg^{2+} cofactor may affect thermostability as the sulfate anion is known to stabilize PDC enzymes [39].

GdPDC was assayed using a range of substrates including 2-ketopropanoate (pyruvate), 2-ketobutanoate, 2-ketopentanoate, 2-keto-4-methylpentanoate, 3-phenyl- 2-oxopropanoate, benzoyl formate, 3-hydroxy-phenyl pyruvate and indole-3-pyruvate. Specific activities for substrates 2-ketobutanoate (12 U/mg), 2-ketopentanoate (0.68 U/mg) and 2-keto-4-methylpentanoate (0.15 U/mg), respectively at 24 mM, are similar to those previously reported for other bacterial PDC's [36,37]. Activities for benzoyl formate, 3-hydroxy-phenyl pyruvate and indole-3-pyruvate, if present, were below detection limits.

G. diazotrophicus PDC crystal structure

GdPDC crystallized in the monoclinic space group C2 with cell dimensions: a = 129.1 Å, b =141.0 Å, c = 91.1 Å, β = 125.8°, with two monomers per asymmetric unit (Table 2; Additional file 3: Figure S3). The crystal structure of the *G. diazotrophicus* PDC was solved by molecular replacement using a side-chain cropped dimer of the *A. pasteurianus* PDC (2VBI) as a search model. The high resolution diffraction data (Table 2) and the good quality of the electron density distribution allowed for facile model building for the major part of the protein (see Methods) and most residues are well-defined.

Table 2 Statistics for data collection, processing and the final model of the GdPDC crystal structure

Statistics of data collection	
Resolution (Å)*	30.0 - 1.69 (1.78-1.69)
Wavelength (Å)(synchrotron and station)	0.980 (SOLEIL Proxima 1)
Total number of reflections*	435137 (61046)
Total number of unique reflections*	146264 (21363)
Multiplicity*	3.0 (2.9)
R_{merge}*	0.080 (0.299)
I/sd(I)*	8.8 (3.3)
Completeness (%)*	99.4 (99.6)
Statistics of refinement and the final model	
Resolution (Å)	84.06 – 1.69
Number of reflections	138948
R_{free}	0.158
R_{work}	0.127
rmsd [bond lengths (Å)/bond angles(°)/chiral volume (A^3)]	0.033/2.48/0.235
Ramachandran plot (preferred/allowed/outlier) (%)	98.1/1.5/0.4
B_{mean} of all atoms ($Å^2$)	14.8

*Values in brackets indicate the shell of highest resolution.

The quaternary structure of GdPDC is a homo-tetramer best described as a dimer of dimers (Figure 3A) as for ZmPDC and ApPDC. The tetramer is generated by applying a crystallographic 2-fold symmetry to the non-crystallographic dimer in the asymmetric unit. The accessible surface area of the monomer-monomer interface amounts to 3740 $Å^2$, somewhat smaller than the 4150 $Å^2$ for ZmPDC [27] but similar to that of ApPDC (3770 $Å^2$). The surface area between the dimers of the tetramer is 2738 $Å^2$ for GdPDC, 3784 $Å^2$ for ZmPDC and 3812 $Å^2$ for ApPDC. GdPDC has 63 hydrogen bonds between monomers, fewer than the 76 for ZmPDC but more than the 60 of ApPDC. Thirteen salt bridges support the monomer-monomer interface (ZmPDC 14, ApPDC 16). There is significantly less hydrogen bonding between dimers which make up a tetramer at 44 compared with ZmPDC-70 and ApaPDC-74, while the number of salt bridges also shows some variation with GdPDC having 26, ZmoPDC-20 and ApaPDC-28.

The refined crystal structure contains two identical chains of 544 amino acids (residues 2–180, 191–555), each binding a ThDP cofactor and a Mg^{2+} ion. The model contains 1167 water molecules. The rmsd for Cα atoms of the two monomers in the asymmetric unit is only 0.088 Å indicating a very high similarity and correspondingly a negligible effect of inter-monomer or crystal packing forces. As for other PDCs, each protein monomer may be thought of consisting of three distinct structural domains: the pyrimidine binding (PYR, residues 1–186), the regulatory (R, 187–349) and the pyrophosphate binding (PP, 350–558) domains. The rmsd between Cα atoms of GdPDC and ApPDC is 0.57 Å implying largely similar structures.

As mentioned, GdPDC demonstrated Michaelis Menten kinetics. Two residues, Tyr157 and Arg224, were shown to be involved in binding a second molecule of the substrate analogue pyruvamide in *Saccharomyces cerevisiae* PDC (ScPDC), and are conserved in SvPDC; both enzymes display substrate activation [35,40]. Arg224 (Arg221 in GdPDC and ZmPDC) is conserved in a range of PDC-like enzymes based on structure- and sequence-based alignments (Figure 4B and Additional file 1: Figure S1), however Tyr157 is not and appears to be unique to the enzymes showing substrate activation.

One of two ThDP molecules in the GdPDC structure appears to be modified as also reported for the ZmPDC, based on weak electron density for the C2 carbon atom of the thiazolium ring. As for ZmPDC, degradation of the cofactor presumably occurs after crystallization [27]; Figure 3B.

Residues 104–113 together with residues 290–304, in the structure of ScPDC (1PYD), are presumably involved in closing the active site during catalysis, as they are disordered, but adopt a stable conformation upon binding

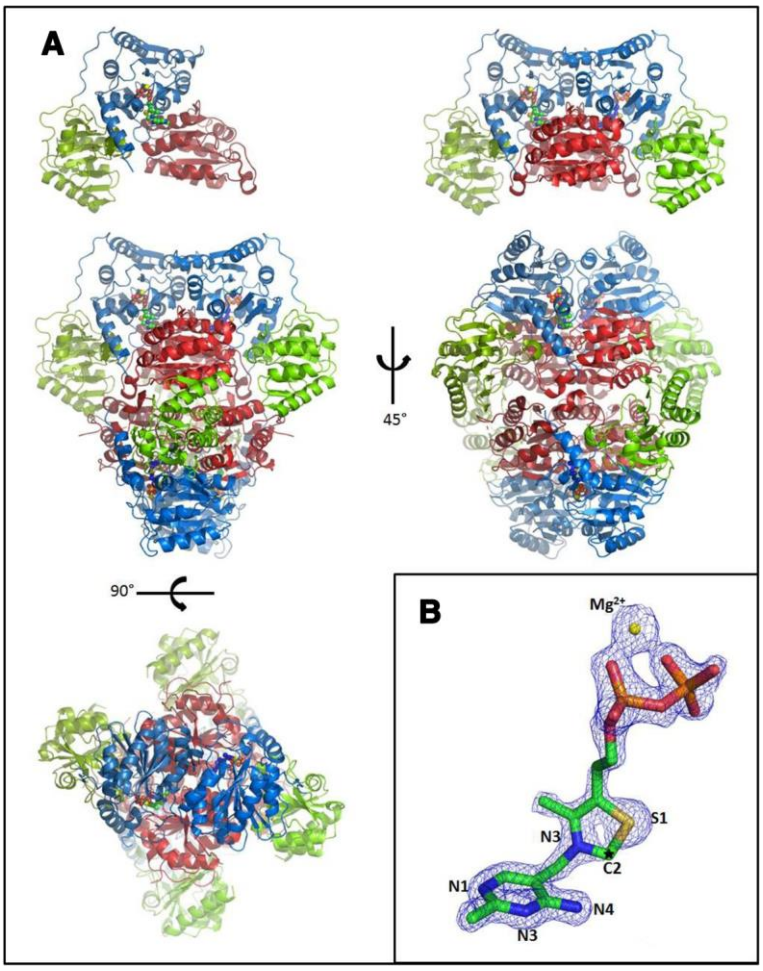


Figure 3 Tertiary and quarternary structure of GdPDC. A) A cartoon representation of GdPDC structure monomer (left) and dimer (right) showing the PYR–domain in red, PP-domain in blue and the R-domain in green. ThDP and Mg^{2+} are shown as space fill models. B) $2F_o$-F_c electron density map (blue, contoured at 2.0 σ) for ThDP. The lack of electron density for the C2 position of the thiazole ring may indicate the loss of this atom.

the substrate analogue pyruvamide (1QPB) [41,42]. In GdPDC these residues are well defined in the electron density map despite the absence of substrate, also as reported for ZmPDC. This may be due to stabilizing interactions with residues of the R- and PP-domains (N288, D289, Q407 and R553). Binding of the inactive ThDP triazole ring analogue and pyruvate induce dramatic conformational changes in ZmPDC [21]. Similar conformational changes would presumably also occur in GdPDC as this region is structurally highly conserved in bacterial PDCs (Figure 4C). A "water tunnel" links the two active sites (Figure 5) presumably to serve as a proton relay system as previously suggested for ZmPDC and the E1 subunit of PDHc [22,43].

Apart from differences in amino acid sequence, ZmPDC and GdPDC differ structurally in several areas (Figure 4A). In ZmPDC a loop of five amino acids (Asn499-Asp503) in the PP-domain extends toward the PP-domain of the second subunit creating a number of stabilizing interactions in particular through $Tyr502^A$ (in monomer A). $Tyr502^A$ intercalates between $Tyr468^B$ and Phe538 involving extensive π-π stacking interactions to the former and van der Waals interactions to the latter. In addition, $Tyr502^A$ forms a C-H· · ·π interaction to $Asn466^B$ and a hydrogen bond between its OH group and both $Asn466^B$-O and $Ile539^B$-N, as well as a hydrogen bond between its main-chain N and $Asn486^B$-$O_{\delta 1}$. Further interactions include a hydrogen bond from $Asp503^A$ to $Tyr468^B$ and a salt bridge between $Asp503^A$ and $Lys485^B$. In GdPDC this loop is shorter by four amino acids, foregoing all the described stabilizing interactions, possibly contributing to the lower thermal stability of this protein. Interestingly, the situation in GdPDC is similar to that in ApPDC (2VBI), which displays higher thermostability (Table 1).

A second region which is clearly different involves the 11 residues linking the PP- and R-domains in GdPDC (residues Thr341 to Thr352, Figure 4A). This stretch is clearly defined in all three structures, however the

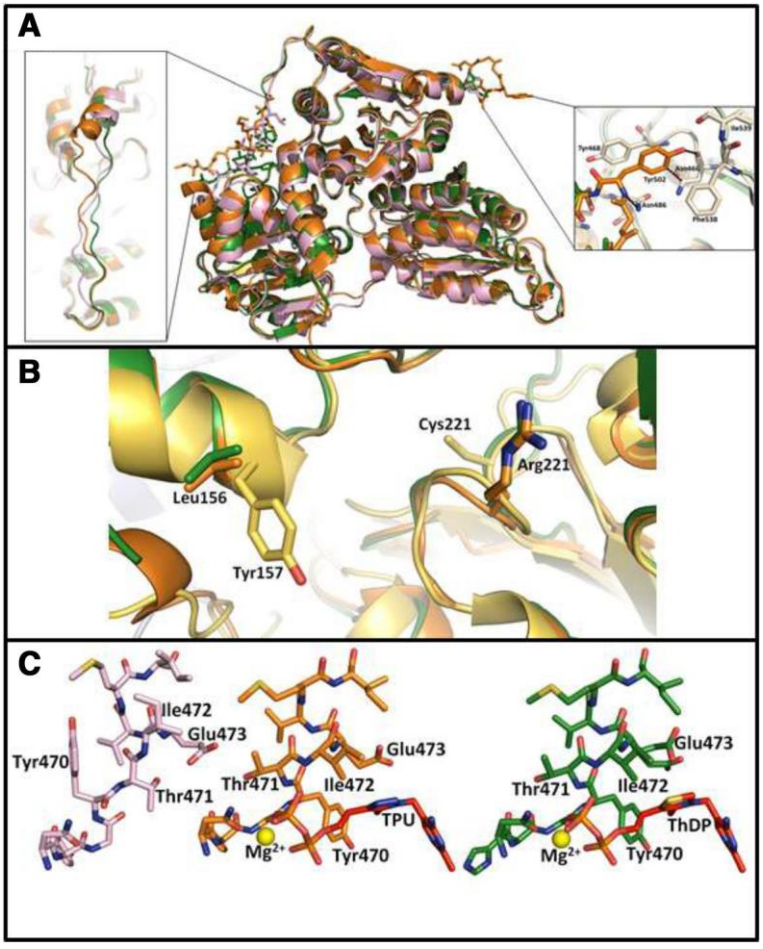


Figure 4 Conformational differences between ZmPDC, ApPDC and GdPDC. A) Superposition of ZmPDC (orange), ApPDC (pink) and GdPDC (green) monomers emphasizing two regions where their conformations differ. Deviating regions are shown as stick models, conserved regions as ribbon diagrams. They extend from T341 to T352 and N499 to D503. Left insert: linker region, right insert: Interactions of Tyr502 B) An alignment of ScPDC (1QPD, yellow), ZmPDC (1ZPD, orange) and GdPDC (4COK, green) showing conserved residues, Arg221, Cys221, Tyr157 and Leu156 on either side of the cleft between PYR and R-domains. Arg221 is conserved but adopts a different conformation in ScPDC compared to ZmPDC and GdPDC. Tyr157 is unique to ScPDC, replaced by Leu156 in the bacterial homologues. Bacterial enzymes lack Cys221, involved in substrate induced allosteric activation in ScPDC C) Conformational change brought about by ThDP cofactor binding. Left to right: apo-ZmPDC (2WVH, pink), ZmPDC, TPU (2WVG, orange) and GdPDC, ThDP (4COK, green).

positioning of this region differs substantially between the three structures implying unique stabilization details in each. The linker can thus potentially affect both enzyme stability and activity, but in a more subtle way.

The linker connecting the R- and PYR-domains of GdPDC (residues 184–191) is not defined in the electron density of both symmetrically independent monomers implying it to be highly disordered. The corresponding residues have therefore not been included in the final model. In crystal structures of ZmPDC and ApPDC these residues are well defined and are stabilized through contacts to other residues in the R- and PYR-domains clearly stabilizing the linker region. Interestingly this seven-residue linker contains three proline residues which likely add rigidity to the region [44]. However, proline has been shown to be one of the preferred amino acids in domain linker regions, and they are thought to structurally isolate the linker from the protein domains as they have no hydrogen bond to donate, perhaps, as in this case, leading to a flexible linker rather than one rigidified by the proline residues [45,46]. Disorder in flexible regions of other PDCs (ScPDC) has been linked to a physiological role, and disorder in linker regions of proteins often indicates a physiological significance [47,48].

Discussion

We have characterized the sixth bacterial PDC, from the acetic acid bacterium *G. diazotrophicus*, and solved its resting state structure. Our analysis indicates the substrate range of the enzyme to be similar to that of

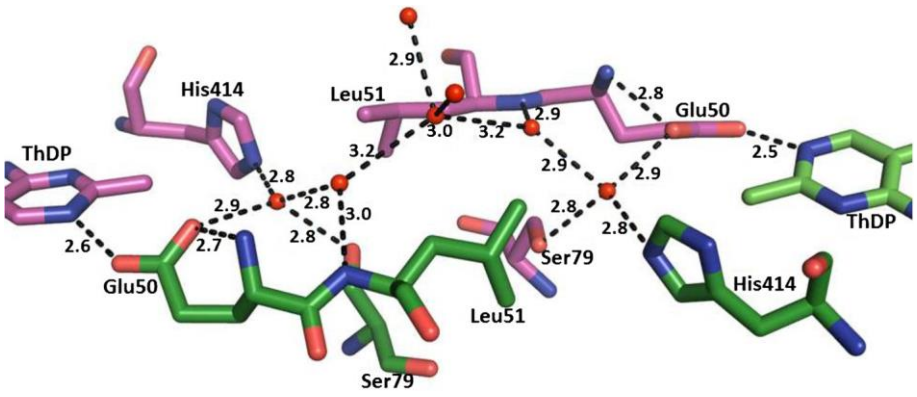


Figure 5 Model of the water tunnel connecting the two active sites in a GdPDC dimer. Water molecules are shown as red spheres. Residues lining the water tunnel are shown as stick models in dark green and labeled. The carbon atoms in the pyrimidine rings of the ThDP cofactor molecules are colored red and nitrogen in blue.

other Gram negative PDCs with regards to substrate recognition and decarboxylation, showing a preference for short-chain aliphatic 2-keto acids [36]. The significantly higher k_{cat}/K_M for pyruvate compared with the nearest analogues 2-ketobutanoate and 2-ketopentanoate, and the retention of Ile468, proposedly crucial for substrate specificity, implies that this enzyme favors pyruvate as its physiological substrate. It can hence be considered a *bona fide* pyruvate decarboxylase [36,37]. Furthermore, as GdPDC does not have any detectable activity on indole-3-pyruvate, it may be ruled out as a contributor to IAA production in *G. diazotrophicus* PAL5.

The pH dependence of K_M and therefore k_{cat}/K_M for this class of enzymes is well documented [5,26,49,50]. GdPDC appears to behave in much the same way as its Gram-negative counterparts in terms of kinetic behavior, displaying the same pH dependence of K_M, with a 20-fold improvement from pH 7 to pH 5, while catalytic efficiency remains largely the same due to only a small change in k_{cat} (2 fold) over the same pH range (Table 1). Although the minimum specific activity for GdPDC with pyruvate as substrate, is nine times lower compared with the maximum specific activity reported for ZmPDC, the lower K_M at pH 5 means that the catalytic efficiency (k_{cat}/K_M) at this pH is comparable to the highest reported values for ZmPDC (Table 1) [38].

A pH optimum of 5.5 for GdPDC (Figure 2A) is similar to those of other bacterial PDCs, and also agrees with the pH optimum for growth of its host [51,52]. *G. diazotrophicus* is an obligate sugarcane endosymbiont which grows optimally at pH 5.5, which is also the pH of sugarcane sap [53]. It seems possible therefore that the GdPDC has evolved to perform best at the physiological pH of the plant sap environment. Whether the *G. diazotrophicus* intracellular pH is similar to that of the sugarcane sap is yet to be determined. However, it has been shown that for other aerobic acetogenic bacteria, such as *Acetobacter aceti*, they are unable to maintain an internal pH above that of its external environment resulting in an acidic intracellular environment [54]. Perhaps a similar scenario is true for *G. diazotrophicus*, applying selective pressure for the PDC to perform at this physiological pH [55]. There are only four other characterized enzymes from *G. diazotrophicus.* One of these is a secreted levansucrase which has an optimal pH at 5, while the other two enzymes, a membrane bound alcohol dehydrogenase has an optimum of 6 and a nitrogenase at pH6 [56]. It has also been shown that plant PDC expression is induced in response to lowered pH caused by oxygen stress [57,58]. In *G. diazotrophicus* the *pdc* is divergently transcribed from a LysR-like regulator with 98 bp between the translational start of both genes, suggesting that *pdc* expression is regulated and is not constitutively expressed. It would therefore be of interest to determine if expression of GdPDC is also pH or oxygen dependent. If *G. diazotrophicus*, however; does not maintain an acidic intracellular environment, then the optimum pH could suggest the possibility that the PDC performs a role outside the bacterial cell in support of plant cell metabolism under oxygen stress.

As discussed, the low K_M for pyruvate at pH 5 suggests that if it functions mainly at or near this pH, GdPDC would be an extremely good pyruvate scavenger under physiological conditions. The structure of GdPDC aligns well to the related PDCs from *A. pasteurianus* and *Z. mobilis* with small rmsd's for Cα positions indicating high structural conservation for these enzymes. The lower thermostability of GdPDC [36] is presumably due to the smaller number of hydrogen bonds and salt bridges between monomers compared to the enzymes from *Z. mobilis* and *A. pasteurianus* [59]. Molecular dynamic studies comparing the structures of the three bacterial PDCs at different temperatures could shed light on the nature of thermostability differences observed [60]. The

enzyme does not exhibit significant biochemical or structural differences to its Gram negative counterparts, and indicates that there may be strong selective pressure to maintain the biochemical and structural properties of these enzymes in a narrow range across the range of microorganisms it has been identified in. Its reduced thermostability and lower T_{opt} likely reflects the physical conditions under which GdPDC has been selected for, resulting from the mesophilic endosymbiotic relationship.

There is obvious biotechnological potential for this class of enzyme in engineering of ethanologenic strains as well as in engineering of transgenic crops capable of surviving adverse conditions [61]. The bacterial enzymes which, apart from the *S. ventriculi* enzyme, are not affected by substrate activation and which have higher thermostabilities and activities compared with their yeast and plant counterparts are particularly attractive. Towards ethanologenesis, the dual function pyruvate ferrodoxin oxidoreductase/pyruvate decarboxylase enzymes from several thermophilic archaea have been described, opening the possibility of using these for thermophilic ethanologenesis. Some of their biochemical characteristics however (low PDC activity, high pH optima and oxygen sensitivity), make them unsuitable for engineering of certain ethanologenic strains that operate under microaerobic conditions (*Geobacillus thermoglucosidasius*) or low temperature (*S. cerevisiae*) [62]. Considering the rarity of true PDCs and their narrow functionality, it seems unlikely that a thermophilic variant exists in nature. We propose that, as with most industrially used enzymes, the ideal PDC can only be generated through engineering, and perhaps these two groups of enzymes represent good starting points.

A picture is emerging that the organisms containing these enzymes are strongly plant associated, in which the environment contains ethanol and a lowered pH; ideal conditions for the PDC to play a key role in metabolism. The rarity of these enzymes therefore appears to be due to the PDC only being of significant metabolic importance in these environments. However, the small range of niches they occupy also puts selective pressure on them to adopt characteristics that fall in a similarly narrow range. *G. diazotrophicus* is an obligate plant endophyte, shown to fix dinitrogen, produce plant growth hormones and protect plants against pathogens such as *Xanthomonas albilineans* [63,64]. It is expected that the role of the PDC enzyme in *G. diazotrophicus* is to convert pyruvate to acetaldehyde. However, the reason for doing so (when and why it's expression is turned on), whether it is part of the central metabolic pathways or selectively expressed under altered physiological states, perhaps in support of its symbiotic host, remains to be determined. The metabolic importance of PDCs in acetic acid bacteria has been described for two of the members from this family, *A. pasteurianus* and *G. oxydans*. In both cases PDC plays an important role in oxidative metabolism [4,16]. The rarity of bacterial PDCs together with their importance in oxidative metabolism in these bacteria, suggests that the enzyme is retained only as a necessity and not as an accessory function. The retention of the enzyme in *G. diazotrophicus* therefore implies importance of the enzyme, however perhaps not in oxidative metabolism. Four proteomic studies looking at global and differential gene expression in *G. diazotrophicus* in pure culture versus when grown in association with sugarcane plantlets did not identify the PDC as an expressed enzyme [65-68]. It could either be that PDC levels are below the detection limit of these experiments, or that the gene is not expressed under the conditions of the experiment (aerobic). It was recently proposed that acetic acid bacteria, although being described as obligate aerobic organisms, have the molecular machinery (ubiquinol oxidases) to enable them to thrive under microaerobic conditions [69]. Although speculative, should the *G. diazotrophicus* PDC be shown to further help plants cope with oxygen stress, by operating in a fermentative manner, this would further deepen the symbiotic relationship between these two organisms to the point where *G. diazotrophicus* could almost be considered a "plant organelle".

Conclusions

Understanding the various roles that pyruvate decarboxylases play in their hosts is of importance not only from a fundamental biology point of view, but as is the case with *G. diazotrophicus*, perhaps also of economic importance. Here we show the enzyme from *G. diazotrophicus* is very similar to those from other Gram negative bacterial hosts, however what role it plays in this host remains to be elucidated. This study opens the door to further exploration of the role the enzyme plays in its host as well as contributing to our knowledge of these rare enzymes.

Methods

Media, bacterial strains and plasmids

Bacterial strains and plasmids used in this study are listed in Table 3. *E. coli* strains were grown in Lysogeny broth (LB) with either ampicillin (200 μg/ml) or kanamycin (50 μg/ml) as required. *G. diazotrophicus* was cultured in medium containing, per liter: 5 g yeast extract, 3 g peptone, 25 g mannitol. All reagents were purchased from Merck. Cultures were incubated at 30°C.

DNA manipulations and sequencing

Plasmid preparation, restriction endonuclease digestion, gel electrophoresis, ligation and Southern/colony blot hybridization were performed using standard methods or manufacturers' recommendations [70]. Ultrapure plasmid DNA was obtained using the Wizard Plus SV miniprep DNA purification system (Promega™). Total DNA from all

Table 3 Bacterial strains, plasmids and primers used in this study

Strain or plasmid	Genotype or description	Source or reference
Strains		
G. diazotrophicus ATCC 49037	Wild type strain PAI 5	American type culture collection
E. coli DH5α	F′/*endA1 hsdR17* ($r_K^- m_K^+$) *supE44 thi-1 reacA1 gyrA* (Nal^r) *relA1* Δ*(lacZYA-argF)U169 (φ80dlac*Δ*(lacZ)M15)*	Promega Corp.
E. coli BL21-DE3	*E. coli* B F$^-$ *dcm ompT hsd*S($r_B^- m_B^-$) *gal* phage Lambda(DE3)	Invitrogen Corp.
Plasmids		
pGEM-T	Ap^r; T-tailed PCR product cloning vector	Promega Corp.
pET17b	Ap^r; ColE1 replicon, HIS-tag expression vector	Novagen Corp.
pET28a	Kan^r; ColE1 replicon, HIS-tag expression vector	Novagen Corp.
pGD	Kan^r; ColE1 replicon; *G. diazatrophicus pdc* gene cloned into pET28a	This study
Primers		
GDPDCpETF	5'-GGAATTC*CATATG*ACCTATACCGTTGGACG-3'	This study
GDPDCpETR	5'-CCG*CTCGAG*TCAGCCCGCGCGCGGC-3'	This study
GDPDCseq	5'-ATCGACGCGCTGCTGAGCCC-3'	This study
T7 promoter	5'-TAATACGACTCACTATAGGG-3'	Promega Corp.
T7 terminator	5'-GCTAGTTATTGCTCAGCGG-3'	Promega Corp.

Italics sections in primer sequences indicate restriction endonuclease sites.

bacterial strains was prepared as described [71]. The QIAGEN plasmid midi kit was used for large-scale plasmid preparations. DNA was sequenced using an ABI Prism 377 automated DNA sequencer and sequences were analyzed with DNAMAN (version 4.1, Lynnon BioSoft). Full length PDC protein sequences were aligned using the full alignment feature of DNAMAN, and the neighbor-joining tree [72] constructed using MEGA6 [73].

Polymerase chain reaction (PCR)

PCR amplifications were performed using KAPA2G Robust DNA polymerase (KAPA BIOSYSTEMS™). Generally, 50 ng DNA were used in a 50 μl reaction volume containing 2 mM $MgCl_2$, 0.125 μM of each primer, 0.2 mM of each deoxynucleoside triphosphate, and 1 U DNA polymerase. Reactions were carried out in a Hybaid Sprint thermocycler, with initial denaturation for 60 s at 94°C, followed by 30 cycles of denaturation (30 s, 94°C), annealing (30 s) and variable elongation (72°C), where annealing temperatures and elongation times were adjusted as required. Primers are also listed in Table 3.

Cloning of the *G. diazotrophicus pdc*

The *pdc* gene from *G. diazotrophicus* (Genbank accession number: KJ746104) was identified by BLASTn search of the genome of this species, using the *Z. mobilis pdc* sequence as a comparator. Primers were designed for its amplification, amplified using Robust DNA polymerase (no 3'-5' exonuclease activity), and cloned into pGEM-T Easy (Promega). To generate an error-free construct, two fragments from two different clones were subcloned into pET17b to reconstruct the original gene. Briefly, the 5' 1320 bp *Nde*I-*Pvu*II fragment, and the 3' 357 bp *Pvu*II-*Xho*I fragment were cloned into pET17b separately, using the *Spe*I (sites in pGEM-T Easy and pET17b) and *Pvu*II (sites in the gene, position 1320 bp, and in pET17b) to clone the 5' fragment into pET17b. The 3' ~560 bp *Pvu*II-*Pvu*II (second *Pvu*II site from pGEM-T Easy vector) fragment was cloned into the pET17b construct using the sole *Pvu*II site. The correct orientation was confirmed by restriction digest with *Pvu*I. The gene was subcloned in pET28a using the *Nde*I and *Xho*I sites, resulting in construct pGD. The final sequence was confirmed as representative of the original gene using primers specific to the T7 promoter, T7 terminator and an internal primer (GDPDCseq).

Purification of PDC protein

An overnight culture of pGD in *E. coli* BL21-DE3 with kanamycin (50 μg/ml) was used to inoculate fresh LB (1% transfer) and incubated overnight at room temperature with aeration (120 rpm) to produce GdPDC without IPTG induction. The cells were collected by centrifugation (3000 × g for 10 min) and lysed with BugBuster™. The suspension was incubated at room temperature for 20 min with shaking. After cell debris removal by centrifugation (7840 × g, 20 min), DNaseI and RNaseA (Fermentas) were added (10 U/ml) to reduce lysate viscosity and the solution incubated at room temperature with shaking for 30 min. HisBind™ resin and buffer kit (Novagen) were used to purify the protein. After elution with 9 ml of 250 mM imidazole buffer (1 M imidazole, 0.5 M NaCl, 20 mM Tris−HCl pH 7.9), the protein was dialyzed against 200 volumes of 200 mM sodium citrate pH 6.0, 1 mM ThDP and 1 mM $MgCl_2$.

The purity was estimated by reducing SDS-PAGE gel (12%) and protein concentrations determined using Bradford reagent (Bio-Rad) with bovine serum albumin as the standard ([74]; Figure 1).

Crystallization and structure determination

Following Ni-NTA/His_6-tag affinity chromatography purification the protein was concentrated to ±4 mg/ml by ultrafiltration using a Vivaspin 20 column (Sartorius). Crystals grew at 25°C without further additives. For cryoprotection 30% (v/v) glycerol was added. X-ray diffraction data was collected at beamline Proxima 1, Soleil Synchrotron, St. Aubin, France at 100 K. Indexing, space group assignment and data integration were performed using iMosflm [75], while data were scaled and merged using SCALA [76]. All further data manipulations were performed using the CCP4 package [77]. MOLREP [78] was used for molecular replacement using 2VBI as molecular model. REFMAC5 was used for structure refinement [79], Coot for graphical model building [80], WHATIF for model validation [81] and PyMOL for molecular depictions (Delano Scientific). The align feature in PyMol was used for structure alignments. The root mean square deviation (rmsd) between two models is calculated using $((\Sigma(d_{ii})^2)/N)^{1/2}$, where d_{ii} is the distance between the i^{th} atom of structure 1 and the i^{th} atom of structure 2, and N is the number of matched atoms. The interface area was calculated and residues in monomer-monomer interfaces identified using the PDBePISA online server (http://tinyurl.com/35w8z7). PDB code 4cok has been assigned to the structure.

Steady state kinetic analysis and determination of substrate range

PDC activity was measured using a coupled assay with baker's yeast ADH (Sigma-Aldrich) as described previously [82]. The reaction mixture (1 ml final volume) contained 0.25 mM NADH, 5 mM $MgCl_2$, 0.1 mM ThDP, 5 mM pyruvate (unless stated otherwise) and 10 U of ADH in 50 mM MES or 200 mM Na citrate buffers, pH 6.4 or 6.0 respectively. For substrate range determination, ADH was replaced with 1 U/ml baker's yeast aldehyde dehydrogenase (ALDH, Sigma-Aldrich) when testing 2-ketobutanoate, 2-ketopentanoate, 2-keto-4-methylpentanoate, 3-phenyl-2-oxopropanoate. β-mercaptoethanol was added to a final concentration of 3 mM and NADH replaced with NAD^+. Assays were performed in 100 mM citric acid/K_2HPO_4 buffer, pH 7 [83]. Activities were recorded at 25°C unless otherwise indicated, using a Cary 50 temperature controlled spectrophotometer (Varian). To determine enzyme activity for benzoyl formate, 3-hydroxy-phenyl pyruvate and indole-3-pyruvate HPLC assays were employed. Reactions were run on a Hypersil Gold C18 250 × 4.6 mm (Thermo Scientific) on a Dionex Ultimate 3000 machine, using 30% MeOH/1% Acetic acid mobile phase as mobile phase under isocratic elution (1 ml/min, 40°C). Twenty μl of each sample was injected by autosampler and the components detected using either a refractive index detector or a UV/Vis photodiode array at 245 nm. To generate kinetic data, initial enzyme velocities were determined over the substrate range 0.1 mM to 30 mM for pyruvate or 24 mM for other 2-keto acids. Kinetic parameters were determined by non-linear data fitting to hyperbolic curves (GraphPad Prism v. 4.00, GraphPad Software, San Diego, CA, USA). k_{cat} values were calculated based on the MW of the tetramer (240 kDa) with four active site.

Availability of supporting data

Supporting data are included as Additional file 1: Figure S1, Additional file 2: Figure S2 and Additional file 3: Figure S3.

Additional files

Additional file 1: Figure S1. Multiple sequence alignment of selected PDC protein sequences generated using DNAman (Lynnon BioSoft). GdiPDC - *G. diazotrophicus* (KJ746104); GoxPDC - *G. oxydans* (KF650839); ApaPDC *Acetobacter pasteurianus* (AF368435.1); ZpaPDC - *Z. palmae* (AF474145); ZmoPDC - *Z. mobilis* (AB359063); ZmaPDC - *Z. mays* (X17555); ScePDC - *S. cerevisiae* (X04675); SvePDC - *S. ventriculi* (AF354297); *Lyngbya aestuarii* (WP023067698); *Acidomonas methanolica* (GAJ29946); *Acetobacter pomorum* (WP006115789); *Acetobacter aceti* (WP010667855); *Microcystis aeruginosa* (WP_0027648); *Moorea producens* (WP008180762); *Microbulbifer variabilis* (WP020414286); *Legionella pneumophila* (YP006505162); MDM (CBI10829); *Ktedonobacter racemifer* (WP007922190); *Komagataeibacter oboediens* (WP010515737); *Komagataeibacter hansenii* (WP003622049); *Komagataeibacter europaeus* (WP010509054); *Granulicella tundricola* (YP004210504); *Gluconobacter thailandicus* (WP007283613); *Gluconobacter morbifer* (WP008852112); *Gluconobacter frateurii* (WP023941876); *Gluconacetobacter xylinus* (AHI26557); *Gluconacetobacter medellinensis* (YP004868149); *Fluoribacter dumoffii* (WP010654974); *Enterobacter cloacae* iPDC (P23234); *Commensalibacter intestini* (WP008853550); *Beijerinckia indica* (YP001834435); *Pseudomonas putida* BFD (YP008115845); MDM-Mine Drainage Metagenome (CBI10829.1). Residues shaded in black are conserved, those in dark grey to 75%, and those in light grey to 50%. The conserved ThDP-binding motif is marked by a solid line, ThDP binding residues by triangles, Mg^{2+}-binding residues by arrows, catalytic pocket residues probably involved in catalysis by circles. An asterisk indicates Ile468 involved in substrate specificity, while a star highlights Ile472 proposed to be involved in substrate positioning. Two squares mark Arg221 located at the same position as Cys221 ScePDC and SvePDC involved in substrate activation.

Additional file 2: Figure S2. A denaturing SDS-PAGE gel showing purified GdiPDC. Lane 1, Molecular weight marker (Fermentas), Lane 2, Ni-NTA purified GdiPDC-His_6 fusion protein. GdiPDC has a mass of ~59 kDa but runs at a slightly smaller size.

Additional file 3: Figure S3. Orthorhombic crystals of GdiPDC. The scale bar indicates 50 μm.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

LVZ performed cloning, purification of the protein, crystallization, sequence alignment, phylogenetic tree construction. WDS collected X-ray data. WDS and LVZ solved and refined the crystal structure. DAC and MIT conceived the study and participated in its design and coordination. All authors participated in preparing the final manuscript. All authors read and approved the final manuscript.

Acknowledgement
This work was supported by the National Research Foundation of South Africa.

Author details
[1]Institute for Microbial Biotechnology and Metagenomics (IMBM), University of the Western Cape, Robert Sobukwe Road, Bellville, Cape Town, South Africa. [2]Department of Biochemistry, University of Pretoria, 2 Lynnwood Road, Pretoria 0002, South Africa. [3]Department of Genetics, University of Pretoria, Pretoria 0002, South Africa.

Received: 8 July 2014 Accepted: 25 September 2014
Published online: 05 November 2014

References

1. Lee K, Jeong C, An YJ, Lee H, Park S, Seok Y, Kim P, Lee J, Lee K, Cha S: FrsA functions as a cofactor-independent decarboxylase to control metabolic flux. *Nat Chem Biol* 2011, 7:434–436.
2. Kellett WF, Brunk E, Desai BJ, Fedorov AA, Almo SC, Gerlt JA, Rothlisberger U, Richards NGJ: Computational, Structural, and Kinetic Evidence That *Vibrio vulnificus* FrsA Is Not a Cofactor-Independent Pyruvate Decarboxylase. *Biochem* 2013, 52:1842–1844.
3. King TE, Cheldelin VH: Pyruvic carboxylase of *Acetobacter suboxydans*. *J Biol Chem* 1954, 208:821–831.
4. Raj KC, Ingram LO, Maupin-Furlow JA: Pyruvate decarboxylase: a key enzyme for the oxidative metabolism of lactic acid by *Acetobacter pasteurianus*. *Arch Microbiol* 2001, 176:443–451.
5. Raj KC, Talarico LA, Ingram LO, Maupin-Furlow JA: Cloning and characterization of the *Zymobacter palmae* pyruvate decarboxylase gene (pdc) and comparison to bacterial homologues. *Appl Environ Microbiol* 2002, 686:2869–2876.
6. Lowe SE, Zeikus JG: Purification and characterization of pyruvate decarboxylase from *Sarcina ventriculi*. *J Gen Microbiol* 1992, 138:803–807.
7. Van Zyl LJ, Taylor MP, Eley K, Tuffin M, Cowan DA: Engineering pyruvate decarboxylase-mediated ethanol production in the thermophilic host *Geobacillus thermoglucosidasius*. *Appl Microbiol Biotechnol* 2013, 98:1247–1259.
8. Candy JM, Duggleby RG: Structure and properties of pyruvate decarboxylase and site directed mutagenesis of the *Zymomonas mobilis* enzyme. *Biochim Biophys Acta* 1998, 1385:323–338.
9. Gold RS, Meagher MM, Tong S, Hutkins RW, Conway T: Cloning and expression of the *Zymomonas mobilis* "production of ethanol" genes in *Lactobacillus casei*. *Curr Microbiol* 1996, 33:256–260.
10. Bongers RS, Hoefnagel MH, Kleerebezem M: High-level acetaldehyde production in *Lactococcus lactis* by metabolic engineering. *Appl Environ Microbiol* 2005, 71:1109–1113.
11. Kaczowka SJ, Reuter CJ, Talarico LA, Maupin-Furlow JA: Recombinant production of *Zymomonas mobilis* pyruvate decarboxylase in the haloarchaeon *Haloferax volcanii*. *Archaea* 2005, 15:327–334.
12. Liu S, Dien BS, Cotta MA: Functional expression of bacterial *Zymobacter palmae* pyruvate decarboxylase gene in *Lactococcus lactis*. *Curr Microbiol* 2005, 50:324–328.
13. Talarico LA, Gil MA, Yomano LP, Ingram LO, Maupin-Furlow JA: Construction and expression of an ethanol production operon in Gram-positive bacteria. *Microbiol* 2005, 151:4023–4031.
14. Liu S, Dien BS, Nichols NN, Bischoff KM, Hughes SR, Cotta MA: Coexpression of pyruvate decarboxylase and alcohol dehydrogenase genes in *Lactobacillus brevis*. *FEMS Microbiol Lett* 2007, 274:291–297.
15. Talarico LA, Ingram LO, Maupin-Furlow JA: Production of the Gram-positive *Sarcina ventriculi* pyruvate decarboxylase in *Escherichia coli*. *Microbiol* 2001, 147:2425–2435.
16. Peters B, Junker A, Brauer K, Mühlthaler B, Kostner D, Mientus M, Liebl W, Ehrenreich A: Deletion of pyruvate decarboxylase by a new method for efficient markerless gene deletions in *Gluconobacter oxydans*. *Appl Microbiol Biotechnol* 2012, 97:2521–2530.
17. Meyer D, Neumann P, Ficner R, Tittmann K: Observation of a stable carbene at the active site of a thiamin enzyme. *Nat Chem Biol* 2013, 9:488–490.
18. Kern D, Kern G, Neef H, Tittmann K, Killenberg-Jabs M, Wikner C, Schneider G, Hubner G: How thiamine diphosphate is activated in enzymes. *Science* 1997, 275:67–70.
19. Zhang S, Liu M, Yan Y, Zhang Z, Jordan F: C2-alpha-lactylthiamin diphosphate is an intermediate on the pathway of thiamin diphosphate-dependent pyruvate decarboxylation. Evidence on enzymes and models. *J Biol Chem* 2004, 279:54312–54318.
20. Baykal AT, Kakalis L, Jordan F: Electronic and nuclear magnetic resonance spectroscopic features of the 1',4'-iminopyrimidine tautomeric form of thiamin diphosphate, a novel intermediate on enzymes requiring this coenzyme. *Biochem* 2006, 4524:7522–7528.
21. Lie MA, Celik L, Jorgensen KA, Schiott B: Cofactor activation and substrate binding in pyruvate decarboxylase. Insights into the reaction mechanism from molecular dynamics simulations. *Biochem* 2005, 4445:14792–14806.
22. Pei XY, Erixon KM, Luisi BF, Leeper FJ: Structural insights into the prereaction state of pyruvate decarboxylase from *Zymomonas mobilis*. *Biochem* 2010, 498:1727–1736.
23. Brandt GS, Kneen MM, Chakraborty S, Baykal AT, Nemeria N, Yep A, Ruby DI, Petsko GA, Kenyon GL, McLeish MJ, Jordan F, Ringe D: Snapshot of a reaction intermediate: analysis of benzoylformate decarboxylase in complex with a benzoylphosphonate inhibitor. *Biochem* 2009, 4815:3247–3257.
24. Chakraborty S, Nemeria NS, Balakrishnan A, Brandt GS, Kneen MM, Yep A, McLeish MJ, Kenyon GL, Petsko GA, Ringe D, Jordan F: Detection and time course of formation of major thiamin diphosphate-bound covalent intermediates derived from a chromophoric substrate analogue on benzoylformate decarboxylase. *Biochem* 2009, 485:981–994.
25. Nemeria NS, Chakraborty S, Balakrishnan A, Jordan F: Reaction mechanisms of thiamin diphosphate enzymes: defining states of ionization and tautomerization of the cofactor at individual steps. *FEBS J* 2009, 276:2432–2446.
26. Meyer D, Neumann P, Parthier C, Friedemann R, Nemeria N, Jordan F, Tittmann K: Double duty for a conserved glutamate in pyruvate decarboxylase: evidence of the participation in stereoelectronically controlled decarboxylation and in protonation of the nascent carbanion/enamine intermediate. *Biochem* 2010, 49:8197–8212.
27. Dobritzsch D, Konig S, Schneider G, Lu G: High resolution crystal structure of pyruvate decarboxylase from *Zymomonas mobilis*. Implications for substrate activation in pyruvate decarboxylases. *J Biol Chem* 1998, 273:20196–20204.
28. Rother D, Kolter G, Gerhards T, Berthold CL, Gauchenova E, Knoll M, Pleiss J, Müller M, Schneider G, Pohl M: *S*-Selective mixed carboligation by structure-based design of the pyruvate decarboxylase from *Acetobacter pasteurianus*. *Chem Cat Chem* 2011, 10:1587–1596.
29. Lee S, Flores-Encarnacion M, Contreras-Zentella M, Garcia-Flores L, Escamilla JE, Kennedy C: Indole-3-acetic acid biosynthesis is deficient in *Gluconacetobacter diazotrophicus* strains with mutations in cytochrome c biogenesis genes. *J Bacteriol* 2004, 186:5384–5391.
30. Saravanan VS, Madhaiyan M, Osborne J, Thangaraju M, Sa TM: Ecological occurrence of *Gluconacetobacter diazotrophicus* and nitrogen-fixing *Acetobacteraceae* members: their possible role in plant growth promotion. *Microb Ecol* 2008, 55:130–140.
31. Bertalan M, Albano R, de Padua V, Rouws L, Rojas C, Hemerly A, Teixeira K, Schwab S, Araujo J, Oliveira A, Franca L, Magalhaes V, Alqueres S, Cardoso A, Almeida W, Loureiro MM, Nogueira E, Cidade D, Oliveira D, Simao T, Macedo J, Valadao A, Dreschsel M, Freitas F, Vidal M, Guedes H, Rodrigues E, Meneses C, Brioso P, Pozzer L, *et al*: Complete genome sequence of the sugarcane nitrogen-fixing endophyte *Gluconacetobacter diazotrophicus* Pal5. *BMC Genomics* 2009, 10:450.
32. Felsenstein J: Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 1985, 39:783–791.
33. Zuckerkandl E, Pauling L: Evolutionary divergence and convergence in proteins. In *Evolving Genes and Proteins*. Edited by Bryson V, Vogel HJ. New York: Academic Press; 1965:97–166.
34. Chang AK, Nixon PF, Duggleby RG: Effects of deletions at the carboxyl terminus of *Zymomonas mobilis* pyruvate decarboxylase on the kinetic properties and substrate specificity. *Biochem* 2000, 3931:9430–9437.
35. König S, Spinka M, Kutter S: Allosteric activation of pyruvate decarboxylases. A never-ending story? *J Mol Cat B: Enzymatic* 2009, 61:100–110.
36. Gocke D, Graf T, Brosi H, Frindi-Wosch I, Walter L, Müller M, Pohl M: Comparative characterization of thiamin diphosphate-dependent decarboxylases. *J Mol Cat B: Enzymatic* 2009, 61:30–35.
37. Siegert P, McLeish MJ, Baumann M, Iding H, Kneen MM, Kenyon GL, Pohl M: Exchanging the substrate specificities of pyruvate decarboxylase from Zymomonas mobilis and benzoylformate decarboxylase from Pseudomonas putida. *Protein Eng Des Sel* 2005, 187:345–357.

38. Bringer-Meyer S, Schimz K-L, Sahm H: Pyruvate decarboxylase from *Zymomonas mobilis*. Isolation and partial characterization. *Arch Microbiol* 1986, 146:105–110.
39. Pohl M, Mesch K, Rodenbrock A, Kula MR: Stability investigations on the pyruvate decarboxylase from *Zymomonas mobilis*. *Biotechnol Appl Biochem* 1995, 22:95–105.
40. Lu G, Dobritzsch D, Baumann S, Schneider G, Konig S: The structural basis of substrate activation in yeast pyruvate decarboxylase. A crystallographic and kinetic study. *Eur J Biochem* 2000, 267:861–868.
41. Arjunan P, Umland T, Dyda F, Swaminathan S, Furey W, Sax M, Farrenkopf B, Gao Y, Zhang D, Jordan F: Crystal structure of the thiamin diphosphate-dependent enzyme pyruvate decarboxylase from the yeast *Saccharomyces cerevisiae* at 2.3 A resolution. *J Mol Biol* 1996, 2563:590–600.
42. Lu G, Dobritzsch D, Konig S, Schneider G: Novel tetramer assembly of pyruvate decarboxylase from brewer's yeast observed in a new crystal form. *FEBS Lett* 1997, 403:249–253.
43. Frank RA, Titman CM, Pratap JV, Luisi BF, Perham RN: A molecular switch and proton wire synchronize the active sites in thiamine enzymes. *Science* 2004, 306:872–876.
44. Couturier M, Roussel A, Rosengren A, Leone P, Stålbrand H, Berrin J-G: Structural and Biochemical Analyses of Glycoside Hydrolase Families 5 and 26 β-(1,4)-Mannanases from *Podospora anserina* Reveal Differences upon Manno-oligosaccharide Catalysis. *J Biol Chem* 2013, 288:14624–14635.
45. George RA, Heringa J: An analysis of protein domain linkers: their classification and role in protein folding. *Prot Eng* 2003, 15:871–879.
46. Poon DKY, Withers SG, McIntosh LP: Direct Demonstration of the Flexibility of the Glycosylated Proline-Threonine Linker in the *Cellulomonas fimi* Xylanase Cex through NMR Spectroscopic Analysis. *J Biol Chem* 2006, 282:2091–2100.
47. Gokhale RS, Khosla C: Role of linkers in communication between protein modules. *Curr Opin Chem Biol* 2000, 4:22–27.
48. Quigley PM, Korotkov K, Baneyx F, Hol WGJ: A new native EcHsp31 structure suggests a key role of structural flexibility for chaperone function. *Protein Sci* 2004, 13:269–277.
49. Schenk G, Leeper FJ, England R, Nixon PF, Duggleby RG: The role of His113 and His114 in pyruvate decarboxylase from *Zymomonas mobilis*. *Eur J Biochem* 1997, 248:63–71.
50. Huang CY, Chang AK, Nixon PF, Duggleby RG: Site-directed mutagenesis of the ionizable groups in the active site of *Zymomonas mobilis* pyruvate decarboxylase: effect on activity and pH dependence. *Eur J Biochem* 2001, 268:3558–3565.
51. Eskin N, Vessey K, Tian L: Research Progress and Perspectives of Nitrogen Fixing Bacterium, *Gluconacetobacter diazotrophicus*, in Monocot Plants. *Int J Agron* 2014, doi:10.1155/2014/208383.
52. James EK, Olivares FL, de Oliveira ALM, dos Reis FB Jr, Aa Silva LG, Reis VM: Further observations on the interaction between sugar cane and *Gluconacetobacter diazotophicus* under laboratory and greenhouse conditions. *J Exp Bot* 2001, 52:747–760.
53. Dong Z, Canny MJ, McCully ME, Roboredo MR, Cabadilla CF, Ortega E, Rodés R: A Nitrogen-Fixing Endophyte of Sugarcane Stems A New Role for the Apoplast. *Plant Physiol* 1994, 105:1139–1147.
54. Menzel U, Gottschalk G: The internal pH of *Acetobacterium wieringae* and *Acetobater aceti* during growth and production of acetic acid. *Arch Microbiol* 1985,143:47–51.
55. Talley K, Alexov E: On the pH-optimum of activity and stability of proteins. *Proteins* 2010, 78:2699–2706.
56. Gomez-Manzo S, Contreras-Zentella M, Gonzalez-Valdez A, Sosa-Torres M, Arreguin-Espinoza R, Escamilla-Marvan E: The PQQ-alcohol dehydrogenase of *Gluconacetobacter diazotrophicus*. *Int J Food Microbiol* 2008, 125:71–78.
57. Kelley PM: Maize PDC mRNA is induced anaerobically. *Plant Mol Biol* 1989, 13:213–222.
58. Mithran M, Paparelli E, Novi G, Perata P, Loreti E: Analysis of the role of the pyruvate decarboxylase gene family in *Arabidopsis thaliana* under low-oxygen conditions. *Plant Biol* 2013, 16:28–34.
59. Bjork A, Dalhus B, Mantzilas D, Sirevag R, Eijsink VG: Large improvement in the thermal stability of a tetrameric malate dehydrogenase by single point mutations at the dimer-dimer interface. *J Mol Biol* 2004, 3415:1215–1226.
60. Paul M, Hazra M, Barman A, Hazra S: Comparative molecular dynamics simulation studies for determining factors contributing to the thermostability of chemotaxis protein "CheY". *J Biomol Struct Dyn* 2013, doi:10.1080/07391102.2013.799438.
61. Tadege M, Brändle R, Kuhlemeier C: Anoxia tolerance in tobacco roots: effect of overexpression of pyruvate decarboxylase. *Plant J* 1998, 14:327–335.
62. Eram MS, Oduaran E, Ma K: The bifunctional pyruvate decarboxylase/pyruvate ferrodoxin oxidoreductase from *Thermococcus guaymasensis*. *Archaea* 2014, doi:10.1155/2014/349379.
63. Arencibia AD, Vinagre F, Estevez Y, Bernal A, Perez J, Cavalcanti J, Santana I, Hemerly AS: *Gluconacetobacter diazotrophicus* Elicits a Sugarcane Defense Response Against a Pathogenic Bacteria *Xanthomonas albilineans*. *Plant Signal Behav* 2006, 1:265–273.
64. Reis VM, Olivares FL, Döbereiner J: Improved methodology for isolation of *Acetobacter diazotrophicus* and confirmation of its endophytic habitat. *World J Microbiol Biotechnol* 1994, 10:401–405.
65. dos Santosa MF, de Páduac VLM, de Matos NE, Hemerlyd AS, Domont GB: Proteome of *Gluconacetobacter diazotrophicus* co-cultivated with sugarcane plantlets. *Proteomics* 2010, 73:917–931.
66. Lery LMS, Hemerly AS, Nogueira EM, von Krüger WMA, Bisch PM: Quantitative Proteomic Analysis of the Interaction Between the Endophytic Plant-Growth-Promoting Bacterium *Gluconacetobacter diazotrophicus* and Sugarcane. *MPMI* 2011, 24:562–576.
67. Lery LMS, von Krüger WMA, Viana FC, Teixeira KRS, Bisch PM: A comparative proteomic analysis of Gluconacetobacter diazotrophicus PAL5 at exponential and stationary phases of cultures in the presence of high and low levels of inorganic nitrogen compound. *Biochim Biophys Acta* 2008, 1784:1578–1589.
68. Lery LMS, Coelho A, von Kruger WMA, Goncalves MSM, Santos MF, Valente RH, Santos EO, Rocha SLG, Perales J, Domont GB, Teixeira KRS, Bisch PM: Protein expression profile of *Gluconacetobacter diazotrophicus* PAL5, a sugarcane endophytic plant growth-promoting bacterium. *Proteomics* 2008, 8:1631–1644.
69. Chouaia B, Gaiarsa S, Crotti E, Comandatore F, Esposti MD, Ricci I, Alma A, Favia G, Bandi C, Daffonchio D: Acetic Acid Bacteria Genomes Reveal Functional Traits for Adaptation to Life in Insect Guts. *Genome Biol Evol* 2014, 6:912–920.
70. Sambrook J, Fritsch EF, Maniatis T: *Molecular Cloning: A Laboratory Manual.* NY: Cold Spring Harbour Laboratory Press; 1989.
71. Kotze AA, Tuffin IM, Deane SM, Rawlings DE: Cloning and characterization of the chromosomal arsenic resistance genes from *Acidithiobacillus caldus* and enhanced arsenic resistance on conjugal transfer of ars genes located on transposon TnAtcArs. *Microbiology* 2006, 12:3551–3560.
72. Saitou N, Nei M: The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987, 4:406–425.
73. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* 2013, 30:2725–2729.
74. Laemmli UK: Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* 1970, 227:680–685.
75. Leslie AG: The integration of macromolecular diffraction data. *Acta Crystallogr D Biol Crystallogr* 2006, 1:48–57.
76. Evans P: Scaling and assessment of data quality. *Acta Crystallogr D Biol Crystallogr* 2006, 1:72–82.
77. Potterton E, Briggs P, Turkenburg M, Dodson E: A graphical user interface to the CCP4 program suite. *Acta Crystallogr D Biol Crystallogr* 2003, 7:1131–1137.
78. Vagin A, Teplyakov A: MOLREP: an Automated Program for Molecular Replacement. *J Appl Cryst* 1997, 30:1022–1025.
79. Murshudov GN, Pavol Skubák P, Lebedev AA, Pannu NS, Steiner RA, Nicholls RA, Winn MD, Long F, Vagin AA: REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D Biol Crystallogr* 2011, 67:355–367.
80. Emsley P, Lohkamp B, Scott WG, Cowtan K: Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* 2010, 66:486–501.
81. Vriend G: WHAT IF: a molecular modeling and drug design program. *J Mol Graph* 1990, 8:52–56.
82. Holzer H, Schultz G, Villar-Palasi C, Juntgen-Sell J: Isolierungder Hefecarboxylase und Untersuchung uher die Aktivität des Enzyms in lebenden Zellen. *Biochem Z* 1956, 327:331–344.
83. Vuralhan Z, Luttik MA, Tai SL, Boer VM, Morais MA, Schipper D, Almering MJ, Kotter P, Dickinson JR, Daran JM, Pronk JT: Physiological characterization of the ARO10-dependent, broad-substrate-specificity 2-oxo acid decarboxylase activity of *Saccharomyces cerevisiae*. *Appl Environ Microbiol* 2005, 71:3276–3284.

doi:10.1186/s12900-014-0021-1
Cite this article as: van Zyl *et al.*: Structure and functional characterization of pyruvate decarboxylase from *Gluconacetobacter diazotrophicus*. *BMC Structural Biology* 2014 14:21.

Permission to reproduce the article here:

Dear Dr. van Zyl,

Thank you for contacting BioMed Central.

The open access articles published in BioMed Central's journals are made available under the Creative Commons Attribution (CC-BY) license, which means they are accessible online without any restrictions and can be re-used in any way, subject only to proper attribution (which, in an academic context, usually means citation).

The re-use rights enshrined in our license agreement (http://www.biomedcentral.com/about/policies/license-agreement) include the right for anyone to produce printed copies themselves, without formal permission or payment of permission fees. As a courtesy, however, anyone wishing to reproduce large quantities of an open access article (250+) should inform the copyright holder and we suggest a contribution in support of open access publication (see suggested contributions at http://www.biomedcentral.com/about/policies/reprints-and-permissions/suggested-contributions).

Please note that the following journals have published a small number of articles that, while freely accessible, are not open access as outlined above: Alzheimer's Research & Therapy, Arthritis Research & Therapy, Breast Cancer Research, Critical Care, Genome Biology, Genome Medicine, Stem Cell Research & Therapy.

You will be able to find details about these articles at http://www.biomedcentral.com/about/policies/reprints-and-permissions

If you have any questions, please do not hesitate to contact me.

With kind regards,

Ricardo Sison Jr.

Global Open Research Support Executive

Global Open Research Support

Springer Nature

T +44 (0)203 192 2009

www.springernature.com

In the US: Springer Customer Service Center LLC, 233 Spring Street, New York, NY 10013

Registered Address: 2711 Centerville Road Wilmington, DE 19808 USA

State of Incorporation: Delaware, Reg. No. 4538065

Rest of World: Springer Customer Service Center GmbH, Tiergartenstraße 15 – 17, 69121 Heidelberg

Registered Office: Heidelberg | Amtsgericht Mannheim, HRB 336546

Managing Directors: Derk Haank, Martin Mos, Dr. Ulrich Vest

----------Your Question/Comment ----------

To whom it may concern,

I would hereby like to ask permission to publish the following article,

authored by me, as part of my doctoral thesis.

"Structure and functional characterization of pyruvate decarboxylase from

Gluconacetobacter diazotrophicus"

doi.org/10.1186/s12900-014-0021-1

Kind Regards

Lonnie (Leonardo Joaquim)

--

Lonnie van Zyl

Senior Researcher

Institute for Microbial Biotechnology and Metagenomics (IMBM)

University of the Western Cape

Bellville

South Africa

<https://www.avast.com/sig-email?utm_medium=email&utm_source=link&utm_campaign=sig-email&utm_content=webmail>

Virus-free.

www.avast.com

<https://www.avast.com/sig-email?utm_medium=email&utm_source=link&utm_campaign=sig-email&utm_content=webmail>

<#DAB4FAD8-2DD7-40BB-A1B8-4E2AA1F9FDF2>

Correspondence with the journal:

MS: 4079795501348797

Structure and Functional Characterization of a Pyruvate decarboxylase from Gluconacetobacter diazotrophicus

Leonardo J van Zyl, Wolf-Dieter Schubert, Marla I Tuffin and Don A Cowan

BMC Structural Biology

Dear Mr van Zyl

Thank you for your recent submission to BMC Structural Biology. I would like to update you regarding your status with respect to the article processing charge that is normally due if a manuscript is accepted.

Submissions from EU countries are subject to VAT at 20.0%.

There is no charge due as the article processing charge for this article is fully covered by the BioMed Central Membership of University of the Western Cape.

Please note that we will check if you are entitled to this Membership in the next few weeks.We will let you know if University of the Western Cape decline to cover this manuscript.If this happens you will need to cover the whole fee of GBP 1,325.00/USD 2,215.00/EUR 1,600.00.

Kind regards,

BioMed Central Accounts Team

236 Gray's Inn Road

London

WC1X 8HB

Tel: +44 (0) 20 3192 2009

e-mail: payment@biomedcentral.com

Dear Mr van Zyl,

The files for your manuscript have been received by BMC Structural Biology. You will shortly receive a further e-mail that will provide you with links to the PDF that is now being generated and will be used for assessing it. At any time, you may log in to My BioMed Central (http://www.biomedcentral.com/my/manuscripts) to view the status of the manuscript, or you can bookmark the URL of your manuscript:

http://www.biomedcentral.com/author/manuscript/details/view.do?txt_nav=man&txt_man_id=4079795501348797&manuscriptId=4079795501348797

If you have any questions, please visit the BioMed Central Support Center (http://www.biomedcentral.com/support/) or e-mail us at editorial@biomedcentral.com.

Questionnaire

Thank you for submitting your research to BMC Structural Biology. To help us understand what is important to our authors, please take a minute to complete our simple online questionnaire by following this link:

http://www.biomedcentral.com/survey/login/2049214874102165/4079795501348797/AS1/

You may be prompted to log on when you follow this link.

Regards

The BMC Structural Biology Editorial Team

Tel: +44 (0) 20 3192 2013

e-mail: editorial@biomedcentral.com

Web: http://www.biomedcentral.com/

Article title: Structure and Functional Characterization of a Pyruvate decarboxylase from Gluconacetobacter diazotrophicus

MS ID : 4079795501348797

Authors : Leonardo J van Zyl, Wolf-Dieter Schubert, Marla I Tuffin and Don A Cowan

Journal : BMC Structural Biology

Dear Mr van Zyl

Thank you for submitting your article. This acknowledgement and any queries below are for the contact author. This e-mail has also been copied to each author on the paper, as well as the person submitting. Please bear in mind that all queries regarding the paper should be made through the contact author.

A pdf file has been generated from your submitted manuscript and figures. We would be most grateful if you could check this file and let us know if any aspect is missing or incorrect. Any additional files you uploaded will also be sent in their original format for review.

http://www.biomedcentral.com/imedia/4079795501348797_article.pdf (4238K)

For your records, please find below link(s) to the correspondence you uploaded with this submission. Please note there may be a short delay in creating this file.

http://www.biomedcentral.com/imedia/1530924331351397_comment.pdf

If the PDF does not contain the comments which you uploaded, please upload the cover letter again, click "Continue" at the bottom of the page, and then proceed with the manuscript submission again. If the letter will not upload, please send a copy to

editorial@biomedcentral.com.

The submitting author can check on the status of the manuscript at any time by logging into 'My BioMed Central' (http://www.biomedcentral.com/my).

In the meantime, if you have any queries about the manuscript you may contact us on editorial@biomedcentral.com. We would also welcome feedback about the online submission process.

Best wishes,

The BMC Structural Biology Editorial Team

Tel: +44 (0) 20 3192 2013

e-mail: editorial@biomedcentral.com

Web: http://www.biomedcentral.com/

MS: 4079795501348797

Research article

Structure and Functional Characterization of a Pyruvate decarboxylase from Gluconacetobacter diazotrophicus

Leonardo J van Zyl, Wolf-Dieter Schubert, Marla I Tuffin and Don A Cowan

BMC Structural Biology (Section: Crystallography)

Dear Mr. Zyl,

Thank you for your recent submission to BMC Structural Biology. Before we can proceed with peer review we will need you to make some changes to your manuscript. We would be very grateful if you could make these changes promptly, as we cannot start the peer review process until we have received a version containing the changes.

**Editor's requests:

-Please include the email addresses of all co-authors in the title page.

-Abstract:

Please include an abstract as the second page in your manuscript file (directly following the title page). Please format your abstract according to the guidelines for authors http://www.biomedcentral.com/authors/authorfaq/format. Potential referees will be asked to review the manuscript having seen only the title and abstract, so it is important that these are both informative and concise.

-Competing interests:

Please include a 'Competing interests' section between the Conclusions and Authors' contributions. If there are none to declare, please write 'The authors declare that they have no competing interests'. Please consider the following questions and include a declaration of competing interests in your manuscript:

Financial competing interests

• In the past five years have you received reimbursements, fees, funding, or salary from an organization that may in any way gain or lose financially from the publication of this manuscript, either now or in the future? Is such an organization financing this manuscript (including the article-processing charge)? If so, please specify.

• Do you hold any stocks or shares in an organization that may in any way gain or lose financially from the publication of this manuscript, either now or in the future? If so, please specify.

• Do you hold or are you currently applying for any patents relating to the content of the manuscript? Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript? If so, please specify.

• Do you have any other financial competing interests? If so, please specify.

Non-financial competing interests

• Are there any non-financial competing interests (political, personal, religious, ideological, academic, intellectual, commercial or any other) to declare in relation to this manuscript? If so, please specify.

-Authors' contributions:

Please include an Authors' contributions section before the Acknowledgements and Reference list.

For the Authors' contributions we suggest the following kind of format (please use initials to refer to each author's contribution): AB carried out the molecular genetic studies, participated in the sequence alignment and drafted the manuscript. JY carried out the immunoassays. MT participated in the sequence alignment. ES participated in the design of the study and performed the statistical analysis. FG conceived of the study, and participated in its design and coordination. All authors read and approved the final manuscript.

An "author" is generally considered to be someone who has made substantive intellectual contributions to a published study. To qualify as an author one should 1) have made substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data; 2) have been involved in drafting the manuscript or revising it critically for important intellectual content; and 3) have given final approval of the version to be published. Each author should have participated sufficiently in the work to take public responsibility for appropriate portions of the content. Acquisition of funding, collection of data, or general supervision of the research group, alone, does not justify authorship.

All contributors who do not meet the criteria for authorship should be listed in an acknowledgements section. Examples of those who might be acknowledged include a person who provided purely technical help, writing assistance, or a department chair who provided only general support.

-Acknowledgements:

We strongly encourage you to include an Acknowledgements section between the Authors' contributions section and Reference list. Please acknowledge anyone who contributed towards the study by making substantial contributions to conception, design, acquisition of data, or analysis and interpretation of data, or who was involved in drafting the manuscript or revising it critically for important intellectual content, but who does not meet the criteria for authorship. Please also include their source(s) of funding. Please also acknowledge anyone who contributed materials essential for the study.

Authors should obtain permission to acknowledge from all those mentioned in the Acknowledgements. Please list the source(s) of funding for the study, for each author, and for the manuscript preparation in the acknowledgements section. Authors must describe the role of the funding body, if any, in study design; in the collection, analysis, and interpretation of data; in the writing of the manuscript; and in the decision to submit the manuscript for publication.

Once you have made these changes, please upload the revised version by following these instructions:

1. Go to 'my BioMed Central' (http://www.biomedcentral.com/my/) and log on with your email address and password. Then click on 'My manuscripts'. You will reach a page giving details of all the manuscripts you have submitted.

2. Click on the 'Submit revision' button next to the title of this manuscript.

3. In the 'Manuscript details' tab, please update the title, abstract and author details if they have changed since the previous version.

4. In the 'Cover letter' tab, please provide a covering letter with a point-by-point description of the changes made.

5. In the 'Upload files' tab, please upload the revised version of the manuscript and press 'Submit new version'. Please wait for the confirmation page to appear - this may take a few moments.

We would be grateful if you could resubmit your files with the changes requested within the next week.

Please do not hesitate to contact us if you have any questions.

Best regards,

Wella Valenzuela

Journal Editorial Office

Biomed Central

E: Editorial@Biomedcentral.com

W: www.biomedcentral.com

Dear Mr van Zyl,

The files for your revised manuscript have been received by BMC Structural Biology. You will shortly receive a further e-mail that will provide you with links to the PDF that is now being generated and will be used for assessing it and, if appropriate, for further peer review. At any time, you may log in to My BioMed Central (http://www.biomedcentral.com/my/manuscripts) to view the status of the manuscript, or you can bookmark the URL of your manuscript:

http://www.biomedcentral.com/author/manuscript/details/view.do?txt_nav=man&txt_man_id=4079795501348797&manuscriptId=4079795501348797

If you have any questions, please visit the BioMed Central Support Center (http://www.biomedcentral.com/support/) or e-mail us at editorial@biomedcentral.com.

Regards

Ms Wella Valenzuela

Tel: +44 (0) 20 3192 2013

e-mail: editorial@biomedcentral.com

Web: http://www.biomedcentral.com/

Article title: Structure and Functional Characterization of a Pyruvate decarboxylase from Gluconacetobacter diazotrophicus

MS ID : 4079795501348797

Authors : Leonardo J van Zyl, Wolf-Dieter Schubert, Marla I Tuffin and Don A Cowan

Journal : BMC Structural Biology

Dear Mr van Zyl

Thank you for submitting a new version of your article.

A pdf file has been generated from your submitted manuscript and figures.

http://www.biomedcentral.com/imedia/4079795501348797_article.pdf (4244K)

For your records, please find below link(s) to the correspondence you uploaded with this submission. Please note there may be a short delay in creating this file.

If the PDF does not contain the comments which you uploaded, please upload the cover letter again, click "Continue" at the bottom of the page, and then proceed with the manuscript submission again. If the letter will not upload, please send a copy to

editorial@biomedcentral.com.

Best wishes,

Ms Wella Valenzuela

Tel: +44 (0) 20 3192 2013

e-mail: editorial@biomedcentral.com

Web: http://www.biomedcentral.com/

MS: 4079795501348797

Structure and Functional Characterization of a Pyruvate decarboxylase from Gluconacetobacter diazotrophicus

Leonardo J van Zyl, Wolf-Dieter Schubert, Marla I Tuffin and Don A Cowan

Dear Mr van Zyl,

Your manuscript has now been peer reviewed and the comments are accessible in PDF format from the links below. Do let us know if you have any problems opening the files.

Referee 2:

http://www.biomedcentral.com/imedia/5535794851379615_comment.pdf

Referee 1:

http://www.biomedcentral.com/imedia/4514156081379452_comment.pdf

**Editor's comments:

Due to the intrinsic value of the work presented I think the manuscript should be accepted after the authors have put in major revision Key among these is the reformatting of the manuscript in the proper format for this journal. Additionally, the comments from both reviewers should be addressed and details about how each comment was responded to must be returned prior to acceptance.

We would be grateful if you could address the comments in a revised manuscript and provide a cover letter giving a point-by-point response to the concerns.

Please also ensure that your revised manuscript conforms to the journal style (http://www.biomedcentral.com/info/ifora/biology_journals). It is important that your files are correctly formatted.

We look forward to receiving your revised manuscript by 3 September 2014. If you imagine that it will take longer to prepare please give us some estimate of when we can expect it.

You should upload your cover letter and revised manuscript through http://www.biomedcentral.com/manuscript/login/man.asp?txt_nav=man&txt_man_id=4079795501348797. You will find more detailed instructions at the base of this email.

Please don't hesitate to contact me if you have any problems or questions regarding your manuscript.

With best wishes,

Ms Wella Valenzuela

on behalf of Dr Oluwatoyin Asojo

e-mail: editorial@biomedcentral.com

Web: http://www.biomedcentral.com/

To submit your revised manuscript

When you have revised your manuscript in light of the reviewers' comments and made any required changes to the format of your paper, please upload the revised version by following these instructions:

1. Go to http://www.biomedcentral.com/manuscript/login/man.asp?txt_nav=man&txt_man_id=4079795501348797 and log on with your email address and password.

2. With the 'Manuscript details' tab, please update the title, abstract and author details if they have changed since the previous version. It is very important that all changes are updated on this page, as well as in the manuscript file as the information on this page will be used in PubMed and on BioMed Central if your manuscript is accepted for publication.

3. With the 'Cover letter' tab, please provide a covering letter with a point-by-point description of the changes made.

4. With the 'Upload files' tab, please upload the revised version of the manuscript and press 'Submit new version'. Please wait for the confirmation page to appear - this may take a few moments.

Article title: Structure and Functional Characterization of a Pyruvate decarboxylase from Gluconacetobacter diazotrophicus

MS ID : 4079795501348797

Authors : Leonardo J van Zyl, Wolf-Dieter Schubert, Marla I Tuffin and Don A Cowan

Journal : BMC Structural Biology

Dear Mr van Zyl

Thank you for submitting a new version of your article.

A pdf file has been generated from your submitted manuscript and figures.

http://www.biomedcentral.com/imedia/4079795501348797_article.pdf (5660K)

For your records, please find below link(s) to the correspondence you uploaded with this submission. Please note there may be a short delay in creating this file.

http://www.biomedcentral.com/imedia/2028332108141460_comment.pdf

http://www.biomedcentral.com/imedia/1017730520141460_comment.pdf

If the PDF does not contain the comments which you uploaded, please upload the cover letter again, click "Continue" at the bottom of the page, and then proceed with the manuscript submission again. If the letter will not upload, please send a copy to

editorial@biomedcentral.com.

Best wishes,

Ms Wella Valenzuela

on behalf of Dr Oluwatoyin Asojo

Tel: +44 (0) 20 3192 2013

e-mail: editorial@biomedcentral.com

Web: http://www.biomedcentral.com/

Dear Mr van Zyl,

The files for your revised manuscript have been received by BMC Structural Biology. You will shortly receive a further e-mail that will provide you with links to the PDF that is now being generated and will be used for assessing it and, if appropriate, for further peer review. At any time, you may log in to My BioMed Central (http://www.biomedcentral.com/my/manuscripts) to view the status of the manuscript, or you can bookmark the URL of your manuscript:

http://www.biomedcentral.com/author/manuscript/details/view.do?txt_nav=man&txt_man_id=4079795501348797&manuscriptId=4079795501348797

If you have any questions, please visit the BioMed Central Support Center (http://www.biomedcentral.com/support/) or e-mail us at editorial@biomedcentral.com.

Regards

Ms Wella Valenzuela

on behalf of Dr Oluwatoyin Asojo

Tel: +44 (0) 20 3192 2013

e-mail: editorial@biomedcentral.com

Web: http://www.biomedcentral.com/

Authors: Leonardo J van Zyl Wolf-Dieter Schubert Marla I Tuffin and Don A Cowan

Title : Structure and Functional Characterization of a Pyruvate decarboxylase from Gluconacetobacter diazotrophicus

Journal: BMC Structural Biology

MS : 4079795501348797

Dear Mr van Zyl,

Peer review of your manuscript (above) is now complete and we are delighted to accept the manuscript for publication in BMC Structural Biology.

Before publication our production team needs to check the format of your manuscript to ensure that it conforms to the standards of the journal. They will get in touch with you shortly to request any necessary changes or to confirm that none are needed.

If you have any problems or questions regarding your manuscript please do get in touch.

Best wishes

Ciara Ní Dhubhghaill, PhD

on behalf of Dr Oluwatoyin Asojo

Tel: +44 (0) 20 3192 2013

e-mail: editorial@biomedcentral.com

Web: http://www.biomedcentral.com/

Author list: Leonardo J van Zyl, Wolf-Dieter Schubert, Marla I Tuffin, Don A Cowan

Title: Structure and functional characterization of pyruvate decarboxylase from Gluconacetobacter diazotrophicus

Journal: BMC Structural Biology

MS ID: 4079795501348797

JWF MS ID: 12900_2014_21

Dear Leonardo J van Zyl,

We are pleased to inform you that a final PDF proof of your article is now available for you to check. The PDF can be found at the following link:

http://biomedcentral.spi-global.com/authorproofs/bmcproofs/index.php?id=lrKTxoctit11052014142647SFmaUFyllp

Please note that this URL is for proofing purposes only and may not be used by third parties.

Please return any corrections within two calendar days via the online submission system, or by email to: bmcproductionteam2@spi-global.com.

Please note that we cannot proceed with publication of your article until we have received a response to this email. Please therefore respond even if you have no corrections to make.

The proof shows the final PDF as it will appear when published, except that:

- the lines are numbered to make it easy to reference any passage that needs to be corrected; and
- an “Uncorrected Proof” watermark is stamped on each page.

Please refer to the proofing procedure which is displayed on the first page of the article. Please note that this will be your final opportunity to make changes to your article. We will not be able to make any further corrections after the manuscript is published. In case of difficulties please contact us at: bmcproductionteam2@spi-global.com.

Thank you very much.

With respect and warm regards,

BioMed Central Production Team 2

Author list: Leonardo J van Zyl, Wolf-Dieter Schubert, Marla I Tuffin, Don A Cowan

Title: Structure and functional characterization of pyruvate decarboxylase from Gluconacetobacter diazotrophicus

Journal: BMC Structural Biology

MS ID: 4079795501348797

JWF MS ID: 12900_2014_21

Dear Author,

Please confirm if you receive our inquiry below for us to proceed

with your article.

Please note that we cannot proceed with publication of your article until we have received a response to this email.

Please return your revisions to us by replying to this email by November 11, 2014.

Thank you very much.

With respect and warm regards,

BioMed Central Production Team 2

From: BioMed Central Production Team 2

To: Leonardo J van Zyl

Cc:

Subject: Final PDF proof: BMC Structural Biology, 12900_2014_21

MIME-Version: 1.0

This is a multi-part message in MIME format.

--172.20.145.178.1.22068.1415168807.261.59

--172.20.145.178.1.22068.1415168807.261.60

Content-Type: text/html

Author list: Leonardo J van Zyl, Wolf-Dieter Schubert,

Marla I Tuffin, Don A Cowan

Title: Structure and functional characterization of pyruvate
decarboxylase from Gluconacetobacter diazotrophicus

Journal: BMC Structural Biology

MS ID: 4079795501348797

JWF MS ID: 12900_2014_21

Dear Leonardo J van Zyl,

We are pleased to inform you that a final PDF proof of your article is
now available for you to check. The PDF can be found at the following
link:

http://biomedcentral.spi-global.com/authorproofs/bmcproofs/index.php?id=lrKTxoctit11052014142647SFmaUFyllp

Please note that this URL is for proofing purposes only and may not be
used by third parties.

Please return any corrections within two calendar days via the online
submission system, or by email to:

bmcproductionteam2@spi-global.com.

Please note that we cannot proceed with publication of your article

until we have received a response to this email. Please therefore respond even if you have no corrections to make.

The proof shows the final PDF as it will appear when published, except that:

- the lines are numbered to make it easy to reference any passage that needs to be corrected; and
- an “Uncorrected Proof” watermark is stamped on each page.

Please refer to the proofing procedure which is displayed on the first page of the article. Please note that this will be your final opportunity to make changes to your article. We will not be able to make any further corrections after the manuscript is published. In case of difficulties please contact us at:

bmcproductionteam2@spi-global.com.

Thank you very much.

With respect and warm regards,

BioMed Central Production Team 2

--172.20.145.178.1.22068.1415168807.261.59--

Thank you very much.

With respect and warm regards,

BioMed Central Production Team 2

Dear BMC production team,

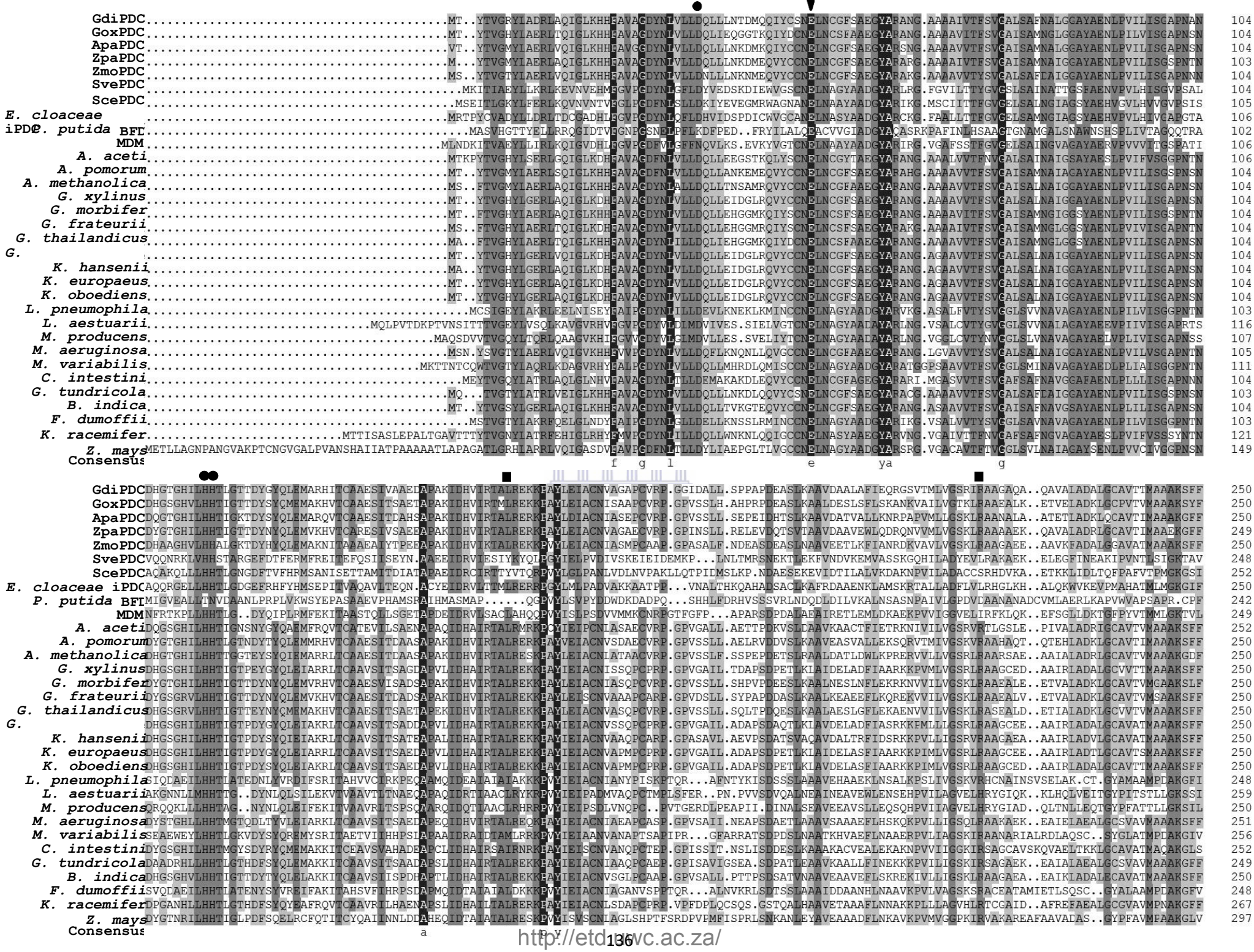
I have replied to this e-mail previously with an edited copy of the manuscript stating that we would like to have two changes made. Line 90 - I would like to have "organisms" changed to "organism" and line 354 the "cat" in kcat needs to be subscript.

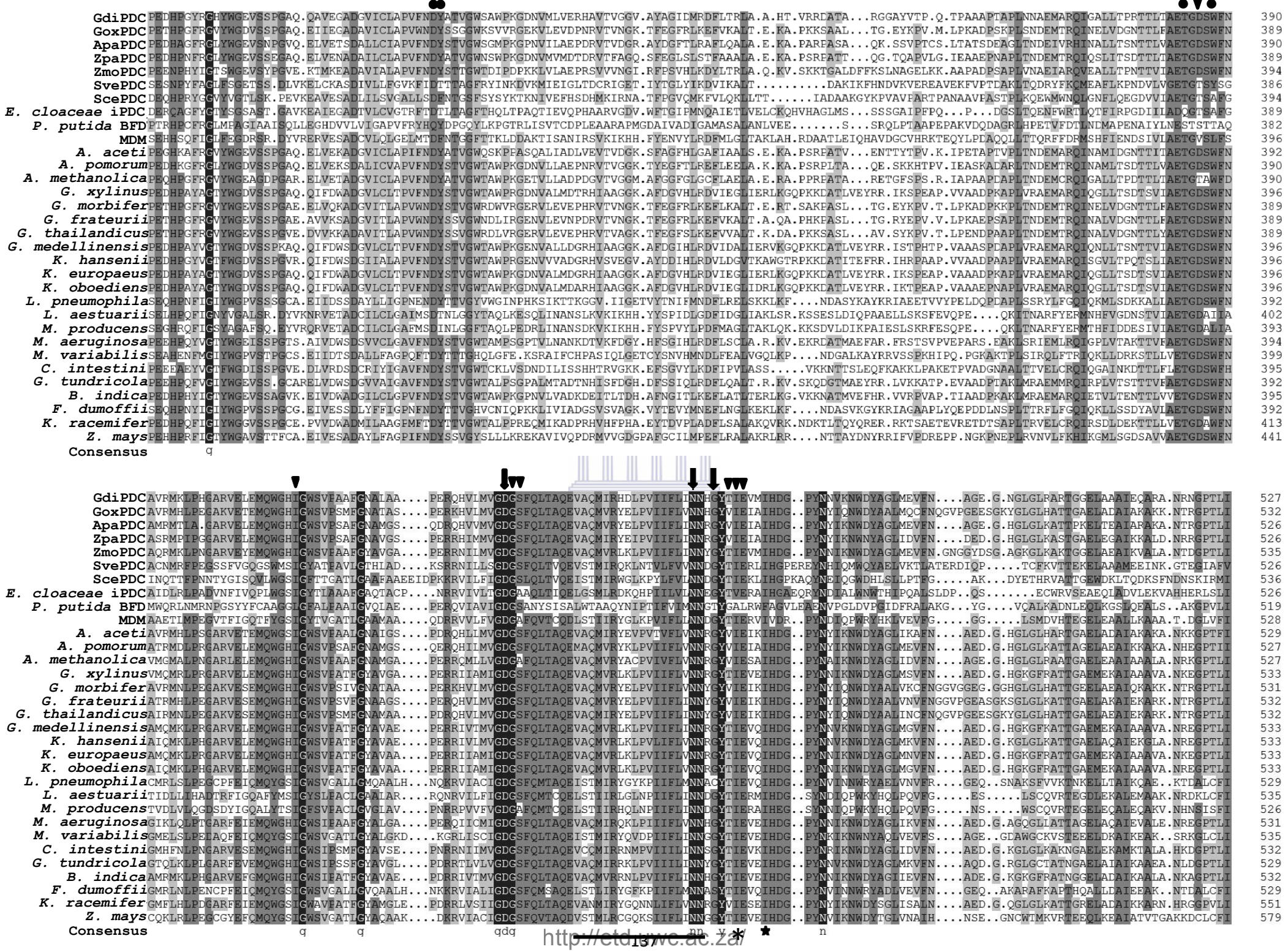
Also there is a query on the author query form asking if the link works. The link does work.

Please find another copy attached to this e-mail, and if possible, send me an indication that this e-mail has reached the relevant individuals.

Kind regards

Appendix:





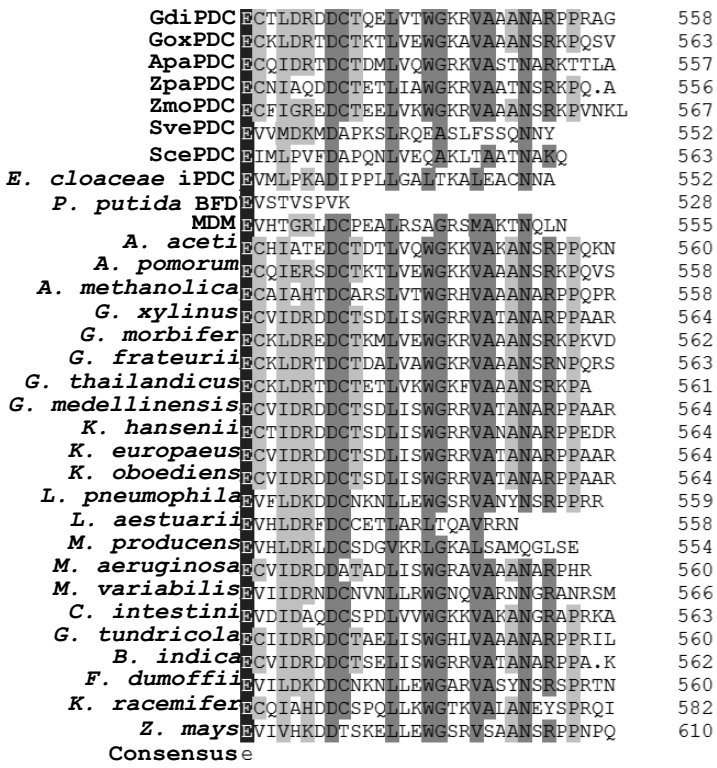


Figure S1: Multiple sequence alignment of selected PDC protein sequences generated using DNAman (Lynnon BioSoft). GdiPDC - *G. diazotrophicus* (KJ746104); GoxPDC - *G. oxydans* (KF650839); ApaPDC *Acetobacter pasteurianus* (AF368435.1); ZpaPDC - *Z. palmae* (AF474145); ZmoPDC - *Z. mobilis* (AB359063); ZmaPDC - *Z. mays* (X17555); ScePDC - *S. cerevisiae* (X04675); SvePDC - *S. ventriculi* (AF354297); *Lyngbya aestuarii* (WP023067698); *Acidomonas methanolica* (GAJ29946); *Acetobacter pomorum* (WP006115789); *Acetobacter aceti* (WP010667855); *Microcystis aeruginosa* (WP_0027648); *Moorea producens* (WP008180762); *Microbulbifer variabilis* (WP020414286); *Legionella pneumophila* (YP006505162); MDM (CBI10829); *Ktedonobacter racemifer* (WP007922190); *Komagataeibacter oboediens* (WP010515737); *Komagataeibacter hansenii* (WP003622049); *Komagataeibacter europaeus* (WP010509054); *Granulicella tundricola* (YP004210504); *Gluconobacter thailandicus* (WP007283613); *Gluconobacter morbifer* (WP008852112); *Gluconobacter frateurii* (WP023941876); *Gluconacetobacter xylinus* (AHI26557); *Gluconacetobacter medellinensis* (YP004868149); *Fluoribacter dumoffii* (WP010654974); *Enterobacter cloacae* iPDC (P23234); *Commensalibacter intestini* (WP008853550); *Beijerinckia indica* (YP001834435); *Pseudomonas putida* BFD (YP008115845); MDM- Mine Drainage Metagenome (CBI10829.1). Residues shaded in black are conserved, those in dark grey to 75%, and those in light grey to 50%. The conserved ThDP-binding motif is marked by a solid line, ThDP binding residues by triangles, Mg^{2+}-binding residues by arrows, catalytic pocket residues probably involved in catalysis by circles. An asterisk indicates Ile468 involved in substrate specificity, while a star highlights Ile472 proposed to be involved in substrate positioning. Two squares mark Arg221 located at the same position as Cys221 ScePDC and SvePDC involved in substrate activation.

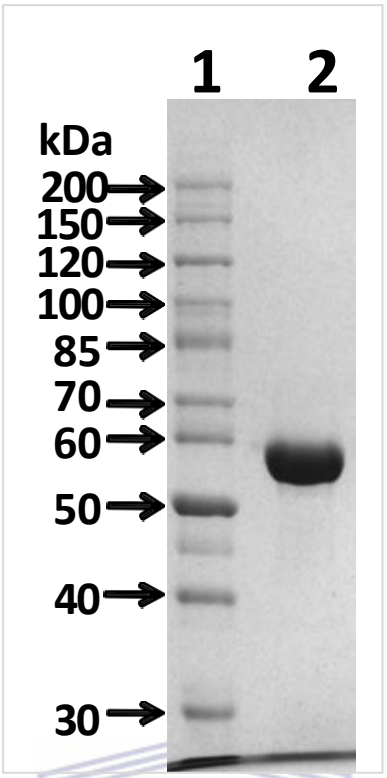


Figure S2: A denaturing SDS-PAGE gel showing purified GdiPDC. Lane 1, Molecular weight marker (Fermentas), Lane 2, Ni-NTA purified GdiPDC-His_6 fusion protein. GdiPDC has a mass of ~59 kDa but runs at a slightly smaller size.

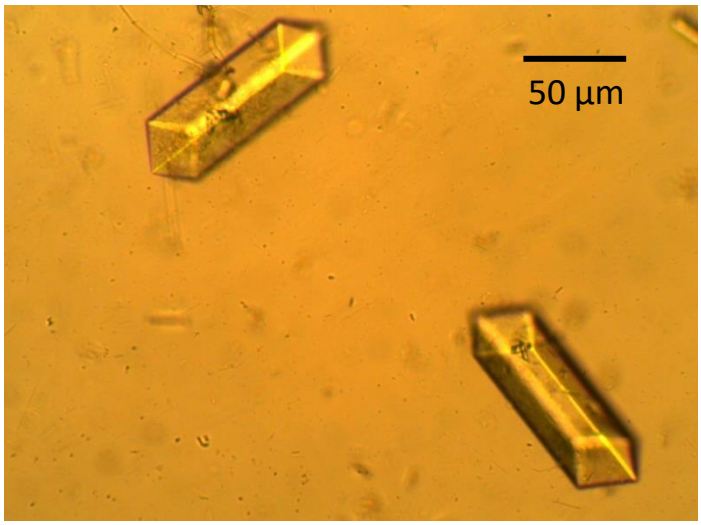


Figure S3: Orthorhombic crystals of GdiPDC. The scale bar indicates 50 µm.

Chapter 4

Author contributions

Marla Trindade, Don Cowan and Mark Taylor conceived the study and participated in its design and coordination. Leonardo Joaquim van Zyl performed all experiments and analysis except for phage proteome experiments performed by Falone Sunda. Leonardo Joaquim van Zyl wrote the bulk of the manuscript. All authors read and approved the final manuscript.

Arch Virol
DOI 10.1007/s00705-015-2497-9

ORIGINAL ARTICLE

Identification and characterization of a novel *Geobacillus thermoglucosidasius* bacteriophage, GVE3

Leonardo Joaquim van Zyl[1] · Falone Sunda[1] · Mark Paul Taylor[2] · Don Arthur Cowan[1,3] · Marla Iris Trindade[1]

Received: 11 March 2015 / Accepted: 12 June 2015

Abstract The study of extremophilic phages may reveal new phage families as well as different mechanisms of infection, propagation and lysis to those found in phages from temperate environments. We describe a novel siphovirus, GVE3, which infects the thermophile *Geobacillus thermoglucosidasius*. The genome size is 141,298 bp (G?C 29.6 %), making it the largest *Geobacillus* spp-infecting phage known. GVE3 appears to be most closely related to the recently described *Bacillus anthracis* phage vB_BanS_Tsamsa, rather than *Geobacillus*-infecting phages described thus far. Tetranucleotide usage deviation analysis supports this relationship, showing that the GVE3 genome sequence correlates best with *B. anthracis* and *Bacillus cereus* genome sequences, rather than *Geobacillus* spp genome sequences.

Electronic supplementary material The online version of this article (doi:10.1007/s00705-015-2497-9) contains supplementary material, which is available to authorized users.

& Leonardo Joaquim van Zyl
vanzyllj@gmail.com

Falone Sunda
falone.sunda@gmail.com

Mark Paul Taylor
marktaylorimbm@gmail.com

Don Arthur Cowan
don.cowan@up.ac.za

Marla Iris Trindade
prof.marlatt@gmail.com

[1] Institute for Microbial Biotechnology and Metagenomics (IMBM), University of the Western Cape, Robert Sobukwe Road, Bellville, Cape Town, South Africa

[2] TMO Renewables Limited, 40 Alan Turing Road, The Surrey Research Park, Guildford, Surrey GU2 7YF, UK

[3] Centre for Microbial Ecology and Genomics, Department of Genetics, University of Pretoria, Pretoria 0002, South Africa

Introduction

The ubiquity of bacteriophages (phages) in nature and their impact on various trophic levels is widely appreciated [58, 76]. As phages directly affect microbial communities that play a pivotal role in biogeochemical cycles, they in turn play a role in altering those cycles [18, 33, 74]. Phages are also known to be prevalent in many extreme environments, including soda lakes, terrestrial hot springs, deep-sea hydrothermal vents, hot/cold deserts and hypersaline systems, with some of the highest phage numbers being recorded in these habitats [40]. However, few studies have investigated the functional relationships between extremophiles and the phages that infect them, compared to the wealth of data that exist for phages and hosts in temperate environments.

Morphological and sequence-based characterization of phages from many temperate environments has shown the predominance of tailed viruses (order *Caudovirales*) with members of the families *Siphoviridae*, *Myoviridae* and *Podoviridae* most often recorded [3, 4, 68, 71]. Morphological characterization of extremophilic phages has led to the introduction of several new families, including *Lipothrixviridae*, *Rudiviridae* and *Fuselloviridae* [6]. The study of extremophilic phages has also revealed new mechanisms for host lysis, as in the case of the deep-sea thermophilic bacteriophage GVE2 [17], and have demonstrated interactions between phage and host proteins that are unlike those normally observed for mesophilic phages [32]. *Thermus thermophilus* phage /YS40 promoters are

thought to be leaderless (i.e., contain no -10 or -35 elements), unlike those found in T4 and many other mesophilic phages, which require phage- and host-encoded sigma factors for transcription [70].

It is therefore likely that the further study of phages infecting extremophiles will reveal new phage families and alternate strategies for infection or the ''decision'' between lysis and lysogeny and will shed further light on the behaviour and the role of host organisms in their natural environments [44, 57, 62]. Extremophilic phages may also provide a source of novel enzymes that are adapted to extreme conditions and serve as the basis for the development of genetic systems by providing strong regulatable promoters, and as vehicles for the introduction of large DNA segments into bacterial hosts for which no genetic tools currently exist [53, 59].

Geobacillus thermoglucosidasius is a Gram-positive thermophile that has been isolated from soil, oil fields, compost heaps, deep-sea sediment and hot springs [54, 67]. This promising ''platform'' organism is capable of producing a range of useful metabolites, including ethanol, isobutanol and polylactic acid [19, 42, 79; http://tinyurl.com/po6a52q]. Several *Geobacillus* species phages have been described (GVE1, GVE2, GBSV1, GBK2, DE6 and / OH2), sequenced and studied [20, 33, 43, 45, 72, 73, 82–84], although none infecting *G. thermoglucosidasius* have been reported. Here, we describe the first phage (GVE3) known to specifically infect *G. thermoglucosidasius*.

Materials and methods

Media, bacterial strains and plasmids

G. thermoglucosidasius strains were cultured in tryptone glycerol pyruvate (TGP) medium. One liter of TGP broth contains 17 g tryptone, 3 g soy peptone, 2.5 g K_2HPO_4 and 5 g NaCl. The pH was adjusted to 7.3 before autoclaving, after which 4 g Na-pyruvate and 4 mL glycerol (filter sterilized) were added. For solid media, 15 g/L agar was added before autoclaving. TGP was used for general maintenance of cultures. Cultures were incubated at 60 °C with vigorous aeration.

DNA manipulations and sequencing

Plasmid preparations, restriction endonuclease digestions, gel electrophoresis and ligations were performed using standard methods or following the manufacturers' recommendations. Total DNA from all bacterial strains was prepared as described [34]. Phage DNA was prepared by first preparing a phage lysate from 1 L of culture as described below. The phage was pelleted by centrifugation at 13000 9 *g* for 30 min after addition of PEG8000 (7.5 ml of 20% PEG8000 per 30 ml lysate) and incubation at 4 °C overnight. The pellet was resuspended in 1 ml of SM buffer (5.8 g of NaCl per liter, 1.2 g of $MgSO_4$ per liter, 50 mL of 1 M Tris-HCl, pH 7.5, 0.1 g of gelatin per liter). The suspension was treated with DNaseI and RNaseA (Fermentas; final concentration, 0.1 1g/ml) at 37 °C for 1 hour. The presence of contaminating bacterial DNA was tested by amplifying the 16S rRNA gene. The suspension was treated with proteinase K (Fermentas; final concentration, 1 1g/ml) at 55 °C for 2 hours before addition of 70 1l of 20 % (wt/vol) SDS and incubation at 37 °C for 1 hour. An equal volume of phenol:chloroform:isoamylalcohol (P:C:I; 25:24:1) was added, the sample was centrifuged (15 ml Sterillin tube, Eppendorf 5810R centrifuge, 5000 rpm for 10 min) to separate the phases, and the top, aqueous phase was removed and transferred to a fresh tube. A second P:C:I extraction was performed. An equal volume of C:I (24:1) was added to the supernatant and re-centrifuged. The top phase was removed and transferred to a fresh tube, and a tenth volume of 3 M sodium acetate (pH 5.2) and two volumes of 100 % ethanol were added. This mixture was incubated at 4 °C to precipitate overnight. The sample was centrifuged at 13,000 rpm for ten minutes to pellet the DNA, and the pellet was resuspended in 40 1l of TE buffer. The phage DNA was electrophoresed on a 1 % low-melting-point agarose gel, excised and purified from the gel using standard agarase (Fermentas) treatment. The pellet was resuspended in 40 1l of TE buffer. The quality and integrity of the DNA was checked using a Bioanalyzer prior to library preparation. Sanger DNA sequencing was performed using an ABI PRISM 377 automated DNA sequencer (University of Stellenbosch Central Analytical Facility), and next-generation sequencing was performed using either a Roche GS Junior with a LibL library preparation kit or an Illumina MiSeq with a Nextera XT 150 bp library kit (Illumina). The raw reads were trimmed and de-multiplexed at the sequencing facility (the University of the Western Cape Next Generation Sequencing facility), resulting in two (2 9 150) paired fastq files. Sequences were analyzed with DNAMAN (version 4.1, Lynnon BioSoft), Newbler (Roche) or CLC Genomics Workbench version 6.5 (CLC Bio). Open reading frames were predicted using the built-in tools in the CLC Genomics workbench and confirmed by BLASTp search against the NCBInr database. Smaller ORFs not identified by the software were assigned through manual translation of DNA sequences and BLASTp analysis of putative ORFs [5]. The complete genome sequence of *G. thermoglucosidasius* bacteriophage GVE3 is available in the GenBank database under accession no. KP144388. RAST [http://rast.nmpdr.org/; 7] and PHAST (http://phast.wishartlab.com/) [87] were used to identify closely related

phages. RADAR was used to identify protein repeat regions (http://www.ebi.ac.uk/Tools/pfa/radar/). Direct repeats were identified using REPFIND [http://zlab.bu.edu/repfind/form.html; 9] with a 15-bp minimum repeat length. Inverted repeats were identified using UGENE (http://ugene.unipro.ru/) with a 20-bp minimum and 80 % similarity as search parameters. tRNA genes were predicted using the tRNAscan-SE program [http://lowelab.ucsc.edu/tRNAscan-SE/; 46] and ARAGORN [http://mbio-serv2.mbioekol.lu.se/ARAGORN/; 39]. Transmembrane regions were predicted using the TMHMM server v2.0 [http://www.cbs.dtu.dk/services/TMHMM/; 36]. Intron prediction was done using the RNAweasel server [http://megasun.bch.umontreal.ca/RNAweasel/; 38]. For phylogenetic tree construction, the full-length amino acid sequences of selected terminase proteins were aligned using MEGA6, and the tree was constructed using the built-in program [24, 88].

Polymerase chain reaction

Polymerase chain reaction (PCR) was performed using Phusion DNA polymerase (New England Biolabs. Generally, 50 ng of DNA was used in a 50-µl reaction volume containing 2 mM $MgCl_2$, 0.125 µM each primer, 0.2 mM each deoxynucleoside triphosphate, and 1 U of DNA polymerase. Reactions were carried out in a Bio-Rad T-100 thermocycler, with an initial denaturation at 98 °C for 3 min, followed by 30 cycles of denaturation (30 s at 98 °C), annealing (30 s), and variable elongation times at 72 °C as required.

Phage purification, maintenance and characterization

Phage lysates were prepared by culturing *G. thermoglucosidasius* to an OD_{600nm} of 0.4 and addition of phage particles at a multiplicity of infection (MOI) of 10. Infected cultures were incubated until complete culture lysis was observed. A 1/10 volume of chloroform was added to lyse residual bacterial cells and release bacteriophage. Cell debris and chloroform were removed by centrifugation (5000 rpm for 10 min), and the supernatant was recovered as the phage stock.

The lysate was diluted in TGP broth and used in standard overlay plaque assays with sloppy agar (0.3 % wt/vol agar). Single plaques from these assays were picked using a cut pipette tip to stab into the agar and lift the plaques from the plate. Plaques were crushed and suspended in 1 ml of TGP broth and then used in subsequent rounds of plaque assays. Three rounds of plaque purification were performed, and the purified phages were used in all subsequent experiments.

Mass spectrometry

Samples were precipitated using five volumes of ice-cold acetone and incubated overnight at -20 °C. Precipitates were pelleted by centrifugation at 12 000 × *g* for 10 min. Supernatants were carefully removed, and pellets were air-dried prior to dissolution in 100 mM triethylammo- nium bicarbonate (TEAB) and determination of protein concentrations (A_{280nm}). Aliquots of 100 µg of solubilized proteins were reduced with 5 mM tris-carboxyethyl phosphine (TCEP; Fluka) for 30 minutes at room temperature. Cysteine residues were methylated by treatment with 10 mM methane methylthiosulfonate (MMTS; Sigma) for 15 minutes at room temperature. After methylation, samples were diluted to 95 µL with 50 mM TEAB before the addition of 5 µL of trypsin (Promega) at 1 mg/mL. Samples were incubated at 37 °C overnight, dried, and resuspended in 30 µL of 2 % acetonitrile:water/0.05 % TFA.

Residual digest reagents were removed using an in-house-manufactured C18 stage tip. The samples were loaded onto the stage tip after activating the C18 membrane with 30 µL of methanol (Sigma) and equilibration with 30 µL of 2% acetonitrile:water/0.05 % TFA. The bound sample was washed with 30 µL of 2 % acetonitrile:water/0.05 % TFA before elution with 30 µL 50 % acetonitrile:water/0.05 % TFA. The eluate was evaporated to dryness. The dried peptides were dissolved in 2 % acetonitrile:water and 0.1 % TFA for LC-MS analysis. Liquid chromatography was performed on a Thermo Scientific Ultimate 3000 RSLC equipped with a 2 cm × 100 µm C18 trap column and a 25 cm × 75 µm Pepmap C18 analytical column. The solvent system employed was as follows: loading, 2 % acetonitrile:water/0.1 TFA; solvent A, 2 % acetonitrile:water/0.1 % TFA; solvent B, 80 % acetonitrile:water. The samples were loaded onto the trap column using loading solvent at a flow rate of 5 µL/min from a temperature-controlled autosampler set at 7 °C. Loading was performed for 10 min before the sample was eluted onto the analytical column. The gradient was generated at 300 nL/min as follows: 0-4 min 2 % A, 4-6 min 6 % A, 6-95 min 6-35 % A (Chromeleon non-linear gradient 6); 95-100 min 35-50 % A. Chromatography was performed at 50°C, and the outflow was delivered to the mass spectrometer through a stainless steel nano-bore emitter. Mass spectrometry was performed on a Thermo Scientific Fusion mass spectrometer. Data were acquired in positive mode using a Nanospray Flex nano-ESI source (Thermo Scientific) with the spray voltage set to 1.7 kV and the ion transfer tube temperature set to 300 °C. MS1 scans were recorded in the Orbitrap mass analyser set to 12 000 resolution over the scan range m/z = 350-1650 with a fill

time of 50 ms or until the adaptive gain control (AGC) target of 4e5 was reached. Ion filter criteria were set to mono-isotopic precursors only with charge state 2-6 and dynamic exclusion of 1 over 40 s with mass tolerance of 10 ppm. Precursor selection was performed in top-speed data-dependent mode with the most intense precursor selected first with a cutoff intensity higher than 50,000. Precursor selection was performed using the quadrupole mass analyser with an isolation window of m/z = 1.5 prior to HCD fragmentation. HCD collision energy was set to 35 %. Detection was performed in the ion trap mass analyser with ion injection time of 40 ms or until an AGC target of 1e4 was reached. The raw files generated by the mass spectrometer were imported into Proteome Discoverer v1.4 (Thermo Scientific) and processed using Sequest HT. Database interrogation was performed against GVE3-predicted ORF sequences with trypsin cleavage, allowing for two missed cleavages. Precursor mass tolerance was set to 10 ppm, and fragment mass tolerance set to 0.8 Da. Deamidation (NQ) and oxidation (M) were allowed as dynamic modifications, and thiomethylation of C as a static modification.

Electron microscopy

Phage suspensions were prepared as described previously [2]. Three microliters of each sample was pipetted onto carbon-coated 200-mesh copper grids and stained with 2 % aqueous uranyl acetate. The samples were viewed using a LEO 912 Omega TEM at 120 kV (Zeiss, Oberkochen, Germany) housed at the University of Cape Town Physics Department. Images were collected using a ProScan CCD camera.

Results and discussion

Isolation, morphology and host range testing

The phage was a donation from TMO Renewables. Transmission electron microscopy indicated that *G. thermoglucosidasius* phage GVE3 had morphological characteristics of the B1 morphotype group of the family *Siphoviridae* [1] with a non-contractile tail (± 210 nm long) and isometric head (90 nm–100 nm in diameter) (Fig. 1B). Despite attempts to image phage with nucleic acid in the head, no clear micrographs could be obtained. Fig. 1A, however, shows some phage particles that may have nucleic acid in the head attached to cell debris. GVE3 was tested for its ability to infect a range of *Geobacillus* species (Table 1), but was only capable of infected *G. thermoglucosidasius*.

The GVE3 genome

The GVE3 genome sequence was determined to be 141,298 bp in length and showed a much lower G?C content (29.6 %) than its *G. thermoglucosidasius* host (44 %), as is typical for most phage host pairs [64]. It has been shown that higher AT content results in lower relative entropy (D_{KL}) of a DNA molecule, which could be associated with structural changes in the molecule [12]. Perhaps the lower than average AT content of GVE3 plays a role in its adaptation to thermophily, or alternatively is a reflection of the energy cost of producing nucleotides for phage genome synthesis [64]. This genome size makes it the largest known *Geobacillus*-infecting phage. Overall, the GVE3 genome shares little nucleotide-level identity with any bacteriophage genome currently in the NCBI database (as of 03-03-2015). However, small sections of the genome share significant nucleotide sequence identity with other phage genomes (vB_BanS_Tsamsa, Spbc2, c-st) and *Geobacillus*, *Bacillus* and *Clostridium* genome sequences (Table S3).

A total of 202 putative open reading frames were identified, 62 of which could be assigned a function based on BLAST similarity to genes of known function. The GVE3 genome displays the classical modular arrangement seen in many other members of the family *Siphoviridae* (Fig. 2). GC skew analysis indicated that a replication terminus could be located between the putative holin/endolysin (ORF53) genes and recombinase (ORF54) [65; c-st], while the origin of replication was predicted to lie at ±3700 bp (Fig. 3). Repeat regions, often \ 10 bp, are associated with regions where DNA replication is initiated, correspond to sites of gene regulation or transcription termination [10, 56, 61]. Depending on the search criteria, hundreds of inverted and direct repeats of\ 10 bp could be identified in the GVE3 genome, although their functional importance, if any, remains to be determined. A search for direct and inverted repeats of [7 bp and no more than 30 bp apart with 100 % nucleotide sequence identity gave a total of 582 repeats. Two of these invert repeats (TATTTTTT/TAATTAT) are located immediately downstream of ORF3 and in the region predicted to be the origin of replication and may play a role in the initiation of replication.

Although GVE3 does not appear to encode any tRNAs, it does encode a putative ADP-ribose-1-monophosphatase (ORF184; Appr-1-p), an enzyme typically involved in tRNA splicing and encoded in a wide variety of phage genomes, including vB_BanS-Tsamsa [25]. The exact role of this phage element is not clearly established [66], although the link with tRNA synthesis suggests that it could function to remove a rate-limiting step in tRNA

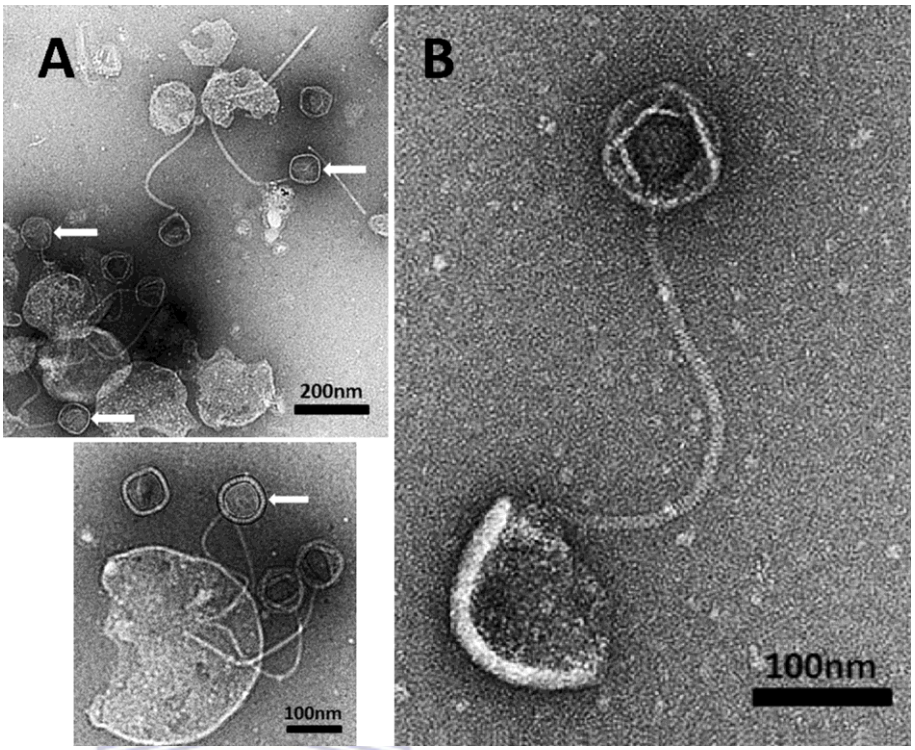


Fig. 1 Bright field TEM of phage GVE3. A) Lower- (top micrograph) and higher-magnification (bottom micrograph) images of several phage attached to cell debris, including some that may still contain nucleic acid in the head (white arrows). B) High magnification image of a single phage particle

Table 1 GVE3 host range

Bacterium	Strain	BGSC no.	Sensitivity to GVE3
Geobacillus stearothermophilus	ATCC 12980^T	9A20^T	-
Geobacillus thermoleovorans	DSM 5366T	96A1^T	-
Geobacillus thermoleovorans	DSM 7263	90A1	-
Geobacillus subterraneus	DSM 13552^T	91A1^T	-
Geobacillus subterraneus	SAM	91A2	-
Geobacillus thermodenitificans	DSM 465^T	94A1^T	-
Geobacillus thermoglucosidans	DSM 2542^T	95A1^T	+
Geobacillus toebii	DSM 14590^T	99A1^T	-
Geobacillus kaue	HU	105A1	-

processing in the host or to aid in recycling of nucleotides [37].

The closest relatives to GVE3, based on subsystems analysis using RAST, appear to be uncharacterized prophages from *Clostridium thermocellum* and *Bacillus* species. The phages predicted to be the most closely related to GVE3, using PHAST (Table S4; http://tinyurl.com/mtg3fbs), are those from *Bacillus* (Spbc2; vB_BanS_Tsamsa) and *Clostridium* (c-st) rather than the known *Geobacillus* phages, an observation that is consistent with an analysis of the terminase large subunit (Fig. 4). GVE3 thus appears to be most closely related to the recently described *B. anthracis*-infecting vB_BanS_Tsamsa [25].

Tetranucleotide usage deviation (TUD) analysis gave a Pearson's correlation coefficient of 0.665 when comparing GVE3 to the genome of *G. thermoglucosidasius*. Interestingly, when comparing the GVE3 sequence to those of *Bacillus anthracis* and *Bacillus cereus*, significantly higher correlation coefficients were obtained (0.796 and 0.797, respectively). TUD analysis using all available *Geobacillus* species genome sequences (*G. kaustophilus, G. toebii, G. thermodinitrificans, G. themoleovorans, G. thermoglucosidasius, G. thermoglucosidans, G. stearothermophilus, G. subterraneus and G. caldoxylosilyticus*) demonstrated that the TUD of GVE3 was most closely matched to that of *G. toebii* (0.705).

Assuming that TUD analysis provides a measure of the adaptation of phage genomes to that of their hosts over time [60], the GVE3 TUD value suggests that *G. thermoglucosidasius* may not be the prevalent host in nature.

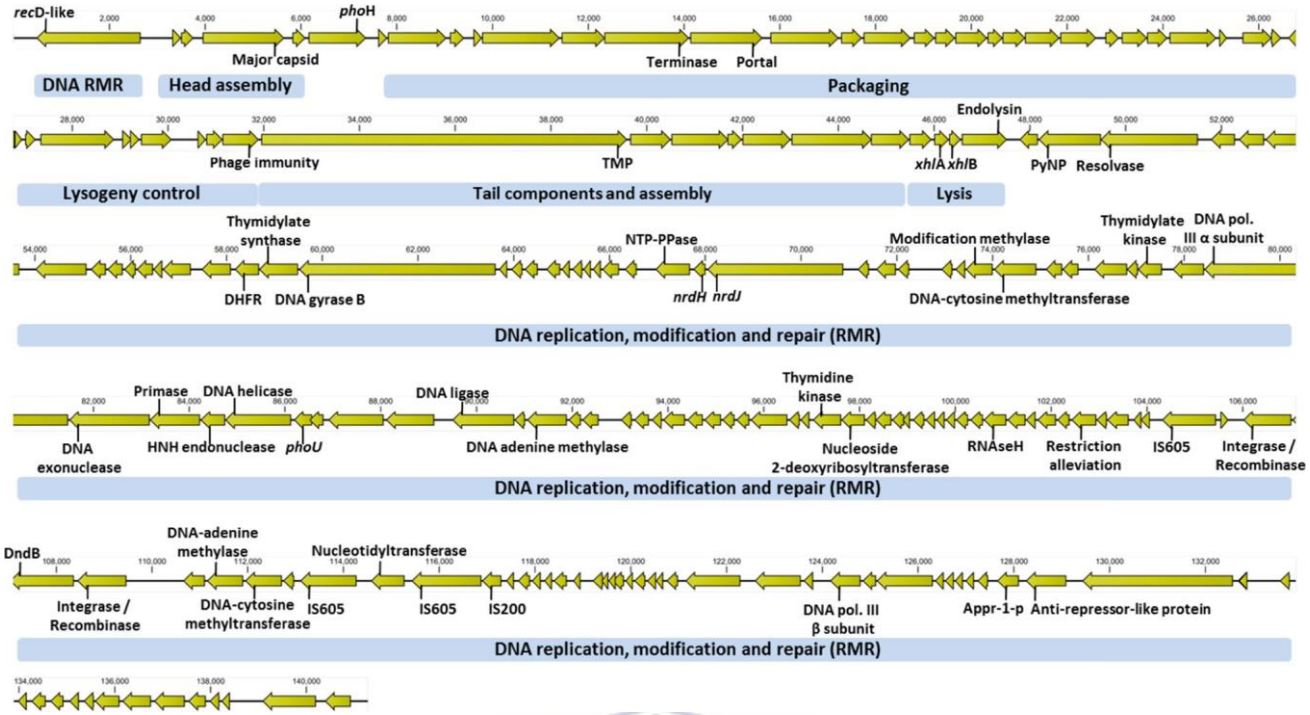


Fig. 2 GVE3 genomic arrangement. Blue boxes indicate modular areas

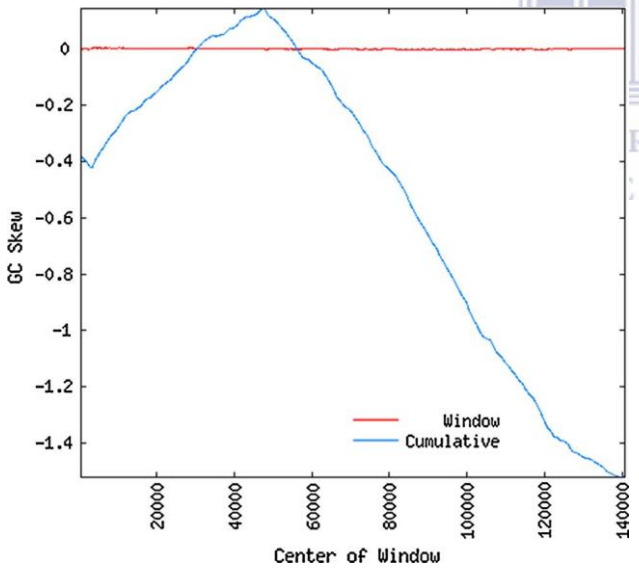


Fig. 3 GC skew analysis of the GVE3 genome showing putative replication origin (*ori*) and termination sites (*ter*) calculated using a window size of 1000 bp and a step size of 100 bp

The higher correlation coefficients of the GVE3 TUD when compared to *B. anthracis* and *B. cereus* (cf. *G. thermoglucosidasius*) suggest that there may be an as yet unidentified *Geobacillus* species with TUD patterns more similar to these two *Bacillus* species that could be the ''natural'' hosts for GVE3. Alternatively, these results could suggest that GVE3 has ''recently'' evolved from a mesophilic counterpart and that the high TUD correlation to mesophilic *Bacillus* species is a genuine indication of its evolutionary heritage. A similar relationship has been observed for GBK2, a *G. kaustophilus*-infecting phage that is most closely related to the *Bacillus subtilis* phage SPP1 [50].

Evolution from mesophily to thermophily should involve the adaptation (in both thermophily and thermostability) of phage proteins, and it is therefore unlikely that the thermophilic GVE3 phage would be capable of replicating effectively in a mesophilic host.

DNA metabolism and replication

GVE3 encodes several proteins associated with nucleotide metabolism, including pyrimidine nucleoside phosphorylase (ORF55; PyNP), thymidylate synthase (ORF69; TS), thymidine kinase (ORF123; TK), ribonucleotide reductase (ORF82/83; RNR), nucleoside triphosphate pyrophosphohydrolase (ORF81; NTP-PPase) and nucleoside-deoxyribosyltransferase (ORF124; ND). Although the ORF encoding the putative RNR is most closely related to class II RNRs, there is a small ORF directly downstream of this gene that shows high homology to a *nrdH*-like gene. Ribonucleotide reductases can be divided into several classes (Ia, Ib, Ic; II and III), of which class II RNRs are usually encoded by a single ORF (*nrdJ*), are oxygen

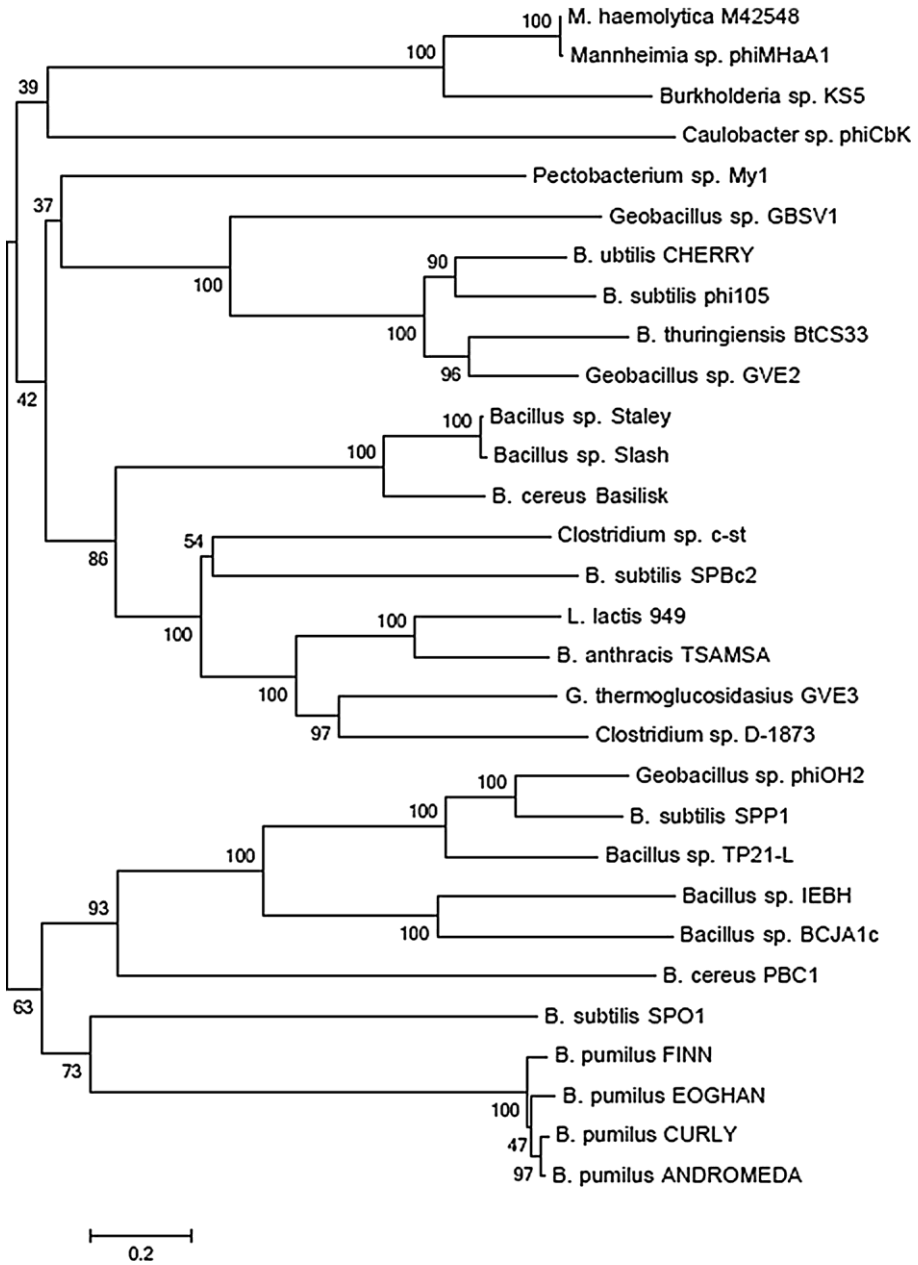


Fig. 4 Neighbor-joining tree comparing full length amino acid sequences of GVE3 terminase large subunit with related proteins. The optimal tree with the sum of branch length = 15.69592415 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method and are in units of the number of amino acid substitutions per site (scale bar). The analysis involved 28 amino acid sequences. All ambiguous positions were removed for each sequence pair. There were a total of 821 positions in the final dataset. GVE3, *G. thermoglucosidasius* (KP144388); IEBH, *Bacillus sp.* (NC_011167); BCJA1c, *Bacillus sp.* (NC_006557); TP21-L, *Bacillus sp.* (NC_011645); SPP1, *B. subtilis* (NC_004166); PBC1, *B. cereus* (NC_017976); phiOH2, *Geobacillus sp*. (NC_021784); D-1873, *Clostridium sp.* (ACSJ01000014); vB_BanS-Tsamsa, *B. anthracis* (NC_023007); 949, *L. lactis* (NC_015263); SPBc2, *B. subtilis* (AF020713); c-st. *Clostridium sp*. (D90210); Basilisk, *B. cereus* (KC595511); SPO1, *B. subtilis* (NC_011421); Slash, *Bacillus sp*. (NC_022774); Staley, *Bacillus sp.* (NC_022767); FINN, *B. pumilus* (NC_020480); EOGHAN, *B. pumilus* (NC_020477); ANDROMEDA, *B. pumilus* (NC_020478); CURLY, *B. pumilus* (NC_020479); BtCS33, *B. thuringiensis* (NC_018085); phi105, *B. subtilis* (NC_004167); CHERRY, *B. anthracis* (NC_007457); GBSV1, *Geobacillus sp*. (NC_008376); My1, *Pectobacterium sp.* (NC_018837); phiCbK, *Caulobacter sp.* (NC_019405); KS5, *Burkholderia sp*. (NC_015265); phiMHaA1, *Mannheimia sp.* (NC_008201); M2548, *M. haemolytica* (CP005383)

independent, and usually rely on vitamin B12 for generation of the tyrosyl radical *in vivo* [21]. Class Ib RNRs, encoded by *nrdHIEF*, rely on the glutaredoxin-like protein encoded by *nrdH* to generate the radical needed for catalysis [80]. GVE3 encodes a class II ribonucleotide reductase (*nrdJ*), as well as a component of a class Ib RNR (*nrdH*). The presence and spatial orientation of both *nrdJ*-like and *nrdH*-like ORFs would suggest that they function together. This unusual arrangement has been described for three *Mycobacterium* siphoviruses: Che12, D29 and L5 [21]. It has been argued that the *nrdH* homologue in these genomes was acquired through horizontal gene transfer. Phage genomes are, however, under strong selective pressure to remain within a strict size limit, and all retained genes are expected to confer some metabolic advantage to the host and the phage [23]. In the case of GVE3, the proximity and spatial arrangement of *nrdH* and *nrdJ* as well as the retention of only the *nrdH* homolog (as opposed to any of the *nrdIEF* genes or gene fragments) would argue that these genes confer an advantage to the phage, perhaps *via* interaction with host-encoded components.

The NTP-PPase contains a MazG domain. MazG belongs to the family of a-nucleoside triphosphate pyrophosphohydrolases, which are thought to be responsible for hydrolysis of all non-canonical nucleoside triphosphates produced as a by-product of metabolism and which are toxic to the host, into monophosphate derivatives, thus playing a house-cleaning role [13]. An alternative hypothesis is that, at least in *E. coli*, the NTP-PPase controls the levels of the global regulator ppGpp, redirecting transcription in favour of genes important for starvation survival [47]. Homologues of these proteins have been identified in many phage genomes [28]. In *E. coli*, *mazG* is co-transcribed with a toxin-antitoxin system (*mazFE*) [27]. It is worth noting that GVE3 encodes several MazF/PemK homologues (ORF38, 40 and 185), although no *mazE* homologues could be identified, and the GVE3 *mazG*-homologue is not co-transcribed with any of these. Whether or not the phage NTP-PPase fulfils multiple roles after host infection, such as regulating the levels of MazF-like toxin produced or eliciting a host survival response to steer its metabolism towards viral production and/or removing toxic nucleoside triphosphates, remains to be determined.

Three DNA-polymerase-like subunits are present on the GVE3 genome. Two of these (ORF97 and 176) are most closely related to the alpha- and beta-clamp subunits of the DNA polymerase III family, similar to those found in *Bacillus* phage vB_BanS_Tsamsa, *Clostridium* phages c-st, D-1873 and *Lactococcus* phage 949. The third subunit, ORF8, shows homology to DNA polymerase A. Other ORFs, the products of which may form part of the DNA Pol III holoenzyme, are a primase (ORF99) and a helicase (ORF101). It has been demonstrated for the *E. coli* DNA polymerase that only the alpha subunit is required for processive replication *in vitro*, although the authors conceded that other subunits, including subunit e, may be required *in vivo* due to the polymerase encountering obstacles such as proteins bound to the DNA, and DNA lesions not taken into account in their *in vitro* assay system [49]. As not all DNA polymerase III holoenzyme components could be identified on the GVE3 genome, it is possible that the phage recruits host-encoded subunits to complete the polymerase holoenzyme assembly to enable the highly processive DNA replication required for fast and accurate replication of the phage genome [14, 69].

Structural proteins

A putative tail tape measure protein (TMP; ORF42/43) appears to be interrupted by a 310-bp insertion (bp 33,537-33,847), most likely a group I self-splicing intron, as predicted by RNAweasel. As for phage JCL1032 from *Lactobacillus delbruckeii* [63], the 3' end of the ORF encoding the N-terminal protein sequence (bp 31,948-33,536) ends with a TAG stop codon followed by the intron. Over the length of the putative TMP, seven large imperfect amino acid repeats could be identified (B 102 aa). The presence of repeat regions in these proteins has been reported previously and is thought to be of structural significance in determining tail length [8].

Mobile elements

Four putative integrase/recombinase genes were identified (ORF28, 54, 147 and 149), none of which share significant amino acid similarity with each other, a feature noted with phage vB_BanS_Tsamsa [25]. The GVE3 phage genome carries three IS605 family OrfB genes (ORF145, 154 and 156). Insertion sequences of this family sometimes comprise two genes encoding an OrfA (IS200 family) and OrfB, together serving as the functional transposon [30]. One OrfB homologue in GVE3 (ORF156) does have an IS200 family gene directly upstream (ORF157), suggesting that they act co-ordinately. The arrangement of the genes is unusual in that they are transcribed in the same direction while most IS200 family transposons, when associated with an OrfB IS605 element, are divergently transcribed. Parts of GVE3 genome have been incorporated into *Geobacillus toebii* WCH70 CRISPR regions (Table S2). One of these spacers (36 bp) is located in the sequence directly downstream of ORF143 on GVE3. Currently, the incorporation of sequences into CRISPR spacer regions is thought to occur through the identification of bi- or trinucleotide sequences found adjacent to the protospacers, which are eventually incorporated in the CRISPR array,

and it is now thought that all type I CRISPR systems target invading DNA for degradation [86]. Interestingly, an IS605/IS200 element (GWCH70_2010 and 2011) is situated directly upstream of the Cas6 (2068682bp-2069410bp) gene in WCH70, probably inactivating this CRISPR array. This CRISPR array also carries the 36-bp spacer, and it is tempting to speculate that a connection exists between these elements. The 36-bp sequence may be important in the ability of the ORF143 transposon to jump, and incorporation of this spacer into a CRISPR cassette may inactivate the transposon, preventing it from inactivating host defence systems.

Nucleotide modifications

Digestion with several restriction endonucleases, including the four-base cutter *RsaI*, for which there are 228 sites on the GVE3 genome, was not successful, whereas treatment with *Alu*I (335 sites) resulted in digestion of the DNA (Table S5). Examples where *Alu*I but not *Rsa*I would digest DNA have been reported and are thought to be due to substitution of thymine with deoxyuridine or substitution of guanine with deoxyinosine [11]. It has also been established that *Alu*I cannot digest the following modified sites: ^{m6}AGCT, AG^{m4}CT, AG^{m5}CT, AG^{hm5}CT [51], and these can probably be excluded as the modifications present in GVE3 DNA. The presence of putative methylases potentially targeting adenine and cytosine residues (ORFs 108, 151 and 152) as well as a DndB domain (ORF146) suggests that the phage DNA is modified to avoid digestion by host-encoded enzymes. For example, *E. coli* T-even phages contain hydroxymethylcytosine (HMC), and *B. subtilis* phage PBS1 contains uracil in place of thymine. The pyrimidine 5-hydroxymethyluracil (HMU) replaces thymine in *B. subtilis* phages SP8, SP5C, SPOl, SP82 and 4e [55]. GVE3 also encodes a putative restriction alleviation protein (bp 102,951-103,163), possibly part of a strategy to avoid host defences.

The presence of restriction endonucleases that inhibit genetic transformation of *Geobacillus* species, and in particular *Hae*III in *G. thermoglucosidasius*, has been reported (WO2006117536A1; [77]). Interestingly, all but one of the *Hae*III sites on GVE3, of which there are only 10, are located within the 3^{0}-terminal 946 bp of the phage genome. They are irregularly spaced and do not appear to form part of conserved repeats. Digestion of phage DNA with *Hae*III could not be detected. The limited number of *Hae*III sites and their location may indicate that the phage genome is under selective pressure to remove such sites. We speculate that, as for phage P1, the 946-bp region containing *Hae*III sites constitutes a *pac* site and that *pac* site cleavage is controlled by the methylation state surrounding the cleavage site [75].

GVE3 proteome

To confirm the expression of predicted ORFs, the complete proteome of GVE3 was determined. The protein products of all predicted ORFs listed in Table S1, except ORF5, 60 and 169, could be identified by at least three unique peptides. The three segments of ORF60, which contains two frameshift mutations, are clearly similar to a hypothetical protein identified in a *Bacillus* species. (WP_028394443.1). However, no peptides similar to any of the three segments of the ORF could be identified, and we conclude that ORF60 is an un-translated region. A peptide corresponding to the PyNP protein was identified, suggesting that this enzyme may play a role in postinfection nucleotide metabolism (see below). No peptide sequences could be identified for the 310-bp region predicted to be a group I self-splicing intron (bp 33,537-33,847) indicating that this is likely to be an untranslated region. If the intron self-excises from this region once inside the host, it is reasonable to expect that a fusion protein, the functional TMP, would be formed by the N- and C-terminal regions of the predicted TMP interrupted by this intron. However, no evidence could be found for the formation of such a fusion between these two terminal regions, and it is likely that each ORF is expressed as a unique protein. The DNA sequence of ORF70 contains a stop codon (TAG) in the reading frame, which translates to VLD*EVK. The identification of a VLDEVK-containing peptide suggests that either readthrough translation or ribosome slippage occurs over this codon. GVE3 structural proteins were also analysed by SDS-PAGE (Fig. 5). Eleven proteins could be identified, of which band 6 corresponds to the size of the predicted major head protein (ORF4), while bands 7 and 8 are likely to correspond to the N-terminal portion of the tape measure protein (ORF42) and the portal protein (ORF14), respectively.

Lysis and lysogeny

There are at least two holin homologues located directly upstream of the endolysin-encoding ORF, the second of these having what appears to be a dual start motif (M-X$_n$-M) with a lysine being one of the two residues separating the methionines. The arrangement of the genes and homology to *xhlA*/*xhlB*/*xlyA* genes from *B. subtilis* phage PBSX suggest that lysis might occur in a manner similar to that system [35]. ORF51 has one predicted transmembrane region (aa 75-97), while ORF52 has two such regions (aa 9-31; aa 41-59).

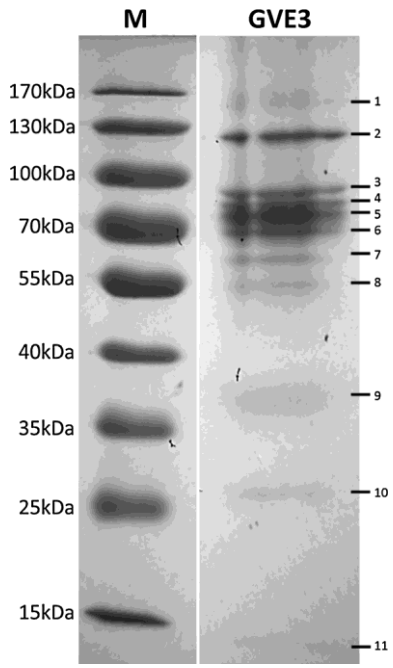


Fig. 5 SDS-PAGE of GVE3 structural proteins. M, Molecular mass marker

Initial plaque assays demonstrated ''bulls-eye'' plaque morphology, suggestive of host lysogeny [41]. Several bacterial colonies could be observed growing inside plaques. These were isolated and tested for their sensitivity to the phage and were found to be resistant to phage infection. The genome sequence of one of these isolates was determined (Illumina MiSeq; 55-fold coverage) and served as confirmation of the GVE3 genome sequence obtained by Roche 454 sequencing. This showed that the phage genome had inserted into the bacterial host genome and that the *attB* site, with a 23-bp sequence (GGTGGCGTCGGCGATACGACGAC) that was duplicated on insertion (Fig. 6). This sequence only occurs once in the *G. thermoglucosidasius* 11955 genome, located 247 bp from the start of the pyrimidine nucleoside phosphorylase gene (*deoA*), a region known to be interrupted by phage insertion in other genomes [16]. The phage encodes a putative PyNP, downstream of a resolvase, in which the *attP* site is situated. Incorporation of the GVE3 genome sequence in CRISPR spacer regions of the lysogen could not be detected, although two spacers with some nucleotide sequence similarity to regions of the GVE3 genome were identified in the *G. thermoglucosidasius* 11955 genome sequence (Table S2). Integration of the GVE3 genome sequence into that of its host is likely to inactivate the host-encoded PyNP. The presence of a phage-encoded PyNP could suggest an obligate requirement to retain this activity, and that once integrated, the host relies on the phage PyNP, making use of a promoter located in the C-terminal region of the integrase (ORF55) or of readthrough transcription from the promoter located upstream of the host-encoded PyNP (Fig. 6). The PyNP on GVE3 does not show 100 % amino acid sequence identity to the gene from *G. thermoglucosidasius*, or any genes in the NCBInr database. If not essential for either the phage or the host (mutation in PyNP is non-lethal), it may suggest that GVE3 is a specialized transducing phage. No *G. thermoglucosidasius* genomic sequences were observed in the GVE3 genome or in the NGS data, suggesting that GVE3 is unlikely to be a generalized transducing phage. A putative anti-repressor protein (ORF183), which contains an ORF6N domain and has amino acid similarity (50 % over 112 aa) to a truncated annotated anti-repressor protein in *Peptoclostridium difficile*, was identified (Table S1; [31]). In phage lambda, this serves as part of the regulatory mechanism to switch between lysis and lysogeny. Early evidence, based on its overexpression in the host prior to infection, suggests that it plays the same role as in phage lambda (van Zyl et al., unpublished data).

Auxiliary metabolic genes

The GVE3 gene carries the auxiliary metabolic gene *phoH*, and a putative regulator of *phoH* expression, *phoU*, is located upstream of the genes for DNA replication and distant (±79 kb apart) from the *phoH* homologue. The phage also encodes a putative ADP-ribose-1-monophosphatase, which catalyses the conversion of ADP-ribose-1-monophosphate to ADP-ribose as part of the tRNA splicing pathway [37]. The role of *phoH* has not been clearly defined, with some studies demonstrating upregulation under phosphate stress or phage infection [26] while others show downregulation or no change. Should the GVE3 *phoH* gene expression be upregulated, this might suggest that, as with other phages, DNA (and RNA) synthesis becomes rate limiting in the host once replication and transcription of the phage genome is initiated.

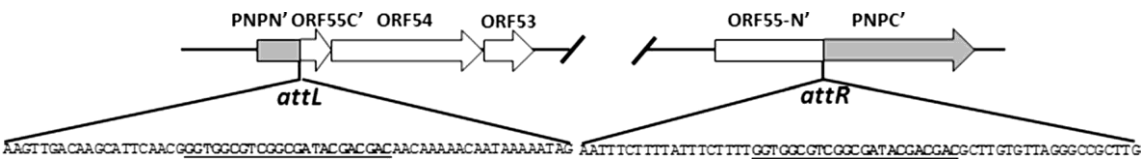


Fig. 6 Layout of the integrated phage. The space between the diagonal lines denotes the rest of the phage genome. The grey box and arrow represent the N- and C-termini, respectively, of the *G. thermoglucosidasius* pyrimidine nucleoside phosphorylase

GVE3 signatures in *Geobacillus* genomes

Two regions of 100 % nucleotide sequence identity to CRISPR spacer regions were found in *G. toebii* WCH70 (Table S2). The presence of these nucleotide sequences suggests that GVE3 or a highly similar phage infected this strain in the past. PCR analysis using four primer pairs targeted to various areas of the GVE3 genome could not detect GVE3 in the chromosome of the *G. toebii* DSM 14590^{T} strain (Table 1), suggesting that superinfection immunity is unlikely to be the cause of failure to infect this strain. Several other putative GVE3-related sequences were identified in CRISPR repeats in a range of *G. thermoglucosidasius* genome sequences (Table S2).

Of the two spacers identified in the *G. toebii* WCH70 genome, one is located at the trailer end of the repeat region, and the other, located in a second CRISPR array, at the leader end in that array, suggesting that this strain has been repeatedly challenged with the same phage [29, 85]. The absence of evidence of lysogenic integration of the GVE3 genome in the WCH70 genome could be due to CRISPR-mediated killing of hosts containing an integrated phage or those that have been infected in the past [22, 48]. Imperfectly matched spacers similar to GVE3 in CRISPR arrays in the 11955 genome could suggest infection by a closely related phage, as seen in polyclonal phage populations during phage blooms or adaptation by the phage to circumvent CRISPR resistance [48, 85]. We suggest that GVE3 represents the latest iteration of a much older version of the phage not currently targeted by the CRISPR system in *G. thermoglucosidacius* 11955. Insertion of spacer sequences based on those identified in WCH70 could be used to engineer resistance by incorporating these into one of the 11955 CRISPR arrays [52].

Conclusion

GVE3, although a member of the well-known family *Siphoviridae* and unremarkable with respect to the overall layout of genes and the genes encoded, appears to have a unique genome sequence, with no close relatives in the current databases. Although there are indications that it may have had the capacity to infect other *Geobacillus* species in the past, the current specificity appears to be restricted to *G. thermoglucosidasius*. The relationships between the GVE3 genome and those of mesophilic phages and bacteria may be a consequence of the small number of thermophilic phage genome sequences in the databases but may reflect the evolutionary history of a phage in transition from mesophily to thermophily. GVE3 encodes a number of enzymes, including ATP-dependent DNA ligase, DNA polymerase III, RNaseH, PyNP, holin and endolysin [78], all of which should be thermostable. These could be of commercial value or employed as research tools, such as in the use of endolysin in the treatment of milk to kill *Geobacillus* species spoilage organisms [15, 81]. *G. thermoglucosidasius* has been engineered as a platform organism for ethanol production and other industrial products, but to date, there is no mechanism for the introduction of large DNA fragments ([12 kb), and GVE3 could potentially be developed as a system for introduction of novel or engineered metabolic and biosynthetic pathways.

Acknowledgments The authors wish to thank TMO Renewables for the gift of the GVE3 phage. This work was funded by the National Research Foundation (NRF) of South Africa. The authors declare no conflict of interest.

References

1. Ackermann HW (2007) 5500 Phages examined in the electron microscope. Arch Virol 152:227–243
2. Ackermann HW, Heldal M (2010) Basic electron microscopy of aquatic viruses. In: Manual of aquatic viral ecology, Chapter 18. American Society of Limnology and Oceanography, Inc., p 182–192
3. Adriaenssens EM, van Zyl LJ, de Maayer P, Rubagotti E, Rybicki E, Tuffin M, Cowan DA (2014) Metagenomic analysis of the viral community in Namib Desert hypoliths. Environ Microbiol. doi:10.1111/1462-2920.12528
4. Ahn D-G, Kim S-I, Rhee J-K, Kim KP, Pan J-G, Oh J-W (2006) TTSV1, a new virus-like particle isolated from the hyperthermophilic crenarchaeote *Thermoproteus tenax*. Virol 351:280–290
5. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410
6. Arnold HP, Zillig W, Ziese U, Holz I, Crosby M, Utterback T, Weidmann JF, Kristjanson JK, Klenk HP, Nelson KE, Fraser CM (2000) A novel lipothrixvirus, SIFV, of the extremely thermophilic crenarchaeon *Sulfolobus*. Virol 267:252–266
7. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O (2008) The RAST server: Rapid annotations using subsystems technology. BMC Genomics 9:75
8. Belcaid M, Bergeron A, Poisson G (2011) The evolution of the tape measure protein: units, duplications and losses. BMC Bioinform 12:S10
9. Betley JN, Frith MC, Graber JH, Choo S, Deshler JO (2002) A ubiquitous and conserved signal for RNA localization in chordates. Curr Biol 12:1756–1761
10. Blatny JM, Godager L, Lunde M, Nes IF (2004) Complete genome sequence of the *Lactococcus lactis* temperate phage uLC3: comparative analysis of uLC3 and its relatives in lactococci and streptococci. Virology 318:231–244
11. Bodnarz JW, Zempsky W, Warder D, Bergson C, Ward DC (1983) Effect of nucleotide analogs on the cleavage of DNA by the restriction enzymes *Alu*I, *Dde*I, *Hin*fI, *Rsa*I, and *Taq*I. J Biol Chem 258:15206–15213
12. Bohlin J, van Passel MWJ, Snipen L, Kristoffersen AB, Ussery D, Hardy SP (2012) Relative entropy differences in bacterial

chromosomes, plasmids, phages and genomic islands. BMC Genomics 13:66–78
13. Bryan MJ, Burroughs NJ, Spence EM, Clokie MRJ, Mann NH, Bryan SJ (2008) Evidence for the intense exchange of MazG in marine cyanophages by horizontal gene transfer. PLOS One. doi:10.1371/journal.pone.0002048
14. Bullard JM, Williams JC, Acker WK, Jacobi C, Janjic N, McHenry CS (2002) DNA polymerase III holoenzyme from *Thermus thermophiles* identification, expression, purification of components, and use to reconstitute a processive replicase. J Biol Chem 277:13401–13408
15. Burgess SA, Lindsay D, Flint SH (2010) Thermophilic bacilli and their importance in dairy processing. Int J Food Microbiol 144:215–225
16. Buxton RS, Hammer-Jespersen K, Hansen TD (1978) Insertion of bacteriophage lambda into the *deo* operon of *Escherichia coli* K-12 and isolation of plaque-forming k*deo*$^{?}$ transducing bacteriophages. J Bacteriol 136:668–681
17. Chen Y, Wei D, Wang Y, Zhang X (2013) The role of interactions between bacterial chaperone, aspartate aminotransferase, and viral protein during virus infection in high temperature environment: the interactions between bacterium and virus proteins. BMC Microbiol 13:48
18. Clokie MRJ, Millard AD, Letarov AV, Heaphy S (2011) Phages in nature. Bacteriophage 1:31–45
19. Cripps RE, Eley K, Leak DJ, Rudd B, Taylor M, Todd M, Boakes S, Martin S, Atkinson T (2009) Metabolic engineering of *Geobacillus thermoglucosidasius* for high yield ethanol production. Metab Eng 11:398–408
20. Doi K, Mori K, Martono H, Nagayoshi Y, Fujino Y, Tashiro K, Kuhara S, Ohshima T (2013) Draft Genome Sequence of Geobacillus kaustophilus GBlys, a Lysogenic Strain with Bacteriophage OH2. Genome Announc 1:e00634–e00713
21. Dwivedi B, Xue B, Lundin D, Edwards RA, Breitbart M (2013) A bioinformatic analysis of ribonucleotide reductase genes in phage genomes and metagenomes. BMC Evol Biol 13:33
22. Edgar R, Qimron U (2010) The *Escherichia coli* CRISPR system protects from lysogenization, lysogens, and prophage induction. J Bacteriol 192:6291–6294
23. Feiss M, Siegele DA (1979) Packaging of the bacteriophage lambda chromosome: dependence of *cos* cleavage on chromosome length. Virology 92:190–200
24. Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791
25. Ganz HH, Law C, Schmuki M, Eichenseher F, Calendar R, Loessner MJ, Getz WM, Korlach J, Beyer W, Klumpp J (2014) Novel giant Siphovirus from *Bacillus anthracis* features unusual genome characteristics. PLoS One 9:e85972
26. Goldsmith DB, Crosti G, Dwivedi B, McDaniel LD, Varsani A, Suttle CA, Weinbauer MG, Sandaa RA, Breitbart M (2011) Development of phoH as a novel signature gene for assessing marine phage diversity. Appl Environ Microbiol 77:7730–7739
27. Gross M, Marianovsky I, Glaser G (2006) MazG—a regulator of programmed cell death in *Escherichia coli*. Mol Microbiol 59:590–601
28. Hargreaves KR, Kropinski AM, Clokie MRJ (2014) Bacteriophage behavioral ecology: how phages alter their bacterial host's habits. Bacteriophage. doi:10.4161/bact.29866
29. Heler R, Marraffini LA, Bikard D (2014) Adapting to new threats: the generation of memory by CRISPR-Cas immune systems. Mol Microbiol. doi:10.1111/mmi.12640
30. Höök-Nikanne J, Berg DE, Peek RM Jr, Kersulyte D, Tummuru MKR, Blaser MJ (1999) DNA sequence conservation and diversity in transposable element IS605 of *Helicobacter pylori*. Helicobacter 3:79–85
31. Iyer LM, Koonin EV, Aravind L (2002) Extensive domain shuffling in transcription regulators of DNA viruses and implications for the origin of fungal APSES transcription factors. Gen. Biol 3:research0012.1–research0012.11
32. Jin M, Ye T, Zhang X (2013) Roles of bacteriophage GVE2 endolysin in host lysis at high temperatures. Microbiology 159:1597–1605
33. Jin M, Chen Y, Xu C, Zhang X (2014) The effect of inhibition of host MreB on the infection of thermophilic phage GVE2 in high temperature environment. Sci Rep 4:4823
34. Kotze AA, Tuffin IM, Deane SM, Rawlings DE (2006) Cloning and characterization of the chromosomal arsenic resistance genes from *Acidithiobacillus caldus* and enhanced arsenic resistance on conjugal transfer of *ars* genes located on transposon TnAtcArs. Microbiology 152:3551–3560
35. Krogh S, Jørgensen ST, Devine KM (1998) Lysis genes of the *Bacillus subtilis* defective prophage PBSX. J Bacteriol 180:2110–2117
36. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580
37. Kumaran D, Eswaramoorthy S, Studier FW, Swaminathan S (2005) Structure and mechanism of ADP-ribose-1-monophosphatase (Appr-1-pase), a ubiquitous cellular processing enzyme. Prot Sci 14:719–726
38. Lang BF, Laforest MJ, Burger G (2007) Mitochondrial introns: a critical view. Trends Genet 23:119–125
39. Laslett D, Canback B (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucl Acids Res 32:11–16
40. Le Romancer M, Gaillard M, Geslin C, Prieur D (2007) Viruses in extreme environments. In: Amils Ricardo, Ellis-Evans Cynan, Hinghofer-Szalkay Helmut (eds) Life in extreme environments. Springer, Netherlands, pp 99–113
41. Levine M, Truesdall S, Ramakrishan T, Bronson MJ (1975) Dual control of lysogeny by bacteriophage P22: an antirepressor locus and its controlling elements. J Mol Biol 91:421–438
42. Lin PP, Rabe KS, Takasumi JL, Kadisch M, Arnold FH, Liao JC (2014) Isobutanol production at elevated temperatures in thermophilic *Geobacillus thermoglucosidasius*. Metab Eng 24:1–8
43. Liu B, Wu S, Song Q, Zhang X, Xie L (2006) Two novel bacteriophages of thermophilic bacteria isolated from Deep-Sea hydrothermal fields. Curr Microbiol 53:163–166
44. Liu B, Zhang X (2008) Deep-sea thermophilic *Geobacillus* bacteriophage GVE2 transcriptional profile and proteomic characterization of virions. Appl Microbiol Biotechnol 80:697–707
45. Liu B, Zhou F, Wu S, Xu Y, Zhang X (2009) Genomic and proteomic characterization of a thermophilic *Geobacillus* bacteriophage GBSV1. Res Microbiol 160:166–170
46. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucl Acids Res 25:955–964
47. Magnusson LU, Farewell A, Nystrom T (2005) ppGpp: a global regulator in *Escherichia coli*. Trends Microbiol 13:236–242
48. Marraffini LA, Sontheimer EJ (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat Rev Genet 11:181–190
49. Marians KJ, Hiasa H, Kim DR, McHenry C (1998) Role of the core DNA polymerase III subunits at the replication fork: ALPHA is the only subunit required for processive replication. J Biol Chem 273:2452–2457
50. Marks TJ, Hamilton PT (2014) Characterization of a thermophilic bacteriophage of *Geobacillus kaustophilus*. Arch Virol 159:2771–2775

51. McClelland M, Nelson M, Raschke E (1994) Effect of site-specific modification on restriction endonucleases and DNA modification methyltransferases. Nucl Acids Res 22:3640–3659
52. Millen AM, Horvath P, Boyaval P, Romero DA (2012) Mobile CRISPR/Cas-mediated bacteriophage resistance in *Lactococcus lactis*. PLoS One 7:e51663
53. Moser MJ, DiFrancesco RA, Gowda K, Klingele AJ, Sugar DR, Stocki S, Mead DA, Schoenfeld TW (2012) Thermostable DNA polymerase from a viral metagenome is a potent RT-PCR enzyme. PLOS One. doi:10.1371/journal.pone.0038371
54. Nazina TN, Tourova TP, Poltaraus AB, Novikova EV, Grigoryan AA, Ivanova AE, Lysenko AM, Petrunyaka VV, Osipov GA, Belyaev SS, Ivanov MV (2001) Taxonomic study of aerobic thermophilic bacilli: descriptions of *Geobacillus subterraneus* gen. nov., sp. nov. and *Geobacillus uzenensis* sp. nov. from petroleum reservoirs and transfer of *Bacillus stearothermophilus*, *Bacillus thermocatenulatus*, *Bacillus thermoleovorans*, *Bacillus kaustophilus*, *Bacillus thermoglucosidasius* and *Bacillus thermodenitrificans* to *Geobacillus* as the new combinations *G. stearothermophilus*, *G. thermocatenulatus*, *G. thermoleovorans*, *G. kaustophilus*, *G. thermoglucosidasius* and *G. thermodenitrificans*. IJSEM 51:433–446
55. Neubort S, Marmur J (1973) Synthesis of the unusual DNA of *Bacillus subtilis* bacteriophage SP-15. J Virol 12:1078–1084
56. Østergaard S, Brøndsted L, Vogensen FK (2001) Identification of a replication protein and repeats essential for DNA replication of the temperate lactococcal bacteriophage TP901-1. Appl Environ Microbiol 67:774–781
57. Payeta JP, Suttle CA (2013) To kill or not to kill: The balance between lytic and lysogenic viral infection is driven by trophic status. Limnol Oceanogr 58:465–474
58. Peduzzi P, Gruber M, Gruber M, Schager M (2014) The virus's tooth: cyanophages affect an African flamingo population in a bottom-up cascade. ISME J 8:1346–1351
59. Plotka M, Kaczorowska A-K, Stefanska A, Morzywolek A, Fridjonsson OH, Dunin-Horkawicz S, Kozlowski L, Hreggvidsson GO, Kristjansson JK, Dabrowski S, Bujnicki JM, Kaczorowskia T (2013) Novel highly thermostable endolysin from *Thermus scotoductus* MAT2119 bacteriophage Ph2119 with amino acid sequence similarity to Eukaryotic peptidoglycan recognition proteins. Appl Environ Microbiol 80:886–895
60. Pride DT, Wassenaar TM, Ghose C, Blaser MJ (2006) Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. BMC Genomics. doi:10.1186/1471-2164-7-8
61. Quiles-Puchalt N, Tormo-Más MA, Campoy S, Toledo-Arana A, Monedero V, Lasa I, Novick RP, Christie GE, Penadés JR (2013) A super-family of transcriptional activators regulates bacteriophage packaging and lysis in Gram-positive bacteria. Nucl Acids Res 41:7260–7275
62. Rice G, Stedman K, Snyder J, Wiedenheft B, Willits D, Brumfield S, McDermott T, Young MJ (2001) Viruses from extreme thermal environments. Proc Nat Acad Sci 98:13341–13345
63. Riipinen KA, Alatossava T (2004) Two self-splicing group I introns interrupt two late transcribed genes of prolate-headed *Lactobacillus delbrueckii* phage JCL1032. Arch Virol 149:2013–2024
64. Rocha EPC, Danchin A (2002) Base composition bias might result from competition for metabolic resources. Trends Genet 18:291–294
65. Sakaguchi Y, Hayashi T, Kurokawa K, Nakayama K, Oshima K, Fujinaga Y, Ohnishi M, Ohtsubo E, Hattori M, Oguma K (2005) The genome sequence of *Clostridium botulinum* type C neurotoxin-converting phage and the molecular mechanisms of unstable lysogeny. Proc Nat Acad Sci 102:17472–17477
66. Savalia D, Westblade LF, Goel M, Florens L, Kemp P, Akulenko N, Pavlova O, Padovan JC, Chait BT, Washburn MP, Ackermann HW, Mushegian A, Gabisonia T, Molineux I, Severinov K (2008) Genomic and proteomic analysis of phiEco32, a novel *Escherichia coli* phage. J Mol Biol 377:774–789
67. Schmidt TR, Scott EJ II, Dyer DW (2011) Whole-genome phylogenies of the family Bacillaceae and expansion of the sigma factor gene family in the *Bacillus cereus* species-group. BMC Genomics 12:430
68. Schoenfeld T, Patterson M, Richardson PM, Wommack KE, Young M, Mead D (2008) Assembly of viral metagenomes from yellowstone hot springs. Appl Environ Microbiol 74:4164–4174
69. Seco E, Zinder J, Manhart CM, Piano AL, McHenry C, Ayora S (2013) Bacteriophage SPP1 in vitro DNA replication strategies promote viral and disable host replication. Nucl Acid Res 41:1711–1721
70. Sevostyanova A, Djordjevic M, Kuznedelov K, Naryshkina T, Gelfand MS, Severinov K, Minakhin L (2007) Temporal regulation of viral transcription during development of *Thermus thermophilus* bacteriophage /YS40. J Mol Biol 366:420–435
71. Sime-Ngando ST, Lucas S, Robin A, Tucker KP, Colombet J, Bettarel Y, Desmond E, Gribaldo S, Forterre P, Breitbart M, Prangishvili D (2010) Diversity of virus–host systems in hypersaline Lake Retba. Environ Microbiol 8:1956–1972
72. Song Q, Zhang X (2008) Characterization of a novel non-specific nuclease from thermophilic bacteriophage GBSV1. BMC Biotechnol 8:43
73. Song Q, Ye T, Zhang X (2011) Proteins responsible for lysogeny of deep-sea thermophilic bacteriophage GVE2 at high temperature. Gene 479:1–9
74. Sorokin DY, Berben T, Melton ED, Overmars L, Vavourakis CD, Muyzer G (2014) Microbial diversity and biogeochemical cycling in soda lakes. Extremophiles 18:791–809
75. Sternberg N, Coulby J (1990) Cleavage of the bacteriophage P1 packaging site (*pac*) is regulated by adenine methylation. Proc Natl Acad. Sci 87:8070–8074
76. Suttle CA (2005) Viruses in the sea. Nature 437:356–361
77. Suzuki H, Yoshida K (2012) Genetic transformation of *Geobacillus kaustophilus* HTA426 by conjugative transfer of host-mimicking plasmids. J Microbiol Biotechnol 22:1279–1287
78. Szekera K, Zhou X, Schwab T, Casanueva A, Cowan D, Mikhailopulo IA, Neubauer P (2012) Comparative investigations on thermostable pyrimidine nucleoside phosphorylases from *Geobacillus thermoglucosidasius* and *Thermus thermophiles*. J Mol Cat B Enzymatic 84:27–34
79. Taylor MP, Eley KL, Martin S, Tuffin MI, Burton SG, Cowan DA (2009) Thermophilic ethanologenesis: future prospects for second-generation bioethanol production. Trends Biotechnol 27:398–405
80. Torrents E (2014) Ribonucleotide reductases: essential enzymes for bacterial life. Front Cell Infect Microbiol. doi:10.3389/fcimb.2014.00052
81. Viedma PM, Abriouel H, Omar NB, Lopez RL, Valdivia E, Gálvez A (2009) Inactivation of *Geobacillus stearothermophilus* in canned food and coconut milk samples by addition of enterocin AS-48. Food Microbiol 26:289–293
82. Wang Y, Zhang X (2008) Identification and characterization of a novel thymidylate synthase from deep-sea thermophilic bacteriophage *Geobacillus* virus E2. Virus Genes 37:218–224
83. Wang Y, Zhang X (2010) Genome analysis of deep-sea thermophilic phage D6E. Appl Environ Microbiol 76:7861–7866
84. Wei D, Zhang X (2010) Proteomic analysis of interactions between a deep-sea thermophilic bacteriophage and its host at high temperature. J Virol 84:2365–2373
85. Weinberger AD, Sun CL, Plucinski MM, Denef VJ, Thomas BC, Horvath P, Barrangou R, Gilmore MS, Getz WM, Banfield JF

(2012) Persisting viral sequences shape microbial CRISPR based immunity. PLoS One 8:e1002475
86. Westra ER, Swarts DC, Staals RHJ, Jore MM, Brouns SJJ, van der Oost J (2012) The CRISPRs, they are A-Changin': How prokaryotes generate adaptive immunity. Annual Rev Genet 46:311–339
87. Zhou Y, Liang Y, Lynch K, Dennis JJ, Wishart DS (2011) PHAST: a fast phage search tool. Nucl Acids Res 39:347–352
88. Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. Edited in Evolving Genes and Proteins by Bryson V and Vogel HJ (eds). Academic Press, New York, pp. 97–166

The final publication is available at Springer via http://dx.doi.org/10.1007%2Fs00705-015-2497-9

Permission to reproduce the article here:

Excerpt From Springer Copyright Transfer Contract:

"Author retains the right to use his/her article for his/her further scientific career by including the final published journal article in other publications such as dissertations and postdoctoral qualifications provided acknowledgement is given to the original source of publication."

License no: 4171350796725

Correspondence with the journal:

Dear Mr Leonardo van Zyl,

The PDF for your manuscript, "Identification and characterization of a novel Geobacillus thermoglucosidasius bacteriophage, GVE3" is ready for viewing.

In order to formally submit your manuscript to the journal, you must approve the PDF.

Please access the following web site:

http://arvi.edmgr.com/

Your username is: lvzyl

Your password is:

Click "Author Login".

In your main menu, you will see there is a category entitled "Submission Waiting for Author's Approval". Click on that category, view your submission and approve it. In the unlikely case of conversion issues you may submit your manuscript data as a PDF file.

Your manuscript will then be formally submitted to the journal.

Thank you very much.

With kind regards,

Springer Journals Editorial Office

Ref.: Ms. No. ARVI-D-15-00182R1

Identification and characterization of a novel Geobacillus thermoglucosidasius bacteriophage, GVE3

Dear Mr van Zyl,

Archives of Virology has received your revised submission.

You may check the status of your manuscript by logging onto Editorial Manager at (http://arvi.edmgr.com/).

Best regards,

Springer Journals Editorial Office

Ref.: Ms. No. ARVI-D-15-00182R1

Identification and characterization of a novel Geobacillus thermoglucosidasius bacteriophage, GVE3

Mr Leonardo Joaquim van Zyl

Archives of Virology

Dear Mr van Zyl,

Dr. Horst Neve, Ph.D. has been assigned for the above mentioned manuscript.

http://arvi.edmgr.com/

Your username is: lvzyl

Your password is:

Regards

Editorial Office

Dear Mr van Zyl,

Comments have been received on your revised paper. You will see that your manuscript is very close to final acceptance. Please add the new reference as suggested. I would also ask you to add one of your new micrographs and present it together with the original micrograph (i.e., a new compilation of 2 micrographs at high and at lower magnification).

Please make sure to submit your editable source files (i. e. Word, TeX).

To submit your (minor)revision, go to http://arvi.edmgr.com/ and log in as an Author. You will see a menu item call Submission Needing Revision. You will find your submission record there.

Yours sincerely

Horst Neve, Ph.D.

Editor

Archives of Virology

COMMENTS TO THE AUTHOR:

—

Reviewer's Responses to Questions

Badly written manuscripts are usually rejected before going out for review. Nevertheless, I would like to receive your opinion. Please shortly comment on English language of the manuscript. English language is

Reviewer #2: good

Reviewer 2: A recent paper on deep-sea bacteriophage (Dahai Wei, Xiaobo Zhang. 2010. Proteomic analysis of interactions between a deep-sea thermophilic bacteriophage and its host at high temperature. Journal of Virology 84: 2365-2373) should be included in the citation of the manuscript.

Dear Dr. van Zyl,

Please be so kind to let me know the exact number of days you require to revise the manuscript so that I could be of better assistance.

Thank you very much.

Best regards, Kamatchi

Kamatchi Ulagappan

Springer

Journals Editorial Office (JEO)

JEO Assistant

---

tel +91 44 42197752

fax + 91 44 42197763

Kamatchi.Ulagappan@springer.com

www.springer.com

From: lonnie van zyl [mailto:vanzyllj@gmail.com]

Sent: 28 April 2015 15:06

To: Ulagappan, Kamatchi

Subject: Re: ARVI-D-15-00182R1

Dear Dr. van Zyl,

I checked the system and find that the revision is due only on 23 June 2015.

Thank you very much.

Best regards,

Kamatchi

Kamatchi Ulagappan

Springer

Journals Editorial Office (JEO)

JEO Assistant

---

tel +91 44 42197752

fax + 91 44 42197763

Kamatchi.Ulagappan@springer.com

www.springer.com

Ref.: Ms. No. ARVI-D-15-00182R2

Identification and characterization of a novel Geobacillus thermoglucosidasius bacteriophage, GVE3

Dear Mr van Zyl,

Archives of Virology has received your revised submission.

You may check the status of your manuscript by logging onto Editorial Manager at (http://arvi.edmgr.com/).

Best regards,

Springer Journals Editorial Office

Ref.: Ms. No. ARVI-D-15-00182R2

Identification and characterization of a novel Geobacillus thermoglucosidasius bacteriophage, GVE3

Mr Leonardo Joaquim van Zyl

Archives of Virology

Dear Mr van Zyl,

Dr. Horst Neve, Ph.D. has been assigned for the above mentioned manuscript.

http://arvi.edmgr.com/

Your username is: lvzyl

Your password is:

Regards

Editorial Office

Dear Mr van Zyl,

Thank you the revised version and the few modifications requested by the reviewers. Before accepting your manuscript, I would like to ask you to add a few more informations to the legend of Fig. 1 (i.e., micrographs in (A) shown at low magnifications, micrograph in (B) at high magnification, meaning of the arrows (as you did in text).

Your revision is due by 04 Jul 2015.

Please make sure to submit your editable source files (i. e. Word, TeX).

To submit a revision, go to http://arvi.edmgr.com/ and log in as an Author. You will see a menu item call Submission Needing Revision. You will find your submission record there.

Yours sincerely

Horst Neve, Ph.D.

Editor

Archives of Virology

COMMENTS TO THE AUTHOR:

Dear Mr van Zyl,

The PDF for your manuscript, "Identification and characterization of a novel Geobacillus thermoglucosidasius bacteriophage, GVE3" is ready for viewing.

In order to formally submit your manuscript to the journal, you must approve the PDF.

Please access the following web site:

http://arvi.edmgr.com/

Your username is: lvzyl

Your password is:

Click "Author Login".

In your main menu, you will see there is a category entitled "Submission Waiting for Author's Approval". Click on that category, view your submission and approve it. In the unlikely case of conversion issues please contact the Journal's Editorial Office by clicking the "CONTACT US" link on the journal EM home page.

Your manuscript will then be formally submitted to the journal.

Thank you very much.

With kind regards,

Springer Journals Editorial Office

Ref.: Ms. No. ARVI-D-15-00182R3

Identification and characterization of a novel Geobacillus thermoglucosidasius bacteriophage, GVE3

Dear Mr van Zyl,

Archives of Virology has received your revised submission.

You may check the status of your manuscript by logging onto Editorial Manager at (http://arvi.edmgr.com/).

Best regards,

Springer Journals Editorial Office

Ref.: Ms. No. ARVI-D-15-00182R3

Identification and characterization of a novel Geobacillus thermoglucosidasius bacteriophage, GVE3

Mr Leonardo Joaquim van Zyl

Archives of Virology

Dear Mr van Zyl,

Dr. Horst Neve, Ph.D. has been assigned for the above mentioned manuscript.

http://arvi.edmgr.com/

Your username is: lvzyl

Your password is:

Regards

Editorial Office

ARVI-D-15-00182R3

Identification and characterization of a novel Geobacillus thermoglucosidasius bacteriophage, GVE3

Mr Leonardo Joaquim van Zyl

Archives of Virology

Dear Mr van Zyl,

I am pleased to tell you that your work has been scientifically approved; before final acceptance and publication in Archives of Virology, we will check your work regarding taxonomy and language. Please allow us additional 14 days for this purpose.

The paper will be thoroughly checked and edited according to the ICTV taxonomy guidelines. Therefore, please do not make any changes to the use of italics or capitalization in virus names or taxonomic terms in the edited manuscript. For information on correct taxonomy usage, you may consult the section "Scientific Style" in the Instructions to Authors of Archives of Virology.

Lastly, please note that no changes in your work are allowed after online publication.

Thank you for submitting your work to this journal.

With kind regards,

Springer Journals Editorial Office

Dear Mr van Zyl,

We are pleased to inform you that your manuscript, "Identification and characterization of a novel Geobacillus thermoglucosidasius bacteriophage, GVE3", has been accepted for publication in Archives of Virology.

You will receive an e-mail from Springer in due course with regards to the following items:

1. Offprints
2. Colour figures
3. Transfer of Copyright

Please remember to quote the manuscript number, ARVI-D-15-00182R3, whenever inquiring about your manuscript.

With best regards,

Kamatchi Ulagappan

JEO Assistant

Appendix:

Table S1. Predicted open reading frames on GVE3 and closest BLASTp hit on the NCBI database

ORF number	Size in amino acids	Start and end positions bp	Selected BLAST hits and comments; accession number; (length of protein on database in aa)	% Identity/Similarity (over number of aa)
1	724	-(478-2649)	RecD/TraA family helicase *Bacillus cereus* WP_016094953.1 (742)	46/66 (338/486)
2	61	+(3313-3495)	Hypothetical protein; phage SPbeta *Bacillus amyloliquefaciens* subsp. plantarum UCMB5033 YP_008413102.1 (59) / SPBc2 prophage-derived protein YonP *Bacillus amyloliquefaciens* LL3 YP_005546102.1 (59)	41/73 (19/34) 42/75 (19/34)
3	87	+(3495-3755)	Hypothetical protein *Bacillus mycoides* WP_003204337.1 (92)	56/73 (48/63)
4	557	+(3947-5647)	Conserved hypothetical *Salsuginibacillus kocurii* WP_018923841.1 (556) / Major head protein and HOOK domain	34/55 (186/300)
5	91	+(5821-6093)	DNA-binding protein HU-alpha *Anoxybacillus flavithermus* WK1 YP_002315451.1 (101)	74/82 (67/74)
6	400	+(6158-7357)	Hypothetical protein *Ruminococcus gnavus* WP_004840076.1 (415) / PhoH family protein *Spirochaeta africana* DSM 8902 YP_005476454.1 (434)	35/56 (144/235) 29/50 (124/217)
7	64	+(7612-7803)	Hypothetical protein *Bacillus licheniformis* WP_021837902.1 (67) / SPBc2 prophage-derived protein YonK *Bacillus amyloliquefaciens* LL3 YP_005546104.1 (63)	57/79 (36/50) 56/77 (34/47)
8	407	+(7816-9036)	Hypothetical protein *Bacillus licheniformis* WP_021837901.1 (405) / DNA polymerase I *Bacillus* phage Troll YP_008430961.1 (431)	38/58 (152/233) 25/49 (63/126)
9	98	+(9110-9406)	Excinuclease ABC subunit A Candidatus *Protochlamydia amoebophila* WP_011176187.1 (1900)	30/52 (22/38)
10	59	+(9604-9780)	Hypothetical protein *Coprococcus sp*. HPP0074 WP_016438666.1 (57)	41/70 (22/38)
11	539	+(9785-11401)	SPBc2 prophage-derived protein YomD *Bacillus megaterium* WSH-002 YP_005494512.1 (486)	28/45 (116/189)
12	301	+(11444-12346)	Hypothetical protein *Bacillus cereus* WP_016094960.1 / SPBc2 prophage-derived protein YonG *Bacillus sonorensis* WP_006640189.1	37/56 (97/149) 28/46 (70/115)
13	589	+(12336-14102)	Hypothetical protein Bacillus cereus WP_016094961.1 (585) / Putative terminase ATPase subunit *Lactococcus* phage 949 YP_004306307.1 (565)	43/63 (245/362) 37/55 (204/308)
14	502	+(14133-15638)	Hypothetical protein SPBc2p055 *Bacillus* phage SPBc2 NP_046607.1 (506) / C-terminal portal protein domain HK97	26/46 (126/230)
15	475	+(15803-17227)	Hypothetical protein CA_C1132 *Clostridium acetobutylicum* ATCC 824 NP_347765.1 (488) / C-terminal Smc domain; cell division and chromosome partitioning protein	27/49 (123/224)
16	145	+(17280-17714)	Hypothetical protein CA_C1131 *Clostridium acetobutylicum* ATCC 824 NP_347764.1 (147)	36/53 (51/76)
17	327	+(17751-18731)	Hypothetical protein CA_C1130 *Clostridium acetobutylicum* ATCC 824 NP_347763.1 (339)	37/58 (124/196)
18	143	+(18801-19229)	Hypothetical protein *Paenibacillus* WP_009671518.1 (165)	27/54 (32/65)
19	138	+(19244-19657)	Hypothetical protein *Blautia hansenii* WP_003020132.1 (142)	44/62 (52/74)
20	220	+(19667-20326)	Hypothetical protein BATR1942_07955 *Bacillus atrophaeus* 1942 YP_003973461.1 (223)	24/48 (52/106)
21	104	+(20341-20652)	Hypothetical protein SERP1632 *Staphylococcus epidermidis* RP62A YP_189197.1 (105)	35/59 (33/55)

22	158	+(20649-21122)	Hypothetical protein *Lysinibacillus sphaericus* WP_010858764.1 (180)	27/46 (46/81)
23	239	+(21122-21838)	Hypothetical protein *Clostridium* phage D-1873 WP_003377629.1 (247)	35/59 (67/113)
24	250	+(21857-22606)	Hypothetical protein *Paenibacillus dendritiformis* WP_006675026.1 (256)	39/59 (98/157)
25	96	+(22796-23083)	Hypothetical protein *Streptococcus suis* WP_024395613.1 (106)	33/51 (29/45)
26	175	+(23143-23667)	SPBc2 prophage-derived protein YomO *Bacillus amyloliquefaciens* LL3 YP_005545386.1 (162) / C-terminal AAK aspartokinase-like domain	24/53 (38/84)
27	142	+(23668-24093)	SPBc2 prophage-derived protein YomN *Bacillus amyloliquefaciens* subsp. plantarum YAU B9601-Y2 YP_005421066.1 (138) / N-terminal Clusterin domain	46/62 (63/86)
28	326	+(24138-25115)	Hypothetical protein *Bacillus sonorensis* WP_006640209.1 (333) / Phage Integrase family protein *Bacillus subtilis* WP_004399568.1 (333) / INT_REC_C domain	84/91 (272/296) 82/91 (266/298)
29	67	+(25175-25336)	Hypothetical protein *Reinekea blandensis* WP_008045292.1 (242)	41/64 (16/25)
30	203	+(25661-26269)	Hypothetical protein *Anoxybacillus sp.* SK3-4 WP_021094316.1 (196) / Sortase *Lactobacillus gasseri* WP_003652011.1 (176)	38/56 (72/108) 32/49 (56/108)
31	54	+(26256-26471)	Hypothetical protein *Lachnobacterium bovis* WP_029067320.1 (74)	37/56 (15/23)
32	58	-(26621-26794)	Hypothetical protein Clopa_1906 *Clostridium pasteurianum* BC1 YP_007940483.1 (52) / Ribbon-helix-helix DNA binding RHH_3 domain	50/75 (26/39)
33	54	+(26795-26956)	Hypothetical protein	N/A
34	74	+(27017-27238)	Hypothetical protein *Bacillus licheniformis* WP_021837869.1 (78)	30/57 (22/42)
35	513	+(27335-28873)	Hypothetical protein *Paenibacillus elgii* WP_010498387.1 (442)	28/45 (127/203)
36	66	+(29045-29242)	Hypothetical protein *Desulfotomaculum ruminis* WP_013840758.1 (65)	39/61 (17/27)
37	65	+(29208-29402)	Ribonucleotide reductase *Clostridium straminisolvens* JCM 21531 GAE90507.1 (354)	34/55 (23/38)
38	213	+(29435-30073)	Transcriptional modulator of MazE / toxin MazF *Desulfotomaculum acetoxidans* DSM 771 YP_003193197.1 (136) / PemK superfamily	44/64 (54/80)
39	58	+(30618-30791)	Hypothetical protein *Bacillus cereus* WP_002164681.1 (56)	67/81 (36/44)
40	114	+(30799-31140)	Hypothetical protein *Bacillus cereus* WP_000073816.1 (115) / PemK-like protein *Desulfitobacterium dichloroeliminans* LMG P-21439 YP_007221476.1 (125) / PemK superfamily	60/78 (68/90) 48/64 (56/75)
41	253	+(31137-31895)	Putative phage immunity protein; phage SPbeta *Bacillus amyloliquefaciens* TA208 YP_005540881.1 (208)	36/59 (72/120)
42	530	+(31948-33537)	SPbeta phage protein *Bacillus sonorensis* WP_006640220.1 (464) / C-terminal tape_meas_TP901 domain	31/48 (159/253)
43	1914	+(33848-39589)	Lytic transglycosylase *Bacillus subtilis* WP_017696892.1 (2296) / Lytic transglycosylase and goose egg white lysozyme domains	41/59 (443/652)
44	283	+(39653-40501)	Conserved domain protein *Paenibacillus* WP_009671521.1 (284)	50/68 (139/190)
45	397	+(40513-41703)	Hypothetical protein *Paenibacillus* WP_009671508.1 (392) / Flagellin *Lactobacillus mucosae* WP_006501042.1 (680)	56/72 (221/288) 24/44 (86/160)
46	100	+(41684-41983)	Hypothetical protein *Paenibacillus* WP_009671531.1 (448)	49/68 (42/58)
47	337	+(41997-43007)	Hypothetical protein *Paenibacillus* WP_009671533.1 (335) / N-terminal GPI_anchored domain	46/70 (152/235)
48	557	+(43023-44696)	Hypothetical protein *Bacillus cereus* WP_016085028.1 (230) / Carbohydrate-binding CenC domain protein *Exiguobacterium sp.* S17 WP_016510078.1 (433)/ two CBM (carbohydrate binding module) domains	38/57 (88/133) 43/61 (66/93)

49	265	+(44683-45477)	Hypothetical protein *Brevibacillus brevis* WP_017251833.1 (204) / C-terminal DUF_4376 domain	32/50 (82/129)
50	155	+(45483-45947)	Hypothetical protein *Bacillus cereus* WP_016085021.1 (171) / N-terminal DUF_830 domain Orthopoxvirus protein of unknown function	41/62 (63/96)
51	102	+(46003-46308)	Hypothetical protein *Bacillus sp*. AP8 WP_019241528.1 (96) / Holin *Enterococcus faecium* WP_002314927.1 (80) / XhlA domain	63/81 (34/44) 38/62 (30/50)
52	82	+(46314-46559)	Holin *Paenisporosarcina sp*. TG-14 WP_017380421.1 (85) / Phage_holin superfamily domain	56/77 (40/56)
53	314	+(46578-47519)	Peptidase M15 *Bacillus megaterium* WP_016763815.1 (231)/ peptidase M15B and M15C DD-carboxypeptidase VanY/endolysin *Bacillus megaterium* WSH-002 YP_005495499.1 (231) / VanY, Peptidase_M15_4, PG_binding_1 domains	57/74 (128/166) 55/74 (122/223)
54	128	-(47783-48166)	Site-specific recombinase, DNA invertase Pin *Clostridium sp*. BNL1100 YP_005148394.1 (515) / N-terminal Zn_ribbon_recom domain	41/57 (28/35)
55	433	-(48186-49484)	Pyrimidine-nucleoside phosphorylase *Clostridium intestinale* WP_021801470.1 (432) / Glycos_transf_3, PYNP_C superfamily domains	65/80 (279/349)
56	676	-(49481-51508)	Resolvase domain-containing protein *Bacillus amyloliquefaciens* subsp. plantarum YAU B9601-Y2 YP_005420773.1 (569) / N-terminal Ser_recombinase, Zn_ribbon_recom domains	37/58 (199/309)
57	167	-(51782-52282)	Hypothetical protein *Paenibacillus polymyxa* WP_019687525.1 (173) / Putative Holliday junction resolvase *Lactococcus* phage 949 YP_004306283.1 (226) / RuvC_resolvase superfamily domain	57/73 (89/115) 40/63 (65/104)
58	175	-(52367-52891)	CtxB *Vibrio* phage CTX AF516344_3 (124) / Tyrosine kinase family protein *Microcystis aeruginosa* WP_002733350.1 (341)	19/32 (34/57) 34/57 (28/47)
59	250	-(52923-53672)	UvrD/REP helicase *Eggerthella sp*. CAG:1427 WP_021899660.1 (1084)	26/42 (38/61)
60	322	-(54011-55074)	Hypothetical protein Bacillus sp. FJAT-14578 WP_028394443.1 (361) / Contains frame shifts	N/A
61	105	-(55168-55482)	Molybdate ABC transporter, periplasmic molybdate-binding protein *Corynebacterium durum* WP_006062076.1 (222)	39/56 (20/29)
62	106	-(55516-55833)	Hypothetical protein *Bacillus cereus* WP_016085003.1 (102)	46/64 (42/59)
63	80	-(55864-56103)	Hypothetical protein *Desulfovirgula thermocuniculi* WP_027718486.1 (143)	24/53 (17/38)
64	113	-(56125-56463)	Aldehyde oxidase-like *Monodelphis domestica* XP_001379598.1 (1342)	35/56 (19/31)
65	63	-(56472-56660)	Putative uncharacterized protein *Clostridium sp*. CAG:451 WP_022469031.1 (66)	68/80 (41/48)
66	194	-(56673-57254)	Hypothetical protein *Mesorhizobium amorphae* WP_006204368.1 (187) / Ntn_hydrolase superfamily protease-like domain	47/67 (85/122)
67	203	-(57484-58092)	Hypothetical protein *Brevibacillus laterosporus* WP_018672623.1 (266)/ VirB10-like	27/50 (41/77)
68	169	-(58175-58681)	Dihydrofolate reductase *Aneurinibacillus aneurinilyticus* WP_021623827.1 (168) / DHFR superfamily domain	49/67 (81/112)
69	271	-(58682-59494)	Thymidylate synthase *Clostridium difficile* CD196 YP_003213108.1 (276) / TS_Pyrimidine_HMase domain	53/68 (145/189)
70	1366	-(59519-63622)	Putative DNA gyrase B subunit *Clostridium botulinum* D str. 16868 KEH96509.1 (1417) / contains an amber mutation	46/64 (620/875)
71	57	-(63695-63865)	Hypothetical protein CHLNCDRAFT_140300 *Chlorella variabilis* XP_005850790.1 (159)	34/55 (13/21)
72	78	-(63950-64183)	Hypothetical protein *Paenibacillus elgii* WP_010497743.1 (75)	36/60 (20/33)
73	91	-(64231-64503)	Conserved hypothetical protein *Bacillus pumilus* WP_003213202.1 (88)	48/66 (43/59)
74	96	-(64681-64968)	Hypothetical protein *Bacillus methanolicus* WP_004438575.1 (111)	61/76 (59/74)

75	55	-(65005-65169)	Hypothetical protein 0305phi8-36p083 *Bacillus* phage 0305phi8-36 YP_001429809.1 (97)	37/67 (17/31)
76	70	-(65224-65433)	Ferredoxin--NADP reductase *Acaryochloris marina* WP_012163372.1 (296)	47/62 (21/28)
77	52	-(65456-65611)	Pyrimidine-nucleoside phosphorylase *Bacillus cereus* WP_014300201.1 (78)	48/54 (16/18)
78	54	-(65656-65817)	Hypothetical protein BCQ_PT52 *Bacillus cereus* Q1 YP_002533118.1 (53) / DUF3797 domain	67/90 (34/46)
79	116	-(65851-66198)	Hypothetical protein HD73_0395 *Bacillus thuringiensis* serovar kurstaki str. HD73 YP_007419496.1 (157)	58/70 (69/84)
80	82	-(66322-66567)	Putative uncharacterized protein *Ruminococcus* sp. CAG:330 WP_022409890.1 (169)	32/51 (18/29)
81	239	-(66963-67682)	Hypothetical protein *Bacillus nealsonii* WP_016203911.1 (170) / NTP phosphohydrolase domain protein *Bacillus subtilis* subsp. subtilis str. NCIB 3610 YP_008244161.1 (174) / NTP-PPase_YP_001813558 MazG-like domain	64/77 (104/127) 61/74 (97/119)
82	84	-(67754-68005)	Thioredoxin *Bacillus sp.* BT1B_CT2 WP_009328268.1 (83) / NrdH-redoxin family domain	47/69 (37/54)
83	946	-(68046-70883)	Vitamin B12-dependent ribonucleotide reductase *Youngiibacter fragilis* WP_023388736.1 (1012) / RNR_II_dimer domain	47/63 (470/637)
84	77	-(71186-71416)	Hypothetical protein *Amycolatopsis orientalis* WP_016330646.1 (162)	33/53 (23/37)
85	135	-(71573-71977)	Hypothetical protein *Bacillus licheniformis* WP_017474291.1 (121)	34/54 (43/68)
86	76	-(72029-72256)	Hypothetical protein *Bacillus ginsengihumi* WP_025731330.1 (60)	57/71 (32/40)
87	75	-(72935-73159)	Rap GTPase activating protein domain-containing protein 1 *Strongyloides ratti* CEF70831.1 (1573)	38/56 (19/28)
88	66	-(73228-73425)	YjgP/YjgQ family permease *Cellulophaga geojensis* KL-A EWH12381.1 (468)	34/60 (20/35)
89	186	-(73443-74000)	Modification methylase CviRI *Roseburia intestinalis* WP_006855705.1 (150) / Methyltransf_26 domain	46/61 (70/94)
90	292	-(74034-74909)	Hypothetical protein *Enterococcus faecalis* WP_010785554.1 (595) / DNA-cytosine methyltransferase *Pseudoramibacter alactolyticus* WP_006598565.1 (582) / N6_N4_Mtase domain	46/64 (139/193) 44/64 (128/189)
91	113	-(75113-75451)	Hypothetical protein *Geobacillus thermoglucosidasius* WP_003253514.1 (245) / Pep_T-like domain	69/84 (59/73)
92	110	-(75461-75790)	Hypothetical protein *Caldalkalibacillus thermarum* WP_007504287.1 (100)	37/56 (27/41)
93	228	-(76126-76809)	Restriction endonuclease, partial Cannes 8 virus AGV01783.1 (516)	28/47 (21/36)
94	77	-(76806-77036)	Hypothetical protein EF87_21880 *Bacillus amyloliquefaciens* KDN88731.1 (103)	29/53 (18/34)
95	167	-(77033-77533)	SPBc2 prophage-derived protein YorR *Bacillus amyloliquefaciens* subsp. plantarum YAU B9601-Y2 YP_005421143.1 (165) / Thymidylate kinase *Methanotorris formicicus* WP_007043565.1 (185) / dNK domain	55/73 (89/119) 27/44 (53/86)
96	218	-(77759-78412)	3D domain protein *Aneurinibacillus aneurinilyticus* WP_021622015.1 (346) / 3D superfamily domain	48/65 (50/68)
97	1014	-(78425-81469)	Hypothetical protein *Bacillus cereus* WP_016094804.1 (1046) / DNA polymerase III alpha subunit *Clostridium botulinum* B str. Osaka05 BAO04764.1 (1031) / PHP_PolIIIA_DnaE3 domain	64/79 (652/808) 53/71 (538/726)
98	560	-(81495-83174)	yorK protein (Fragment) *Clostridium botulinum* B str. Osaka05 BAO04763.1 (561) / single-stranded DNA exonuclease *Bacillus vallismortis* WP_010331034.1 (576) / DHH family domain	49/67 (275/379) 46/63 (257/355)
99	349	-(83179-84225)	DNA primase *Faecalibacterium prausnitzii* SL3/3 YP_007800891.1 (340) / TOPRIM_DnaG_primases domain	40/59 (136/202)
100	170	-(84240-84749)	numod4 motif family protein *Ruminococcus* sp. CAG:9 WP_022380330.1 (241) / HNH endonuclease *Streptococcus dysgalactiae* subsp. equisimilis AC-2713 YP_006905070.1 (196) / HNH_3 domain	55/72 (88/117) 40/60 (62/93)

101	458	-(84754-86127)	Hypothetical protein *Clostridium bolteae* WP_002573415.1 (482) / Replicative DNA helicase *Paenibacillus elgii* WP_010497788.1 (529) / DnaB domain	40/61 (190/295) 41/60 (195/288)
102	126	-(86188-86565)	Hypothetical protein *Subdoligranulum* sp. 4_3_54A2FAA WP_009323218.1 (235) / phosphate uptake regulator, PhoU *Thermoplasmatales archaeon* SCGC AB-539-N05 WP_008441552.1 (232)	34/52 (40/63) 33/60 (23/42)
103	97	-(86513-86803)	Unique cartilage matrix-associated protein-like *Macaca mulatta* XP_001087282.2 (132)	28/42 (30/46)
104	384	-(86909-88060)	Hypothetical protein *Subdoligranulum* sp. 4_3_54A2FAA WP_009323219.1 (385)	36/58 (137/227)
105	335	-(88113-89117)	Hypothetical protein *Clostridium bolteae* WP_002573419.1 (355)	42/63 (144/215)
106	430	-(89499-90788)	Hypothetical protein *Bacillus cereus* WP_016094813.1 (433) / ATP-dependent DNA ligase *Paenibacillus alvei* WP_005552334.1 (430) / Adenylation_kDNA_ligase_like domain	48/68 (208/296) 48/69 (207/300)
107	74	-(90794-91015)	Hypothetical protein TCA2_4414 *Paenibacillus* sp. TCA20 GAK41922.1 (80)	73/86 (16/19)
108	276	-(91058-91885)	DNA adenine methylase family protein *Clostridium difficile* WP_021425109.1 (276) / Dam domain	61/78 (166/214)
109	85	-(91925-92179)	Aspartate carbamoyltransferase *Pandoraea* sp. B-6 WP_026131998.1 (427)	32/50 (24/37)
110	112	-(92214-92549)	Hypothetical protein *Brevibacillus brevis* WP_017248624.1 (92) / fliH domain	39/59 (30/46)
111	83	-(92772-93020)	Neurabin-1-like *Lepisosteus oculatus* XP_006636653.1 (1370)	32/55 (23/40)
112	74	-(93020-93241)	Resolvase *Salinicoccus carnicancri* WP_017549374.1 (194)	40/60 (20/30)
113	101	-(93298-93600)	Hypothetical protein BMQ_3493 *Bacillus megaterium* QM B1551 YP_003563949.1 (100)	37/48 (34/45)
114	68	-(93663-93866)	Hypothetical protein *Bacillus* phage vB_BanS-Tsamsa YP_008873365.1 (114)	42/60 (25/36)
115	151	-(93901-94353)	Hypothetical protein *Aneurinibacillus aneurinilyticus* WP_021624839.1 (146) / GIY-YIG_UvrC_Cho domain	31/55 (44/79)
116	107	-(94405-94725)	Hypothetical protein C623_0204625 *Bacillus thuringiensis* serovar aizawai str. Hu4-2 ETE99341.1 (100)	28/59 (29/61)
117	115	-(94763-95107)	Hypothetical protein *Bacillus macauensis* WP_007201879.1 (242)	33/58 (38/67)
118	84	-(95137-95388)	Hypothetical protein *Bacillus cereus* WP_001020283.1 (87)	53/74 (42/59)
119	89	-(95430-95696)	Hypothetical protein *Aneurinibacillus aneurinilyticus* WP_021621784.1 (121)	47/68 (40/58)
120	254	-(95731-96492)	Hypothetical protein BRADO3889 *Bradyrhizobium sp.* ORS 278 YP_001205874.1 (102) /	46/66 (30/43)
121	73	-(96531-96749)	PTS mannose transporter subunit IID *Clostridium novyi* B str. ATCC 27606 KEI13252.1 (139)	30/50 (24/40)
122	60	-(96773-96952)	Hypothetical protein BCP78_0087 *Bacillus* phage BCP78 YP_006907922.1 (59)	34/66 (18/35)
123	195	-(97030-97614)	Thymidine kinase *Carnobacterium* sp. AT7 WP_007720733.1 (193) / PRK04296 domain	49/66 (93/125)
124	162	-(97625-98110)	Nucleoside 2-deoxyribosyltransferase *Lysinibacillus boronitolerans* WP_016993282.1 (163) / Nuc_deoxyrib_tr domain	42/65 (68/105)
125	79	-(98132-98343)	Hypothetical protein *Bacillus* phage vB_BanS-Tsamsa YP_008873397.1 (95)	57/78 (39/54)
126	108	-(98340-98663)	Conserved protein of unknown function *Bacillus amyloliquefaciens* subsp. plantarum UCMB5033 YP_008413036.1 (124)	37/57 (38/59)
127	84	-(98680-98931)	Putative RNaseH uncultured marine crenarchaeote HF4000_ANIW97P9 ABZ07137.1 (118)	29/48 (22/37)
128	50	-(98901-99050)	Hypothetical protein	N/A
129	78	-(99116-99349)	Hypothetical protein H839_15993 *Geobacillus stearothermophilus* NUB3621 EZP75017.1 (74)	64/68 (16/17)

130	65	-(99381-99575)	Ribosome small subunit-dependent GTPase A *Acinetobacter calcoaceticus* WP_016139273.1 (353)	47/72 (17/26)
131	56	-(99608-99775)	Hypothetical protein *Bacillus flexus* WP_025909346.1 (135)	67/76 (37/42)
132	63	-(99805-99993)	XRE family transcriptional regulator Serratia sp. Ag2 KFK95694.1 (178)	34/65 (16/31)
133	82	-(100019-100264)	Putative small protein *Oscillatoriales cyanobacterium* JSC-12 WP_009556859.1 (84)	37/82 (21/30)
134	85	-(100318-100575)	Hypothetical protein V529_20360 *Bacillus amyloliquefaciens* SQR9 AHZ16062.1 (87)	31/49 (26/42)
135	155	-(100597-101061)	Ribonuclease H *Sporolactobacillus vineae* WP_010631855.1 (149) / Rnase_HI_prokaryote_like domain	67/82 (99/122)
136	124	-(101095-101466)	Hypothetical protein *Eubacterium plexicaudatum* WP_004067758.1 (107)	44/69 (27/43)
137	65	-(101478-101672)	Cobalt-precorrin-4 C(11)-methyltransferase *Thiocystis violascens* WP_014778112.1 (271)	43/71 (15/25)
138	127	-(101719-102099)	Hypothetical protein GBVE2_gp051 *Geobacillus* virus E2 YP_001285857.1 (128) / YopX family protein *Staphylococcus* phage vB_SepiS-phiIPLA5 YP_006560999.1 (129) / YopX domain	61/75 (80/99) 39/53 (52/71)
139	85	-(102130-102384)	Cytochrome P450 *Fischerella* sp. PCC 9605 WP_026734540.1 (437)	33/44 (25/33)
140	173	-(102426-102944)	AbrB family transcriptional regulator *Lentibacillus jeotgali* WP_010532485.1 (106)	66/75 (29/33)
141	71	-(102951-103163)	Hypothetical protein *Bacillus flexus* WP_025909277.1 (66) / HTH and Lar_restr_allev domains	62/70 (37/42)
142	148	-(103175-103618)	Histidine kinase *Streptomyces megaspores* WP_031506246.1 (420)	30/47 (21/33)
143	55	-(103721-103885)	Hypothetical protein	N/A
144	73	-(103872-104060)	Hypothetical protein *Brevibacillus brevis* WP_017248777.1 (65)	29/59 (18/37)
145	373	-(104319-105440)	Transposase *Desmospora* sp. 8437 WP_009710148.1 (377) / OrfB_IS605 domain	69/83 (262/314)
146	55	+(105536-105700)	Hypothetical protein *Clostridium perfringens* WP_003458470.1 (48) / RHH_3 domain	57/80 (24/34)
147	334	-(106015-107016)	Integrase *Anoxybacillus* sp. DT3-1 WP_009362130.1 (334) / INT_REC_C domain	46/68 (152/227)
148	445	-(107034-108368)	SPBc2 prophage-derived protein YopQ *Anoxybacillus* sp. DT3-1 WP_009362131.1 (445) / DndB superfamily domain	55/74 (244/329)
149	342	-(108440-109465)	Phage integrase family site specific recombinase *Staphylococcus epidermidis* RP62A YP_189166.1 (347) / INT_REC_C domain	33/54 (113/185)
150	151	-(110652-111104)	Putative uncharacterized protein *Firmicutes* bacterium CAG:449 WP_022266857.1 (141)	37/51 (47/66)
151	252	-(111149-111904)	DNA adenine methylase *Alicyclobacillus hesperidum* WP_006446198.1 (267) / Dam domain	43/62 (110/159)
152	255	-(111947-112711)	DNA-cytosine methyltransferase *Bacillus cereus* WP_000934366.1 (248) / N6_N4_Mtase domain	61/77 (154/197)
153	75	-(112743-112967)	Hypothetical protein *Bacillus subtilis* WP_019712282.1 (77)	45/68 (33/50)
154	392	-(113099-114274)	Hypothetical protein *Brevibacillus brevis* WP_017248728.1 (397) / Transposase, IS605 OrfB family *Geobacillus* sp. WCH70 YP_002951068.1 (392) / OrfB_IS605 domain	63/78 (248/309) 41/61 (157/237)
155	233	-(114576-115274)	Hypothetical protein *Bacillus azotoformans* WP_003329177.1 (241) / Nucleotidyltransferase *Bacillus* phage Grass AGY47305.1 (246)	51/68 (115/154) 31/54 (73/126)
156	486	-(115424-116881)	IS transposase *Geobacillus* sp. WCH70 YP_002948967.1 (487) / OrfB_IS605 domain	75/87 (365/428)
157	134	-(116896-117297)	Transposase *Anoxybacillus* sp. SK3-4 WP_021094952.1 (133) / Y1_Tnp domain	84/95 (112/127)
158	76	-(117404-117559)	Hypothetical protein *Bacillales* WP_015252758.1 (172)	28/60 (16/35)
159	84	-(117643-117894)	Hypothetical protein *Anoxybacillus* sp. SK3-4 WP_021095442.1 (83)	87/92 (72/77)

160	63	-(117932-118120)	Flagellar protein FliT *Virgibacillus halodenitrificans* CDQ30829.1 (117)	38/64 (20/34)
161	54	-(118193-118354)	Uncharacterized protein LOC100785018 *Glycine max* XP_006599955.1 (337)	44/66 (16/24)
162	91	-(118392-118664)	Hypothetical protein *Shigella* phage Shf125875 AIM50726.1 (120)	41/57 (22/31)
163	51	-(118797-118952)	Hypothetical protein CPR_C0019 *Clostridium* phage phiSM101 YP_699948.1 (44)	32/75 (13/31)
164	71	-(119206-119418)	Hypothetical protein *Amycolatopsis azurea* WP_005166724.1 (271) / Orthopox_35kD domain	36/58 (18/29)
165	49	-(119387-119533)	Hypothetical protein DJ51_5110 *Bacillus cereus* KFL86206.1 (39)	61/75 (22/27)
166	49	-(119538-119684)	Hypothetical protein JCM16418_5101 *Paenibacillus pini* JCM 16418 GAF10872.1 (75)	47/68 (21/31)
167	82	-(119614-119856)	Hypothetical protein *Kineococcus radiotolerans* WP_011981686.1 (85)	29/56 (16/31)
168	52	-(119890-120045)	Hypothetical protein CRE_08251 *Caenorhabditis remanei* XP_003109456.1 (241)	47/61 (16/21)
169	76	-(120076-120303)	Hypothetical protein *Bacillus thuringiensis* WP_030030167.1 (109)	49/73 (37/55)
170	61	-(120344-120526)	Hypothetical protein *Catenulispora acidiphila* WP_012787201.1 (61)	42/50 (25/30)
171	54	-(120526-120678)	CRISPR-associated protein Csm1 *Thermococcus onnurineus* WP_012571853.1 (777) / Cas10_III domain	38/55 (17/25)
172	83	-(120747-120986)	Hypothetical protein *Paenibacillus elgii* WP_010499937.1 (86)	36/65 (28/51)
173	382	-(121143-122288)	Hypothetical protein *Paenibacillus polymyxa* WP_019687721.1 (340)	25/44 (55/96)
174	320	-(122590-123549)	Hypothetical protein GWCH70_2831 *Geobacillus* sp. WCH70 YP_002950780.1 (321) / HTH_36 domain	65/78 (211/254)
175	65	-(123612-123806)	Hypothetical protein Bsph_1942 *Lysinibacillus sphaericus* C3-41 YP_001697661.1 (61)	52/80 (32/49)
176	212	-(124164-124799)	Hypothetical protein *Staphylococcus aureus* WP_016187599.1 (226) / Beta_clamp superfamily domain	34/51 (41/63)
177	94	-(124842-125123)	Hypothetical protein *Alicyclobacillus contaminans* WP_026973934.1 (159)	57/71 (20/25)
178	390	-(125137-126306)	Hypothetical protein *Paenibacillus elgii* WP_010499918.1 (376) / COG6 domain	38/63 (76/126)
179	61	-(126354-126536)	Hypothetical protein *Bacteroides* WP_004325856.1 (92)	35/58 (16/27)
180	62	-(126559-126744)	Hypothetical protein *Desulfotomaculum kuznetsovii* WP_013823457.1 (116)	39/61 (14/22)
181	55	-(126772-126936)	Hypothetical protein IscW_ISCW014631 *Ixodes scapularis* XP_002414485.1 (369) / COG3264 domain	41/56 (19/26)
182	60	-(126990-127169)	Hypothetical protein	N/A
183	88	-(127240-127461)	Hypothetical protein CANTEDRAFT_116679 *Candida tenuis* ATCC 10573 XP_006689815.1 (236)	38/57 (17/26)
184	151	-(127652-128104)	Appr-1-p processing protein *Paenibacillus lactis* WP_007128434.1 (149) / Macro_Poa1p_like domain	58/74 (87/111)
185	281	-(128253-129095)	Hypothetical protein *Clostridium botulinum* WP_003374316.1 (254) / toxin-antitoxin system, toxin component, Bro family *Clostridium hathewayi* WP_006775691.1 (279) / ORF6N domain	49/67 (59/82) 40/60 (60/90)
186	1048	-(129429-132572)	Hypothetical protein *Bacillus licheniformis* WP_017474472.1 (644)	43/65 (111/170)
187	59	-(132699-132875)	Hypothetical protein *Desulfotomaculum alcoholivorax* WP_027363725.1 (79)	30/48 (15/24)
188	64	-(132673-132864)	Hypothetical protein *Acinetobacter junii* WP_004954691.1 (420)	42/60 (16/23)
189	70	-(133553-133762)	Arginyl-tRNA synthetase *Pseudomonas syringae* KFE56075.1 (578)	38/56 (18/27)
190	61	-(133979-134161)	Uncharacterized protein LOC101122157 *Ovis aries* XP_004003698.1 (250)	30/60 (18/36)
191	98	-(134262-134555)	Transitional endoplasmic reticulum ATPase TER94-like *Apis florea* XP_003692162.1 (893)	32/47 (23/34)

192	103	-(134657-134929)	Luciferase *Mycobacterium* sp. 360MFTsu5.1 WP_029105512.1 (282)	36/55 (20/31)
193	67	-(135051-135251)	Unknown protein Candidatus *Kuenenia stuttgartiensis* CAJ74213.1 (66)	38/49 (21/27)
194	70	-(135354-135563)	Uncharacterized lipoprotein ymbA *Xenorhabdus poinarii* G6 CDG21808.1 (206)	32/50 (17/27)
195	172	-(135584-136099)	5 nucleotidase deoxy cytosolic type C Firmicutes bacterium CAG:582 WP_022178382.1 (208)	36/58 (26/42)
196	199	-(136158-136754)	Hypothetical protein BAUCODRAFT_38792 *Baudoinia compniacensis* UAMH 10762 EMC91680.1 (339)	33/49 (24/36)
197	212	-(136824-137459)	Hypothetical protein *Brevibacillus brevis* WP_017248816.1 (219)	45/63 (98/138)
198	121	-(137538-137900)	Hypothetical protein *Bacillus cereus* WP_000787323.1 (75)	46/68 (31/46)
199	68	-(137981-138184)	ATP-binding cassette sub-family A member 3-like *Trichechus manatus latirostris* XP_004373510.1 (1758)	33/58 (22/39)
200	59	-(138229-138405)	DNA topoisomerase I *Bacteroides uniformis* WP_005834124.1 (715)	45/55 (22/27)
201	372	-(139086-140201)	Hypothetical protein CTC01563 *Clostridium tetani* E88 NP_782174.1 (419)	43/56 (165/217)
202	179	-(140391-140927)	Phosphate ABC transporter permease *Thioalkalivibrio* sp. ARh3 WP_018864092.1 (555)	26/54 (22/46)

Table S2

Bacterial strain / note	CRISPR spacer sequence	CRISPR repeat unit sequence	Location on GVE3
G. thermoglucosidasius **-11955**			
	(C)ATTCAACAACAGG[G/A]GAA[G/-]AAAAAGA[A/C]CTTCACACAA	GTTTCAATTCCTTATAGGTAAGATACAAAC	Intergenic 135200bp
	(GTGG)GATGGC[G/A]AC[C/T]T[A/G]C[G/T]AC[G/]CTGATGATGGCAAGCC(GA)	GTTTCAATTCCTTATAGGTAAGATACAAAC	Intergenic 138460bp
G. thermoglucosidasius **-C56-YS93**	AACAA[G/T]CGCAAAGGTTT[A/C]A[A/CG]-TTTTCCTTTTT[C/T]AA[G/A]CGT	GTTTGTATCTTACCTATGAGGAATTGAAAC	Inside ORF184
	ATAC[T/C]A[G/A]ACTTTCTTT[G/A]TATTGTGCGTATGGCTCGT	GTTTGTATCTTACCTATGAGGAATTGAAAC	Inside ORF195
	GCCAAATTTTTTATCTATCCAAGAGTAGCACCTT(TCC)	GTTTGTATCTTACCTATGAGGAATTGAAAC	Intergenic 139000bp
	TCGACATCAGGAATTTCGTCGATAAATACTTTGAA(TAT)	GTTTGTATCTTACCTATGAGGAATTGAAAC	Intergenic 134700bp
	TGAGAACATAAGCGAATTTTCCATTCGAG[A/C]A[T/A]ATT	GTTTGTATCTTACCTATGAGGAATTGAAAC	Start of ORF8
	(TAA)TAA[T/C]TGTAAAATCT[AC/GT]TTTAATACTGGT(GCGCC)	GTTTGTATCTTACCTATGAGGAATTGAAAC	Inside ORF196
	TTATATCTTC[GC/AT]T[A/G]TTAAAG[T/C]AGTCATGCCATCTG	GTTTCAATTCCTTATAGGTAAGATAAAACC	Inside ORF196
	(TA)TA[T/C]TTTGC[G/A]CAAATGAATACAATTGAA[T/C]C[A/T]ATTGG(G)	GTTTCAATTCTTCATAGGTAAGATAAAAAC	Inside ORF146
	CTTTTAG[C/T]TTCATATTGCTTGA[G/A]CCACG[A/G]AG[A/C]GAAGT	GTTTCAATTCCTCATAGGTAAGATACAAAC	Inside ORF69
	GCTCA[T/A]A[TTT/AAC]GCC[TT/CA]ATTTTGCGTTCTAGATG[A/G]CTTCC	GTTTCAATTCCTCATAGGTAAGATACAAAC	Intergenic 133400bp
	AGCGTCTG[GG/AA]AGCGTGTGAAG[T/G]TGATAGG[A/T]AAAG(GA)	GTTTCAATTCCTCATAGGTAAGATACAAAC	Intergenic 133420bp
	(A)TG[G/A]TGAAA[C/A]GAAA[AATA/GGAG]TTGTGAGAAGAGT(CTTGA)	GTTTCAATTCCTCATAGGTAAGATACAAAC	Inside ORF4
	TATGATCCTCCCTTTTCTGTACAATACCTTAAACTT	GTTTTATCTGAACGTAGTGGGATATAAAG	Start of ORF32
	ATGAC[T/C]GC[TG/AA]TTGATTGGAAAGC[A/C]GG[C/T]AT[T/C]CCAAA(TG)	GTTTTATCTGAACGTAGTGGGATATAAAG	ORF42
	AAAGAAGCTTT[G/A]CAAGAATATATTGGAAAAATGGA	GTTTTATCTGAACGTAGTGGGATATAAAG	Inside ORF40
G. thermoglucosidasius **- TNO-09.020**	GCGTGTGAAGTGG[G/A]TAGGT[G/-]AA[G/A]GA[G/T]AAAA[C/T]AAAA	GTTTCAATTCCTTATAGGTAAGATACAAAC	Intergenic 133420bp

	G[G/A]ACACCTGCAACCTAACTAAAT[A/-]AAA[C/T]GAATGGAGGAA	GTTTCAATTCCTCATAGGTAAGATACAAAC	Intergenic, beginning of ORF193
	AT[A/C]GCATCCCAACGATTATCATCACCACTATAAGT	GTTTCAATTCCTCATAGGTAAGATACAAAC	Inside ORF44
	CTACATACTTTTTGTGACTCCATGACTT[T/C][C/GG][TA/GG]CGTT	GTTTCAATTCCTCATAGGTAAGATACAAAC	Inside ORF193
***G. toebii* - WCH70**	ATACTGG[CG/TA]CTCCACCGTTATCC[AT/--][A/C]TATTTTTGT	GTTTTATCTTACCTATGAGGAATTGAAAC	Inside ORF196
	CCATCATCAGGTAGCAAGTTGCCATCTTGCTACGACAAG	GTTTGTATCTTAACTATGAGGAATTGAAAC	Intergenic 138330
	TTGACAGGATATTGACCAAGCTCACCCCGTCTGCCCG	CTTTATATCCCACTACGTTCAGATAAAAC	End of ORF 143
***G. thermoglucosidasius* - Y4.1MC1**	GGATTAGTTGGC[G/A][C/TAG]---[TG/G][T/TTTAG][G/T][AC/A]ATT	GTTTGTATCTTACCTATGAGGAATTGAAAC	Inside ORF38
	[T/C]TTTTCACTAAT[A/G]AA[A/G]AAGTCTGGATAGGATTGTTG	GTTTCAATTTTCCTTATAGGTAAGATAAAAAC	Inside ORF50
	GTCGAATGACGAACG[A-/CC]AGTGAGGAATGAGACAAGCA	GTTTCAATTTTCCTTATAGGTAAGATAAAAAC	Intergenic 138960
	(AT)GGTGGAGT[G/A]CCAGTATTAAA[G/A]CAGAT[AC/TT]TACA(AT)	GTTTCAATTCCTCATAGGATACAAAC	Inside ORF196
	TT[T/C]CCGTGTAT[C/G][CG/GA]TTATTTACATATGCACGAAACT	GTTTCAATTCCTCATAGGATACAAAC	Inside ORF136
	GAATTTGGAAAATCAAGAGCGCAATA[T/C]TCAGCAGA	GTTTCAATTCCTCATAGGATACAAAC	Inside ORF95

	GCTCA(T/A)A(TTT/AAC)GCC(TT/CA)ATTTTGCGTTCTAGATG(A/G)CTTCC	GTTTCAATTCCTCATAGGTAAGATACAAAC	Intergenic 133400bp
	AGCGTCTG(GG/AA)AGCGTGTGAAG(T/G)TGATAGG(A/T)AAAG(GA)	GTTTCAATTCCTCATAGGTAAGATACAAAC	Intergenic 133420bp
	(A)TG(G/A)TGAAA(C/A)GAAA(AATA/GGAG)TTGTGAGAAGAGT(CTTGA)	GTTTCAATTCCTCATAGGTAAGATACAAAC	Inside ORF4
	TATGATCCTCCCTTTTCTGTACAATACCTTAAACTT	GTTTTATCTGAACGTAGTGGGATATAAAG	Start of ORF32
	ATGAC(T/C)GC(TG/AA)TTGATTGGAAAGC(A/C)GG(C/T)AT(T/C)CCAAA(TG)	GTTTTATCTGAACGTAGTGGGATATAAAG	ORF42
	AAAGAAGCTTT(G/A)CAAGAATATATTGGAAAAATGGA	GTTTTATCTGAACGTAGTGGGATATAAAG	Inside ORF40
***G. thermoglucosidasius* - TNO-09.020**	GCGTGTGAAGTGG(G/A)TAGGT(G/-)AA(G/A)GA(G/A)AAAA(C/T)AAAA	GTTTCAATTCCTTATAGGTAAGATACAAAC	Intergenic 133420bp
	G(G/A)ACACCTGCAACCTAACTAAAT(A/-)AAA(C/T)GAATGGAGGAA	GTTTCAATTCCTCATAGGTAAGATACAAAC	Intergenic, beginning of ORF193
	AT(A/C)GCATCCCAACGATTATCATCACCACTATAAGT	GTTTCAATTCCTCATAGGTAAGATACAAAC	Inside ORF44
	CTACATACTTTTTGTGACTCCATGACTT(T/C)(TA/GG)CGTT	GTTTCAATTCCTCATAGGTAAGATACAAAC	Inside ORF193
***G. toebii* - WCH70**	ATACTGG(CG/TA)CTCCACCGTTATCC(AT/--)A(A/C)TATTTTTGT	GTTTTATCTTACCTATGAGGAATTGAAAC	Inside ORF196
	CCATCATCAGGTAGCAAGTTGCCATCTTGCTACGACAAG	GTTTGTATCTTAACTATGAGGAATTGAAAC	Intergenic 138330
	TTGACAGGATATTGACCAAGCTCACCCCGTCTGCCCG	CTTTATATCCCACTACGTTCAGATAAAAC	End of ORF 143

***G. thermoglucosidasius* - Y4.1MC1**	GGATTAGTTGGC$^{G}_{C}$ $^{---}$ ^{T} TG TTTAG$^{G}_{A}$C$^{C}_{T}$ATT A TAG G T A	GTTTGTATCTTACCTATGAGGAATTGAAAC	Inside ORF38
	$^{T}_{C}$TTTTCACTAAT$^{A}_{G}$AA$^{A}_{G}$AAGTCTGGATAGGATTGTTG	GTTTCAATTTTCCTTATAGGTAAGATAAAAAC	Inside ORF50
	GTCGAATGACGAACG$^{A-}_{CC}$AGTGAGGAATGAGACAAGCA	GTTTCAATTTTCCTTATAGGTAAGATAAAAAC	Intergenic 138960
	(AT)GGTGGAGT$^{G}_{A}$CCAGTATTAAA$^{G}_{A}$CAGAT$^{AC}_{TT}$TACA(AT)	GTTTCAATTCCTCATAGGATACAAAC	Inside ORF196
	TT$^{T}_{C}$CCGTGTAT$^{C}_{G}$G$^{CG}_{GA}$TTATTTACATATGCACGAAACT	GTTTCAATTCCTCATAGGATACAAAC	Inside ORF136
	GAATTTGGAAAATCAAGAGCGCAATA$^{T}_{C}$TCAGCAGA	GTTTCAATTCCTCATAGGATACAAAC	Inside ORF95

- Subscript sequence corresponds to phage DNA sequence

Table S3

Bacterium / Phage	Areas of similarity (length in bp)		Function encoded on GVE3	Percentage identity/gaps
	GVE3	Bacterium / Phage		
***Geobacillus toebii* WCH70**	115262-115412 (150)	2854467-2854315 (152)*	Sequence associated with IS elements	96/0
		2218799-2218949 (150)		96/0
		666917-666768 (149)		96/0
		965841-965692 (149)		96/0
		1675325-1675474 (149)		96/0
		1799931-1800080 (149)*		96/0
		1997961-1998110 (149)*		96/0
		2036474-2036623 (149)		96/0
		2699836-2699687 (149)		96/0
		2930818-2930669 (149)		96/0
		1258531-1258682 (151)*		94/1
		882630-882784 (154)*		87/0
		1582295-1582146 (149)		96/0
		1541709-1541560 (149)		96/0
		1427576-1427426 (150)*		96/0
		1675325-1675474 (149)		96/0
		537200-537048 (152)*		95/0
	115261-117428 (2167)	884159-886181 (2022)	IS elements	77/0
		896311-894137 (2174)		74/0
		1988484-1989797 (1313)		77/0
		2069551-2071616 (2065)		72/2
	115262-115827 (565)	506720-506156 (564)	IS element	90/0
	122344-123507 (1163)	2863055-2864228 (1173)	HTH containing ORF	72/5

	48714-49484 (770)	2306426-237196 (770)	N-terminal of Pyrimidine nucleoside phosphorylase	70/0
	5914-6093 (179)	2219251-2219072 (179)	IHF-like protein	81/0
	5499-5698 (199)	2869300-2869094 (206)	C-terminal of cons. hypo. protein	78/3
	104247-104483 (236)	2150056-2150293 (237)	C-terminal of ORF100 (IS605) and sequence downstream of it	73/0
	138448-138486 (38)	360836-360874 (38)	Upstream of four conserved hypothetical proteins	100/0
	104298-104334 (36)	2061807-2061771 (36)	Sequence downstream of ORF100	100/0
***Clostridium botulinum* B str. Osaka05 DNA, contig: Osaka05p1_contig002, extrachromosome 1**	78895-82892 (3997)	19479-23480 (4001)	DNA pol III alpha subunit and Single stranded exonuclease	65/4
***Bacillus* phage vB_BanS-Tsamsa**	79026-81468 (2442)	97875-100338 (2463)	DNA pol III alpha subunit	69/2
	59838-62400 (2562)	61548-64155 (2607)	DNA gyrase A and B	67/4
	33358-33413 (55)	127875-127930 (55)	Tape measure protein	80/0
***Bacillus* phage SPBc2**	24098-25107 (1009)	40307-41318 (1011)	Integrase-like	75/0
	36326-36985 (659)	32687-33350 (663)	Tape measure protein	71/4
	5927-6090 (163)	60843-61007 (164)	IHF-like protein	77/2
	82724-83165 (441)	105199-105640 (441)	N-terminal of Single stranded exonuclease	64/4
***Clostridium* phage c-st**	115428-116985 (1557)	4081-5668 (1587)	IS element	70/3
		16518-18105 (1587)		70/3
		158521-160108 (1587)		70/3
	115615-116519 (904)	168684-169567 (883)	IS element	69/6
	79711-81489 (1778)	51559-53317 (1758)	DNA pol III alpha subunit	66/3

	78922-79273 (351)	53761-54112 (351)	DNA pol III alpha subunit	68/2
	60218-60300 (82)	90017-89935 (82)	DNA gyrase	76/0

Table S4. PHAST annotation of three regions of phage GVE3

Region 1, total: 35 CDS.

ORF #	CDS_POSITION	BLAST_HIT	E-VALUE
1	3313..3495	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p064; PP_00004; phage(gi9630189)	3e-06
2	3495..3755	PHAGE_Staphy_Twort_NC_007021: ORF137; PP_00005; phage(gi66391382)	4e-08
3	3947..5647	hypothetical protein SERP1620 [Staphylococcus epidermidis RP62A] gi|57867549|ref|YP_189185.1|; PP_00006	4e-56
4	5354..5365	attL ATTCGGGATATG	0.0
5	5821..6093	PHAGE_Bacill_SPBc2_NC_001884: histone-like prokaryotic DNA-binding protein family; PP_00007; phage(gi9630187)	7e-29
6	6164..7357	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00008; phage(gi564292570)	1e-72
7	7612..7803	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p061; PP_00009; phage(gi9630186)	7e-14
8	7816..9036	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p060; PP_00010; phage(gi9630185)	2e-68
9	9110..9406	hypothetical; PP_00011	0.0
10	9589..9780	phage protein [Bacillus licheniformis DSM 13 = ATCC 14580] gi|404488823|ref|YP_006712929.1|; PP_00012	3e-05
11	9830..11401	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p027; PP_00013; phage(gi9630152)	2e-27
12	11444..12346	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p057; PP_00014; phage(gi9630182)	2e-13
13	12339..14102	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p056; PP_00015; phage(gi9630181)	6e-58
14	14133..15638	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p055; PP_00016; phage(gi9630180)	2e-34
15	15649..15771	hypothetical; PP_00017	0.0
16	15803..17227	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p054; PP_00018; phage(gi9630179)	2e-23
17	17280..17714	hypothetical protein CEA_G1142 [Clostridium acetobutylicum EA 2018] gi|384457855|ref|YP_005670275.1|; PP_00019	3e-13
18	17751..18731	PHAGE_Bacill_Slash_NC_022774: hypothetical protein; PP_00020; phage(gi609217106)	6e-10
19	18801..19229	hypothetical protein CEA_G1140 [Clostridium acetobutylicum EA 2018] gi|384457853|ref|YP_005670273.1|; PP_00021	1e-05
20	19244..19657	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00022; phage(gi564292646)	3e-15
21	19667..20326	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p048; PP_00023; phage(gi9630173)	3e-06

22	20341..20652	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00024; phage(gi564292696)	7e-09
23	20649..21122	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p047; PP_00025; phage(gi9630172)	3e-06
24	21122..21838	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p046; PP_00026; phage(gi9630171)	5e-05
25	21857..22606	PHAGE_Lactoc_949_NC_015263: putative phage structural protein; PP_00027; phage(gi327197979)	2e-35
26	22670..23083	hypothetical; PP_00028	0.0
27	23143..23667	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p038; PP_00029; phage(gi9630163)	1e-06
28	23668..24093	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p037; PP_00030; phage(gi9630162)	1e-21
29	23960..23971	attR ATTCGGGATATG	0.0
30	24114..25115	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p036; PP_00031; phage(gi9630161)	2e-164
31	25136..25336	hypothetical; PP_00032	0.0
32	25517..25639	hypothetical; PP_00033	0.0
33	25661..26269	hypothetical protein Tresu_1654 [Treponema succinifaciens DSM 2489] gi|328948512|ref|YP_004365849.1|; PP_00034	2e-16
34	26256..26471	hypothetical; PP_00035	0.0
35	complement(26621..26794)	hypothetical protein Clopa_1906 [Clostridium pasteurianum BC1] gi|488770689|ref|YP_007940483.1|; PP_00036	2e-06
36	27017..27238	hypothetical; PP_00037	0.0
37	27350..28873	PHAGE_Staphy_CNPH82_NC_008722: conserved phage protein; PP_00038; phage(gi119953709)	5e-07

Region 2, total : 83 CDS.

ORF #	CDS_POSITION	BLAST_HIT	E-VALUE
1	complement(56673..57254)	PHAGE_Vibrio_12B8_NC_021073: hypothetical protein; PP_00075; phage(gi481019685)	2e-21
2	complement(57303..57482)	hypothetical; PP_00076	0.0
3	complement(57484..58092)	hypothetical protein ERIC2_c26990 [Paenibacillus larvae subsp. larvae DSM 25430] gi|568264958|ref|YP_008968433.1|; PP_00077	1e-11
4	complement(58175..58681)	PHAGE_Bacill_phiAGATE_NC_020081: putative dihydrofolate reductase; PP_00078; phage(gi448260875)	1e-35

5	complement(58682..59494)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00079; phage(gi564292594)	2e-45
6	complement(59519..61651)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00080; phage(gi564292556)	0.0
7	complement(61664..63622)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00081; phage(gi564292558)	4e-160
8	complement(63695..63865)	hypothetical; PP_00082	0.0
9	complement(63950..64183)	hypothetical; PP_00083	0.0
10	complement(64231..64503)	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p155; PP_00084; phage(gi9630280)	8e-15
11	complement(64524..64679)	hypothetical; PP_00085	0.0
12	complement(64681..64968)	PHAGE_Bacill_SP10_NC_019487: hypothetical protein; PP_00086; phage(gi418489661)	4e-15
13	complement(65005..65160)	hypothetical; PP_00087	0.0
14	complement(65224..65388)	hypothetical; PP_00088	0.0
15	complement(65456..65593)	hypothetical; PP_00089	0.0
16	complement(65656..65817)	PHAGE_Clostr_CDMH1_NC_024144: conserved hypothetical protein; PP_00090; phage(gi640884924)	2e-07
17	complement(65851..66198)	PHAGE_Strept_phiBHN167_NC_022791: phage protein; PP_00091; phage(gi557745672)	1e-15
18	complement(66322..66567)	hypothetical; PP_00092	0.0
19	complement(66602..66856)	hypothetical; PP_00093	0.0
20	complement(66963..67682)	PHAGE_Cyanop_NATL1A_7_NC_016658: gp32; PP_00094; phage(gi372217788)	6e-08
21	complement(67754..68005)	PHAGE_Bacill_SPBc2_NC_001884: thioredoxin; PP_00095; phage(gi9630289)	2e-14
22	complement(68046..70883)	PHAGE_Halovi_HVTV_1_NC_020158: ribonucleotide reductase alpha subunit; PP_00096; phage(gi443404588)	2e-38
23	complement(70913..71161)	hypothetical; PP_00097	0.0
24	complement(71186..71416)	hypothetical; PP_00098	0.0
25	complement(71573..71977)	PHAGE_Bacill_BCP78_NC_018860: hypothetical protein; PP_00099; phage(gi410492830)	4e-13
26	complement(72029..72214)	hypothetical; PP_00100	0.0
27	complement(72405..72572)	hypothetical; PP_00101	0.0
28	complement(72606..72857)	hypothetical; PP_00102	0.0
29	complement(72935..73159)	hypothetical; PP_00103	0.0
30	complement(73228..73416)	hypothetical; PP_00104	0.0
31	complement(73443..74000)	PHAGE_Cronob_CR9_NC_023717: putative DNA methyltransferase; PP_00105; phage(gi593777337)	1e-18
32	complement(74034..74909)	PHAGE_Cellul_phi12:1_NC_021791: DNA methylase; PP_00106; phage(gi526177136)	2e-55
33	complement(74906..75094)	hypothetical; PP_00107	0.0

34	complement(75113..75451)	hypothetical; PP_00108	0.0
35	complement(75461..75790)	hypothetical; PP_00109	0.0
36	complement(75910..76077)	PHAGE_Halovi_HCTV_1_NC_021330: hypothetical protein; PP_00110; phage(gi509140762)	1e-06
37	complement(76126..76809)	hypothetical; PP_00111	0.0
38	complement(76806..77036)	hypothetical; PP_00112	0.0
39	complement(77033..77533)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00113; phage(gi564292628)	1e-24
40	complement(77759..78412)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00114; phage(gi564292603)	9e-08
41	complement(78425..81469)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00115; phage(gi564292551)	0.0
42	complement(81495..83174)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00116; phage(gi564292561)	2e-52
43	complement(83179..84225)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00117; phage(gi564292575)	4e-39
44	complement(84240..84749)	PHAGE_Xantho_Xp10_NC_004902: endonuclease of the HNH family with predicted DNA-binding module at C-terminus; PP_00118; phage(gi32128470)	3e-24
45	complement(84754..86175)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00119; phage(gi564292562)	2e-12
46	complement(86188..86562)	hypothetical; PP_00120	0.0
47	complement(86513..86848)	hypothetical; PP_00121	0.0
48	complement(86909..88060)	PHAGE_Clostr_c_st_NC_007581: hypothetical protein CST056; PP_00122; phage(gi80159742)	1e-09
49	complement(88113..89117)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00123; phage(gi564292571)	5e-18
50	complement(89499..90782)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: DNA ligase, ATP-dependent; PP_00124; phage(gi564292567)	1e-74
51	complement(90794..91015)	hypothetical; PP_00125	0.0
52	complement(91058..91885)	PHAGE_Parame_bursaria_Chlorella_virus_NY2A_NC_009898: hypothetical protein NY2A_B774R; PP_00126; phage(gi157953078)	3e-29
53	complement(91925..92164)	hypothetical; PP_00127	0.0
54	complement(92214..92549)	hypothetical; PP_00128	0.0
55	complement(92578..92745)	hypothetical; PP_00129	0.0
56	complement(92772..93020)	hypothetical; PP_00130	0.0
57	complement(93020..93241)	hypothetical; PP_00131	0.0
58	complement(93298..93600)	hypothetical protein BMD_3488 [Bacillus megaterium DSM 319] gi|295705603|ref|YP_003598678.1|; PP_00132	2e-07
59	complement(93663..93860)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00133; phage(gi564292684)	2e-06
60	complement(93901..94323)	hypothetical; PP_00134	0.0
61	complement(94405..94725)	PHAGE_Bacill_Spock_NC_022763: hypothetical protein; PP_00135; phage(gi568190861)	5e-05

62	complement(94763..95107)	hypothetical; PP_00136	0.0
63	complement(95137..95388)	hypothetical; PP_00137	0.0
64	complement(95430..95696)	hypothetical; PP_00138	0.0
65	complement(95731..96492)	PHAGE_Acanth_mimivirus_NC_014649: hypothetical protein; PP_00139; phage(gi311978204)	4e-08
66	complement(96531..96749)	hypothetical; PP_00140	0.0
67	complement(96773..96919)	hypothetical; PP_00141	0.0
68	complement(97030..97614)	PHAGE_Entero_vB_EcoM_VR7_NC_014792: Tk thymidine kinase; PP_00142; phage(gi314121676)	9e-30
69	complement(97625..98110)	PHAGE_Lactoc_949_NC_015263: putative nucleoside-2-deoxyribosyltransferase; PP_00143; phage(gi327197942)	1e-23
70	complement(98107..98343)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00144; phage(gi564292710)	1e-18
71	complement(98340..98663)	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p126; PP_00145; phage(gi9630251)	2e-10
72	complement(98680..98931)	hypothetical; PP_00146	0.0
73	complement(98901..99050)	hypothetical; PP_00147	0.0
74	complement(99116..99349)	hypothetical; PP_00148	0.0
75	complement(99381..99575)	hypothetical; PP_00149	0.0
76	complement(99608..99775)	hypothetical; PP_00150	0.0
77	complement(99805..99993)	hypothetical; PP_00151	0.0
78	complement(100019..100264)	hypothetical; PP_00152	0.0
79	complement(100318..100575)	hypothetical; PP_00153	0.0
80	complement(100597..101061)	PHAGE_Ostreo_OlV1_NC_014766: hypothetical protein; PP_00154; phage(gi313844138)	2e-26
81	complement(101095..101466)	hypothetical; PP_00155	0.0
82	complement(101478..101672)	hypothetical; PP_00156	0.0
83	complement(101719..102099)	PHAGE_Geobac_virus_E2_NC_009552: hypothetical protein GBVE2_gp051; PP_00157; phage(gi148747778)	2e-42

Region 3, total : 22 CDS.

ORF #	CDS_POSITION	BLAST_HIT	E-VALUE
1	89396..89412	attL TTTTATATTTTATTTAA	0.0

2	complement(104319..105440)	PHAGE_Clostr_c_st_NC_007581: putative IS transposase (OrfB); PP_00163; phage(gi80159731)	4e-99
3	105536..105700	hypothetical protein [Acetohalobium arabaticum DSM 5501] gi|302391636|ref|YP_003827456.1|; PP_00164	8e-05
4	complement(106015..107016)	PHAGE_Clostr_c_st_NC_007581: conserved hypothetical phage-related protein; PP_00165; phage(gi80159716)	2e-21
5	complement(107034..108377)	PHAGE_Clostr_c_st_NC_007581: conserved hypothetical phage-related protein; PP_00166; phage(gi80159715)	3e-15
6	complement(108440..109465)	PHAGE_Lactoc_phiL47_NC_023574: putative integrase-recombinase; PP_00167; phage(gi589890760)	2e-31
7	complement(109487..109648)	hypothetical; PP_00168	0.0
8	complement(109725..109907)	hypothetical; PP_00169	0.0
9	complement(110501..110674)	hypothetical; PP_00170	0.0
10	complement(110652..111104)	PHAGE_Strept_K13_NC_024357: phage protein; PP_00171; phage(gi658307253)	2e-05
11	complement(111149..111904)	PHAGE_Clostr_phiSM101_NC_008265: putative modification methylase dpniia; PP_00172; phage(gi110804053)	7e-52
12	complement(111947..112711)	PHAGE_Geobac_GBK2_NC_023612: DNA methylase; PP_00173; phage(gi589893811)	5e-82
13	complement(112743..112967)	hypothetical; PP_00174	0.0
14	complement(113099..114274)	PHAGE_Staphy_vB_SauM_Remus_NC_022090: transposase; PP_00175; phage(gi530787614)	1e-11
15	complement(114352..114498)	hypothetical; PP_00176	0.0
16	complement(114576..115274)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00177; phage(gi564292596)	8e-29
17	complement(115424..116881)	PHAGE_Clostr_c_st_NC_007581: putative IS transposase (OrfB); PP_00178; phage(gi80159857)	6e-175
18	complement(116896..117297)	PHAGE_Clostr_c_st_NC_007581: putative IS transposase (OrfA); PP_00179; phage(gi80159868)	2e-26
19	complement(117386..117598)	hypothetical; PP_00180	0.0
20	complement(117643..117894)	hypothetical; PP_00181	0.0
21	complement(117932..118120)	hypothetical; PP_00182	0.0
22	complement(118193..118354)	hypothetical; PP_00183	0.0
23	complement(118392..118664)	PHAGE_Staphy_GH15_NC_019448: hypothetical protein; PP_00184; phage(gi418488124)	2e-05
24	127541..127557	attR TTTTATATTTTATTTAA	0.0

Table S5

Enzyme- recognition sequence (digestion detected Y/N)	Number of sites	Fragment sizes expected
*Nde*I - CATATG (N)	85	Too numerous, evenly spread over genome
*Sph*I - GCATGC (N)	6	4, 2535 (if linear), 3754, 11669, 12321, 40638, 70377 (if linear)
*Bst*EII - GGTNACC (N)	4	3278, 13846, 33925 (if linear), 35927, 54322 (if linear)
*Bgl*II - AGATCT (Y)	8	1125, 1836, 2543 (linear), 3260, 8546, 15880, 18533 (if linear), 28753, 60822
*Dra*III - GACNNNGTG (N)	6	600 (if linear), 7300, 13782, 23155 (if linear), 27163, 34442, 34856
*Sma*I - CCCGGG (N)	1	55122 and 86176 (if linear)
*Eco*RI - GAATTC (Y)	27	Too numerous, evenly spread over the genome
*Eco*RV - GATATC (N)	12	498, 1099, 1573, 2887, 4125, 5147, 10180 (if linear), 12539, 14904, 16207, 20493, 21079, 30747 (if linear)
*Pvu*II - CAGCTG (N)	2	4782bp, 136516bp (circular) or 34452 (if linear,) 102064 (if linear)
*Hin*dIII - AAGCTT (N)	23	Too numerous, evenly spread over genome
*Rsa*I - GTAC (N)	228	Too numerous, evenly spread over the genome
*Alu*I - AGCT (Y)	345	Too numerous, evenly spread over the genome
*Hae*III - GGCC (N)	11	Eight small fragments < 200bp, 36918 (if linear), 104380 (if linear)

Chapter 5

Author contributions

Marla Trindade and Leonardo Joaquim van Zyl conceived the study and participated in its design and coordination. Leonardo Joaquim van Zyl performed all experiments and analysis. Leonardo Joaquim van Zyl wrote the bulk of the manuscript. All authors read and approved the final manuscript.

Appl Microbiol Biotechnol
DOI 10.1007/s00253-015-7109-9

Engineering resistance to phage GVE3 in *Geobacillus thermoglucosidasius*

Leonardo Joaquim van Zyl [1] & Mark Paul Taylor[2] & Marla Trindade [1]

Received: 18 September 2015 / Revised: 13 October 2015 / Accepted: 16 October 2015

Abstract *Geobacillus thermoglucosidasius* is a promising platform organism for the production of biofuels and other metabolites of interest. *G. thermoglucosidasius* fermentations could be subject to bacteriophage-related failure and financial loss. We develop two strains resistant to a recently described *G. thermoglucosidasius*-infecting phage GVE3. The phage-encoded immunity gene, *imm*, was overexpressed in the host leading to phage resistance. A phage-resistant mutant was isolated following expression of a putative anti-repressor-like protein and phage challenge. A point mutation was identified in the polysaccharide pyruvyl transferase, *csaB*. A double crossover knockout mutation of *csaB* confirmed its role in the phage resistance phenotype. These resistance mechanisms appear to prevent phage DNA injection and/or lysogenic conversion rather than just reducing efficiency of plating, as no phage DNA could be detected in resistant bacteria challenged with GVE3 and no plaques observed even at high phage titers. Not only do the strains developed here shed light on the biological relationship between the GVE3 phage and its host, they could be employed by those looking to make use of this organism for metabolite production, with reduced occurrence of GVE3-related failure.

Keywords Bacteriophage · *Geobacillus* · Resistance · Polysaccharide pyruvyl transferase · Immunity

* Leonardo Joaquim van Zyl
vanzyllj@gmail.com

Mark Paul Taylor
marktaylorimbm@gmail.com

Marla Trindade
prof.marlatt@gmail.com

[1] Institute for Microbial Biotechnology and Metagenomics (IMBM), University of the Western Cape, Robert Sobukwe Road, Bellville, Cape Town, South Africa

[2] TMO Renewables Limited, 40 Alan Turing Road, The Surrey Research Park, Guildford, Surrey GU2 7YF, UK

Introduction

Geobacillus thermoglucosidasius is a promising ʙplatform^ organism to use in the production of a range of useful metabolites with demonstrated ability to produce ethanol, isobutanol, and polylactic acid for biodegradable plastics (Cripps et al. 2009; Taylor et al. 2009; Lin et al. 2014; http://tinyurl.com/po6a52q). Extensive work has been done on engineering the organism for enhanced ethanol production, made possible through the development of a genetic system (Taylor et al. 2008). Bacteriophages, or phages, are viruses that specifically infect bacteria, and they are thought to be the most abundant biological entities on the planet with roughly ten virus particles to every bacterial cell and a total estimate of 1×10^{31} virus particles (Breitbart and Rohwer, 2005). It is well known that commercial bacterial fermentations are prone to bacteriophage-related failure with reports dating back to the 1920s in acetone/butanol (AB) fermentation processes (Jones et al. 2000) and subsequently well documented in dairy fermentations (Marco et al. 2012). These infections usually result in longer fermentation times with reduced yields or complete loss of the fermentation. This is costly due to the loss of product and feedstock, as well as the down time to disinfect the facility (Jones et al. 2000; Mahony et al. 2012).

There are several ways in which phage resistance can manifest itself in nature. Phages may be blocked from adsorbing to the host cell through mutation of gene-encoding phage receptors on the cell surface. They can be prevented from injecting their nucleic acid, the hosts may destroy the nucleic acid on injection, or resistance is gained through abortive infection

Published online: 04 November 2015

(ABI) mechanisms (Durmaz and Klaenhammer, 2007; Örmälä and Jalasvuori, 2013). Nucleic acid destruction can be mediated by restriction enzymes or may be targeted for degradation by clustered regularly interspaced short palindromic repeats (CRISPR) (Coffey and Ross, 2002; Bhaya et al. 2011). Over 20 ABI systems have been described and these affect nearly every aspect of the phage's ʙlife cyle^ (Samson et al. 2013). These range from genes to induce early lysis (abiZ or toxin/antitoxin-like systems, abiQ), thereby stopping the infected cell from producing more progeny, and those that interfere with phage replication (Abia, F, K and R) as well as RNA transcription (abiB and G) (Coffey and Ross, 2002). Irreversible binding of the phage to a cell is often mediated through a receptor protein. In Gram-negative bacteria, several membrane proteins have been described as receptors including OmpA (F, C, T, X), LamB, BtuB, TolC, and a range of flagella proteins (Chaturongakul and Ounjai, 2014). Only a few membrane protein receptor proteins have been described for Gram-positive bacteria including YueB in *Bacillus subtilis* and phage infection protein (PIP) in *Lactobacillus lactis* (São-José et al. 2004; Jakutytė et al. 2011). Mutation of these receptors has also been used in engineering of phage resistance (Dupont et al. 2004; Clément et al. 1983).

Many of these mechanisms have been seen to be selected for in dairy starter cultures and then used to rationally engineer resistance in commercially used strains. Although several of these strategies have been employed, some are preferred. In the case of AB fermentations, it was found that using a lysogenic strain, resistant due to superinfection immunity, produced less product per unit molasses used and had significantly longer doubling times than the wild-type organism, making it less desirable as production strain (Jones et al. 2000). Phage resistance systems are also not equally efficient, with some merely reducing the number of phage progeny produced compared with the wild type-system as measured by the efficiency of plating (EOP = phage titer on the resistant host containing the anti-phage system divided by the phage titer on the sensitive host), with industry naturally preferring higher potency systems (Moineau, 1999).

It is generally accepted that fermentations employing thermophilic bacteria such as *G. thermoglucosidasius* and *Bacillus coagulans* are less prone to contamination and even immune to phage-related failure or it may be that they are not often reported (Su and Xua, 2014). However, phages that infect thermophiles are known and are expected to eventually become a problem for industries making use of such organisms (Moineau, 1999). As a commercial platform, *G. thermoglucosidasius* would be expected to ferment a range of globally sourced feedstocks and could be exposed to phages from a variety of environments which may lead to failed or stuck fermentations and associated financial loss. The study of how phages infect and propagate inside the host cell should enable the development of phage-resistant strains and strain rotation strategies analogous to those developed for the dairy industry (Brűssow, 2001). We recently described a new *G. thermoglucosidasius*-infecting phage, GVE3, isolated from a pilot scale fermentation (van Zyl et al. 2015), and here, we describe the generation of strains resistant to the GVE3 phage by employing two strategies.

Materials and methods

Media, bacterial strains, plasmids, and phage used

Bacterial strains and plasmids used in this study are shown in Table 1. *Escherichia coli* strains were cultured in lysogeny broth (LB) broth, with 200 μg/ml ampicillin or 50 μg/ml kanamycin added as required. *G. thermoglucosidasius* strains were cultured in TGP medium. In general, *E. coli* DH5α was used for plasmid construction. One liter of TGP broth contains 17 g tryptone, 3 g soy peptone, 2.5 g K_2HPO_4, and 5 g NaCl. The pH was adjusted to 7.3 before autoclaving, after which 4 g Na-pyruvate and 4 ml glycerol were added in the form of filter-sterilized 10× concentrates. For solid media, 15 g/l agar was added before autoclaving. TGP was used during genetic manipulation and general maintenance of cultures. Cultures were incubated 60 °C with vigorous aeration. Phage GVE3 (GenBank accession no.: KP144388) is from the IMBM lab collection.

DNA manipulations and sequencing

Plasmid preparation, restriction endonuclease digestion, gel electrophoresis, and ligation were performed using standard methods or following the manufacturers' recommendations. Total DNA from all bacterial strains was prepared as described (Kotze et al. 2006). The QIAGEN Plasmid Midi Kit was used for large-scale plasmid preparations. Phage DNAwas prepared by first preparing a phage lysate from 1 l of culture as described below. The phage was pelleted by centrifugation, 13,000×*g* for 30 min, after addition of PEG8000 (7.5 ml of 20 % PEG8000 per 30 ml lysate) and incubation at 4 °C overnight. The pellet was resuspended in 1 ml SM buffer. The suspension was treated with DNaseI and RNaseA (Fermentas; final concentration of 0.1 μg/ml) at 37 °C for 1 h (DNaseI). The presence of free or background contaminating bacterial DNAwas checked by amplifying the 16 rDNA gene. The suspension was treated with ProteinaseK (Fermentas—final concentration 1 μg/ml) at 55 °C for 2 h. To this was added 70 μl 20 % (wt/vol) SDS and incubated at 37 °C for 1 h. An equal volume of phenol, chloroform, and isoamylalcohol (25:24:1) was added, the sample spun (15 ml Sterilin tube, Eppendorf 5810R centrifuge, 5000 RPM for 10 min) to separate the phases, and the top aqueous phase removed to a fresh tube. A second P:C:I extraction was performed. To this supernatant was added an equal volume of C:I (24:1) and spun again. The aqueous phase was

Table 1 Bacterial strains and plasmids used in this study

Strain, plasmid, or primer	Genotype or description	Source or reference
Strains		
G. thermoglucosidasius TM242	*ldh*A⁻*pfl*⁻P_*ldh*(NCA1503)/*pdh*^{up} variant of *G. thermoglucosidasius* NCIMB 11955	(Cripps et al. 2009)
E. coli JM109	F′ *traD*36 *proA* + *B*+ lacI^{q} Δ(*lacZ*)*M*15/Δ(*lac-proAB*) *glnV*44 e14− *gyrA*96 *recA*1 *relA*1 *endA*1 *thihsdR*17	
Plasmids		
pJET1.2	Amp^{r}; ColE1 replicon; cloning vector	Fermentas
pUCG18	Kan^{r} (in *G. thermoglucosidasius*), Amp^{r}; thermostable *Geobacillus* spp.—*E. coli* shuttle/expression vector	(Taylor et al. 2008)
pGR002	pUCG18 with P$_{ldh}$-*pheB* cloned into the MCS	(Bartosiak-Jentys et al. 2012)
pG18imm	pGR002 with *imm* replacing *pheB* between *Xba*I and *Mlu*I sites and P$_{idh}$ in place of P$_{ldh}$ between *Pst*I and *Xba*I sites	This study
pG18AR	pGR002 with the putative GVE3 anti-repressor cloned downstream of P$_{idh}$ using *Xba*I and *Mlu*I sites	This study
pTMO111	Amp^{r}; Kan^{r} (in *G. thermoglucosidasius*); ColE1 replicon, pUB110 IncA replicon, *E. coli* - *G. thermoglucosidasius* shuttle/suicide (>55 °C) vector containing a truncated *pflB* gene	(Cripps et al. 2009)
pImm111	pTMO111 with P$_{idh}$-*imm* cloned into *Not*I site	This study
pcsaB111	pTMO111 with a truncated and disrupted *csaB* cloned into the *Hin*dIII and *Eco*RI sites	This study
Primers		
#4F	5′-GAAATATTCCCTAATAATCC-3′	This study
#4R	5′-TAAACGATATGCACTATCTGCCG-3′	This study
#5F	5′-ATGGAGATAGAATTGACAAGC-3′	This study
#5R	5′-TTTGTTCATCAGTAACACGGGC-3′	This study
ImmF	5′-CGCGAGTCTAGAATGACGGTTTTTCTTGTG-3′	This study
ImmR	5′-GCGCGCACGCGTTTAAGCATTATTTTTAATTA-3′	This study
idh-immF	5′-TATATATGCGGCCGCCGATTTTTGCCGTAAGCCGC-3′	This study
idh-immR	5′-CGCGCGCGGCCGCTTAAGCATTATTTTTAATTA-3′	This study
AntF	5′-GCTAATCTAGAATGAACAAAAAGGAATTGGT-3′	This study
AntR	5′-ATCGAACGCGTTTATTTAACCGCATCTTTTAACG-3′	This study
csaBup	5′-ACGCTTGAGGAGCGAGTGCA-3′	This study
csaBdown	5′-CGCAGCGCTTTCTCGCTTCCT-3′	This study
csabF1	5′-GGCGGGAATTCTGTTTGTCAGCATCCATGTCA-3′	This study
csaBR1	5′-GACAGATTTAAATGGCACAGCGGTAACGGCTACTT-3′	This study
csaBF2	5′-CGGGCATTTAAATCCAGGATGTGCAAGTAACAAAG-3′	This study
csaBR2	5′-TACGTAAGCTTGCTGAAATATGCGGCGGTTAAC-3′	This study
pflup	5′-AAGGGCCTACAGAAGCAACG-3′	This study
pfldown	5′-GACAGAGCTTAGCGAAGCGAGC-3′	This study

removed to a fresh tube and a tenth volume 3 M sodium acetate (pH 5.2) and two volumes 100 % ethanol added. This was left at 4 °C to precipitate overnight. The sample was spun at 13,000 RPM for 10 min to pellet the DNA, and the pellet resuspended in 40 μl of TE buffer. The phage DNA was electrophoresed on a 1 % low melting point agarose gel, excised, and purified from the gel using standard agarase (Fermentas) treatment. The pellet was resuspended in 40 μl TE buffer. The quality and integrity of the DNA was checked using a Bioanalyzer prior to library preparation. Sanger DNA sequencing was performed using an ABI Prism 377 automated DNA sequencer (University of Stellenbosch central analytical facility) while next-generation sequencing was performed using an Illumina Miseq. Sequences were analyzed with DNAMAN (version 4.1, LynnonBioSoft) and CLC Genomics Workbench version 6.5 (CLC Bio).

Construction of plasmids for the disruption of the *csaB* and insertion of the *imm* gene and phage resistance testing

In general, *E. coli* JM109 was used for plasmid construction. Single and double crossover knockout and knock-fsin mutants

were constructed as previously described with modifications (Cripps et al. 2009).When transformed with pTMO111-based constructs, *G. thermoglucosidasius* transformants are incubated at 52 °C after transformation to first establish transformants and subsequently cultured at 60 °C (see below). To generate the *csaB* knockout, two gene fragments from each end of the *csaB* gene were ligated to replace a central 180 bp region with a *Swa*I site using the csaBF1, csaBF2, csaBR1, and csaBR2 primer set. For generation of single crossover (SCOs) mutants, following confirmation of transformation of a TM242 culture with the construct of interest, it is transferred to 50 ml TGP broth without kanamycin and cultured at 60 °C for 12 h. One milliliter of this culture is then transferred to fresh 50 ml TGP broth without selection and again cultured for 12 h. The strain is serially cultured as described without selection for 2 weeks. A serial dilution of the culture is then plated on non-selective media and 100 colonies picked and patched to selective plates (TGP-Kan). If the majority of colonies (90–100 %) are kanamycin resistant, the serial culture is continued without selection. If the majority of the colonies are kanamycin sensitive, 1 ml of the culture is transferred to fresh 50 ml TGP broth with kanamycin selection and cultured for 12 h. A serial dilution of this culture is plated on selective plates and 100 colonies picked for crossover analysis using polymerase chain reaction (PCR). For generation of double crossover (DCO) mutants, the SCO would be serially cultured, until again the majority of the culture is kanamycin sensitive, and in the case of the *csaB* knockout, a phage challenge was used to eliminate WT revertants when selecting the DCO from a pool of SCOs. For generating the *imm* gene knock-in, the fusion of P_{idh} and *imm* gene from the pG18imm was amplified with primers idh-immF and idh-immR containing *Not*I restriction endonuclease sites, and the product cloned directly into pTMO111 after digestion with *Not*I of both vector and insert.

To test for phage resistance, the test and control strains were cultured on fresh TGP agar plates overnight. The cultures would be streaked to a fresh TGP plate the next morning and incubated for 2 h at 60 °C. Fifty milliliters TGP medium in a 250-ml Erlenmeyer flask would be inoculated from the 2-h-old culture and incubated at 60 °C with vigorous aeration. Once the culture reached OD600nm of 1.0, 1 ml of the culture would be used to inoculate three 50-ml TGP cultures in a 250-ml Erlenmeyer flask and the growth monitored until the OD600nm reached 0.4–0.5 and the phage was added.

Polymerase chain reaction

Polymerase chain reaction (PCR) was performed using Phusion DNA polymerase (New England Biolabs™). Generally, 50 ng DNA were used in a 50-μl reaction volume containing 2 mM $MgCl_2$, 0.125 μM of each primer, 0.2 mM of each deoxynucleoside triphosphate, and 1 U DNA polymerase.

Reactions were carried out in a BioRad T100 thermocycler, with an initial denaturation at 94 °C for 60 s, followed by 30 cycles of denaturation (30 s at 94 °C), annealing (30 s), and variable elongation (72 °C), where annealing temperatures and elongation times were adjusted as required.

Fermentative product profile quantification

G. thermoglucosidasius strain-resistant to GVE3 were cultured at 60 °C for 16 h and 200 rpm TGP medium. A volume of 0.5 ml of this culture was transferred to 10 ml of USMYE media in 15-ml screw-cap universal tubes (Cripps et al. 2009). These cultures were grown for 16 h at 60 °C and 200 rpm. Cells were removed from the fermentation by centrifugation (2057×*g* for 10 min) and the supernatant was used for determination of metabolite concentrations by high-performance liquid chromatography (HPLC). Reactions were analyzed on a Rezex RHM-Monosaccharide column (Phenomenex), using 5 mM H_2SO_4 as mobile phase under isocratic elution (0.6 ml/min, 48 °C) on a Dionex UltiMate 3000. Samples (20 μl) were injected by autosampler and the components detected using refractive index detector and UV/Vis photodiode array. Products were compared to suitable standards of known concentration and against the media in which the cultures were grown.

Results

Phage resistance through expression of the GVE3 immunity gene

Characterization of the phage GVE3 genome sequence (van Zyl et al. 2015) identified a possible phage immunity gene (ORF 41; *imm*), showing 36 % amino acid identity to the immunity protein (*yomJ*; d gene; AAC13006.1) from *B. subtilis* phage SPbeta. Expression of immunity proteins has proved effective in engineering resistance against closely related phage (McLaughlin et al. 1986; McGrath et al. 2002). ORF 41 was overexpressed under control of the constitutive *idh* promoter in the TM242 strain and tested for phage sensitivity. When compared with the infected control, the strains expressing the proposed immunity protein did not suffer culture collapse. Plaque assays were performed to determine if expression of the immunity protein abolishes infection or whether it reduces the efficiency of plating. No plaques were observed for strains expressing the immunity protein even at the highest phage titer (1×10^8 pfu/ml). The immunity gene, again expressed from the *idh* promoter, was integrated on the *G. thermoglucosidasius* genome by homologous recombination using the pyruvate formate lyase gene (*pfl*) as the integration site. The double crossover (DCO) mutant integration was confirmed by PCR amplification of the region where

integration was targeted to (*pfl*), using primers *pflup* and *pfldown* which target to the genome outside the area used to construct pTMO111. TM242-*pfl*::*imm* DCO mutants were expected to give a 2.3-kb product as opposed to 1.4 kb for TM242, as well as sensitivity to kanamycin and the ability to amplify the *imm* gene from genomic DNA (Fig. 1). The knock-in strain (TM242-*pfl*::*imm*) was assayed for phage resistance and found to be resistant (Fig. 2). TM242-*pfl*::*imm* was assayed to see if it is still capable of producing ethanol to the same level as the parent strain (TM242) in 10/15 model fermentations. The knock-in strain produced 0.497 ± 0.011 (n = 5) g of ethanol produced per gram glucose consumed while TM242 produced 0.474 g/g ± 0.007 (n = 5) showing that the strain still performs well in this regard.

Identification of a phage-resistant mutant

Initial attempts to identify naturally phage-resistant mutants were hampered by the background of lysogens produced during phage infection which masked the presence of resistant individuals. It was reasoned that overexpression of an anti-repressor-like protein, similar to that described for other phages (Shearwin et al. 1998; Mardanov and Ravin 2007; Fogg et al. 2010), would force induction of the lytic cycle postinfection by not allowing repression of the genes responsible for lytic conversion. A possible anti-repressor-like protein was identified in the GVE3 genome (ORF 184; van Zyl et al. 2015), which showed 48 % amino acid identity over 115aa to a putative anti-repressor protein from *Peptoclostridium difficile* (WP_021424465) (Iyer et al. 2002).

The TM242 strain expressing the putative anti-repressor (TM242-pG18AR) was challenged with GVE3 at a multiplicity of infection of 100. Following culture crash, a colony count was performed and the total number of bacteria calculated at ±4000 cfu. Sixteen of these colonies were streaked several times to get rid of any free GVE3 and checked by PCR for the presence of the GVE3 genome as a lysogen. GVE3 was detected in eight of these colonies suggesting that half of these isolates represented natural phage-resistant mutants and that the anti-repressor worked as envisioned reducing the background of lysogens. To confirm phage resistance, one of the isolates was subjected to phage challenge, which demonstrated that the strain was resistant, with no lysis observed post infection. The genome sequence of the resistant isolate TM242-*csaB** was determined using Roche 454 sequencing. The average coverage of the TM242-*csaB** genome was low (8 fold); thus, a literature survey was performed to identify potential target genes for investigation including sortase A/B (Davison et al. 2005), OmpA, LamB, phage infection protein (PIP; *yueB*), and polysaccharide pyruvyl transferase—*csaB* (Bishop-Lilly et al. 2012). None of these targets except *csaB* appeared to be altered in the TM242-*csaB** genome when compared to the reference TM242 genome sequence. Manual inspection of the sequence indicated a cytosine to thymine transition which resulted in an amber mutation (TGG to TAG) in the *csaB* reading frame. The gene was PCR amplified

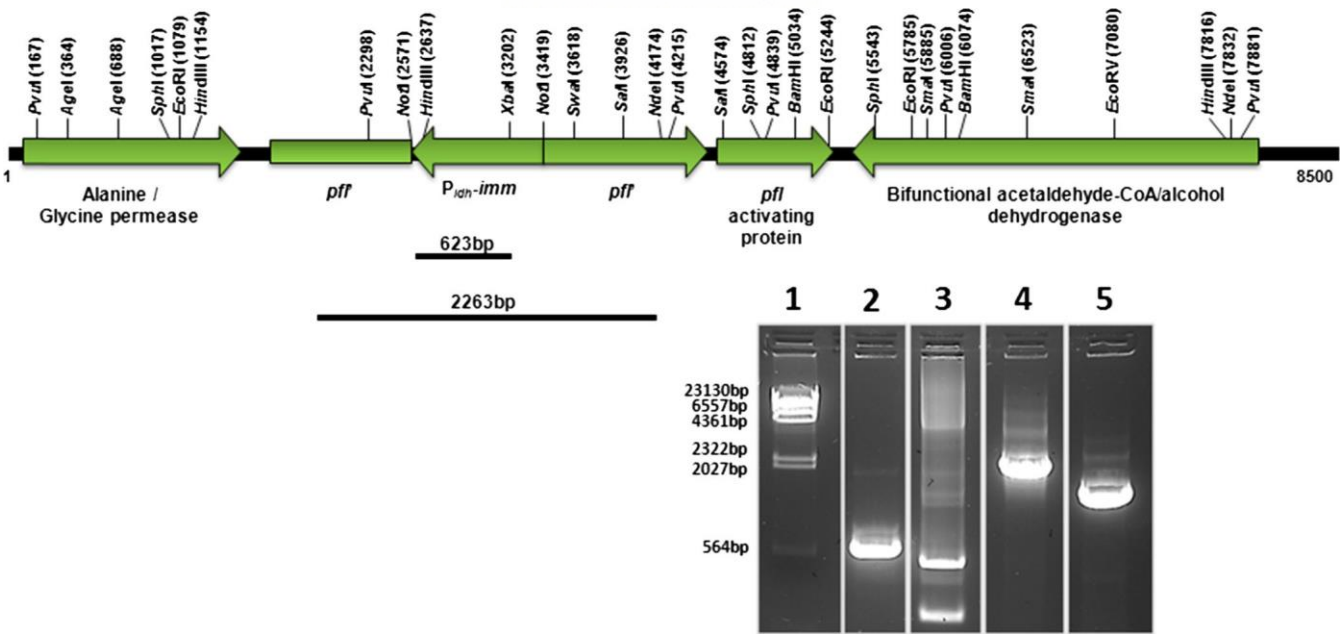


Fig. 1 Confirmation of *imm* insertion DCO mutant. Lane *1*—Molecular weight marker (phage phage Lambda DNA digested with *Hin*dIII), lane *2*—623-bp PCR product using ImmF and ImmR primers and genomic DNA extracted from DCO mutant as template, lane *3*—non-specific amplification products when using ImmF and ImmR and genomic DNA from TM242 as template, lane *4*—2263-bp PCR product when using pflup and pfldown primer set on genomic DNA from DCO mutant, and lane *5*—1415-bp PCR product using pflup and pfldown primer and TM242 genomic DNA as template

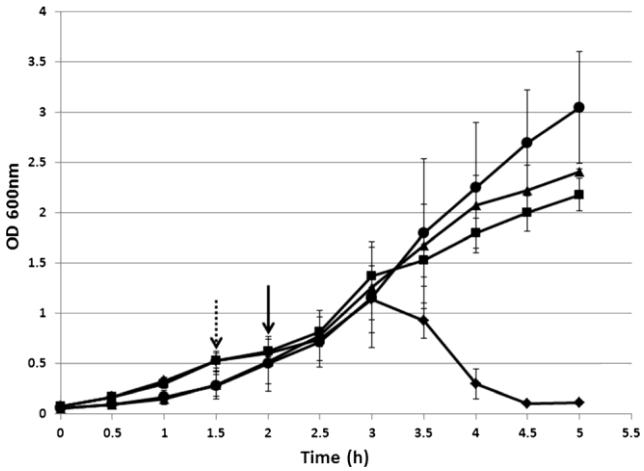


Fig. 2 Comparison of growth curves of GVE3-infected (*black diamonds*) and uninfected (*black squares*), TM242 cultures with infected (*black triangles*) and uninfected (*black circles*) TM242-P*idh-imm*. *Dashed arrow* indicates the time at which GVE3 was added for TM242 cultures and the *solid arrow* for TM242-P*idh-imm* cultures

from the resistant TM242-*csaB** isolate and the mutation confirmed by Sanger sequencing (Fig. 3). To complement the mutation, the WT *csaB* was overexpressed in the resistant TM242-*csaB** strain and assayed for phage resistance. Overexpression of *csaB* appeared to be toxic to cell growth; however, culture crash was observed on several occasions (data not shown). A *csaB* knockout in TM242 was generated to give TM242-Δ*csaB*. Primers (csaBup and csaBdown) designed to anneal to the chromosomal DNA directly up- and downstream of the DNA fragment used to generate the knockout construct for *csaB* were used to amplify this region from the chromosome to test whether it contained the engineered *Swa*I site. *Swa*I digests of these PCR amplicons resulted in the two products of expected size (Fig. 4), and this DCO mutant was assayed for phage resistance (Fig. 5). Compared with the control strains, no culture lysis was observed in Δ*csaB* strains when challenged with GVE3 phage. Plaque assays were performed to determine if the mutation results in reduced efficiency of plating. No plaques could be observed for the Δ*csaB* mutant strain even at the highest phage titer (1×10^8 pfu/ml). A PCR check to confirm that phage resistance was not due to superinfection immunity (lysogeny) was performed using five primer sets targeting GVE3, including a set targeted to the immunity protein, and no amplification was observed for any of the primer sets (data not shown). Phage pull-down assays using the TM242-Δ*csaB* mutant showed that it binds 19 % (±19 %) of the phage while TM242 pulls down 89 % (±6.5 %), which suggests that there may still be reversible binding of phage particles to TM242-Δ*csaB*. The TM242-Δ*csaB* strain was tested for its ability to produce ethanol compared to the TM242 strain in 10/15 model fermentations at 60 °C in USM medium. The TM242-Δ*csaB* strain produced 0.46 g/g on average ± 0.02 g/g ($n = 7$) which is comparable to that of TM242 0.45 g/g ± 0.02 ($n = 9$).

Overexpression of ʙantiholin^, putative regulator, and *yueB* knockout

Chang and co-workers demonstrated that overexpression of the antiholin (S107) of phage phage Lambda abolished lysis in *E. coli* (Chang et al. 1995; Raab et al. 1988). ORF52, a holin-like gene with a potential dual start motif (M-T-K-M), including two putative ribosomal binding sites located upstream of each start, was identified during GVE3 genome characterization and the full length gene product (82aa) was overexpressed in *G. thermoglucosidasius* using the *idh* promoter. Phage resistance testing showed that expression of this ORF did not give resistance to phage infection.

A second putative regulator (ORF174; HTH-motif containing) related to hypothetical proteins in *Anoxybacillus gonensis* (79 % identity over 322 amino acids of 315) and weak identity

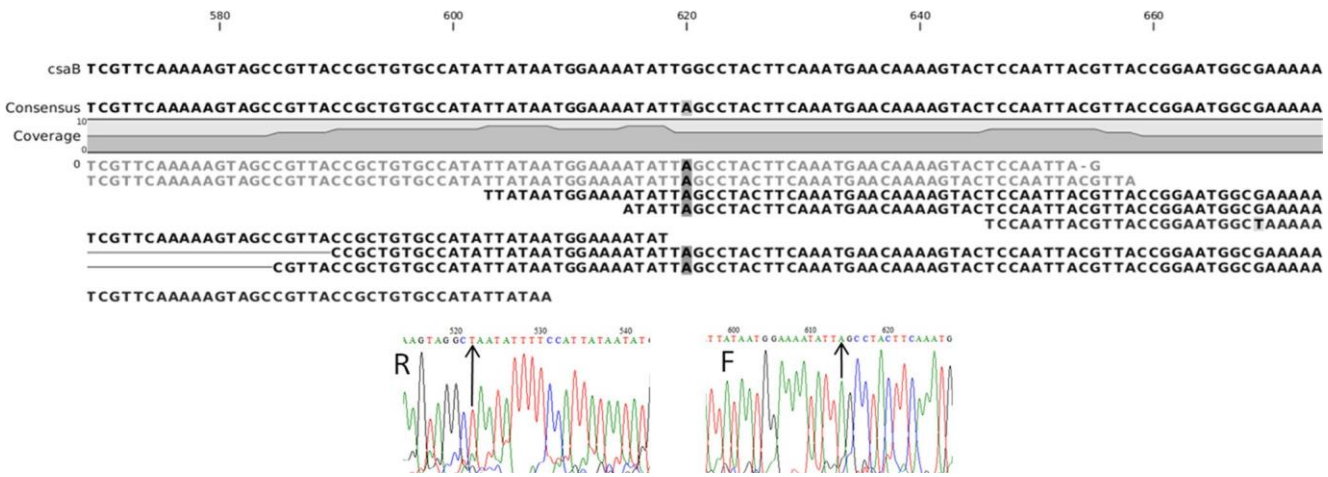


Fig. 3 Mapping of reads generated from TM242-*csaB** genome sequencing to WT *csaB*, and Sanger sequence (*R*—reverse; *F*—forward) of the same region confirming the mutation

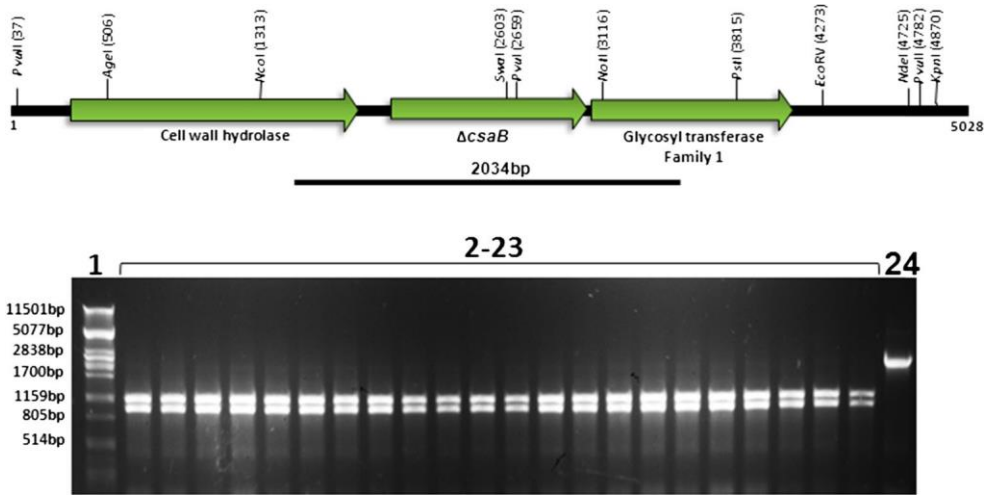


Fig. 4 Confirmation of *csaB* inactivation double crossover mutants. Lane *1*—Molecular weight marker (phage phage Lambda DNA digested with *Pst*I), lane *2–23*, 2034-bp PCR product from possible DCO clones generated using csaBup and csaBdown primers digested with *Swa*I giving the 1108- and 926-bp products demonstrating inactivation of *csaB* through deletion of 127 bp and introduction of a frame shift, lane *24*—undigested 2034-bp PCR product using csaBup and csaBdown primers from TM242 strain digested with *Swa*I

(28 % identity over 104 amino acids of 1956) to tail fiber proteins from *Bacillus* phages Stills and Stahl was identified. An attempt was made to overexpress it in TM242; unfortunately, the construct could not be transformed into TM242, suggesting that overexpression of this ORF may be lethal to the host.

Disruption of YueB, a membrane protein that is part of an Esat-6 or type VII secretion system and related to phage infection protein (PIP) of *L. lactis*, has also been shown to result in a phage resistance phenotype in *B. subtilis* against phage SPP1 (Baptista et al. 2013; São-José et al. 2004). A *yueB* homolog (Geoth_0462; 51 % identity over 700 amino acids of 1056 of *yueB* from *Bacillus atrophaeus*), which has the same predicted transmembrane regions as PIP and YueB, was identified in *G. thermoglucosidasius* and a single crossover knockout generated. This mutation also did not lead to a phage resistant phenotype.

Discussion

Here, we described the generation of two GVE3-resistant strains of *G. thermoglucosidasius* through the overexpression of the phage immunity gene (*imm*) and knocking out of host-encoded polysaccharide pyruvyl transferase. Both of these have previously been shown to result in phage resistance in *B. subtilis* and *Bacillus anthracis*, respectively (McLaughlin et al. 1986; Bishop-Lilly et al. 2012). Both mechanisms completely abolish phage-induced lysis of cells as opposed to merely reducing the efficiency of plating. The absence of phage DNA in the host post infection, in both engineered strains, suggests that these mechanisms either prevent phage DNA entry or lysogenic conversion. The TM242-*pfl*::*imm* integrant showed resistance, indicating that it is not necessary for *imm* to be expressed from a multicopy plasmid for effective resistance, but that enough is produced from the *idh* promoter to give resistance. The *Imm* protein is predicted to have one transmembrane region, suggesting that this may be the site where it acts to prevent phage infection.

To our knowledge, this is only the second report of the involvement of *csaB* in phage attachment to the cell, and the first for the thermophilic host *G. thermoglucosidasius*. The role of CsaB is to attach pyruvyl moieties to peptidoglycan-associated polysaccharides which, in turn, are used as anchor points for S-layer homology (SLH) domain containing proteins for display on the cell surface (Mesnage et al. 2000). Thus, it is possible that one of the SLH proteins could be the target for phage attachment to the cell and that inactivation of

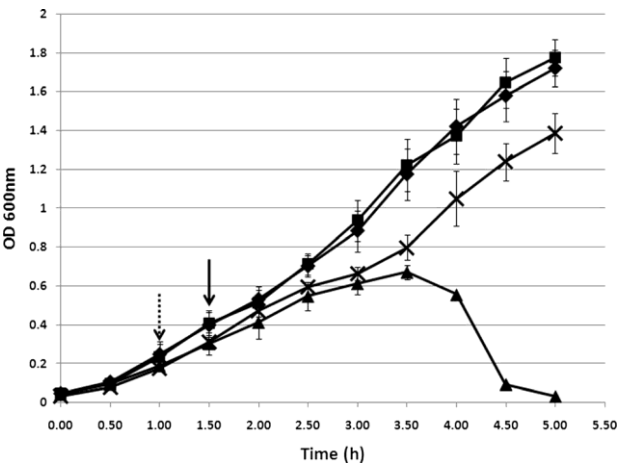


Fig. 5 Comparison of growth curves of TM242-Δ*csaB* infected (*black diamonds*) and uninfected (*black squares*) with TM242-infected (*black triangles*) and uninfected (*crosses*)

csaB no longer allows the protein to be displayed on the surface, therefore rendering the cell immune to the phage. A search for SLH domain-containing proteins on the *G. thermoglucosidasius* genome indicated that there are nine proteins containing this domain, and *csaB* is located at one end of a 75-kb region (±300,000–375,000 bp on NC_015660) of the *G. thermoglucosidasius* genome containing eight of these proteins. Sequential knockout of these should answer the question as to whether or not one of them is the target for phage attachment. This study also suggests that Firmicute-infecting phages may target SLH domain-containing proteins in general, for attachment. It also implies that this GVE3 targets similar proteins for infection as does its mesophilic counterpart AP50c. Faster lysis together with smaller numbers of viable cells, recovered from cultures over-expressing the anti-repressor-like protein following phage challenge, suggests that the product of this ORF plays a similar role to the anti-repressor of *E. coli* phage lambda (Nijkamp et al. 1971; Reichardt 1975; Shearwin et al. 1998). However, a detailed study of the gene regulation enabling the switch between lysis and lysogeny in GVE3 has yet to be carried out. Both phage-resistant strains produced ethanol at levels comparable to the parent strains; thus, the inactivation of *csaB* and introduction of *imm* do not negatively affect ethanol production.

The demonstration of effective engineering of phage resistance in a thermophile using techniques applied to mesophilic organisms bodes well for the use of thermophiles in industrial fermentations. Together with potential improvements in ethanol yields (Cripps et al. 2009; van Zyl et al. 2014), resistance against phage infection should make *G. thermoglucosidasius* a more productive and robust platform for not only biofuel production but also other metabolites of interest.

Acknowledgments The authors wish to thank TMO Renewables for the gift of the GVE3 phage. This work was funded by the National Research Foundation (NRF) of South Africa.

Compliance with ethical standards

Conflict of interest The authors declare that they have no competing interest.

References

Baptista C, Barreto HC, São-José C (2013) High levels of Deg U-P activate an Esat-6-like secretion system in *Bacillus subtilis*. PLOS One. doi:10.1371/journal.pone.0067840

Bartosiak-Jentys J, Eley K, Leak DJ (2012) Application of *pheB* as a reporter gene for *Geobacillus* spp., enabling qualitative colony screening and quantitative analysis of promoter strength. Appl Environ Microbiol 78:5945–5947

Bhaya D, Davison M, Barrangou R (2011) CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. Annu Rev Genet 45:273–297

Bishop-Lilly KA, Plaut RD, Chen PE, Akmal A, Willner KM, Butani A, Dorsey S, Mokashi V, Mateczun AJ, Chapman C, George M, Luu T, Read TD, Calendar R, Stibitz S, Sozhamannan S (2012) Whole genome sequencing of phage-resistant *Bacillus anthracis* mutants reveals an essential role for cell surface anchoring protein CsaB in phage AP50c adsorption. Virol J 9:246

Breitbart M, Rohwer F (2005) Here a virus, there a virus, everywhere the same virus? Trends Microbiol 13:278–284

Brűssow H (2001) Phages of dairy bacteria. Annu Rev Microbiol 55: 283–303

Chang C, Nam K, Young R (1995) S gene expression and the timing of lysis by bacteriophage phage Lambda. J Bacteriol 177:3283–3294

Chaturongakul S, Ounjai P (2014) Phage–host interplay: examples from tailed phages and Gram-negative bacterial pathogens. Front Microbiol 5:442

Clément JM, Lepouce E, Marchal C, Hofnung M (1983) Genetic study of a membrane protein: DNA sequence alterations due to 17 *lamB* point mutations affecting adsorption of phage lambda. EMBO J 2: 77–80

Coffey A, Ross RP (2002) Bacteriophage-resistance systems in dairy starter strains: molecular analysis to application. Antonie Van Leeuwenhoek 82:303–321

Cripps RE, Eley K, Leak DJ, Rudd B, Taylor M, Todd M, Boakes S, Martin S, Atkinson T (2009) Metabolic engineering of *Geobacillus thermoglucosidasius* for high yield ethanol production. Metab Eng 11:398–408

Davison S, Couture-Tosi E, Candela T, Mock M, Fouet A (2005) Identification of the *Bacillus anthracis* Ƴ phage receptor. J Bacteriol 187:6742–6749

Dupont K, Janzen T, Vogensen FK, Josephsen J, Stuer-Lauridsen B (2004) Identification of *Lactococcus lactis* genes required for bacteriophage adsorption. Appl Environ Microbiol 70:5825–5832

Durmaz E, Klaenhammer TR (2007) Abortive phage resistance mechanism AbiZ speeds the lysis clock to cause premature lysis of phage-infected *Lactococcus lactis*. J Bacteriol 189:1417–1425

Fogg PCM, Rigden DJ, Saunders JR, McCarthy AJ, Allison HE (2010) Characterization of the relationship between integrase, excisionase and antirepressor activities associated with a super infecting Shiga toxin encoding bacteriophage. Nuc Acids Res 39:2116–2129

Iyer LM, Koonin EV, Aravind L (2002) Extensive domain shuffling in transcription regulators of DNA viruses and implications for the origin of fungal APSES transcription factors. Gen Biol 3:research0012.1–0012.11

Jakutytė L, Baptista B, São-José C, Daugelavičius R, Carballido-López R, Tavares P (2011) Bacteriophage infection in rod-shaped Gram-positive bacteria: evidence for a preferential polar route for phage SPP1 entry in *Bacillus subtilis*. J Bacteriol 193:4893–4903

Jones DT, Shirley M, Wu X, Keis S (2000) Bacteriophage infections in the industrial acetone butanol (AB) fermentation process. J Mol Microbiol Biotechnol 2:21–26

Kotze AA, Tuffin IM, Deane SM, Rawlings DE (2006) Cloning and characterization of the chromosomal arsenic resistance genes from *Acidithiobacillus caldus* and enhanced arsenic resistance on conjugal transfer of *ars* genes located on transposon TnAtcArs. Microbiology 152:3551–3560

Lin PP, Rabe KS, Takasumi JL, Kadisch M, Arnold FH, Liao JC (2014) Isobutanol production at elevated temperatures in thermophilic *Geobacillus thermoglucosidasius*. Metab Eng 24:1–8

Mahony J, Murphy J, van Sinderen D (2012) Lactococcal 936-type phages and dairy fermentation problems: from detection to evolution and prevention. Front Microbiol 3:335

Marco MB, Moineau S, Quiberoni A (2012) Bacteriophages and dairy fermentations. Bacteriophage 2:149–158

Mardanov AV, Ravin NV (2007) The antirepressor needed for induction of linear plasmid-prophage N15 belongs to the SOS regulon. J Bacteriol 189:6333–6338

McGrath S, Fitzgerald GF, van Sinderen D (2002) Identification and characterization of phage-resistance genes in temperate lactococcal bacteriophages. Mol Microbiol 43:509–520

McLaughlin JR, Wong HC, Ting YE, Van Arsdell JN, Chang S (1986) Control of lysogeny and immunity of *Bacillus subtilis* temperate bacteriophage SPβ by its d gene. J Bacteriol 167:952–959

Mesnage S, Fontaine T, Mignot T, Delepierre M, Mock M, Fouet A (2000) Bacterial SLH domain proteins are non-covalently anchored to the cell surface via a conserved mechanism involving wall polysaccharide pyruvylation. EMBO J 19:4473–4484

Moineau S (1999) Applications of phage resistance in lactic acid bacteria. Antonie Van Leeuwenhoek 76:377–382

Nijkamp HJJ, Szybalski W, Calef E (1971) Antirepressor controls the transcription of the repressor operon of lambda prophage (L. G. H. Ledoux, Ed.), Informative molecules in biological systems, p. 241–248. Amsterdam: North-Holland Publishing Co

Örmälä A-M, Jalasvuori M (2013) Phage therapy: should bacterial resistance to phages be a concern, even in the long run? Bacteriophage 3(1):e24219

Raab R, Neal G, Sohaskey C, Smith J, Young R (1988) Dominance in lambda S mutations and evidence for translational control. J Mol Biol 199:95–105

Reichardt LF (1975) Control of bacteriophage lambda repressor synthesis: regulation of the maintenance pathway by the *cro* and *cI* products. J Molec Biol 93:289–305

Samson JE, Magadán AH, Sabri M, Moineau S (2013) Revenge of the phages: defeating bacterial defences. Nat Rev Microbiol 11:675–687

São-José C, Baptista C, Santos MA (2004) *Bacillus subtilis* operon encoding a membrane receptor for bacteriophage SPP1. J Bacteriol 186:8337–8346

Shearwin KE, Brumby AM, Egan JB (1998) The Tum protein of coliphage 186 is an antirepressor. J Bacteriol 273:5708–5717

Su F, Xua P (2014) Genomic analysis of thermophilic *Bacillus coagulans* strains: efficient producers for platform bio-chemicals. Sci Rep 4:3926

Taylor MP, Esteban CD, Leak DJ (2008) Development of a versatile shuttle vector for gene expression in *Geobacillus* spp. Plasmid 60: 45–52

Taylor MP, Eley KL, Martin S, Tuffin MI, Burton SG, Cowan DA (2009) Thermophilic ethanologenesis: future prospects for second-generation bioethanol production. Trends Biotechnol 27:398–405

Van Zyl LJ, Taylor MP, Eley K, Tuffin M, Cowan DA (2014) Engineering pyruvate decarboxylase-mediated ethanol production in the thermophilic host *Geobacillus thermoglucosidasius*. Appl Microbiol Biotechnol 98:1247–1259

Van Zyl LJ, Sunda F, Taylor MP, Cowan DA, Trindade MI (2015) Identification and characterization of a novel *Geobacillus thermoglucosidasius* bacteriophage, GVE3. Arch. Virol: 2269-2282

The final publication is available at Springer via http://dx.doi.org/10.1007/s00253-015-7109-9

Permission to reproduce the article here:

Excerpt From Springer Copyright Transfer Contract:

"Author retains the right to use his/her article for his/her further scientific career by including the final published journal article in other publications such as dissertations and postdoctoral qualifications provided acknowledgement is given to the original source of publication."

License no: 4171351184781

Correspondence with the journal:

Dear Mr. van Zyl,

The PDF for your manuscript, "Engineering resistance to phage GVE3 in Geobacillus thermoglucosidasius" is ready for viewing.

In order to formally submit your manuscript to the journal, you must approve the PDF.

Please access the following web site:

http://amab.edmgr.com/

Your username is: lonnievanzyl

Your password is:

Click "Author Login".

In your main menu, you will see there is a category entitled "Submission Waiting for Author's Approval".

Click on that category, view your submission and approve it. In the unlikely case of conversion issues you may

submit your manuscript data as a PDF file.

Your manuscript will then be formally submitted to the journal.

Thank you very much.

With kind regards,

Springer Journals Editorial Office

Applied Microbiology and Biotechnology

Dear Mr. van Zyl,

Dear Co-Author(s),

Your submission entitled "Engineering resistance to phage GVE3 in Geobacillus thermoglucosidasius" has been received by

Applied Microbiology and Biotechnology

You will be able to check on the progress of your paper by logging on to Editorial Manager as an author. The URL is http://amab.edmgr.com/. (This applies to the corresponding author only.)

Your manuscript will be given a reference number once an Editor has been assigned.

Thank you for submitting your work to our journal.

Kind regards,

Editorial Office

Applied Microbiology and Biotechnology

PS: If there would be any concern regarding authorship, please contact the Managing Editor (Dr. Dorothea Kessler) at AMBoffice@gmx.de

Now that your article will undergo the editorial and peer review process, it is the right time to think about publishing your article as open access. With open access your article will become freely available to anyone worldwide and you will easily comply with open access mandates. Springer's open access offering for this journal is called Open Choice (find more information on www.springer.com/openchoice). Once your article is accepted, you will be offered the option to publish through open access. So you might want to talk to your institution and funder now to see how payment could be organized; for an overview of available open access funding please go to www.springer.com/oafunding.

Although for now you don't have to do anything, we would like to let you know about your upcoming options.

Dear Mr. van Zyl,

Your submission entitled "Engineering resistance to phage GVE3 in Geobacillus thermoglucosidasius" has been assigned the following manuscript number: AMAB-D-15-02215.

You will be able to check on the progress of your paper by logging on to Editorial Manager as an author.

The URL is http://amab.edmgr.com/.

Thank you for submitting your work to this journal.

Kind regards,

Editorial Office

Applied Microbiology and Biotechnology

Dear Mr. van Zyl,

The PDF for your manuscript, "Engineering resistance to phage GVE3 in Geobacillus thermoglucosidasius" is ready for viewing.

In order to formally submit your manuscript to the journal, you must approve the PDF.

Please access the following web site:

http://amab.edmgr.com/

Your username is: lonnievanzyl

Your password is: available at this link

Click "Author Login".

In your main menu, you will see there is a category entitled "Submission Waiting for Author's Approval".

Click on that category, view your submission and approve it. In the unlikely case of conversion issues please

contact the Journal's Editorial Office by clicking the "CONTACT US" link on the journal EM home page.

Your manuscript will then be formally submitted to the journal.

Thank you very much.

With kind regards,

Springer Journals Editorial Office

Applied Microbiology and Biotechnology

Ref.: Ms. No. AMAB-D-15-02215R1

Engineering resistance to phage GVE3 in Geobacillus thermoglucosidasius

Dear Mr. van Zyl,

Applied Microbiology and Biotechnology has received your revised submission.

You may check the status of your manuscript by logging onto Editorial Manager at http://amab.edmgr.com/.

Kind regards,

Editorial Office

Applied Microbiology and Biotechnology

Dear Mr. van Zyl,

The manuscript cannot start the review process until the following corrections are made to meet the journal's requirements (see Instructions for authors, as well):

1. Reference list entries must not be numbered
2. Table is labeled and cited as 'Table I' -> must be changed to Table 1
3. The GenBank accession number for the GVE3 genome sequence should be mentioned in the Materials & Method section

Please edit your submission and make the necessary changes by logging into the Editorial Manager at:

http://amab.edmgr.com/

and clicking on "Submissions Sent back to Author".

You must then click on "Edit Submission", make the necessary changes, upload your revised manuscript, remove your old manuscript, and approve your submission.

If you have any questions, please do not hesitate to contact me.

Kind regards,

Ethel Dionela

JEO Assistant

Applied Microbiology and Biotechnology

Dear Mr. van Zyl,

The PDF for your manuscript, "Engineering resistance to phage GVE3 in Geobacillus thermoglucosidasius" is ready for viewing.

In order to formally submit your manuscript to the journal, you must approve the PDF.

Please access the following web site:

http://amab.edmgr.com/

Your username is: lonnievanzyl

Your password is: available at this link

Click "Author Login".

In your main menu, you will see there is a category entitled "Submission Waiting for Author's Approval".

Click on that category, view your submission and approve it. In the unlikely case of conversion issues please

contact the Journal's Editorial Office by clicking the "CONTACT US" link on the journal EM home page.

Your manuscript will then be formally submitted to the journal.

Thank you very much.

With kind regards,

Springer Journals Editorial Office

Applied Microbiology and Biotechnology

Dear Mr. van Zyl,

Re: Your manuscript entitled "Engineering resistance to phage GVE3 in Geobacillus thermoglucosidasius"

NOTE: This letter applies to the corresponding author. All co-authors are cc:'d on this email for information purposes only.

Thank you for approving the changes that the Editor made to your submission or updating your submission according to the requested changes.

You will be able to check on the progress of your paper by logging on to Editorial Manager as an author. The URL is http://amab.edmgr.com/.

Thank you for submitting your work to this journal.

Kind regards,

Editorial Office

Applied Microbiology and Biotechnology

Ref.: Ms. No. AMAB-D-15-02215

Engineering resistance to phage GVE3 in Geobacillus thermoglucosidasius

Applied Microbiology and Biotechnology

Dear Mr. van Zyl,

Your manuscript AMAB-D-15-02215 entitled Engineering resistance to phage GVE3 in Geobacillus thermoglucosidasius which you submitted to "Applied Microbiology and Biotechnology", has been reviewed.

The reviewers' comments can be found at the end of this email or can be accessed by following the provided link.

The reviewer(s) have recommended publication, but also suggest major revisions to your manuscript. Therefore, I invite you to respond to the reviewer(s)' comments and revise your manuscript.

This is your login information:

Your username is: lonnievanzyl

Your password is: available at this link

When revising your work, please submit a list of changes or a rebuttal against each point which is being raised when you submit the revised manuscript.

Please upload the revised version within the next six weeks.

To submit a revision, go to http://amab.edmgr.com/ and log in as an Author. You will see a menu item called 'Submissions Needing Revision'. You will find your submission record there.

Beside the comments of the reviewers, please make sure to address the points listed in the following checklist before you submit your revised version (not all items of this checklist apply to Mini-Reviews):

- Have all in the study described strains been deposited in a public strain collection? Has the collection number been mentioned in the manuscript?

- Does the manuscript meet the guidelines for authors of AMB?

- Have "Results" and "Discussion" sections been separated? - (For Original Paper articles only)

- Are the conclusions integrated in the discussion? (There is no separate "Conclusion" section allowed!) - (For Original Paper articles only)

- Have the references been prepared according to the format requested in the guideline to authors?

- Have all taxa names (species names, genus names, and names of higher categories) been italicized?

- Has an Ethical Statement/Conflict of Interest statement been inserted before the list of References?

- It is not allowed to submit the text of your revised version as pdf-file. Please make sure to submit your editable source files (i.e. Word, Tex)

Once again, thank you for submitting your manuscript to "Applied Microbiology and Biotechnology" and I look forward to receiving your revision.

Yours sincerely,

Arnold Demain

International Editor

Applied Microbiology and Biotechnology

ademain@drew.edu;pmaire@optonline.net

Reviewers' comments:

Reviewer #1: This study reports the construction of phage-resisitant strains of Geobacillis thermoglucosidasius, a thermophile that can be cultured at 60 degrees Centigrade. An intention is to target these hosts or their derivatives for production of ethanol or other bioproducts.

Figure 2 nicely displays that one of the constructed strains is resistant to phage infection by GVE3, and likewise for Figure 5.

In "references," #12 omits the journal name. This should be corrected.

Reviewer #2: Geobacillis thermoglucosidasius is a potential organism for the production of biofuels and other metabolites of interest from lignocellulosic biomass. However, this strain is susceptible to specific bacteriophage GVE3 infection. In the present work the authors have developed two G. thermoglucosidasius resistant strains to the infection phage GVE3. The phage encoded an immunity (imm) gene that when overexpressed in the host led to phage resistance in one of the strains. Another resistant strain was obtained by expression of a putative anti-repressor like protein. Both solutions prevented lytic phage infection as judged by the absence of phage DNA inside the host, post infection and an inability to form plaques when using the recombinant strains as host. A point mutation was identified in the polysaccharide pyruvyl transferase, csaB as responsible for the phage resistant phenotype. The experimental evidence suggested that the phage resistant mechanism seems to be related to phage injection prevention and/or lysogenic conversion.

I think the authors approach in this study with G. thermoglucosidasius was acceptable in order to understand the phage resistance mechanism in a thermophile bacteria involved in biofuel production. However, most of the generated knowledge and conclusions have previously been published for Lactococcus, Bacillus and the thermophile Geobacillus sp. E263 (McLaughlin et al. 1986; Bishop-Lilly et al. 2012; Mesnage et al. 2000, Jin et al., 2013). Therefore, the authors must emphasize their main contributions to the basic knowledge of GVE3 bacteriophage infection and go further with experimental evidence to analize the role of SLH proteins in the target for phage attachment.

In two occasions, the authors refer to experiment results not shown in the results section: lines 2012 to 214: "The knock-in strain produced 0.497±0.011 (n=5) grams of ethanol produced per gram glucose consumed while TM242 produced 0.474 g/g ±0.007 (n=5) showing that the strain still performs well in this regard." and lines 254-255. It is advisable to write "not shown" after these sentences.

The manuscript presents a tendency to write joined words like 30ml or plasmidconstruction. This tendency was observed 38 times.

References: Several cites were not included in the list of references like Bhaya et al., 2011; Jakutyte et al., 2011; and Samson et al., 2013. Other references were not cited appropriately like: Durmaz and Klaenhammer, 2007; Ormala and Jalavuori, 2013; Shearwin et al., 1998; and Chang et al., 1995;

Please enter your comments to the Author below:

Chapter 6

General Discussion

P. thermoglucosidans is a promising "platform" organism which has been engineered to produce a range of useful metabolites; initially for bioethanol, but subsequently for isobutanol and polylactic acid for biodegradable plastics (Cripps et al. 2009; Lin et al. 2014). The aim of this study was to i) modify the organism *P. thermoglucosidans* to produce more ethanol, and do so ii) without the threat of phage mediated lysis. We therefore had applied microbiology aspects to the study, however it also allowed us to describe the first virus to infect *P. thermoglucosidans*, and reveal the first clues about its interaction with its host as well as gain insight into the processes governing protein folding in this thermophile. This allowed us to add new knowledge about how thermophilic phages and their hosts interact broadening our basic understanding of biology. The work presented in this thesis is over three years old now and covers a wide range of rather obscure topics in microbiology: conversion of *thermophiles* for ethanol production, *rare* bacterial enzymes and *high temperature* bacteriophages. It is interesting to note that in the years since the publication of the articles not much new progress has been made in these respective fields (Buddrus et al., 2016, Jiang et al., 2017). This shows the pace at which research in these fields moves and that they should perhaps enjoy more attention in future.

The two objectives mentioned above were achieved, although there are caveats. I) In terms of ethanol production, the modifications which had been made to *P. thermoglucosidans* (Δ*ldh*, Δ*pfl*, ↑*pdh*) previously had already produced an organism capable of producing near theoretical amounts of ethanol per unit glucose consumed. Thus, whether or not the development of the Pdc pathway was necessary or currently relevant for ethanol production using this organism is debatable. The development of a Pdc pathway may however still be useful in adjusting carbon flux in the organism as well as for the generation of other metabolites not yet envisioned by those wishing to utilize the organism as a platform for metabolite production.

Probably the most useful output from this study was the demonstration that codon harmonization allowed the successful expression of a mesophilic gene in a thermophile. This showed the ability of a mesophilic protein to fold correctly under elevated temperatures, guided by the stability of its final correctly folded structure under high temperature, if co-translational folding was better controlled. It also showed that codon harmonization, as opposed to the industry standard codon optimization, might serve to optimize other genes for expression in this Gram-positive thermophile. This further adds to the work done by Angov and co-workers to show that codon harmonization is a general codon optimization strategy to enable the expression of various proteins in heterologous hosts.

Since our publication, work on codon harmonization is a field that has moved forward considerably (Quax et al., 2015; Athey et al., 2017; Tian et al., 2017; Claassens et al., 2017). Although Claassens and co-workers demonstrated that, for membrane proteins, transcriptional levels were probably more important for correct protein folding (so as not to overload folding chaperones), codon harmonization proved to be a very effective tool to ensure correct protein folding in heterologous hosts. Depending on the protein, it vastly outperformed the classic optimization strategies. Tian and co-workers have now developed a method for codon harmonization, independent of the need to know the codon usage profile of the native host and used it to greatly improve expression of fluorescent proteins in *E. coli*. Claassens and co-workers have further developed software to enable codon harmonization to be easily applied by researchers looking to improve protein expressions and folding in non-native hosts, showing their confidence in the technique (Claassens et al., 2017). The studies by Tian and Claassens therefore vindicates the original concept proposed by Angov and co-workers demonstrating that rare codons do play a role in correct protein processing and needs to be considered when optimizing proteins for heterologous expression. So as to almost cement codon harmonization as a technique to enable heterologous expression and specifically for *Geobacillus*/*Parageobacillus* species, Buddrus showed improved protein expression for the *Z. palmae* Pdc in *Parageobacillus thermoglucosidans* by using codon harmonization, following our success with GoPDC (Buddrus, 2016). These studies vindicate our decision to apply codon harmonization, as opposed to standard codon optimization algorithms, when trying to improve GoPDC expression in *P. thermoglucosidans*.

The changes in nucleotide sequence made during codon harmonization did change the potential for secondary structures to form at the start of the mRNA molecule, however, according to work done previously, this is not expected to have an effect on translation initiation as it was shown that initiation is only inhibited when the first codon and Shine-Dalgarno sequences form part of the secondary structure (**Figure 6.1**; Plotkin and Kudla 2011). Indeed, a hairpin with lower ΔG value was created starting at the 11th nucleotide. However, just as thermophiles show a very different codon usage pattern compared with thermophiles, their translation initiation may also differ markedly which may have implications for the expression of this gene and others (Ma et al., 2002; Singer and Hickey 2003).

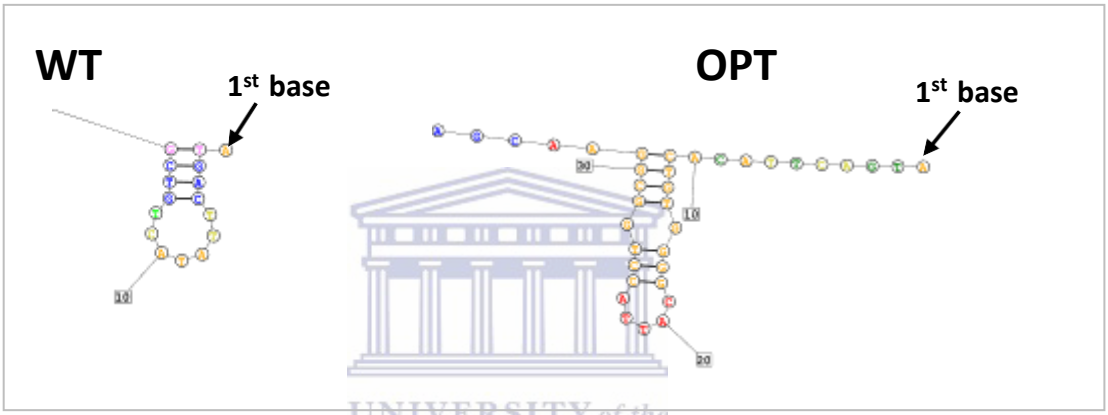


Figure 6.1. Secondary structure prediction at the start of *Gluconobacter oxydans* pyruvate decarboxylase mRNA.

Here I employed the lactate dehydrogenase promoter to drive expression of Pdc genes. Under microaerobic conditions, promoters in the glycolytic and fermentative pathways in *P. thermoglucosidans* are upregulated between 2 to 6-fold on average with the exception of pyruvate formate lyase upregulated 280-fold (Cripps et al., 2009; Loftie-Eaton et al., 2013). The *ldh* promoter is upregulated 3.4-fold which would suggest an "average" induction for the switch from aerobic growth to microaerobic conditions. As little to no lactate production is seen during aerobic growth, the 3.4-fold induction suggests low basal transcription from this promoter under aerobic conditions and an enzyme that is highly efficient, when produced, at converting pyruvate to lactate. The Pdc is therefore also expected to be expressed at these average levels upon induction and it is questionable whether or not chaperone overloading is taking place during its expression. However, in the work done by Buddrus,

they found a 100-fold reduction in codon harmonized ZmPDC expression by RT-qPCR, when moving from 50℃ to 60℃ with the ZmPDC, also expressed from the *ldh* promoter. This suggests a temperature-dependent expression from this promoter, perhaps mediated through decreased dissolved oxygen at elevated temperature, and that at our fermentation temperature of 45℃, we had substantially higher expression of GoPDC from this promoter. This may point to a particular oxygen tension being needed for maximal induction from this promoter as enzymes such as RNA polymerase are expected to work optimally at ±60℃. Perhaps better ethanol yields could be reached through overexpression of the Pdc, including the use of GVE3 promoter sequences, which may be much stronger than host promoters. The availability of promoters such as that driving *pfl* expression, or the development of a series of semi-synthetic promoter sequences which allow tuning of the steady constitutive expression of genes of interest in *Parageobacillus* species, should, in combination with codon harmonization, enable the functional expression of thermophilic, and perhaps some mesophilic proteins in this promising platform organism (Pogrebnyakov et al., 2017).

The only thermophilic GroEL/GroES complex studied from *Geobacillus thermopakistaniensis* shows that it is a group I chaperonin with a temperature optimum for binding of ATP at 65℃ (Ashraf et al., 2017). As binding and hydrolysis of ATP is necessary to advance the complex through the various stages of binding, internalization, and release of folded protein, this suggests that it may have minimal activity or much slower cycling at lower temperatures. During expression of *G. oxydans* Pdc in *P. themoglucosidans* at 45℃, protein monomers may have benefitted from a longer residence time inside the GroEL/ES complex.

Although codon harmonization improved expression of GoxPDC in *P. thermoglucosidans*, it could not overcome the protein's inherent inability to fold correctly at the organism's optimum growth temperature. As eluded to in the literature review, several other adjustments can be made to improve protein expression in heterologous hosts, including modification of the promoter sequence, expression of chaperones and enzyme engineering. A fourth bacterial PDC crystal structure, that of ZpPDC, was reported subsequent to that of GdPDC (Buddrus et al., 2016). Again, the gross structure was no different to those previously described (homo-tetramer), however analysis of the monomer and dimer interfaces

pointed to a greater number of salt bridges and interface area in ZpPDC compared with the other known structures. This is thought to be the source of its superior thermostability compared with the other enzymes. It was recently shown that highly expressed thermophilic proteins have larger subunit interface areas than lowly expressed proteins when compared to mesophilic counterparts (He and Ma 2016). What exactly this relationship means for the evolution of thermophilic proteins is not yet clear, however, this may be indicative of the performance of heterologously expressed proteins (high expression, low subunit interface area = poor performance). Thus, reduced expression levels through promoter modification as mentioned earlier, could have two benefits: i) reduce the load on the folding chaperones and ii) improve performance of native protein subunit interfaces.

Investigators have sought a thermostable Pdc for some time, however there appears to be no natural source of such an enzyme, and engineering efforts are made more difficult by the lack of a suitable assay for protein folding and activity under elevated temperatures. There is still interest in using ThDP-dependent enzymes for production of alcohols in thermophiles. Soh and co-workers recently demonstrated the successful engineering of the *L. lactis* ketoisovalerate decarboxylase (Kivd) to be thermostable (Soh et al., 2017) by using a screen of a mutant library of *kivd* in *E. coli*, lysing cells and assaying at 50℃. Most recently, Tian and co-workers evaluated the effect of expression of four Pdc's (ZpPDC, ZmPDC, GoPDC and ApPDC) in *Clostridium thermocellulum* at 55℃ where they demonstrated that co-expression of ApPDC and *adhA* from *Thermoanaerobacterium saccharolyticum* resulted in a 54% increase in ethanol yield over the parent strain (Tian et al., 2017). Interestingly, GoPDC was functional and resulted in improved ethanol production when expressed alongside the *adhA* at this elevated temperature, indicating that folding of the enzyme was either not a problem at higher temperatures, or that it is assisted in some way in *C. thermocellulum*. As discussed in the literature review and in Chapter 2, the inability of GoPDC to be produced in *P. thermoglucocidans* at temperatures >45℃ likely shows a limitation in its ability to fold under high temperature rather than the thermostability of the homotetramer being the limiting factor, as this complex purified from *E. coli* was relatively stable at 60℃. For Pdc thermo-folding engineering we propose the use of polyclonal antibodies raised against the final correctly folded quaternary structure of Pdc (expressed in *E. coli*) to

screen a random Pdc mutant library in *T. thermophilus* as a way to first establish if the protein folded correctly (more correctly folded protein will recruit more fluorescently labelled antibody), then proceeding with enzymatic assay of the positive clones. In the absence of a complete understanding of all factors involved in protein expression and the sheer diversity of interactions possible in biological systems, the decision of which host/promoter/codon optimization method to employ to ensure successful heterologous protein expression dictates, that currently, these parameters need to be determined empirically. The addition of the fourth Pdc crystal structure, and that of the most thermostable known to date, should guide researchers in rational engineering of the protein for improved thermostability, and molecular dynamic simulations at various temperatures may guide their efforts in improving the folding of the monomers at higher temperatures.

Given the numerous cellular processes that play a role in effective protein expression, there is still much to be explored to enable high level expression of Pdc, or other genes of interest, in *P. thermoglucosidans*.

The second aim of this study was to describe a novel phage that infects *P. themoglucosidans* and to develop strains resistant to infection by it. There is currently a debate over the classification of the genus *Geobacillus* and whether it should be split into two genera (Habibu et al., 2016; Burgess et al., 2017) namely *Geobacillus* and *Parageobacillus*. If the split is accepted, it changes the importance of the phage discovered here, in relation to what was said in our earlier publication (van Zyl et al., 2015). It would make GVE3 only the third phage known to infect a *Parageobacillus* species as well as being the only lysogenic phage known for this genus, the largest phage known for the genus, and still the only known phage to infect *P. thermoglucosidans.*

II) Although I successfully developed phage resistant isolates, the dairy industry experience teaches us that without a range of strategies and strains to prevent further infections, these strains will eventually succumb to new phages. Therefore, although ours is a good first step, it will only protect this organism against GVE3 but not the full, as yet undiscovered, range of *Parageobacillus*-infecting phages in nature and those that evolve to infect this organism which currently infects other thermophilic hosts. Neither

will it protect against versions of GVE3 that are selected for, due to the use of the phage resistant strains produced in this work. Given the efficacy of the CRISPR system to protect against phage infection, it could be employed as a strategy to protect *P. thermoglucosidans* against GVE3. As discussed in chapter 4 two regions of 100% nucleotide identity were identified in a *Parageobacillus* WCH70 CRISPR array. This suggests that these two CRISPR spacers may provide protection against GVE3, and their incorporation into a *P. thermoglucosidans* CRISPR array as shown in **Figure 6.2** could provide another mechanism of protection against this phage. The identification of the SLHD-containing protein, which is expected to be the actual binding target of the phage, will add yet another level of protection.

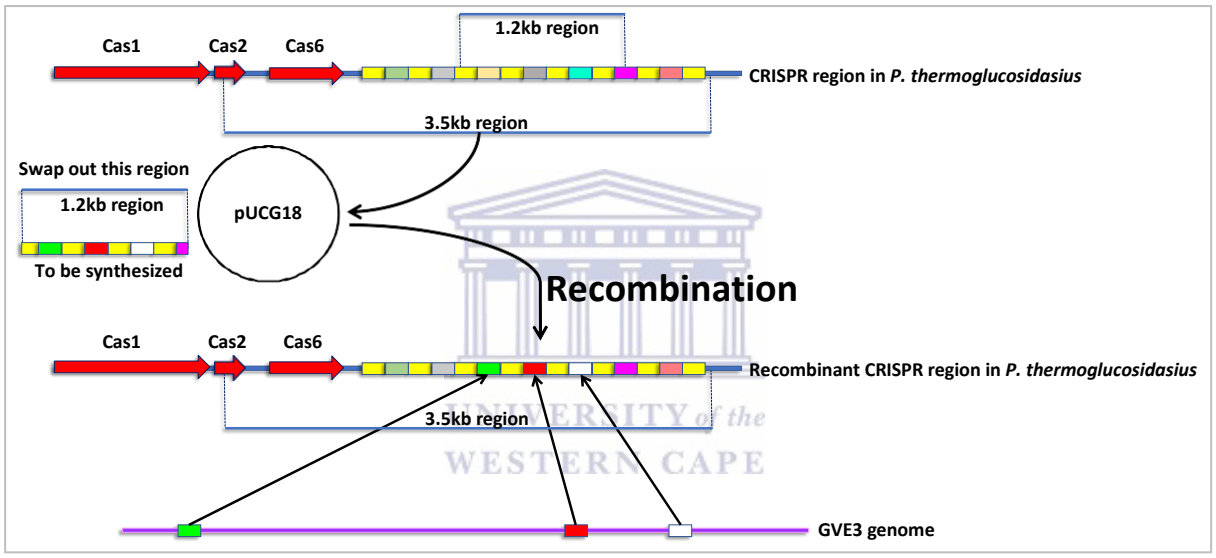


Figure 6.2. Strategy to generate additional GVE3 phage resistance phenotypes in *P. thermoglucosidans*.

Although it is expected that many of the mechanisms of infection, replication, assembly and progeny release will be shared between GVE3 and mesophilic relatives such as phage lambda, as evidenced in GVE2, there will undoubtedly be subtle or perhaps even major, differences. One aspect of thermophilic phage evolution which warrants further investigation is how phages evolve from mesophily to thermophily or *vice versa*. It is easier to understand for lysogenic phages such as GVE3 which can integrate and evolve alongside the host genome, however it is more difficult perhaps for lytic phages which can only be selected for on infection. Do they perhaps switch hosts by moving through a moderately thermophilic host, or are they tied to their current hosts' evolutionary path? Much more

work is needed to characterize the interactions between phages and their hosts, and in particular the thermophilic bacterial viruses for which little is known. Although the evolution of a mesophilic bacterium to a thermophilic version has been demonstrated (Blaby et al., 2012), it is generally accepted that life evolved from high temperature microorganisms (Weiss et al., 2016; Gogarten and Deamer 2016). It is possible that the diversity of viral morphologies observed for thermophilic hosts are remnants of the first experimental designs adopted by microbial viruses prior to going through a selection bottleneck which saw the tailed phages win out in temperate environments.

As mentioned in Chapter 1, lysogenic conversion of bacteria can often offer advantages to the host. Although not reported in the preceding Chapters, the performance of several confirmed GVE3 lysogens were assessed in terms of their growth characteristics and ethanol producing ability. None of the lysogens showed altered growth characteristics compared to both wild type and engineered strains of *P. thermoglucosidans*. All but one lysogen, performed equally well in producing ethanol. Thus, there appears to be no immediate indication of the phage affecting fermentation, other than the lysis of hosts which, obviously, dramatically impacts on the ability of the organism to be used as a cell factory. It may be that in a natural setting the lysogeny of the host by this phage, may offer a distinct advantage, such as providing additional DNA replication proteins (pyrimidine nucleoside phosphorylase, thymidylate synthase, thymidine kinase, ribonucleotide reductase, nucleoside triphosphate pyrophosphohydrolase and nucleoside-deoxyribosyltransferase) or the phosphate starvation protein (PhoH).

The discovery of GVE3 further opens the possibility for the introduction of large metabolic pathways into *P. thermoglucosidans*. This would be advantageous in the case where cellulosomes, for example, would have to be engineered in the organism for improved lignocellulose breakdown. More work is needed to identify the true nature of the ends of the phage genome to determine if it truly has a PAC site and where it is located on the phage genome as speculated on in Chapter 3. Two factors may hamper the use of this phage as vector for the introduction of large genomic fragments is: i) no vector capable of replicating ±140kb of DNA in *P. thermoglucosidans* has been identified and ii) the ability to work with and clone such large inserts is not easy. If the phage replication requirements are established, this

together with the delineation of the physical ends may enable the construction of a vector capable of accepting and replicating such a large fragment of DNA in *P. thermoglucosidans*. This may have to be married with a technique such as transformation assisted recombination (TAR cloning) to make the system work. The discovery of *P. thermoglucosidans*-infecting phages with smaller genomes would make this a more tractable problem and provides an incentive for continued screening to identify such phages.

As said earlier, in this body of work I've attempted to bring together a range of topics in microbiology under the heading "Engineering *P. thermoglucosidans* as a robust platform for bioethanol production". The successful modification of a thermophilic bacterial strain capable of producing more ethanol than its parent strain and doing so while resistant to the only phage currently known to infect it would, in the opinion of the author, qualifies as giving credence to this title.

References

1) Abbani MA, Papagiannis CV, Sam MD, Cascio D, Johnson RC, Clubb RT. 2007. Structure of the cooperative Xis–DNA complex reveals a micronucleoprotein filament that regulates phage lambda intasome assembly. Proc. Natl. Acad. Sci. 104: 2109-2114
2) Abdel-Banat BMA, Nonklang S, Hoshida H, Akada R. 2010. Random and targeted gene integrations through the control of non-homologous end joining in the yeast *Kluyveromyces marxianus*. Yeast 27: 29–39
3) Ackermann HW. 2003. Bacteriophage observations and evolution. Res. Microbiol. 154: 245-51
4) Ackermann H-W. 2007. 5500 Phages examined in the electron microscope. Arch. Virol. 152: 227-243
5) Adams MJ, Lefkowitz EJ, King AMQ, Harrach B, Harrison RL, Knowles NJ, Kropinski AM, Krupovic M, Kuhn JH, Mushegian AR, Nibert ML, Sabanadzovic S, Sanfacon H, Siddell SG, Simmonds P, Varsani A, Zerbini FM, Orton RJ, Smith DB, Gorbalenya AE, Davison AJ. 2017. 50 years of the International Committee on Taxonomy of Viruses: Progress and prospects. Arch. Virol. 162: 1441-1446
6) Adriaenssens EM, Brister JR. 2017. How to name and classify your phage: An informal guide. Viruses 9: 70
7) Adriaenssens EM, Kramer R, Van Goethem MW, Makhalanyane TP, Hogg I, Cowan DA. 2017. Environmental drivers of viral community composition in Antarctic soils identified by viromics. Microbiome 5: 83
8) Adriaenssens EM, Van Zyl L, De Maayer P, Rubagotti E, Rybicki E, Tuffin M, Cowan DA. 2014. Metagenomic analysis of the viral community in Namib Desert hypoliths. Environ. Microbiol. 17: 480-495
9) Akhtar N, Gupta K, Goyal D, Goyal A. 2015. Recent advances in pretreatment technologies for efficient hydrolysis of lignocellulosic biomass. Environmental Progress and Sustainable Energy. 35: 489-511
10) Ali Y, Koberg S, Heßner S, Sun X, Rabe B, Back A, Neve H, Heller KJ. 2014. Temperate *Streptococcus thermophilus* phages expressing superinfection exclusion proteins of the Ltp type. Front. Microbiol. 5: 98
11) Aliyu H, Lebre P, Blom J, Cowan D, De Maayer P. 2016. Phylogenomic re-assessment of the thermophilic genus *Geobacillus*. Syst. App. Microbiol. 39: 527-533
12) Altintas MM, Eddy CK, Zhang M, McMillan JD, and Kompala DS. 2006. Kinetic modeling to optimize pentose fermentation in *Zymomonas mobilis*. Biotechnol Bioeng 94: 273–295

13) An YJ, Rowland SE, Na J-H, Spigolon D, Hong SK, Yoon YJ, Lee J-H, Robb FT, Cha S-S. 2017. Structural and mechanistic characterization of an archaeal-like chaperonin from a thermophilic bacterium. Nature Communications 8: 827
14) Andrews FH, Tom AR, Gunderman PR, Novak WRP, McLeish MJ. 2013. A bulky hydrophobic residue is not required to maintain the v-conformation of enzyme-bound thiamin diphosphate. Biochemistry 52: 3028-3030
15) Andrews FH, Wechsler C, Rogers MP, Meyer D, Tittmann K, McLeish MJ. 2016. Mechanistic and structural insight to an evolved benzoylformate decarboxylase with enhanced pyruvate decarboxylase activity. Catalysts 6: 190
16) Anfinsen CB. 1973. Principles that govern the folding of protein chains. Science 181: 223-230
17) Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P, Felts B, Nulton J, Mahaffy J, Rohwer F. 2005. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics. 6: 41
18) Angov E, Hillier CJ, Kincaid RL, Lyon JA. 2008. Heterologous protein expression is enhanced by harmonizing the codon usage frequencies of the target gene with those of the expression host. PLoS One. 3: e2189.
19) Ardell DH, Kirsebom LA. 2005. The genomic pattern of tDNA operon expression in *E. coli*. PLoS Comput Biol. 1: e12
20) Aro E-M. 2016. From first generation biofuels to advanced solar biofuels. Ambio 2016, 45: S24–S31
21) Ashraf R, Muhammad MA, Rashid N, Akhtar M. 2017. Cloning and characterization of thermostable GroEL/GroES homologues from *Geobacillus thermopakistaniensis* and their applications in protein folding. J. Biotechnol. 254: 9-16
22) Atadashi IM, Aroua MK, Abdul Aziz AR, Sulaiman NMN. 2013. The effects of catalysts in biodiesel production: A review. Journal of Industrial and Engineering Chemistry 19: 14-26
23) Athey J, Alexaki A, Osipova E, Rostovtsev A, Santana-Quintero LV, Katneni U, Simonyan V, Kimchi-Sarfaty C. 2017. A new and updated resource for codon usage tables. BMC Bioinformatics. 18: 391.
24) Åvall-Jääskeläinen S, Palva A. 2005. *Lactobacillus* surface layers and their applications. FEMS Microbiol. Rev. 29: 511-529
25) Baeshen NA, Baeshen MN, Sheikh A, Bora RS, Ahmed MMM, Ramadan HAI, Saini KS, Redwan EM. 2014. Cell factories for insulin production. Microb Cell Fact. 13: 141.
26) Baker-Austin C, Dopson M. 2007. Life in acid: pH homeostasis in acidophiles. TRENDS in Microbiology 15: 165-171
27) Ball CA, Johnson RC. 1991. Efficient excision of phage lambda from the *Escherichia coli*

chromosome requires the Fis protein. J. Bacteriol. 173: 4027-4031

28) Baneyx F, Palumbo JL. 2003. Improving heterologous protein folding via molecular chaperone and foldase co-expression. In: Vaillancourt P.E. (eds) *E. coli* Gene Expression Protocols. Methods in Molecular Biology™, vol 205. Humana Press

29) Barber JM. 1977. Studies on the fermentation of molasses by *Clostridium acetobutylicum*. M.Sc. thesis. Rhodes University, Grahamstown, South Africa.

30) Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. Science 315: 1709-1712

31) Bartosiak-Jentys J, Hussein AH, Lewis CJ, Leak DJ. 2013. Modular system for assessment of glycosyl hydrolase secretion in *Geobacillus thermoglucosidasius*. Microbiology 159: 1267-1275

32) Baykal AT, Kakalis L, Jordan F. 2006. Electronic and nuclear magnetic resonance spectroscopic features of the 1',4'-iminopyrimidine tautomeric form of thiamin diphosphate, a novel intermediate on enzymes requiring this coenzyme. Biochemistry 45: 7522-7528

33) Bebeacua C, Fajardo JCL, Blangy S, Spinelli S, Bollmann S, Neve H, Cambillau C, Heller KJ. 2013. X-ray structure of a superinfection exclusion lipoprotein from phage TP-J34 and identification of the tape measure protein as its target. Mol. Microbiol. 89: 152-165

34) Bhatia SK, Kim S-H, Yoon J-J, Yang Y-H. 2017. Current status and strategies for second generation biofuel production using microbial systems. Energy Conversion and Management 148: 1142-1156

35) Bidnenko E, Ehrlich SD, Chopin MC. 1998. *Lactococcus lactis* phage operon coding for an endonuclease homologous to RuvC. Mol. Microbiol. 28: 8238-34

36) Blaby IK, Lyons BJ, Wroclawska-Hughes E, Phillips GCF, Pyle TP, Chamberlin SG, Benner SA, Lyons TJ, de Crécy-Lagard V, de Crécy E. 2012. Experimental evolution of a facultative thermophile from a mesophilic ancestor. Appl. Environ. Microbiol. 78: 144-155

37) Bläsi U, Young R. 1996. Two beginnings for a single purpose: The dual-start holins in the regulation of phage lysis. Mol. Microbiol. 21: 675-682

38) Bokinsky G, Peralta-Yahya PP, George A, Holmes BM, Steen EJ, Dietrich J, Lee TS, Tullman-Ercek D, Voigt CA, Simmons BA, Keasling JD. 2011. Synthesis of three advanced biofuels from ionic liquid-pretreated switchgrass using engineered *Escherichia coli*. Proc Natl Acad Sci 108: 19949-19954.

39) Bolduc B, Jang HB, Doulcier G, You Z-Q, Roux S, Sullivan MB. 2017. vConTACT: An iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. PeerJ. 5: e3243

40) Bondy-Denomy J, Garcia B, Strum S, Du M, Rollins MF, Hidalgo-Reyes Y, Wiedenheft B, Maxwell KL, Davidson AR. 2015. Multiple mechanisms for CRISPR–Cas inhibition by anti-CRISPR proteins. Nature 526: 136-139

41) Bondy-Denomy J, Pawluk A, Maxwell KL, Davidson AR. 2013. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. Nature. 493: 429-432

42) Bonhivers M, Letellier L. 1995. Calcium controls phage T5 infection at the level of the *Escherichia coli* cytoplasmic membrane. FEBS Lett. 374: 169-173

43) Bothast R, Schlicher M. 2005. Biotechnological processes for conversion of corn into ethanol. Appl. Microbiol. Biotechnol. 67: 19-25.

44) Braatsch S, Helmark S, Kranz H, Koebmann B, Jensen PR. 2005. Rapid fine tuning of *Eschericia coli* gene expression. Biotechniques. 45: 1-4.

45) Brandt GS, Kneen MM, Chakraborty S, Baykal AT, Nemeria N, Yep A, Ruby DI, Petsko GA, Kenyon GL, McLeish MJ, Jordan F, Ringe D. 2009. Snapshot of a reaction intermediate: analysis of benzoylformate decarboxylase in complex with a benzoylphosphonate inhibitor. Biochemistry 48: 3247-3257

46) Breitbart M, Rohwer F. 2005. Here a virus, there a virus, everywhere the same virus? Trends Microbiol. 13: 278-284

47) Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F. 2002. Genomic analysis of uncultured marine viral communities. Proc. Natl. Acad. Sci. 99: 14250-14255

48) Broecker F, Klumpp J, Schuppler M, Russo G, Biedermann L, Hombach M, Rogler G, Moelling K. 2016. Long-term changes of bacterial and viral compositions in the intestine of a recovered *Clostridium difficile* patient after fecal microbiota transplantation. Cold Spring Harb. Mol. Case Stud. 2: a000448

49) Brouns SJJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RKH, Snijders APL, Dickman MJ, Makarova KS, Koonin EV, van der Oost J. 2008. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science 321: 960-963

50) Browning DF, Busby SJW. 2016. Local and global regulation of transcription initiation in bacteria. Nature Reviews Microbiology. doi:10.1038/nrmicro.2016.103

51) Brum JR, Hurwitz BL, Schofield O, Ducklow HW, Sullivan MB. 2008. Seasonal time bombs: Dominant temperate viruses affect Southern Ocean microbial dynamics. ISME J. 10: 437-449

52) Bruttin A, Desiere F, Lucchini S, Foley S, Brüssow H. 1997. Characterization of the lysogeny DNA module from the temperate *Streptococcus thermophilus* bacteriophage Sfi21. Virology 233: 136-148

53) Bruttin A, Foley S, Brüssow H. 2002. DNA-binding activity of the *Streptococcus thermophilus* phage Sfi21 repressor. Virology 303: 100-109

54) Bryan D, El-Shibiny A, Hobbs Z, Porter J, Kutter EM. 2016. Bacteriophage T4 infection of stationary phase *E. coli*: Life after log from a phage perspective. Front. Microbiol. 7: 1391

55) Buddrus L, Andrews ESV, Leak DJ, Danson M, Arcus VL, Crennell SJ. 2016. Crystal structure of pyruvate decarboxylase from Zymobacter palmae. Acta. Crystallogr. F Struct.

Biol. Commun. 72: 700-706

56) Buddrus L. 2016. Creation and evaluation of a pyruvate decarboxylase dependent ethanol fermentation pathway in *Geobacillus thermoglucosidasius*. PhD thesis University of Bath.

57) Burgess SA, Flint SH, Lindsay D, Cox MP, Biggs PJ. 2017. Insights into the *Geobacillus stearothermophilus* species based on phylogenomic principles. BMC Microbiology. 17: 140 doi.org/10.1186/s12866-017-1047-x

58) Casjens SR, Hendrix RW. 2015. Bacteriophage lambda: Early pioneer and still relevant. Virology. 479-480:310-330

59) Chakraborty S, Nemeria NS, Balakrishnan A, Brandt GS, Kneen MM, Yep A, McLeish MJ, Kenyon GL, Petsko GA, Ringe D, Jordan F. 2009. Detection and time course of formation of major thiamin diphosphate-bound covalent intermediates derived from a chromophoric substrate analogue on benzoylformate decarboxylase. Biochemistry 48: 981-994

60) Chhibber S, Kaur T, Kaur S. 2013. Essential role of calcium in the infection process of broad-spectrum methicillin-resistant *Staphylococcus aureus* bacteriophage. J. Basic Microbiol. 54: 775-780

61) Claassens NJ, Silia MF. 2017. Improving heterologous membrane protein production in *Escherichia coli* by combining transcriptional tuning and codon usage algorithms. PLoS ONE 12: e0184355

62) Clare DK, Vasishtan D, Stagg S, Quispe J, Farr GW, Topf M, Horwich AL, Saibil HR. 2012. ATP-triggered conformational changes delineate substrate-binding and -folding mechanics of the GroEL chaperonin. Cell. 149: 113-123.

63) Clausen T, Southan C, Ehrmann M. 2002. The HtrA Family of Proteases: Implications for Protein Composition and Cell Fate. Cell 10: 443-455.

64) Clokie MRJ, Millard AD, Letarov AV, Heaphy S. 2011. Phages in nature. Bacteriophage 1: 31-45

65) Cook GM, Russell JB, Reichert A, Wiegel J. 1996. The Intracellular pH of *Clostridium paradoxum*, an Anaerobic, Alkaliphilic, and Thermophilic Bacterium. Applied and Environmental Microbiology 62: 4576-4579

66) Costa DA, de Souza CJA, Costa PS, Rodrigues MQRB, dos Santos AF, Lopes MR, Genier HLA, Silveira WB, Fietto LG. 2014. Physiological characterization of thermotolerant yeast for cellulosic ethanol production. Appl. Microbiol. Biotechnol. 98: 3829-3840

67) Couturier E, Rocha EP. 2006. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. Mol Microbiol. 59:1506-1518

68) Cripps RE, Eley K, Leak DJ, Rudd B, Taylor M, Todd M, Boakes S, Martin S, Atkinson T. 2009. Metabolic engineering of *Geobacillus thermoglucosidasius* for high yield ethanol production. Metab. Eng. 11: 398-408

69) Danovaro R, Corinaldesi C, Dell'Anno A, Fuhrman JA, Middelburg JJ, Noble RT, Suttle CA. 2011. Marine viruses and global climate change. FEMS Microbiol. Rev. 35: 993-1034

70) De Maayer P, Brumm PJ, Mead DA, Cowan DA. 2014. Comparative analysis of the *Geobacillus* hemicellulose utilization locus reveals a highly variable target for improved hemicellulolysis. BMC Genomics 15: 836 doi: 10.1186/1471-2164-15-836

71) Demir AS, Ayhan P, Sopaci SB. 2007. Thiamine pyrophosphate dependent enzyme catalyzed reactions: Stereoselective c–c bond formations in water. Clean Soil Air Water 35: 406-412

72) den Haan R, Kroukamp H, Mert M, Bloom M, Görgens JF, van Zyl WH. 2013. Engineering *Saccharomyces cerevisiae* for next generation ethanol production. J Chem Technol Biotechnol. 88: 983-991

73) Dilucca M, Cimini G, Semmoloni A, Deiana A, Giansanti A. 2015. Codon Bias Patterns of *E. coli*'s Interacting Proteins. PLoS ONE 10: e0142127.

74) Diruggiero J, Robb FT. 1995. Expression and *in vitro* assembly of recombinant glutamate dehydrogenase from the hyperthermophilic archaeon *Pyrococcus furiosus*. Appl. Environ. Microbiol. 61: 159-164.

75) Djordjevic GM, O'Sullivan DJ, Walker SA, Conkling MA, Klaenhammer, TR. 1997. A triggered-suicide system designed as a defense against bacteriophages. J. Bacteriol. 179: 6741-6748

76) Dodd IB, Shearwin KE, Egan JB. 2005. Revisited gene regulation in bacteriophage phage Lambda. Curr. Opin. Genet. Dev. 15: 145-152

77) Doi K, Mori K, Martono H, Nagayoshi Y, Fujino Y, Tashiro, K, Kuhara S, Ohshimaa T. 2013. Draft Genome Sequence of *Geobacillus kaustophilus* GBlys, a Lysogenic Strain with Bacteriophage OH2. Genome Announc. 1: e00634-13

78) Doyle SM, Genest O, Wickner S. 2013. Protein rescue from aggregates by powerful molecular chaperone machines. Nature Rev. 14: 617-629

79) Dulermo R, Brunel F, Dulermo T, Ledesma-Amaro R, Vion J, Trassaert M, Thomas S, Nicaud J-M, Leplat C. 2017. Using a vector pool containing variable-strength promoters to optimize protein production in *Yarrowia lipolytica*. Microbial Cell Factories 6: 31

80) Durmaz E, Klaenhammer TR. 2007. Abortive phage resistance mechanism AbiZ speeds the lysis clock to cause premature lysis of phage-infected *Lactococcus lactis*. J. Bacteriol. 189: 1417-1425

81) Dy RL, Przybilski R, Semeijn K, Salmond GPC, Fineran PC. 2014b. A widespread bacteriophage abortive infection system functions through a Type IV toxin–antitoxin mechanism. Nucleic Acids Res. 42: 4590-4605

82) Dy RL, Richter C, Salmond GPC, Fineran PC. 2014a. Remarkable Mechanisms in Microbes to Resist Phage Infections. Annu. Rev. Virol. 1: 307-331

83) Echols H, Court D, Green L. 1976. On the nature of *cis*-acting regulatory proteins and genetic

organization in bacteriophage: The example of gene Q of bacteriophage lambda. Genetics 83: 5-10

84) Egbert LN, Mitchell HK. 1967. Characteristics of Tϕ3, a bacteriophage for Bacillus stearothermophilus. Journal of Virology 1: 610-616

85) Elbreki M, Ross RP, Hill C, O'Mahony J, McAuliffe O, Coffey A. 2014. Bacteriophages and their derivatives as biotherapeutic agents in disease prevention and treatment. J. Viruses 2014: 382539

86) Englaender JA, Jones JA, Cress BF, Kuhlman TE, Linhardt RJ, Koffas MAG. 2017. Effect of genomic integration location on heterologous protein expression and metabolic engineering in *E. coli*. ACS Synth. Biol. 6: 710-720

87) Eram MS, Ma K. 2013. Decarboxylation of pyruvate to acetaldehyde for ethanol production by hyperthermophiles. Biomol. 3: 578-596

88) Eram MS, Oduaran E, Ma K. 2014. The bifunctional pyruvate decarboxylase/pyruvate ferrodoxin oxidoreductase from *Thermococcus guaymasensis*. Archaea 2014: 349379 doi:10.1155/2014/349379.

89) Eram MS, Oduaran E, Ma K. 2014. The bifunctional pyruvate decarboxylase/ pyruvate ferrodoxin oxidoreductase from *Thermococcus guaymasensis*. Archaea doi:10.1155/2014/349379.

90) Fagan RP, Fairweather NF. 2014. Biogenesis and functions of bacterial S-layers. Nature Rev. Microbiol. 12: 211-222

91) Fineran PC, Blower TR, Foulds IJ, Humphreys DP, Lilley KS, Salmond GP. 2009. The phage abortive infection system, ToxIN, functions as a protein-RNA toxin-antitoxin pair. Proc. Natl Acad. Sci. 106: 894-899

92) Fonseca GG, Heinzle E, Wittmann C, Gombert AK. 2008. The yeast *Kluyveromyces marxianus* and its biotechnological potential. Appl. Microbiol. Biotechnol. 79: 339-354

93) Frank RAW, Titman CM, Pratap JV, Luisi BF, Perham RN. 2004. A molecular switch and proton wire synchronize the active sites in thiamine enzymes. Science 306: 872-876

94) Frappier V, Najmanovich R. 2014. Vibrational entropy differences between mesophile and thermophile proteins and their use in protein engineering. Prot Sci 24: 474-483

95) Fuhrman JA. 1999. Marine viruses and their biogeochemical and ecological effects. Nature. 399: 541-548

96) Gabardo S, Pereira GF, Rech R, Ayub MAZ. 2015. The modeling of ethanol production by *Kluyveromyces marxianus* using whey as substrate in continuous A-Stat bioreactors. J. Ind. Microbiol. Biotechnol. 42: 1243-1253

97) Galiez C, Siebert M, Enault F, Vincent J, Söding J. 2017. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs Bioinformatics. 33: 3113–3114

98) Gallagher PW, Yee WC, Baumes ES. 2015. 2015 Energy Balance for the Corn-Ethanol

Industry. USDA

99) Geller BL, Ivey RG, Trempy JE, Hettinger-Smith B. 1993. Cloning of a chromosomal gene required for phage infection of *Lactococcus lactis* subsp. *lactis* C2. J. Bacteriol. 175: 5510-5519

100) Gírio FM, Fonseca C, Carvalheiro F, Duarte LC, Marques S, Bogel-Łukasik R. 2010. Hemicelluloses for fuel ethanol: A review. Bioresour Technol. 101: 4775-4800

101) Gogarten JP, Deamer D. 2016. Is LUCA a thermophilic progenote? Nature Microbiol. 1: 16229, doi: 10.1038/NMICROBIOL.2016.229

102) Gonzalez R, Tao H, Purvis JE, York SW, Shanmugam KT, Ingram LO. 2003. Gene array based identification of changes that contribute to ethanol tolerance in ethanologenic *Escherichia coli*: comparison of KO11 (parent) to LY01 (resistant mutant). Biotechnol Prog 19: 612-623

103) Gorski SA, Vogel J, Doudna JA. 2017. RNA-based recognition and targeting: Sowing the seeds of specificity. Nat. Rev. Mol. Cell Biol. 18: 215-228

104) Grose JH, Jensen GL, Burnett SH, Breakwell DP. 2014. Genomic comparison of 93 *Bacillus* phages reveals 12 clusters, 14 singletons and remarkable diversity. BMC Genomics. 15: 855 doi.org/10.1186/1471-2164-15-855

105) Gründling A, Manson MD, Young R. 2001. Holins kill without warning. Proc. Natl. Acad. Sci. 98: 9348-9352

106) Gulstrom TJ, Pearce LE, Sandine WE, Elliker PR. 1979. Evaluation of commercial phage inhibitory media. J. Dairy Sci. 62: 208-221

107) Gustafsson C, Govindarajan S, Minshull J. 2004. Codon bias and heterologous protein expression. Trends Biotechnol. 22: 346-353.

108) Hack CJ, Marchant R. 1998. Ethanol adaptation in a thermotolerant yeast strain *Kluyveromyces marxianus* IMB3. J. Ind. Microbiol. Biotechnol. 20: 227-231

109) Halfmann C, Gu L, Gibbons W, Zhou R. 2014. Genetically engineering cyanobacteria to convert CO2, water, and light into the long-chain hydrocarbon farnesene. Appl Microbiol Biotechnol 98: 9869

110) Haq A. 1984. Occurrence of pyruvate decarboxylase in *Erwinia amylovora*. Pak. J. Sci. Ind. Res. 27: 8-13

111) Hartinger D, Heinl S, Schwartz H, Grabherr R, Schatzmayr G, Haltrich D, Moll W-D. 2010. Enhancement of solubility in *Escherichia coli* and purification of an aminotransferase from *Sphingopyxis* sp. MTA144 for deamination of hydrolyzed fumonisin B1. Microb Cell Factories 9: 62.

112) Hatano M, Nakamura K, Kurokawa M. 1959. Isolation of a new temperature phage causing the lysogenic conversion in *Corynebacterium diphtheriae*. Jpn. J. Microbiol. 3: 301-311Haugen SP, Ross W, Gourse RL. 2008. Advances in bacterial promoter recognition and

its control by factors that do not bind DNA. Nature Reviews Microbiology. doi:10.1038/nrmicro1912

113) He Y-M, Ma B-G. 2016. Abundance and temperature dependency of protein-protein interaction revealed by interface structure analysis and stability evolution. Scientific Reports 6: 26737

114) Hidalgo A, Betancor L, Moreno R, Zafra O, Cava F, Fernández-Lafuente R, Guisán JM, Berenguer J. 2004. *Thermus thermophilus* as a cell factory for the production of a thermophilic Mn-dependent catalase which fails to be synthesized in an active form in *Escherichia coli*. Appl. Environ. Microbiol. 70: 3839-3844

115) Hill J, Nelson E, Tilman D, Polasky S, Tiffany D. 2006. Environmental, economic, and energetic costs and benefits of biodiesel and ethanol biofuels. Proc. Natl. Acad. Sci. 103: 11206-11210

116) Houry WA, Frishman D, Eckerskorn C, Lottspeich F, Hartl FU. 1999. Identication of in vivo substrates of the chaperonin GroEL. Nature 402: 147-154

117) Hu B, Margolin W, Molineux IJ, Liu J. 2015. Structural remodeling of bacteriophage T4 and host membranes during infection initiation. Proc. Natl. Acad. Sci. 112: E4919-E4928

118) Huang S, Wang K, Jiao N, Chen F. 2011. Genome sequences of siphoviruses infecting marine *Synechococcus* unveil a diverse cyanophage group and extensive phage-host genetic exchanges. Environ. Microbiol. 14: 540-558

119) Huerta-Beristain G, Cabrera-Ruiz R, Hernandez-Chavez G, Bolivar F, Gosset G, Martinez A. 2016. Metabolic engineering and adaptive evolution of *Escherichia coli* KO11 for ethanol production through the Entner–Doudoroff and the pentose phosphate pathways. J. Chem. Technol. Biotechnol. doi: 10.1002/jctb.5138

120) Hyman P, Abedon ST. 2012. Smaller fleas: Viruses of microorganisms. Scientifica 2012: 734023

121) Ikemura T. 1985. Codon Usage and tRNA Content in Unicellular and Multicellular Organisms. Mol. Bid. Evol. 2: 13-34.

122) Ilmen M, Den Haan R, Brevnova E, McBride J, Wiswall E, Froehlich A, ´Koivula A, Voutilainen SP, Siika-aho M, Lagrange DC, Thorngren N, Ahlgren S, Mellon M, Deleault K, Rajgarhia V, Van Zyl WH, Penttila M. 2011. High level secretion of cellobiohydrolases by *Saccharomyces cerevisiae*. Biotechnol. Biofuels 4: 30.

123) Ingram LO, Aldrich HC, Borges ACC, Causey TB, Martinez A, Morales F, Saleh A, Underwood SA, Yomano LP, York SW, Zaldivar J, Zhou S. 1999. Enteric bacterial catalysts for fuel ethanol production. Biotechnol. Prog. 15: 855-86

124) Jakutytė L, Baptista C, São-José C, Daugelavičius R, Carballido-López R, Tavares P. 2011. Bacteriophage infection in rod-shaped Gram-Positive bacteria: Evidence for a preferential polar route for phage SPP1 entry in *Bacillus subtilis*. J. Bacteriol. 193: 4893-4903

125) Jeffries TW. 1983. Utilization of xylose by bacteria, yeasts, and fungi. Adv. Biochem. Eng. Biotechnol. 27: 1-32

126) Jensen PR, Hammer K. 1998. The sequence of spacers between the consensus sequences modulates the strength of prokaryotic promoters. Appl Env Microbiol. 64: 82-87

127) Jiang Y, Xin F, Lu J, Dong W, Zhang W, Zhang M, Wu H, Ma J, Jiang M. 2017. State of the art review of biofuels production from lignocellulose by thermophilic bacteria. Bioresource Technology. doi.org/10.1016/j.biortech.2017.05.142

128) Jin M, Ye T, Zhang X. 2013. Roles of bacteriophage GVE2 endolysin in host lysis at high temperatures. Microbiology 159: 1597-1605

129) Joh RI, Weitz JS. 2011. To lyse or not to lyse: Transient-mediated stochastic fate determination in cells infected by bacteriophages. PLOS Comput. Biol. 7: e1002006

130) Jones DT, Shirley M, Wu X, Keis S. 2000. Bacteriophage infections in the industrial acetone butanol (AB) fermentation process. J. Mol. Microbiol. Biotechnol. 2: 21-26

131) Jones DT, Woods DR. 1986. Acetone-Butanol fermentation revisited. Microbiol. Rev. 50: 484-524

132) Jönsson LJ, Alriksson B Nilvebrant N-O. 2013. Bioconversion of lignocellulose: inhibitors and detoxification. Biotechnology for Biofuels 6: 16

133) Kern D, Kern G, Neef H, Tittmann K, Killenberg-Jabs M, Wikner C, Schneider G, Hubner G. 1997. How thiamine diphosphate is activated in enzymes. Science 275: 67-70

134) Kern J, Ryan C, Faull K, Schneewind O. 2010. *Bacillus anthracis* surface-layer proteins assemble by binding to the secondary cell wall polysaccharide in a manner that requires *csaB* and *tagO*. J. Mol. Biol. 401: 757-775

135) Kim Y, Ingram LO, Shanmugam KT. 2007. Construction of an *Escherichia coli* K-12 mutant for homoethanologenic fermentation of glucose or xylose without foreign genes. Appl. Environ. Microbiol. 73: 1766-1771.

136) Klein-Marcuschamer D, Oleskowicz-Popiel P, Simmons BA, Blanch HW. 2012. The challenge of enzyme cost in the production of lignocellulosic biofuels. Biotechnology and Bioengineering. 109: 1083-1087

137) Knowles B, Silveira CD, Bailey BA, Barott K, Cantu VA, Cobián-Güemes AG, Coutinho FH, Dinsdale EA, Felts B, Furby KA, George EE, Green KT, Gregoracci GB, Haas AF, Haggerty JM, Hester ER, Hisakawa N, Kelly LW, Lim YW, Little M, Luque A, McDole-Somera T, McNair K, de Oliveira LS, Quistad SD, Robinett NL, Sala E, Salamon P, Sanchez SE, Sandin S, Silva GGZ, Smith J,

Sullivan C, Thompson C, Vermeij MJA, Youle M, Young C, Zgliczynski B, Brainard R, Edwards RA, Nulton J, Thompson F, Rohwer F. 2016. Lytic to temperate switching of viral communities. Nature 531: 466-470

138) Kobiler O, Rokney A, Friedman N, Court DL, Stavans J, Oppenheim AB. 2005. Quantitative kinetic analysis of the bacteriophage l genetic network. Proc. Natl. Acad. Sci. 102: 4470-4475

139) Kommireddy Vasu, Valakunja Nagaraja. 2013. Diverse functions of restriction-modification systems in addition to cellular defense. J. Bacteriol. 77: 53-72

140) Konig S. 1998. Subunit structure, function and organisation of pyruvate decarboxylases from various organisms. Biochim. Biophys. Acta 1385: 271-286

141) Koser, S. A. 1926. Action of the bacteriophage on a thermophilic *Bacillus*. Proc. Soc. Exptl. Biol. Med. 24: 109-111

142) Kourilsky P. 1973. Lysogenization by bacteriophage lambda. I. Multiple infection and the lysogenic response. Mol. Gen. Genet. 122: 183-195

143) Krupovic M, Prangishvili D, Hendrix RW, Bamford DH. 2011. Genomics of bacterial and archaeal viruses: Dynamics within the prokaryotic virosphere. Microbiol. Mol. Biol. Rev. 75: 610-635

144) Kumar AK, Sharma S. 2017. Recent updates on different methods of pretreatment of lignocellulosic feedstocks: a review. Bioresources and Bioprocessing. 4: 7

145) Kumar RR, Prasad S. 2011. Metabolic engineering of bacteria. Indian J. Microbiol. 51: 403-409.

146) Lambertz C, Garvey M, Klinger J, Heesel D, Klose H, Fischer R, Commandeur U. 2014. Challenges and advances in the heterologous expression of cellulolytic enzymes: a review. Biotechnology for Biofuels 7: 135

147) Le Romancer M, Gaillard M, Geslin C, Prieur D. 2007. Viruses in extreme environments. p99-113. *In* Life in Extreme Environments, Eds. Ricardo Amils, Cynan Ellis-Evans, Helmut Hinghofer-Szalkay. Springer, Netherlands

148) Lee AF, Bennett JA, Manayil JC, Wilson K. 2014. Heterogeneous catalysis for sustainable biodiesel production via esterification and transesterification. Chem. Soc. Rev. 43: 7887-7916

149) Li X, Gerlach D, Du X, Larsen J, Stegger M, Kühner P, Peschel A, Xia G, Winstel V. 2015. An accessory wall teichoic acid glycosyltransferase protects *Staphylococcus aureus* from the lytic activity of *Podoviridae*. Scientific Reports 5: doi:10.1038/srep17219

150) Lie MA, Celik L, Jorgensen KA, Schiott B. 2005. Cofactor activation and substrate binding in pyruvate decarboxylase. Insights into the reaction mechanism from molecular dynamics simulations. Biochemistry 44: 14792-14806

151) Lin PP, Rabe KS, Takasumi JL, Kadisch M, Arnold FH, Liao JC. 2014. Isobutanol production at elevated temperatures in thermophilic *Geobacillus thermoglucosidasius*. Metab. Eng. 24: 1-8

152) Lin Z, Rye HS. 2004. Expansion and Compression of a Protein Folding Intermediate by GroEL. Molecular Cell 16: 23-34

153) Liu B, Wu S, Song Q, Zhang X, Xie L. 2006. Two novel bacteriophages of thermophilic bacteria isolated from deep-sea hydrothermal fields. Curr. Microbiol. 53: 163-166

154) Liu B, Zhang X. 2008. Deep-sea thermophilic *Geobacillus* bacteriophage GVE2 transcriptional profile and proteomic characterization of virions. Appl. Microbiol. Biotechnol. 80: 697-707

155) Liu B, Zhou F, Wu S, Xu Y, Zhang X. 2009. Genomic and proteomic characterization of a thermophilic *Geobacillus* bacteriophage GBSV1. Res. Microbiol. 160: 166-170

156) Loftie-Eaton W, Taylor M, Horne K, Tuffin MI, Burton SG, Cowan DA. 2013. Balancing redox cofactor generation and ATP synthesis: key microaerobic responses in thermophilic fermentations. Biotechnol. Bioeng. 110: 1057-1065

157) Łoś M, Węgrzyn G. 2012. Pseudolysogeny. Adv. Virus Res. 82: 339-349

158) Lowe SE, Zeikus JG. 1992. Purification and characterization of pyruvate decarboxylase from *Sarcina ventriculi*. J. Gen. Microbiol. 138: 803-807

159) Lüdtke S, Neumann P, Erixon KM, Leeper F, Kluger R, Ficner R, Tittmann K. 2013. Sub-ångström-resolution crystallography reveals physical distortions that enhance reactivity of a covalent enzymatic intermediate. Nature Chemistry. 5: 762-767.

160) Luke K, Radek A, Liu X, Campbell J, Uzan M, Haselkorn R, Kogan Y. 2002. Microarray analysis of gene expression during bacteriophage T4 infection. Virology. 299: 182-191

161) Luria SE, Human ML. 1952. A nonhereditary, host-induced variation of bacterial viruses. J. Bacteriol. 64: 557-569

162) Ma J, Campbell A, Karlin S. 2002. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. J. Bacteriol. 184: 5733-5745.

163) Mahony J, Tremblay DM, Labrie SJ, Moineau S, van Sinderen D. 2015. Investigating the requirement for calcium during lactococcal phage infection. Int. J. Food Microbiol. 201: 47-51

164) Mallamace F, Corsaro C, Mallamace D, Vasi S, Vasi C, Baglioni P, Buldyrev SV, Chen S-H, Stanley HE. 2016. Energy landscape in protein folding and unfolding. PNAS 113: 3159-3163

165) Manning AJ, Kuehn MJ. 2011. Contribution of bacterial outer membrane vesicles to innate bacterial defense. BMC Microbiol. 11: 258

166) Marco MB, Moineau S, Quiberoni A. 2012. Bacteriophages and dairy fermentations. Bacteriophage 2: 149-158

167) Marcó MB, Moineau S, Quiberoni A. 2012. Bacteriophages and dairy fermentations. Bacteriophage 3: 149-158

168) Marschall L, Sagmeister P, Herwig C. 2017. Tunable recombinant protein expression in *E. coli*: Promoter systems and genetic constraints. Applied Microbiology and Biotechnology 101: 501-512

169) Maslov S, Sneppen K. 2017. Population cycles and species diversity in dynamic Kill-the-Winner model of microbial ecosystems. Sci. Rep. 7: 39642

170) Mayor U, Guydosh NR, Johnson CM, Grossmann JG, Sato S, Jas GS, Freund SMV, Alonsok DOV, Daggettk V, Fersht AR. 2003. The complete folding pathway of a protein from nanoseconds to microseconds. Nature 421: 863-867

171) MC Jarvis. 2013. Cellulose Biosynthesis: Counting the Chains. Plant Physiol. 163: 1485-1486

172) McGovern PE, Voigt MM, Glusker DL, Exner LJ. 1986. Neolithic resinated wine. Nature 381: 480-481

173) Merrill BD, Ward AT, Grose JH, Hope S. 2016. Software-based analysis of bacteriophage genomes, physical ends, and packaging strategies. BMC Genomics. 17: 679

174) Meyer D, Neumann P, Ficner R, Tittmann K. 2013. Observation of a stable carbine at the active site of a thiamin enzyme. Nat. Chem. Biol. 9: 488-490

175) Meyer D, Neumann P, Parthier C, Friedemann R, Nemeria N, Jordan F, Tittmann K. 2010. Double duty for a conserved glutamate in pyruvate decarboxylase: Evidence of the participation in stereoelectronically controlled decarboxylation and in protonation of the nascent carbanion/ enamine intermediate. Biochemistry 49: 8197-8212

176) Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R; Wilke A, Wilkening J, Edwards RA. 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 9: 386

177) Miller ES, Kutter E, Mosig G, Arisaka F, Kunisawa T, Rüger W. 2003. Bacteriophage T4 genome. Microbiol. Mol. Biol. Rev. 67: 86-156

178) Mir BA, Mewalal R, Mizrachi E, Myburg AA, Cowan DA. 2014. Recombinant hyperthermophilic enzyme expression in plants: A novel approach for lignocellulose digestion. Trends in Biotechnology 32: 281-289

179) Mir BA, Myburg AA, Mizrachi E, Cowan DA. 2017. In planta expression of hyperthermophilic enzymes as a strategy for accelerated lignocellulosic digestion Scientific Reports 7: 11462

180) Mizuno CM, Rodriguez-Valera F, Kimes NE, Ghai R. 2013. Expanding the marine virosphere using metagenomics. PLoS Genet. 9: e1003987

181) Modarres HP, Mofradab MR, Sanati-Nezhad A. 2016. Protein thermostability engineering. RSC Adv. 6: 115252-115270

182) Mohr A, Raman S. 2013. Lessons from first generation biofuels and implications for the sustainability appraisal of second generation biofuels. Energy Policy. 63: 114–122

183) Moineau S, Lévesque C. 2004. Control of bacteriophages in industrial fermentations, *In* Bacteriophages: Biology and Applications. Eds Elizabeth Kutter and Alexander Sulakvelidze, CRC Press, https://doi.org/10.1201/9780203491751.ch10

184) Moineau S, Pandian S, Klaenhammer TR. 1993. Restriction/Modification systems and restriction endonucleases are more effective on lactococcal bacteriophages that have emerged recently in the dairy industry. Appl. Environ. Microbiol. 59: 197-202

185) Molineux IJ, Panja D. 2013. Popping the cork: Mechanisms of phage genome ejection. Nat. Rev. Microbiol. 11: 194-204

186) Monier A, Claverie J-M, Ogata H. 2008. Taxonomic distribution of large DNA viruses in the sea. Genome Biol. 9: R106 doi:10.1186/gb-2008-9-7-r106

187) Mosig, G. 1998. Recombination and recombination-dependent DNA replication in bacteriophage T4. Annu. Rev. Genet. 32: 379-413

188) Moussa SH, Lawler JL, Young R. 2014. Genetic Dissection of T4 Lysis. J. Bacteriol. 196: 2201-2209

189) Muhammed MK, Kot W, Neve H, Mahony J, Castro-Mejía JL, Krych L, Hansen LH, Nielsen DS, Sørensen SJ, Heller KJ, van Sinderen D, Vogensen FK. 2017. Metagenomic analysis of dairy bacteriophages: Extraction method and pilot study on whey samples derived from using undefined and defined mesophilic starter cultures. Appl. Environ. Microbiol. 83: e00888-17

190) Muller YA, Lindqvist Y, Furey W, Schulz GE, Jordan F, Schneider G. 1993. A thiamin diphosphate binding fold revealed by comparison of the crystal structures of transketolase, pyruvate oxidase and pyruvate decarboxylase. Structure 1: 95-103

191) Munsch-Alatossava P, Alatossava T. 2013. The extracellular phage-host interactions involved in the bacteriophage LL-H infection of *Lactobacillus delbrueckii* ssp. *lactis* ATCC 15808. Front. Microbiol. 4: 408

192) Murphy J, Bottacini F, Mahony J, Kelleher P, Neve H, Zomer A, Nauta A, van Sinderen D. 2016. Comparative genomics and functional analysis of the 936 group of lactococcal *Siphoviridae* phages. Sci. Rep. 6: 21345

193) Nagayoshi Y, Kumagae K, Mori K, Tashiro K, Nakamura A, Fujino Y, Hiromasa Y, Iwamoto T, Kuhara S, Ohshima T, Doi K. 2016. Physiological properties and genome structure

of the hyperthermophilic filamentous phage φOH3 which infects *Thermus thermophilus* HB8. Front. Microbiol. 7: 50. doi: 10.3389/fmicb.2016.00050

194) Naik SN, Goud VV, Rout PK, Dalai AK. 2010. Production of first and second-generation biofuels: A comprehensive review. Renew. Sustainable Energy Rev. 14: 578-597

195) Nauton L, Hélaine V, Théry V, Hecquet L. 2016. Insights into the thiamine diphosphate enzyme activation mechanism: Computational model for transketolase using a quantum mechanical/molecular mechanical method. Biochemistry 55: 2144-2152

196) Nazina TN, Tourova TP, Poltaraus AB, Novikova EV, Grigoryan AA, Ivanova AE, Lysenko AM, Petrunyaka VV, Osipov GA, Belyaev SS, Ivanov MV. 2001. Taxonomic study of aerobic thermophilic bacilli: Descriptions of *Geobacillus subterraneus* gen. nov., sp. nov. and *Geobacillus uzenensis* sp. nov. from petroleum reservoirs and transfer of *Bacillus stearothermophilus*, *Bacillus thermocatenulatus*, *Bacillus thermoleovorans*, *Bacillus kaustophilus*, *Bacillus thermoglucosidasius* and *Bacillus thermodenitrificans* to *Geobacillus* as the new combinations *G. stearothermophilus*, G. thermocatenulatus, *G. thermoleovorans*, *G. kaustophilus*, *G. thermoglucosidasius* and *G. thermodenitrificans*. Int. J. Syst. Evol. Microbiol. 51: 433-446

197) Nemeria NS, Chakraborty S, Balakrishnan A, Jordan F. 2009. Reaction mechanisms of thiamin diphosphate enzymes: Defining states of ionization and tautomerization of the cofactor at individual steps. FEBS J. 276: 2432-2446

198) Neubort S, Marmur J. 1973. Synthesis of the unusual DNA of *Bacillus subtilis* bacteriophage SP-15. J. Virol. 12: 1078-1084

199) Nevoigt E, Kohnke J, Fischer CR, Alper H, Stahl U, Stephanopoulos G. 2006. Engineering of promoter replacement cassettes for fine-tuning of gene expression in *Saccharomyces cerevisiae*. Appl Environ Microbiol. 72: 5266-5273

200) Numrych TE, Gumport RI. Gardner JF. 1990. A comparison of the effects of single-base and triple-base changes in the integrase arm-type binding sites on the site-specific recombination of bacteriophage phage Lambda. Nucleic Acids Res. 18: 3953-3959

201) Oberman H, Libudzisz Z. 1998. Fermented milks. Pages 308-350 in Microbiology of Fermented Foods. 2nd ed. B. J. Wood, ed. Blackie Academic & Professional, London, UK.

202) Ohta K, Beall DS, Mejia JP, Shanmugam KT, Ingram LO. 1991a. Genetic improvement of *Escherichia coli* for ethanol production: Chromosomal integration of *Zymomonas mobilis* genes encoding pyruvate decarboxylase and alcohol dehydrogenase. Applied Environmental Microbiology 57: 893-900.

203) Ohta K, Beall DS, Mejia JP, Shanmugam KT, Ingram LO. 1991b. Metabolic engineering of *Klebsiella oxytoca* M5A1 for ethanol production from xylose and glucose. Applied Environmental Microbiology 57: 2810-2815.

204) Oliveira H, Melo LDR, Santos SB, Nóbrega FL, Ferreira EC, Cerca N, Azeredo J, Kluskens LD. 2013. Molecular aspects and comparative genomics of bacteriophage endolysins. J. Virol. 87: 4558-4570

205) Oren, A. 2003. Intracellular Salt Concentrations and Ion Metabolism in Halophilic Microorganisms. *In*: Halophilic Microorganisms and their Environments. Cellular Origin, Life in Extreme Habitats and Astrobiology, vol 5. Springer, Dordrecht

206) Orlova EV. 2012. Bacteriophages and their structural organisation, Bacteriophages, Dr. Ipek Kurtboke (Ed.), InTech, doi: 10.5772/34642

207) Otsuka Y, Yonesaki T. 2012. Dmd of bacteriophage T4 functions as an antitoxin against *Escherichia coli* LsoA and RnlA toxins. Mol. Microbiol. 83: 669-681

208) Pang Z-W, Liang J-J, Qin X-J, Wang J-R, Feng J-X, Huang R-B. 2010. Multiple induced mutagenesis for improvement of ethanol production by *Kluyveromyces marxianus*. Biotechnol. Lett. 32: 1847-1851

209) Panja AS, Bandopadhyay B, Maiti S. 2015. Protein thermostability is owing to their preferences to non-polar smaller volume amino acids, variations in residual physico-chemical properties and more salt-bridges. PLoS ONE 10: e0131495.

210) Park E-J, Kim K-H, Abell GCJ, Kim M-S, Roh SW, Bae J-W. 2011. Metagenomic analysis of the viral communities in fermented foods. Appl. Environ. Microbiol. 77: 1284-1291

211) Parma DH, Snyder M, Sobolevski S, Nawroz M, Brody E, Gold L. 1992. The Rex system of bacteriophage phage Lambda: Tolerance and altruistic cell death. Genes Dev. 6: 497-510

212) Paul JH. 2008. Prophages in marine bacteria: Dangerous molecular time bombs or the key to survival in the seas? ISME J. 2: 579-589

213) Paulikat M, Wechsler C, Tittmann K, Mata RA. 2017. Theoretical studies of the electronic absorption spectra of thiamin diphosphate in pyruvate decarboxylase. Biochemistry 56: 1854-1864

214) Pawlowski A, Rissanen I, Bamford JKH, Krupovic M, Jalasvuori M. 2014. *Gammasphaerolipovirus*, a newly proposed bacteriophage genus, unifies viruses of halophilic archaea and thermophilic bacteria within the novel family *Sphaerolipoviridae*. Arch Virol. 159: 1541-1554

215) Pedersen JF, Funnell DL, Toy JJ, Oliver AL, Grant RJ. 2006. Registration of seven forage sorghum genetic stocks near-isogenic for the brown midrib genes *bmr-6* and *bmr-12*. Crop Sci. 46: 490

216) Pederson CS. 1971. Microbiology of Food Fermentations 2nd Ed. Wesport CT, AVIS, Pp. 153-172.

217) Pei XY, Erixon KM, Luisi BF, Leeper FJ. 2010. Structural insights into the prereaction state of pyruvate decarboxylase from *Zymomonas mobilis*. Biochemistry 49: 1727-1736

218) Pfanner N. 1999. Protein folding: Who chaperones nascent chains in bacteria? Cell 9: R720-R724

219) Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. Nature Rev. 12: 32-42.

220) Pogrebnyakov I, Jendresen CB, Nielsen AT. 2017. Genetic toolbox for controlled expression of functional proteins in *Geobacillus* spp. PLoS ONE 12: e0171313.

221) Pope WH, Haase-Pettingell C, King J. 2004. High-temperature limit on growth of phage P22 in *Salmonella enterica* serovar typhimurium. Applied Environmental Microbiology, 70: 4840-4847.

222) Quax TEF, Claassens NJ, Söll D, van der Oost J. 2015. Codon Bias as a Means to Fine-Tune Gene Expression. Mol Cell. 59: 149-161.

223) Rabelo SC, Carrere H, Filho RM, Costa AC. 2011. Production of bioethanol, methane and heat from sugarcane bagasse in a biorefinery concept. Bioresource Technology 102: 7887-7895

224) Rabinovitch A, Hadas H, Einav M, Melamed Z, Zaritsky A. 1999. Model for Bacteriophage T4 Development in *Escherichia coli*. J. Bacteriol. 181: 1677-1683

225) Rajaure M, Berry J, Kongari R, Cahill J, Young R. 2015. Membrane fusion during phage lysis. 112: 5497-5502

226) Rakhumba DV, Kolomiets EI, Dey ES, Novik GI. 2010. Bacteriophage receptors, mechanisms of phage adsorption and penetration into host cell. Polish J. Microbiol. 59: 145-155

227) Rampelli S, Soverini M, Turroni S, Quercia S, Biagi E, Brigidi P, Candela M. 2016. ViromeScan: a new tool for metagenomic viral community profiling. BMC Genomics. 17: 165

228) Rath D, Amlinger L, Rath A, Lundgren M. 2015. The CRISPR-Cas immune system: Biology, mechanisms and applications. Biochimie. 117: 119-128

229) Rice G, Stedman K, Snyder J, Wiedenheft B, Willits D, Brumfield S, McDermott T, Young MJ. 2001. Viruses from extreme thermal environments. Proc. Natl. Acad. Sci. 98: 13341-13345

230) Ries W, Hotzy C, Schocher I, Sleytr UB, Sára M. 1997. Evidence that the N-terminal part of the s-layer protein from *Bacillus stearothermophilus* PV72/p2 recognizes a secondary cell wall polymer. J. Bacteriol. 179: 3892-3898

231) Rocha–Meneses L, Raud M, Orupõld K, Kikas T. 2017. Second-generation bioethanol production: A review of strategies for waste valorisation. Agronomy Research 15: 830-847

232) Rodrigues EP, Soares CdP, Galvão PG, Imada EL, Simões-Araújo JL, Rouws LFM, Oliveira ALMd, Vidal MS, Baldani JI. 2016. Identification of Genes Involved in Indole-3-Acetic Acid Biosynthesis by Gluconacetobacter diazotrophicus PAL5 Strain Using Transposon Mutagenesis. Front. Microbiol. 7: 1572. doi: 10.3389/fmicb.2016.01572

233) Ross W, Landy A. 1983. Patterns of phage Lambda Int recognition on the regions of strand exchange. Cell 33: 261-272

234) Rothenberg E, Sepu´lveda LA, Skinner SO, Zeng L, Selvin PR, Golding I. 2011. Single-virus tracking reveals a spatial receptor-dependent search mechanism. Biophys. J. 100: 2875-2882

235) Roux S, Enault F, Hurwitz BL, Sullivan MB. 2015. VirSorter: mining viral signal from microbial genomic data. PeerJ. 3: e985

236) Roux S, Tournayre J, Mahul A, Debroas D, Enault F. 2014. Metavir 2: New tools for viral metagenome comparison and assembled virome analysis. BMC Bioinformatics. 15: 76

237) Rubin E. 2008. Genomics of cellulosic biofuels. Nature 454: 841-845

238) Rulli MC, Bellomi D, Cazzoli A, De Carolis G, D'Odorico P. 2016. The water-land-food nexus of first generation biofuels. Sci. Rep. 6: 22521

239) Samson JE, Magadán AH, Sabri M, Moineau S. 2013. Revenge of the phages: Defeating bacterial defences. Nat. Rev. Microbiol. 11: 675-687 doi:10.1038/nrmicro3096

240) Samsygina GA, Boni EG. 1984. Bacteriophages and phage therapy in pediatric practice. Pediatriia 4: 67-70

241) Sao-Jose C, Parreira R, Vieira G, Santos MA. 2000. The N-terminal region of the *Oenococcus oeni* bacteriophage fOg44 lysin behaves as a *bona fide* signal peptide in *Escherichia coli* and as a cis inhibitory element, preventing lytic activity on oenococcal cells. J. Bacteriol. 182: 5823-5831

242) Sára M, Sleytr UB. 2000. S-Layer proteins. J. Bacteriol. 182: 859-868

243) Sarmiento F, Peralta R, Blamey JM. 2015. Cold and Hot Extremozymes: Industrial Relevance and Current Trends. Front Bioeng Biotechnol. 3: 148

244) Schell D, Nguyen Q, Tucker M, Boynton B. 1998. Pretreatment of softwoodby acid-catalyzed steam explosion followed by alkali extraction. Appl. Biochem. Biotechnol. 70-72: 17-24

245) Schmidt TR, Scott II EJ, Dyer DW. 2011. Whole-genome phylogenies of the family Bacillaceae and expansion of the sigma factor gene family in the *Bacillus cereus* species-group. BMC Genomics 12: 430

246) Schoenfeld T, Patterson M, Richardson PM, Wommack KE, Young M, Mead D. 2008. Assembly of viral metagenomes from Yellowstone hot springs. Appl. Environ. Microbiol. 74: 4164-4174

247) Schrö der-Tittmann K, Meyer D, Arens J, Wechsler C, Tietzel M, Golbik R, Tittmann K. 2013. Alternating sites reactivity is a common feature of thiamin diphosphate-dependent enzymes as evidenced by isothermal titration calorimetry studies of substrate binding. Biochemistry 52: 2505-2507

248) Schultes V, Jaenicke R. 1991. Folding intermediates of hyperthermophilic D-glyceraldehyde-3-phosphate dehydrogenase from *Thermotoga maritima* are trapped at low temperature. FEBS Letters. 290: 235-238.

249) Seed KD. 2015. Battling phages: How bacteria defend against viral attack. PLoS Pathog. 11: e1004847

250) Semwal S, Arora AK, Badoni RP, Tuli DK. 2011. Biodiesel production using heterogeneous catalysts. Bioresource Technology. 102: 2151-2161

251) Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M. 2007. CAMERA: A community resource for metagenomics. PLoS Biol. 5: e75

252) Sessions AL, Doughty DM, Welander PV, Summons RE, Newman DK. 2009. The continuing puzzle of the great oxidation event. Current Biology 19: R567-R574

253) Shao Q, Trinh JT, McIntosh CS, Christenson B, Balázsi G, Zeng L. 2016. Lysis-lysogeny coexistence: Prophage integration during lytic development. Microbiol. Open 6: e00395

254) Sharp RJ, Ahmad SI, Munster A, Dowsett B, Atkinson T. 1986. The isolation and characterization of bacteriophages infecting obligately thermophilic strains of *Bacillus*. J. Gen. Microbiol. 132: 1709-1722

255) Siddiqui MA, Fujiwara S, Takagi M, Imanaka T. 1998. *In vitro* heat effect on heterooligomeric subunit assembly of thermostable indolepyruvate ferredoxin oxidoreductase. FEBS Letters 434: 372-376

256) Siegert P, McLeish MJ, Baumann M, Iding H, Kneen MM, Kenyon GL, Pohl M. 2005. Exchanging the substrate specificities of pyruvate decarboxylase from *Zymomonas mobilis* and benzoylformate decarboxylase from *Pseudomonas putida*. Protein Engineering, Design and Selection. 18: 345-357

257) Silhavy TJ, Kahne D, Walker S. 2010. The bacterial cell envelope. Cold Spring Harb. Perspect. Biol. 2: a000414

258) Silva JB, Sauvageau D. 2014. Bacteriophages as antimicrobial agents against bacterial contaminants in yeast fermentation processes. Biotechnol. Biofuels 7: 123

259) Simmonds P, Adams MJ, Benko M, Mya Breitbart, J. Rodney Brister, Eric B. Carstens, Andrew J. Davison, Eric Delwart, Alexander E. Gorbalenya, Balázs Harrach, Roger Hull, Andrew M.Q. King, Eugene V. Koonin, Mart Krupovic, Jens H. Kuhn, Elliot J. Lefkowitz, Max L. Nibert, Richard Orton, Marilyn J. Roossinck, Sead Sabanadzovic, Matthew B. Sullivan, Curtis A. Suttle, Robert B. Tesh, René A. van der Vlugt, Arvind Varsani, Zerbini FM. 2017. Virus taxonomy in the age of metagenomics. Nature Reviews. 15: 161-168

260) Singer GAC, Hickey DA. 2003. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. Gene 317: 39-47

261) Siringan P, Connerton PL, Cummings NJ, Connerton IF. 2014. Alternative bacteriophage life cycles: The carrier state of *Campylobacter jejuni*. Open Biol. 4: 130200

262) Smith DE, Tans SJ, Smith SB, Grimes S, Anderson DL, Bustamante C. 2001. The bacteriophage f29 portal motor can package DNA against a large internal force. Nature 413: 748-752

263) Snyder L, Gold L, Kutter E. 1976. A gene of bacteriophage T4 whose product prevents true late transcription on cytosine-containing T4 DNA. Proc. Natl. Acad. Sci. 73: 3098-3102

264) Sommer P, Georgieva T, Ahring BK. 2004. Potential for using thermophilic anaerobic bacteria for bioethanol production from hemicellulose. Biochem Soc Trans.32: 283-289.

265) Song Q, Ye T, Zhang X. 2011. Proteins responsible for lysogeny of deep-sea thermophilic bacteriophage GVE2 at high temperature. Gene. 479: 1-9

266) Song Q, Zhang X. 2008. Characterization of a novel non-specific nuclease from thermophilic bacteriophage GBSV1. BMC Biotechnol. 8: 43

267) Spencer PS, Barral JM. 2012. Genetic code redundancy and its influence on the encoded polypeptides. Computational and Structural Biotechnology Journal 1: e201204006.

268) St Pierre F, Endy D. 2008. Determination of cell fate selection during phage lambda infection. Proc. Natl. Acad. Sci. 105: 20705-20710

269) Storms ZJ, Sauvageau D. 2015. Modeling tailed bacteriophage adsorption: Insight into mechanisms. Virology 485: 355-362

270) Studholme DJ. 2014. Some (bacilli) like it hot: Genomics of *Geobacillus* species. Microb. Biotechnol. 8: 40-48

271) Sturino JM, Klaenhammer TR. 2004. Antisense RNA targeting primase interferes with bacteriophage replication in *Streptococcus thermophilus*. Appl. Environ. Microbiol. 70: 1735-1743

272) Suárez VB, Capra ML, Rivera M, Reinheimer JA. 2007. Inactivation of calcium-dependent lactic acid bacteria phages by phosphates. J. Food Prot. 70: 1518-1522

273) Sun S, Kondabagil K, Draper B, Alam TI, Bowman VD, Zhang Z, Hegde S, Fokine A, Rossmann MG, Rao VB. 2008. The structure of the phage T4 DNA packaging motor suggests a mechanism dependent on electrostatic forces. Cell 135: 1251-1262

274) Suzuki H, Yoshida K-I, Ohshima T. 2013. Polysaccharide-degrading thermophiles generated by heterologous gene expression in *Geobacillus kaustophilus* HTA426. Appl. Environ. Microbiol. 79: 5151-5158.

275) Svenningsen SL, Semsey S. 2014. Commitment to lysogeny is preceded by a prolonged period of sensitivity to the late lytic regulator Q in bacteriophage. J. Bacteriol. 196: 3582-3588

276) Sykes RW, Gjersing EL, Foutz K, Rottmann WH, Kuhn SA, Foster CE, Ziebell A, Turner GB, Decker SR, Hinchee MAW, Davis MF. 2015. Down-regulation of p-coumaroyl quinate/shikimate 3′-hydroxylase (C3′H) and cinnamate 4-hydroxylase (C4H) genes in the

lignin biosynthetic pathway of *Eucalyptus urophylla* × *E. grandis* leads to improved sugar release. Biotechnol. Biofuels 8: 128

277) Tadege M, Dupuis I, Kuhlemeier C. 1999. Ethanolic fermentation: new functions for an old pathway. Trends in Plant Science Reviews 4: 320-325

278) Tanimura A, Kikukawa M, Yamaguchi S, Kishino S, Ogawa J, Shima J. 2015. Direct ethanol production from starch using a natural isolate, *Scheffersomyces shehatae*: Toward consolidated bioprocessing. Sci Rep. 5: 9593

279) Tauer C, Heinl S, Egger E, Heiss S. Grabherr R. 2014. Tuning constitutive recombinant gene expression in *Lactobacillus plantarum*. Microbial Cell Factories 13:150.

280) Taylor MP, Eley KL, Martin S, Tuffin MI, Burton SG, Cowan DA. 2009. Thermophilic ethanologenesis: Future prospects for second-generation bioethanol production. Trends Biotechnol. 27: 398-405

281) Taylor MP, Esteban CD, Leak DJ. 2008. Development of a versatile shuttle vector for gene expression in *Geobacillus* spp. Plasmid 60: 45-52

282) Teusink B, Molenaar D. 2017. Systems biology of lactic acid bacteria: For food and thought. Curr. Opin. Syst. Biol. 6: 7-13

283) Thomas KC, Ingledew WM. 1992. Production of 21% (v/v) ethanol by fermentation of very high gravity (VHG) wheat mashes. J Ind Microbiol 10: 61-68

284) Thomas LH, Forsyth VT, Šturcová A, Kennedy CJ, May RP, Altaner CM, Apperley DC, Wess TJ, Jarvis MC. 2013. Structure of cellulose microfibrils in primary cell walls from collenchyma[1[C][W][OA]]. Plant Physiol. 161: 465-476

285) Thompson AH, Studholme DJ, Green EM, Leak DJ. 2008. Heterologous expression of pyruvate decarboxylase in *Geobacillus thermoglucosidasius*. Biotechnol Lett. 30: 1359-1365

286) Thompson JF, Moitoso de Vargas L, Koch C, Kahmann R, Landy A. 1987. Cellular factors couple recombination with growth phase: Characterization of a new component in the lambda site-specific recombination pathway. Cell 50: 901-908

287) Thompson TL, Shafia F. 1962. Energy requirement for adsorption of bacteriophage ϕ-4′. Biochemical and Biophysical Research Communications 8: 467-470.

288) Tian J, Yan Y, Yue Q, Liu X, Chu X, Wu N, Fan Y. 2017. Predicting synonymous codon usage and optimizing the heterologous gene for expression in *E. coli*. Sci Rep. 7: 9926

289) Tian L, Perot SJ, Hon S, Zhou J, Liang X, Bouvier JT, Guss AM, Olson DG, Lynd LR. 2017. Enhanced ethanol formation by *Clostridium thermocellum via* pyruvate decarboxylase. Microb Cell Fact. 16: 171.

290) Tittmann K. 2009. Reaction mechanisms of thiamin diphosphate enzymes: Redox reactions. FEBS Journal 276: 2454–2468

291) To KH, Young R. 2014. Probing the structure of the S105 hole. J. Bacteriol. 196: 3683-3689

292) Torsvik T, Dundas ID. 1974. Bacteriophage of *Halobacterium salinarium*. Nature, vol. 248: 680–681

293) Trinh JT, Székely T, Shao Q, Balázsi G, Zeng L. 2017. Cell fate decisions emerge as phages cooperate or compete inside their host. Nature Comm. 8: 14341

294) Turoverov KK, Kuznetsova IM, Uversky VN. 2010. The protein kingdom extended: ordered and intrinsically disordered proteins, their folding, supramolecular complex formation, and aggregation. Prog Biophys Mol Biol. 102: 73–84.

295) Tzul FO, Vasilchuk D, Makhatadze GI. 2017. Evidence for the principle of minimal frustration in the evolution of protein folding landscapes. PNAS | Published online E1627–E1632

296) Updegraff DM. 1969. Semimicro determination of cellulose in biological materials. Anal. Biochem. 3: 420-424

297) Vale PF, Little TJ. 2010. CRISPR-mediated phage resistance and the ghost of coevolution past. Proc. Biol. Sci. 277: 2097-2103

298) Van Dyk JS, Pletschke BI. 2012. A review of lignocellulose bioconversion using enzymatic hydrolysis and synergistic cooperation between enzymes-Factors affecting enzymes, conversion and synergy. Biotechnol. Adv. 30: 1458-1480

299) van Rooyen R, Hahn-Hagerdal B, La Grange DC and Van Zyl WH. 2005. Construction of cellobiose-growing and fermenting *Saccharomyces cerevisiae* strains. J Biotechnol 120: 284-295.

300) van Waarde A. 1991. Alcoholic fermentation in multicellular organisms. Physiological Zoology 64: 895-920

301) van Zyl WH, Lynd LR, den Haan R, McBride JE. 2007. Consolidated bioprocessing for bioethanol production using *Saccharomyces cerevisiae*. Adv Biochem Eng Biotechnol. 108: 205-35

302) Venkatesh Balan. 2014. Current challenges in commercially producing biofuels from lignocellulosic biomass. ISRN Biotechnology. 2014: 463074

303) Ververis C, Georghiou K, Christodoulakis N, Santas P, Santas R. 2004. Fiber dimensions, lignin and cellulose content of various plant materials and their suitability for paper production. Ind. Crops Prod. 19: 245-254

304) Villegas-Silva PA, Toledano-Thompson T, Canto-Canché BB, Larqué-Saavedra A, Barahona-Pérez LF. 2014. Hydrolysis of *Agave fourcroydes* Lemaire (henequen) leaf juice and fermentation with *Kluyveromyces marxianus* for ethanol production. BMC Biotechnol. 14:14

305) Walker SA, Klaenhammer TR. 2000. An explosive antisense RNA strategy for inhibition of a lactococcal bacteriophage. Appl. Environ. Microbiol. 66: 310-319

306) Wang R, Li L, Zhang B, Gao X, Wang D, Hong J. 2013. Improved xylose fermentation of *Kluyveromyces marxianus* at elevated temperature through construction of a xylose isomerase pathway. J Ind Microbiol Biotechnol. 40: 841-854.

307) Wang R, Wang D, Gao X, Hong J. 2014. Direct fermentation of raw starch using a *Kluyveromyces marxianus* strain that expresses glucoamylase and alpha-amylase to produce ethanol. Biotechnol. Prog. 30: 338-347

308) Wang Y, Zhang X. 2008. Identification and characterization of a novel thymidylate synthase from deep-sea thermophilic bacteriophage *Geobacillus* virus E2. Virus Genes 37: 218-224

309) Weiss MC, Sousa FL, Mrnjavac N, Neukirchen S, Roettger M, Nelson-Sathi S, Martin WF. 2016. The physiology and habitat of the last universal common ancestor. Nature Microbiol. 1: 16116

310) Welker NE, Campbell L. 1965. Induction and properties of a temperate bacteriophage from *Bacillus stearothermophilus*. Journal of Bacteriology 89: 175-189

311) Werts C, Michel V, Hofnung M, Charbit A. 1994. Adsorption of bacteriophage lambda on the LamB protein of *Escherichia coli* K-12: Point mutations in gene J of lambda responsible for extended host range. 176: 941-947

312) White R, Chiba S, Pang T, Dewey JS, Savva CG, Holzenburg A, Pogliano K, Young R. 2010. Holin triggering in real time. Proc. Natl. Acad. Sci. 108: 798-803

313) White R, Georgi CE, Militzer WE. 1955. Characteristics of a thermophilic bacteriophage. Proc. Soc. Exp. Biol. Med. 80: 373-377

314) Whitehead HR, Hunter GJE. 1945. Bacteriophage infection in cheese manufacture. J. Dairy Res. 14: 64-80

315) Wilkins MR, Suryawati L, Maness NO, Chrz D. 2007. Ethanol production by *Saccharomyces cerevisiae* and *Kluyveromyces marxianus* in the presence of orange-peel oil. World J. Microbiol. Biotechnol. 23: 1161-1168

316) Williamson SJ, Rusch DB, Yooseph S, Halpern AL, Heidelberg KB, et al. 2008. The Sorcerer II Global Ocean Sampling Expedition: Metagenomic characterization of viruses within aquatic microbial samples. PLOS One 3: e1456

317) Wommack KE, Bhavsar J, Polson SW, Chen J, Dumas M, Srinivasiah S, Furman M, Jamindar S, Nasko DJ. 2012. VIROME: A standard operating procedure for analysis of viral metagenome sequences. Stand. Genomic Sci. 6: 427-439

318) Wu C-C, Naveen V, Chien C-H, Chang Y-W, Hsiao C-D. 2012. Crystal structure of DnaK protein complexed with nucleotide exchange factor GrpE in DnaK chaperone system insight into intermolecular communication. J. Biol. Chem. 287: 21461-21470

319) Wulff DL, Rosenberg M. 1983. Establishment of repressor synthesis. Lambda II. Cold Spring Harbor Laboratory (pp. 53-73). Cold Spring Harbor, New York: Cold Spring Harbor Laboratory

320) Wyman CE. 1999. Biomass ethanol: Technical progress, opportunities and commercial challenges. Ann. Rev. Ener. Environ. 24: 189-226

321) Wynn RM, Davie JR, Cox RP, Chuang DT. 1992. Chaperonins GroEL and GroES promote assembly of heterotetramers ($\alpha_2\beta_2$) of mammalian mitochondrial branched-chain a-keto acid decarboxylase in *Escherichia coli*. J Biol. Chem. 267: 12400-12403

322) Xu J, Xiang Y. 2017. Membrane penetration by bacterial viruses. J. Virol. 91: e00162-17

323) Yang Li, Hao Wang, Kai Nie, Chen Zhang, Yi Zhang, Ji Wang, Peihua Niu & Xuejun Ma. 2016. VIP: an integrated pipeline for metagenomics of virus identification and discovery. Sci. Rep. 6: 23774

324) Yang S, Fei Q, Zhang Y, Contreras LM, Utturkar SM, Brown SD, Himmel ME, Zhang M. 2016. *Zymomonas mobilis* as a model system for production of biofuels and biochemicals. Microb Biotechnol. 9: 699–717.

325) Yep A, McLeish MJ. 2009. Engineering the Substrate Binding Site of Benzoylformate Decarboxylase. Biochemistry 48: 8387-8395

326) Yona AH, Bloom-Ackermann Z, Frumkin I, Hanson-Smith V, Charpak-Amikam Y, Feng Q, Boeke JD, Dahan O, Pilpel Y. 2013. tRNA genes rapidly change in evolution to meet novel translational demands. eLife. 2: e01339.

327) Yu C-Y, Jiang B-H, Duan K-J. 2013. Production of bioethanol from carrot pomace using the thermotolerant yeast *Kluyveromyces marxianus*. Energies 6: 1794-1801

328) Yuan WJ, Zhao XQ, Ge XM, Bai FW. 2008. Ethanol fermentation with *Kluyveromyces marxianus* from Jerusalem artichoke grown in salina and irrigated with a mixture of seawater and freshwater. J. Appl. Microbiol. 105: 2076-2083

329) Zablocki O, Van Zyl L, Adriaenssens EM, Rubagotti E, Tuffin M, Cary SC, Cowan D. 2014. High-level diversity of tailed phages, Eukaryote-associated viruses, and virophage-like elements in the metaviromes of Antarctic soils. Appl. Environ. Microbiol. 80: 6888-6897

330) Zeng L, Skinner SO, Zong C, Sippy J, Feiss M, Golding I. 2010. Decision making at a subcellular level determines the outcome of bacteriophage infection. Cell 141: 682-691

331) Zhang L, Xu D, Huang Y, Zhu X, Rui M, Wan T, Zheng X, Shen Y, Chen X, Ma K, Gong Y. 2017. Structural and functional characterization of deep-sea thermophilic bacteriophage GVE2 HNH endonuclease. Sci. Rep. 7:42542 doi:10.1038/srep42542

332) Zhang N, Pan X-M, Ge M. 2012. Without salt, the 'thermophilic' protein mth10b is just mesophilic. PLoS ONE 7: e53125.

333) Zhang S, Liu M, Yan Y, Zhang Z, Jordan F. 2004. C2-alpha-lactylthiamin diphosphate is an intermediate on the pathway of thiamin diphosphate dependent pyruvate decarboxylation. Evidence on enzymes and models. J. Biol. Chem. 279: 54312–54318

334) Zhong Z, Hou Q, Kwok L, Yu Z, Zheng Y, Sun Z, Menghe B, Zhang H. 2016. Bacterial microbiota compositions of naturally fermented milk are shaped by both geographic origin and sample type. J. Dairy Sci. 99: 1-10

335) Zhou H, Cheng J-S, Wang BL, Fink GR, Stephanopoulos G. 2012. Xylose isomerase overexpression along with engineering of the pentose phosphate pathway and evolutionary engineering enable rapid xylose utilization and ethanol production by *Saccharomyces cerevisiae*. Metabolic Engineering 14: 611-622

336) Ziegler DR. 2013. The *Geobacillus* paradox: why is a thermophilic bacterial genus so prevalent on a mesophilic planet? Microbiology. 160: 1-11 doi: 10.1099/mic.0.071696-0

337) Zong C, So L-H, Sepúlveda LA, Skinner SO, Golding I. 2010. Lysogen stability is determined by the frequency of activity bursts from the fate-determining gene. Mol. Syst. Biol. 6: 440