Computational and experimental analysis of horizontally-acquired, secreted enzymes for plant carbohydrate degradation - from hemibiotrophic oomycete, *Phytophthora sojae*

Submitted by

Victoria Attah

to the **University of Exeter** as a thesis for the degree of **Doctor of Philosophy** in **Biological Sciences - September 2020**

I certify that all material in this thesis which is not my own work has been identified and that any material that has previously been submitted and approved for the award of a degree by this or any other University has been acknowledged.

Signature ……………………………………………………………………………………

Abstract

The oomycetes are heterotrophic, filamentous protists that share morphological similarities with some fungi, but according to phylogenetic evidence, are actually relatives of diatoms and brown algae, within a different eukaryotic supergroup (Stramenopiles) (Förster et al., 1990; Leclerc et al., 2000; Hudspeth et al., 2000; Hudspeth et al., 2003; Thines et al., 2007; McCarthy and Fitzpatrick., 2017). Horizontal Gene Transfer (HGT) from fungi to oomycetes has been previously suggested to have played a role in the evolution of osmotrophic and phytopathogenic traits during divergence of the oomycete lineage; such transfers include secreted proteins predicted to degrade plant cell wall-specific substrates – these are the initial barriers to pathogen invasion as well as an abundant source of fixed carbon (Torto et al., 2002; Belbahri et al., 2008; Richards et al., 2011; Savory et al., 2015).

The HGTs are largely expanded in hemibiotrophic oomycetes, and their selective benefit is suggested by subsequent gene duplications following acquisition (Richards et al., 2011; Savory et al., 2015) - giving rise to paralogs hypothesised to possess functional differences. Bioinformatics and computational analysis of laterally-transferred Glycoside Hydrolase 10 (GH10) and GH12 in *Phytophthora sojae* (oomycete hemibiotrophic parasite of soybean) discovered unique differences amongst paralogs – *P. sojae*_482953 (GH12 xyloglucanase) encodes a significantly disordered, 186 amino acid 'tail', which improves the enzymatic activity of the protein towards xyloglucan when heterologously-expressed in *Saccharomyces cerevisiae*. Interestingly, knockout

of the gene encoding *P. sojae*_482953 *in vivo* did not affect the ability of *P. sojae* to utilise xyloglucan as a sole carbon source, indicating that gene duplication events are also an important mechanism for transcriptional fidelity and maintenance of function. A second GH12 paralog, *P. sojae*_559651 was predicted to encode a ‘second’ carbohydrate-binding site, with the prediction being conserved for orthologs encoded by *P. cactorum* and *P. nicotiniae*; interestingly, two indels (coding for alanine and serine) were identified as being important for the prediction, and their removal abolished the second binding site prediction for all three proteins. Functional diversification of variant paralogs post-HGT was also suggested by differences in oligosaccharides released from xyloglucan breakdown by *P. sojae*_559651 and *P. sojae*_482953 proteins using mass spectrometry. Taken together, this work extends our understanding of the functional significance of gene duplication post-HGT in hemibiotrophic oomycetes.

Of further interest is how N-terminal signal peptide sequences affect the transferability of secreted proteins by cross-phylum HGT – this work has optimised a functional agar plate screen (and demonstrated a promising micro-droplet approach) in order to explore signal peptide evolution in future work.

Acknowledgements

“It is our choices that show what we truly are, far more than our abilities”

-Albus Dumbledore.

Thank you to Professor Tom Richards for your ideas, support, and for helping me to put the pieces together during my PhD. Thank you to both Professor Richards and the late Professor Ken Haynes, for giving me the opportunity back in 2014 to be part of your research groups, and for considering me for a PhD project two years later. Thank you to everyone that has let me borrow equipment, reagents, or taken the time to troubleshoot experiments with me (notably, Dr David Milner (and thanks to Dr Guy Leonard for help with the bioinformatics)). To everyone I’ve met at the University of Exeter – thank you for your friendship, motivation and trips to the pub! Thank you to my Dad for always wanting the best for me, to my sister for always speaking her mind, and to my mum – luckily you won’t have to read this huge thing, but I hope you look on proudly. I did it. Thank you to the friends I’ve made over the last 6 years living in Devon, and thank you to my friends from home – for all of the support and laughs that kept me pushing through to the end.

Table of contents

Chapter 5: Development of methods to investigate the mutational landscape from non-functional to functional signal peptide sequences for a secreted protein 'acquired' by *S. cerevisiae*

List of tables

Page

List of figures

Page

List of abbreviations

3’ Three prime

3D Three-dimensional

5’ Five prime

Å Angstrom

AAA α-aminoadipate

AADS Absorbance-activated droplet sorting

ampR ampicillin resistance

Avr Avirulence

BLAST Basic Local Alignment Search Tool

BLASTp Protein Basic Local Alignment Search Tool

Bp Base pair

BUSCO Benchmarking Universal Single-Copy Orthologs

CAZyme Carbohydrate-Active enZyme

CBM Carbohydrate-binding module

CE Carbohydrate Esterase

Cis Latin **-** on this side

CMC Carboxymethylcellulose

CRISPR Clustered Regularly Interspaced Short Palindromic Repeats

crRNAs CRISPR RNAs

DAP Diaminopimelate

db Decibel

DDC Duplication-degeneration-complementation

DHB 2,5-Dihydroxybenzoic acid

DNA Deoxyribonucleic acid

dNTP Deoxynucleoside triphosphate

DNS 3,5-Dinitrosalicylic acid

dsDNA double-stranded DNA

DTT Dithiothreitol

$_{d}H_2O$ Distilled water

EAC Escape from Adaptive Conflict

EAM Energy-Absorbing Matrix

EC Experimentally characterised (CAZy reference)

EDTA Ethylenediaminetetraacetic acid

e.g. Latin: *exempli gratia* – for example

ER Endoplasmic reticulum

EST Expressed sequence tag

EXPASY EXpert Protein Analysis SYstem

et al. Latin: *et alii*, *et aliae*, *et alia* – and others

FADS Fluorescence-Activated Droplet Sorting

FB Fastidious bacteria

FPKM Fragments per kilobase of exon model

G418 Geneticin

g Grams

g/L Grams per litre

GH Glycoside Hydrolase

GT Glycosyl Transferase

Glu Glutamic acid, E

GPD Glyceraldehyde-3-phosphate dehydrogenase

gRNA guide RNA

GTP Guanosine triphosphate

i.e. Latin: *id est* – that is

H_2O Water

HCl Hydrochloric acid

HDR Homology directed repair

HGT Horizontal Gene Transfer

HMM Hidden Markov Model

HRP Horseradish Peroxidase

IAD Innovation, Amplification, Divergence

in vivo Latin – within the living

kHz Kilohertz

kV Kilo volt

L Litre

LB Luria broth

LED Light Emitting Diode

LGT Lateral Gene Transfer

LiAc Lithium acetate

Ma Million years ago

MALDI Matrix Assisted Laser Desorption/Ionisation

Mb Megabase

MDN Mutation During Non-functionality

MeCN Acetonitrile

MFA Mating factor alpha

mg Milligram

mg/mL Milligram per millilitre

MIC Minimum Inhibitory Concentration

mL Millilitre

MS Mass spectrometry

mV Millivolts

m/z Mass-to-charge ratio

NHEJ Non-homologous end joining

NLR Nucleotide binding site and leucine-rich repeat

nm Nanometer

NMR Nuclear Magnetic Resonance

NUMPTs NUclear MiTochondrial DNA

NaCl Sodium Chloride

OD Optical density

OD$_{600}$ Optical density at 600nm

OLIMP OLIgosaccharide Mass Profiling

ORF Open Reading Frame

PAM Protospacer Adjacent Motif

PBS Phosphate Buffered Saline

PCR Polymerase chain reaction

PDMS polydimethylsiloxane

PEG Polyethylene glycol

Pfam Protein families' database

pH Acidity or alkalinity of a solution

PheDH Phenylalanine dehydrogenase

Phyre2 The protein homology/analogY recognition engine V 2.0

pmol picomole

PL Polysaccharide lyase

qPCR

rDNA Ribosomal DNA

RNA Ribonucleic acid

Rps Major resistance gene

SAR Stramenopiles-Alveolata-Rhizaria

SCM Synthetic complete medium

SDS Sodium dodecyl sulphate

SDS-PAGE SDS- polyacrylamide gel electrophoresis

SGD *Saccharomyces* genome database

sgRNA single guide RNA

SP Signal peptide

spp. Species

SRP Signal Recognition Particle

TAE Tris-acetate-EDTA

TE Transposable element

TOF Time of Flight

TRIS Trisaminomethane

tRNA Transfer RNA

UV Ultraviolet

WT wild-type

w/v Weight to volume ration

YPD Yeast peptone dextrose

U units

µF Microfarad

µg Micrograms

µg/mL Micrograms per millilitre

µL Microliter

µM Micromolar

µM s^{-1} Micromolar per second

µs Microsecond

V volts

α Alpha

β Beta

Ω Ohm

% Percentage

°C Degrees Celsius

~ Approximately

Chapter 1

Introduction

1.1 What are the oomycetes?

The oomycetes are a diverse group of heterotrophic, filamentous protists that include plant, fungal and animal parasites, as well as saprotrophs that acquire nutrients from decaying matter (Beakes et al., 2012). They share biological and morphological similarities with some fungi (such as filamentous growth and an absorptive mode of osmotrophic nutrition), and as a result, were once classified in the same kingdom (Ainsworth., 1961). Later, analysis of conserved DNA and protein sequences (including the nuclear-encoded small-subunit ribosomal DNA (rDNA) (Förster et al., 1990), large-subunit rDNA (Leclerc et al., 2000), and the mitochondrial-encoded COX2 (Thines et al., 2007; Hudspeth et al., 2000; Hudspeth et al., 2003)), as well as analysis of expressed sequence tag (EST) data (McCarthy and Fitzpatrick., 2017), all provided strong phylogenetic evidence that the oomycetes branch separately to the “true” fungi. With accepted divergent ancestry, it was later believed that oomycetes and fungi independently evolved similar biological and morphological traits, and so some of their similarities may be due to convergent evolution (Latijnhouwers et al., 2003; Money et al., 2004). Now oomycetes are classified within the diverse stramenopiles lineage (also known as heterokonts), within the Stramenopiles-Alveolata-Rhizaria (SAR) eukaryotic supergroup (Riisberg et al., 2009).

Oomycetes are characterised by bi-flagellated motile zoospores - with one of the flagella, the anterior “tinsel” flagellum, being a characteristic of the stramenopiles/heterokonts (Cavalier-Smith., 1986). The stramenopiles once belonged to the polyphyletic Chromista kingdom (protists possessing plastids (the photosynthetic organelles) that contain chlorophyll *c*) (Cavalier-Smith., 1981), but when the kingdom was later refined, they were assigned to the Chromalveolata supergroup (including organisms that are proposed to be descended from a single secondary endosymbiosis involving a red alga and a bikont) (Cavalier-Smith., 1999; Keeling., 2009). However, contrary to the early analyses (e.g. Yoon et al., 2002), monophyly for the Chromalveolata has since been rejected (Harper., 2005; Rice and Palmer., 2006; Janouskovec et al., 2010; Gaston and Roger., 2013) – but multiple phylogenetic analyses currently support monophyly of the SAR supergroup (Moreira et al., 2007; Hackett et al., 2007; Gaston and Roger., 2013). When the draft genomes of two oomycetes (*Phytophthora* spp.) were published (Tyler et al., 2006), it was suggested that the stramenopiles descended from a photosynthetic ancestor (based on identification of encoded genes of ‘algal’ origin in the oomycete genomes) (Tyler et al., 2006). However, the oomycetes are aplastidic and, in contrast to the earlier chromalveolate hypothesis (proposing a common ancestor that engulfed a red alga, followed by multiple secondary plastid losses within descendant heterotrophic lineages (Cavalier-Smith., 1999)), there is currently insufficient evidence for relic plastid genes within available oomycete genomes (Stiller et al., 2009; Wang et al., 2017).

Oomycetes branch sister to the hyphochytriomycetes (also stramenopiles), which likewise share similarities with fungi, but unlike oomycetes, some have undergone secondary loss of their posterior flagellum (Cooney et al., 1985). Together, the oomycetes and the hyphochytriomycetes are classified in the phylum Pseudofungi (Cavalier-Smith., 2006); phylogenies of nuclear-encoded genes show that the Pseudofungi form a sister group to photosynthetic stramenopiles, including diatoms and ‘golden-brown’ algae (Moreira and Lopez-Garcia., 2002; Cavalier-Smith and Chao., 2006; Derelle et al., 2016). Although, this is conflicted by phylogenies using mitochondrial-encoded genes, which suggest the Pseudofungi form a sister group to other heterotrophic stramenopiles (Secq et al., 2006; Tsui et al., 2009; Leonard et al., 2018). Taken together, the oomycetes photosynthetic ancestry remains ambiguous.

The earliest known evidence of oomycete (or ‘oomycete-like’) structures in the fossil record dates back to the Devonian period (between 400 and 360 million years ago (Ma)) (Krings et al., 2011) – although molecular clock estimates suggest oomycetes could date back further to the Silurian period (the shortest of the Paleozoic Era; between 440 and 410 Ma) (Matari and Blair., 2014). The Silurian period was characterised by a significant evolutionary transition of multicellular life from marine to terrestrial environments - after which came huge diversification of terrestrial life into the Carboniferous period (between 360 and 300 Ma), and a major shift in atmospheric oxygen levels (up to 35%). Strullu-Derrien et al. (2011) provide evidence of oomycete plant parasitism during this time - identifying fossilised *Combresomyces williamsonii* within tissues of the now extinct seed fern, *Lyginopteris oldhamia* (Strullu-Derrien et al., 2011). It is estimated that the two major oomycete lineages (the saprolegnialean and peronosporalean branches) diverged in the early Mesozoic Era between 225 and 190 Ma, which was ~200 Ma after the first appearance of oomycetes in the fossil record (Matari and Blair., 2014). Current phylogenetic evidence demonstrates the presence of some marine species within the predominantly terrestrial oomycete branches (e.g. *Pythium porphyrae*, which infects red seaweeds, and *Myzocytiopsis vermicola*, which infects marine nematodes (Newell et al., 1977)). Therefore, it has been proposed that the oomycetes could have transitioned from marine environments to the land via parasitism of nematodes in terrestrial soil, or through colonisation of early coastal plants (Beakes et al., 2012). However, it should be noted that the two deeply-diverging peronosporalean clades, the *Rhipidiales* and the *Albuginales*, are (according to current sampling) exclusively terrestrial – therefore the origins of the 'higher' marine oomycete species is as yet not fully understood (Beakes et al., 2012).

1.2 Oomycetes and Fungi

Although traditionally included in the Kingdom Fungi, molecular phylogenetics has demonstrated a divergent and paraphyletic evolutionary relationship between the oomycetes and fungi (Förster et al., 1990; Leclerc et al., 2000; Thines et al., 2007; Hudspeth et al., 2000; Hudspeth et al., 2003; Baldauf and Palmer., 1993). Morphologically, the oomycetes appear similar to fungi; for example, both grow filamentously during vegetative lifecycle phases, and the vegetative bodies of both fungi and oomycetes are composed of mycelium (the complex networks of hyphae that are used for dispersal, reproduction, and nutrient acquisition (Richards et al., 2006)). Being heterotrophic organisms, most fungi and oomycetes feed exclusively by an osmotrophic mode of nutrition (Richards et al., 2017); this involves the secretion of depolymerising enzymes outside of the cell, with the function of breaking down complex molecules into smaller units (e.g. monomers that include simple sugars, amino acids, and fatty acids), which are subsequently transported into the cell to derive energy for important cellular processes. Interestingly, *Rozella*, the as-yet only culturable member of the deeply-diverging fungal clade, *Cryptomycota* (Jones et al., 2011), feeds osmotrophically but lacks the hallmark characteristic fungal cell wall during its trophic phase (conceptually referred to as 'no jacket required') – a phenomenon thought to be due to convergence and adaptation to life as an intracellular parasite (James and Berbee., 2011). Whilst earlier described as 'intermediates' between fungi and ancestral protists (Jones et al., 2011), it is possible that that the cryptomycota are a deeply-diverging 'enigmatic' fungal lineage (capable of producing chitin-rich cell walls during immature spore growth (James and Berbee., 2011)) – blurring our understanding of the early evolution of fungal characteristics. Nevertheless, both fungi and oomycetes obtain nutrients by absorption from external environments (or by invading the body of another organism), and they have in common some membrane transporter families for nutrient uptake (e.g. Chothia et al., 2003; Savory et al., 2018).

Despite sharing similar morphological features, there are also important differences between the oomycetes and fungi. Early biochemical studies suggested that oomycete cell walls are more similar to those of *Vaucheria* (yellow-green alga) than to fungi (Parker et al., 1963), and it has been demonstrated that oomycete cell walls are predominantly composed of cellulose, with some β-1,3 and β-1,6-linked glucose polymers, and small amounts of chitin (<1%) (Bartnicki-Garcia., 1966; 1968 and 1969). Conversely, chitin is the major structural component of most fungal cell walls. Interestingly, oomycetes do express chitin synthases during hyphal tip morphogenesis (Levesque et al., 2010; Guerrieo et al., 2010), suggesting that the chitin polysaccharide plays a role during oomycete-host (or environment) interactions. Unlike fungi, which grow as haploid vegetative hyphae (although some exceptions do exist) (Levesque et al., 2010; Emerson., 1941), the oomycetes' vegetative phase is diploid, and the cells only become haploid nuclei transiently in the gametangia during sexual reproduction. There is no dikaryotic phase in the oomycete lifecycle, as gametic nuclei fuse to form a diploid zygote and oomycete sexual spores (known as oospores) are formed on terminal hyphae - each containing one viable zygotic nucleus. Oomycete asexual spores (known as zoospores) are biflagellated (with the anterior tinsellated flagellum being a defining characteristic of the heterokonts (Cavalier-Smith., 1986)) - this is unlike many members of the Kingdom Fungi, which lack flagella (with an exception being the 'chytrids', which produce singular-flagellated gametes (zoospores)) (Van der Auwera et al., 1995). Other major morphological distinctions between the oomycetes and fungi include the structure of their inner mitochondrial membranes, as the oomycetes possess tubular cristae (like animals), whereas fungi possess flattened cristae (like many other eukaryotic groups) (Taylor., 1978), and the structure of their hyphae (where oomycete hyphae are always non-septate/aseptate, i.e. not divided into separate cells, as in most fungi (Latijnhouwers et al., 2003)).

Differences in biochemical pathways have also been identified between the oomycetes and fungi (Vogel., 1960 & 1961; LeJon., 1971), including in enzymes involved in the lysine biosynthetic pathway, which, in oomycetes, is more similar to higher plants than to fungi. It should be noted that the two described lysine biosynthesis pathways (the diaminopimelate (DAP) pathway (present in bacteria, plants and algae (Velasco et al., 2002; Hudson et al., 2005)), and the α-aminoadipate (AAA) pathway (present in fungi (Miyazaki et al., 2004)), have complex evolutionary histories (likely due in part to horizontal/lateral gene transfers) – for example, a core DAP gene (*lysA*) is encoded by multiple opisthokonts, and a core AAA gene (*AAR*) is encoded by a marine protist (Sumathi et al., 2006; Torruella et al., 2009).

1.3 Current division of the oomycetes

The oomycetes are composed of multiple ecologically-destructive species (Beakes et al., 2012). Although many species remain unsampled, based on current data the oomycetes can be divided into two broad subclasses – the deeply diverging, Saprolegniomycetidae (also known as the “water molds”, all known members of this group are aquatic in habitat (with the exception of nematode-infecting species (Hakariya et al., 2007))), and the Peronosporomycetidae. The Saprolegniomycetidae include the orders *Eurychasmales*, *Leptomitales*, and *Saprolegniales*; the *Saprolegniales* include ecologically-destructive fish pathogens of both marine and freshwater environments – diseases caused by *Saprolegnia parasitica*, for example, contribute to major losses in aquaculture (van West., 2006)). Also included in the saprolegnian branches are animal and plant pathogens of the *Aphanomyces* genus (Jiang and Tyler., 2012; Kamoun., 2003). Within the genus *Albugo*, which forms part of the *Albuginales* order, are obligate biotrophic microbes that cause “white blister rusts” in crops of the Brassicaceae family (these are flowering plants also known as the mustards, crucifers, or cabbages) (Kemen et al., 2011; Links et al., 2011). The Peronosporomycetidae include the orders *Rhipidiales*, *Pythiales*, and *Peronosporales* (with the *Pythiales* and *Peronosporales* being mainly associated with terrestrial environments (Jiang and Tyler., 2012)). The *Peronosporales* include some of the best-characterised oomycetes to date - the diversity in oomycete microbial forms includes obligate and non-obligate parasites of the order *Pythiales*, including plant pathogens of the *Pythium* genus (containing necrotrophic pathogens that cause disease by killing host plant cells), and of the *Phytophthora* genus (containing hemibiotrophic pathogens that live biotrophically with their plant hosts before switching to a necrotrophic phase). The majority of the *Peronosporales* are phytopathogens (with the exception of some *Pythium* species, which infect animals and some fungi (Gaastra et al., 2010; Benhamou et al., 2012)).

Plant pathogenicity is thought to have evolved independently in multiple oomycete lineages (Thines and Kamoun., 2010), and plant host ranges vary widely; for example, *Pythium* spp. (the causative agents of root rot and 'damping off' of seedlings) can have a broad host range, (such as *P. ultimum*, which infects soybean, corn, and wheat), or can show some host preferences (such as *P. arrhenomanes*) (Adhikari et al., 2013). Interestingly, differences in temperature 'preferences' amongst *Pythium* spp. have also been observed, such as increased virulence of *P. ultimum* and *P. aphinidermatum* associated with higher temperatures, whilst *P. iwayamai and P. irregulare* seem to be more virulent at lower temperatures (Adhikari et al., 2013). The order *Peronosporales* includes phytopathogens of the genus *Phytophthora* (from the Greek meaning 'plant destroyer'), which is the most studied oomycete genus to date, and so far contains over 100 known species affecting a range of agriculturally- and ecologically-important plants). Again, host ranges vary within the genus - some important diseases caused by *Phytophthora* spp. include late blight of potato (*P. infestans* (causal agent of the Irish potato famine in the 1840s)) (Van West and Vleeshouwers., 2004), sudden oak death (*P. ramorum*) (Tyler., 2006), and root rot of soybean (*Glycine max*) (*P. sojae*) (Tyler., 2007). International transportations of plants has spread many species of *Phytophthora* away from natural areas of origin, and invasive species have been known to cause devastation to natural ecosystems (Hansen., 2008). In particular, the hemibiotroph, *P. sojae* is among the most important species of phytopathogenic oomycetes of economic importance (Kamoun et al., 2015), with around 20.5 million metric ton losses of crops due to root and stem rot since 1996 (Dorrance., 2018).

1.4 The oomycete life cycle

Oomycetes predominantly grow as branched aseptate hyphae, about 4-8 µM in diameter, collectively known as mycelia. The hyphae differentiate into sporangia, which are vessels containing motile spores, or zoospores; each sporangium can release 20-30 single-celled zoospores, which are the oomycetes' asexual means of dispersal. Spores possess dual flagella (one anterior 'whiplash' flagella, and one posterior ciliated flagella) – both thought to contribute to tighter control of motility. In the case of oomycetes that form close associations with host plants, it is thought that zoospores can 'swim' through soil or water films for hours (at 125-150 µm s^{-1}) (Ho and Hickman., 1967), until chemotactically attracted to host cells through recognition of secreted, specific host molecules (isoflavones released by the roots, in the case of *P. sojae*) (Morris & Ward., 1992; Morris et al., 1998). Once at the surface of a host cell, the flagella retract and zoospores attach to the surface, form a cell wall, and encyst. Germination of cysts leads to filamentous growth (i.e. extensions of hyphae), that penetrates the host cell. Hyphae extend throughout host cells, whilst digestive enzymes are secreted to break down and utilise host-derived substrates for growth (e.g. carbohydrates of the plant cell wall that provide a source of fixed carbon). Oomycetes have evolved sophisticated systems for interacting with their plant hosts, including the release of various classes of proteins that promote virulence - of particular interest to this work is how oomycete systems have evolved for breaking down and feeding on plant-specific substrates. Other important systems include those that trigger the oomycete sexual cycle, for example - when host substrates have been digested and nutrient sources have been depleted, oomycete cells undergo sexual reproduction - initiated when meiosis occurs in terminal hyphal cells, differentiating into female 'oogonium' and male 'antheridium' haploid gametes (coincident with stage-specific gene expression (Fabritius et al., 2002; Lamour and Kamoun., 2009). The cells mate via a fertilisation tube, giving rise to the zygote, known as an oocyte. It has been observed that oospores (derived from oocytes) have a thicker cell wall and are significantly more resistant to desiccation than zoospores (Lamour and Kamoun., 2009), suggesting that sexual reproduction in oomycetes plays an important role in ensuring survival - by allowing the organism to remain dormant during times when nutrients are limited.

1.5 What are hemibiotrophs?

Phytopathogens utilise a wide variety of infection strategies to penetrate host cells and avoid immune recognition. Biotrophs, for example, must maintain a close relationship with host plants without disrupting cells - instead extracting essential nutrients from the host cell environment, whilst withstanding (or avoiding triggering) the host immune response. Biotrophs take up nutrients from living cells through structures called haustoria, which are differentiated hyphae that penetrate the plant cell wall, but do not enter the cytoplasm (Mendgen and Hahn., 2002; Latijnhouwers et al., 2003; de Writ., 2007; Kabbage et al., 2015) –

for example, the obligate biotrophic fungus, *Blumeria gramminis* (powdery mildew of barley), extracts glucose from host cells and uses it to synthesise glycogen (Both et al., 2015). Interestingly, biotrophs lack some genes involved in the biosynthesis of metabolites, such as the obligate biotrophic oomycete, *Hyaloperonospora arabidopsidis* and basal-branching oomycete, *Albugo laibachii*, which lack some genes involved in sulphur and nitrogen synthesis pathways (Meadows., 2011). For example, *A. laibachii* has lost an entire molybdopterin biosynthesis pathway – molybdopterin is an essential cofactor for sulphite oxidase and nitrate reductase function, whilst *A. laibachii* and *H. arabidopsidis* have both lost a molybdenum-dependent nitrate reductase, involved in the reduction of nitrate to nitrite (Kemen et al., 2011).

Generally, many biotrophs have also lost genes associated with pathogenesis, including reduced numbers of plant cell wall-degrading enzymes and genes required for toxin production (Kabbage et al., 2015), suggesting the genomes of biotrophs reflect their lifestyle, i.e. maintenance of long-term interactions with hosts, as well as the requirement to be largely ‘undetected’ as a pathogen (Micali et al., 2008; Kemen and Jones., 2012). On the other hand, hemibiotrophs maintain only an initial biotrophic existence with their hosts, causing very little damage and suppress host cell death during this phase. Most hemibiotrophs also produce haustoria, although some may develop intracellular hyphae, to facilitate early spreading throughout intact plant tissues, asymptomatically (Oliver and Ipcho., 2004; Divon and Fluhr., 2006). However, the biotrophic lifestyle later switches to necrotrophy, where necrotrophic microbes kill host cells and extract nutrients from dead tissues; necrotrophy is characterised by the development of secondary extracellular hyphae that rapidly spread throughout tissues, and the breakdown of host cell walls through increased secretion of plant cell wall-degrading enzymes - facilitating both nutrient uptake and parasitic colonisation (Latijnhouwers et al., 2003; Oliver and Ipcho., 2004; Kabbage et al., 2015). There is some ambiguity around the term ‘hemibiotroph’ in the literature (specifically, whether biotrophic and necrotrophic phases are distinct or overlapping); studies by Dufresne et al. (2000), involving random mutagenesis of the hemibiotrophic fungus, *Colletotrichum lindemuthianum*, identified a GAL4-like transcriptional activator, named CLTA1, which was shown to play a role in reprogramming host cell metabolism and proposed to be a gene involved in facilitating the hemibiotrophic ‘switch’ in lifestyle from biotrophy to necrotrophy (Dufresne et al., 2000; Oliver and Ipcho., 2004; Krola et al., 2015). Conversely, studies involving the oomycete *P. infestans* suggest that an effector, SNE1 (expressed during biotrophic growth), acts as an antagonist for other necrosis-inducing effectors, regulating the change from biotrophy to necrotrophy (Kelley et al., 2010). Despite their evolutionary divergence from one another, due to some shared biological similarities between the oomycetes and fungi, it is interesting to compare their systems for interacting with plants, and more specifically to this work - to better understand how osmotrophic phytopathogenic oomycetes evolve to feed on their hosts.

Currently, over 20 ‘complete’ oomycete genomes spanning a range of lifestyles are publically available for analysis. Of the hemibiotrophic genus, *Phytophthora*, these include *P. sojae* (95 Mb; Tyler et al., 2006), *P. ramorum* (65 Mb; Tyler et al., 2006), *P. infestans* (240 Mb; Haas et al., 2009), *P. capsici* (Lamour et al., 2012), and *P. lateralis* (44 Mb; Quinn et al., 2013); of the necrotrophic *Pythium* genus, these include *P. ultimum var. ultimum* (Levesque et al., 2010), *P. vexans*, *P. ultimum var. sporangiiferum*, *P. aphinidermatum*, *P. arregenomanes*, *P. irregulare* and *P. iwayamai* (Adhikari et al., 2013), and of the obligate biotrophic oomycetes, these include *A. laibachii* (Kemen et al., 2011), *H. arabidopsis* (Baxter et al., 2010), and *A. candida* (Links et al., 2011). It is thought that around 8000-9500 'core' genes are conserved among oomycete species (Haas et al., 2009; Seidl et al., 2012), therefore, up to half of some genomes are variable – largely due to gene loss, horizontal gene transfers, and/or gene family expansion, resulting in novel genome partitioning into regions of low or high gene density. Size expansion of *Phytophthora* spp. genomes is thought to be caused by repetitive DNA sequences (Tyler et al., 2006; Haas et al., 2009) - some sequences in gene-sparse, repeat-rich regions have been hypothesised to be under higher selection pressures (and therefore evolve 'faster') - such dynamic regions are also associated with higher numbers of pathogenicity-related genes (Haas et al., 2009; Raffaele et al., 2010). Furthermore, Dong et al (2015) discuss the concept of 'two-speed genomes' of filamentous pathogens, where the bipartite structure (gene-rich vs gene-sparse regions) is hypothesised to foster a higher opportunity for adaptive evolution (in the gene-sparse regions) (Dong et al., 2015).

Leonard et al. (2018) have recently published the draft genome of and demonstrated that many gene families associated with plant parasitism in the oomycetes appear to be absent in free-living sister *Hyphochytrium catenoides* (Richards et al., 2011; Leonard et al., 2018) - consistent with independent evolution of these traits within the lineage (e.g. Thines and Kamoun., 2010). It has been proposed that multiple Horizontal Gene Transfers (HGTs) from fungi have contributed to the evolution of plant parasitic traits in the oomycetes (Torto et al., 2002; Richards et al., 2006; Belbahri et al., 2008; Richards et al., 2011; Misner et al., 2015; Savory et al., 2015).

1.6 What is Horizontal Gene Transfer (HGT)?

Organisms evolve or adapt to new environments by natural selection, traditionally involving spontaneous genetic mutations that are transferred vertically from parents to progeny (via sexual or asexual reproduction). In this case, a genome's evolutionary history would be expected to reflect that of the organism's lineage. Consequently, acquisition of novel traits is limited by the constraints of the existing genetic variation, as well as the rate at which new mutations are introduced into a lineage. So specific traits may be restricted to lineages that encode the necessary genetic precursors, as well as requiring some evolutionary time for 'meaningful' mutations to accumulate and undergo selection – allowing novel, adaptive traits to arise. However, analysis of genomes reveal they are dynamic in size and content, reflecting the genes that have been lost, gained, as well as differences in evolution rates (due to genes having differing selective pressures). Genes can be acquired through duplication events or from foreign transfers from 'donor' organisms. The transfer of genes among groups of organisms with divergent evolutionary histories can enable organisms to hypothetically 'bypass' some of the constraints associated with vertical evolution, and lead to the acquisition of novel genetic material for natural selection to act on - and also has the evolutionary advantage of increasing the overall genetic diversity of species pan-genomes (the total set of genes across all of the strains of a single species). The reshaping of a species genome (or genotype), and the resulting phenotype, can therefore allow recipients of gene transfers to explore novel ways of functioning in otherwise unexplored/diverse sequence space (E.g. Soanes and Richards., 2014).

Horizontal gene transfer (HGT), also known as Lateral Gene Transfer (LGT), involves the transmission of genetic material between unrelated species, in the absence of sexual reproduction – in other words, gene evolution which appears to contradict the known or accepted species phylogeny. In principle, this could occur between any two organisms that encode DNA, and the recipient of a HGT may have a very different evolutionary history to that of the donor (Doolittle., 1999; Syvanen., 1985). In 1960, Akiba et al. demonstrated transfer of drug-resistance phenotypes between *Escherichia coli* and *Shigella* species, suggesting the possibility of HGT occurrence between bacteria in the human intestinal environment (Akiba et al., 1960); subsequently, mechanisms of prokaryote exchange of genetic material have been well described, including transformation, transduction and conjugation (Thomas and Nielsen., 2005; Zhaxybayeva and Doolittle., 2011). It is generally accepted that HGT is common and frequent between prokaryotes, and that these recurrent events of gene gains (and losses) contribute to speciation, genomic plasticity and environmental adaptation (Ochman et al., 2000; Boucher et al., 2003; Jain et al., 2003). A large proportion of genes in some prokaryote genomes (up to 80%) are thought to have been transferred laterally (intradomain HGT (i.e. prokaryote to prokaryote)) during their evolution (Dagan et al., 2008). HGT is thought to be so ubiquitous in prokaryotes, that it is often the main source of ambiguous species phylogenies (which has been suggested to limit our current inference and understanding of organism evolutionary relationships for bacteria (and archaea)) (Doolittle et al., 2016).

Although once thought to be infrequent, HGT into the eukaryotes has also become increasingly recognised as an important evolutionary mechanism driving adaptation (e.g. Hirt et al., 2015; Soanes and Richards., 2014) – although its study has, perhaps in part, been previously overshadowed by the study of HGT amongst bacteria, for which there are considerable sequenced genomes available for analysis. Over evolutionary time, bacteria have been associated with eukaryotes (for example, as parasites or symbionts), but have also influenced eukaryote biology through transfers of novel DNA. However, despite increasing evidence, there has been a problematic history of prokaryote-eukaryote HGT - due to past misattributions of HGT into the human genome (Salzberg et al., 2001; Salzberg et al., 2017), and into a tardigrade genome (Arakawa., 2016). This reflects the importance of robust analysis for the hypothesis of prokaryote-eukaryote HGT (and consideration of microbial contamination in eukaryote genome projects by, for example, looking at hypothesised HGT gene linkage to vertically-derived genes). Nevertheless, the recent increase in genomic datasets has made it possible to extensively screen more eukaryote genomes for evidence of horizontally-transferred genes, and has demonstrated that HGT (once considered a rare occurrence in eukaryotes), has also been common in these organisms.

HGT from bacteria to eukaryotes has been described previously, most significantly being endosymbiotic gene transfers from organellar genomes to the nuclear genome (Margulis., 1970; Martin et al., 2001) - improving our understanding of the origin and evolution of eukaryotic genomes (Keeling and Palmer., 2008). Horizontally-acquired genes have been found in both unicellular and multicellular eukaryote genomes (Andersson., 2009; Richards et al., 2009), HGT of metabolic genes have been identified in parasitic microbial eukaryotes (Hirt et al., 2015; Alsmark et al., 2013), and pathogenicity genes (for example, those encoding secreted cell wall-degrading enzymes) of soil-borne bacterial origin, have been found in the genome of plant-parasitic nematode, *Meloidogyne incognita* (Danchin et al., 2010). Marcet and Gabaldon. (2010) suggest that more than 700 genes of prokaryote origin are present in fungi, mostly in a subdivision of the *Ascomycota* (the *Pezizomycotina*) (Marcet and Gabaldon., 2010) - suggesting bacteria have been an important source of genetic material throughout eukaryote evolution. HGT of ice-binding proteins to the green alga, *Pyraminmonas gelidicola* (Raymond et al., 2012), and to the diatom, *Fragilariopsis cylindrus* (Mock et al., 2017), also provide good examples of gene transfer events from bacteria that have enabled eukaryotes to colonise variant environments.

Interestingly, eukaryotes have also acquired genes from bacteria that function to target other bacteria, such as a bacterial lysozyme family with antibacterial function (Metcalf et al., 2014) – an evolutionary phenomenon described as ‘the eukaryotes strike back’ (Dunning et al., 2014). Other interesting examples of bacterial defence-related genes acquired by eukaryotes and re-purposed for eukaryote (fungal) attack have also been described, such as secreted proteins in the oomycete, *Aphanomyces euteiches* (AeCRN13) and the fungus, *Batrachochytrium dendrobatidis* (BdCRN13), which were shown to induce DNA damage in host nuclei (Ramirez-Garces et al., 2016). In the oomycetes, bacterial metabolic genes involved in amino acid metabolism have been identified in *Phytophthora* spp. (Whitaker et al., 2009), in addition to secreted cutin-degrading enzymes that are proposed to have Actinobacterial origin (Belbahri et al., 2008). HGT of a secreted endoglucanase from bacteria has also been identified in basal Saprolegniales, *Achlya hypogyna* and *Traustotheca clavata* (Misner et al., 2015). Additionally, McCarthy and Fitzpatrick (2016) demonstrate five well-supported examples of HGT into the oomycetes from bacterial species, including metabolic genes, a secreted protein, and enzymes involved in degrading xenobiotics (McCarthy and Fitzpatrick., 2016). Fungal-oomycete HGTs have also been described. Of significant interest to this work are the secreted proteins acquired by oomycetes with a strong functional trend towards degradation of plant cell wall-specific substrates (Torto et al., 2002; Richards et al., 2006; Belbahri et al., 2008; Richards et al., 2011; Misner et al., 2015; Savory et al., 2015). Laterally-transferred oomycete transporter proteins with proposed fungal origin and involved in metabolic uptake have also been described and characterised previously (Richards et al., 2011; Savory et al., 2015, 2018).

HGT can lead to the acquisition of new genes, paralogs of existing genes (known as ‘xenologues’ (Fitch., 1970)), or the displacement of existing genes by laterally-transferred orthologs. As such, HGT events can be broadly split into two categories, ‘maintenance’ transfers and ‘innovative’ transfers (Husnik and McCutcheon., 2018). The definition of maintenance transfers are those that maintain (or replace) a pre-existing function (such as those involving functions encoded by endosymbionts (e.g. Husnik and McCutcheon., 2016)), whilst innovative transfers are those in which the HGT encodes a new function to the recipient (a novel innovation), providing an ability to colonise additional environments (e.g. new nutritional sources, protection from other organisms, or acquisition of genes that may facilitate parasitism (Andersson et al., 2005; Keeling and Palmer., 2008)).

Horizontally-transferred genes have been identified between species of the same genus (Silva and Kidwell., 2000), of different genus (Won and Renner., 2003), and between organisms of different domains (with, for example, some interesting insights into dynamic viral-eukaryote associations (Monier et al., 2017). Among plants, HGT events have been identified involving mitochondrial genes (Bergthorsson et al., 2003; Won and Renner., 2003; Rice et al., 2013), nuclear genes (Li et al., 2014; Mahelka et al., 2017), and genes involved in plant-parasitic interactions (Davis and Wurdack., 2004; Mower et al., 2004). Recently, Dunning et al. (2019) identified fifty-nine HGT events (from multiple donors) into the grass *Alloteropsis semialata*, and demonstrated (by RNA-seq analysis) that the acquisitions added functional diversity to the recipient genome (Dunning et al., 2019). Richards et al. (2009) used a comprehensive bioinformatics approach to identify five HGTs from fungi into plants, and four HGTs from plants into fungi (including, for example, a putative two-domain zinc binding alcohol dehydrogenase (Richards et al., 2009)). Eukaryote acquisitions of putative virulence-associated genes (e.g. from one species of fungal pathogen to another (Friesen et al., 2006; Gardiner et al., 2012)) also demonstrates the impact of HGT to biological adaptation during eukaryotic evolution (and demonstrate that eukaryote features, such as the presence of introns and the nuclear envelope, are not barriers to HGT as once thought).

Whilst a large proportion of published HGT events involve the transfer of single genes between species, there are also examples of transfers of

Transposable Elements (TEs) (which can contain multiple genes). TEs are widespread amongst most organisms (Craig and Craigie., 2002) and are naturally mobile sequences of DNA that can move within a genome to different positions – sometimes introducing sequence mutations, and often resulting in duplication of their DNA at other locations within a genome, generating large numbers of copies. TEs belong to two classes; Class I (retrotransposons), which are reverse-transcribed by an RNA intermediate and integrated into the genome as a duplicated copy, and Class II (transposons), which are excised and directly integrated back into another part of the genome. TEs, therefore, play important roles in modifying genetic variation and genome size, and because they exist as ‘free molecules’ during transposition, they may be more amenable for HGT (e.g. through close contact with vectors (e.g. viruses) that could mobilise them (Piskurek and Okada., 2007)). Horizontal TE transfer has been described for the *P-element* from *Drosophila willistoni*, ‘laterally’ transferred to *D. melanogaster* and then to other species of *Drosophila* (Daniels et al., 1990; Haring et al., 2000). In plants, a few cases of TE HGT have been described (e.g. Diao et al., 2006; Fortune et al., 2008), expanding our understanding of eukaryote genome evolution.

1.7 Mechanisms of HGT

For eukaryotes, transfer mechanisms of HGT depend on contact between the donor and the recipient organisms (or the frequency of donor DNA in the environment), the ability of a recipient to ‘take up’ and integrate foreign DNA into its genome (i.e. recombination rates, as well as the genomic location of integration), and the type of selection (e.g. positive, purifying, neutral) imposed on the newly-acquired DNA (i.e. the rates of transcription and translation, that would then impact the functionality of the transferred gene) (Baltrus., 2013). Interestingly, for phagotrophic eukaryotes, the ‘you are what you eat’ hypothesis has been proposed to play a role in increasing the *opportunity* of HGT in predator-prey relationships (Doolittle., 1998) - consistent with studies that have identified prokaryote genes in phagotrophic eukaryote genomes (e.g. Andersson et al., 2003; Archibald et al., 2003). However, evidence of HGT involving non-phagotrophic fungi and oomycetes (and plants (Richards et al., 2009)), suggests that overlapping ecologies or shared associations may also play an important role in the frequency of HGT. In addition to proposed donor to recipient HGT transfers, Richards et al. (2009) also demonstrate prokaryote to eukaryote gene transfer followed by a eukaryote to eukaryote (plant to fungi, or fungi to oomycetes) transfer – referred to as a ‘two-step serial gene transfer (Richards et al., 2009; Richards et al., 2011).

Principally, foreign DNA could be integrated anywhere into a genome where it does not disrupt an already existing functional genomic element (only if that functional genomic element is essential). Previous studies have shown that transferred genes are generally integrated into dynamic (gene-sparse) genomic regions, often rich in transposons and retrotransposons, with the major mechanism of integration in eukaryotes proposed to be non-homologous end joining (NHEJ) (Leiber., 2010). As previously described, it is thought that these dynamic, gene-sparse, repeat-rich regions of pathogen genomes are under higher selection pressures and consistently, are associated with a greater number of pathogenicity-related genes (Haas et al., 2009; Raffaele et al., 2010).

Another important factor in acquisition of foreign DNA that eventually gains functionality in the recipient could be the size of the acquired genetic material - where shorter DNA sequences may be more amenable for integration and expression in recipient genomes. For example, analyses of entire *Wolbachia* spp. genomes that have been transferred to arthropod and nematode genomes, show many transfers of DNA fragments resulting in high levels of 'junk' DNA (i.e. pseudogene formation and eventual DNA deletion), as well as low expression levels (Kumar et al., 2012; Choi et al., 2015). This is also true of fragments of mitochondrial DNA in eukaryote nuclear genomes (called NUMTs, or NUclear MiTochondrial DNA) (Lopez et al., 1994). Interestingly, on the other hand, a large ~1.5 Mb DNA fragment transferred from *Wolbachia* spp. to *Armadillidium vulgare*, has been important in evolving a new female sex chromosome (successively duplicated to 3 Mb) (Leclercq et al., 2016). Increased expression of integrated, transferred DNA in eukaryotes can result from intron gain (Le Hir et al., 2003), and/or subsequent tandem duplications. The latter is of significant interest to this work – particularly how the increase in paralog number (resulting in multi-gene families arising from single HGT events) influences both transcriptional dosage and function (i.e. where some duplicates may be functional or non-functional, and some may gain new or additional functions not encoded by the original HGT). Functions of duplicated HGT gene families has not been extensively studied previously, although clearly this phenomenon is evident (e.g. Richards et al., 2009; Richards et al., 2011; Savory et al., 2015). Of the fungal-oomycete HGTs relevant to this work (with putative functions in degradation of plant cell wall carbohydrates), most have undergone subsequent gene duplication events – demonstrated by identification of paralogous proteins in recipient genomes (Richards et al., 2011., Savory et al., 2015). In this case, it is interesting to consider how the acquisitions by HGT have been under positive selection to evolve the paralogs in recipient oomycetes, in order to digest host-specific substrates with greater efficiency.

1.8 Barriers of HGT

Principally, there are physical and biological barriers to HGT (Thomas and Nielsen., 2005). The maintenance (and fixation) of a newly acquired gene in a recipient genome is limited by both the fitness advantage of the expressed gene, and the fitness cost that is associated with expressing that gene – and so there must exist a balance between both factors. Mechanistic costs can include, for example, differences in **codon bias**[1] between the donor and recipient organisms, or how complex the resulting protein-protein interactions are by expression of the gene (Francino., 2012). To expand on the latter, Jain et al. (1999) propose 'the complexity hypothesis', in which more complex genes (i.e. 'informational' genes, such as those encoding functions related to transcription and translation) have lower frequencies of HGT, when compared to 'operational' genes (encoding proteins that interact with fewer gene products), which have higher observed rates of HGT (Jain et al., 1999). Differences in codon bias between the donor and recipient genomes is also a key factor in the efficiency of horizontally-transferred gene expression. For example, using a heterologous expression approach, Amoros et al. (2010) expressed three synonymous versions of the chloramphenicol acetyl transferase (cat) gene in *E. coli*, and used this experiment to demonstrate that the gene with the highest positive impact to chloramphenicol

[1] **Codon bias:** the *preferred* (or *optimised*) usage of codons over alternate synonymous codons amongst different types of genes in an organism's genome (Behura and Severson., 2012).

resistance in *E. coli* was the one which mirrored the bacterium's own preferred codon frequency (Amoros et al., 2010). It is assumed that rare (or mismatched) codon usage slows down protein translation by introducing incorrect amino acids into the nascent polypeptide, which then leads to subsequent misfolding of proteins (Sorensen et al., 1989; Buchan., 2006; Hershberg and Petrov., 2008; Qian et al., 2012). Misfolded, non-functional protein accumulated inside a cell would have negative impacts to fitness, and force higher energy consumption to activate the required protein degradation pathways to clear the protein (Drummond and Wilke., 2008). The normal translation of other cellular proteins may also be affected, through tRNA competition between multiple transcripts (Frumkin et al., 2018). It has been suggested that tRNA copy number and codon usage preferences co-evolved to optimise translation and protein folding (Sharp et al., 2010; Gingold and Pilpel., 2011). It has been proposed that HGT genes with mismatched codon usage can be maintained in recipient genomes by the accumulation of synonymous mutations that drive codon bias in the favour of the recipients codon preferences (Lawrence and Ochman., 1997, 1998; Garcia-Vallve., 2003) - alternatively, synonymous mutations that increase the overall efficiency of transcription and translation may also prove to be beneficial for the acquisition of new genes. As such, HGT genes that are maintained and provide functionality to the recipient will eventually follow the same fate as other evolved eukaryotic genes – i.e. DNA sequence features will be ameliorated to match that of the host genome.

Experimental evolution studies have also provided interesting insights into recipient adaption to the acquisition of orthologous genes (by gene displacement); for example, Lind et al. (2010) 'transferred' orthologous ribosomal genes into *Salmonella typhimurium*, and demonstrated an initial fitness cost to the organism for their expression (demonstrated by low levels of protein translation), however, experimental evolution resulted in increased expression of the transferred genes, as well as increased translation quantity (via gene duplications or mutations in promoter sequences) (Lind et al., 2010). Similarly, Bershtein et al. (2015) 'transferred' an orthologous dihydrofolate reductase gene into *E. coli* with comparable consequences (Bershtein et al., 2015).

1.9 Identification of HGTs

HGT can be inferred by detecting conflicting gene ancestries; this can be done explicitly by reconstructing the gene phylogeny (sampling a diversity of taxa that allows for comparison of gene phylogenetic relationships with established species phylogenetic relationships). HGT can be hypothesised if the gene of one species appears to share ancestry with genes from distantly related species (Keeling and Palmer., 2008; Andersson., 2009), i.e. if the two genes share the most recent ancestral node, with strong statistical support for the HGT gene branching both **with** and **within** the donor group (Figure 1) (Richards et al., 2011; Soanes and Richards., 2014). Alternative topology tests (e.g. approximately unbiased (AU) test based on Efron et al. (1996) (Shimodaira., 2002)) can be used to detect the robustness of established phylogenetic trees (suggestive of inconsistent evolutionary histories), and bootstrap tests (Felsenstein., 1985) can be used to evaluate the statistical support for nodes in a phylogenetic tree, to assess confidence in the clade suggestive of HGT (i.e. what percentage of bootstrap trees show the same clade). Doolittle and Bapteste (2007) describe the concept of pattern pluralism, i.e. different evolutionary models will be appropriate for different taxa (Doolittle and Bapteste., 2007); model-based phylogenetic methods can test multiple evolutionary scenarios resulting in the conflicting gene phylogeny, such as ancestral gene duplication and gene losses (hidden paralogy), representing an alternative hypothesis to HGT; rejection of alternative evolutionary scenarios can be based on parsimonious criteria (e.g. if the total number of gene duplications/losses required to explain the resulting tree topology is high and overly complex) (Soanes and Richards., 2014). It is also important to exclude DNA contamination from putative HGT genome projects (for example, resulting from DNA extraction methods), which can be done by looking at the linkage of the hypothesised transfers to vertically-derived genes (Richards et al., 2009). Taken together, HGT can be suggested if it is the most consistent hypothesis for the gene phylogeny/taxonomic distribution. Phylogenetic-based methods for HGT identification have the advantage of putatively detecting the direction of transfer, as well as donor and recipient species, allowing further study within the context of recipient evolution. As mentioned previously, comparative phylogenetic analyses have identified multiple cases of HGT into the oomycete lineage (Torto et al., 2002; Richards et al., 2006; Belbahri et al., 2008; Richards et al., 2011; Misner et al., 2014; Savory et al., 2015).

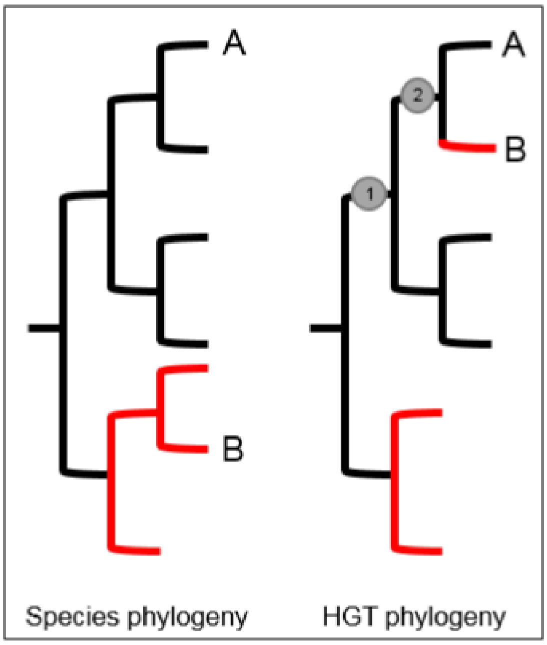


Figure 1.1: HGT can be hypothesised if the gene of one species appears to share ancestry with distantly related species, i.e. the red gene (B) branching with the black gene (A), which conflicts the known or accepted species phylogeny. Statistical support must be strong for the HGT gene branching both **with** and **within** the donor group (nodes numbered 1 and 2, respectively), and alternative evolutionary scenarios to explain the conflicting gene phylogeny (e.g. gene duplication and gene loss) must be rejected based on adequate genome sampling and comparisons of parsimonious criteria.

HGT can enable recipients to better adapt to a novel lifestyle or an ecological niche, leading to improved fitness (Hirt et al., 2015; Soanes and Richards., 2014). Therefore, further support for HGT may include consideration of the acquired function. By coupling HGT identification with knowledge of function, we can better understand the selective benefit of the transfer - informative for understanding how HGT genes are maintained (or fixed) within recipient genomes. To date, there have been few functional studies to confirm the roles of horizontally-transferred genes in eukaryotes (Friesen et al., 2006; Hall and Dietrich., 2007; Gardiner et al., 2012; Savory et al., 2018), although it is possible to infer putative protein functions based on conserved sequence identity in comparison with characterised genes (e.g. Pfam (Finn et al., 2016) and Interpro (Finn et al., 2017) database searches). However, it is important to note that the reliability of functional annotation may be limited if characterised proteins are distantly related, or are paralogous. High sequence similarity also might not (confidently) account for differences in broad, variable, or 'promiscuous' protein functions. 'Enzyme promiscuity' for example, describes an enzyme that can catalyse reactions outside of its normal biological function (i.e. those that are under selection), including degradation of different types of substrates using the same active site.

As previously mentioned, one of the mechanisms that can shape the resulting gene function and transcriptional dosage is gene duplication (Long et al., 2003). Gene expansion by duplication events can be a source of phenotypic diversity, by generating variant protein-coding genes (neofunctionalization (Ohno., 1970), by maintaining mutational robustness, or by partitioning ancestral functions (subfunctionalization (Stoltzfus., 1999; Force et al., 1999). Diversity and expansion in genes associated with phytopathogenicity likely reflect their importance; although phylogenetic identification of gene duplications have improved (e.g. by performing Hidden Markov Model (HMM) searches against genome databases (Finn et al., 2015), using raw profile HMM training sets for putative protein families (Pfam: Finn et al., 2016)), subtle functional differences between paralogs may be difficult to elucidate in the absence of experimental analysis (e.g. Zallot et al., 2016), limiting our understanding of their wider impact to the evolution of plant pathogenic oomycetes.

1.10 Oomycete secreted proteins

Although the oomycetes appear to be descended from a phagotrophic (and possibly photosynthetic) ancestor (Cavalier-Smith., 2006), they have lost phagotrophic capability (i.e. feeding by engulfing prey), and feed exclusively by osmotrophy. As mentioned previously, this involves secretion of depolymerising enzymes into the environment to break down complex biological molecules into monomers, which are transported back into the cell (Richards and Talbot., 2013). Interestingly, the majority of identified HGT events into the oomycete lineage encode putative secreted proteins (Torto et al., 2002; Belbahri et al., 2008; Richards et al., 2011; Misner et al., 2015; Savory et al., 2015). These secreted proteins are hypothesised to play important roles in digesting plant cell wall carbohydrates (enabling entry to plant tissues and for scavenging nutrients), and also in mediating oomycete interactions with their plant hosts, as well as in environmental social interactions (through generation and acquisition of ‘public goods’ in diverse microbial communities) (West et al., 2007; Kamoun., 2009; Richards and Talbot., 2013).

The soluble secreted molecules of an organism can be collectively referred to as a secretome, and can be determined computationally for eukaryotes by identifying proteins with an N-terminal secretion signal peptide (SignalP: Bendtsen et al., 2004), and those predicted to have extracellular localisation (WoLFPSORT: Horton et al., 2007). The eukaryotic secretion pathway (involving protein co-translational translocation to the Endoplasmic reticulum (ER), and vesicular transport to the Golgi (Lodish et al., 2000)) is initiated by recognition of a short N-terminal peptide sequence (16-30 amino acids), and is therefore a good indicator for identifying putative secreted proteins amongst proteomic datasets. Sperschneider et al. (2015) compared differences in secretion prediction sensitivities across experimentally-validated fungal and oomycete effectors, and suggested that previous versions of SignalP (v3, v2) may be the most sensitive in detecting oomycete signal peptides (Sperschneider et al., 2015). However, most bioinformatics tools are generally trained using secreted proteins from model species (for eukaryotes, the yeast *Saccharomyces cerevisiae* and the plant, *Arabidopsidis thaliana*), so it is important to be cautious in the absence of experimental evidence of secretion.

1.11 The *P. sojae* secretome

Parasitic microbes have evolved specialised strategies to gain entry into host cells, and induce host cell susceptibility - one such mechanism being the secretion of proteins to degrade host cell components, or otherwise interact with/disrupt host innate immunity. The *P. sojae* putative secretome is predicted to include 1464-1756 proteins (Tyler et al., 2006; Richards et al., 2011; Adhikari et al., 2013; McGowan and Fitzpatrick., 2017). As with other hemibiotrophic phytopathogens, *Phytophthora* spp*.* secrete diverse proteins to interact with plant hosts or manipulate host defences (plant defences that may otherwise stall successful infection by parasites) (Halder et al., 2006; Kamoun., 2006). Some of these are so-called ‘effector’ proteins, which selectively bind to and alter the activity of host proteins; they are usually small proteins, and they can either act within the host apoplast (apoplastic effectors) or the host cytoplasm (cytoplasmic effectors) (Kamoun., 2006). Effector proteins have been studied in great detail amongst phytopathogens, in addition to their counterpart receptor proteins in plant host species. This has particularly been significant for disease management in crops – for example, *P. sojae* avirulence (Avr) genes (typically identified by N-terminal ‘RxLR’ and ‘DEER’ amino acid motif sequences) are known to be recognised by host soybean proteins (major resistance (Rps) genes) that have nucleotide binding site and leucine-rich repeat (NLR) receptors (e.g. Dong et al., 2011). This recognition induces a plant defence response, therefore the upregulation of Rps in soybean cultivars has been a strategy used to increase host resistance to infection (e.g. Sahoo et al., 2017). Interestingly, the apicomplexans (chromalveolate obligate parasites) have evolved a similar ‘RxLR’ motif (Bhattacharjee at al., 2006; Whisson et al., 2007), although it is unclear if it is a result of convergent evolution or inherited evolution from a common ancestor in this line. Mutations in *Avr* genes, or changes in expression demonstrate that these genes evolve rapidly, resulting in an evolutionary ‘arms race’ between pathogens and hosts, as per the Red Queen Hypothesis (Dybdahl and Storfer., 2003). This is further suggested by the genomic locations of Avr genes – again found in dynamic areas containing duplications and repetitive sequences prone to chromosomal rearrangements (Tyler and Gijzen., 2014).

Recently, Rodenburg et al. (2019) reconstructed an integrated metabolic model of *P. infestans* and *Solanum lycopersicum* (tomato) to simulate metabolic fluxes from host to pathogen, highlighting the complexity of interactions between them both. It is important to understand the evolutionary origins of the secreted proteins involved in pathogenicity (and the selective pressures they are under), as well as how novel genes and/or protein functions could be evolved from existing genetic material (i.e. the significance of paralog evolution for pathogenicity-related functions) (Rodenburg et al., 2019). Understanding these functions and targets of the full arsenal of a pathogens secreted proteins might also impact the study of plant resistance – with implications on reducing crop loss due to biotic stressors (e.g. due to root and stem rot caused by *P. sojae* (Dorrance., 2018).

1.12 Summary

Previously identified HGT events from fungi are thought to have played a role in the evolution of osmotrophic and plant-parasitic traits in the oomycetes. Many of the HGTs encode secreted proteins with putative functions involved in plant cell wall digestion (Torto et al., 2002; Belbahri et al., 2008; Richards et al., 2011; Savory et al., 2015) – facilitating entry into plant tissues, as well as providing a source of fixed carbon. Many of the transferred protein families have undergone subsequent gene duplication events, and it is currently unclear whether the paralogs possess functional differences (e.g. Ohno., 1970; Stoltzfus., 1999; Force et al., 1999).

1.13 Aims of study

This work aims to characterise the functional significance of HGT followed by gene duplication in the oomycetes, by investigating multi-paralog enzyme families in the genome of *P. sojae* - a widespread hemibiotrophic pathogen of soybean (and one of the current model species for the *Phytophthora* genus). Functional investigation of paralogs in a GH12 enzyme family, putatively involved in cellulose degradation (endo-1,4-β-glucanase (EC 3.2.1.4), xyloglucan endo-hydrolase (EC 3.2.1.151) and endo-1,3-1,4-β-glucanase (EC 3.2.1.73) activities), and a GH10 enzyme family, putatively involved in xylan (hemicellulose) degradation (endo-1,4- β-xylanase (EC 3.2.1.8), endo-1,3- β-xylanase (EC 3.2.1.32) and xylan endotransglycosylase (EC 2.4.2.-) activities), will provide functional evidence of the phenotypic fate (post-acquisition) of horizontally-transferred, expanded enzyme families in this organism.

It is hypothesised that whilst HGT of plant cell wall-degrading enzymes into the oomycetes facilitated novel gain of function, successive gene expansion of HGT families has given rise to further enhancement of function - by allowing i) transcriptional amplification, ii) transcriptional fidelity, and iii) subfunctionalization or neofunctionalization - and contributed to adaptation to an osmotrophic and plant-parasitic lifestyle. *P. sojae* is an economically-significant pathogen of soybean, so it is likely to benefit from the expression and secretion of multiple paralogs for the breakdown of specific plant carbohydrates. It is also possible that selection following subsequent gene duplication events has widened the functions of some of the paralogs from the functions originally acquired by HGT - including a gain in novel substrate specificity or improved activity under unique biological conditions (e.g. temperature and pH).

It is conceivable that subtle differences in protein sequences and three-dimensional protein structures between the paralogs could result in differences in enzyme activity or in novel ways to interact with carbohydrate substrates (or plant hosts). Therefore in this work, bioinformatics and computational methods will be used to characterise sequence and structural differences between *P. sojae* GH12 and GH10 paralogs. This will be followed by experimental investigation of paralogous proteins - using enzyme assays under different pH and temperatures, and mass spectrometry to compare differences in products released from carbohydrate breakdown. Gene-deletion tools will also be used to investigate carbohydrate utilisation by *P. sojae in vivo*. By better understanding any functional differences between the paralogs of HGT, we can gain a better insight into the dynamics of phytopathogenicity, as well as a better understanding of gene flow in phytopathogenic oomycetes.

In addition to improving the understanding of the functions of the enzyme paralogs, this work will also seek to experimentally investigate N-terminal signal peptide sequences as a potential barrier to HGT. This will be achieved by generating large libraries of signal peptide sequences and using high-throughput enzyme assays as a proxy to identify the positive sequences for secretion - enabling the construction of an N-terminal signal peptide *mutational landscape* (from non-functional to functional), improving our understanding of the dynamics of integration of horizontally-acquired secreted proteins.

The yeast *S. cerevisiae* will be used as a model for heterologous protein expression and secretion analysis, using a synthetic gene synthesis and molecular cloning approach. *S. cerevisiae* is a widely used eukaryotic model organism (Karathia et al., 2011); it is non-pathogenic, simple to culture, and has a generation time of ~90 minutes. The *S. cerevisiae* BY4742 (derivative of S288C) genome sequence (Winston et al., 1995; Brachmann et al., 1998) is freely available from the Saccharomyces Genome Database (SGD), and a number of genetic tools have been established for this organism. Whilst the bacterium, *E. coli* is a widely used as a heterologous host for producing many recombinant proteins, it was not preferable for this study due to its inability to perform post-translational protein modifications associated with eukaryotic proteins (e.g. glycosylation). *S. cerevisiae* was chosen over other eukaryotic hosts due to its ease of culture (when compared with mammalian cell lines, for example).

The aim is to use yeast culture supernatants to build a partial-synthetic *P. sojae* secretome for study. Oomycete proteins have been successfully expressed in *S. cerevisiae* in the past, including a *P. sojae* pleitropic drug resistance transporter (Connolly et al., 2005). To date, only one oomycete secreted protein has been expressed in *S. cerevisiae* - a β-amylase from *S. ferax* (Kim et al., 2000; Choi et al., 2002), which was successfully secreted into the culture medium using the proteins native signal peptide. To date, no secreted GH from *P. sojae* has been expressed in *S. cerevisiae*; however cellulose- and xylan-degrading enzymes from fungi, insects and bacteria have been successfully secreted in this yeast (Haan et al., 2007; Shirley et al., 2014) (Fang et al., 2017a).

Chapter 2

General Materials and Methods

2.1 Bioinformatics Analysis

For identification of HGTs, a bioinformatics pipeline consisting of a series of PERL scripts (Richards et al., 2009) was used to generate phylogenetic trees for predicted protein sequences. BUSCO (Benchmarking Universal Single-Copy Orthologs) was used to measure genome *completeness* across oomycete genomes used for further analysis (Simao et al., 2015). HGT paralogs were confirmed by Hidden Markov Model (HMM) searches against Ensembl genomes with default parameters (Finn et al., 2015), using raw profile HMM training sets for putative protein families (Pfam; Finn et al., 2016). *P. sojae* transcriptome data for candidate paralogs was compared during three lifecycle stages (normalised from paired-end RNA-seq data; FungiDB; Stajich et al., 2012; Basenko et al., 2018)). SignalP 3.0 (Bendtsen et al., 2004) with default eukaryote parameters, and WoLFPSORT (Horton et al., 2007) with default fungi parameters, were used to identify putative N-terminal secretion signals in the amino acid sequences, and predicted protein localisation, respectively. The CAZy database was mined for GH families present in *P. sojae* with predicted activities similar to those acquired by HGT (http://www.cazy.org; Lombard et al., 2013). Three-dimensional protein structures were obtained with Phyre2 (Protein Homology/analogY Recognition Engine v2.0 (Kelly and Sternberg., 2009)). Clustal Omega was used to align amino acid sequences (Larkin et al., 2007; Madeira et al., 2019). Molecular weight of the proteins was calculated with Compute pI/Mw tool (available at

ExPASY; http://web.expasy.org/compute_pi/). N-glycosylation sites were predicted with NetNGlc 1.0 (Gupta et al., 2004). Phosphorylation sites (serine, threonine and tyrosine) were predicted with NetPhos 3.1 (Blom et al., 1999). Ligand binding sites were predicted by 3DLigandSite (Wass et al., 2010).

2.2 Media and culture conditions for yeast and bacteria

Bacterial strains were grown in rich Luria Broth (LB), with appropriate antibiotic selection (kanamycin or ampicillin at 100 µg/mL). Bacteria were routinely cultured in 5 mL at 37°C (180 rpm).

Yeast strains were grown in rich Yeast Peptone Dextrose (YPD) (1% (w/v) yeast extract, 2% (w/v) bactopeptone, 2% (w/v) glucose) or Synthetic Complete Media minus uracil (SCM-URA) (0.67% (w/v) yeast nitrogen base without amino acids, 2% (w/v) glucose or other appropriate sugar, and 0.077% (w/v) drop-out URA amino acid mixture) for plasmid maintenance. Yeast strains were routinely cultured in 5-20 mL at 30°C (180 rpm).

Solid media was prepared by adding 1.5-1.8% (w/v) agar. All media was autoclaved using a standard autoclave cycle (120°C for 15 minutes for <100 mL or 30 mins for >100 mL), or filter-sterilised through a 0.22 µM pore.

2.3 Synthesis of *P. sojae* GH12 and GH10 paralogs

Candidate nucleotide sequences were yeast codon optimised to accommodate high expression levels in heterologous host, *S. cerevisiae* (http://www.genscript.com/tools/rare-codon-analysis (selection of *S. cerevisiae* as the expression host organism)). The gene sequences were externally synthesised by Synbio Technologies into standard pUC57_Amp or p426-GPD vectors, received as 4 µg lyophilized plasmid and diluted to 100 ng/µL (stored at -20°C). *P. sojae* N-terminal signal peptide sequences were replaced with *S. cerevisiae* mating factor α (MFA) – a *pre-pro* signal sequence including Kex (KR) and Ste13 (EAEA) protease cleavage sites (267 bp)), to direct export from the cell. *S. cerevisiae* MFA sequence as in pGAPZ-alpha *Pichia pastoris* expression vector (Thermo Fisher)).

2.4 Polymerase Chain Reaction (PCR)

Unless otherwise stated, all PCR reactions were set up in a total of 25 µL with Q5 High-fidelity DNA polymerase (NEB), using a standard protocol: 5 µL 5x Q5 Reaction Buffer, 0.5 µL 10 mM dNTPs, 1.25 µL 10 µM forward primer, 1.25 µL 10 µM reverse primer, 0.25 µL Q5 polymerase, 1-2 µL template DNA, and H_2O (up to 25 µL). PCR conditions were optimised based on a recommended protocol for the polymerase (initial denaturation at 98°C (30 seconds), followed by 25-30 cycles of denaturation at 98°C (5-10 seconds), annealing at 50-72°C* (10-30 seconds), and extension at 72°C (20-30 seconds per kb), with a final extension at 72°C (2 minutes)). *annealing temperatures for primer sets were adjusted using NEBTm Calculator. PCR products were separated by gel electrophoresis using 1.8% (w/v) agarose (Sigma) in TAE buffer (4.84 g/L Tris base, 1.14 mL/L glacial acetic acid, 0.37 g/L EDTA). A 1 kb DNA ladder was used in parallel (NEB), and samples were mixed with loading dye containing GelRed (Biotium; a safer alternative to ethidium bromide for staining DNA), run at 100 V for ~ 30-45 minutes, and visualised by UV exposure. PCR purification was carried out using Thermo Scientific GeneJET PCR purification kit according to the manufacturer's instructions, and quantified using a Nanodrop spectrophotometer (ND-2000; Thermo Fisher).

2.5 Cloning - restriction digests

Unless otherwise stated, 30 µL DNA (plasmid or insert) was digested overnight at 37°C, with 1 µL restriction enzyme (a total of 10 units) in 5 µL 10x buffer and 14 µL H_2O. Each digested DNA insert was PCR-purified as previously described, and each digested plasmid (vector) backbone was confirmed by running the entire reaction on agarose gel (100 V for > 1 hour). A linear band (representing the digested vector) was excised and purified using Thermo Scientific GeneJET Gel Extraction kit, according to the manufacturer's instructions. Phosphatase treatment was carried out by incubating 10 µL of the digested, purified vector with 1 µL 10x Antarctic Phosphatase reaction buffer and 1 µL Antarctic Phosphatase at 37°C for 30 minutes, followed by 70°C for 5 minutes. All samples were stored on ice or at -20°C for long-term storage.

Unless otherwise stated, a ratio of DNA insert to vector backbone of 4:1 was used to set up ligation reactions, i.e. 8 µL of insert was mixed with 2 µL of vector, with 2 µL T4 DNA ligase buffer (10x), 0.4 µL T4 DNA ligase (400 units), in a total of 20 µL. Ligations were carried out overnight at room temperature.

2.6 *E. coli* transformation

A total of 2 µL of plasmid was incubated with 50 µL of chemically-competent *E. coli* (Thermo Fisher) in a 1.5 mL sterile tube on ice for 30 minutes, followed by a heat shock of 42°C for 45 seconds. The tube was briefly incubated on ice for 2 minutes, followed by addition of 250 µL sterile LB, and incubated at 37°C for 1 hour (180 rpm). Volumes of 20 µL or 100 µL were plated on LB agar supplemented with an appropriate antibiotic, and incubated at 37°C overnight.

2.7 *E. coli* plasmid purification

Three single *E. coli* colonies were re-streaked onto appropriate agar, and inoculated into 5 mL LB broth (with the appropriate antibiotic selection) in 3x sterile 50 mL tubes. The cells were grown overnight at 37°C (180 rpm). Plasmid purification was carried out according to the manufacturer's instructions (GeneJET plasmid miniprep kit, Thermo Fisher). Plasmid DNA was quantified using a Nanodrop, and stored at -20°C. DNA sequences were confirmed by Sanger Sequencing using an appropriate sequencing primer.

2.8 *S. cerevisiae* transformation (electroporation method)

A yeast colony was inoculated into 5 mL YPD broth in a sterile 50 mL tube. The cells were grown overnight at 30°C (180 rpm). A fresh tube of 15 mL YPD was prepared, and 250 µL of the overnight yeast culture was added. The culture was grown until it reached OD_{600} ~1.5, and the cells collected by centrifugation at 2300 xg for 5 minutes. The supernatant was removed, the cells were washed in 50 mL sterile $_dH_2O$ (centrifugation as before), and resuspended in 8 mL $_dH_2O$, 1 mL 10xTE and 1 mL LiAc by gentle shaking. The cells were incubated at 30°C for 30 minutes (with shaking), after which 250 µL cold 1M Dithiothreitol (DTT) was added and the cells were incubated at 30°C for a further 60 minutes (with shaking). 40 mL of sterile $_dH_2O$ was added to the tube and the cells were harvested by centrifugation at 2300 xg for 5 minutes (at 4°C). The cells were washed once in 25 mL cold sterile $_dH_2O$, and once in 5 mL cold 1 M sorbitol (centrifugation as before), and finally resuspended in 550 µL 1 M sorbitol. The competent cells were stored on ice; 45 µL of cells were added to a sterile 1.5 mL tube; 5 µL of plasmid DNA was added to the tube and incubated on ice for 5-10 minutes. The transformation mixture was transferred to a sterile pre-chilled electroporation cuvette (2 mm), which was subject to electroporation (BioRad) with the following pulse settings for yeast: 200 Ω, 1.5 kV, 25 µF; then, 950 µL of YPD was added to the cuvette and the entire mixture was transferred back into the original 1.5 mL tube, and incubated at 30°C overnight. The transformation mixture was plated on SCM-URA agar to select for plasmid maintenance, and incubated at 30°C for 2 days.

2.9 Plasmids used for this study

Plasmid	Description	Source
p426-GPD	Yeast high copy vector Promoter: constitutive GPD *P. sojae* sequences (including MFA signal peptide) cloned via 5’ *SmaI* and 3’ *HindIII* Selectable markers: *ura3*, *ampR* No tag	Addgene
p426-GPD-6xHis	Yeast high copy vector Promoter: constitutive GPD *P. sojae* sequences (including MFA signal peptide) cloned via 5’ *SmaI* and 3’ *HindIII* Selectable markers: *ura3*, *ampR* C-terminal 6xHis tag	Addgene
pYF515	CRISPR vector (encoding Cas9 nuclease) Promoter: *P. sojae* RPL41 gRNA cloned via *NheI* and *BsaI* Selectable markers: *ampR* No tag	Dr Fang (Duke University)
pBluescript-KS(-)	CRISPR vector (for construction of HDR template)	Dr Fang (Duke University)

Table 2.1. Plasmids used for this study.

2.10 Plasmids constructed for this study

Plasmid	Description	Source
p426-GPD-MFA	Yeast high copy vector Promoter: constitutive GPD MFA signal peptide sequence cloned via 5' *SpeI* and 3' *SmaI* *P. sojae* sequences cloned via 5' *SmaI* and 3' *HindIII* Selectable markers: *ura3*, *ampR* No tag	This study
p426-GPD-MFA-6xHis	Yeast high copy vector Promoter: constitutive GPD MFA signal peptide sequence cloned via 5' *SpeI* and 3' *SmaI* *P. sojae* sequences cloned via 5' *SmaI* and 3' *HindIII* Selectable markers: *ura3*, *ampR* C-terminal 6xHis tag	This study
p426-GPD_*P. sojae*_482953_noSP	Yeast high copy vector Promoter: constitutive GPD *P. sojae*_482953 missing its N-terminal signal peptide (noSP – no signal peptide) cloned via 5' *HindIII* and 3' *ClaI* Selectable markers: *ura3*, *ampR* No tag	This study
pYF515_482953	CRISPR vector (encoding Cas9 nuclease and gRNA for *P. sojae*_482953 Promoter: *P. sojae* RPL41 gRNA cloned via *NheI* and *BsaI* Selectable markers: *ampR* No tag	This study
pBluescript-KS(-) HDR_482953	CRISPR vector with GFP HDR template for replacement of *P. sojae*_482953	This study

Table 2.2. Plasmids constructed for this study.

2.11 General sequencing primers used in this study

Primer	Sequence 5' to 3'
M13 forward (F)	ACTGGCCGTCGTTTTAC
M13 reverse (R)	GGAAACAGCTATGACCATG
GPD_Pro forward (F)	CGGTAGGTATTGATTGTAATTCTG

Table 2.3. General sequencing primers used in this study.

2.12 Primer sequences used to clone *P. sojae* genes

Primer	Sequence 5' to 3'
SmaI MFA forward	TAA GCA CCC GGG ATG AGA TTT CCT TCA
520599 *HindIII* C-tag Reverse	TGC TTA AAG CTT GTT CGC ACC CGC CGT
520924 *HindIII* C-tag Reverse	TGC TTA AAG CTT AGC ACT ATT AAC TGC
355355 *HindIII* C-tag Reverse	TGC TTA AAG CTT TGC AGT ATG GAC TGC
260883 *HindIII* C-tag Reverse	TGC TTA AAG CTT CTT TAC CTG GAC AGA
338064 *HindIII* C-tag Reverse	TGC TTA AAG CTT CTT GAC CTG CAC TGA
559651 *HindIII* C-tag Reverse	TGC TTA AAG CTT ATT GAC CGC AGC AGA
360375 *HindIII* C-tag Reverse	TGC TTA AAG CTT CGT TGA TAG GGG TAG
482953 *HindIII* C-tag Reverse	TGC TTA AAG CTT TTC ACG CCT CAC CCT
247788 *HindIII* C-tag Reverse	TGC TTA AAG CTT ATT GAC AGC AGC TGA

338074 *SmaI* forward	TAA GCA CCC GGG CAG GAG GAG TTC TGT
338074 *HindIII* No tag Reverse	TGC TTA AAG CTT CTA TTG TTG TTG TAC
338074 *HindIII* C-tag Reverse	TGC TTA AAG CTT TTG TTG TTG TAC
520248 *SmaI* forward	TAA GCA CCC GGG GAT AAG ATG TGT GGG
520248 *HindIII* No tag Reverse	TGC TTA AAG CTT CTA GTT TTG GGA GAA
520248 *HindIII* C-tag Reverse	TGC TTA AAG CTT GTT TTG GGA GAA
527497 *HindIII* C-tag Reverse	TGC TTA AAG CTT ATT TTG CAG GTT CAT
519234 *HindIII* C-tag Reverse	TGC TTA AAG CTT CAA TAT CCA ACC GGC
489338 *HindIII* C-tag Reverse	TGC TTA AAG CTT GGA TTG CCA AGC CTG
518763 SmaI Forward	TAA GCA CCC GGG GAC CCT CGT GTG ATG
518763 *HindIII* No tag Reverse	TGC TTA AAG CTT CTA TAA AAT CCA TCC
518763 *HindIII* C-tag Reverse	TGC TTA AAG CTT TAA AAT CCA TCC
482953 No tag forward	TAA GCA AAG CTT GCT GAA TTC TGT GAT CAG
482953 *ClaI* No tag Reverse	TGC TTA ATC GAT CTA TTC ACG CCT CAC

Table 2.4. Primer sequences used to clone *P. sojae* genes.

2.13 Bacterial strains used in this study

Strain	Genotype	Source
DH5 α	F$^-$ Φ80*lac*ZΔM15 Δ(*lac*ZYA-*arg*F) U169 *rec*A1 *end*A1 *hsd*R17(r$_K^-$, m$_K^+$) *pho*A *sup*E44 *thi*-1 *gyr*A96 *rel*A1 λ$^-$	ThermoFisher Scientific (18265017)

Table 2.5. Bacterial strains used in this study.

2.14 Yeast strains used in this study

Strain	Genotype	Source
S. cerevisiae BY4742	MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0	Haynes group (University of Exeter)

Table 2.6. Yeast strains used in this study.

2.15 Other microbial strains used in this study

Strain	Source
P. sojae P6497 (Race 2)	Dr Fang (Duke University)

Table 2.7. Other microbial strains used in this study.

Chapter 3

Using bioinformatics, protein structure- and substrate binding-prediction tools to investigate *P. sojae* paralogs putatively involved in plant cell wall degradation

3.1 Overview

The plant cell wall is a heterogeneous structure representing the first barrier to abiotic and biotic stresses including pathogen attack (Hamann., 2012). As a result, plant parasites must be able to digest or break through the external barrier - both as a means to scavenge nutrients from degraded cell wall carbohydrates, and as part of a process to gain entry to host cells to facilitate infection. The genomes of plant parasites are abundant in genes encoding secreted digestive enzymes for this purpose, including previously identified HGTs into oomycete genomes from fungi (Torto et al., 2002; Richards et al., 2006; Belbahri et al., 2008; Richards et al., 2011; Misner et al., 2015; Savory et al., 2015). The selective benefit of maintaining the HGT genes is suggested by evidence of widespread gene duplication (post-acquisition) of HGT-acquired genes - however, knowledge is limited in understanding the functional significance of the paralogous proteins for the degradation of plant carbohydrates in phytopathogenic oomycetes. Further characterisation is therefore crucial - especially as most proteins of interest may only be distantly-related to biochemically-characterised enzymes. This chapter aims to characterise putative functional differences between *P. sojae* paralogs of horizontally-acquired enzymes for plant cell wall degradation, using available computational methods.

3.2 Introduction

3.2.1 The plant cell wall

The plant cell wall is a polysaccharide-rich structure formed from cellulose microfibrils cross-linked by glycans, and embedded within a 3D matrix composed of pectin, hemicellulose, lignin and other various aromatic polymers, and proteins, which together provide structural integrity (McNeil et al., 1984; Schindler., 1998). Whilst cell wall composition varies among plant species (and across different tissues within the same plant, as well as throughout various life stages), cellulose is always the most abundant polysaccharide and makes up 20-50% (w/v) of a typical plant cell wall, whilst hemicellulose (including all of the cellulose-binding polysaccharides) constitutes around 15-35% (w/v), and primarily includes xyloglucan, xylan, arabinoxylan, glucuronoxylan, mannan and glucomannan (Kubicek., 2013; Alvarez et al., 2016). Cellulose microfibrils are synthesised in the plasma membrane (Mueller et al., 1976; Mueller and Brown., 1980), whilst hemicelluloses (and non-hemicellulosic sugars) are synthesised in the Golgi and delivered to the surface of the cell wall by vesicles (Driouich et al., 1993; Lerouxel et al., 2006; Driouich et al., 2012). Pectin is the major non-hemicellulosic polysaccharide, formed by polymerisation of D-galacturonic acid (an oxidised form of D-galactose) (Atmodjo et al., 2013), with complex functions involved in cell wall porosity and cell wall thickness, as well as in connecting cells in the middle lamella (Iwai et al., 2002). It is thought that pectin binds to the hemicellulose xyloglucan to strengthen the cell wall (Rizk., 2000; Cumming et al., 2005), and the degradation products of pectin are also important elicitors of plant immune responses (Hahn et al., 1981; Ferrari et al., 2013). Many plants also possess an extracellular lipid (hydrophobic) membrane called the cuticle, consisting of cutin with embedded waxes (Jeffree., 1996).

Cellulose is a simple polymer of D-glucose units linked by β-1,4-glycosidic bonds; polymerisation occurs when the anomeric carbon (C1) of a glucose residue covalently joins to the hydroxyl oxygen (O4) of the next glucose residue (resulting in the net release of one water molecule for each glucose residue added to the chain) (Sandgren et al., 2001). Together, these cellulose polymers form microfibrils that are linked by the hemicelluloses, which are divided into classes based on the main sugar in the backbone, and are often more complex in structure than cellulose due to the presence of multiple side groups. For example, xyloglucan consists of a backbone of D-glucose units linked by β-1,4-glycosidic bonds (like cellulose), however, most are substituted with α-1,6-linked xylose (up to 75%), which are often in turn capped with a galactose, arabinose, or sometimes a fucose residue. Xyloglucan is the most abundant hemicellulosic polysaccharide, however, owing to the different possible variations of the sugar side chains on the xylosyl residues, its structure varies significantly between plant species (York et al., 1990, 1996; Hoffman et al., 2005; Pena et al., 2008) – a nomenclature-based system has been described to help to define the side chain variants (Fry et al., 1993). Whilst xyloglucan is thought to stabilise the adjacent cellulose fibrils, the importance of xyloglucan for the formation of the cross-linked cellulose network is unclear – for example, there is a xyloglucan-deficient *Arabidopsis* mutant that is still able to form cross-linked cellulose (albeit more irregularly spaced than a wild-type strain) (Cavalier et al., 2008; Anderson et al., 2010). Some studies suggest xyloglucan creates important tensile strength in hardwoods (Nishikubo et al., 2007; Mellerowicz et al., 2008) - nevertheless, degradation of xyloglucan is necessary for complete lignocellulosic decomposition (e.g. Hayashi and Kaida., 2011).

Another major component of hemicellulose is xylan, which is formed from β-1,4-linked D-xylose (a pentose sugar), with many side chains substituted with α-arabinofuranose and α-glucuronic acids. *Arabidopsis* mutants deficient in xylan have weakened cell walls and cannot develop a vascular system (Brown et al., 2007; Wu et al., 2009), similarly suggesting the importance of close interactions between xylan within the hemicellulose structure, and cellulose microfibrils, for plant cell wall integrity and strength (e.g. Busse-Wicher et al., 2014; Simmons et al., 2016) – a tough external barrier that plant pathogens must overcome.

3.2.2 Plant cell wall-degrading enzymes

Cellulose as the major structural component of plant cell walls was discovered in 1837 (Hon., 1994), and by extension of its abundance in plant cell walls, it is considered the most abundant polysaccharide on the planet that is not easily broken down in nature. Therefore, cellulolytic organisms that can release carbon from its degradation (i.e. for nutrient sources), play a significant role in cellulose recycling in the biosphere, contributing to global carbon fluxes (Levesque et al., 2010; de Vries and Visser., 2001). Whilst some phytopathogens can partly overcome the plant barrier by formation of mechanical structures such as appressoria (e.g. Grenville-Briggs et al., 2008; Wilson and Talbot., 2009), in order to completely breach the plant cell wall, release sugars, and gain entry to cells, many phytopathogens synthesise and secrete proteins to degrade cellulose, as well as the other components of the cuticle and cell wall layers (Mueller et al., 2008; Raffaele et al., 2010). Across the tree of life, eukaryotes have evolved to break down living and dead plant biomass and many of the genes involved are considered pathogenicity factors; for example, *Fusarium solani* deficient in cutin-degrading enzymes has significantly reduced virulence (Dantzig et al., 1986), as is the case for pectinase-deficient *Nectria hematococca* (Rogers et al., 2000). At the molecular level, secretion of plant cell wall-degrading enzymes weakens the plant cell wall structure and often facilitates successful parasite infection. Such extracellular enzymes must withstand external conditions (variations in temperature or pH) at the plant cell wall surface, as well as tolerate plant proteolytic attack (Jones and Dangl., 2006; Schwessinger and Ronald., 2012).

Cellulases are well-studied enzymes that degrade cellulose; these are naturally expressed and secreted by many microorganisms including the filamentous ascomycete fungi, *Tricoderma reesei*, which is the best studied cellulolytic organism to date, and is often exploited in industry practises for cellulase production as a biological tool to degrade plant biomass for biofuel production (see historical review of *T. reesei* as an enzyme producer - Bischof et al., 2016). Many industrial processes require elevated temperatures or non-physiological pHs; significant research has improved the understanding of *T. reesei* biology (as well as other cellulolytic microbes), leading to successes in protein engineering of cellulases with enhanced biochemical characteristics, such as improved enzyme stability at high temperatures and across broad pH ranges, as well as enriched heterologous secretion of enzymes from suitable recombinant hosts (e.g. Day et al., 2007; Lantz et al., 2010). Extending our understanding of the cellulolytic activity of other plant-pathogenic microbes therefore has the potential to identify other novel protein features for enhanced cellulose breakdown across various industrial processes.

Cellulose degradation by anaerobic and aerobic microbes follows two distinct mechanisms; whilst anaerobes use multi-bound enzyme complexes (cellulosomes; Bae et al., 2013), aerobes use combinations of single enzymes (named the ‘free enzyme paradigm’; Gupta et al., 2016) – the latter of which contributes to the highest plant polysaccharide breakdown in nature (Kubicek., 2013). Due to the tough crystalline nature of cellulose microfibrils in plant cell walls (which makes cellulose extremely resistant to digestion), many cellulolytic microorganisms secrete multiple ‘types’ of enzymes that target different parts of the structure (Wilson., 2009). Complete cellulose degradation by single enzymes involves the synchronised activity of endoglucanases (which cleave amorphous sites of the cellulose backbone), exoglucanases (or cellobiohydrolases, which cleave the new side chains produced from endoglucanase activity, yielding mainly cellobiose (a disaccharide of D-glucose)), and β-glucosidases (which cleave cellobiose and liberate D-glucose). The synergy between the different types of enzymes enables more rapid and efficient cellulose breakdown (Henrissat et al., 1985; Nidetzky et al., 1994). Few cellulose-degrading fungi (which share some biological similarities with some oomycetes) can also combine the activities of cellobiose dehydrogenase and copper-dependent oxidases (Langston et al., 2011; Quinlan et al., 2011), whilst many cellulases have carbohydrate-binding modules (CBMs) (Tomme et al., 1995) - which are not required for enzymatic activity, but play a role in promoting the binding of the enzyme to its substrates (Boraston et al., 2004; Tomme et al., 1995). Enzymatic degradation of xylan follows a similar synergy of three hydrolytic enzymes involving endoxylanases, exoxylanases, and β-xylosidases (review by Biely et al., 2016) – and together, these groups of enzymes are known as Carbohydrate-Active enZymes (CAZymes).

3.2.3 Carbohydrate-Active enZymes (CAZymes)

CAZymes are enzymes involved in the synthesis, metabolism and transport of carbohydrates; CAZy (http://www.cazy.org; Lombard et al., 2013) is a specialist database for these enzymes, which are divided into families based on sequence and structural similarities. New enzymes are assigned to a CAZyme family based on significant amino acid similarity with at least one biochemically characterised member of the family. As a result, predicted substrates for enzymes are based on assignment to a family (this has the potential to limit functional annotation amongst proteomic datasets in the absence of experimental characterisation – biochemically characterised enzymes with substrate specificities that lie outside of the CAZyme family they are assigned to are often later grouped into 'subfamilies' to try and account for this). Currently, the CAZy database describes Glycoside Hydrolases (GH), Glycosyl Transferases (GT), Polysaccharide Lyases (PL) and Carbohydrate Esterases (CE), grouped according to the type of chemical bond they attack and mechanism of action (http://www.cazy.org; Lombard et al., 2013).

GH enzymes hydrolyse glycosidic bonds in complex sugars, and are the most represented group in the database so far, with 148 families described, spanning a range of substrate preferences and enzymatic activities. Distribution of GH families in oomycete genomes appears to be species-specific, with expansion in *Phytophthora* spp. (Adhikari et al., 2013; McGowan and Fitzpatrick., 2017), suggesting an importance of these types of digestive enzyme activities (at least for **hemibiotrophic**[2] lifestyles). Consistently, all CAZy-encoded genes reported as horizontally-acquired in the oomycetes are GHs (Torto et al., 2002; Richards et al., 2011; Savory et al. 2015).

3.3 Aims of chapter

Comparative phylogenetic analyses identified multiple cases of HGT into the oomycete lineage from fungi (Torto et al., 2002; Richards et al., 2006; Belbahri et al., 2008; Richards et al., 2011; Misner et al., 2015; Savory et al., 2015). Of these transfer events, many are predicted to encode secreted proteins putatively associated with degradation of plant cell wall-specific substrates – providing a means of entry into plant tissues for parasitic oomycetes, as well as an abundant source of fixed carbon during hyphal growth. As previously mentioned, there has been widespread gene duplication following lateral acquisition of GH enzymes in the oomycetes (Richards et al., 2011; Savory et al., 2015), giving rise to paralogs that could possess as yet unknown significant differential or overlapping functions important for plant carbohydrate digestion (Ohno., 1970; Stoltzfus., 1999; Force et al., 1999).

This chapter aims to re-confirm 11 previously-identified HGT events associated with breakdown of the plant cell wall, including updated taxon sampling to investigate the evolutionary ancestry of the putative transfer, and confirm the total numbers of paralogs across representative *complete* oomycete genomes (i.e. spanning different ecological lifestyles). Two HGT events will then be selected for further analysis – a GH12 enzyme family putatively involved in

[2] Hemibiotrophic pathogens live biotrophically with their plant hosts before switching to a necrotrophic phase.

cellulose degradation, and a GH10 enzyme family putatively involved in xylan degradation. Using the hemibiotroph, *P. sojae* (a widespread pathogen of soybean), as a model species for analysis, the putative protein functions acquired by HGT will be compared to alternative cellulose- and xylan-degrading capabilities across the oomycetes (i.e. by identifying other GH gene families in oomycete genomes that putatively perform the same enzyme functions as well). As HGT events can be broadly separated into 'maintenance' transfers and 'innovative' transfers (as described in Chapter One) (Husnik and McCutcheon., 2018), it is important to consider the wider functional impact of the HGTs across oomycete species - it is hypothesised that GH12- and GH10-like activities should be more widespread amongst oomycetes with close plant associations that feed on host-derived substrates (therefore the abundance of such secreted proteins would be higher than (for example) saprotrophic oomycetes that feed on decaying matter).

All confirmed paralogs of both GH12 and GH10 in *P. sojae* (11 and 4 protein sequences, respectively) will be investigated using bioinformatics tools - using phylogenetic tree-building methods to confirm the evolutionary history of *P. sojae* homologous sequences (and their respective orthologs across other oomycetes). The *P. sojae* protein sequences will be further investigated using protein three-dimensional structure-prediction tools, post-translational modification site-prediction tools, and ligand (carbohydrate) binding-prediction tools - in order to elucidate putative functional differences between the paralogs, as well as identify amino acid sequence and structural features that could theoretically alter the putative substrate interaction sites *in vivo*. Publically-available transcriptome (RNA-sequencing) data will be used to explore the differences in gene expression across *P. sojae* life stages (including during infection of its soybean host), and relative genomic locations of the GH12 and GH10 genes will be investigated - it is hypothesised that such analysis could uncover duplicated paralogs with putative functional differences *in vivo*.

3.4 Methods

3.4.1 Identification of putative HGTs

11 HGT events were selected from Torto et al. (2002), Belbahri et al. (2008), Richards et al. (2011) and Savory et al. (2015), representing acquired oomycete gene families associated with plant cell wall breakdown (See Appendix I for putative protein functions). Representative oomycete protein sequences (in FASTA format) were selected as seed sequences for analysis, and a bioinformatics pipeline consisting of a series of PERL scripts (Richards et al., 2009) was used to process each seed sequence separately. This pipeline was used to generate phylogenetic trees for predicted protein sequences: Protein Basic Local Alignment Search Tool (BLASTp) (Altschul et al., 1990) was used to search for protein homologs against a custom-built database of ~1200 prokaryote and eukaryote taxa (e-value 1e^{-10}), with no top hit limit for *P. sojae*, in order to retrieve all possible paralogous sequences for this species (maximum of three sequences for all other taxa). Recovered sequences were aligned using MUSCLE (Edgar., 2004) and uninformative alignment sites (e.g. gaps) removed (masked) using GBLOCKS (Castresana, 2000; Talavera & Castresana., 2007). Phylogenetic tree construction was carried out by PhyML (Guindon and Gascuel., 2003), and resulting tree image files were edited to include conserved protein domains (Pfam; Finn et al., 2016). Phylogenies were scrutinised for tree

topologies suggestive of HGT into the oomycetes – that is, oomycete sequences branching within a clade of fungal (or bacterial) sequences.

Phylogenetic trees were used to annotate the presence or absence of each HGT across representative oomycete genomes, including *Phytophthora* (hemibiotrophs), *Pythium* (**necrotrophs**[3]), saprophytic and non-pathogenic oomycetes, and *H. catenoides* (free-living representative of the sister group to the oomycetes). Thirty-five genomes spanning diverse oomycete species (see McCarthy and Fitzpatrick., 2017) were selected for analysis of *completeness* using BUSCO (Benchmarking Universal Single-Copy Orthologs) (Simao et al., 2015) in order to be selected for analysis. BUSCO provides a quantitative measure of genome *completeness*, based on expected gene content in comparison with a 'core' gene list derived from multiple eukaryotic genomes (note that for BUSCO v.2, the core gene list is only derived from fungal and metazoan genomes, which may limit estimates of completion outside of these groups). The output of BUSCO analysis is split into four categories: i) Complete and single-copy, ii) Complete and duplicated, iii) Fragmented, or iv) Missing BUSCOs (Simao et al., 2015).

3.4.2 Identification of HGT paralogs

For phylogenies suggestive of HGT into the oomycetes, all paralogous sequences in *P. sojae* were selected for further analysis. Sequence homologs (paralogs) for each HGT family were confirmed by performing Hidden Markov Model (HMM) searches against Ensembl genomes with default parameters (Finn et al., 2015), using raw profile HMM training sets for putative protein families

[3] Necrotrophic pathogens kill host plant cells and feed on dead/dying tissues.

(Pfam; Finn et al., 2016). Profile HMMs use multiple sequence alignments gained from training data sets to compute a (position-specific) scoring matrix - the score at each position is determined by how conserved each amino acid is in the alignment, and takes into account biases such as which amino acids are most likely to occur at which positions, as well as the probability of insertions/deletions at specific positions in the sequence. When the profile HMM is compared to a genome database, high-scoring sequences (compared to the 'null') are considered homologous to the sequences used to construct the profile HMM. E-values (i.e. the number of hits expected by chance alone) can be set as default or user-defined (Finn et al., 2015). BLAST-based methods do not distinguish insertions and deletions based on position (i.e. they are all scored equally), therefore HMM searches were chosen to enable accurate identification of the true protein homologs.

3.4.3 Identification of N-terminal secretion signals

SignalP 3.0 (Bendtsen et al., 2004) with default eukaryote parameters, and WoLFPSORT (Horton et al., 2007) with default fungi parameters, were used to identify putative N-terminal secretion signals in the amino acid sequences, and predicted protein localisation, respectively[4]. SignalP-based predictions use HMM algorithms (v2 and v3; Nielsen and Krogh., 1998; Bendtsen et al., 2004) – the training data sets enable identification of differences in amino acid composition from the N-terminus to the mature protein (allowing the prediction, presence and

[4] It is important to note that a positive secretion prediction does not mean that a protein is extracellular, as proteins with N-terminal signal sequences have been shown to be retained intracellularly (e.g. in the ER or Golgi), and similarly, a negative secretion prediction does not mean that a protein is intracellular, as some proteins can be secreted via non-classical pathways. (e.g. Rubartelli., 1997).

location of signal peptide cleavage sites). Each amino acid position is given three scores (0-1): the C-score, which recognises a cleavage site vs no cleavage site, the S-score, which recognises a signal peptide vs no signal peptide, and the Y-score, which takes into consideration both the C-score and the S-score to optimise the overall prediction (Nielsen and Krogh., 1998; Bendtsen et al., 2004).

WoLFPSORT predicts protein localisation to 10+ sites based on a *k*-nearest neighbour algorithm, i.e. prediction is based on the closest similarity to proteins (of known localisation) in the existing data set within the model. The scoring matrix is based on features such as amino acid composition and sorting signals, and the output can include dual localisation prediction. Results are given as a list of proteins in order of similarity to the query sequence, with details of their localisation features (Horton et al., 2007).

3.4.4 Further analysis of candidate HGT families

GH12 and GH10 HGT candidates were taken forward for further analysis. GH12 enzymes were annotated as having endo-1,4-β-glucanase (EC 3.2.1.4), xyloglucan endo-hydrolase (EC 3.2.1.151) and endo-1,3-1,4-β-glucanase (EC 3.2.1.73) activities (http://www.cazy.org; Lombard et al., 2013); other GH families putatively annotated with one or more of the same activities were identified by searching the CAZy database, identifying GH5, -6, -7 and -17 present in *P. sojae* (http://www.cazy.org; Lombard et al., 2013). The presence of the GH families across the oomycetes, and the total numbers of paralogs, were confirmed by carrying out profile HMM searches as previously described (Finn et al., 2015).

GH10 enzymes were annotated as having endo-1,4- β-xylanase (EC 3.2.1.8), endo-1,3- β-xylanase (EC 3.2.1.32) and xylan endotransglycosylase (EC 2.4.2.-) activities (http://www.cazy.org; Lombard et al., 2013); both the latter activities were not found in any other GH family in *P. sojae*, however endo-1,4-β-xylanase activity was found to be associated with GH5, 30 and 43. Similarly, the presence of these GH families and the total numbers of paralogs across the oomycetes was explored as detailed previously (Finn et al., 2015).

P. sojae GH12 and GH10 protein sequences were aligned with Clustal Omega (Larkin et al., 2007; Madeira et al., 2019), using *T. reesei*_Cel12a (fungal GH12) and *T. reesei*_Xyn3 (fungal GH10) for comparison. Three-dimensional structures of paralogous proteins of GH12 and GH10 were obtained with Phyre2 (Protein Homology/analogY Recognition Engine v2.0 (Kelly and Sternberg., 2009)), using protein sequences without their predicted N-terminal signal peptide. N-glycosylation sites were predicted with NetNGlc 1.0 (Gupta et al., 2004), and phosphorylation sites (serine, threonine and tyrosine) were predicted with NetPhos 3.1 (Blom et al., 1999). Predictions of carbohydrate-binding sites for paralogous proteins were generated by 3DLigandSite (Wass et al., 2010) – an automated pipeline based on a previously used human method for predicting binding sites in CASP8 (Wass and Sternberg., 2009; Wass et al., 2010). The server was used to identify high-scoring homologous protein structures with ligands bound (by comparative MAMMOTH score) (Wass et al., 2010).

For each paralog of GH12 and GH10, *P. sojae* transcriptome (RNA-sequencing) data was compared during three lifecycle stages: mycelial, cyst and 3 days post-infection (soybean hypocotyls infected with *P. sojae* strain P6497), expressed as transcript levels of fragments per kilobase of exon model per million mapped reads (FPKM - sequencing depth and gene depth normalised from paired-end RNA-seq data; FungiDB; Stajich et al., 2012; Basenko et al., 2018). *P. sojae* transcriptome data was additionally used to scrutinise HGT gene sequences from published gene models. The FungiDB genome browser tool was additionally used to investigate the genomic locations of (and distances between) the *P. sojae* GH12 and GH10 genes (Stajich et al., 2012; Basenko et al., 2018).

A phylogenetic tree was constructed to confirm the evolutionary relationships between *P. sojae* GH12 paralogs, and orthologs in other oomycetes as follows: GH12 protein homologs were identified from a selection of eukaryotic (with a focus on oomycete and fungi) and prokaryotic genomes using (BLASTp) (Altschul et al., 1990); from these hits a multiple sequence protein alignment was constructed and aligned using automated methods in Seaview, it was then edited and masked by eye using Seaview (Galtier et al., 1996). A preliminary phylogenetic tree was calculated using the built-in PhyML option within Seaview, this was then checked manually (enabling distantly-related fungal and prokaryotic outgroups, and partial sequences to be removed) and re-masked (Guindon and Gascuel., 2003), finally resulting in an alignment with 161 sequences and 211 amino acid sites. A second PhyML tree was calculated and checked, and the final ML tree was constructed with IQ-Tree v2.0.3 to take advantage of the automatic model selection criteria (type of analysis: ModelFinder (Kalyaanamoorthy et al., 2017) with tree reconstruction and non-parametric bootstrap (200 replicates) (Minh et al., 2020)). As part of the run, WAG+R5 was predicted as the best-fit model of evolution (Whelan and Goldman., 2001). The tree was rooted with a fungal outgroup because (although distantly related to the oomycetes), fungi were previously identified as the putative donor group of the GH12 HGT (Richards et al., 2011; Savory et al., 2015).

3.5 Results

3.5.1 Presence of HGT families associated with plant cell wall degradation across oomycetes

Eleven previously-identified HGT events from fungi to the oomycetes (Torto et al., 2002; Belbahri et al., 2008; Richards et al., 2011; Savory et al., 2015) were re-confirmed through phylogenetic reconstruction, and the presence or absence of each were verified across thirty-five oomycete genomes (McCarthy and Fitzpatrick., 2017) - including hemibiotrophs (*Phytophthora* spp.), necrotrophs (*Pythium* spp.), saprotrophs, and non-pathogenic species, in addition to the sister group to the oomycetes (*H. catenoides*) (Figure 3.1). The coloured grey boxes indicate the presence of the HGT gene in the genome of that organism, whilst numbers indicate the total sequence homologs (paralogs) of that gene family in each organism – confirmed by performing profile HMM searches against Ensembl genomes (Finn et al., 2015). Results of BUSCO analysis indicates that all genomes included scored >80% for 'complete and single-copy BUSCOs' (Simao et al., 2015) (Base of Figure 3.1).

None of the HGT genes were identified in the genome of *H. catenoides*, the sister group to the oomycetes (Figure 3.1). The HGTs identified were found in higher numbers across members of the hemibiotrophic *Phytophthora* genus, including *P. infestans* T30-4, *P. nicotiniae*, *P. parasitica* P1569, *P. lateralis* MPF-4, *P. ramorum*, *P. sojae* and *P. kernoviae* 00238/432 – contributing between >50 and >150 proteins to the total putative secretomes of these organisms. Across all of the oomycete genomes sampled, GH88 (encoding a putative α-L-rhamnosidase) was only confirmed to be present in *Phytophthora* spp., and GH28 (encoding a putative pectin hydrolase) was only confirmed in genomes of *Peronosporales*, i.e. hemibiotrophs and obligate biotrophs. All of the HGTs were found in reduced paralog numbers within the genomes of the two obligate biotrophs within the order *Peronosporales*, contributing a total of 19 proteins to the *P. halstedii* putative secretome, and 27 proteins to the *H. arabidopsidis* (Emoy-2) putative secretome. GH78 and GH88 were not found encoded in either obligate biotrophs - both with putative functions in the degradation of pectin (α-L-rhamnosidase and d-4,5-unsaturated beta-glucuronyl hydrolase activities, respectively). *P. halstedii* was not found to encode a horizontally-transferred pectate lyase gene – involved in the breakdown of pectin polysaccharides (Figure 3.1).

3.5.1 Presence of HGT families associated with plant cell wall degradation across oomycetes

	Peronosporales																				Pythiales										Saprolegniales					
	Phytophthora infestans T30-4	*Phytophthora nicotianae*	*Phytophthora parasitica* P1569	*Plasmopara viticola*	*Plasmopara halstedii*	*Phytophthora multivora*	*Phytophthora capsici*	*Phytophthora pluvialis*	*Hyaloperonospora arabidopsidis* Emoy2	*Phytophthora agathidicida*	*Phytophthora pinifolia*	*Phytophthora cryptogea*	*Phytophthora lateralis* MPF4	*Phytophthora ramorum*	*Phytophthora fragariae*	*Phytophthora rubi*	*Phytophthora sojae* P6497	*Phytophthora pisi*	*Phytophthora cinnamomi*	*Phytophthora kernoviae* 00238/432	*Phytopythium vexans*	*Plasporangium apinafurcum*	*Pythium irregulare*	*Pythium iwayami*	*Pythium ultimum* var. *ultimum*	*Pythium ultimum* var. *sporangiiferum*	*Pythium aphanidermatum*	*Pythium arrhenomanes*	*Pythium insidiosum*	*Pythium oligandrum*	*Albugo laibachii*	*Saprolegnia diclina* VS20	*Saprolegnia parasitica* CBS 223.65	*Aphanomyces invadans*	*Aphanomyces astaci*	*Hyphochytrium catenoides*
CO-esterase	5	24	19		3				4				5	18			14			7	9		8	2			8	20				4	3	7	6	
Cutinase	2	8	6		2				2				4	4			14			2							9	7				1	1			
GH12	9	17	14		3				3				6	8			11			8								2								
GH10	4	9	4		1				2				7	7			5			2	1							3								
GH43	3	8	5		1				1				6	6			4			6	3		1				1	1								
GH78	4	9	4										1	3			5			2	2			1			1									
GH53	3	6	6		1				1				3	6			5			3	1							1							1	
Pectate lyase	7	30	15						4				6	11			15			12	3		4	1			6						1			
GH88	2	1	1										1	1			1			4																
FAD binding	13	17	16		5				7				3	11			15			7	7		6	4			6	4				8	8	7	9	
GH28	23	36	17		3				3				13	17			21			9																
	75	165	107		19				27				55	92			110			62	26		19	8			31	38				13	13	14	16	

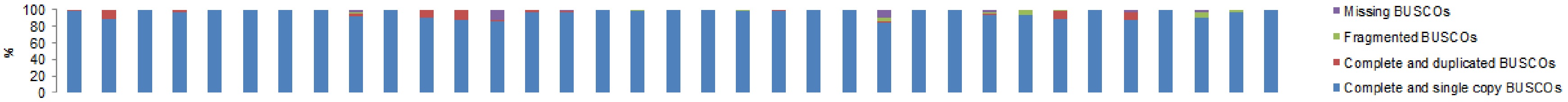


Figure 3.1. Presence of HGTs across oomycete genomes. 11 HGT events were selected from Torto et al (2002), Belbahri et al (2008), Richards et al (2011) and Savory et al (2015), representing acquired oomycete gene families associated with plant cell wall breakdown. Phylogenies generated using protein sequences were used to annotate the presence or absence of each HGT across oomycete genomes, including *Phytophthora* (hemibiotrophs), *Pythium* (necrotrophs), saprophytic and non-pathogenic oomycetes, and *H. catenoides* (the sister group to oomycetes). Significance E-values: 0.01 (sequence). BUSCO (Simao et al., 2015) was used to assess the genome completeness for each oomycete genome used for analysis; results are shown at the base of the figure. The results demonstrate that the HGTs are largely retained amongst hemibiotrophic oomycetes, with higher total paralog numbers across *Phytophthora* spp., compared to oomycetes occupying other ecological niches.

3.5.2 Presence of cellulose and xylan-degrading activities across oomycetes

GH12 enzymes include endo-1,4-β-glucanases (EC 3.2.1.4), xyloglucan endo-hydrolases (EC 3.2.1.151) and endo-1,3-1,4-β-glucanases (EC 3.2.1.73) (http://www.cazy.org; Lombard et al., 2013). Mining the CAZy database for other GH families with the above activities identified **GH5**[5], -6, -7 and -17 (that are encoded by *P. sojae*) (Figure 3.2). The coloured grey boxes indicate the presence of the GH gene, whilst numbers indicate the total number of putative homologs (paralogs) of that gene family in each organism – confirmed by performing profile HMM searches as previously described (Finn et al., 2015).

Higher total numbers of CAZymes associated with the activities of GH12 enzymes were identified in the genomes of hemibiotrophic oomycetes - including *P. infestans* T30-4 (23 proteins), *P. nicotiniae* (50 proteins), *P. parasitica* P1569 (34 proteins), *P. lateralis* MPF-4 (24 proteins), *P. ramorum* (26 proteins), *P. sojae* (34 proteins) and *P. kernoviae* 00238/432 (26 proteins). Conversely, for obligate biotrophic oomycetes, total CAZymes conferring putative cellulose-degrading activities (associated with the enzyme activities related to GH12) were found in reduced numbers - 14 proteins in *P. halstedii*, and 11 proteins in *H. arabidopsidis* genomes. Within the order *Pythiales*, *P. arrhenomanes* was the only sampled genome found to encode horizontally-transferred GH12 - however *P. vexans*, *P. irregulare*, *P. iwayami* and *P. aphanidermatum* were all confirmed to encode paralogs of GH6, 7 and 17, and total numbers of CAZymes across the GH families varied between 8 and 22 proteins. None of the oomycetes sampled from

[5] As GH5 is associated with activities not involved in cellulose/xylan degradation (http://www,cazy.org), for clarity it is not included in Figure 3.2 or Figure 3.3.

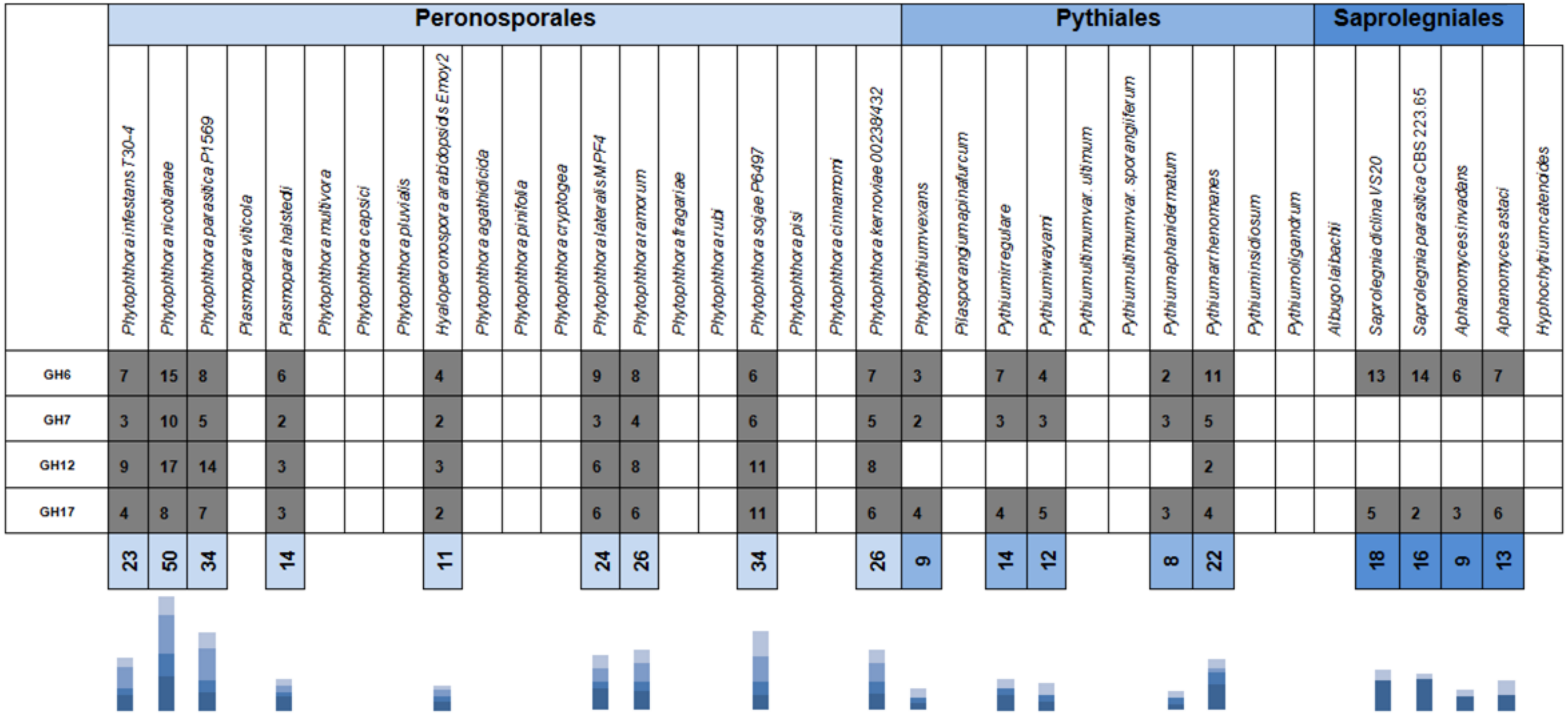
the order *Saprolegniales* were found to encode GH7 or GH12 enzymes (by HMM search methods), however, high paralog numbers of GH6 were identified amongst aquatic parasites *S. diclina* VS20 (13 proteins) and *S. parasitica* CBS 223.65 (14 proteins) - contributing to a total of 18 and 16 proteins associated with putative cellulose-degradation to each genome, respectively (Figure 3.2).

GH10 enzymes include endo-1,4- β-xylanases (EC 3.2.1.8), endo-1,3- β-xylanases (EC 3.2.1.32) and xylan endotransglycosylases (EC 2.4.2.-) (http://www.cazy.org; Lombard et al., 2013). Other GH families putatively encoding one or more of these activities were identified by mining the CAZy database. Endo-1,3- β-xylanase and xylan endotransglycosylase activities were not found in any other GH family, but endo-1,4- β-xylanase activity was found to be associated with **GH5**[5], 30 and 43 (that are encoded by *P. sojae*) (Figure 3.3). The coloured grey boxes indicate the presence of the HGT gene, whilst numbers indicate the total number of putative homologs (paralogs) of that gene family in each organism – confirmed by performing profile HMM searches as previously described (Finn et al., 2015).

Higher total numbers of CAZymes associated with the activities of GH10 enzymes were identified in the genomes of hemibiotrophic oomycetes - including *P. infestans* T30-4 (24 proteins), *P. nicotiniae* (33 proteins), *P. parasitica* P1569 (23 proteins), *P. lateralis* MPF-4 (21 proteins), *P. ramorum* (22 proteins), *P. sojae* (21 proteins) and *P. kernoviae* 00238/432 (16 proteins). Conversely, for obligate biotrophic oomycetes, total CAZymes conferring xylan-degrading activities (associated with the enzyme activities related to GH10) were found in reduced numbers - 7 proteins in *P. halstedii*, and 6 proteins in *H. arabidopsidis* genomes.

None of the oomycetes sampled from the order *Saprolegniales* were found to encode GH43, or horizontally-transferred GH10. Comparatively lower total numbers of xylanase-associated CAZymes were identified in these organisms (only associated with GH30) - *S. diclina* VS20 (2 proteins), *S. parasitica* CBS 223.65 (4 proteins), *A. invadans* (2 proteins) and *A. astaci* (2 proteins) (Figure 3.3).

3.5.2 Presence of cellulose-degrading activities across oomycetes



	Peronosporales																			
	Phytophthora infestans T30-4	Phytophthora nicotianae	Phytophthora parasitica P1569	Plasmopara viticola	Plasmopara halstedii	Phytophthora multivora	Phytophthora capsici	Phytophthora pluvialis	Hyaloperonospora arabidopsidis Emoy2	Phytophthora agathidicida	Phytophthora pinifolia	Phytophthora cryptogea	Phytophthora lateralis MPF4	Phytophthora ramorum	Phytophthora fragariae	Phytophthora rubi	Phytophthora sojae P6497	Phytophthora pisi	Phytophthora cinnamomi	Phytophthora kernoviae 00238/432
GH6	7	15	8		6				4				9	8			6			7
GH7	3	10	5		2				2				3	4			6			5
GH12	9	17	14		3				3				6	8			11			8
GH17	4	8	7		3				2				6	6			11			6
	23	50	34		14				11				24	26			34			26

	Pythiales									
	Phytopythium vexans	Pilasporangium apinafurcum	Pythium irregulare	Pythium iwayami	Pythium ultimum var. ultimum	Pythium ultimum var. sporangiiferum	Pythium aphanidermatum	Pythium arrhenomanes	Pythium insidiosum	Pythium oligandrum
GH6	3		7	4			2	11		
GH7	2		3	3			3	5		
GH12								2		
GH17	4		4	5			3	4		
	9		14	12			8	22		

	Saprolegniales					
	Albugo laibachii	Saprolegnia diclina VS20	Saprolegnia parasitica CBS 223.65	Aphanomyces invadans	Aphanomyces astaci	Hyphochytrium catenoides
GH6		13	14	6	7	
GH7						
GH12						
GH17		5	2	3	6	
		18	16	9	13	

Figure 3.2. Presence of GH12-like activities across oomycete genomes. GH12 is annotated as having endo-1-4-β-glucanase (EC 3.2.1.4), xyloglucan endo-hydrolase (EC 3.2.1.151) and endo-1,3-1,4-β-glucanase (EC 3.2.1.73) activities. Other *P. sojae* GH families putatively associated with one or more of the same activities are GH5*, 6, 7 and 17 (http://www.cazy.org; Lombard et al., 2013); presence and expansion of all GH families across oomycete genomes is shown (the bars at the base of the figure represent the total numbers of paralogs across all GH families). The results show diversity and expansion in GH12-like activities amongst hemibiotrophic oomycetes. ***GH5 is one of the most widespread GH families present in archaea, bacteria and eukaryotes, and is associated with activities not involved in cellulose/hemicellulose degradation (http://www,cazy.org; Lombard et al., 2013); for clarity, GH5 is not included in the figure.**

3.5.2 Presence of xylan-degrading activities across oomycetes

	Peronosporales									
	Phytophthora infestans T30-4	*Phytophthora nicotianae*	*Phytophthora parasitica P1569*	*Plasmopara viticola*	*Plasmopara halstedii*	*Phytophthora multivora*	*Phytophthora capsici*	*Phytophthora pluvialis*	*Hyaloperonospora arabidopsidis Emoy2*	*Phytophthora agathidicida*
GH10	4	9	4		1				2	
GH30	17	16	14		5				3	
GH43	3	8	5		1				1	
	24	33	23		7				6	

	Peronosporales									
	Phytophthora pinifolia	*Phytophthora cryptogea*	*Phytophthora lateralis MPF4*	*Phytophthora ramorum*	*Phytophthora fragariae*	*Phytophthora rubi*	*Phytophthora sojae P6497*	*Phytophthora pisi*	*Phytophthora cinnamomi*	*Phytophthora kernoviae 00238/432*
GH10			7	7			5			2
GH30			8	9			12			8
GH43			6	6			4			6
			21	22			21			16

	Pythiales										Saprolegniales					
	Phytopythium vexans	*Pilasporangium apinafurcum*	*Pythium irregulare*	*Pythium iwayami*	*Pythium ultimum var. ultimum*	*Pythium ultimum var. sporangiiferum*	*Pythium aphanidermatum*	*Pythium arrhenomanes*	*Pythium insidiosum*	*Pythium oligandrum*	*Albugo laibachii*	*Saprolegnia diclina VS20*	*Saprolegnia parasitica CBS 223.65*	*Aphanomyces invadans*	*Aphanomyces astaci*	*Hyphochytrium catenoides*
GH10	1							3								
GH30	9		6	6			2	6				2	4	2	2	
GH43	3		1				1	1								
	13		7	6			3	10				2	4	2	2	

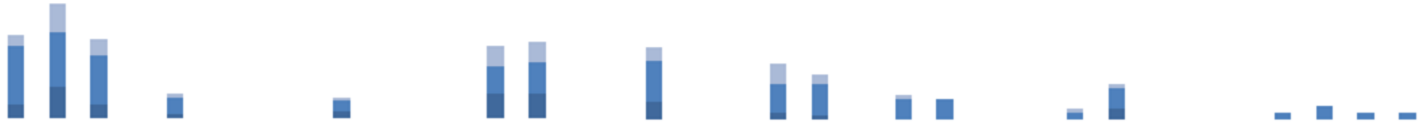


Figure 3.3. Presence of GH10-like activities across oomycete genomes. GH10 is annotated as having endo-1-4-β-xylanase (EC 3.2.1.8), endo-1-3-β-xylanase (EC 3.2.1.8) and xylan endotransglycosylase (EC 2.4.2.-) activities; both the latter activities were not found in any other GH family, however endo-1-4-β-xylanase activity was also found to be associated with GH5*, 30, and 43 (http://www.cazy.org; Lombard et al., 2013); presence and expansion of all GH families across oomycete genomes is shown (the bars at the base of the figure represent the total numbers of paralogs across all GH families). The results show diversity and expansion in GH10-like activities amongst hemibiotrophic oomycetes. ***GH5 is one of the most widespread GH families present in archaea, bacteria and eukaryotes, and is associated with activities not involved in cellulose/hemicellulose degradation (http://www,cazy.org; Lombard et al., 2013); for clarity, GH5 is not included in the figure.**

3.5.3.1 Sequence and structural features of *P. sojae* GH12 paralogs

P. sojae GH12 protein sequences were aligned with Clustal Omega (Larkin et al., 2007; Madeira et al., 2019), using *T. reesei*_Cel12a (Fungal GH12) as a comparison (Sandgren et al., 2001) (Figure 3.4, Table 3.1 for protein alignment key).

*	Fully conserved residue
:	Conservation between groups of strongly similar properties
.	Conservation between groups of weakly similar properties
-	Gap

Table 3.1. Protein alignment key (Clustal Omega (Larkin et al., 2007; Madeira et al., 2019)).

The *P. sojae* GH12 protein alignment was manually annotated to include predicted 3D structural features (α-helices and β-sheets) (Phyre2; Kelly and Sternberg., 2009), N-glycosylation sites (NetNGlc 1.0; Gupta et al., 2004), phosphorylation sites (serine, threonine and tyrosine) (NetPhos 3.1; Blom et al., 1999), predicted ligand (carbohydrate)-binding sites (3DLigandSite; Wass et al., 2010), as well as glutamic acid (Glu; E) catalytic residues (conserved amongst GH12 members (Sandgren et al., 2001)) (Figure 3.4).

All *P. sojae* GH12 proteins except *P. sojae*_360375 were found to have conserved both glutamic acid (Glu; E) residues theoretically required for enzymatic activity (shown by the red vertical arrows (Figure 3.4)). Total putative N-glycosylation sites were variable amongst the GH12 protein sequences (*P.*

*sojae*_520924: 6 sites; *P. sojae*_520248: 5 sites; *P. sojae*_338074 and _520599: 4 sites; *P. sojae*_338064: 3 sites; *P. sojae*_355355, _482953, _247788, _559651 and _360375: 1 site; *P. sojae*_260883: 0 sites). Predicted phosphorylation sites (serine, threonine and tyrosine) were between >20 and 38 for most *P. sojae* GH12 paralogs, however, a significantly higher number of putative phosphorylation sites were identified in *P. sojae*_482953 (97 sites) and *P. sojae*_247788 (107 sites) (Figure 3.4).

3.5.3.2 *P. sojae* paralog _559651 is predicted to have a 'second' carbohydrate-binding site

Predictions of carbohydrate-binding sites were generated by 3DLigandSite (Wass et al., 2010). *P. sojae*_559651 was the only *P. sojae* GH12 paralog in which a putative 'second' carbohydrate-binding site was predicted. The residues putatively involved in this additional binding site are highlighted by the red boxes in the GH12 protein alignment for that sequence (Figure 3.4) – these are Gly55, Ala56, Ala57, Thr58, Val97, Phe205 and Val206 (residue position numbers given for the protein sequence in the absence of its N-terminal signal peptide).

3D structures of all *P. sojae* GH12 paralogous proteins are shown in Figure 3.5. Predictions of the amino acid residues putatively involved in binding the cellulose backbone were generated by 3DLigandSite (Wass et al., 2010), and are labelled in blue. Known glutamic acid (Glu, E) catalytic residues are shown by the red arrows. For *P. sojae*_559651, two carbohydrate-binding sites were predicted – both are shown (labelled (1) and (2)). The black arrow indicates the Thr58 residue common in both predicted binding sites for this paralog.

Orthologous proteins of *P. sojae* paralog _559651 in *P. cactorum* and *P. nicotiniae* were also predicted to have a putative ‘second’ carbohydrate-binding site using 3DLigandSite (Wass et al., 2010) (Figure 3.6). Whilst all of the ‘second’ binding site residues were found to be conserved amongst the three orthologous protein sequences, for *P. cactorum*, Ala57 was not predicted to form part of its putative ‘second’ binding site (Figure 3.6). Inspection of the *P. sojae* GH12 protein alignment (Figure 3.4) revealed two putative indels coding for alanine (Ala, A) and serine (Ser, S) – of all the *P. sojae* sequences, these amino acids are present only in *P. sojae*_559651 and *P. sojae*_360375 protein sequences (Figure 3.7). Removal of both amino acids from the *P. sojae*_559651 protein sequence abolished the prediction of the putative ‘second’ binding site; the two indels were also found to be conserved in orthologs in *P. cactorum* and *P. nicotiniae* (Figure 3.7), and likewise, their removal from those respective protein sequences eliminated both predictions of a putative ‘second’ binding site.

3.5.3.1 Sequence and structural features of *P. sojae* GH12 paralogs

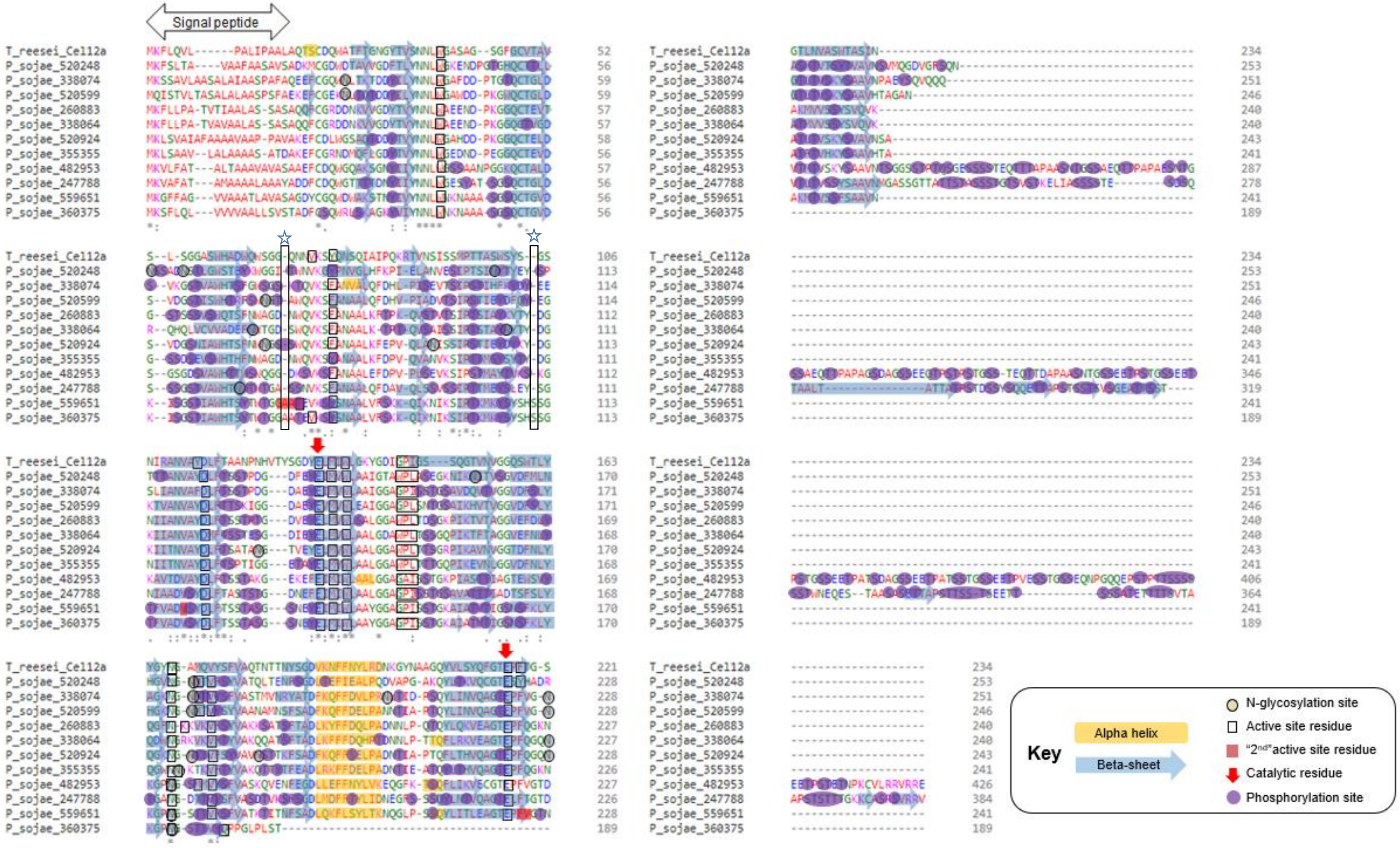
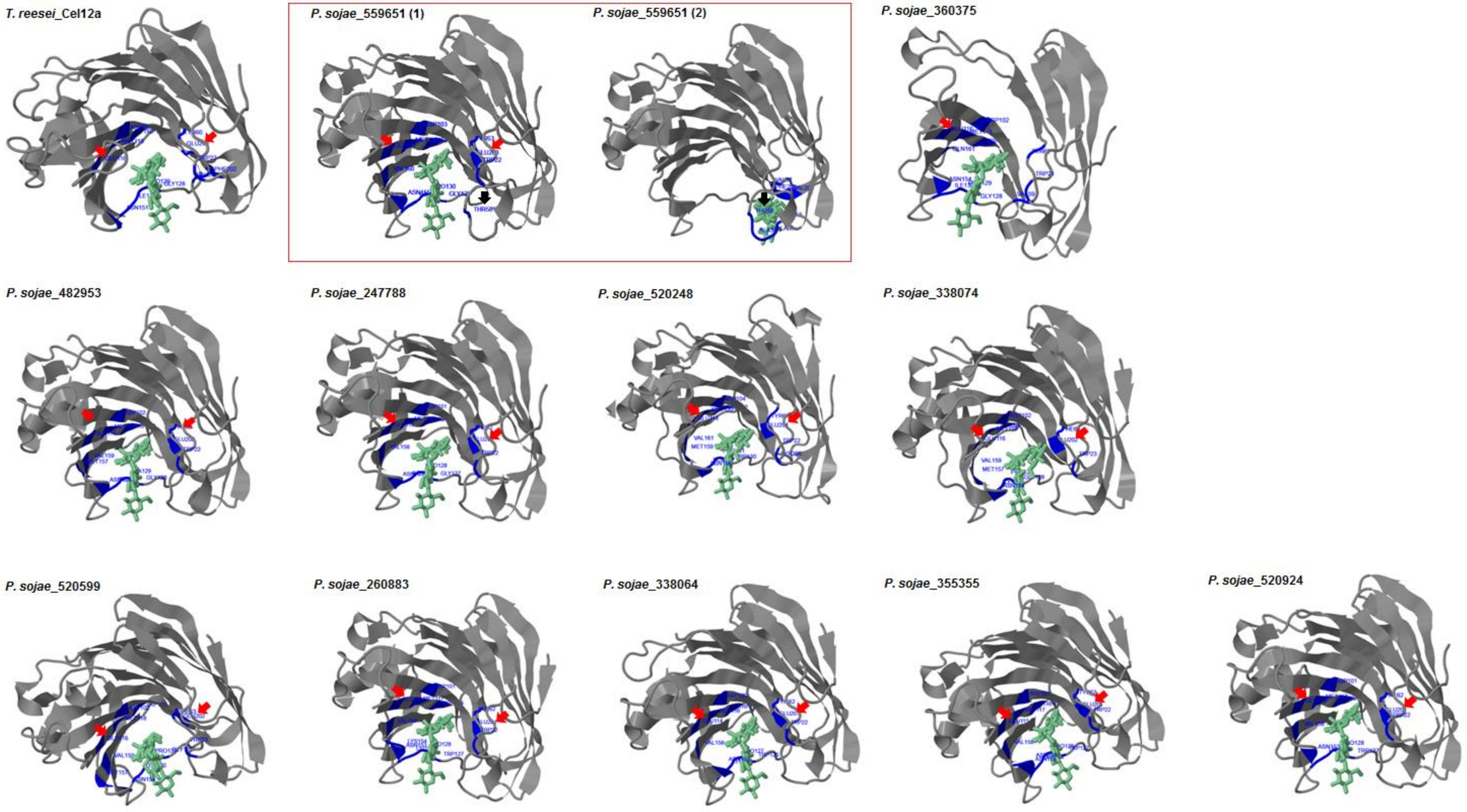


Figure 3.4. Protein alignment of *P. sojae* GH12 paralogs, generated by Clustal Omega (Larkin et al., 2007; Madeira et al., 2019), aligned with *T. reesei* Cel12a (fungal) protein for comparison. N-glycosylation sites were predicted with NetNGlc 1.0 (Gupta et al., 2004); three-dimensional structures of paralogous proteins were obtained with Phyre2 (Protein Homology/analogY Recognition Engine v2.0 (Kelly and Sternberg., 2009)); phosphorylation sites (serine, threonine and tyrosine) were predicted with NetPhos 3.1 (Blom et al., 1999); predictions of carbohydrate-binding sites (active site residues) were generated by 3DLigandSite (Wass et al., 2010). All *P. sojae* paralogs except _360375 possess both glutamic acid residues theoretically required for enzymatic activity (red arrows); *P. sojae*_482953 and _247788 possess long, significantly phosphorylated unique C-terminal 'tail' regions; *P. sojae*_559651 possesses a putative 'second' substrate binding site (as predicted by 3DLigandSite (Wass et al., 2010) – see Figure 3.5), with two putative indels (alanine and serine) that are important for the second binding site prediction (blue stars in the alignment).

3.5.3.2 *P. sojae* paralog _559651 is predicted to have a 'second' carbohydrate-binding site



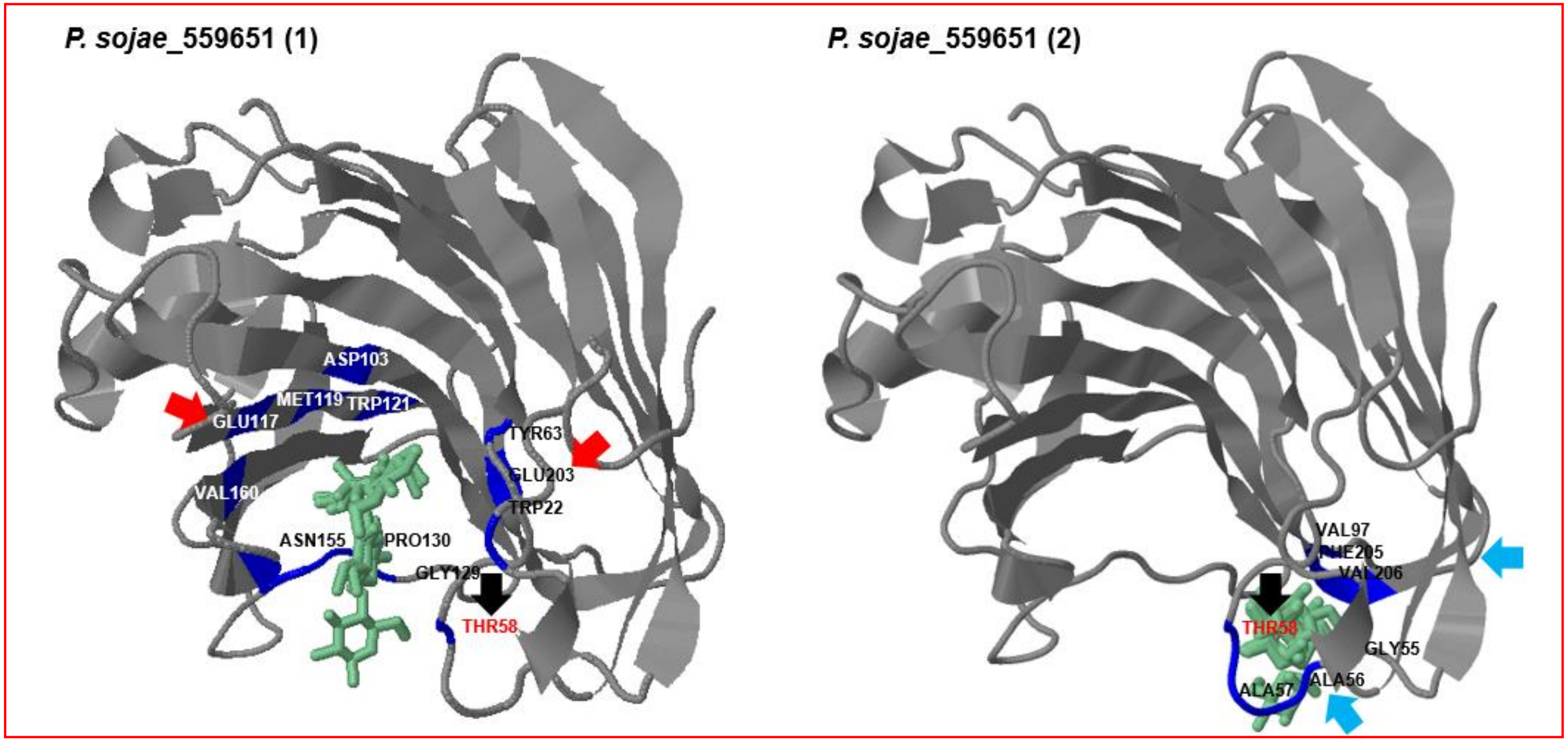


Figure 3.5. Three-dimensional structures of *P. sojae* GH12 paralogs were obtained with Phyre2 (Protein Homology/analogY Recognition Engine v2.0 (Kelly and Sternberg., 2009)); predictions of carbohydrate-binding sites (active site residues) were generated by 3DLigandSite (Wass et al., 2010), and are labelled in the figure. Known glutamic acid [Glu] catalytic residues are shown by the red arrows. For *P. sojae* paralog 559651, two carbohydrate-binding sites were predicted by 3DLigandSite – amino acids predicted for each binding site are labelled (1) and (2). The black arrow indicates the threonine Thr58 residue common in both predicted sites; two putative indels (alanine and serine) that are important for the 'second' binding site prediction for this paralog are indicated by the blue arrows.

Orthologs of *P. sojae* paralog _559651 in *P. cactorum* and *P. nicotiniae* are also predicted to have a ‘second’ carbohydrate-binding site

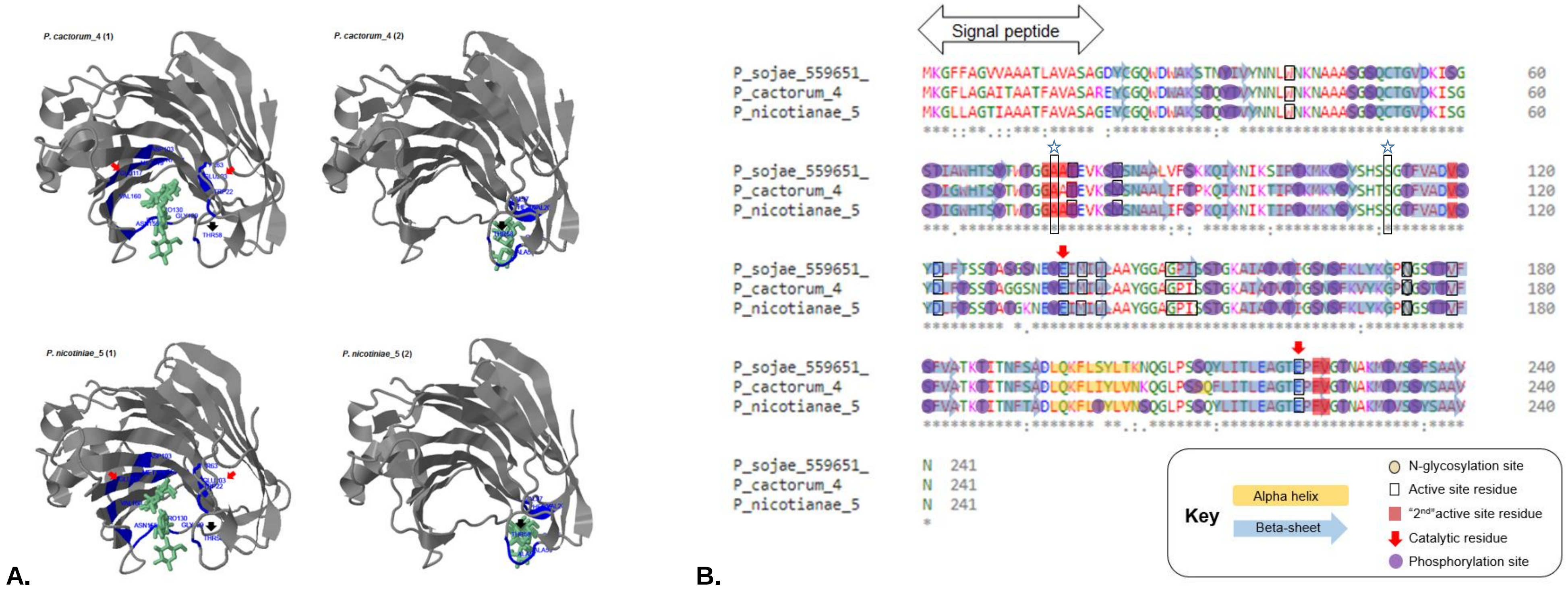
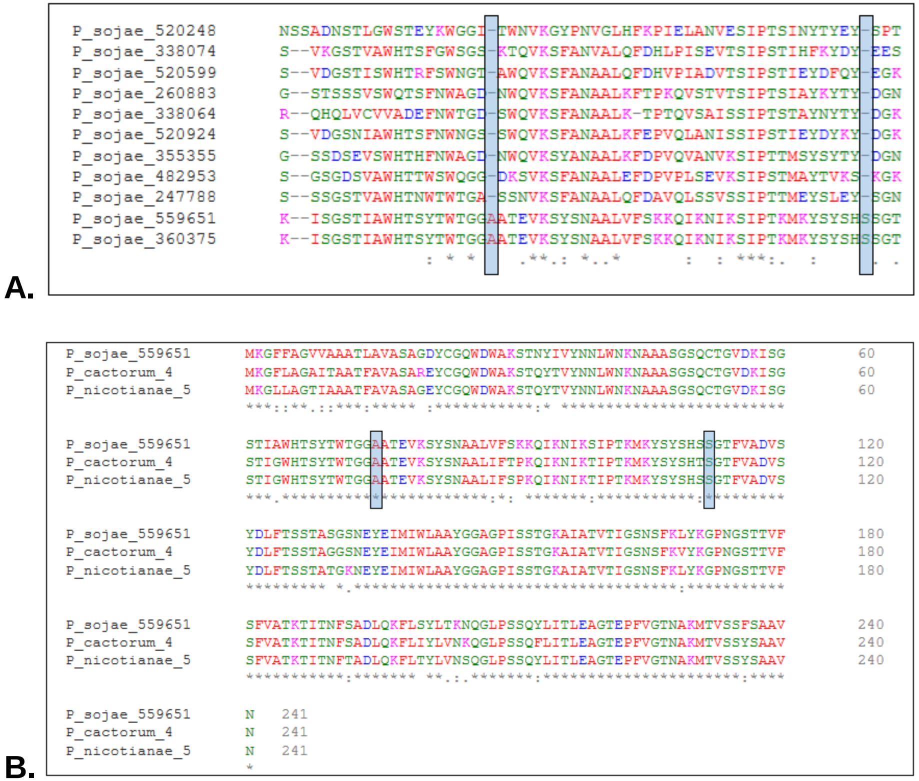


Figure 3.6. *P. sojae*_559651 and orthologs in *P. cactorum* and *P. nicotiniae* are predicted to encode a putative 'second' carbohydrate binding site. **A.** Three-dimensional structures of *P. sojae*_559651 orthologs in *P. cactorum* and *P. nicotianae* were obtained with Phyre2 (Protein Homology/analogY Recognition Engine v2.0 (Kelly and Sternberg., 2009)); predictions of carbohydrate-binding sites (active site residues) were generated by 3DLigandSite (Wass et al., 2010), and are labelled in blue. Known glutamic acid [Glu] catalytic residues are shown by the red arrows. Two carbohydrate-binding sites were predicted by 3DLigandSite – both are shown (labelled (1) and (2)). The black arrow indicates the threonine Thr58 residue common in both predicted sites. **B.** Protein alignment of *P. sojae*_559651 and orthologs in *P. cactorum* and *P. nicotiniae*, generated by Clustal Omega (Larkin et al., 2007; Madeira et al., 2019). N-glycosylation sites were predicted with NetNGlc 1.0 (Gupta et al., 2004); phosphorylation sites (serine, threonine and tyrosine) were predicted with NetPhos 3.1 (Blom et al., 1999); two putative indels (alanine and serine) that are important for the second binding site prediction are indicated by the blue stars in the alignment.

Two indels, coding for alanine and serine, are important for the ‘second’ carbohydrate-binding site prediction



```
P_sojae_520248      NSSADNSTLGWSTEYKWGGI-TWNVKGYPNVGLHFKPIELANVESIPTSINYTYEY-SPT
P_sojae_338074      S--VKGSTVAWHTSFGWSGS-KTQVKSFANVALQFDHLPISEVTSIPSTIHFKYDY-EES
P_sojae_520599      S--VDGSTISWHTRFSWNGT-AWQVKSFANAALQFDHVPIADVTSIPSTIEYDFQY-EGK
P_sojae_260883      G--STSSSVSWQTSFNWAGD-NWQVKSFANAALKFTPKQVSTVTSIPTSIAYKYTY-DGN
P_sojae_338064      R--QHQLVCVVADEFNWTGD-SWQVKSFANAALK-TPTQVSAISSIPTSTAYNYTY-DGK
P_sojae_520924      S--VDGSNIAWHTSFNWNGS-SWQVKSFANAALKFEPVQLANISSIPSTIEYDYKY-DGK
P_sojae_355355      G--SSDSEVSWHTHFNWAGD-NWQVKSYANAALKFDPVQVANVKSIPTTMSYSYTY-DGN
P_sojae_482953      S--GSGDSVAWHTTWSWQGG-DKSVKSFANAALEFDPVPLSEVKSIPSTMAYTVKS-KGK
P_sojae_247788      S--SSGSTVAWHTNWTWTGA-SSNVKSFANAALQFDAVQLSSVSSIPTTMEYSLEY-SGN
P_sojae_559651      K--ISGSTIAWHTSYTWTGGAATEVKSYSNAALVFSKKQIKNIKSIPTKMKYSYSHSSGT
P_sojae_360375      K--ISGSTIAWHTSYTWTGGAATEVKSYSNAALVFSKKQIKNIKSIPTKMKYSYSHSSGT
                                  : * *    .**.: *..*      :  : ***:.  :     . .
```

```
P_sojae_559651      MKGFFAGVVAAATLAVASAGDYCGQWDWAKSTNYIVYNNLWNKNAAASGSQCTGVDKISG     60
P_cactorum_4        MKGFLAGAITAATFAVASAREYCGQWDWAKSTQYTVYNNLWNKNAAASGSQCTGVDKISG     60
P_nicotianae_5      MKGLLAGTIAAATFAVASAGEYCGQWDWAKSTQYTVYNNLWNKNAAASGSQCTGVDKISG     60
                    ***::**.::***:***** :***********:* *************************

P_sojae_559651      STIAWHTSYTWTGGAATEVKSYSNAALVFSKKQIKNIKSIPTKMKYSYSHSSGTFVADVS     120
P_cactorum_4        STIGWHTSYTWTGGAATEVKSYSNAALIFTPKQIKNIKTIPTKMKYSYSHTSGTFVADVS     120
P_nicotianae_5      STIGWHTSYTWTGGAATEVKSYSNAALIFSPKQIKNIKTIPTKMKYSYSHSSGTFVADVS     120
                    ***.***********************:*: *******:***********:*********

P_sojae_559651      YDLFTSSTASGSNEYEIMIWLAAYGGAGPISSTGKAIATVTIGSNSFKLYKGPNGSTTVF     180
P_cactorum_4        YDLFTSSTAGGSNEYEIMIWLAAYGGAGPISSTGKAIATVTIGSNSFKVYKGPNGSTTVF     180
P_nicotianae_5      YDLFTSSTATGKNEYEIMIWLAAYGGAGPISSTGKAIATVTIGSNSFKLYKGPNGSTTVF     180
                    ********* *.************************************:***********

P_sojae_559651      SFVATKTITNFSADLQKFLSYLTKNQGLPSSQYLITLEAGTEPFVGTNAKMTVSSFSAAV     240
P_cactorum_4        SFVATKTITNFSADLQKFLIYLVNKQGLPSSQFLITLEAGTEPFVGTNAKMTVSSYSAAV     240
P_nicotianae_5      SFVATKTITNFTADLQKFLTYLVNSQGLPSSQYLITLEAGTEPFVGTNAKMTVSSYSAAV     240
                    ***********:******* **.:.*******:**********************:****

P_sojae_559651      N   241
P_cactorum_4        N   241
P_nicotianae_5      N   241
                    *
```

Figure 3.7. A. Protein alignment for *P. sojae* GH12 generated by Clustal Omega (Larkin et al., 2007; Madeira et al., 2019); blue boxes highlight the positions of two indels (Alanine; A, and Serine, S), only present in *P. sojae*_559651 and _360375 - when the two amino acid residues are removed from the protein sequence of _559651, the ‘second’ carbohydrate-binding site is no longer predicted by 3DLigandSite (Wass et al., 2010). **B**. Protein alignment for *P. sojae*_559651 and orthologous proteins in *P. cactorum* and *P. nicotianae*, generated by Clustal Omega; blue boxes highlight the positions of the two indels (Alanine; A, and Serine, S) that are conserved between the proteins, and also required for the ‘second’ carbohydrate-binding site predictions in the orthologs.

3.5.3.3 *P. sojae* paralog _482953, and orthologs in *P. cactorum* and *P. nicotiniae* have a highly phosphorylated, significantly disordered C-terminus 'tail'

The protein sequence of *P. sojae*_482953 is predicted to encode **97** putative phosphorylation sites (serine, threonine and tyrosine) (NetPhos 3.1; Blom et al., 1999) – a high proportion of which are clustered towards the C-terminus region of the protein (Figure 3.4). Phyre2 (Kelly and Sternberg., 2009) was unable to confidently assign the terminal 186 amino acids of this protein sequence to an appropriate structural prediction, due to the disordered nature of the 'tail' sequence – consistent with structural predictions of protein orthologs in *P. cactorum* and *P. nicotiniae*. Figure 3.8 (A) shows the results of Phyre2 analysis for the three orthologous proteins (Kelly and Sternberg., 2009) – the colours represent Phyre2 disorder scores (blue to red, from order to disorder); non-modelled C-terminal sequences are shown below the predicted structures with a red star indicating the start positions of the 'tail' sequences on the structures that were unable to be modelled. Figure 3.8 (B) shows the protein alignment of the orthologous sequences (Clustal Omega (Larkin et al., 2007; Madeira et al., 2019)), manually edited for features as described previously (NetNGlc 1.0 (Gupta et al., 2004); 3DLigandSite (Wass et al., 2010); NetPhos 3.1 (Blom et al., 1999)).

3.5.3.3 *P. sojae* paralog _482953, and orthologs in *P. cactorum* and *P. nicotiniae* have a highly phosphorylated, significantly disordered C-terminus ‘tail’

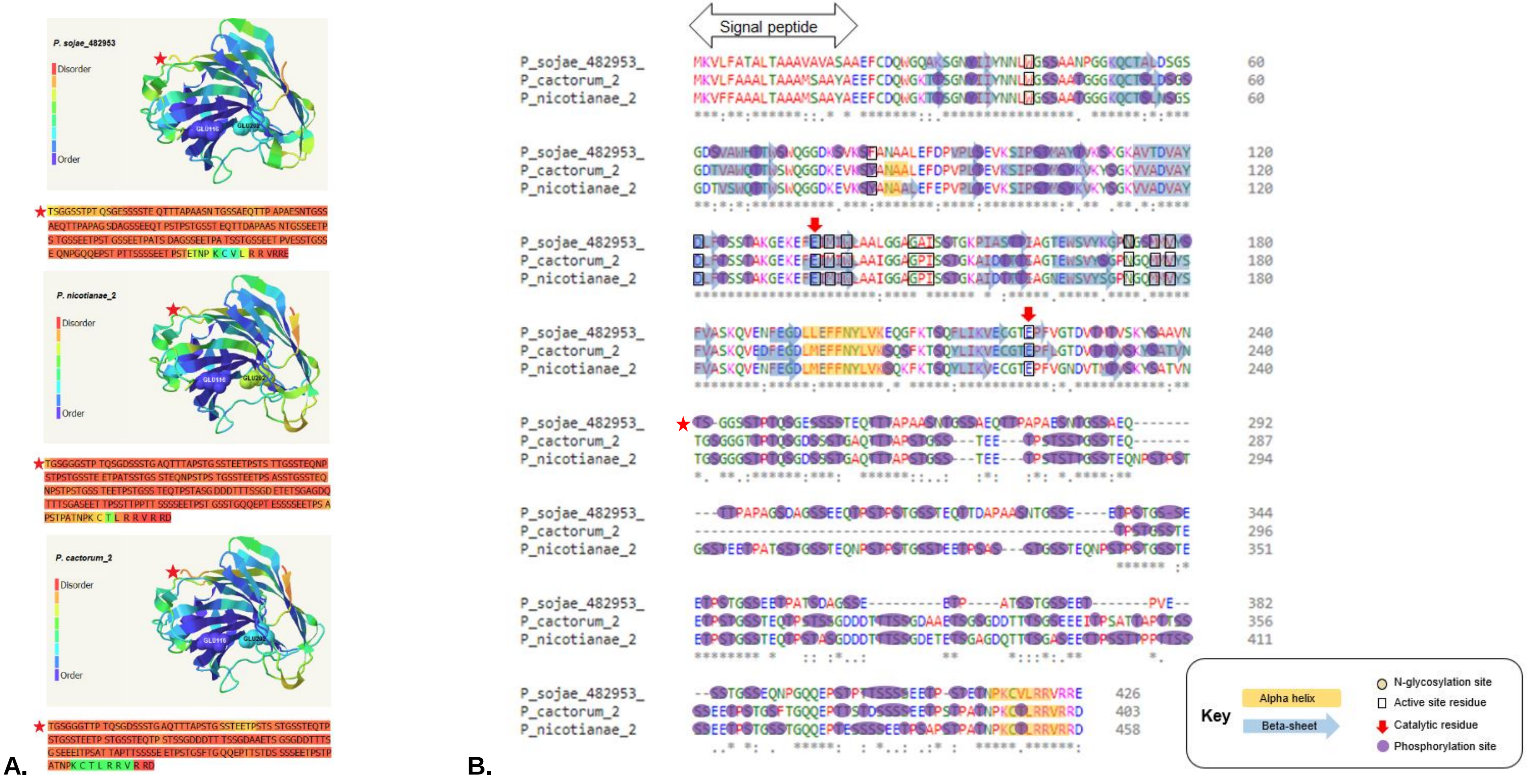


Figure 3.8. A. Three-dimensional structures of *P. sojae*_482953 and orthologs in *P. nicotiniae* and *P. cactorum* were obtained with Phyre2 (Protein Homology/analogY Recognition Engine v2.0 (Kelly and Sternberg., 2009)). C-terminal 'tail' sequences were not able to be modelled; tail sequences are shown below the structures (the red star indicates where on the structure the tail sequence begins). Colours represent disorder scores generated by Phyre2 and demonstrate significant disorder in the C-terminal extensions. **B.** Protein alignment of *P. sojae*_482953 and orthologs in *P. cactorum* and *P. nicotiniae*, generated by Clustal Omega (Larkin et al., 2007; Madeira et al., 2019). N-glycosylation sites were predicted with NetNGlc 1.0 (Gupta et al., 2004); predictions of carbohydrate-binding sites (active site residues) were generated by 3DLigandSite (Wass et al., 2010); phosphorylation sites (serine, threonine and tyrosine) were predicted with NetPhos 3.1 (Blom et al., 1999). The data indicates that the significant disorder and phosphorylation of the C-terminal 'tail' sequences is conserved amongst orthologs of *P. sojae*_482953.

3.5.4 Gene expression profiles of GH12 paralogs

For each *P. sojae* GH12 paralog, transcriptome data was compared during three lifecycle stages: mycelial, cyst and 3 days post-infection (soybean hypocotyls infected with *P. sojae* strain P6497) (FungiDB; Stajich et al., 2012; Basenko et al., 2018). Fragments per kilobase of exon model per million mapped reads (FPKM) values were used to calculate Log2 values (plotted in Figure 3.9).

Genes coding for *P. sojae_* 247788 and *P. sojae_*520599 did not appear to be expressed under any of the conditions tested. For the remaining GH12 paralogs, all show expression during infection (*P. sojae_*338074 and *P. sojae_*520248 are exclusively expressed during infection only (expressed in Log2(FPKM)), at 5.08 and 3.12, respectively). *P. sojae_*355355 is highly expressed across all three lifecycle stages compared to other GH12 genes (4.66 mycelial, 9.23 cyst, and 8.68 infection). Expression of the gene encoding *P. sojae_*482953 is highly upregulated from mycelial to cyst and infection stages (0.95 mycelial, 6.66 cyst, 5.51 infection) (Figure 3.9).

3.5.5 Genomic locations of GH12 paralogs

For each *P. sojae* GH12 gene, the genome browser tool within FungiDB (Stajich et al., 2012; Basenko et al., 2018) was used to locate the coordinates of the eleven GH12 genes within the *P. sojae* genome. Approximate distances between each of the GH12 genes is shown in Figure 3.10. Notably, *P. sojae* tandem repeat pairs _559651 and _360375, and _482953 and _247788 (putatively based on protein sequence similarity between the sequences, see Figure 3.4), have unique genomic locations, 2278 kb and 5868 kb, upstream and downstream, respectively, from the remaining GH12 paralogs (Figure 3.10).

3.5.4 Gene expression profiles of GH12 paralogs

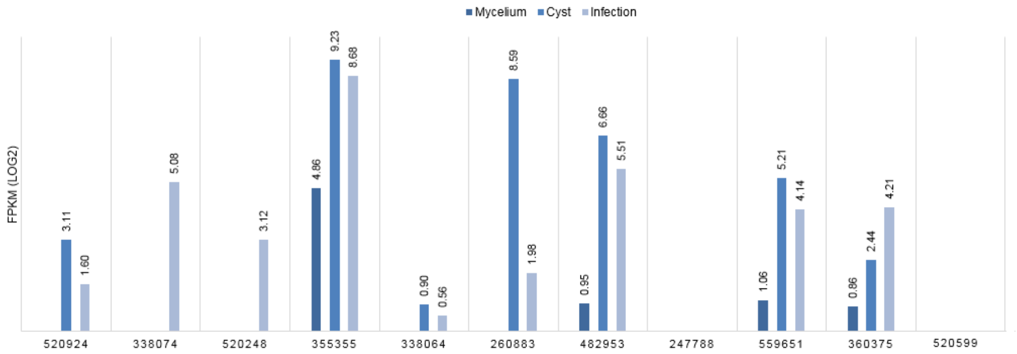


Figure 3.9. For each GH12 gene, *P. sojae* transcriptome data was compared during three lifecycle stages: mycelial, cyst and 3 days post-infection of soybean hypocotyls, expressed as transcript levels of fragments per kilobase of exon model per million mapped reads (FPKM - sequencing depth and gene depth normalised from paired-end RNA-seq data; FungiDB; Stajich et al., 2012; Basenko et al., 2018). FPKM values were used to calculate Log2 (plotted). The data indicates that all paralogs have unique expression profiles during all lifecycle stages, with many upregulated during infection (i.e. close interactions with the plant host and increased capacity for plant cell wall digestion).

3.5.5 Genomic locations of GH12 paralogs

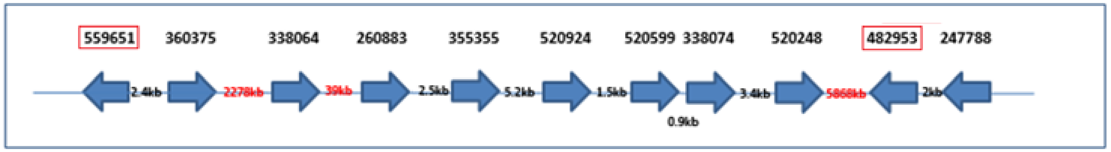


Figure 3.10. For each GH12 gene, FungiDB (Stajich et al., 2012; Basenko et al., 2018) was used to locate genomic coordinates - the relative approximate distances between the paralogs are shown in kilobases (kb). The red boxes highlight the two *P. sojae* paralogs (_559651 and _482953) that were shown to have unique structural features, and notably, these paralogs (with tandem repeat duplicates), are more distantly located to the other GH12 genes.

3.5.6 Evolutionary history of oomycete GH12

A phylogenetic tree was constructed to confirm the evolutionary relationships between *P. sojae* GH12 paralogs (and their orthologs in other oomycetes) (Figure 3.11). The black stars indicate *P. sojae* GH12 paralogs; the red stars indicate GH12 paralogs *P. sojae*_482953 and *P. sojae*_559651, which form distinct clades with orthologous proteins (blue boxes). This is consistent with the unique sequence and structural features of the two paralogs described in this chapter (C-terminal 'tail' of *P. sojae*_482953 and a putative 'second' carbohydrate-binding site of *P. sojae*_559651).

3.5.6 Evolutionary history of oomycete GH12

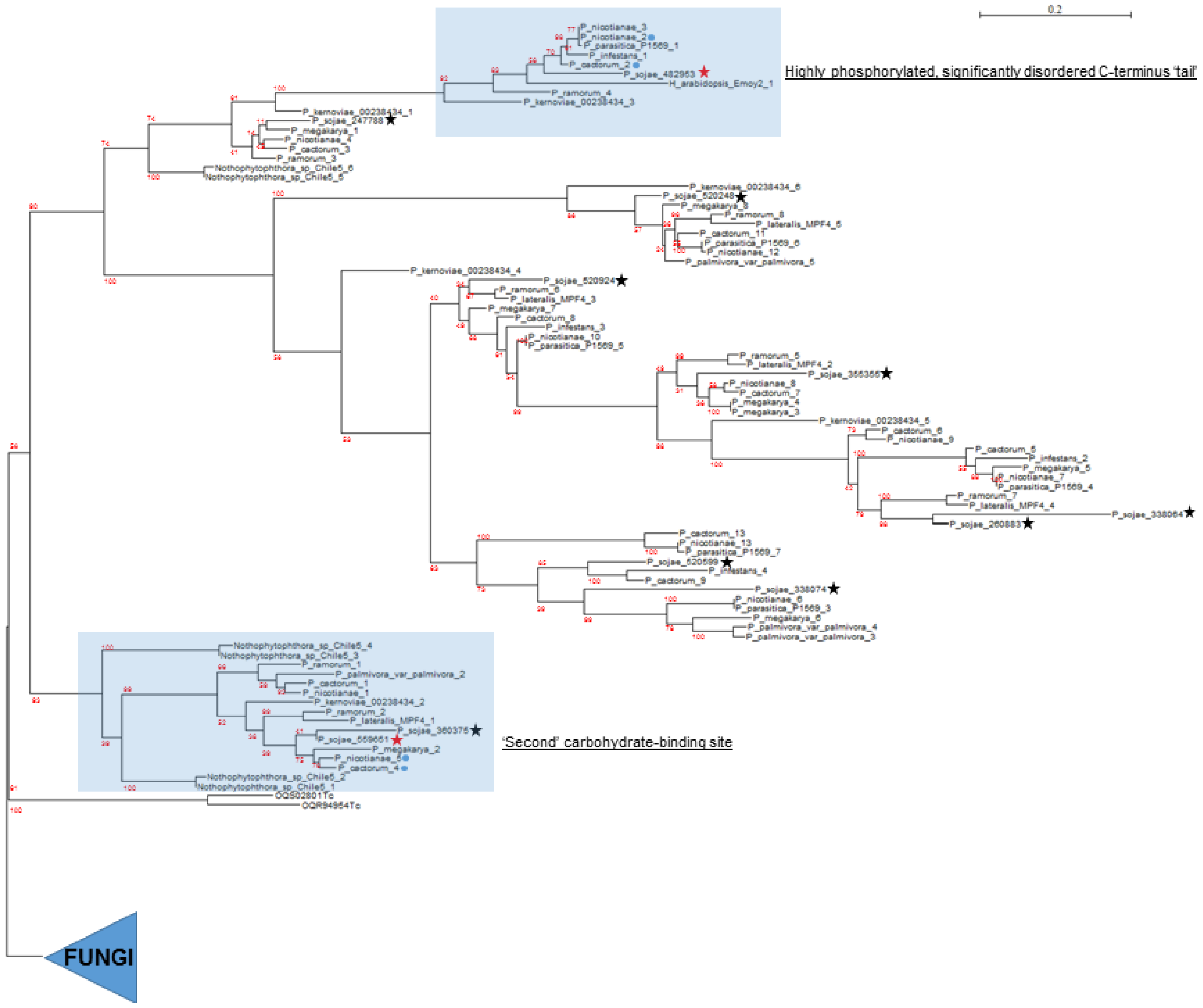


Figure 3.11. Evolutionary history of oomycete GH12. ML tree constructed with IQ-Tree v2.0.3 (WAG+R5 model of evolution (Whelan and Goldman., 2001) selected as the best-fit model by ModelFinder (Kalyaanamoorthy et al., 2017)). The tree was constructed with an alignment of 161 sequences comprising 211 amino acids; node values indicate results of non-parametric bootstrap (200 replicates) (Minh et al., 2020). The tree is rooted with a fungal outgroup because fungi were previously identified as the putative donor group of the GH12 HGT into the oomycetes (Richards et al., 2011; Savory et al., 2015). The black stars indicate *P. sojae* GH12 paralogs; the red stars indicate GH12 paralogs *P. sojae*_482953 and *P. sojae*_559651, which form distinct clades with orthologous proteins (blue boxes; blue dots indicate the orthologs characterised in Figures 3.6, 3.7 and 3.8) - consistent with the unique sequence and structural features of the two paralogs described in this chapter (C-terminal 'tail' of *P. sojae*_482953 and a putative 'second' carbohydrate-binding site of *P. sojae*_559651).

3.5.7 Sequence and structural features of *P. sojae* GH10 paralogs

P. sojae GH10 protein sequences were aligned[6] with Clustal Omega (Larkin et al., 2007; Madeira et al., 2019), using *T. reesei*_Xyn3 (Fungal GH10) as a comparison (Figure 3.12, Table 3.1 for protein alignment key). The *P. sojae* GH10 protein alignment was manually annotated to include predicted 3D structural features (α-helices and β-sheets) (Phyre2; Kelly and Sternberg., 2009), N-glycosylation sites (NetNGlc 1.0; Gupta et al., 2004), phosphorylation sites (serine, threonine and tyrosine) (NetPhos 3.1; Blom et al., 1999), predicted ligand (carbohydrate)-binding sites (3DLigandSite; Wass et al., 2010), as well as glutamic acid (Glu; E) catalytic residues (Figure 3.12).

All *P. sojae* GH10 proteins were found to have conserved both glutamic acid (Glu; E) residues theoretically required for enzymatic activity (shown by the red vertical arrows (Figure 3.12)). Total putative N-glycosylation sites were as follows: *P. sojae*_519234: 3 sites; *P. sojae*_518763: 3 sites; *P. sojae*_527497: 1 site; *P. sojae*_489338: 1 site. Predicted phosphorylation sites (serine, threonine

[6] Whilst 5 sequence homologs for GH10 were identified in *P. sojae* (Figure 3.1), available transcriptome data for the paralogs (FungiDB; Stajich et al., 2012; Basenko et al., 2018) allowed scrutinisation of gene model predictions (using gene expression data for correction). For GH10 gene *P. sojae*_254209, RNA-seq data did not support the prediction of two introns called in the genome-derived sequence, and additionally suggested that a portion of the 5’ sequence was missing (including a putative N-terminal signal sequence, which was not predicted in the genome-derived sequence). Using the gene expression data, within the upstream 5’ sequence, there was an intron of 78bp over-predicted by 1 nucleotide (resulting in a frame shift in the translated protein sequence), in addition to many stop codons preceding the mis-annotated 5’ intron – such that it was not possible to confidently assemble the sequence before 162bp upstream of the original transcriptional start site. As a result, this paralog was omitted from further analysis as a putative pseudogene – the protein sequence is not included in the *P. sojae* alignment (Figure 3.12), however, it’s expression profile and genomic location are included in Figures 3.13 and 3.14, respectively.

and tyrosine) were between 36 and 59 for all paralogs (59 in *P. sojae*_527497, which also has an additional ~63 amino acid ‘tail’ sequence at its C-terminus (Figure 3.12)). Predictions of carbohydrate-binding sites were generated by 3DLigandSite (Wass et al., 2010); all GH10 paralogs except *P. sojae*_527497 possessed a putative additional carbohydrate-binding site – although, predicted amino acid residues for both sites varied in some instances between the protein sequences (Figure 3.12). Whilst *P. sojae*_519234, _527497 and _489338 were predicted to include 12, 11 and 11 amino acids in one binding site (respectively), *P. sojae*_518763 was only predicted to include 5 amino acid residues in the same site (although the site prediction did include both glutamic acid (Glu; E) residues as expected) (Figure 3.12).

3.5.7 Sequence and structural features of *P. sojae* GH10 paralogs

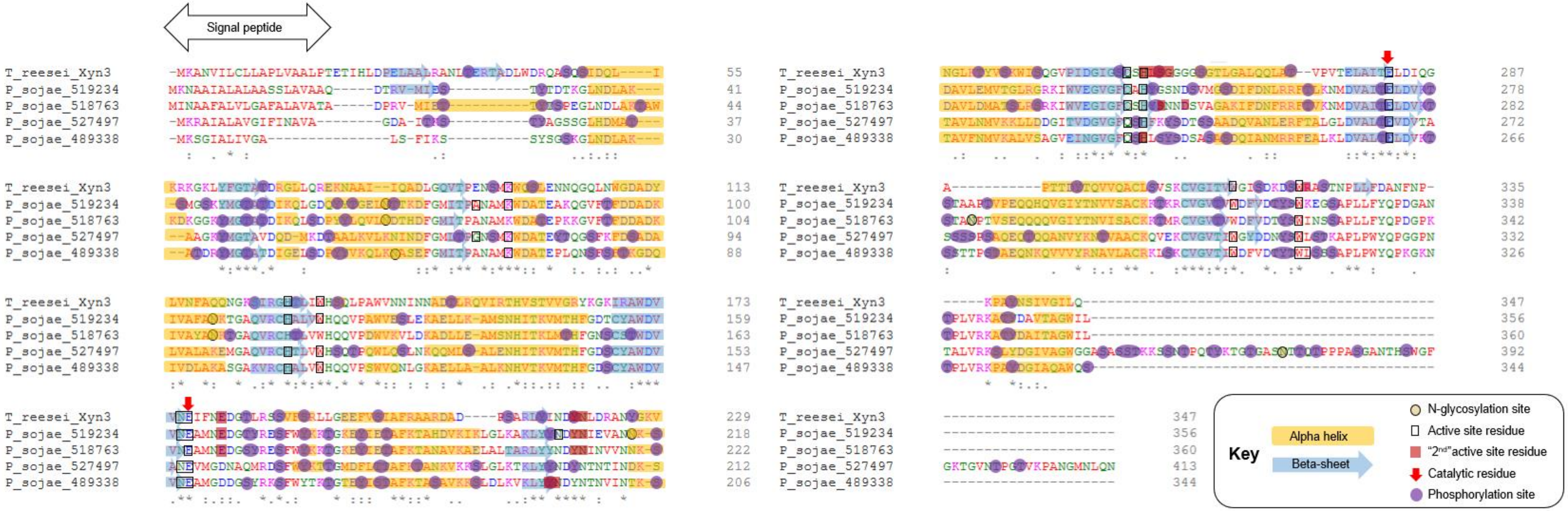


Figure 3.12. Protein alignment of *P. sojae* GH10, generated by Clustal Omega (Larkin et al., 2007; Madeira et al., 2019), aligned with *T. reesei* Xyn3 (fungal) protein for comparison. N-glycosylation sites were predicted with NetNGlc 1.0 (Gupta et al., 2004); three-dimensional structures of paralogous proteins were obtained with Phyre2 (Protein Homology/analogY Recognition Engine v2.0 (Kelly and Sternberg., 2009)); predictions of carbohydrate-binding sites (active site residues) were generated by 3DLigandSite (Wass et al., 2010). All *P. sojae* paralogs possess both glutamic acid residues theoretically required for enzymatic activity (red arrows); *P. sojae*_527497 possesses a long, phosphorylated unique C-terminal 'tail' region; interestingly, carbohydrate-binding site residues were variable amongst paralogs and *P. sojae*_527497 was the only paralog not to predict a 'second' binding site.

3.5.8 Gene expression profiles of GH10 paralogs

For each *P. sojae* GH10 paralog, transcriptome data was compared during three lifecycle stages as previously described (FungiDB; Stajich et al., 2012; Basenko et al., 2018) (Figure 3.13).

Genes coding for *P. sojae*_ 518763 and *P. sojae*_527497 were not shown to be expressed during mycelial growth, in contrast to *P. sojae*_519234 and *P. sojae*_489338 genes - (expressed in Log2(FPKM): 1.93 and 3.22, respectively). All paralogs were shown to be expressed during infection, and interestingly, putative pseudogene (*P. sojae*_254209) is highly upregulated during infection (Figure 3.13).

3.5.9 Genomic locations of GH10 paralogs

For each *P. sojae* GH10 gene, the genome browser tool within FungiDB (Stajich et al., 2012; Basenko et al., 2018) was used to locate the coordinates of the five GH10 genes within the *P. sojae* genome. Approximate distances between each of the GH10 genes is shown in Figure 3.14. Interestingly, *P. sojae*_519234 and *P. sojae*_518763 are arranged ‘head to head’ with 0.3 kb between the genes, whilst *P. sojae*_489338 is located 209 kb upstream of *P. sojae*_519234. Putative pseudogene, *P. sojae*_254209 and *P. sojae*_527497 are also orientated in a ‘head to head’ arrangement, 41 kb from one another (Figure 3.14).

3.5.8 Gene expression profiles of GH10 paralogs

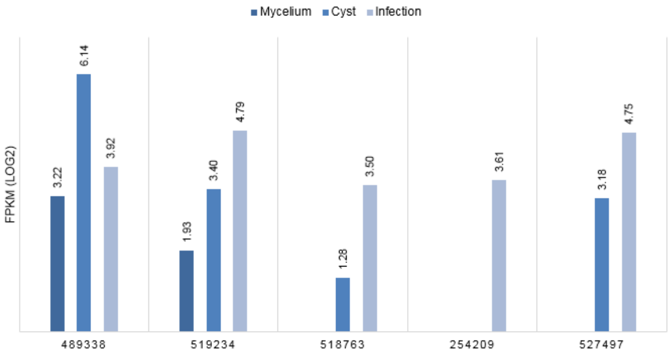


Figure 3.13. For each GH10 gene, *P. sojae* transcriptome data was compared during three lifecycle stages: mycelial, cyst and 3 days post-infection of soybean hypocotyls, expressed as transcript levels of fragments per kilobase of exon model per million mapped reads (FPKM - sequencing depth and gene depth normalised from paired-end RNA-seq data; FungiDB; Stajich et al., 2012; Basenko et al., 2018). FPKM values were used to calculate Log2 (plotted). The data indicates that all paralogs have unique expression profiles during all lifecycle stages, with many upregulated during infection (i.e. close interactions with the plant host and increased capacity for plant cell wall digestion).

3.5.9 Genomic locations of GH10 paralogs

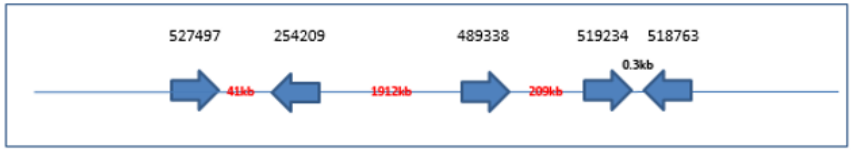


Figure 3.14. For each GH10 gene, FungiDB (Stajich et al., 2012; Basenko et al., 2018) was used to locate genomic coordinates - the relative approximate distances between the paralogs are shown in kilobases (kb). *P. sojae*_519234 and *P. sojae*_518763 are arranged 'head to head' with 0.3 kb between the genes, whilst *P. sojae*_489338 is located 209 kb upstream of *P. sojae*_519234. Putative pseudogene, *P. sojae*_254209 and *P. sojae*_527497 are also orientated in a 'head to head' arrangement, 41 kb from one another.

3.6 Discussion

3.6.1 HGT families associated with plant cell wall degradation have been largely retained by hemibiotrophic oomycetes

Previously reported HGT events from fungi to oomycetes include putatively secreted enzymes predicted to degrade plant cell wall-specific substrates (Torto et al., 2002; Belbahri et al., 2008; Richards et al., 2011; Savory et al., 2015). Eleven of the HGT events are absent in *H. catenoides* (the sister group to the oomycetes), and are largely expanded in hemibiotrophic oomycetes in the *Phytophthora* genus (Figure 3.1) - suggesting HGT has had a lifestyle-specific influence to oomycete evolution. Whilst the HGTs appear to have been lost by some hemibiotrophs, the hypothesis is consistent with the detection of fewer HGTs in obligate biotrophs (*P. halstedii* and *H. arabidopsidis*), as well as in necrotrophs, saprotrophs, and non-pathogenic oomycetes - although this could reflect a bias in genome sampling for the original analysis (i.e. predicted proteomes were selected from *P. ramorum*, *P. sojae*, *P. infestans* and *H. parasitica* - therefore limiting identification of putative HGTs to these taxa only (Richards et al., 2011)). *Pythium vexans* has been renamed *Phytopythium vexans*, an intermediate between *Phytophthora* and *Pythium* (de Cock et al., 2015), which correlates with the distribution of the HGTs in this organism (Figure 3.1).

Previous studies indicate that obligate biotrophs express and secrete fewer CAZymes than hemibiotrophs (Baxter et al., 2010; Kemen et al., 2011) – indicative of their requirement to maintain an intimate association with their host without stimulating an immune response from cell wall maceration (presumably the host association having a greater importance than increased carbohydrate availability for biotrophs). Also consistent with observations of this study are previous genome analyses of *Phytophthora* spp., which have shown that they generally encode more GH genes than *Pythium* spp. (e.g. Zerillo et al., 2013) – in part due to the significant expansion (by gene duplication) of GH families in *Phytophthora* species (Levesque et al., 2010; Zerillo et al., 2013), but also likely due to differences in the enzyme families expressed by necrotrophic pathogens. *P. sojae* (95 Mb genome) has been previously annotated to encode a total of 309 GH enzymes (Tyler et al., 2006), whilst *P. ultimum* (42 Mb genome), a total of 183 GH enzymes (Levesque et al., 2010) - higher selection pressures for the HGT genes (and other CAZymes) in *Phytophthora* spp. and their rapid evolution (i.e. through subsequent gene duplication events leading to multiple paralogs, inflating the total genome size) reflect their importance for phytopathogenicity in hemibiotrophic oomycetes – for more efficient substrate breakdown, or a means to evade host defences.

Interestingly, a HGT-acquired GH88 (encoding a putative α-L-rhamnosidase), putatively involved in pectin degradation was only confirmed to be present in hemibiotrophic *Phytophthora* spp., and GH28 (encoding a putative pectin hydrolase) was only confirmed in genomes of *Peronosporales*, i.e. hemibiotrophs and obligate biotrophs (Figure 3.1). GH78 was not found encoded in either obligate biotroph included in the analysis – again, with a putative function in the degradation of pectin (D-4,5-unsaturated beta-glucuronyl hydrolase). As pectin is the major non-hemicellulosic polysaccharide of plant cell walls, formed from a backbone of D-galacturonic acid (Atmodjo et al., 2013) rather than D-glucose, it represents an important structural component for plant parasites to overcome - in addition to the cellulose and hemicellulose layers. Therefore, the retention of pectin-specific HGT genes largely by hemibiotrophic oomycetes indicates the importance of these enzymes for efficient and complete cell wall digestion. Furthermore, the transferred genes likely contribute to an arsenal of alternative pectinase genes in these organisms, although that was not the focus of this work. As pectin degradation products are also important elicitors of plant immune responses (Hahn et al., 1981; Ferrari et al., 2013), it is feasible for phytopathogens to secrete a battery of digestive enzymes as a major mode of attack to counteract host defence (by driving complete and rapid breakdown of the substrate).

P. sojae is currently one of the model organisms for the *Phytophthora* genus, and the selective benefit of maintaining the HGTs is further suggested by widespread gene expansion (by duplication) following acquisition. For each HGT paralog in *P. sojae* (110 in total, across all HGT families) (Figure 3.1), this has putatively contributed to >6% of the *P. sojae* predicted secretome (calculated from 1659 total secreted proteins predicted in McGowan and Fitzpatrick (2017)) (>6% of 1756 total secreted proteins predicted in Adhikari et al. (2013), >6% of 1586 total secreted proteins predicted in Richards et al. (2011), >7% of 1464 total secreted proteins predicted in Tyler et al. (2006)).

3.6.2 There is diversity and expansion of cellulose- and xylan-degrading activities among hemibiotrophic oomycetes

To identify other GH families present in *P. sojae* putatively involved in cellulose (GH12) and xylan (GH10) metabolism, the CAZy database was searched, identifying GH5, 6, 7 and 17 (for GH12) and GH5, 30 and 43 (for GH10) (http://www.cazy.org; Lombard et al., 2013). Interestingly, GH6 has been previously published as a HGT into the oomycetes and proposed to have a bacterial origin (Misner et al., 2015; Savory et al., 2015), and GH43 is a fungal-oomycete HGT putatively involved in the breakdown of arabinan in hemicellulose layers (Richards et al., 2011; Savory et al., 2015).

The presence and expansion of GH families putatively annotated for cellulose and xylan breakdown across oomycetes was investigated (Figure 3.2, Figure 3.3). GH5 is one of the largest and most widespread GH families present in archaea, bacteria and eukaryotes, and is associated with a significant variety of activities, including those not involved in cellulose and xylan degradation (http://www.cazy.org; Lombard et al., 2013). As it was not possible to confidently assign carbohydrate-degrading functions to all predicted proteins for GH5, for clarity it is not included in the figures. Interestingly, GH5 enzymes of *Phytophthora* spp. have been cloned and investigated previously, but substrate specificity was not investigated (McLeod et al., 2003; Costanzo et al., 2007).

Consistent with the distribution of the eleven HGT events across the oomycetes (Figure 3.1), higher numbers of CAZymes associated with the activities of horizontally-acquired GH12 were identified in the genomes of hemibiotrophic oomycetes in the *Phytophthora* genus (Figure 3.2). Interestingly, of the necrotrophic oomycetes sampled, *P. arrhenomanes* was the only *Pythium* species found to encode horizontally-transferred GH12 (Figure 3.1) - initially suggesting reduced secreted cellulase activity within the genus. However, when investigating the wider GH families putatively associated with the same activities, *P. vexans*, *P. irregulare*, *P. iwayami* and *P. aphanidermatum* were all confirmed to encode paralogs of GH6, 7 and 17 (in reduced paralog numbers as compared with *Phytophthora* spp.) (Figure 3.2).

None of the oomycetes sampled from the order *Saprolegniales* were found to encode GH7 or GH12 enzymes by the methods used in this study, however, high paralog numbers of GH6 were identified amongst aquatic parasites *S. diclina* VS20 and *S. parasitica* CBS 223.65 (Figure 3.2). Whilst removal of GH5 enzymes from the analysis aimed to limit the inclusion of enzymes not involved in cellulose degradation from plant host substrates, it is important to note that putative cellulose-digesting enzymes of the GH families identified may not always be host-targeted, and could play a role in remodelling the oomycete's own cell wall (which is rich in cellulose (Bartnicki-Garcia., 1968; Grenville-Briggs et al., 2008)). In the absence of experimental evidence, it is unclear how the wider GH families identified are involved in cellulose digestion. That said, cellulases have also been identified from (for example) *Arabidopsis* (Williamson et al., 2002) and crayfish (Bryne et al., 1999), suggesting they have diverse roles across the tree of life.

As with GH12-associated enzymes, higher total numbers of CAZymes associated with the activities of GH10 enzymes were identified amongst hemibiotrophic oomycetes – with comparatively lower total enzymes amongst obligate biotrophs (Figure 3.3). None of the oomycetes sampled from the order *Saprolegniales* were found to encode GH43 or horizontally-transferred GH10, and only 2-4 paralogs of xylanase-associated CAZymes were identified in these organisms (only associated with GH30) (Figure 3.3).

Overall, the data is consistent with previous studies indicating significant expansion of GH enzymes in *Phytophthora* species (e.g. Adhikari., 2013; McGowan and Fitzpatrick., 2017). Although many oomycete genomes encode gene families putatively involved in cellulose and xylan metabolism, it is likely that the diversity in enzyme families, as well as their expansion by subsequent gene duplications in *Phytophthora* species, correlates with the importance of the specific functions for hemibiotrophic lifestyles – functions that HGT and gene duplication are hypothesised to have contributed to. It is therefore important to better understand the functional significance of the paralogous proteins in hemibiotrophic oomycetes.

3.6.3.1 Ten out of eleven *P. sojae* GH12 paralogs possess both catalytic amino acids theoretically required for enzymatic activity

The protein sequence of *T. reesei*_Cel12a was used for comparison of *P. sojae* GH12 sequence and structural features, because its structure has already been elucidated by Sandgren et al. (2001) using x-ray crystallography (Sandgren et al., 2001), whereas no crystal structure of a *P. sojae* GH12 is currently available.

T. *reesei*_Cel12a is made up of 15 β-strands that fold into two twisted, largely anti-parallel β-sheets that pack on top of one another (Sandgren et al., 2001). The only two cysteine residues in *T. reesei*_Cel12a are also conserved amongst the *P. sojae* GH12 paralogs (Cys4 and Cys32 for *T. reesei* (numbers not including the N-terminal signal sequence) (Figure 3.4) – these are proposed to form a disulphide bridge between two of the β-strands of the first β-sheet (Sandgren et al., 2001).

Interestingly, Sandgren et al. (2003) mutated Ala35 in *T. reesei*_Cel12a to Val35 and recorded an increase in thermal stability of the protein of 7.7°C, and the authors later went on to crystallise the structure of this variant (Sandgren et al., 2003). The alanine residue at this position is not conserved amongst *P. sojae* GH12 proteins (Figure 3.4), however, a valine residue at the position for *P. sojae*_260883, _355355, _559651 and _360375 (whilst would require further testing), could be a target for modification to study improved thermostability of the *P. sojae* paralogs.

Ten of the *P. sojae* GH12 paralogs possess both glutamic acid (Glu, E) residues required for catalytic activity for this family (Figure 3.4) (Okada et al., 2000). In *T. reesei*_Cel12a, the catalytic nucleophile Glu116 is in close proximity to residues Asp99 and Met118, which are widely conserved amongst GH12 proteins (including the *P. sojae* GH12 paralogs (Figure 3.4)).

The protein sequence of *P. sojae*_360375 is missing a significant portion of its C-terminus (including the second conserved glutamic acid residue) – previous *in vivo* work by Ma et al. (2017) demonstrated that this *P. sojae* protein transiently expressed in *N. bethamiana* leaves does not result in hydrolytic activity (although, the authors note that the preferred substrates could have been absent in the assay; Ma et al., 2017) - which is consistent with both glutamic acid residues being required for catalytic activity. Interestingly, despite loss of enzymatic activity, the researchers demonstrate that this protein has a strong binding affinity to a host immune protein (named GmGIP1) (Ma et al., 2017) – demonstrating a role for non-active isozymes in *P. sojae* virulence as putative ‘stealth factors’. Strong selection pressures resulting from intimate pathogen-plant interactions suggests that evolution of multiple enzyme paralogs (whilst many are likely to contribute to the overall function through transcriptional redundancy and/or neo/sub-functionalisation (Ohno., 1970; Stoltzfus., 1999; Force et al., 1999; Long et al., 2003)), are also important for subverting or exploiting host defences - whether binding directly to host immune proteins to free active paralogs from targeting (as shown by Ma et al., 2017), or even (hypothetically) binding plant cell wall degradation products (i.e. oligosaccharides, di- or monosaccharide units), in order to limit activation of host immune receptors. Whilst this work aimed to characterise multiple enzyme paralogs on the basis of evolved functions gained by oomycete parasites through a combination of HGT and subsequent gene duplication events, it is also important to consider how both phenomena have additionally contributed to host-pathogen co-evolution as well.

3.6.3.2 *P. sojae* paralog _559651 is predicted to have a 'second' carbohydrate-binding site that could alter enzymatic activity

Paralogous enzymes with conserved catalytic sites could have unique sequence or structural features that are important for their activity and/or their interactions with substrates. Predictions of carbohydrate-binding sites (active site residues) for the GH12 paralogs were generated by 3DLigandSite (Wass et al., 2010), and suggested that *P. sojae*_559651 was a unique GH12 paralog in which a 'second' carbohydrate binding site was predicted (Figure 3.5). The putative amino acid residues for this site are Gly55, Ala56, Ala57, Thr58, Val97, Phe205, Val206 (residue position numbers given for the protein sequence in the absence of its N-terminal signal peptide) – Thr58 is the common residue predicted in both ligand sites for this paralog. Orthologous proteins in *P. cactorum* and *P. nicotiniae* were also predicted to have a ‘second’ carbohydrate-binding site (Figure 3.6), and interestingly, two putative indels coding for alanine (Ala, A) and Serine (Ser, S) that are conserved amongst the three sequences (Figure 3.7), are important for the ‘second’ binding site prediction. Removal of both amino acids abolished the prediction of the binding site for the orthologous proteins (whilst the binding site common amongst all GH12 paralogs was left intact). The two indels are also conserved in *P. sojae*_360375, however, as this paralog is significantly truncated at the C-terminus (see Figure 3.4), a second putative binding site was not originally predicted by 3DLigandSite (Wass et al., 2010) (Figure 3.5). Whilst further experimental characterisation of *P. sojae*_559651 would be required to better understand the functional significance of an additional (putative) ligand-binding pocket in the context of plant carbohydrate degradation, it is feasible that this isoenzyme interacts with its substrate by a novel mechanism in the family – putatively able to bind *more* of the cellulose backbone, or putatively able to digest its substrates more efficiently. Sandgren et al. (2001) note that the *T. reesei*_Cel12a cellulose-binding site putatively binds at least six glucose residues (Sandgren et al., 2001), so it would be interesting to investigate how this compares with the *P. sojae* proteins.

3.6.3.3 *P. sojae* paralog _482953, and orthologs in *P. cactorum* and *P. nicotiniae*, have a significantly disordered, highly phosphorylated C-terminus ‘tail’ that could alter enzymatic activity

As previously mentioned, isozymes of the same enzyme could have evolved under strong selection pressures driving unique activities with possibly untested adaptive consequences. Another interesting feature of two of the *P. sojae* GH12 paralogs (_482953 and _247788), is that they both possess long, significantly phosphorylated C-terminal tails (Figure 3.4). *P. sojae*_482953 was taken forward for further analysis and (in addition to its orthologs in *P. cactorum* and *P. nicotianae*) the C-terminal tail regions of each protein were unable to be accurately modelled to a protein structure using Phyre2 (Kelly and Sternberg., 2009). Amounts of disorder for the protein sequences (calculated by Phyre2) were 57% (*P. sojae*_482953), 51% (*P. cactorum*_2), and 64% (*P. nicotianae*_2) - the disordered sequences are shown in Figure 3.9. It is currently unclear how the 186 amino acid 'tail' of *P. sojae*_482953 affects the proteins function, however it is conceivable that it could affect (for example) binding affinity, catalytic activity, or tolerance across biological conditions (e.g. temperature, pH).

The C-terminus tail of *P. sojae*_482953 is highly phosphorylated (serine and threonine (Figure 3.4; Figure 3.9)). Interestingly, other families of cellulases with CBMs are usually connected to the CBM by a flexible linker sequence – also rich in serines, threonines and prolines, but highly glycosylated (Harrison et al., 1998). Li et al (2014) also describe a C-terminal proline-rich sequence of xylanase XynA, and its removal affects the protein function (Li et al., 2014), and Wen et al (2005) describe a truncated glucanase that displays improved enzymatic activity (Wen et al., 2005). It is therefore possible that the *P. sojae*_482953 tail region has a positive or negative effect to the overall function of the enzyme – which would require further experimental investigation to elucidate.

3.6.4 *P. sojae* GH12 paralogs have unique RNA-seq profiles during three different stages of the organism’s lifecycle

The transcription of multiple isozymes of the same enzyme may be independently regulated and expressed in different relative amounts, therefore *P. sojae* transcriptome data was compared for all GH12 paralogs. Interestingly, *P. sojae*_247788 and *P. sojae*_520599 did not appear to be expressed under any of the conditions tested, suggesting these paralogs could be putatively inactive (Figure 3.9). *P. sojae*_247788 is most closely related to *P. sojae*_482953; both possess long, significantly disordered C-terminal tails (as previously described; Figure 3.4), however of the two, *P. sojae*_482953 is the only paralog expressed during mycelial, cyst and 3 days post-infection (soybean hypocotyls infected with *P. sojae* strain P6497) (FungiDB; Stajich et al., 2012; Basenko et al., 2018) (Figure 3.9) - suggestive of a functional role of this paralog with this particular sequence feature in this organism. In contrast, there is currently no evidence for expression of *P. sojae*_247788 during any of these phases.

For the remaining GH12 paralogs, all show expression during infection (*P. sojae*_338074 and *P. sojae*_520248 are exclusively expressed during infection only) (Figure 3.9), suggestive of an important role during *P. sojae* interactions with its host. It is unclear how GH12 expression is induced during this stage, however, it is possible that other functional paralogs constitutively expressed recognise the cellulose backbone, and early enzymatic cleavage of the chain generates oligosaccharides that further induces GH12 expression. Similarly, when *T. reesei* is given cellulose as the sole carbon source, multiple cellulase genes are induced, suggestive of recognition in response to the presence of cellulose (Kubicek and Penttila., 1998). Interestingly, growth on glucose as the sole carbon source (i.e. the monomer that makes up the cellulose chain), represses cellulase expression in *T. reesei* (Ilmen et al., 1997), but varying lengths of oligosaccharides are known to induce cellulase expression in *T. reesei* (Sternberg and Mandels., 1980). Ma et al. (2015) demonstrated expression of the gene encoding *P. sojae*_559651 at 12 time points during early infection (up to 48 hours), and showed high expression up to 2 hours followed by a decline – it would be interesting to know how the other functional paralogs compare to this pattern of expression during very early points of infection (Ma et al., 2015).

3.6.5 *P. sojae* paralogs _559651 and _482953 have distinctive structural features, and unique genomic locations with tandem-repeat paralogs

Genomic locations of the *P. sojae* GH12 paralogs were identified using an available FungiDB genome browsing tool (Stajich et al., 2012; Basenko et al., 2018). Whilst paralogs _338064, _260883, _355355, _520924, _520599, _338074 and _520248 were found largely clustered together (Figure 3.10), tandem repeat pairs _559651 and _360375, and _482953 and _247788 were found located 2278 kb and 5868 kb upstream and downstream of the clustered genes, respectively. Interestingly, *P. sojae*_559651 and *P. sojae*_482953 possess unique structural features, as previously described.

3.6.6 Evolutionary history of oomycete GH12

To confirm the evolutionary relationships between *P. sojae* GH12 paralogous protein sequences, and their orthologs amongst oomycete taxa, a phylogenetic tree was constructed on the basis of multiple sequence alignment of the proteins (Figure 3.11). *P. sojae*_482953 and *P. sojae*_559651 form distinct clades with orthologous proteins (blue dots highlight the orthologs characterised in this chapter), consistent with the observed unique sequence and structural features of the paralogs. As previously mentioned, *P. sojae*_247788 (most closely related to *P. sojae*_482953) also possesses a significantly phosphorylated C-terminal tail (Figure 3.4); however, this protein and orthologs form a clade distinct to that containing *P. sojae*_482953 (Figure 3.11). Interestingly, there is no evidence that *P. sojae*_247788 is expressed during *P. sojae* life stages (FungiDB; Stajich et al., 2012; Basenko et al., 2018) (Figure 3.9), so it is intriguing to consider whether the evolved 'tail' sequence of *P. sojae*_247788 could have a negative or toxic effect to protein function – this would be interesting to explore further through experiments involving the full-length and truncated proteins.

3.6.7 Sequence and structural features of *P. sojae* GH10 paralogs

All *P. sojae* GH10 protein sequences were found to possess both conserved glutamic acid (Glu; E) residues theoretically required for enzymatic activity (shown by the red vertical arrows (Figure 3.12)). Both glutamic acid residues were also predicted to form part of one (of two) predicted carbohydrate-binding sites amongst the GH10 proteins (predictions generated by 3DLigandSite (Wass et al., 2010)). Interestingly, predicted amino acid residues for both putative binding sites varied between the protein sequences (Figure 3.12), suggesting that (despite conservation of amino acids in the protein alignment), the respective side chains within the binding pocket of the assembled proteins could be differentially exposed between the paralogs. *P. sojae*_527497 was not predicted to possess an additional binding site, however, this paralog has an additional ~63 amino acid phosphorylated 'tail' sequence at its C-terminus (Figure 3.12)) – which, could alter its interaction with putative substrates.

3.6.8 Gene expression profiles of GH10 paralogs

As for GH12, *P. sojae* transcriptome data was additionally compared for GH10 genes. As seen for some GH12 members, two of the GH10 genes were not shown to be expressed during mycelial growth (*P. sojae*_518763 and *P. sojae*_ 527497) (Figure 3.13). As mentioned previously, it is possible that constitutively-expressed paralogs (in this case, *P. sojae*_519234 and *P. sojae*_489338) could play an important role in substrate recognition during *P. sojae* infection (which would benefit from further experimentation to identify the importance of the genes in expression signalling). All GH10 genes are expressed during infection, interestingly, including putative pseudogene *P. sojae*_254209, which is not induced at any other stage of *P. sojae* growth (Figure 3.13). It could be possible that this gene also plays some sort of role in regulation of the gene family (if indeed it is non-functional - however, as discussed earlier in the chapter, it was not possible to confidently assign a nucleotide sequence, therefore it is not included in the experimental analysis of the paralogs in Chapter 4).

3.6.9 Genomic locations of GH10 paralogs

Genomic locations of the *P. sojae* GH10 paralogs were identified using an available FungiDB genome browsing tool (Stajich et al., 2012; Basenko et al., 2018). Interestingly, *P. sojae*_519234 and *P. sojae*_518763 are arranged 'tail to tail' (as are orthologs in *P. parasitica* (Lai et al., 2018), with just 0.3 kb between the genes – though the two have varying levels of transcription (Figure 3.13). Putative pseudogene, *P. sojae*_254209 and *P. sojae*_527497 are also orientated in a 'tail to tail' arrangement, but 41 kb from one another (Figure 3.14) – between the two are 17 genes including a predicted ABC transporter and two ferric reductases (FungiDB; Stajich et al., 2012; Basenko et al., 2018).

3.7 General Conclusion

The genomes of plant parasites are abundant in genes encoding secreted digestive enzymes to break down the plant cell wall, of particular interest to this work are previously identified HGTs into oomycetes from fungi (Torto et al., 2002; Richards et al., 2006; Belbahri et al., 2008; Richards et al., 2011; Misner et al., 2015; Savory et al., 2015). This chapter aimed to re-confirm eleven HGT events and identify the total numbers of paralogs across oomycetes with diverse ecological lifestyles; the results suggest that the HGTs have been largely retained by hemibiotrophic oomycetes, and widespread gene duplication amongst transferred gene families indicates strong selection for the maintenance of the genes. However, knowledge is limited in understanding the functional significance of the paralogous proteins for the degradation of plant carbohydrates in phytopathogenic oomycetes - therefore, this chapter also aimed to investigate *P. sojae* GH12 and GH10 paralogs using bioinformatics and available computational methods. Interestingly, GH12 paralog *P. sojae*_482953 (and orthologs in *P. cactorum* and *P. nicotiniae*) were found to possess a significantly disordered and phosphorylated C-terminal ‘tail’ which could putatively alter the activity of this paralog (or its interactions with substrates) *in vivo*. Another GH12 paralog, *P. sojae*_559651 (and orthologs in *P. cactorum* and *P. nicotiniae*) were predicted to encode a putative ‘second’ substrate binding site, and interestingly, two putative indels coding for alanine and serine are important for this binding site prediction (their removal from amino acid sequences abolishes the prediction). Further experimental characterisation will be crucial to further understand such putative functional differences between paralogs of horizontally-acquired enzymes for plant cell wall degradation.

Chapter 4

Using experimental methods to investigate functional characteristics of *P. sojae* GH12 and GH10 paralogs putatively involved in plant cell wall degradation

4.1 Overview

Previously identified fungal-oomycete HGTs encoding putative secreted enzymes for plant cell wall degradation (Torto et al., 2002; Richards et al., 2006; Belbahri et al., 2008; Richards et al., 2011; Misner et al., 2015; Savory et al., 2015) show evidence of gene duplication (post-acquisition) - most significantly among hemibiotrophic *Phytophthora* spp. (as explored in Chapter 3). Secretion of multiple paralogs of GH12 and GH10 proteins in *P. sojae* is suggestive of important roles in transcriptional dosage and/or represent functional divergence – it is therefore hypothesised that some duplicates may be functional, some non-functional, and some may have evolved new or additional functions (i.e. those not encoded by the original HGT (e.g. evidence of neofunctionalization (Ohno., 1970) or subfunctionalization (Stoltzfus., 1999; Force et al., 1999)). Chapter 3 aimed to explore this hypothesis through the use of bioinformatics and available computational methods to characterise putative functional differences between *P. sojae* GH12 and GH10 paralogs. Of particular interest were two GH12 paralogs (*P. sojae*_482953 and *P. sojae*_559651), which, along with their orthologs in other hemibiotrophic *Phytophthora* spp., possess unique structural features (a significantly disordered C-terminus ‘tail’ extension, and a putative ‘second’ substrate binding site, respectively), which could be responsible for altered biological activity *in vivo*.

The current chapter aims to experimentally characterise *P. sojae* HGT paralogs through heterologous gene expression and secretion of the proteins in eukaryotic host, *S. cerevisiae*. Yeast culture supernatants will be used as crude protein extracts for enzyme activity assays across different biological conditions (temperature and pH), as well as for analysis of oligosaccharide degradation products by mass spectrometry, in order to explore putative functional differences between the paralogs. A truncated version of GH12 paralog *P. sojae*_482953 will be engineered and expressed in *S. cerevisiae* BY4742, to better understand the importance of the 186 amino acid C-terminal ‘tail’ (see Chapter 3) for the proteins’ function, as well as providing useful insights to paralog evolution (i.e. the significance of gene duplication and mutation for the evolution of protein function). Additionally, *P. sojae* genome editing methods developed by Fang and Tyler (2016; 2017) will be used to *knock-out* the gene encoding enzymatically-active *P. sojae*_482953 (the full-length protein) *in vivo*[7], to better understand the functional significance of individual GH12 paralogs for carbohydrate utilisation in phytopathogenic oomycetes (as well as establishing another *P. sojae* mutant for the research field for future biological experiments).

[7] Generation of the *P. sojae* mutants was carried out in the research group of Professor Joseph Heitman (Duke University, North Carolina), using methods implemented in this organism (Fang and Tyler., 2016); Fang et al., 2017), Thanks to Professor Heitman and Dr Fang for their support during this part of the project.

4.2 Introduction

4.2.1 Catalytic mechanism of GH12 endo-β-1,4-glucanases

CAZymes assigned to GH12 are currently associated with endo-β-1,4-glucanase (EC 3.2.1.4), xyloglucan endo-hydrolase (EC 3.2.1.151), and endo-β-1,3-1,4-glucanase (EC 3.2.1.73) activities - with the assignment of proteins based on similarity with resolved structures and characterised functions of family members (http://www.cazy.org; Lombard et al., 2013). The first X-ray crystallography structure of a cellulase (*T. reesei* Cel6A (PDB: 3CBH)) was elucidated in 1990 (Rouvinen et al., 1990), and the first structure of a GH12 enzyme (bacterial *Streptomyces lividans* CelB2 (PDB: 1NLR)) was elucidated in 1997 (Sulzenbacher et al., 1997) - since then, many more cellulase structures have been published for a wide range of microorganisms, enabling their catalytic mechanisms to be elucidated.

Endo-β-1,4-glucanases hydrolyse β-1,4-glycosidic bonds between the glucose residues of cellulose polymers – hydrolysis can be achieved via two distinct mechanisms (known as *retaining* or *inverting* mechanisms, which are distinguished from one another by the resulting configuration of the anomeric carbon (C1) of the oligosaccharide following hydrolysis). Nuclear Magnetic Resonance (NMR) studies, including those of endoglucanase 3 from *Humicola insolens* (Schou et al., 1993), have shown that GH12 enzymes use a *retaining* mechanism during digestion, i.e. the C1 configuration is retained following the hydrolysis (Koshland., 1953; Sinnott., 1990; Davies and Henrissat., 1995) (Figure 4.1). The retaining reaction follows a so called ‘classical Koshland mechanism’, in which there are two-steps (glycosylation and deglycosylation), involving double displacement (i.e. the hydrolysis of the glycosidic bond creates a product with the same configuration at C1 that the substrate had before the hydrolysis). For GH12 enzymes, the mechanism involves two glutamic acid (Glu, E) residues, located ~5.5 Å apart – usually localised at opposite ends of the carbohydrate-binding pocket of the secreted protein (McCarter and Withers., 1994) (Also see Chapter 3, Figure 3.5, for localisation of Glu residues in *P. sojae* GH12 predicted protein structures (indicated by the red arrows)). During the glycosylation step, one of the glutamic acid residues acts as a catalytic nucleophile (attacking the C1 of the substrate), and the other acts as a general acid catalyst (protonating the oxygen of the next glucose residue in the chain as the bond cleaves) – this generates a glycosyl enzyme intermediate. The intermediate is then hydrolysed by water during the deglycosylation step - the other glutamic acid residue now acts as a general base catalyst (removing a proton from the water molecule as it attacks). The resulting C1 that is no longer linked to the next glucose residue (but has a free hydroxyl group) is now located at the supposed 'reducing' end of the carbohydrate chain (Figure 4.1). Conversely, *Inverting* enzymes typically function via a one-step, single displacement mechanism (Koshland., 1953), but will not be discussed further in this thesis.

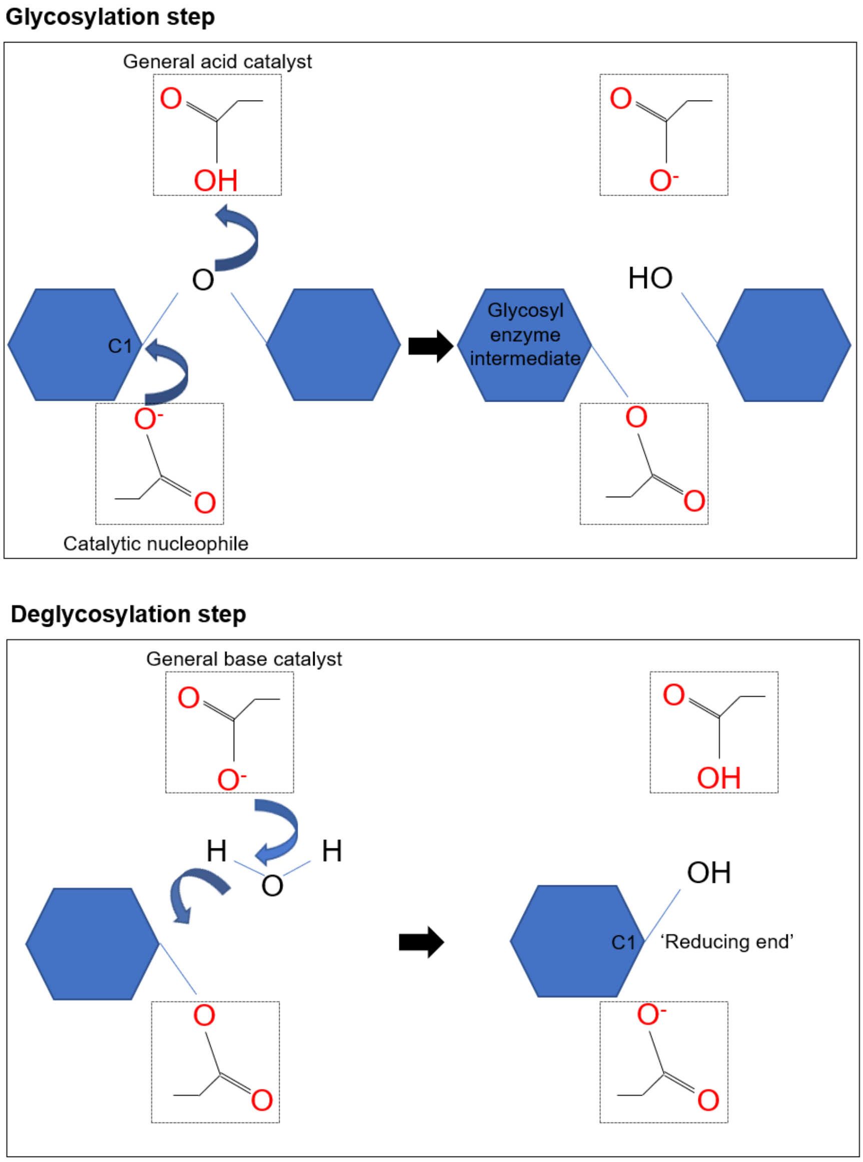


Figure 4.1. Retaining mechanism (glycosylation and deglycosylation) involving a double displacement: one of the catalytic residues acts as a catalytic nucleophile (attacking the C1 of the substrate), whilst the other catalytic residue acts as a general acid catalyst (protonating the oxygen of the next glucose residue). The glycosyl enzyme intermediate is hydrolysed by water: one of the catalytic residues now acts as a general base catalyst (removing a proton from the water molecule). The C1 has a free hydroxyl group and is known as the 'reducing end'. Figure adapted from Sandgren (2003).

4.2.2 Xyloglucan-specific endo-β-1,4-glucanases

Many enzymes assigned to GH12 possess xyloglucan endo-hydrolase activity (EC 3.2.1.151) (http://www.cazy.org; Lombard et al., 2013). This activity is also associated with GH5 and GH74 CAZyme families, the latter of which includes well characterised bacterial and fungal members (e.g. Hasper et al., 2002; Yaoi et al., 2004; Yaoi et al., 2005), and some crystal structures (e.g. *Clostridium thermocellum* (PDB: 2CN2) (Fleites et al., 2006)). Xyloglucan endo-hydrolases are encoded across the tree of life in bacteria, plants and fungi, with many demonstrating hydrolase activity towards unbranched polymers (Sandgren et al., 2005; Sinnott et al., 1990). Interestingly, studies including those by Grishutin et al. (2004) show that xyloglucanases from *Aspergillus japonicas*, *Chrysosporium luckenowense* and *T. reesei* have high specific activities towards xyloglucan (backbone of β-1,4 linkages), but low to zero activity against barley β-glucan (β-1,4 and β-1,3 linkages), and carboxymethylcellulose (CMC; an artificially derived cellulose with a backbone of β-1,4 linkages) (Grishutin et al., 2004).

As previously mentioned, plant xyloglucans are complex and heterogeneous in structure; they have a backbone of β-1,4-linked glucose residues (like that of cellulose polymers), but most of the residues are substituted with side chains of α-1,6-linked xylose, which may in turn have substitutions of α-linked arabinose or fucose, or β-linked galactose attached. In *Arabidopsis thaliana* (and other vascular plants), every fourth glucose residue in the xyloglucan backbone is unsubstituted (Vincken et al., 1997). The nomenclature system introduced by Fry et al. (1993) assigns letters to the possible xyloglucan side chain variants, therefore it is possible to describe the oligosaccharides released from xyloglucan breakdown by xyloglucan endoglucanases (Fry et al., 1993). With no other hydrolase activity, the result of endoglucanase treatment of xyloglucan is the release of oligosaccharides with unsubstituted glucose residues at their reducing end (the end of the chain with an anomeric carbon (C1) not linked to another glucose residue). For example, endoglucanase digestion of *Arabidopsis* spp. xyloglucan releases the oligosaccharides XXG, GXXG, XXXG, XXLG, XLXG, XLLG, XXFG, and XLFG (Madson et al., 2003; Obel et al., 2009). **G** refers to an unsubstituted glucose residue, **X** refers to a glucose residue substituted with an α-linked xylose, **L** refers to a glucose residue substituted with an α-linked xylose further substituted with a β-linked arabinose, and **F** refers to a glucose residue substituted with an α-linked xylose further substituted with a fucose residue (Fry et al., 1993) (Table 4.1).

G	Unsubstituted glucose residue
X	Glucose residue substituted with an α-linked xylose
L	Glucose residue substituted with an α-linked xylose further substituted with a β-linked arabinose
F	Glucose residue substituted with an α-linked xylose further substituted with a fucose residue

Table 4.1. Example of nomenclature system for xyloglucan oligosaccharides (Fry et al., 1993).

Cellulase digestion of xyloglucans from tamarind seeds has been shown to release four types of oligosaccharides - XXXG, XLXG, XXLG and XLLG (Buckeridge et al., 1992; Marry et al., 2003; Grishutin et al., 2004). More widely, structural studies of monocotyledon xyloglucans by Hsieh and Harris (2009) have also revealed interesting differences in oligosaccharides that the authors discuss within phylogenetic context of monocotyledon species evolution (Hsieh and Harris., 2009).

The nomenclature system developed by Fry et al. (1993) is useful for investigating putative differences in xyloglucan binding (by identifying the oligosaccharides released) between paralogous proteins of enzyme families. Enzymatic activity has been studied previously – for example, Yaoi et al. (2005) isolated a GH74 xyloglucanase from *Paenibacillus* sp., demonstrating putative dual endo- and exo-xyloglucanase activity (or processive endo-activity), by the release of XXX, XXXG and GXXXG from xyloglucan oligosaccharides (Yaoi et al., 2005).

4.2.3 Catalytic mechanism of GH10 endo-β-1,4-xylanases

Whilst xylanases have been assigned to a range of CAZyme families, the most common families are GH10 and GH11 (http://www.cazy.org; Lombard et al., 2013). Proteins from both families are structurally distinct - for example, GH10 enzymes fold into a $(\beta/\alpha)_8$ barrel ((Harris et al., 1994), and see Chapter 3, Figure 3.12 for predicted structural features of *P. sojae* GH10 paralogs), with smaller substrate binding sites (suggested by their high activity towards short oligosaccharides, as well as crystal structure studies that suggest they are able to bind 4-5 substrate units (Biely et al., 1997; Biely et al., 1981)). Both families have varying substrate specificities (Biely et al., 2016), as well as different targets for catalysis – GH10 enzymes cleave non-reducing ends of substituted xylose residues (resulting in the release of shorter oligosaccharides), whilst GH11 enzymes cleave unsubstituted xylose residues (Collins et al., 2005). Interestingly, some GH10 enzymes are also enzymatically-active towards cellulose – Chu et al. (2017) investigated *Caldicellulosiruptor bescii* GH10 substrate promiscuity and demonstrated 6 and 10 amino acid residues involved in the xylanase and cellulase activity, respectively (Chu et al., 2017), also highlighting the importance of non-catalytic residues within substrate binding pockets for activity. Similar to the mode of action described for endo-glucanases, endo-xylanases hydrolyse the β-1,4 linkages between xylose residues by a double displacement mechanism - retaining the configuration of C1 at the site of cleavage (Davies and Henrissat., 1995).

4.2.4 Duplication and divergence in the evolution of enzyme functions

For proteins of interest, identifying putative functions (including assignment to a CAZyme family) is possible through conserved sequence identity (e.g. Pfam (Finn et al., 2016) and Interpro (Finn et al., 2017) searches). However, reliability is limited especially if characterised proteins are only distantly related, or if proteins of interest are paralogous with the potential for derived functions and neofunctionalization. It is possible that many enzymes possess 'promiscuous' functions, i.e. those that lie outside of their previously classified range. For example, a fourth function of GH12 enzymes has been described, a xyloglucan endo-transcosylase (EC 2.4.1.207), but this is so far only associated with a single fungal GH12 protein (GenBank AAN89225.1; Nielsen., 2002). It is therefore feasible that unique sequence and structural features of GH12 paralogs would play a role in broad or variable enzymatic functions (e.g. across different biological conditions), the efficiency of substrate binding (e.g. *how much of* or *how strongly* a substrate is able to be bound), or in the substrate range itself (e.g. the ability to digest multiple types of substrates using modifications of the same active site). Different properties associated with amino acid side chains around the substrate binding sites are likely to play an important role here, as previously mentioned - particularly as CAZymes secreted to digest plant cell walls must be able to bind and function efficiently within a complex of milieu of plant substrates (McNeil et al., 1984; Schindler., 1998).

Promiscuous enzyme activities can contribute to an organism’s fitness - creating new selective advantages for the new activities (Ohno., 1970). This is consistent with other early work (for example) by Ycas (1974) and Jensen (1976), who hypothesised the evolution of modern metabolism through duplication of *primitive* broadly-specific enzymes – with gene duplication events creating novel genetic material for selection to act on (leading to the optimisation of the activity and independent regulation of each enzyme (Ycas., 1974; Jensen., 1976)).

It is interesting to consider the problem of how gene sequence evolution leads to protein change, which in turn leads to adaptation to novel functions (innovation), whilst their original functions are maintained (conservation). Evolution of new enzyme functions has been represented by various models – including neofunctionalization, where gene duplication is proposed to yield redundant gene copies (Ohno., 1970; Force et al., 1999). In the original model of neofunctionalization, one gene copy is maintained under purifying selection to retain the ancestral function, leaving the duplicated gene hypothetically subject to genetic drift (neutral selection) – meaning it is theoretically able to accumulate mutations over time that could eventually confer a new activity (in addition to or in the absence of the original function). This model of gene duplication implies that the ancestral enzyme was unable to perform the new activity prior to duplication, and is also referred to as ‘mutation during non-functionality’ (MDN) (Hughes., 1994). The novel function (if advantageous to the organisms’ fitness) is then maintained through positive selection. However, the original neofunctionalization model received some criticism as there is little experimental support – some studies show that both duplicated genes are under purifying selection (e.g. demonstrated by analysis of duplicate genes in *Xenopus laevis* (Hughes and Hughes., 1993)), and some studies show that in the absence of selection, deleterious mutations (rather than advantageous ones) are more likely to accumulate - leading to non-functionalization, rather than neofunctionalization. For example, only when a selective pressure was applied to a library of TEM-1 β-lactamase variants (i.e. ampicillin), was there an increase in the accumulation of beneficial mutations eventually leading to a new function (cefotaxime resistance), however, in the absence of selection, many enzymes lost their function (Bershtein et al., 2008). It is also possible that ‘relaxed’ selection pressures on individual enzymes (by increasing gene/transcriptional dosage through gene duplication events) might also protect the original activity from putatively deleterious mutations, whilst increasing the accumulation of beneficial mutations and thus the probability of neofunctionalization.

Conversely, the Innovation, Amplification, Divergence (IAD) model of neofunctionalization involves single gene evolution towards one or more additional ‘weaker’ or ‘side’ activities (i.e. novel protein functions are evolved in the ancestral gene before gene duplication), meaning that subsequent gene duplications have the potential to evolve mutations favouring a side activity becoming the main activity of the protein (Nasvall et al., 2012).

Subfunctionalization refers to a model of gene evolution in which multiple ancestral functions encoded by a single gene/protein become sub-divided or partitioned between duplicated genes (Force et al., 1999). Subfunctionalization has been proposed to occur by duplication-degeneration-complementation (DDC), or by escape from adaptive conflict (EAC). In the DDC model (as assumed in Ohno's model) the duplicated gene is maintained at low frequency, putatively leading to the compartmentalisation of functions over the duplicated gene and its ancestor - strong selection pressure for maintenance of both genes would be expected in this model, as together, both genes encode the full set of functions required. The EAC model (Hittinger and Carroll., 2007) is similar in scope, however, there is a focus on the multi-functionality of a single gene before the duplication event (similar to what is assumed of the ancestral gene in the IAD model of neofunctionalization) - over evolutionary time, the gene evolves towards multiple advantageous functions, but each are unlikely to be carried out with optimal efficiency (creating an 'adaptive conflict'). Gene duplication therefore allows the functions to be partitioned (as in the DDC model), however, in the EAC model, after a gene duplication it is assumed that the genes are both under positive selection (to maintain the full repertoire of functions) (Hittinger and Carroll., 2007).

Although hypothesising a model of gene evolution by duplication for *P. sojae* GH10 and GH12 paralogs was not the main focus of this thesis, it is interesting to consider whether experimental characterisation of the gene families could uncover putative neofunctionalised or subfunctionalised functions – functions that have evolved from an HGT event from fungi, therefore enabling us to better consider the significance of both evolutionary phenomena in the evolution of plant-parasitic oomycetes. Further work would involve using ancestral gene reconstruction methods and heterologous characterisation of the putative functions of the fungal zenologue and the ancestral acquired HGT – this approach has been previously used to demonstrate the evolutionary history of oomycete transporter proteins acquired from fungi (Savory et al., 2018).

4.2.5 Genome-editing tools to investigate enzyme paralogs *in vivo*

The CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) system is found across diverse species of bacteria and archaea (Makarova et al., 2006), and plays an important role in adaptive immunity against bacteriophage invasion (e.g. demonstrated experimentally in *Streptococcus thermophilus* (Barrangou et al., 2007)). During an initial invasion, the host cell stores fragments of viral genomes within loci known as CRISPR arrays – the arrays consist of repetitive sequences, interspaced by short non-repetitive sequences (these are the fragments that match viral sequences from previous attacks (short segments of bacteriophages and conjugative plasmids – see Mojica et al., 2005)), acting as a type of molecular memory for the host cell.

Three types of CRISPR systems have been identified so far across a range of hosts, with the type II system of *Streptococcus pyrogenes* being the best-studied to date. Upon subsequent viral attacks, two non-coding CRISPR RNAs (crRNAs) (pre-crRNA and tracrRNA) are transcribed - the pre-crRNA is processed into mature-crRNA, which then acts as a ‘guide’ to direct an associated nuclease (known as Cas; CRISPR-associated protein) to the target DNA. A sequence known as the Protospacer Adjacent Motif (PAM) is required for target DNA recognition by the Cas nuclease (which is directed to the protospacer on the target DNA next to the PAM sequence) (Mojica et al., 2009). The PAM sequence is organism-specific, for example the sequence ‘NGG’ is recognised in *S. pyrogenes* (Mojica et al., 2009). At the target site, the Cas nuclease creates a DNA double-strand break in the DNA (thereby inactivating the virus).

The study of diverse CRISPR/Cas systems has led to novel Cas proteins being discovered with new enzymatic activities, functioning as single proteins or within a complex; of the microbial nucleases discovered so far, Cas9 is able to create double-strand DNA cleavage through the use of a single guide RNA (sgRNA). The basic components of CRISPR have been used to alter DNA in diverse organisms, including fungi, plants, mammals (including in multiplex, e.g. Cong et al. (2013); in *S. cerevisiae* e.g. Mali et al. (2013), and in the oomycetes (*P. sojae* - Fang and Tyler (2016); Fang et al; (2017), *P. capsici* - Wang et al (2018), *P. infestans* – Hoogen and Govers (2018)). By using CRISPR as a genome editing tool to precisely knockout or replace genes in target organisms, we can better understand their impact *in vivo.* The technique requires a nuclease, a 20 nucleotide gRNA and a tracrRNA expressed within a plasmid – upon the targeted dsDNA break by the nuclease, host DNA repair pathways are initiated – cleaved ends are ligated by Non-Homologous End Joining (NHEJ), or Homology Directed Repair (HDR), if a HDR template is provided in the experiment (e.g. as previously demonstrated for *P. sojae* (Fang and Tyler., 2016)).

4.3 Aims of chapter

Chapter 3 aimed to use computational methods to explore putative functional differences between multiple paralogs of two previously-identified HGT events associated with breakdown of the plant cell wall (Torto et al., 2002; Richards et al., 2006; Belbahri et al., 2008; Richards et al., 2011; Misner et al., 2015; Savory et al., 2015) - a GH12 enzyme family (putatively involved in cellulose degradation), and a GH10 enzyme family (putatively involved in xylan degradation). Endoglucanases and endoxylanases contribute to an important repertoire of enzymes for parasitic microbes to degrade plant-specific substrates, providing a means of entry into plant tissues, as well as a source of fixed carbon. As discussed in the previous chapter, ten out of eleven of the *P. sojae* GH12 paralogs possess both glutamic acid residues theoretically required for enzymatic activity (Chapter 3, Figure 3.4) - therefore, putative differences in the *in vivo* functions (including interactions with their substrates) could arise from diverging amino acid sequences, giving rise to distinct structural properties of the paralogous proteins. For example, the protein sequence of *P. sojae*_482953 includes a significantly disordered C-terminus 'tail' of 186 amino acid residues, present in multiple orthologs across multiple species – potentially altering the enzymatic activity of this GH12 paralog.

In the absence of direct experimental evidence, it is difficult to appreciate the functional significance of HGT followed by gene duplication and mutation for the evolution of paralog functions. Therefore, this chapter aims to investigate the functions of *P. sojae* GH12 and GH10 protein paralogs using experimental tools. Publically-available *P. sojae* transcriptome (RNA-sequencing) data will be used to scrutinise HGT gene sequences from published gene models (FungiDB; Stajich et al., 2012; Basenko et al., 2018); nucleotide sequences will then be codon-optimised for expression in *S. cerevisiae* BY4742, and protein secretion will be optimised by replacing native N-terminal signal peptides with a yeast-specific sequence. Yeast culture supernatants will be used as crude protein extracts for testing differences in enzymatic activity across pH 5-10 and temperatures 20-30°C, as well as exploring differences in oligosaccharide breakdown products using mass spectrometry. Additionally, the protein sequence of *P. sojae*_482953 will be truncated (i.e. it's significantly phosphorylated, disordered C-terminal 'tail' will be removed) in order to compare its enzymatic activity to that of the full-length protein - allowing us to better appreciate the importance of the C-terminal sequence for protein function. Lastly, a CRISPR/Cas9 genome editing technique (Fang and Tyler., 2016; Fang et al., 2017), will be used to replace the gene encoding *P. sojae*_482953 with a GFP HDR template *in vivo*[8]; verified *P. sojae* mutants will then be tested for differences in carbon utilisation using an enzymatic substrate (xyloglucan) as a sole carbon source (alongside a wild-type control) – this will enable investigation of the impact of single HGT paralogs to *P. sojae* carbon utilisation and phytopathogenicity.

4.4 Methods

4.4.1 Analysis of enzymatic activity by DNS reducing sugar detection

3,5-Dinitrosalicylic acid (DNS) is an aromatic compound that is reduced in the presence of reducing sugars (released during the breakdown of carbohydrates), to 3-amino-5-nitrosalicylic acid, which absorbs light at 540 nm (Miller., 1959). Therefore, an increase in absorbance at 540 nm over time can be used to infer enzymatic activity towards specific carbohydrate substrates.

[8] Generation of the *P. sojae* mutants was carried out in the research group of Professor Joseph Heitman (Duke University, North Carolina), using methods implemented in this organism (Fang and Tyler., 2016); Fang et al., 2017), Thanks to Professor Heitman and Dr Fang for their support during this part of the project.

Yeast transformants were screened for released reducing sugars from carbohydrate breakdown as follows: recombinant *S. cerevisiae* strains (in biological triplicate) were cultured in 20 mL for 7 days at 30°C (with shaking), the supernatants were removed by centrifugation at 4°C and concentrated 10x (Corning Spin-X UF; Sigma Aldrich) (stored on ice). Total protein concentrations were measured using a Qubit Fluorometer (Thermo Scientific). Concentrated supernatants (at 100 µg/mL) were incubated with 1% (w/v) carbohydrate substrate (see Table 4.2 for the carbohydrates tested in this study) in citrate buffer (50 mM, pH 5, 7 and 10: citric acid monohydrate dissolved in H_2O, with sodium hydroxide (NaOH) added to pH 5, 7 or 10) or buffer only (supernatant control) at 20 and 30°C. At each time point sampled, released reducing sugars were measured by removing 60 µL of each sample into a sterile 1.5 mL tube, and adding 60 µL of DNS reagent (30 g K-Na tartrate dissolved in 50 mL H_2O, with 1 g DNS and 20 mL NaOH added and made up to 100 mL with H_2O), followed by incubation at 95°C for 5 minutes to allow colour development. Samples were transferred to a sterile 96-well plate (clear, flat bottom), and the absorbance was measured at 540 nm using an absorbance microplate reader (CLARIOstar; BMG LABTECH).

Reducing sugars released over the incubation time were calculated based on reference to a 0-5 mg/mL standard curve (glucose or xylose), prepared during each experiment.

Xyloglucan (Tamarind)	Backbone of β-1,4-glucan; most substituted (endo activity)
Carboxymethylcellulose (CMC)	Backbone of β-1,4-glucan (cellulose derivative; endo activity)
Avicel	Backbone of β-1,4-glucan (microcrystalline cellulose; exo activity)
Laminarin	Backbone of β-1,3-glucan with some β-1,6-glucan branches
Xylan	Backbone of β-1,4-linked xylose

Table 4.2. Carbohydrates tested during this study.

4.4.2 Analysis of enzymatic activity by 'halo' screening

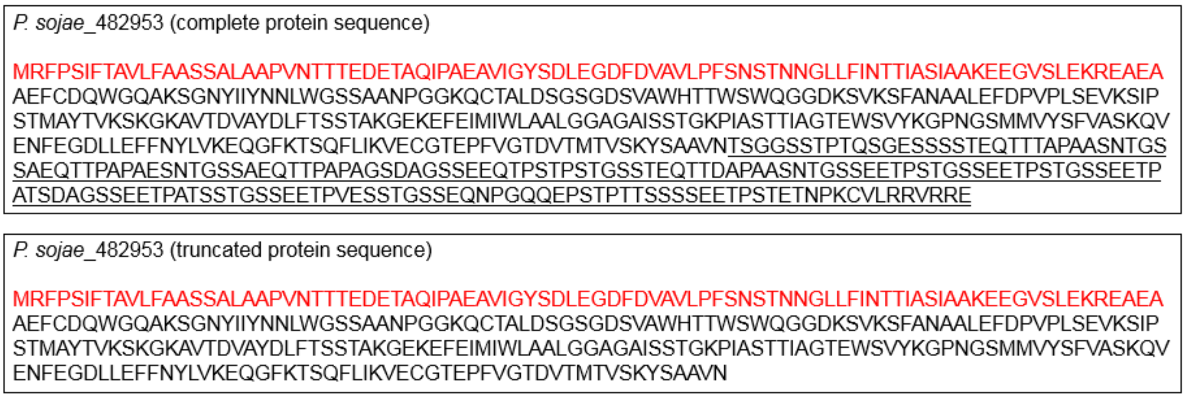
Yeast transformants were screened for carbohydrate-degrading activity as follows: recombinant *S. cerevisiae* strains (in biological triplicate) were cultured in 5 mL for 24 hours at 30°C (with shaking), and the cell pellets were washed once and resuspended in H_2O. 10 µL spots of OD_{600} –matched cells (0.1-1) were spotted onto SCM-URA agar containing 0.2% (w/v) substrate (with 2% (w/v) glucose as an additional carbon source). Additionally, concentrated supernatants at 100 µg/mL total protein were spotted on agar plates as described, (alongside performing DNS assays with the same supernatant sample). Agar plates were incubated at 30°C for 24-48 hours.

Yeast colonies were washed off the plates (not applicable to the concentrated supernatant spots), and the remaining intact polysaccharide on the plates was stained with 0.2% (w/v) Congo red (Sigma) for 30 minutes at room temperature, and de-stained with 1 M (w/v) sodium chloride (NaCl) for 30 minutes at room temperature (Wood and Weisz., 1987). H_2O adjusted to pH2 with hydrochloric acid (HCl) was used to intensify the stain. Extracellular enzyme activity was indicated by a clearing or 'halo' around colonies of enzyme-secreting *S. cerevisiae* strains.

For recombinant yeast strains secreting GH10 enzymes, screening was carried out as above, using 0.2% (w/v) Azo-Xylan (Birchwood) (Megazyme) as the substrate; incubation of agar plates resulted in visible halos around colonies without the requirement for further staining.

4.4.3 Removal of the disordered C-terminus 'tail' of *P. sojae*_482953

As described in Chapter 3, the protein sequence of GH12 paralog *P. sojae*_482953 includes a significantly disordered C-terminus 'tail' of 186 amino acid residues. Using computational tools alone (see Chapter 3), it was unclear how the additional amino acids affected the protein structure, or its interactions and enzymatic function towards putative substrates. To better understand the functional significance of the C-terminus tail of *P. sojae*_482953, a truncated version of the gene was engineered and expressed in *S. cerevisiae* BY4742 (Figure 4.2).



*P. sojae*_482953 (complete protein sequence)

```
MRFPSIFTAVLFAASSALAAPVNTTTEDETAQIPAEAVIGYSDLEGDFDVAVLPFSNSTNNGLLFINTTIASIAAKEEGVSLEKREAEA
AEFCDQWGQAKSGNYIIYNNLWGSSAANPGGKQCTALDSGSGDSVAWHTTWSWQGGDKSVKSFANAALEFDPVPLSEVKSIP
STMAYTVKSKGKAVTDVAYDLFTSSTAKGEKEFEIMIWLAALGGAGAISSTGKPIASTTIAGTEWSVYKGPNGSMMVYSFVASKQV
ENFEGDLLEFFNYLVKEQGFKTSQFLIKVECGTEPFVGTDVTMTVSKYSAAVNTSGGSSTPTQSGESSSSTEQTTTAPAASNTGS
SAEQTTPAPAESNTGSSAEQTTPAPAGSDAGSSEEQTPSTPSTGSSTEQTTDAPAASNTGSSEETPSTGSSEETPSTGSSEETP
ATSDAGSSEETPATSSTGSSEETPVESSTGSSEQNPGQQEPSTPTTSSSSEETPSTETNPKCVLRRVRRE
```

*P. sojae*_482953 (truncated protein sequence)

```
MRFPSIFTAVLFAASSALAAPVNTTTEDETAQIPAEAVIGYSDLEGDFDVAVLPFSNSTNNGLLFINTTIASIAAKEEGVSLEKREAEA
AEFCDQWGQAKSGNYIIYNNLWGSSAANPGGKQCTALDSGSGDSVAWHTTWSWQGGDKSVKSFANAALEFDPVPLSEVKSIP
STMAYTVKSKGKAVTDVAYDLFTSSTAKGEKEFEIMIWLAALGGAGAISSTGKPIASTTIAGTEWSVYKGPNGSMMVYSFVASKQV
ENFEGDLLEFFNYLVKEQGFKTSQFLIKVECGTEPFVGTDVTMTVSKYSAAVN
```

Figure 4.2. *P. sojae*_482953 full (upper) and truncated (lower) protein sequences tested during this study. The underlined sequence was removed in order to generate the truncated version of the protein. The red sequence is the *S. cerevisiae*-specific MFA signal peptide sequence used for efficient protein secretion in this heterologous host.

Primers were designed to amplify a truncated *P. sojae*_482953 from the full gene sequence (i.e. 558bp removed from the 3' end), with 5' *XmaI* and 3' *HindIII* restriction site sequences. The PCR master mix was prepared as follows (1x): in a sterile 1.5 mL tube, 5 µL Q5 Reaction Buffer (5x), 0.5 µL DNTPs (10 mM), and 1.25 µL of each forward and reverse primers were added to 15.75 µL H_2O. Q5 High-Fidelity DNA Polymerase was thawed on ice and added (0.25 µL) to the PCR mix. 1 µL plasmid DNA (5 ng/µL) containing the full-length gene was used as the template for PCR amplification. PCR conditions were as follows: denaturation at 98°C (5 Minutes), annealing at 98°C (10 seconds) + 50-70°C (30 seconds) + 72°C (2.5 minutes) (30 cycles), followed by extension at 72°C for 10 minutes. Amplification of the truncated 933bp gene product (plus the additional bases required for restriction enzyme cloning) was confirmed by gel electrophoresis, and the PCR product was purified using Thermo Scientific GeneJET PCR purification kit according to the manufacturer's protocol.

The purified PCR product (insert) and the plasmid backbone (vector) were prepared as follows: 30 µL insert and plasmid p426-GPD were digested with 1 µL *XmaI* and *HindIII* (each 10 units in total) in 5 µL Cutsmart Buffer (10x) and 14 µL H_2O, overnight at 37°C. The digested insert was PCR-purified as previously described. The digested plasmid was confirmed by gel electrophoresis, and a linear band was excised for gel purification using Thermo Scientific GeneJET Gel Extraction kit, according to the manufacturer's instructions. Purified plasmid DNA was quantified using a Nanodrop. The vector was additionally phosphatase-treated to prevent self-ligation as follows: 10 µL of the vector was incubated with 1 µL Antarctic Phosphatase reaction buffer (10x) and 1 µL Phosphatase at 37°C for 30 minutes, followed by enzyme inactivation at 70°C for 5 minutes, and subsequently stored on ice.

The DNA insert (8 µL) and the vector (2 µL) were ligated in 2 µL T4 DNA ligase buffer (10x), 0.4 µL T4 DNA ligase (400 U), in a total of 20 µL. The ligation reaction was incubated at room temperature overnight, and transformed into chemically-competent *E. coli* – 2 µL of ligation reaction was added to 50 µL competent cells, incubated on ice for 30 minutes, heat-shocked at 42°C for 30 seconds, and incubated on ice for 2 minutes. Cells were recovered by adding 250 µL LB and were incubated at 37°C for 1 hour with shaking. 100 µL of cells were inoculated onto LB agar (supplemented with 100 µg/mL ampicillin for plasmid selection). Agar plates were incubated at 37°C overnight. Three transformants were re-streaked on fresh agar and used to prepare liquid cultures (one colony inoculated into 5 mL LB-amp broth in sterile 50 mL tubes, incubated at 37°C overnight with shaking). Plasmid extraction was carried out using Thermo Scientific GeneJET plasmid miniprep kit according to the manufacturer's protocol, and purified plasmid DNA was quantified using a Nanodrop. The correct truncated *P. sojae*_482953 gene sequence in the vector was confirmed by Sanger Sequencing, as previously described.

Purified plasmid DNA containing the truncated *P. sojae*_482953 gene was transformed into competent *S. cerevisiae* BY4742 using a basic electroporation protocol, as previously described, and transformation mutants were selected on SCM-URA agar. Three independent *S. cerevisiae* mutants were re-stocked and maintained as the three biological replicates used in this study of this enzyme.

The truncated *P. sojae*_482953 protein (secreted into *S. cerevisiae* culture supernatants) was assayed for DNS reducing sugar activity, and enzymatic halo detection using the same methods as described in sections 4.4.1 and 4.4.2 of this chapter.

4.4.4 Analysis of oligosaccharides released during enzymatic degradation of xyloglucan by Mass Spectrometry

Mass Spectrometry (MS) was utilised in order to investigate the differences in degradation products released by active *P. sojae* GH12 paralogs from xyloglucan breakdown. MS involves the ionisation of samples into charged molecules, allowing their mass-to-charge ratio (*m/z*) to be determined; it was expected that digestion of xyloglucan would result in the release of shorter oligosaccharides (and putatively some smaller sugars (e.g. mono- and disaccharides), however it was not possible to detect mono- and di-saccharides using the method described below)).

Sample ionisation by Matrix Assisted Laser Desorption/Ionisation (MALDI) is useful for carbohydrates and other large biomolecules that are otherwise likely to fragment or decompose when heated. In MALDI, the sample is dissolved by an organic solvent, mixed with an energy-absorbing matrix (EAM), and spotted on a target plate. When the spot is irradiated by a laser, the sample ionises, producing ions in the gas phase - an electric field is applied, causing the ions to move towards a detector which records the time of flight (the ionisation system is MALDI and the detection system is Time of Flight (TOF), hence this technique is referred to as MALDI-TOF), and the mass-to-charge ratio (*m/z*) is computed. The resulting spectra includes the *m/z* and the intensity (numbers of ions detected) – therefore, specific molecular species present in the sample result in peaks of intensity at specific *m/z*.

Yeast transformants were prepared for MS analysis as follows: recombinant *S. cerevisiae* strains (in biological triplicate) were cultured in 20 mL for 7 days at 30°C, the supernatants were removed by centrifugation at 4°C and concentrated 10x (Corning Spin-X UF) (stored on ice). Protein concentrations were measured using a Qubit Fluorometer (Thermo Scientific). Concentrated supernatants (at 100 µg/mL) were incubated with 1% (w/v) xyloglucan (tamarind) in citrate buffer (50 mM, pH 7: citric acid monohydrate dissolved in H_2O) or buffer only (supernatant control) for 0-90 hours at 30°C.

At 0 and 90 hours, 60 µL of each sample was removed into a sterile 1.5 mL tube), and the reactions terminated by heating at 95°C for 5 minutes. The samples were air-dried using a centrifugal vacuum concentrator (SpeedVac), and sent for MS analysis to the National Mass Spectrometry facility (NMSF) at

Swansea University. Another 60 µL of the reactions were also removed at each time point (0-90 hours), to approximate the reducing sugars released by the DNS method described above.

4.4.5 Detection of N-terminal *his*-tagged proteins in *S. cerevisiae* culture supernatants by western blotting

All gene sequences under study were additionally cloned into a p426-GPD-6x*His* vector (resulting in a 6x*His* tag fused to the C-terminus of the secreted protein), in order to detect and purify recombinant proteins from *S. cerevisiae* culture supernatants. Molecular cloning methods used were as previously described, using *XmaI* and *HindIII* restriction enzymes for cloning into p426-GPD-6x*His*.

Detection of heterologously-expressed GH12 proteins in yeast culture supernatants was carried out by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) as follows: yeast strains expressing genes with an N-terminal *6xHis* tag were cultured in 20 mL for 7 days at 30°C, to allow high density growth and maximal protein secretion. The supernatants were removed by centrifugation at 4°C, and concentrated approximately 10x (Corning Spin-X UF). Samples were subsequently stored on ice or at 4°C. Proteins in the culture supernatants were separated by molecular mass on TruPAGE 10% polyacrylamide gels (Sigma-Aldrich), alongside the Chameleon Duo ladder (LI-COR). Electrophoresis was carried out at 180 V for ~45 minutes at room temperature. To confirm supernatant-localised heterologous proteins, fluorescence-based western blotting techniques were used as follows: after electrophoresis, proteins were transferred onto an Immobilon-FL membrane (Merck Millipore) (transfer carried out on ice at 100 V for 1-4 hours). The membrane was then washed in 1x phosphate buffered saline (PBS), and blocked overnight in 0.1% Alkali-soluble Caesin in 0.2% PBS (with shaking, in the dark). After incubation, the membrane was washed in PBS, and incubated in 1 µg primary antibody (Rabbit anti-6x His; Abcam ab9108) in blocking buffer + 0.1% Tween20 for 1 hour with shaking. The membrane was then washed four times in PBS + 0.1% Tween20, and incubated in 2 µg secondary antibody (Goat Anti-Rabbit; Abcam) in blocking buffer + 0.1% Tween20 for 1 hour with shaking (in the dark). The membrane was washed four times in PBS + 0.1% Tween20, and once in PBS to remove residual Tween20, and subsequently stored in the dark - imaging was carried out on a LI-COR Odyssey Fc imaging system.

4.4.6 Using the CRISPR/Cas9 system to disrupt an enzymatically active GH12 paralog in *P. sojae*

CRISPR/Cas9 methods for *P. sojae* were developed by Fang and Tyler (2016); Fang et al; (2017), and previously used for gene disruption by non-homologous end joining (NHEJ) and homology-directed repair (HDR). The methods were used in this study to disrupt the gene encoding *P. sojae*_482953 as per Fang et al (2017).

Guide RNA (gRNA) was designed using the Eukaryotic Pathogen CRISPR guide RNA/DNA Design Tool (EuPaGDT); the nuclease selected was ‘SpCas9’ – i.e. Cas9 from *S. pyrogenes*, which recognises the PAM sequence ‘NGG’ (chosen here because it was the Cas9 nuclease used in the work by Fang et al to develop CRISPR plasmids for *P. sojae*). The gRNA length was selected as 20 nucleotides, and the *P. sojae*_482953 gene sequence was used as an input in plain/FASTA format. From the long-list of potential gRNAs, several scoring criteria were used to evaluate gRNA efficiency and off-target potential (i.e. alignment of the gRNA sequences to the rest of the genome). Off-target sites were also manually checked using the BLASTN function within FungiDB (Stajich et al., 2012; Basenko et al., 2018); short-listed gRNAs were then also checked for their secondary structures, to avoid self-complementarity (preventing hybridization with the target sequence).

The gRNA linear insert was constructed using sense and antisense oligonucleotides (100 µM), with *NheI* and *BsaI* restriction sites. The oligonucleotides were annealed together to form dsDNA as follows: In a sterile 0.2 mL PCR tube, 3 µL of each oligonucleotide was mixed with 3 µL 10x T4 DNA ligase buffer, 4 µL 0.5 M NaCl and 21 µL H_2O. The reactions were incubated at 100°C for 2 minutes in a thermocycler, and cooled slowly to room temperature for 3-4 hours. The annealed oligonucleotides were diluted 1 µL into 499 µL H_2O and stored at 4°C.

The plasmid backbone for the gRNA was prepared as follows: 3 µg plasmid pYF515 (13032 bp, encoding the Cas9 nuclease) was digested with *NheI* and *BsaI* (each 15 units in total) in 1x CutSmart buffer, for 5 hours at 37°C. Digested plasmid DNA was confirmed by gel electrophoresis, and linear bands were excised for gel purification. Purified plasmid DNA was quantified using a Nanodrop. Diluted, annealed gRNA oligonucleotides (4 µL) were ligated with the purified digested pYF515 plasmid (50 ng) in 1x T4 DNA ligase buffer, and 1 µL T4 DNA ligase (400 U), in a total of 20 µL. The reaction was incubated for 30 minutes at room temperature. The ligation mixture was transformed into chemically-competent *E. coli* as follows: 2 µL of ligation product was added to 50 µL competent cells, incubated on ice for 30 minutes, heat-shocked at 42°C for 45 seconds, and incubated on ice for a further 2 minutes. Cells were recovered by adding 800 µL FB media (for 1 litre: 25 g tryptone, 7.5 g yeast extract, 1 g glucose, 6 g NaCl, 50 mL TRIS (pH 7.6), made up to 1 litre with dH_2O, followed by incubation at 37°C for 45 minutes with shaking. 200 µL of cells were inoculated onto YT Amp agar, and the plates were incubated at 37°C overnight. Three transformants were re-steaked onto fresh media and used to set up overnight cultures for plasmid purification. The gRNA insert in pYF515 was confirmed by Sanger sequencing. Positive transformants were used to set up overnight cultures in 50 mL for larger-scale plasmid DNA isolation (midi-prep), in order to obtain ~200 µg of the plasmid DNA.

A Homology Directed Repair (HDR) template was constructed in order to produce a *P. sojae* gene-deletion mutant with a precise genome mutation (i.e. insertion of a specific (known) DNA sequence by homologous recombination). Fang et al. (2017) note that HDR occurs efficiently in *P. sojae* when 1 kb of homologous flanking sequences are included in the HDR DNA template (Fang and Tyler., 2016; Fang et al., 2017). The plasmid pBlueScript KS(-) was used as the vector backbone for the HDR template, and the donor DNA sequence used was GFP, with 1 kb of ‘left’ and ‘right’ flanking arms – homologous to the flanking sequences of the target gene (encoding *P. sojae*_482953) in the *P. sojae* genome. For assembly of the HDR template, NEBuilder was used to design PCR primers for single-reaction cloning (4 fragments: the vector backbone, GFP, 1 kb ‘left arm’ and 1 kb ‘right arm’ – each with 25 bp overlap to adjacent fragments, for efficient assembly).

The upstream and downstream homologous sequences of the gene encoding *P. sojae*_482953 were PCR-amplified from a *P. sojae* DNA extraction (gDNA), and the donor DNA fragment (GFP) was PCR-amplified from a plasmid DNA preparation. The PCR master mix was prepared as follows (1x): in a sterile 1.5 mL tube, 5 µL High Fidelity (HF) buffer (5x), 1 µL DNTPs (10 mM) and 0.5 µL of each forward and reverse primers were added to 17.25 µL H_2O. Phusion Polymerase was thawed on ice and added (0.25 µL) to the PCR mix. 2 µL *P. sojae* gDNA, and 1 µL GFP plasmid DNA were used as the templates for PCR amplification. PCR conditions were as follows: denaturation at 98°C (5 minutes), annealing at 98°C (10 seconds) + 61.2°C (left arm), 67.8°C (GFP) or 61.4°C (right arm) (30 seconds) + 72°C for 2.5 minutes (30 cycles), followed by extension at 72°C for 10 minutes. PCR products were confirmed by gel electrophoresis, and linear bands were excised for gel purification.

The plasmid backbone was prepared as follows: 3 µg plasmid pBlueScript KS(-) (2958bp) was digested with *XbaI* and *XmaI* (each 15 units in total) in 1x CutSmart buffer, for 5 hours at 37°C. Digested plasmid DNA was confirmed by gel electrophoresis, and linear bands were excised for gel purification. Purified plasmid DNA was quantified using a Nanodrop. NEBuilder HiFi DNA Assembly cloning kit was used to clone the 4 DNA fragments - a total of 0.2-0.5 pmols of DNA was used in the reaction; pmols per DNA fragment was calculated using the following equation: pmols = (weight in ng x 1000) / (base pairs x 650 Daltons). For example, 50 ng of 500 bp DNA is 0.15 pmols. As per the manufacturers protocol, the recommended ratio for DNA fragments in a 4-6 fragment assembly (1:1) were used. Reactions were prepared in a sterile 1.5 mL tube as follows: 10 µL of NEBuilder Assembly master mix (2x), DNA fragments (total of 0.2-0.5

pmols), with dH_2O added to a total volume of 20 µL. The reaction was incubated at 50°C for 60 minutes, and subsequently stored on ice or at -20°C. The assembled fragment mixture was transformed into chemically-competent *E. coli* using a standard transformation procedure as previously described.

Thirty-two transformants were re-streaked onto fresh media and used as template DNA for colony PCRs. Two primer pairs were used to verify correct assembly of the DNA fragments: (i) M13_Fw and GFP_Rv (expected amplimer: 1821 bp if the left homology arm and GFP were correctly inserted), and (ii) GFP_Fw and M13_Rv (expected amplimer: 1861 bp if GFP and the right homology arm were correctly inserted). The PCR master mix was prepared as follows (1x): in a sterile 1.5 mL tube, 12.5 µL GoTaq Green Master Mix (2x) (pre-mixed solution containing *Taq* DNA polymerase, dNTPs, $MgCl_2$ and reaction buffers – stored at -20°C and thawed on ice), and 0.5 µL of each forward and reverse primers were added to 10.5 µL H_2O. Colonies were picked into 10 µL dH_2O in a sterile 1.5 mL tube, and 1 µL from each tube was used as the DNA template for PCR amplification. PCR conditions were as follows: denaturation at 98°C (2 minutes), annealing at 98°C (30 seconds) + 50°C (primer pair i) or 46.8°C (primer pair ii) + 72°C for 2.5 minutes (35 cycles), followed by extension at 72°C for 10 minutes. PCR products were confirmed by gel electrophoresis.

Three transformants that were positive using both PCR primer pairs were selected to set up overnight cultures for plasmid DNA isolation (mini-prep). The HDR template was confirmed by Sanger sequencing (GeneWiz) using GFP_Fw, GFP_Rv, M13_Fw and M13_Rv primers.

gRNA plasmid DNA and HDR plasmid DNA were introduced into *P. sojae* by Polyethylene glycol (PEG) mediated protoplast transformation, as per Fang et al. (2017) (optimised from previously described methods (Dou et al., 2008; Mcleod et al., 2008)). Briefly; wild-type (WT) *P. sojae race 2* (*p6497*) was cultured on V8 agar at 25°C (in the dark without shaking) for 3-7 days, and then cultured on nutrient pea agar at 25°C (in the dark without shaking) for 3-7 days (V8 and nutrient pea media prepared as per Fang et al. (2017)). Five mycelial discs were extracted from the edge of mycelia growing on the nutrient pea agar, and inoculated into 50 mL nutrient pea broth (in sterilised 250 mL flasks). Flasks were incubated at 25°C (in the dark without shaking) for 2-4 days. *P. sojae* growth was visualised as 'clumps' of mycelia in the liquid media. The mycelia were collected using sterilised tweezers, and rinsed with 50 mL sterile dH_2O over (autoclaved-sterile) Miracloth and cheese cloth. The mycelia were rinsed with 50 mL 0.8 M mannitol, and then removed into a sterile tall petri dish and covered with 0.8 M mannitol for plasmolysis (incubation for 10 minutes at room temperature). The mycelia were then transferred back to the Miracloth and cheese cloth layers, and rinsed once with 0.8 M mannitol, before being removed into a sterile tall petri dish and covered with an enzyme solution (see Fang et al., 2017) for digestion. The petri dish was sealed and covered with foil, then placed on a shaking platform set to 50 rpm for 1 hour at room temperature, to allow for enzymatic digestion of the hyphae.

The dish was subject to light microscopy, to check for the release of round protoplasts (as an indicator of efficient enzymatic digestion). Using a Pasteur pipette to transfer the mycelia, the mycelial debris were removed by filtering through a 70 µM Falcon nylon mesh cell strainer (BD Biosciences) – collecting the flow through. *P. sojae* protoplasts were pelleted by centrifugation (1200 xg for 2 minutes at room temperature), washed in 30 mL W5 solution (see Fang et al., 2017) and resuspended in 10 mL W5 solution. The protoplasts were kept on ice for 20 minutes (or overnight, depending on time availability), and then pelleted again by centrifugation (1200 xg for 2 minutes at room temperature). All of the W5 solution was removed and the protoplasts were gently resuspended in 6 mL MMG solution (see Fang et al., 2017). The protoplasts were kept for 10 minutes at room temperature. In triplicate, 1 mL of protoplasts were added to 30 µg of the constructed sgRNA plasmid (in pYF515) and 13 µg of the constructed HDR plasmid (in PBS-K-) (1:1 equal molar ratio between the two plasmids) in 50 mL tubes, and incubated for 20 minutes on ice. 1.74 mL of freshly-prepared PEG-calcium solution was added to the tubes (with each tube at a 90° angle, 580 µL aliquots were added and allowed to slide down the tube, the tube was then gently rotated to mix). Transformation reactions were incubated for 15 minutes on ice, 2 mL of cold regeneration media (as in Fang et al., 2017) was added, followed by incubation for 2 minutes on ice. 8 mL of cold regeneration media was added and the reactions incubated for 2 minutes on ice again. Transformation reactions were each poured into a tall petri dish containing 10 mL of cold regeneration media, and ampicillin (50 µg/mL) was added to control bacterial contamination. Petri dishes were incubated at room temperature (in the dark without shaking) for 18 hours.

Regenerated hyphae were harvested by centrifugation (2000 xg for 5 minutes) in 50 mL tubes, and the supernatants discarded until ~3 mL remained in each tube. Each pellet was resuspended and divided evenly between three 50 mL tubes (i.e. 1 mL in each tube). 45 mL of liquid regeneration agar (containing

50 µg/mL G418; agar prepared and incubated at 42°C on the day of testing) was added to each of the three tubes, which were inverted to mix and poured across 3 petri dishes (i.e. 15 mL in each). For each transformation, there were a total of 9 petri dishes. Plates were incubated at 25°C (in the dark without shaking) for 2 days until mycelial colonies were visible. Twenty-four *P. sojae* colonies were transferred into 2 mL of liquid V8 media supplemented with 50 µg/mL G418, in sterile 12-well microplates. Plates were incubated at 25°C (in the dark without shaking) for 2-3 days.

gDNA was isolated at per Fang et al. (2017). Briefly; individual *P. sojae* transformants (and the wild-type strain) were grown in 2 mL V8 liquid medium in a 12-well plate for 2-3 days in the dark, without shaking. Fingertip-sized clumps of *P. sojae* hyphae were picked using 2x 200 µL sterile pipette tips and excess liquid removed on Kimwipe paper. Hyphae were transferred to a sterile 1.5 mL tube and snap-frozen in liquid nitrogen. Frozen hyphae were ground using autoclave-sterile pestles, and resuspended in 500 µL lysis buffer (as in Fang et al., 2017). The tubes were vortexed for 30 seconds, and incubated for 30 minutes at 37°C for RNA digestion. 500 µL of phenol-chloroform-isoamyl alcohol (25:24:1; saturated with TE) was added to the tubes and they were inverted five times. The tubes were centrifuged for 15 minutes at 12,000 xg at room temperature, and the upper clear layer was removed into a new sterile 1.5 mL tube. An equal volume of chloroform was added to the tubes and they were inverted five times. The tubes were centrifuged for 15 minutes at 12,000 xg at room temperature, and the aqueous layer was removed into a new sterile 1.5 mL tube. gDNA was precipitated by adding 0.5 volumes of isopropanol to the tubes and they were inverted five times, before placing at -20°C for 15 minutes. The tubes were centrifuged for 15 minutes at 12,000 xg at 4°C, and the supernatants were carefully removed. The pellet containing gDNA was dried, and dissolved in 40 µL sterile deionized water. Quality and quantity of *P. sojae* gDNA was analysed by gel electrophoresis.

After genotyping by PCR, single *P. sojae* zoospores were isolated (these are >95% mononucleate). Positive mutants were inoculated onto V8 agar supplemented with 50 µg/mL G418 and incubated at 25°C (in the dark without shaking) for 7-10 days. Plates were then flooded with sterile dH_2O every 30 minutes for ~4 hours – this allowed for nutrient depletion and sporangia production. Zoospore release from sporangia was monitored using a light microscope. The dH_2O was removed and the plates sealed and incubated at 14°C (in the dark without shaking) for 16 hours. Zoospores were counted using a haemocytometer, and ~15 zoospores were plated on V8 agar. Plates were incubated at 25°C (in the dark without shaking) for 2 days. Individual mycelial colonies were grown in 2 mL V8 liquid medium in a 12-well tissue culture plate for 2-3 days in the dark, without shaking. gDNA was extracted as previously described, and transformants were again genotyped by PCR. Verified homokaryotic *P. sojae* mutants were maintained on V8 agar (sub-cultured weekly for all downstream experiments).

Mutants with verified knockout of the gene encoding *P. sojae*_482953 (replaced with GFP by HDR), were tested for their ability to utilise xyloglucan as a sole carbon source. *P. sojae* mutants (and the wild type) were initially cultured on V8 medium as previously described; an agar plug taken from the edge of fast-growing mycelium was then transferred to the centre of minimal media agar (pH

6 (per litre): 0.5 g KH2PO4, 0.5 g K2HPO4, 4 x 10^{-4} g $MnSO_4$, 4 x 10^{-4} g $ZnSO_4$, 1.05 g NH_4Cl, 6.8 mL 1 M $CaCl_2.2H_2O$*****, 2 mL 1 M $MgSO_4.7H_2O$*****, 4 x 10^{-3} g $FeSO_4$******, 1% (w/v) substrate (*****Autoclaved separately, ******Sterile-filtered). The agar plates were incubated at 25°C (in the dark without shaking). Colony morphology pictures were taken at 7 days.

4.5 Results

4.5.1 Analysis of enzymatic activity by DNS reducing sugar detection and 'halo' screening

Reducing sugars released from the incubation of recombinant *P. sojae* proteins (secreted into *S. cerevisiae* concentrated culture supernatants) with carbohydrate substrates was investigated using DNS reagent (Miller., 1959). Secretion of all proteins was directed by the *S. cerevisiae* MFA N-terminal signal peptide, therefore it was expected that relative secretion rates between strains would be comparable. The crude enzyme extracts were matched to a total protein concentration of 100 µg/mL and incubated with 1% (w/v) substrate in Citrate buffer (pH5, 7, 10; 20°C and 30°C).

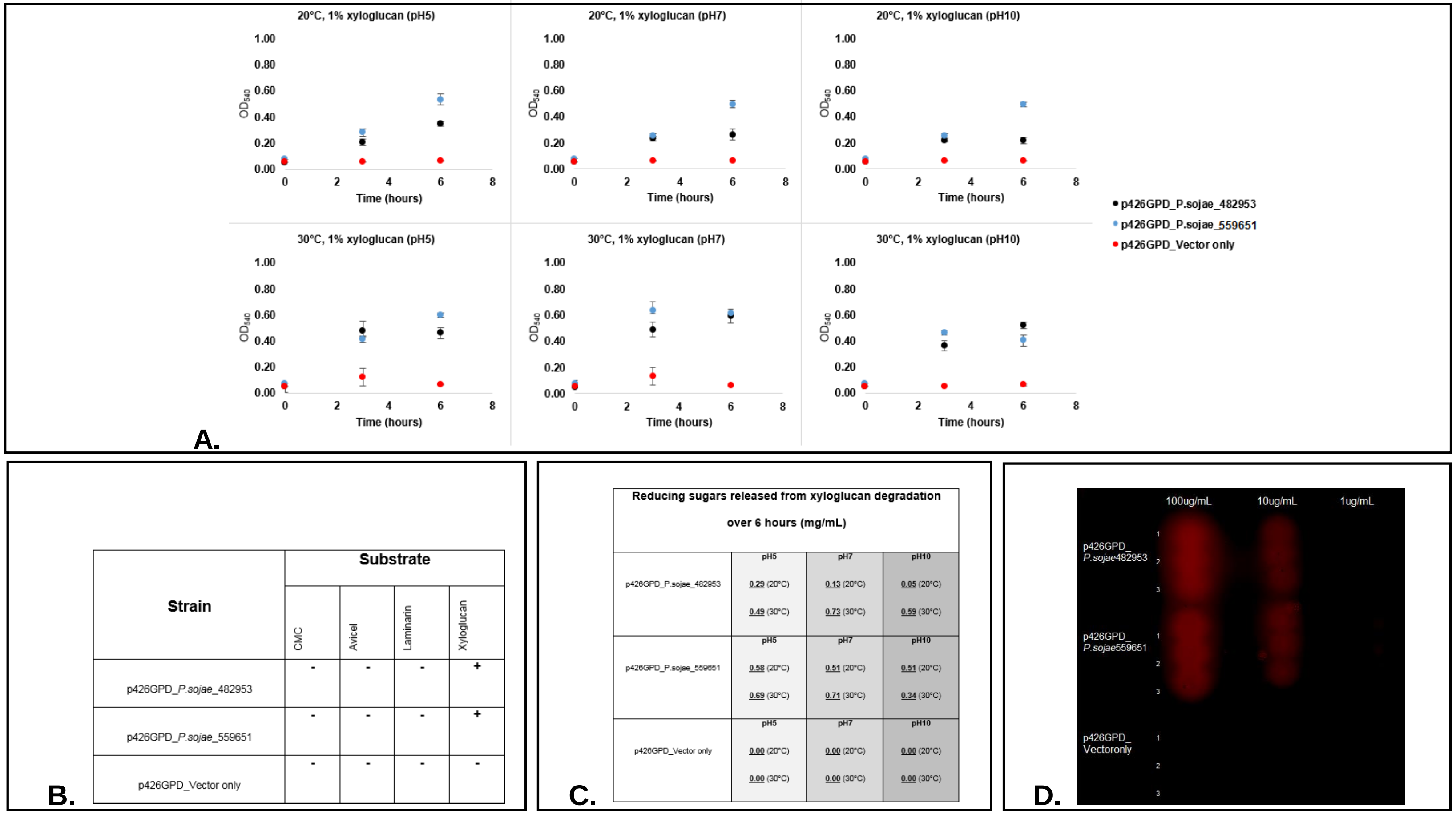
4.5.1.1 GH12

Of the *P. sojae* GH12 paralogs, it was only possible to detect activity of *P. sojae*_482953 and *P. sojae*_559651 using the DNS method; both paralogs demonstrated activity towards 1% xyloglucan up to 6 hours of incubation under all conditions tested (Figure 4.3). The reducing sugars released over 6 hours were calculated based on reference to a 0-4 mg/mL standard curve (glucose); at 20°C, *P. sojae*_482953 released 0.29 mg/mL, 0.13 mg/mL and 0.05 mg/mL reducing sugars at pH5, pH7 and pH10, respectively. *P. sojae*_559651 released

0.58 mg/mL, 0.51 mg/mL and 0.51 mg/mL at pH5, pH7 and pH10, respectively. At 30°C, *P. sojae*_482953 released 0.49 mg/mL, 0.73 mg/mL and 0.59 mg/mL reducing sugars at pH5, pH7 and pH10, respectively. *P. sojae*_559651 released 0.69 mg/mL, 0.71 mg/mL and 0.34 mg/mL at pH5, pH7 and pH10, respectively. Interestingly, at 20°C, *P. sojae*_482953 displayed highest activity at pH5 (0.29 mg/mL reducing sugar detection), whilst at 30°C, *P. sojae*_482953 displayed highest activity at pH7 (0.73 mg/mL detected). *P. sojae*_559651 released higher detectable reducing sugars during incubation at 20°C (compared with *P. sojae*_482953), again with highest detection at pH5 (0.58 mg/mL), whilst at 30°C, the highest detection of reducing sugars was at pH7 (0.71 mg/mL) for this paralog.

No activity was detected against CMC, Avicel or Laminarin for the two *P. sojae* paralogs (i.e. the proteins are xyloglucan-specific endoglucanases). The same concentrated yeast supernatants used in the DNS assay were additionally spotted onto SCM-URA agar plates containing 0.2% (w/v) xyloglucan; consistently, halos were visible after staining for both *P. sojae*_482953 and *P. sojae*_559651 GH12 paralogs, in comparison with a p426-GPD vector only sample (negative control) (Figure 4.3).

4.5.1.1 *P. sojae* GH12 paralogs _482953 and _559651 are active against xyloglucan (pH5, 7 and 10, at 20°C and 30°C), but show no activity against CMC, Avicel or Laminarin



Strain	Substrate			
	CMC	Avicel	Laminarin	Xyloglucan
p426GPD_*P.sojae*_482953	-	-	-	+
p426GPD_*P.sojae*_559651	-	-	-	+
p426GPD_Vector only	-	-	-	-

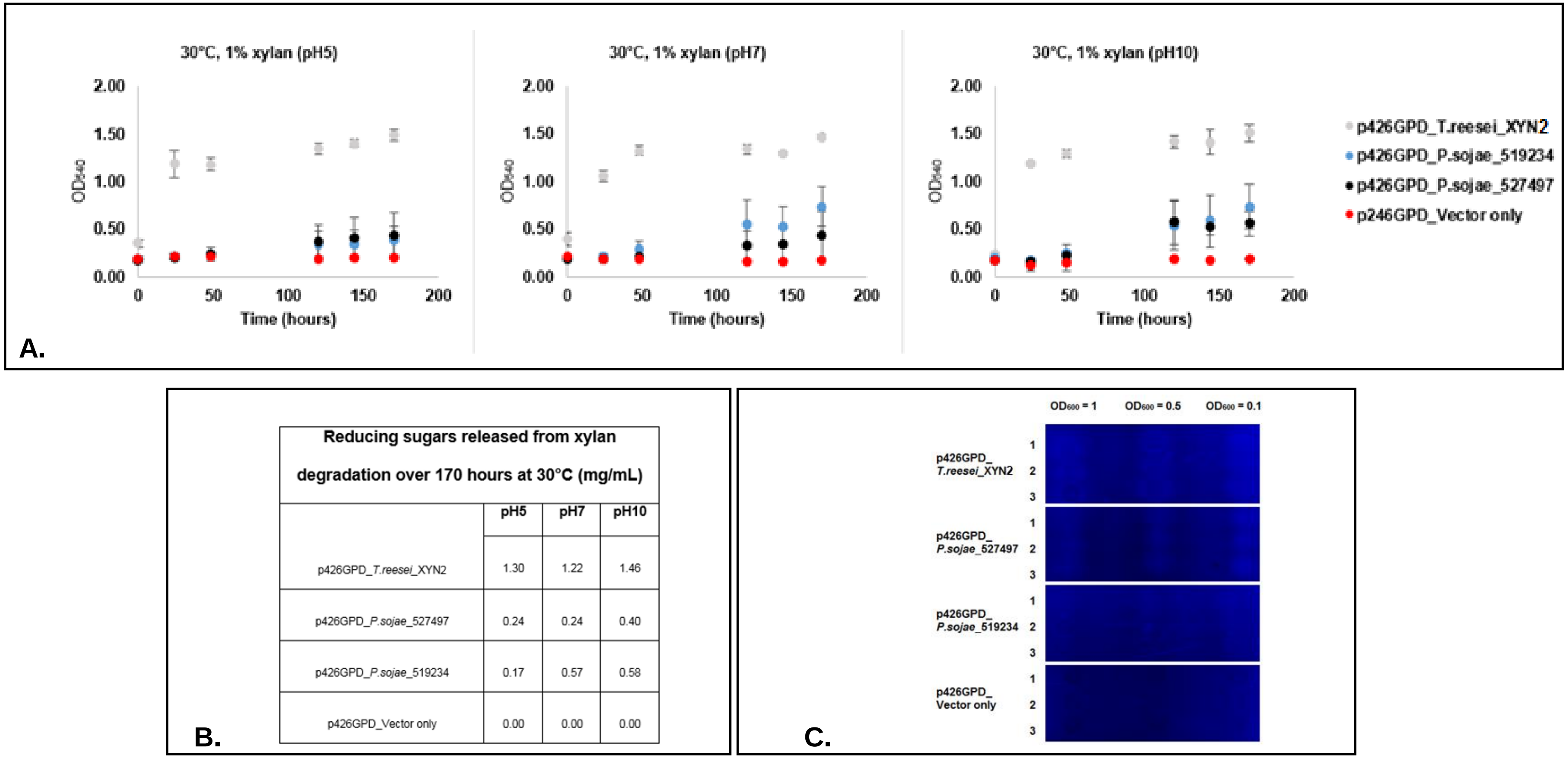
Reducing sugars released from xyloglucan degradation over 6 hours (mg/mL)			
	pH5	pH7	pH10
p426GPD_P.sojae_482953	0.29 (20°C) 0.49 (30°C)	0.13 (20°C) 0.73 (30°C)	0.05 (20°C) 0.59 (30°C)
	pH5	pH7	pH10
p426GPD_P.sojae_559651	0.58 (20°C) 0.69 (30°C)	0.51 (20°C) 0.71 (30°C)	0.51 (20°C) 0.34 (30°C)
	pH5	pH7	pH10
p426GPD_Vector only	0.00 (20°C) 0.00 (30°C)	0.00 (20°C) 0.00 (30°C)	0.00 (20°C) 0.00 (30°C)

Figure 4.3. A. *P. sojae*_482953 and *P. sojae*_559651 secreted into *S. cerevisiae* BY4742 culture supernatants were incubated with 1% xyloglucan; the increase in absorbance (OD_{540}) of DNS reagent added to the samples is suggestive of an increase in the reducing sugars released (i.e. from the breakdown of xyloglucan). Enzymatic activity was detected up to 6 hours of incubation under all conditions tested. No reducing sugars were detected in the p426-GPD vector only sample (N=3, +/- SD). **B.** No activity of *P. sojae*_482953 and *P. sojae*_559651 was detected against CMC, Avicel or Laminarin (i.e. the paralogs are xyloglucan-specific endoglucanases). **C.** The reducing sugars released over 6 hours (A) were calculated based on reference to a 0-4 mg/mL standard curve (glucose). **D.** Concentrated supernatants were spotted onto SCM-URA agar plates containing 0.2% (w/v) xyloglucan; consistently, halos were visible after staining for both *P. sojae*_482953 and *P. sojae*_559651 GH12 paralogs, in comparison with a p426-GPD vector only sample (negative control).

4.5.1.2 GH10

Of the *P. sojae* GH10 paralogs, it was only possible to detect activity of *P. sojae*_527497 and *P. sojae*_519234 at 30°C (pH5, pH7 and pH10); however, a longer incubation period was required to demonstrate activity under all pH tested (compared to GH12 paralogs - up to 170 hours) (Figure 4.4). Enzymatic activity of the secreted GH10 paralogs was compared to a fungal xylanase, *T. reesei*_Xyn2, which was heterologously expressed in *S. cerevisiae* using the native signal peptide sequence of the protein (as per Grange et al., 1996). The reducing sugars released over the incubation time were calculated based on reference to a 0-4 mg/mL standard curve (xylose); for *T. reesei*_Xyn2, 1.30 mg/mL, 1.22 mg/mL and 1.46 mg/mL reducing sugars were released after 170 hours at pH5, pH7 and pH10, respectively. For *P. sojae*_527497, 0.24 mg/mL, 0.24 mg/mL and 0.40 mg/mL, and *P. sojae*_519234, 0.17 mg/mL, 0.57 mg/mL and 0.58 mg/mL were released after 170 hours at pH5, pH7 and pH10, respectively (Figure 4.4). Enzymatic activity for *P. sojae*_519234 and *P. sojae*_527497 was also detected by the presence of halos around recombinant yeast strains spotted onto SCM-URA (+ 0.2% (w/v) Azo-xylan).

4.5.1.2 *P. sojae* GH10 paralogs _527497 and _519234 are active against xylan (pH5, 7 and 10, at 30°C)



Reducing sugars released from xylan degradation over 170 hours at 30°C (mg/mL)			
	pH5	pH7	pH10
p426GPD_*T.reesei*_XYN2	1.30	1.22	1.46
p426GPD_*P.sojae*_527497	0.24	0.24	0.40
p426GPD_*P.sojae*_519234	0.17	0.57	0.58
p426GPD_Vector only	0.00	0.00	0.00

Figure 4.4. A. *P. sojae*_527497 and *P. sojae*_519234 secreted into *S. cerevisiae* BY4742 culture supernatants were incubated with 1% xylan; the increase in absorbance (OD_{540}) of DNS reagent added to the samples is suggestive of an increase in the reducing sugars released (i.e. from the breakdown of xylan). Both proteins demonstrated activity towards xylan at 30°C under all pH tested. Enzymatic activity was screened alongside fungal xylanase, *T. reesei*_Xyn2 (heterologously expressed in *S. cerevisiae* using the native signal peptide sequence of the protein (as per Grange et al., 1996)). No reducing sugars were detected in the p426-GPD vector only sample (N=3, +/- SD). **B.** The reducing sugars released over the incubation time were calculated based on reference to a 0-4 mg/mL standard curve (xylose). **C.** Enzymatic activity was also detected by the presence of halos around recombinant yeast strains spotted onto SCM-URA (+ 0.2% (w/v) Azo-xylan).

4.5.2 Removal of the disordered C-terminus 'tail' of *P. sojae*_482953

To investigate the functional significance of the 186 amino acid 'tail' sequence at the C-terminus of GH12 paralog, *P. sojae*_482953 (as discussed in Chapter 3), a truncated version of the protein was engineered and expressed in *S. cerevisiae*. The full length and truncated proteins were expressed with the N-terminal MFA secretion signal, driving secretion into yeast culture supernatants, and then the culture supernatant was subject to DNS reducing sugar and halo assays as previously described.

*P. sojae*_482953 (truncated) was found to be enzymatically-active towards xyloglucan, although displayed significantly reduced activity in comparison with the full-length protein (Figure 4.5). All reactions were maintained at 20°C and 30°C (pH5, pH7 and pH10) for 22 hours, which did not change this observation. The reducing sugars released over 22 hours were calculated based on reference to a 0-5 mg/mL standard curve (glucose); at 20°C, *P. sojae*_482953 (full length) released a total of 1.39 mg/mL, 1.15 mg/mL and 1.16 mg/mL reducing sugars at pH5, pH7 and pH10, respectively, whilst *P. sojae*_482953 (truncated) released a total of 0.19 mg/mL, 0.41 mg/mL and 0.41 mg/mL, at the corresponding pHs. At 30°C, *P. sojae*_482953 (full length) released a total of 2.53 mg/mL, 1.83 mg/mL and 2.11 mg/mL reducing sugars at pH5, pH7 and pH10, respectively, whilst *P. sojae*_482953 (truncated) released a total of 0.69 mg/mL, 1.02 mg/mL and 1.14 mg/mL reducing sugars at the corresponding pHs.

Incubation of concentrated supernatants at 100 µg/mL, 10 µg/mL and 1 µg/mL total protein, with 1% (w/v) xyloglucan for 144 hours (pH7, 30°C) suggested that *P. sojae*_482953 (truncated) was less enzymatically active compared to the full-length protein; total reducing sugars released after 144 hours (at 100 µg/mL starting concentration of total protein) were 2.03 mg/mL and 1.54 mg/mL for full length and truncated proteins, respectively. Proteins spotted onto SCM-URA agar plates containing 0.2% (w/v) xyloglucan gave consistent results; halos were visible after staining for both *P. sojae*_482953 (full length) and *P. sojae*_482953 (truncated) – however, with reduced halo sizes (diameters) for the truncated version of this enzyme (Figure 4.5).

4.5.2 Removing the disordered C-terminus ‘tail’ from *P. sojae* _482953 reduces its enzymatic activity against xyloglucan

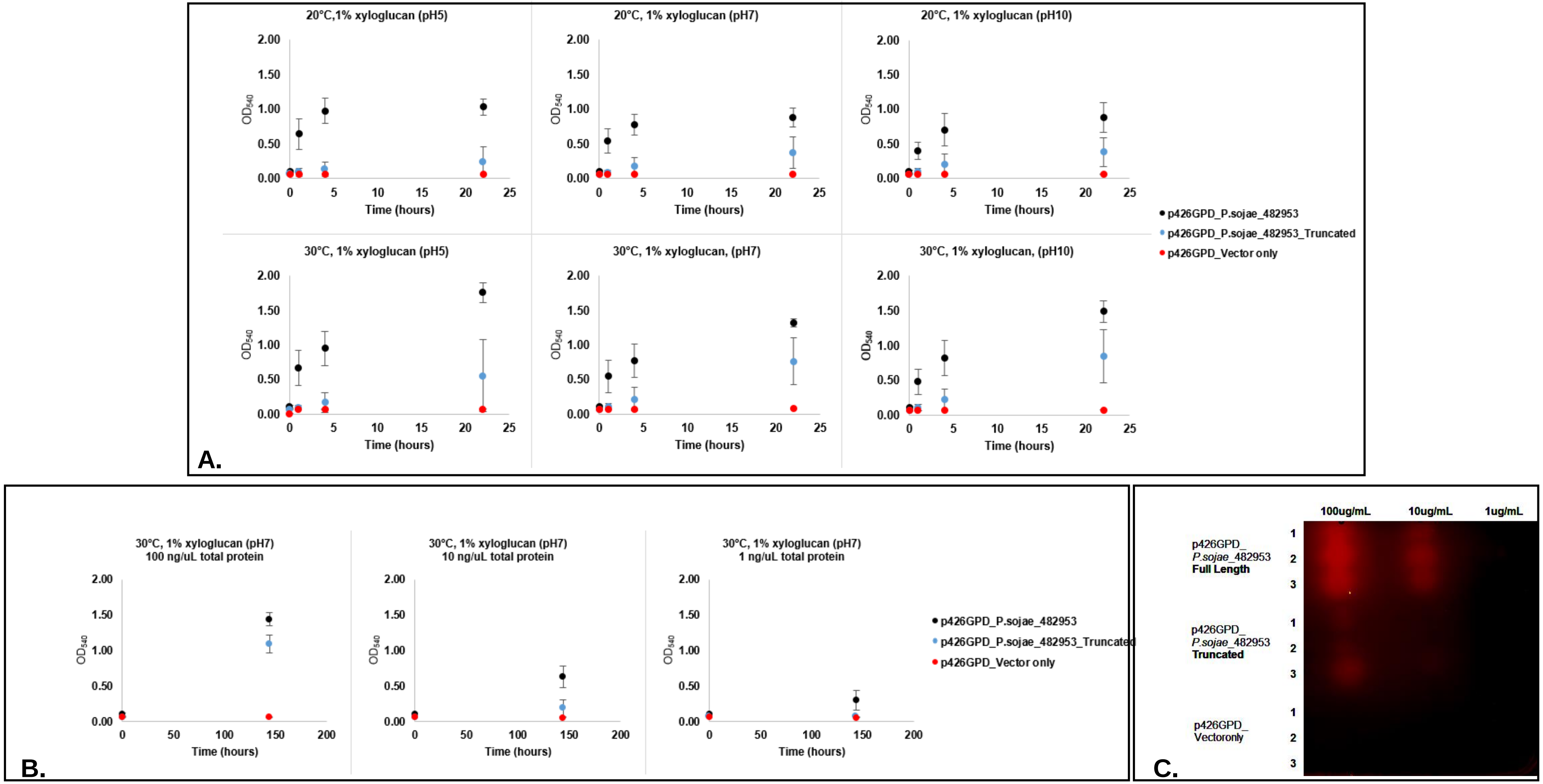


Figure 4.5. A. *P. sojae*_482953 (truncated) and *P. sojae*_482953 (full-length) secreted into *S. cerevisiae* BY4742 culture supernatants were incubated with 1% xyloglucan; the increase in absorbance (OD_{540}) of DNS reagent added to the samples is suggestive of an increase in the reducing sugars released (i.e. from the breakdown of xyloglucan). *P. sojae*_482953 (truncated) displayed significantly reduced activity towards xyloglucan in comparison with the full-length protein, up to 22 hours of incubation under all conditions tested. No reducing sugars were detected in the p426-GPD vector only sample (N=3, +/- SD). **B.** Incubation of concentrated supernatants at 100 µg/mL, 10 µg/mL and 1 µg/mL total protein, with 1% (w/v) xyloglucan for 144 hours (pH7, 30°C) suggested that *P. sojae*_482953 (truncated) was active at a slower rate compared to the full length protein; total reducing sugars released after 144 hours (at 100 µg/mL starting concentration of total protein) were 2.03 mg/mL and 1.54 mg/mL for full length and truncated proteins, respectively. **C.** Proteins spotted onto SCM-URA agar plates containing 0.2% (w/v) xyloglucan gave consistent results; halos were visible after staining for both *P. sojae*_482953 (full length) and *P. sojae*_482953 (truncated) – however, with reduced halo sizes (diameter) for the truncated version of this enzyme.

4.5.3 Analysis of oligosaccharides released during enzymatic degradation of xyloglucan by Mass Spectrometry

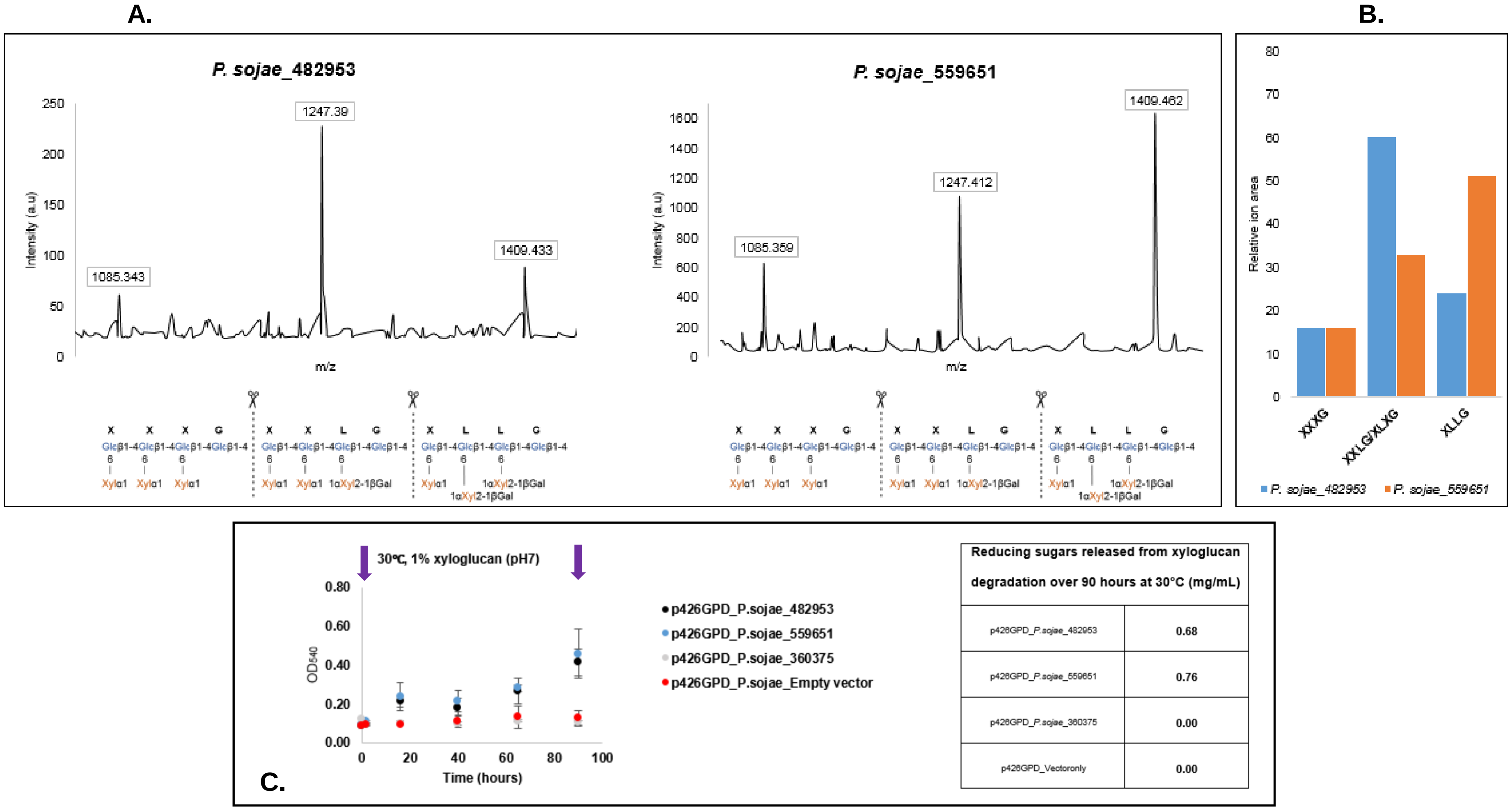
The oligosaccharides cleaved and generated during 90 hours incubation of 1% (w/v) xyloglucan with *S. cerevisiae* culture supernatants containing *P. sojae*_482953, *P. sojae*_559651, *P. sojae*_360375, and a p426-GPD vector only control (30°C, pH7), were sent for MS analysis to the NMSF at Swansea University. The reducing sugars generated by 90 hours were estimated by DNS assay (calculated based on reference to a glucose standard curve); for *P. sojae*_482953, a total of 0.68 mg/mL, and for *P. sojae*_559651, a total of 0.76 mg/mL were released, whereas 0.00 mg/mL reducing sugars were released by *P. sojae*_360375 (catalytically-inactive GH12 paralog, as discussed in Chapter 3; Ma et al., 2017) and p426-GPD vector-only negative control (Figure 4.6).

Using MALDI-MS, samples were analysed in (1) positive-reflectron and (2) positive-linear, and for each there were two preparations of the 2,5-Dihydroxybenzoic acid (DHB) matrix, (a) 1:1 MeOH:H_2O and (b) 2:1 MeCN:H_2O. The optimal combination was found to be positive-reflectron mode (1) with 2:1 MeCN:H_2O buffer for the matrix (b). For each sample, blank matrix reflectron spectra were included, as the low mass region (including putative mono- and disaccharide species) was also the region expected to show blank matrix and other background species.

MS spectra are shown in Figure 4.6 for *P. sojae*_482953 and *P. sojae*_559651 – the species (*m/z*) are indicated on the x-axis, with the relative intensity of the ions on the y-axis. Focussing on the *m/z* >1000 range (shown in the figure), three main peaks were observed for both samples after 90 hours incubation with 1% (w/v) xyloglucan (30°C, pH7); ions with the m/z of 1085, 1247 and 1409 – putatively corresponding to the oligosaccharides XXXG, XXLG (or XLXG), and XLLG, respectively (Fry et al., 1993). These species were not observed in the equivalent 0 hour samples for these enzymes, or within the 90 hour samples for *P. sojae*_360375 (catalytically-inactive GH12 paralog (Ma et al., 2017)) and the p426-GPD vector only control (spectra not shown in the figure) – consistent with the observed enzymatic activity of paralogs *P. sojae*_482953 and *P. sojae*_559651 by DNS reducing sugar assays and halo assays as previously described. Unfortunately, it was not possible to confidently identify low *m/z* species within the samples that could putatively correspond to mono- and di-saccharides (or smaller sugars) released from random or further processing of longer chain oligosaccharides by *P. sojae* GH12 paralogs - this was due to blank matrix and background ion species located in the lower *m/z* region for all samples.

The relative intensity of species identified at *m/z* 1085, 1247 and 1409 after 90 hours incubation (within single spectra for each GH12 paralog) were compared by calculating the ratios between the areas under the peaks (Figure 4.6). The results suggest that there may be putative differences in preferential binding of the xyloglucan backbone by the two enzymatically-active GH12 paralogs – *P. sojae*_482953 and *P. sojae*_559651 released similar relative ratios of XXXG, whilst there was an increase in oligosaccharides XXLG/XLXG generated by *P. sojae*_482953, and increased XLLG generated by *P. sojae*_559651.

4.5.3 *P. sojae* paralogs _482953 and _559651 cleave the same glycosidic linkages of the xyloglucan backbone, but show some preferences in the oligosaccharides bound/released



Reducing sugars released from xyloglucan degradation over 90 hours at 30°C (mg/mL)	
p426GPD_*P.*sojae_482953	0.68
p426GPD_*P.*sojae_559651	0.76
p426GPD_*P.*sojae_360375	0.00
p426GPD_Vectoronly	0.00

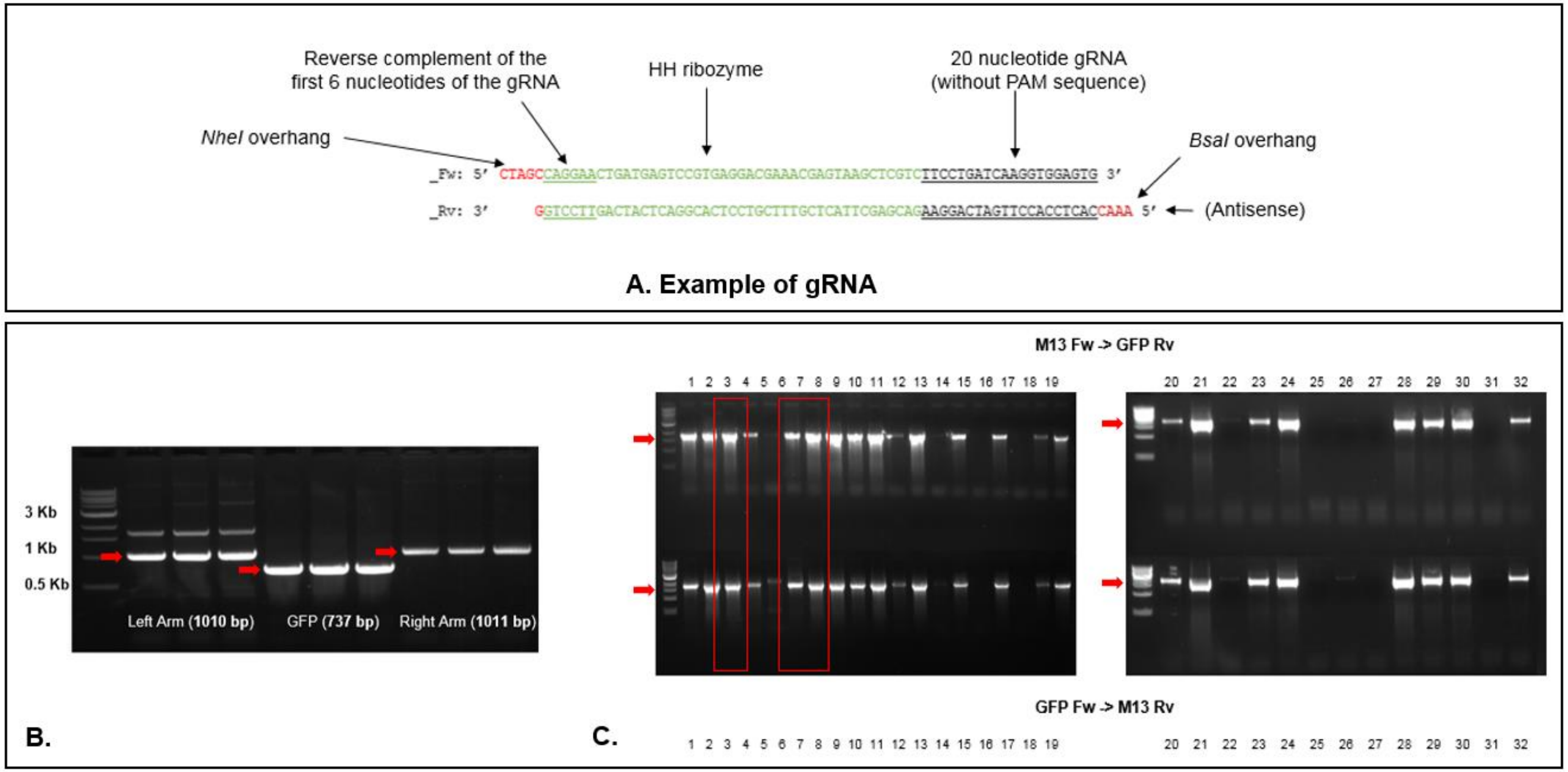
Figure 4.6. A. MALDI-MS spectra for *P. sojae*_482953 and *P. sojae*_559651 after xyloglucan incubation suggests both paralogs putatively bind to different regions of the xyloglucan chain. The species (*m/z*) are indicated on the x-axis, with the relative intensity of the ions on the y-axis. Focussing on the *m/z* >1000 range, three main peaks were observed for both samples after 90 hours incubation with 1% (w/v) xyloglucan (30°C, pH7); ions with the m/z of 1085, 1247 and 1409 – putatively corresponding to the oligosaccharides XXXG, XXLG (or XLXG), and XLLG, respectively (Fry et al., 1993). **B.** The relative intensity of species identified at *m/z* 1085, 1247 and 1409 were compared by calculating the ratios between the areas under the peaks, and suggest there may be putative differences in preferential binding of the xyloglucan backbone. **C.** The reducing sugars generated by 90 hours were quantified by DNS (calculated based on reference to a glucose standard curve); for *P. sojae*_482953, a total of 0.68 mg/mL, and for *P. sojae*_559651, a total of 0.76 mg/mL were released, whereas 0.00 mg/mL reducing sugars were released by *P. sojae*_360375 (catalytically-inactive GH12 paralog, as discussed in Chapter 3; Ma et al., 2017) and a p426-GPD vector-only negative control sample.

4.5.4 Using the CRISPR/Cas9 system to disrupt an enzymatically-active GH12 paralog in *P. sojae*

The gene encoding *P. sojae*_482953 was knocked out *in vivo*, using the CRISPR/Cas9 methods published by Fang et al. (2017) – involving design of efficient gRNA and a HDR template in order to replace the target gene sequence with GFP (Figure 4.7). PCR was used to amplify 1 kb of flanking homology arms to the target gene, as well as the GFP sequence, and the fragments were assembled by HiFi cloning into pBlueScript KS(-) – with verification of the HDR template by multiple PCR screening (Figure 4.7).

Three independent *P. sojae* mutants were tested for their ability to utilise xyloglucan as a sole carbon source by incubating hyphal plugs on minimal agar containing 1% (w/v) xyloglucan at 25°C (for 7 days, in the dark), after which colony morphology pictures were taken. Phenotypic analysis of the knock-out strains indicated that there was no significant effect to growth on xyloglucan as the sole carbon source, based on comparisons with the wild-type strain (i.e. there were no differences in the diameters of mycelial growth between isolates that would suggest an altered capacity to utilise xyloglucan resulting from the gene knockout) (Figure 4.7).

4.5.4 Knock-out of the gene encoding *P. sojae*_482953 does not affect xyloglucan utilisation as a sole carbon source *in vivo*



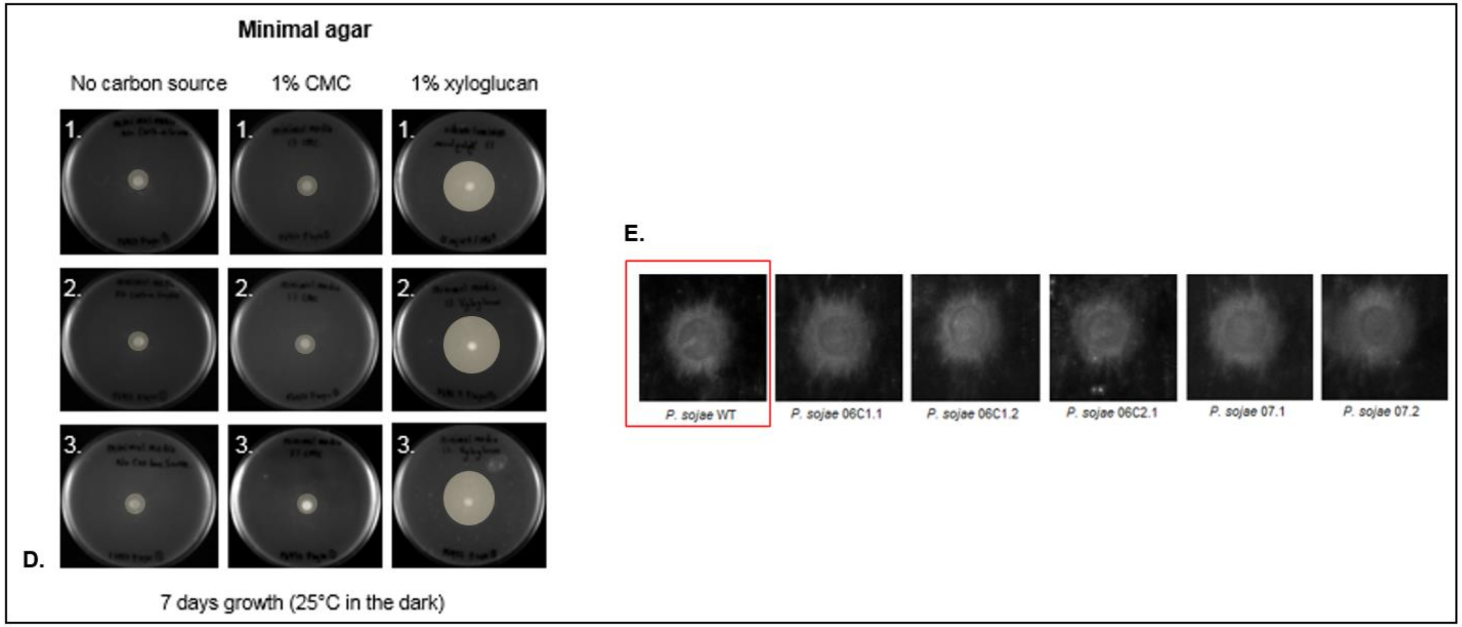


Figure 4.7. CRISPR/Cas9 methods for knockout of *P. sojae*_482953 followed as per Fang et al (2017). **A.** Example of gRNA used to direct Cas9 to the target gene sequence. **B.** A HDR template was constructed to produce a precise genome mutation (insertion of GFP sequence by homologous recombination). The upstream and downstream homologous sequences of the gene encoding *P. sojae*_482953 (~1 kb) were PCR-amplified from a *P. sojae* DNA extraction (gDNA), and the donor DNA fragment (GFP) was PCR-amplified from a plasmid DNA preparation. Red arrows indicate the expected product sizes. **C.** Two primer pairs were used to verify correct assembly of the HDR DNA fragments: (i) M13_Fw and GFP_Rv (expected amplimer: 1821 bp if the left homology arm and GFP were correctly inserted), and (ii) GFP_Fw and M13_Rv (expected amplimer: 1861 bp if GFP and the right homology arm were correctly inserted). Red arrows indicate the expected product sizes; red boxes highlight the verified products corresponding to the three transformants taken forward for plasmid DNA isolation (mini-prep) and Sanger sequencing. **D.** Growth of WT *P. sojae* on minimal media (no carbon source, 1% CMC and 1% xyloglucan); highlights were added to the images to show clearly the extent of growth for the WT strain on each media type. **E.** Growth of *P. sojae* CRISPR mutants on minimal media (1% xyloglucan) (06C1, 06C2 and 07 refer to individual transformants; '1' and '2' refer to the number of technical repeats per transformant); no differences in mycelial growth (diameter) were observed in comparison with the WT strain.

4.6 Discussion

4.6.1 *P. sojae* GH12 paralogs _482953 and _559651 are active against xyloglucan, but show no activity against CMC, Avicel or Laminarin

Hemicelluloses (including all of the cellulose-binding polysaccharides) make up around 15-35% (w/v) of a typical plant cell wall (Kubicek., 2013; Alvarez et al., 2016), and xyloglucan is an important component that plays a role in the cross-linking of the cellulose microfibrils, increasing strength (Bauer et al., 1973). As a result, plant parasites must overcome the xyloglucan layers in order to gain entry to plant tissues; xyloglucan-specific endoglucanases (xyloglucanases) are widespread amongst plant pathogens, and have been previously associated with CAZyme families GH5, GH12 and GH74 (http://www.cazy.org; Lombard et al., 2013).

Some xyloglucanases have been shown to lack activity (or are active at significantly lower levels) to CMC and other polysaccharides with β-1,4 and β-1,3 linkages (e.g. Grishutin et al., 2004; Hasper et al., 2002) – consistently, this study found that GH12 paralogs, *P. sojae*_559651 and *P. sojae*_482953 both displayed activity towards xyloglucan, with no demonstrable activity towards CMC or Laminarin (β-1,3, with some β-1,6 linkages). Enzyme activity against Avicel was also investigated – this substrate is typically used to demonstrate exo-glucanase activity (although studies such as those by Cohen et al. (2005) describe a 42 kDa endoglucanase (Cel5A) in *Gloeophyllum trabeum* with avicelase activity (Cohen et al., 2005)). However, activity towards Avicel was not demonstrated by the GH12 paralogs. A CBD has been previously shown to aid Avicel digestion (e.g. Ahn et al., 1997), which could explain the lack of exoglucanase activity amongst some endoglucanases. Nevertheless, the narrow substrate range for xyloglucan-specific endoglucanases suggests their importance for the targeted digestion of an important hemicellulosic component involved in cross-linking and strengthening the plant cell wall (Bauer et al., 1973).

The reducing sugars released from the incubation of *P. sojae*_482953 and *P. sojae*_559651 (secreted into *S. cerevisiae* concentrated culture supernatants) with carbohydrate substrates was investigated using DNS reagent (Miller., 1959). After 6 hours of incubation with xyloglucan, *P. sojae*_482953 and *P. sojae*_559651 displayed higher activity at 30°C under all pH tested, compared to incubation at 20°C – this is consistent with the optimum temperature for disease development by *P. sojae* (25-30°C; warm soil (Dorrance and Mills., 2012)). For *P. sojae*_482953, the biggest difference between temperatures was observed at pH7, where there was an increase of 0.60 mg/mL detected reducing sugars when the temperature was increased to 30°C (Figure 4.3). The same concentrated yeast supernatants used in the DNS assay were additionally spotted onto SCM-URA agar plates containing 0.2% (w/v) xyloglucan; consistently, halos were visible after staining for both *P. sojae*_482953 and *P. sojae*_559651 GH12 paralogs, in comparison with a p426-GPD vector only sample (negative control), confirming the observed enzymatic activity towards xyloglucan (Figure 4.3).

4.6.2 Removing the disordered C-terminus 'tail' from *P. sojae* _482953 reduces its enzymatic activity against xyloglucan

As previously mentioned, isozymes of the same enzyme could display unique activities with possibly untested adaptive consequences. Chapter Three demonstrated that GH12 paralog, *P. sojae*_482953, and orthologs in *P. cactorum* and *P. nicotiniae*, possess a disordered, significantly phosphorylated C-terminus ‘tail’ (unable to be modelled to available protein structures using Phyre2 (Kelly and Sternberg., 2009)) – therefore, using computational tools alone, it was not possible to identify the functional significance of the evolved ‘tail’. It was hypothesised that the unique sequence (of 186 amino acids) could putatively play a role in the enzymatic activity of the paralog *in vivo* (i.e. improving catalytic activity, substrate binding, or tolerance across broader temperature and pH ranges), therefore a truncated version of the protein was engineered and expressed in *S. cerevisiae*.

Results of the DNS reducing sugar assays and the halo assays suggest that the disordered C-terminal sequence improves enzymatic activity of the protein towards xyloglucan (Figure 4.5). Incubation of the full-length protein with the substrate resulted in the release of a significantly higher concentration of reducing sugars compared to the truncated protein. It was supposed that *P. sojae*_482953 (truncated protein) could be active at a slower rate compared to the full-length protein (and could indicate that the ‘tail’ sequence is putatively involved in mediating efficient or prolonged substrate binding, for example), therefore an extended incubation time of 144 hours (pH7, 30°C) was included to confirm whether a similar concentration of reducing sugars could be eventually detected by both versions of the protein. Total reducing sugars released after 144 hours (at 100 µg/mL starting concentration of total protein) were 2.03 mg/mL and 1.54 mg/mL for full length and truncated proteins, respectively. Results of halo assays were consistent with enzymatic assays – whilst both the full length and truncated proteins resulted in visible halos after staining, there were reduced halo sizes (diameters) observed for the truncated version of this enzyme (Figure 4.5). Interestingly, Li et al. (2014) describe a C-terminal extension of a xylanase (XynA

– from sheep rumen), which improves catalytic efficiency as well as increasing the range of pH and temperature in which the protein retains activity (Li et al., 2014) - consistent with the hypothesis and results for *P. sojae*_482953. The data also demonstrate that HGT followed by subsequent gene duplication can provide novel genetic material for selection to act on and increase functionality within recipient proteomes – putatively resulting in expanded protein functions between paralogs of the same gene family (in this case, an evolved 'tail' sequence which appears to enhance xyloglucan degradation). This also highlights the importance of characterising multiple paralogs of the same gene families of parasite-secreted CAZymes, in order to better understand their roles within the complex microenvironment of plant cell walls – i.e. how each protein putatively adds to the efficiency of digestive functions. Future experiments will involve characterising C-terminal tailed orthologs in *P. cactorum* and *P. nicotianae* (both full-length and truncated proteins), to confirm their significance to enzymatic activity. It would also be interesting to further investigate how the C-terminal tail affects the substrate binding affinity of the proteins, as well as their putative interactions with host immune proteins. As previously discussed, Ma et al. (2017) demonstrated a difference in host protein-binding between catalytically-inactive *P. sojae*_360375 and its closest-related (catalytically active) GH12 paralog (Ma et al., 2017) – therefore, it could be feasible to suggest that an (as yet, unmodelled) 186 amino acid C-terminal sequence encoded by *P. sojae*_482953 could result in altered interactions with putative host immune proteins, compared to a truncated protein.

4.6.3 *P. sojae* paralogs _482953 and _559651 cleave the same glycosidic linkages of the xyloglucan backbone, but show some preferences in the oligosaccharides bound/released

As previously mentioned, xyloglucan is a highly-branched polysaccharide with different side chain variants branching from the main glucose backbone (Fry et al., 1993). Previous analysis of the oligosaccharides released by endoglucanase digestion of the xyloglucan chain has come from (for example) MALDI-TOF screening methods to profile the units released (Oligosaccharide Mass Profiling; OLIMP (e.g. Lerouxel et al., 2002)) - this has also enabled the characterisation of oligosaccharides that make up plant cell wall polysaccharides including xyloglucan (e.g. Obel et al., 2009; Westphal et al., 2010).

In this study, the oligosaccharides released from xyloglucan degradation by GH12 paralogs, *P. sojae*_482953 and *P. sojae*_559651, were investigated by MALDI-MS analysis carried out at the NMSF at Swansea University. Following incubation with xyloglucan for 90 hours, three peaks were observed in the *P. sojae*_559651 spectra correlating to the oligosaccharides XXXG, XXLG (or XLXG), and XLLG, respectively (Fry et al., 1993)), and for *P. sojae*_482953, one significant peak was observed correlating to XXLG (or XLXG)), with smaller peaks of XXXG and XLLG. These peaks were not present at 0 hours, consistent with their release over the duration of the experiment (90 hours), and in line with observable xyloglucanase activity by DNS assay for these samples. The putative oligosaccharide peaks are not observable for *P. sojae*_360375, consistent with this being a catalytically inactive GH12 paralog (Ma et al., 2017), or the p426-GPD vector only control. It was not possible to distinguish between the oligosaccharides XXLG and XLXG in this study, as they are structural isomers and remain unresolved (although other studies have used further MALDI-PSD or –TOF/TOF analysis to distinguish them (e.g. Yamagaki et al., 1998)).

Relative intensities of the oligosaccharide species identified suggest putative differences in binding to the xyloglucan backbone by the two enzymatically-active GH12 paralogs. *P. sojae*_482953 and *P. sojae*_559651 released similar relative ratios of XXXG[9], whilst there was an increase in oligosaccharides XXLG/XLXG generated by *P. sojae*_482953, and increased XLLG generated by *P. sojae*_559651 (Figure 4.6). Therefore, it is possible that the two GH12 paralogs display binding preferences for the xyloglucan backbone (by, for example, recognising specific side chains (Fry et al., 1993)). Whilst it was not possible to optimise a suitable substrate-binding assay during this study to investigate differences in oligosaccharide binding, further work could explore the relative binding affinities of *P. sojae*_482953 and *P. sojae*_559651 to different oligosaccharide units, in order to identify putative differences amongst the paralogs. It would be feasible to hypothesise that in addition to increasing transcriptional dosage by the expression and secretion of multiple enzyme paralogs (as well as putatively widening resulting functions), a greater efficiency of *in vivo* xyloglucan digestion could be achieved through the secretion of two catalytically-active GH12 paralogs with preferences in binding to different parts of the xyloglucan chain.

[9] G refers to an unsubstituted glucose residue; X refers to a glucose residue substituted with an α-linked xylose; L refers to a glucose residue substituted with an α-linked xylose further substituted with a β-linked arabinose; F refers to a glucose substituted with an α-linked xylose further substituted with a fucose residue (Fry et al., 1993).

The MS spectra generated in this study were not appropriate to detect the release of mono- or disaccharide units resulting from xyloglucan degradation, therefore it would also be interesting to explore the further processing of the larger oligosaccharide chains into simpler sugars using alternative spectrometry techniques.

4.6.4 Knock-out of the gene encoding *P. sojae*_482953 does not affect xyloglucan utilisation as a sole carbon source *in vivo*

The gene encoding *P. sojae* _482953 was knocked out *in vivo*, using the CRISPR/Cas9 methods published by Fang et al. (2017). Phenotypic analysis of the knock-out strains indicated that there was no significant effect to growth on xyloglucan as the sole carbon source, based on comparisons with the wild-type strain (i.e. there were no differences in the diameters of mycelial growth between isolates that would suggest an altered capacity to utilise xyloglucan resulting from the gene knockout) (Figure 4.7). The results are unsurprising, as it is likely that other functional *P. sojae* proteins (of the GH12 family, or other associated GH families with similar enzymatic activities (see Chapter 3)) are able to compensate for the deletion of *P. sojae*_482953 and conceal the loss of function. Whilst it was not possible within the duration of this study, it would be interesting to further explore the effect of the knock-out to the expression of the remaining GH12 paralogs (and other putatively associated GH genes) to confirm the transcriptional redundancy of the paralogs – this could be achieved through qPCR, although the experiment would require dedicated design of appropriate qPCR primers for all genes to be tested. Alternatively, RNA of the mutant *P. sojae* strains grown on minimal media substituted with xyloglucan could be extracted at multiple time points and subject to RNA (transcriptome) sequencing – this would enable the study of the wider impact of the gene deletion for secreted xyloglucanase activity in this organism. Multiplexed CRISPR/Cas methods (i.e. knocking out expression of multiple genes) would also be a significant step to better understanding the functional significance of the horizontally-transferred GH12 (and GH10) duplicated paralogs in *P. sojae* – to date, there are no reported studies of large gene family knock-outs in this organism, although multiplexed CRISPR methods have been described in (for example) *Trypanosoma cruzi* (Peng et al., 2015) and *Candida albicans* (Vyas et a., 2015). The key strategy in this case would be in the design of the gRNA (which could ideally target homologous regions of the paralogous sequences) (e.g. see Hyams et al., 2018 for developments in algorithms for efficient gRNA design).

Whilst host virulence was not assayed during the study, it would be interesting to test pathogenicity of the knockout mutant of *P. sojae*_482953 – particularly as previous studies by Ma et al. (2015) demonstrated that overexpression and silencing of the paralog _559651 both reduced *P. sojae* virulence *in vivo* (Ma et al., 2015).

4.6.5 *P. sojae* GH10 paralogs _527497 and _519234 are active against xylan

Two of the four *P. sojae* GH10 paralogs secreted into *S. cerevisiae* culture supernatants displayed activity towards xylan at 30°C under all pH tested. Enzymatic activity (by released reducing sugars) was detected up to 170 hours, alongside *T. reesei*_Xyn2 (Grange et al., 1996). The *T. reesei* enzyme displayed significantly higher activity (by detected reducing sugar and halo assays (Figure 4.4), suggesting increased capacity for xylan digestion (although this could alternatively be related to improved secretion resulting from expression of the proteins' native signal peptide (as per Grange et al., 1996)).

Amino acid sequence and structural analysis of *P. sojae*_527497 (Chapter Three) demonstrated that this is the only GH10 paralog not predicted to encode a putative additional carbohydrate-binding site (by 3DLigandSite (Wass et al., 2010)) – however, it does have an additional ~63 amino acid 'tail' sequence at its C-terminus (Chapter 3, Figure 3.12)), which could compensate for the lack of additional binding site by altering the proteins interaction with putative substrates (and explain the similar concentrations of reducing sugars released from xylan degradation by *P. sojae*_527497 and *P. sojae*_519234 (Figure 4.4)). Further work could involve truncating *P. sojae*_527497 to elucidate the functional significance of its C-terminal extension to enzymatic activity of the protein. Interestingly, the gene encoding *P. sojae*_519234 is expressed during *P. sojae* mycelial growth, whilst *P. sojae*_527497 is not expressed at this stage (both are expressed during infection) (See previous chapter (FungiDB; Stajich et al., 2012; Basenko et al., 2018)). This could indicate that secretion of *P. sojae*_527497 is specifically associated with the later necrotrophic switch in the *P. sojae* lifestyle, therefore it is possible that the additional C-terminal tail sequence could also play a role in 'masking' the protein from host immune detection at this stage – this would be interesting to explore further in future work.

4.7 General Conclusion

This chapter aimed to experimentally investigate *P. sojae* HGT paralogs through heterologous gene expression and secretion of GH12 and GH10 proteins in *S. cerevisiae*. GH12 paralogs, *P. sojae*_482953 and *P. sojae*_559651 are enzymatically active towards xyloglucan, with higher reducing sugars released from substrate breakdown at 30°C (compared to 20°C) (i.e. at the optimum temperature for disease development by *P. sojae* (25-30°C; warm soil (Dorrance and Mills., 2012)). MS analysis of the oligosaccharides released from xyloglucan digestion by *P. sojae*_559651 and *P. sojae*_482953 suggest there could be putative differences in binding of the paralogs to the xyloglucan chain (putatively leading to enhanced breakdown of the substrate *in vivo*). Further experiments will be useful to investigate such putative binding differences between the multi-paralog families.

GH12 paralog *P. sojae*_482953, and orthologs in other hemibiotrophic *Phytophthora* spp., possess a disordered, significantly phosphorylated C-terminal 'tail' (unable to be modelled to available protein structures (see Chapter Three; Phyre2 (Kelly and Sternberg., 2009)). The native protein and a truncated version (missing the 186 amino acid C-terminal 'tail') were both expressed and secreted into *S. cerevisiae* BY4742 culture supernatants – interestingly, the full-length protein resulted in a higher concentration of reducing sugars released during incubation with xyloglucan, indicating that the tail sequence enhances substrate breakdown. It is interesting that knock-out of the same paralog *in vivo* has no detectable effect to the utilisation of xyloglucan as a sole carbon source for *P. sojae* – although it is likely that the remaining paralogs (or alternative genes) provide transcriptional redundancy (highlighting the functional significance of multiple paralog expression for oomycete parasites that rely on liberation of carbon from the plant cell wall (particularly where most secreted enzymes would themselves be subject to digestion by host immune responses)).

Enzymatic activity was not demonstrated for many of the GH12 and GH10 paralogs, therefore their functions in the digestion of plant-specific substrates remain unclear. The secretion strategy utilized for this study was optimized using a yeast-specific N-terminal secretion signal peptide, however, it is important to note that protein secretion in heterologous hosts can be limited by a number of factors (and remains a challenge – e.g. Schroder and Friedl., 1997; Hodgson., 2003; Idiris et al., 2010), and therefore, the methods described may not have been optimal for the study of the remainder of the proteins.

Chapter 5

Development of methods to investigate the mutational landscape from non-functional to functional signal peptide sequences for a secreted protein ‘acquired’ by *S. cerevisiae*

5.1 Overview

Previously-identified HGTs transferred from fungi into the oomycete lineage (Torto et al., 2002; Richards et al., 2006; Belbahri et al., 2008; Richards et al., 2011; Misner et al., 2015; Savory et al., 2015) include a suite of secreted proteins, specifically those targeted for transit through the ER and Golgi networks into the extracellular environment by recognition of an N-terminal secretion signal peptide sequence. Whilst the previous chapters aimed to improve understanding of the paralogous functions gained from HGT and subsequent gene duplications, the present chapter seeks to develop methods that will experimentally investigate the putative degeneracy of N-terminal signal peptide sequences, alongside the hypothesis that signal peptide sequences represent a putative barrier to HGT that must be overcome - i.e. if a donor secretion sequence is incompatible with the recipients secretory pathway (or results in inefficient secretion), the signal peptide must evolve to gain compatibility with the hosts cellular requirements allowing secretion. The work here seeks to understand if a gain of a functional signal peptide is likely (i.e. there is a wide diversity of sequence variants that allow secretion), or alternatively, it is unlikely (i.e. there are few sequence variants that allow secretion).

The acquired protein function and the complexity of subsequent protein interactions are important determinants that affect the transferability of genes by HGT (e.g. Rivera et al., 1998; Jain et al., 1999; Cohen et al., 2011). High gene expression (Park and Zhang., 2012), as well as differences in codon usage between donor and recipient also represent putative mechanistic costs to the acquisition of foreign genes (due to high demand for cellular energy, the increased likelihood of aberrant amino acid translation, and the cytotoxic misfolding of proteins (Sorensen et al., 1989; Buchan., 2006; Hershberg and Petrov., 2008; Qian et al., 2012), leading to further downstream fitness costs). It has been proposed that DNA sequence features will eventually be driven to match that of the host genome (Lawrence and Ochman., 1997, 1998; Garcia-Vallve., 2003), or otherwise transferred genes with similar codon usage to the recipient may be selectively retained by the host (Callens et al., 2020 (preprint)).

What determines the transferability of genes encoding secreted proteins is less well understood. Eukaryote N-terminal signal peptides represent an interesting paradox, because they are heterogeneous (there is no consensus sequence, even amongst highly similar paralogs of the same gene family), but intracellular transit of proteins is complex and can be highly specific (Lodish et al., 2000) and therefore the sequences must successfully target protein secretion. As a result, evolution towards a ‘positive’ signal peptide is not likely to be linear but rather a landscape of putative sequence space containing sequences that positively encode signal peptide function and those that don’t, recognised by translocation machinery with varying efficiency. Whilst other studies have focussed on amino acid sequence requirements for efficient signal peptides (e.g. Ryan and Edwards., 1995; Rothe and Lehle., 1998; Nilsson et al., 2015), there is currently limited experimental data assessing the functionality of all possible N-terminal sequences for the secretion of a protein. Such data would allow us to investigate how a protein acquired laterally could theoretically traverse mutational sequence space in order to acquire, improve, or lose a functional signal peptide.

This chapter aims to develop methods that will ultimately be used in combination with high-throughput microfluidic technologies to screen large libraries of randomised N-terminal peptide sequences, using secreted enzyme activity in a heterologous yeast host as a proxy for identifying signal peptide function. This will enable the construction of a mutational sequence landscape (from non-functional to functional signal peptide sequences) for the 'acquisition' of a secreted protein by *S. cerevisiae* – improving our understanding of the evolutionary dynamics of eukaryotic horizontally-acquired secreted proteins, i.e. the ability of a protein to acquire, lose, or augment its extracellular localisation, and more generally – how malleable the yeast is as a cell factory for production of secreted proteins.

5.2 Introduction

5.2.1 Protein secretion

To be able to carry out encoded functions effectively, assembled proteins must be directed to their correct locations - either to intracellular organelles (e.g. membrane-bound, mitochondrial, ER, Golgi, or endosome-localised proteins), or the extracellular space (i.e. secreted proteins) (Bendtsen et al., 2004; Horton et al., 2007). Therefore, the correct protein localisation is central to protein function across the tree of life - so it is important to understand the molecular features that enable different types of proteins to reach their target destinations. In 1971, Gunter Blobel and David Sabatini hypothesised the presence of cis-encoded targeting sequences for the translocation of proteins (Blobel and Sabatini., 1971). The presence of such sequences was demonstrated experimentally by Milstein et al. (1972), who showed that the light chain of IgG becomes a smaller mature protein in the ER vesicles (Milstein et al., 1972) – suggesting prior cleavage of a putative targeting sequence from the nascent polypeptide. The signal hypothesis suggests that different proteins encode variant targeting sequences depending on their final destination (Blobel and Dobberstein., 1975) – these sequences are commonly referred to as N-terminal signal sequences (signal peptides or leader peptides). They are short N-terminal sequences with a cleavage site upstream of the nascent polypeptide (von Heijne., 1983, 1985, 1990), which are responsible for sorting proteins into the secretory pathway (or for transport to other organelles inside the cell). In 1985, Takahara et al. took the signal peptide from *E. coli* major outer membrane protein, OmpA, and fused it upstream of *Staphylococcus aureus* nuclease A – the signal peptide was cleaved as expected, demonstrating that a viable N-terminal signal sequence is an important determinant for secretion (Takahara et al., 1985).

In eukaryotes, protein secretion generally involves co-translational translocation of a precursor polypeptide to the ER membrane (i.e. translocation is coupled to protein synthesis), followed by vesicular transport to the Golgi (Lodish et al., 2000). This is initiated by recognition of the N-terminal signal peptide by a cytoplasmic, GTP-hydrolysing, ribonucleoprotein complex called the

Signal Recognition Particle (SRP) – comprised of one or more protein components and an RNA (Walter and Blobel., 1980; Walter and Johnson., 1994; Plath et al., 1998; Keenan et al., 2001; Egea et al., 2005; Rapoport., 2008; Akopian et al., 2013; Elvekrog and Walter., 2015). As the nascent polypeptide emerges from the ribosome, SRP binds to the hydrophobic core of the signal sequence (with high affinity - in the range of 0.05-0.4 nM (Flanagan et al., 2003)), transiently stalling translation. The SRP-polypeptide complex then binds to the SRP receptor on the ER membrane, which targets the elongating polypeptide to the ER membrane-spanning translocon, Sec61 (composed of Sec61, Sec62 and Sec63) – which acts as the translocation channel (Van den Berg et al., 2004; Crowley et al., 1994; Rapoport., 2008; Nyathi et al., 2013). The newly-translated polypeptide crosses the ER membrane in an unfolded state, the N-terminal signal sequence is enzymatically cleaved by a serine protease (signal peptidase) on the luminal side of the ER membrane (Evans et al., 1986), and the nascent polypeptide is then properly folded and may undergo further modifications in the ER, such as glycosylation (the addition of sugar chains), before being transported to the Golgi. Here, the protein can undergo further processing, before being packaged into secretory vesicles for cellular export (Lodish et al., 2000).

Co-translational translocation to the ER (or to the plasma membrane of prokaryotes) is an evolutionarily conserved mechanism involving N-terminal signal peptide recognition, but remarkably, signal peptide sequences are largely heterogeneous (even within a single species), varying in length and amino acid composition. For example, the signal peptides of *P. sojae* GH12 paralogs generally vary between 18-21 amino acids (see Chapter 3, Figure 3.4). However, N-terminal signal peptides do have in common a basic tripartite architecture -

observed for both prokaryote (Gennity et al., 1990) and eukaryote sequences. Generally, the N-terminus region consists of a short stretch of positively-charged basic amino acids (an average length of 5 residues (von Heijne., 1990)), followed by a core of ~10-15 hydrophobic amino acids (which form an α-helix that enters the lipid bilayer of the ER membrane and interacts with the Sec61 translocon), and a C-terminus that includes the signal peptide cleavage site for signal peptidase (and forms a β-sheet) (von Heijne., 1984, 1990). Experimental determination of signal peptide sequences has come from taxonomic sampling; interestingly, von Heijne and Abrahmsen (1989) compared N-terminal signal sequences across animals, plants and bacteria, and found that eukaryote signal peptides possess longer stretches of hydrophobicity within their core region than those of bacteria (von Heijne and Abrahmsen., 1989). It is widely accepted that the amino acid residues within the hydrophobic core play an important role in translocation and secretion efficiency (including in bacteria (Gennity et al., 1990) and mammals (Bird et al., 1990)). For example, deletions within this region have been shown to have detrimental effects to mammalian SRP binding (Nilsson et al., 2015), as well as impairing sorting of proteins into the SRP-dependent secretory pathway in yeast (Rothe and Lehle., 1998). Structural determination and studies of *Thermus aquaticus* Ffh (the prokaryotic counterpart to SRP) are also consistent with the importance of the hydrophobic interactions with SRP in prokaryotes as well (Keenan et al., 1998). Ryan and Edwards (1995) demonstrated that proline substitutions at most positions along the hydrophobic core of the eukaryotic glycoprotein C signal peptide (whilst maintaining near identical hydrophobicity), have differing effects on translocation (Ryan and Edwards., 1995), and Duffy et al. (2010) suggest that the hydrophobic cores between diverse species are unique and consist of unique so-called *h*-motifs that are not random (Duffy et al., 2010). Equally, other studies have shown that an excess of hydrophobicity within the sequence is also detrimental for signal peptides (e.g. Huber et al., 2005) – presumably affecting the ability of the core residues to fold into the predicted α-helix and interact with the correct cellular machinery at the ER surface.

At the N-terminus of the signal peptide, mutation of the first three amino acids of the yeast invertase N-terminal signal peptide from MRF (methionine – arginine – phenylalanine) to MFR or MFK (methionine – phenylalanine - lysine) was previously shown to reduce translocation by up to 75% (Green et al., 1989), and interestingly, an increase in protein secretion when lysine or arginine are the second residues has also been described (Puziss et al., 1992; Sugai and Tsumoto., 2010).

It is also important to consider that downstream features of the polypeptide could also play a role in efficient protein secretion – for example, yeast acid phosphatase secretion was partially disrupted when part of the N-terminus sequence following the signal peptide was deleted (Haguenauer-Tsapis and Hinnen., 1984) - although it is unclear if this was due to aberrant misfolding of the protein. Conversely, Silve et al. (1987) demonstrated that deletion of the signal peptide resulted in yeast acid phosphatase still (inefficiently) entering the secretory pathway (Silve et al., 1987), and whilst depletion of SRP abolishes the translocation of dipeptidyl diaminopeptidase B, the translocation of carboxypeptidase Y is unaffected (Hann and Walter., 1991; Brown et al., 1994) - suggesting some protein features may be recognised by an alternative mechanism (in some species at least), or that cells putatively use the extracellular space as a default ‘dumping ground’ for proteins not properly targeted to other cellular locations. It is also interesting to consider that not all eukaryotic secreted proteins possess an N-terminal secretion signal peptide, and yet many are still secreted by so called non-classical mechanisms (e.g. some secreted effectors of *P. sojae* and the filamentous fungus *Verticillium dahliae* (Liu et al., 2014)).

5.2.2 Degeneracy of protein targeting sequences

Protein transport is specific and involves complex interactions between receptors and proteins at different stages (e.g. during or after protein synthesis), as well as post-protein folding (e.g. peroxisomal-targeted proteins). N-terminal signal peptides attached to proteins destined for secretion must be recognised by the correct machinery, and avoid mis-targeting to other cellular locations – they have a loosely-defined architecture with no consensus sequence, so it is interesting to consider whether non-conventional or randomised N-terminal secretion sequences can also enable proteins to be secreted (albeit with varying specificity and efficiency of secretion). For example, previous experiments in *S. cerevisiae* revealed that around 20% of random signal peptide sequences could facilitate secretion of yeast invertase, suggesting lowered specificity for recognition by translocation machinery (Kaiser et al., 1987). Of 19,610 random transformants, the authors found that 164 inserts facilitated protein secretion (0.8%), however the frequency of inserts in frame with no stop codons was suggested to be 4% of the 19,610 sequences (i.e. 784 sequences), meaning that the frequency of signal peptide sequences that facilitated secretion was estimated at 164/784 (or 20%) in the study.

The putative degeneracy of protein-targeting sequences has been described recently by Dunn and Paavilainen (2019) - in the context of both intracellular organelle targeting and extracellular secretion (involving conceptually similar protein recognition events) (Dunn and Paavilainen., 2019). The authors suggest that degeneracy of protein targeting sequences lowers the requirements for successful protein targeting – therefore putatively enabling many eukaryotic proteins to sample multiple subcellular localisations over the course of their evolution (Bogorad., 2008; Dunn and Paavilainen., 2019) – although it is also important to note that this can lead to negative consequences as well e.g. cellular toxicity. Evolutionary retargeting of eukaryotic protein families across multiple subcellular locations has also been described in review by Gabaldon and Pittis (2015); for extant proteins, the dual-targeting (or bi-localisation) of protein disulfide isomerases has been demonstrated between the secretory pathway and other organelles - including *A. thaliana* ATPD12, which has been shown to localise independently to the ER, Golgi, vacuole and the nucleus (Cho et al., 2011; also see review by Porter et al., 2015). Moreover, in yeast under stress, non-specific targeting of intracellular proteins from the cytosol to the mitochondria has been observed (Ruan et al., 2017). Non-specific targeting of proteins to different cellular locations has obvious advantages for genomic efficiency, and in the case of novel proteins, could buffer against possible toxicity or foster subsequent adaption to new cellular compartments (e.g. the targeting of bacterial-derived HGT genes to plant plastids (Llorente., 2016)) (although it could lead to negative effects if there are not clear boundaries, as mentioned above).

Mechanisms for multiple protein targeting can include receptor competition for recognition of signal sequences, splice variants with alternative targeting sequences (e.g. the *Drosophila* protein tyrosine phosphate gene has two splice variants – one with an ER-targeting signal and the other with a nuclear localisation signal (Buszard et al., 2013)), or by process of co-translation (i.e. piggyback transport) (Porter et al., 2015). It has also previously been hypothesised that some eukaryotic proteins could have migrated from the cytoplasmic proteome to the secretome by 'acquiring' a viable N-terminal signal sequence - this could occur by (for example) extension of an open reading frame (ORF) (Dunn and Paavilainen., 2019), and in *Plasmodium falciparum*, it has been shown that signal peptides can arise from exon shuffling and random amino acid shuffling (Tonkin et al., 2008). Gene duplication events can therefore provide the raw material for enabling, disabling, or altering signal peptide function over the course of protein evolution (as also demonstrated by the work in the previous chapters of this thesis). Honigschmid et al. (2018) explored losses and gains of signal peptide sequences amongst orthologous *Enterobacterales* proteins, finding related genes both with and without signal peptides - including, for example, a pectin methylesterase, which harbours a signal peptide in two *Dickeya* species, but not in four *Pectobacteria* (Honigschmid et al., 2018), suggesting that signal peptide loss is also an important evolutionary mechanism for regulating secretory protein function.

It is interesting to consider the putative degeneracy of protein-targeting sequences in the context of HGT of secreted proteins, because it is difficult to understand how a recipient organism would initially benefit from a lateral gene transfer if the proteins N-terminal signal peptide was not recognised by the cells translocation machinery due to a mismatch. Such transfers could be selected against (unless selectively neutral or they elicit other responses that enhance their integration), therefore it is possible that liberal recognition of a newly acquired signal peptide would be advantageous for initial maintenance (albeit, putatively resulting in lower secretion efficiency). It is also possible that transferred genes with inefficient signal peptides, could themselves ‘acquire’ functional N-terminal sequences as a result of gene duplication and mutation. However, another possibility is simply the selective retention of HGTs with more compatible N-terminal signal peptide sequences i.e. those sequences that are initially recognised by the recipient’s translocation machinery. As previously mentioned, genes acquired by HGT can contain codons that do not match the host genome preferred usage – over evolutionary time, synonymous mutations (or changes in tRNA content) can better adapt the genes to host translation (e.g. Bulmer., 1991). However, there may be initial strong selection for better adapted genes - interestingly, Callens et al. (2020) recently explored the evolutionary significance of codon usage post-HGT, using *P. aeruginosa* as a recipient species and a comparative genomics approach involving ancestral gene reconstruction – finding that maintained transfers largely displayed an initial bias to the host’s own codon usage – i.e. there is evidence of selective retention of genes that more closely matched the ancestral genome at the time of transfer, with eventual convergence towards the core genome (Callens et al., 2020 (preprint)). It would be interesting to explore such selective retention in the context of acquired signal peptide sequences by HGT events - although this would initially require comprehensive investigation of functional and non-functional signal peptide sequences for a given protein in a given organism.

5.2.3 Directed evolution and DNA mutagenesis

Directed evolution is a useful tool for protein engineering, enabling a protein to be removed from its natural context to observe how alternative adaptive scenarios might affect its function (or contribute to reproductive fitness). This is an artificial process that imitates natural selection to evolve proteins, enabling us to better consider how natural adaptation to functional challenges can occur (and is therefore also a useful tool for studying the underlying principles of protein evolution itself (Peisajovich and Tawfik., 2007)). Directed evolution has enabled the discovery of novel therapeutics (Carter., 2006) (including enhanced activity of therapeutically-relevant antibodies (Barbas et al., 1994)), and improved enzymes (reviewed by Bershtein and Tawfik., 2008; Kaur and Sharma., 2008), as well as optimised strains of microorganisms for industrial applications or further downstream experiments (see review by Keasling., 2008).

DNA mutagenesis is a tool involving deliberate mutation of a nucleotide sequence, creating a population of variants, and aims to elucidate the phenotypes resulting from those mutations – e.g. the impact of altered amino acid residues to the resulting protein function. DNA mutagenesis allows us to disconnect a protein from its natural environment and has been beneficial for many biotechnology processes, enabling proteins with desired or improved properties (compared to their wild-type, extant counterparts) to be exploited for medical and biofuel industries. Variant proteins with improved stability (e.g. thermal tolerance), catalytic activity, substrate specificity, and immunogenicity have all previously been harnessed by this tool (e.g. Sandgren et al., 2003; Sriprang et al., 2006). The early mutagenesis techniques were entirely random, involving exposure of cells or organisms to chemical mutagens, or UV radiation, and selecting for the desired phenotypes (such as altered growth under specific nutrient conditions). Later development of site-specific mutagenesis then enabled researchers to make more precise nucleotide edits, offering the potential to narrow the sequence space of potential variants for screening (based on what was already know about the sequence of interest). For example, structural information could inform researchers of the amino acid positions most likely to affect function or protein stability etc. - enabling single or multiple precise amino acid substitutions to be made (the latter being known as combinatorial site-directed mutagenesis). For example, Sandgren et al. (2003) mutated Ala35 in *T. reesei*_Cel12a endo-glucanase to Val35 and recorded an increase in thermal stability of the protein of 7.7°C (Sandgren et al., 2003). This technique has also led to the characterisation of a number of modified enzymes with improved activities (site-directed mutagenesis discussed in Plapp., 2005), including the increased thermostability of *A. niger* xylanase (Sriprang et al., 2006).

As earlier mentioned, previous studies have used DNA mutagenesis tools to alter signal peptide sequences, demonstrating that the hydrophobic core plays an important role in efficient translocation (e.g. Ryan and Edwards., 1995; Rothe and Lehle., 1998; Huber et al., 2005; Nilsson et al., 2015), as well as the impact of key residues at the N-terminus for secretion (e.g. Green et al., 1989; Puziss et al., 1992; Tsumoto., 2010). However, there is currently no comprehensive data set to investigate how all possible combinations of amino acid variants of a given signal peptide sequence systematically affects protein secretion efficiency; such information would be useful to better understand the putative sequence space of signal peptide evolution, and in particular, how a secreted protein acquired laterally would need to traverse this theoretical space to gain, lose, or augment its signal peptide.

Saturation mutagenesis (or oligonucleotide-directed randomisation) refers to a method of randomising sets of codons in order to generate a library of all possible amino acids at every desired position. Targeted positions can reflect those most likely to affect protein characteristics of interest, and this technique has been previously used to enhance enzyme activity, binding affinity and thermostability (e.g. Reetz et al., 2006; Yeung et al., 2009). Saturation mutagenesis can be achieved using degenerate sequences where each base ('N') can be A, T, C or G with equal probability, i.e. NNN for a codon. To minimise stop codons, reduced codon sets such as NNK (where K can be 'G' or 'T') or NNS (where S can be 'C' or 'G') can be used (which allow all 20 amino acids to be encoded by 31 possible codons, but limits the possible stop codons to TAG only, therefore a stop codon has a 3% probability of being encoded by each set of randomised triplicate nucleotides) (e.g. Reetz et al., 2008). The randomised library can then be screened for variants of interest (e.g. improved secretion of an enzyme as a proxy for signal peptide function). The larger the library size of distinct variants, the higher the probability of exploring more of the sequence space and discovering distinct variants. Screening can be achieved through agar or microplate-based methods that detect substrate degradation, or higher-throughput methods that enable intensive screening of very large populations or libraries.

5.2.4 High-throughput library screening using droplet-based microfluidics

In order to successfully screen large libraries of variants (e.g. enzymes), efficient quantitative tools must be used to effectively sample sequence space. Traditional methods for screening randomised enzyme libraries include microplate and agar-based assays, although these can be laborious and time-consuming - reducing the throughput and potentially the total number of variants that can be sampled in the time available. Additionally, such methods may require high reagent volumes, as well as excessive consumption of single-use laboratory plasticware. Large amounts of purified proteins (and large cultures of the organism used for heterologous expression of the enzymes) are usually needed for functional assays, overall increasing the time and cost footprint for each library study. Therefore, the development of rapid, cost-effective, and high-throughput tools for diverse library characterisation has been a necessity; such developments have been valuable across diverse scientific fields including protein engineering, synthetic biology, and evolutionary biology (Peisajovich and Tawfik., 2007).

Current high-throughput, microplate-based screening methods involve advanced robots that can process ~1 assay per second (up to 100,000 a day), on microplates with up to 1536 wells (Wang et al., 2016). However, robotic approaches are limited by physical constraints such as the risk of evaporation of small (µL) volumes, capillary forces (Dove., 2003), and financial constraints (such as the high cost to set up, maintain, and run large robotics equipment). Therefore, miniaturisation of high-throughput screening with microfluidic lab-on-a-chip approaches have provided a convenient format for functional analysis with significant improvements; such tools are rapid, have high reproducibility (and repeatability), and they significantly reduce reagent consumption per experiment, as each screen is carried out on a microfluidic scale – all at a lower cost compared with microplate-based screening methods (e.g. Agresti et al., 2010; reviews by Theberge et al., 2010 and Colin et al., 2015; Gielen et al., 2016). Data are generated rapidly, therefore large sequence diversities can be explored in a considerably shorter time-frame. For example, Agresti et al. (2010) describe the screening of horseradish peroxidase (HRP) mutants in yeast (where the enzyme is displayed on the cells surface), using droplet-based microfluidics; in total, ~10^8 reactions were characterised in 10 hours using less than 150 µL total reagent volume. The authors describe the library screen as having a 1,000-fold increase in speed, a 10 million-fold decrease in reagent volume, and a 1-million-fold reduction in cost, compared to sophisticated robotic microplate screening (Agresti et al., 2010).

For enzyme screening, the basic microfluidic workflow involves encapsulating an enzyme with its substrate inside a water-in-oil droplet, implying that the product of the reaction will also be retained within the droplet (therefore each genotype and phenotype are confined within a single droplet boundary). This is a crucial concept, as it means that the phenotype is immediately linked to the proteins' genetic identity. Droplet volumes can be in the femto-nanoliter scale (e.g. for single bacteria or yeast, they are typically 10 pico-liter droplets), and each one conceptually represents a single miniature experiment that can be manipulated in its own 'micro-environment' within the droplet. For example, the droplets can be fused together, or reagents can be injected - reproducing the equivalent pipetting actions, but with tighter control over timing and with greater efficiency. The droplets created are monodisperse (the desired size can be achieved by manipulating the flow rate during droplet formation), and they are generated quickly (at kHz frequencies); commonly up to 1000+ droplets can be produced per second, meaning that large libraries (e.g. up to 10^6 clones) – ultimately representing 10^6 single and distinct mini experiments - can be encapsulated in only 20 minutes (Gielen et al., 2013, 2016).

Recombinant expression of randomised enzyme libraries by a suitable heterologous host (e.g. bacteria or yeast), theoretically results in encapsulation of single cells expressing a single variant enzyme within each droplet. As well as increasing the diversity of clones that can be sampled, this is additionally useful as single-cell phenotypes can be monitored away from heterogeneous populations – informative for characterising isolated enzyme responses. Recombinant protein targeting, either to the cell surface or extracellularly (via secretion), are beneficial for characterisation because the enzyme has to be accessible to the substrate (which is generally encapsulated around the cell within each droplet). Studies in *E. coli* involving proteins expressed and localised to the cytoplasm or periplasm relied on an additional lysis step following encapsulation, in order to release intracellular enzymes and enable detection of reaction products (e.g. Romero et al., 2015), (or instead were based on assumption that the substrate and product could traverse the cellular membrane (e.g. Beneyton et al., 2014; Hosokawa et al., 2015)). However, heterologous secretion of enzymes by well-characterised eukaryote model systems removes the need for an additional lysis step; instead, single cells can be incubated following encapsulation, allowing extracellular enzyme secretion. For example, Beneyton et al. (2017), expressed five hydrolytic genes from *Aspergillus niger* (2 endoxylanases, a cellobiohydrolase, an endoglucanase, and an aspartic protease) in *Yarrowia lipolytica*; single yeast cells were encapsulated and grown successfully in droplets (Beneyton et al., 2017).

Following incubation with the appropriate enzyme substrate, a certain proportion of the substrate is broken down and turned into a detectable product (depending on the functionality of the enzyme variant). Fluorescence-activated droplet sorting (FADS)-based methods have been used previously (e.g. Baret., 2009; Agresti et al., 2010), or else traditional flow cytometers could be used for final readouts after initial microfluidic preparations (Zinchenko et al., 2014; Fischlechner et al., 2014). Enzymatic characterisation based on absorbance detection has traditionally relied on microplate assays, or agar-based colony screens (e.g. staining insoluble substrate), the latter of which is limited by the dynamic range of reactions - reducing the ability to confidently discriminate between subtly different phenotypes. Absorbance measurements in droplets (e.g. Srinivasan et al., 2004; Trvedi et al., 2010; Gielen et al., 2013, 2016), have been useful for enzyme reactions that lack a fluorescent product for detection. In 2016, Gielen et al. reported a microfluidic device successfully used for absorbance detection with a connection to a sorting module (absorbance-activated droplet sorting (AADS)), validating it through directed evolution experiments involving a phenylalanine dehydrogenase (PheDH) (Gielen et al., 2016). Droplets of interest (i.e. those that lie above or below a threshold of interest) can be sorted by a bespoke microfluidic device (similar to a flow cytometer or FACS) - sorting droplets of phenotypic interest enables rapid recovery of their genetic fingerprints (i.e. the amino acid sequences of enzymes of interest). In the context of signal peptide sequences, rapid screening of randomised libraries would result in a comprehensive dataset linking the sequences to function (i.e. improved secretion), which can be displayed as a mutational sequence landscape.

5.2.5 Fitness landscapes

Understanding the effects of amino acid mutations to protein function (and organism fitness) is a central part of evolutionary biology, as it allows us to appreciate the complex driving forces that shape the evolutionary fates of populations. Adaptation occurs through mutation and selection (or genetic drift), therefore it is important to understand the molecular impact of different mutations to a theoretical path of protein evolution (and their selective constraints). Conceptually, proteins evolve across a sequence space where each 'step' represents a mutation that can affect its function (and therefore potentially the fitness of the organism) (Wright., 1932; Maynard-Smith., 1970; Eigen and Schuster., 1977; Sasaki and Nowak., 2003). Therefore, in sequence space, neighbouring sequences would differ by a single amino acid substitution (i.e. encoded by one codon (a set of three nucleotides)); each sequence is also associated with a fitness score (i.e. a relative value of fitness – which could relate to function, reproductive or mutation rate). Ultimately, this gives rise to a fitness landscape (Wright., 1932; Maynard-Smith., 1970) – allowing us to visualise the interplay between mutation, protein function, organism fitness, as well as epistasis (the effect of combined mutations that may stabilise or compensate for otherwise deleterious effects (e.g. Poon et al., 2005)). A combinatorial complete fitness landscape (where all possible combinations of mutations are considered) is a good strategy to fully understand all of the driving forces that underpin protein evolution (e.g. review by de Visser et al., 2011) - the resulting landscape can be *smooth* (with a single peak), suggesting narrow protein evolvability (i.e. most mutations or combinations thereof have a neutral or negative effect to function or fitness), or *rugged* (with many peaks and valleys), suggesting multiple theoretical routes for protein evolution are possible that maintain or lead to functionality (Romero and Arnold., 2009; Carneiro and Hartl., 2010).

Considering the effect of all possible mutations on an amino acid sequence as a landscape is useful, because it allows us to consider all of the possible paths in protein evolution and how they would theoretically relate to protein function. It is also interesting to appreciate the number of neutral mutations that maintain function, as well as the mutations that could putatively give rise to novel phenotypes within the landscape. Cory and Dunn (2019) describe the concept of 'evolvability' as the capacity of a population to find its way across a phenotypic environment – whether through a process of sequential mutations (as is the case for proteins inherited vertically) (Romero and Arnold., 2009), or through the phenomena of HGT and gene duplication, which enables organisms to theoretically jump across large spaces of the sequence landscape (Cory and Dunn., 2019). Protein evolution across sequence space has been studied previously for a number of different proteins, e.g. Jacquier et al. (2013) demonstrate the mutational landscape of a beta-lactamase (TEM-1) using 10,000 mutants and an enzyme activity screening by minimum inhibitory concentration (MIC) approach (Jacquier et al., 2013), whilst Meini et al. (2015) investigated the path of an alternative metallo-beta-lactamase (Meini et al., 2015). Interestingly, in a study of the local fitness landscape of *Aequorea victoria* GFP using tens of thousands of mutants, Sarkisyan et al. (2016) discovered that ¾ of proteins with one point mutation displayed reduced fluorescence, and a complete loss of function was associated with four mutations in half of the mutants tested

(Sarkisyan et al., 2016). However, there are fewer studies regarding the evolution of organelle-targeting sequences, including signal peptides, which are also required to traverse phenotypic space during evolution. Williams et al. (2000) analysed the signal peptides of 76 mouse-rat orthologs, suggesting that many mutations are under stabilising selection, evolving half as fast as neutral sequences (Williams et al., 2000), but there is currently no experimental data exploring a complete sequence landscape of signal peptides sequence and their functional consequences to the secretion of a protein. This limits our knowledge about whether signal peptide sequences represent a putative barrier to the cross-phylum acquisition of secreted proteins by HGT in eukaryotes.

5.3 Aims of chapter

Signal peptides play an important role in determining translocation and the entry of proteins into the secretory pathway (and therefore also affect protein-folding and other modifications required prior to secretion). Eukaryote signal peptides share a basic tripartite structure, but the sequences are heterogeneous, even amongst paralogs of the same gene family – and yet, protein targeting to the secretory pathway is a complex and specific process involving recognition of the N-terminal signal peptide by translocation machinery. Therefore, it is interesting to consider how the function of a secreted protein acquired laterally through cross-phylum HGT could be restricted by signal peptide differences - in particular, whether signal peptides represent a putative barrier for HGT into eukaryotes (i.e. if a donor secretion sequence is incompatible with the recipient's secretory pathway (or results in inefficient secretion)).

The aim of this chapter is to develop gene-to-phenotype methods for a mutational perturbation study that will be used for a larger effort to assess the efficacy of signal peptides in the secretion of a xylo-glucanase synthetically 'acquired' by *S. cerevisiae*. Ultimately, this will enable screening of large libraries of randomised signal peptide sequences (cloned upstream of the xylo-glucanase missing its native secretion sequence), using secreted enzyme activity in the recombinant yeast host as a proxy for signal peptide function and efficacy. Assessment of all possible variant signal peptides will facilitate the construction of a mutational sequence landscape – a putative sequence space containing both positive and negative signal peptide sequences for the secretion of a heterologously-expressed ('acquired') xylo-glucanase in *S. cerevisiae*. This landscape will be informative for exploring how a protein acquired laterally could theoretically traverse sequence space in order to evolve, improve, or lose a functional signal peptide, as well as improving insights into the evolution of secretion sequences. By inferring how costly different mutations may be, this will improve our understanding of the evolutionary dynamics relating to successful acquisition and processing of foreign secreted proteins gained through HGT events, as well as provide a better understanding of the relationship between sequence variation and protein coding.

Degenerate signal peptide libraries will be generated and transformed into *S. cerevisiae*, using two different approaches to screen the secreted xylo-glucanase activities: (i) Congo red staining of agar plates and halo detection around recombinant *S. cerevisiae* colonies, followed by PCR and Sanger Sequencing to identify the corresponding signal peptide sequences, and (ii) Congo red absorbance detection in liquid droplets, using microfluidics to encapsulate single recombinant *S. cerevisiae* cells in media containing xyloglucan (inferring the breakdown of the substrate by secreted xylo-glucanase from the absorbance shift of Congo red)[10].

5.4 Methods

5.4.1 Expression of the gene encoding *P. sojae*_482953 without an N-terminal signal sequence

Primers were designed to amplify *P. sojae*_482953 without its N-terminal signal sequence, with 5' *HindIII* and 3' *ClaI* restriction site sequences. The PCR master mix was prepared as follows (1x): in a sterile 1.5 mL tube, 5 µL Q5 Reaction Buffer (5x), 0.5 µL DNTPs (10 mM), and 1.25 µL of each forward and reverse primers were added to 15.75 µL H_2O. Q5 High-Fidelity DNA Polymerase was thawed on ice and added (0.25 µL) to the PCR mix. 1 µL plasmid DNA (5 ng/µL) containing the full-length gene was used as the template for PCR amplification. PCR conditions were as follows: denaturation at 98°C (5 Minutes), annealing at 98°C (10 seconds) + 50-70°C (30 seconds) + 72°C (2.5 minutes) (30 cycles), followed by extension at 72°C for 10 minutes. Amplification of the 1221 bp (plus the additional bases required for restriction enzyme cloning) was confirmed by gel electrophoresis, and the PCR product was purified using Thermo Scientific GeneJET PCR purification kit according to the manufacturer's protocol.

The purified PCR product (insert) and the p426-GPD plasmid backbone (vector) were prepared as follows: 30 µL insert and plasmid p426-GPD were digested with 1µL *HindIII* and *ClaI* (each 10 units in total) in 5 µL Cutsmart Buffer

[10] Thanks to Dr Fabrice Gielen and Vasilis Anagnostidis (University of Exeter) for their help and support during the micro-droplet component of the project.

(10x) and 14 µL H_2O, overnight at 37°C. The digested insert was PCR-purified as previously described. The digested plasmid was confirmed by gel electrophoresis, and a linear band was excised for gel purification using Thermo Scientific GeneJET Gel Extraction kit, according to the manufacturer's instructions. Purified plasmid DNA was quantified using a Nanodrop. The vector was additionally phosphatase-treated to prevent self-ligation as follows: 10 µL of the vector was incubated with 1 µL Antarctic Phosphatase reaction buffer (10x) and 1 µL Phosphatase at 37°C for 30 minutes, followed by enzyme inactivation at 70°C for 5 minutes, and subsequently stored on ice (Figure 5.2).

The DNA insert (8 µL) and the vector (2 µL) were ligated in 2 µL T4 DNA ligase buffer (10x), 0.4 µL T4 DNA ligase (400 U), in a total of 20 µL. The ligation reaction was incubated at room temperature overnight, and transformed into chemically-competent *E. coli* DH5α– 2 µL of ligation reaction was added to 50 µL competent cells, incubated on ice for 30 minutes, heat-shocked at 42°C for 30 seconds, and incubated on ice for 2 minutes. Cells were recovered by adding 250 µL LB and were incubated at 37°C for 1 hour with shaking. 100 µL of cells were inoculated onto LB agar (supplemented with 100 µg/mL ampicillin for plasmid selection). Agar plates were incubated at 37°C overnight. Three transformants were re-streaked on fresh agar and used to prepare liquid cultures (one colony inoculated into 5 mL LB-amp broth in sterile 50 mL tubes, incubated at 37°C overnight with shaking). Plasmid extraction was carried out using Thermo Scientific GeneJET plasmid miniprep kit according to the manufacturer's protocol, and purified plasmid DNA was quantified using a Nanodrop. The correct *P. sojae*_482953 gene sequence (without its N-terminal signal sequence) in the vector was confirmed by Sanger Sequencing, as previously described (Figure 5.2).

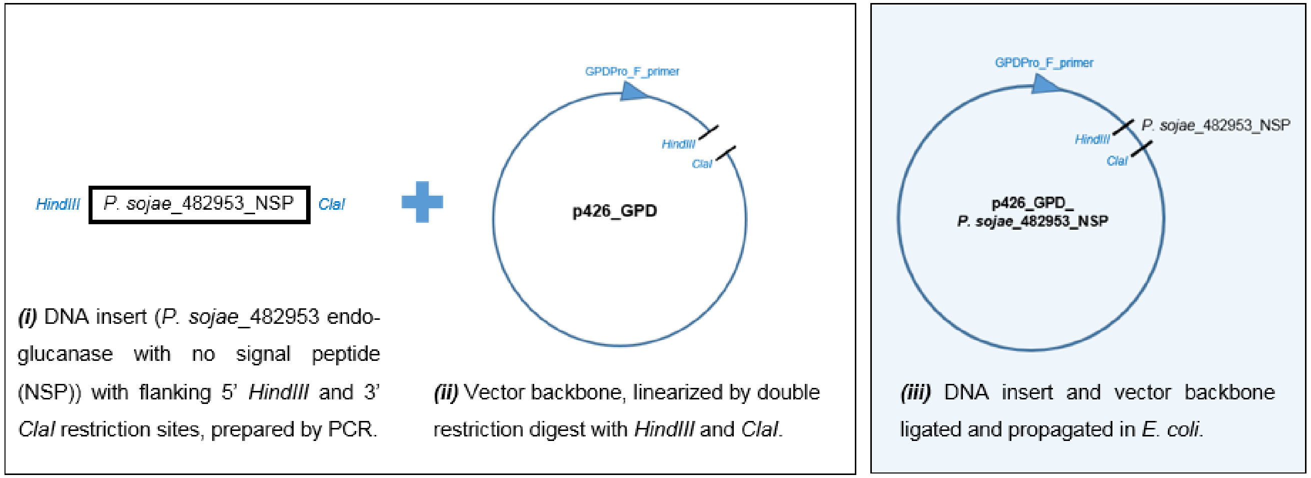
5.4.2 Generation of the degenerate N-terminal signal peptide library

Single-tube HiFi reactions were prepared in order to generate the degenerate signal peptide library (upstream of *P. sojae*_482953 missing its native signal peptide in plasmid p426-GPD). A forward degenerate oligonucleotide primer (117 bp) was designed to insert random N-terminal signal peptide sequences (ATG followed by NNK^{18} (where K can be ‘G’ or ‘T’ - allowing all 20 amino acids to be encoded but limits the stop codons to TAG only)) upstream of the gene encoding *P. sojae*_482953, with 5’ *XmaI* and 3’ *HindIII* restriction sites (including 24 bp overlaps to the adjacent plasmid and gene sequences, for efficient assembly) (Figure 5.3).

5’ acggattctagaactagtggatcc-cccggg-ATG(NNK)^{18} -aagctt-gctgaattctgtgatcagtggggc 3’

Figure 5.1. Degenerate oligonucleotide primer (117 bp); blue boxes indicate the 24 bp 5’ and 3’ overlaps to the adjacent plasmid p426-GPD and gene sequences, respectively, red lettering indicates the 5’ *XmaI* and 3’ *HindIII* restriction sites for cloning; yellow box indicates the degenerate signal peptide sequence (ATG start codon followed by NNK^{18}).

The plasmid backbone (encoding the gene for *P. sojae*_482953 without an N-terminal signal sequence) was prepared as follows: 3 µg plasmid was digested with *XmaI* and *HindIII* (each 15 units in total) in 1x CutSmart buffer, overnight at 37°C. Digested plasmid DNA was confirmed by gel electrophoresis, and linear bands were excised for gel purification. Purified plasmid DNA was quantified using a Nanodrop. NEBuilder HiFi DNA Assembly cloning kit was used to clone the 2 DNA fragments together – a total of 0.03-0.2 pmols of DNA was used in the reaction; pmols per DNA fragment was calculated using the following equation: pmols = (weight in ng x 1000) / (base pairs x 650 Daltons). For example, 50 ng of 500 bp DNA is 0.15 pmols. A total of 0.02 pmols of the vector and a total of 0.1 pmol of the degenerate insert (i.e. a 1:5 ratio) was used. Multiple reactions were prepared in sterile 1.5 mL tubes as follows: 10 µL of NEBuilder Assembly master mix (2x), DNA fragments (total of 0.03-0.2 pmols), with dH_2O added to a total volume of 20 µL. The reactions were incubated at 50°C for 60 minutes, and subsequently stored on ice or at -20°C (Figure 5.3).

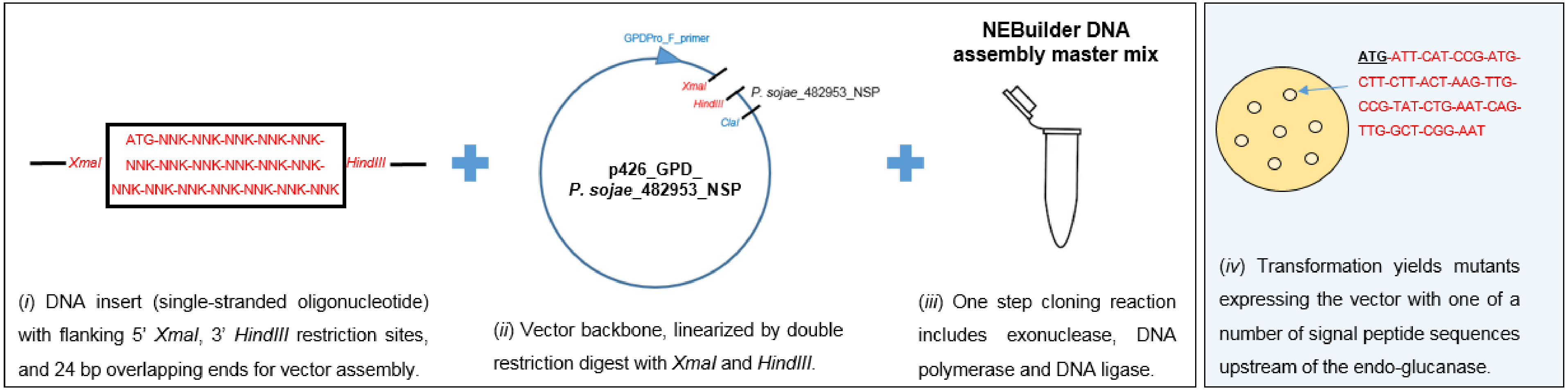


(i) DNA insert (*P. sojae*_482953 endo-glucanase with no signal peptide (NSP)) with flanking 5' *HindIII* and 3' *ClaI* restriction sites, prepared by PCR.

(ii) Vector backbone, linearized by double restriction digest with *HindIII* and *ClaI*.

(iii) DNA insert and vector backbone ligated and propagated in *E. coli*.

Figure 5.2. Construction of the p426-GPD vector expressing *P. sojae*_482953 missing its native N-terminal signal peptide. *P. sojae*_482953 (missing its native signal peptide) was amplified by PCR and digested with 5' *HindIII* and 3' *ClaI*. The plasmid backbone (p426-GPD) was digested with 5' *HindIII* and 3' *ClaI*. The insert and vector were ligated with T4 DNA ligase, and the constructed plasmid was propagated in *E. coli*. Three independent transformants were confirmed to harbour the correct sequence by Sanger Sequencing.



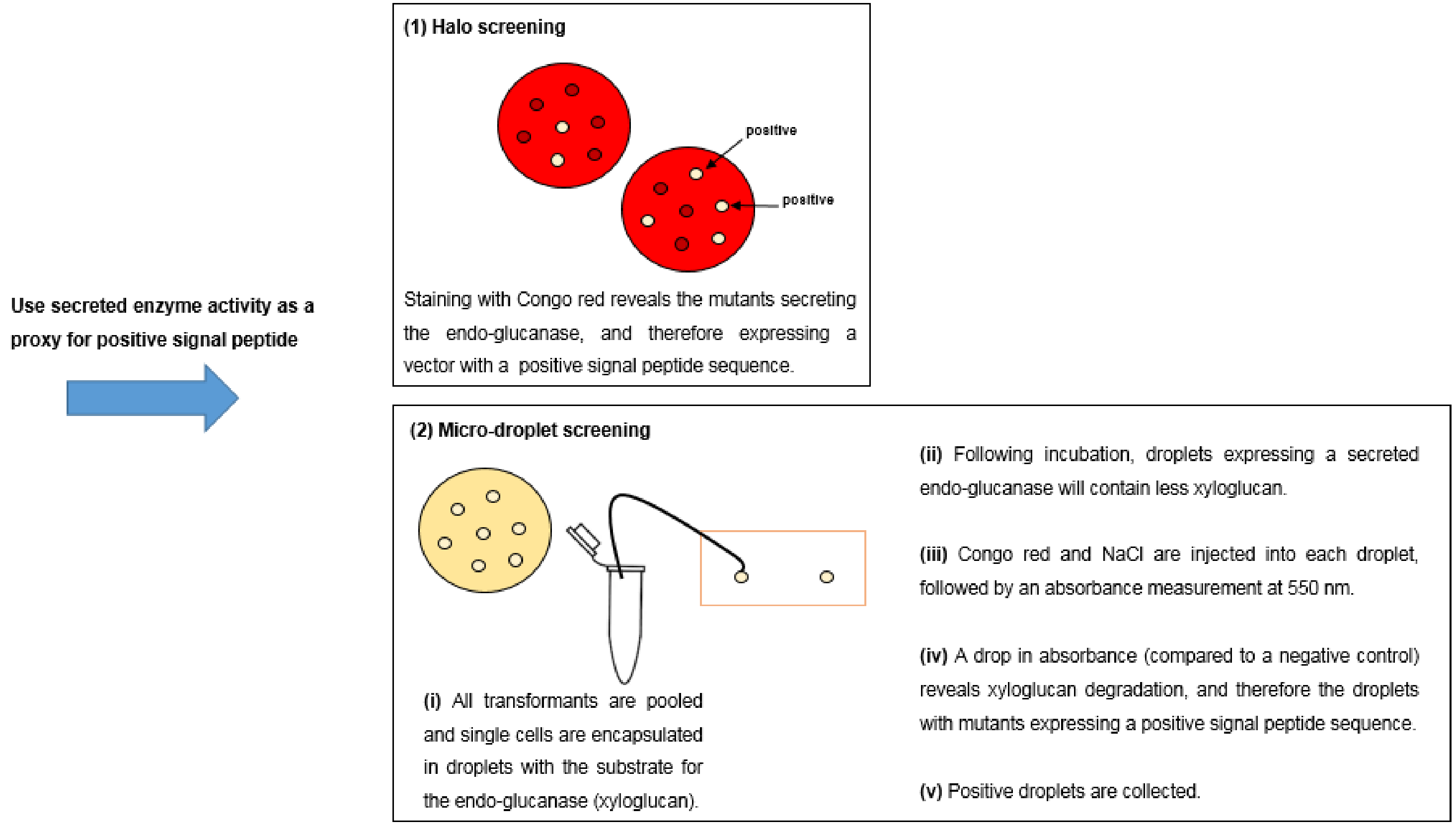
(*i*) DNA insert (single-stranded oligonucleotide) with flanking 5' *XmaI*, 3' *HindIII* restriction sites, and 24 bp overlapping ends for vector assembly.

(*ii*) Vector backbone, linearized by double restriction digest with *XmaI* and *HindIII*.

(*iii*) One step cloning reaction includes exonuclease, DNA polymerase and DNA ligase.

(*iv*) Transformation yields mutants expressing the vector with one of a number of signal peptide sequences upstream of the endo-glucanase.

Figure 5.3. Construction of the signal peptide library. A forward degenerate oligonucleotide primer (117 bp) was designed to insert random N-terminal signal peptide sequences (ATG followed by NNK^{18}) upstream of the gene encoding *P. sojae*_482953 in the plasmid constructed in Figure 5.2. The DNA insert and the plasmid backbone were digested with 5' *XmaI* and 3' *HindIII* and the DNA fragments were assembled using a NEBuilder HiFi DNA Assembly cloning kit (a one-step cloning reaction containing exonuclease, DNA polymerase and DNA ligase). Cloning reactions were transformed into *E. coli* and *S. cerevisiae* (each transformant theoretically expresses one of a number of possible signal peptide sequences upstream of the gene encoding *P. sojae*_482953).



Use secreted enzyme activity as a proxy for positive signal peptide

(1) Halo screening

Staining with Congo red reveals the mutants secreting the endo-glucanase, and therefore expressing a vector with a positive signal peptide sequence.

(2) Micro-droplet screening

(i) All transformants are pooled and single cells are encapsulated in droplets with the substrate for the endo-glucanase (xyloglucan).

(ii) Following incubation, droplets expressing a secreted endo-glucanase will contain less xyloglucan.

(iii) Congo red and NaCl are injected into each droplet, followed by an absorbance measurement at 550 nm.

(iv) A drop in absorbance (compared to a negative control) reveals xyloglucan degradation, and therefore the droplets with mutants expressing a positive signal peptide sequence.

(v) Positive droplets are collected.

Figure 5.4. Overview of the two screening approaches for the signal peptide libraries. Transformants generated in Figure 5.3 can be screened for secreted enzyme activity as a proxy for signal peptide function. Method (1) involves staining agar plates with 0.2% (w/v) Congo red (Sigma), and de-staining with 1 M (w/v) NaCl for 30 minutes (e.g. Wood and Weisz., 1987). A clearing (or halo) around colonies indicates extracellular enzyme activity. Method (2) involves using a micro-droplet screen to identify secreted enzyme activity of single cells in droplets, by a shift in Congo red absorbance (indicating it is no longer bound to soluble substrate - e.g. Haft et al., 2012).

5.4.3 Pilot study of the N-terminal signal peptide library in *E. coli*

A pilot study in *E. coli* aimed to validate the HiFi cloning method and use of a degenerate oligonucleotide primer as the insert DNA upstream of the gene encoding *P. sojae*_482953 (without it's N-terminal signal peptide). Three assembled HiFi fragment mixtures were transformed into chemically-competent *E. coli* using a standard transformation procedure as previously described. Each transformation reaction was inoculated across three 12 cm x 12 cm LB agar plates (supplemented with 100 µg/mL ampicillin for plasmid selection). Agar plates were incubated at 37°C overnight (Figure 5.3). Across each of the three sets of HiFi reaction plates, a total of 50 individual colonies were re-streaked on fresh agar and used to prepare liquid cultures (one colony inoculated into 5 mL LB-amp broth in sterile 50 mL tubes, incubated at 37°C overnight with shaking). Plasmid extraction was carried out using Thermo Scientific GeneJET plasmid miniprep kit according to the manufacturer's protocol, and purified plasmid DNA was quantified using a Nanodrop. Cloned N-terminal signal peptide sequences were identified using Sanger Sequencing, as previously described (using a M13_Rv primer).

5.4.4 Pilot study of the N-terminal signal peptide library in *S. cerevisiae* (halo screening method)

5 µL per HiFi reaction (from section 5.4.2) was transformed directly into *S. cerevisiae* BY4742 using a basic electroporation method, as previously described. Each entire transformation reaction was inoculated onto a SCM-URA agar plate (+ 0.2% (w/v) xyloglucan) and incubated at 30°C for 48 hours. This resulted in approximately 200 individual *S. cerevisiae* transformants per HiFi reaction transformed. A positive control (*P. sojae*_482953 expressing the *S.*

cerevisiae MFα N-terminal signal peptide (as used for experiments in the previous chapters of this thesis)), and a negative control (*P. sojae*_482953 expressing no signal peptide) were also included alongside the yeast transformations.

Colonies were picked into 96-well plates containing ~200 µL SCM-URA liquid media to maintain a stock, and the original plates stained with Congo red - colonies were washed off the plates, and the remaining intact polysaccharide stained with 0.2% (w/v) Congo red (Sigma) for 30 minutes at room temperature, and de-stained with 1 M (w/v) NaCl for 30 minutes at room temperature (Wood and Weisz., 1987). H_2O adjusted to pH 2 with hydrochloric acid (HCl) was used to intensify the stain. Extracellular enzyme activity was indicated by a clearing or 'halo' around colonies of enzyme-secreting strains (Figure 5.4 (1)). Additionally, 31 of the stocked *S. cerevisiae* transformants were cultured individually in 5 mL for 24 hours at 30°C (with shaking), and the cell pellets were washed once and resuspended in H_2O. 10 µL spots of OD_{600} –matched cells (0.5-1) were spotted onto SCM-URA agar containing 0.2% (w/v) xyloglucan (with 2% (w/v) glucose as an additional carbon source). Agar plates were incubated at 30°C for 48 hours. Yeast colonies were washed off the plates, and the remaining intact polysaccharide on the plates was stained with 0.2% (w/v) Congo red as before to confirm extracellular enzyme activities of recombinant *S. cerevisiae* strains.

The 31 *S. cerevisiae* transformants were subject to colony PCR using GPD pro-F and 482953_NT_ClaI_Rv primers (PCR master mix and thermocycler conditions as previously described in this chapter). Amplification of the 1420 bp gene product was confirmed by gel electrophoresis, and the PCR product was purified using Thermo Scientific GeneJET PCR purification kit according to the manufacturer's protocol. Cloned N-terminal signal peptide sequences were identified by Sanger Sequencing, as previously described (using a GPD pro-F primer).

5.4.5 Screening secreted xylo-glucanase activities by recombinant *S. cerevisiae* using microfluidics

In order to screen large numbers of N-terminal signal peptides by secreted enzyme activity in *S. cerevisiae*, it would be advantageous to develop a high-throughput method that can simultaneously phenotype very large libraries of yeast variants. Haft et al. (2012) developed a simple colorimetric microplate assay to quantify cellulose degradation by the shift in Congo red absorption when no longer bound to soluble cellulose - this method was adapted for the present study to infer xyloglucan degradation by single *S. cerevisiae* mutants compartmentalised in water-in-oil droplets. The ultimate aim is to develop a high-throughput microdroplet tool to rapidly screen and quantify secreted xylo-glucanase activities in *S. cerevisiae* based on Congo red absorbance shift (Haft et al., 2012) (Figure 5.4 (2)). Firstly, it was important to establish that droplets containing *S. cerevisiae* cells secreting an active enzyme could be distinguished from droplets containing cells not secreting an active enzyme (i.e. a vector-only control) - by the shift in Congo red absorbance, indicating substrate breakdown. The appropriate starting concentration of xyloglucan and the concentration of Congo red required for detection were confirmed by a microplate assay: SCM-URA media containing 1%, 0.5%, 0.25% and 0% (w/v) xyloglucan were mixed with 0.4%, 0.3%, 0.2% and 0.1% (w/v) Congo red (+ 1 M NaCl for colour development), and the absorbance measured at 550 nm (Figure 5.7). This revealed the optimal combination of 1% xyloglucan with 0.1% Congo red – the biggest difference was observed between Congo red absorbance of 1% xyloglucan, and the absorbance of 0% xyloglucan (i.e. SCM-URA only), which would indicate complete substrate breakdown in a hypothetical sample (Figure 5.7).

S. cerevisiae BY4742 samples were prepared for the microdroplet study as follows: *P. sojae*_482953 expressed in plasmid p426-GPD downstream of the *S. cerevisiae* MFα N-terminal signal peptide sequence was used as the positive control because secreted enzyme activity was already demonstrated using this construct (see Chapter 4), and a p426-GPD vector-only was used as a negative control. *S. cerevisiae* transformants were cultured individually in 5 mL for 24 hours at 30°C (with shaking), and the cell pellets were washed once and resuspended in SCM-URA (+ 1% (w/v) xyloglucan) at an OD_{600} of 0.1 (0.34 X 10^{7} cells per mL) (Beneyton et al., 2017)).

The polydimethylsiloxane (PDMS) microfluidic chip was designed and fabricated by Dr Fabrice Gielen and Vasilis Anagnostidis (University of Exeter); chips were operated using syringe pumps (neMESYS) using 1 mL disposable or glass syringes; chips were visualised under a light microscope (LED device: coolLED pE 100 (38% intensity)), using Vimba Viewer 2.1.3 (brightness/exposure of 40-100 µs and gain to 5-10 db). Droplet generation and collection were operated with two liquid streams – (i) 1% oil and (ii) the cell suspension. Positive and negative *S. cerevisiae* samples were encapsulated using flow rates for oil and cell preparation of 15 µL/min and 3 µL/min, respectively; droplets were collected in 0.2 mL tubes and incubated at 30°C for 24 hours. Following incubation, droplets were reinjected into the microfluidic chip; equal volumes of 0.1% Congo red and 1 M NaCl were premixed in a 1.5 mL tube and injected into each droplet passing alongside channels filled with salt solution (5 M NaCl) - salt water electrodes have been previously used to substitute metal electrodes for droplet pico-injection (e.g. Sciambi and Abate., 2014). The voltage signal was recorded for a series of droplets read in sequence using a custom Labview program, and data analysed using a custom MatLab script developed by Dr Gielen.

5.5 Results

5.5.1 Pilot study of the N-terminal signal peptide library in *E. coli*

Three independent HiFi reactions were transformed into chemically-competent *E. coli*; a total of 50 colonies were selected for plasmid extraction and Sanger Sequencing in order to identify putative signal peptide sequences.

All 50 sequences were analysed manually and were all unique to one another – 42% had at least one stop codon within the sequence (21 sequences), 10% were an ‘unexpected length’ (i.e. not 57 bp) (5 sequences), and 48% were expected 57 bp sequences with no stop codons, and therefore had the potential to act as viable signal peptides (24 sequences). This latter set of 24 sequences were taken forward for further analysis. WebLogo (Crooks et al., 2004) was used to show the frequency of nucleotides at each position amongst the putative signal peptide sequences (Figure 5.5). The 24 nucleotide sequences were analysed for G/C content (21.8-61.4%), and the translated amino acid sequences were analysed for % hydrophobicity (26-74%); SignalP 3.0 (Bendtsen et al., 2004) with default eukaryote parameters was used to identify putative N-terminal cleavage sites when the putative signal peptide amino acid sequences were pasted upstream of *P. sojae*_482953 (missing its native signal peptide) (Table 5.1). Two of the sequences generated a prediction of signal peptide cleavage (sequences 2.16 and 3.18, which predicted cleavage sites +1 and +10 amino acids following the signal peptide sequence, respectively Table 5.1).

5.5.1 Pilot study of the N-terminal signal peptide library in *E. coli*

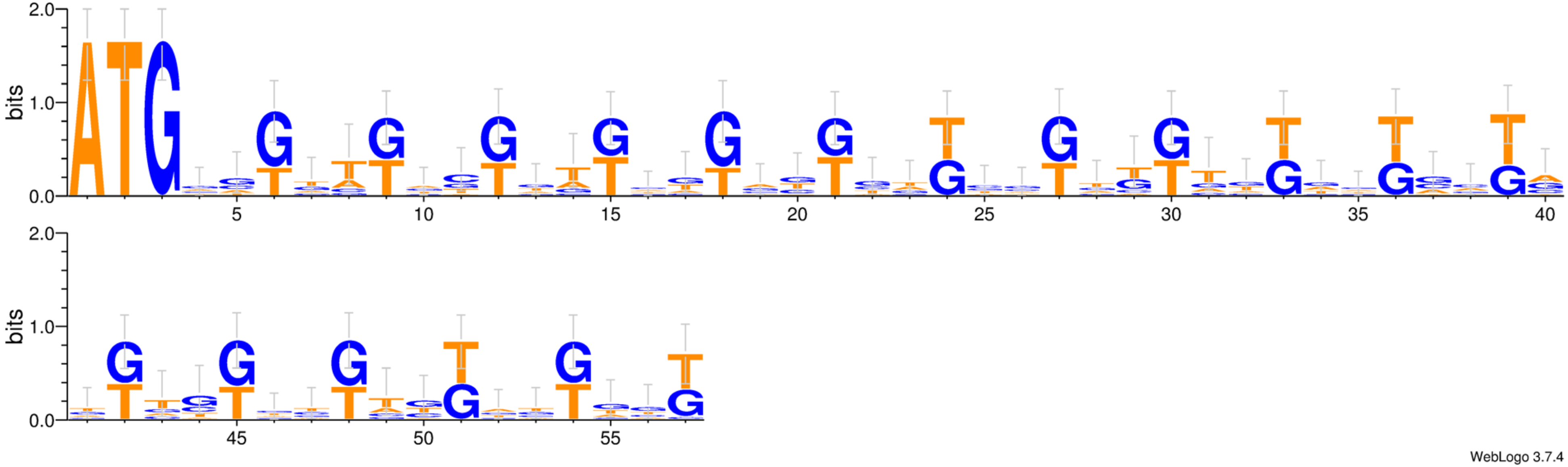


Figure 5.5. WebLogo (Crooks et al., 2004) was used to show the frequency of nucleotides at each position amongst the putative signal peptide sequences isolated from *E. coli*. The logo includes the 24 sequences which were expected 57 bp in length with no stop codons (therefore had the potential to act as viable N-terminal signal peptide sequences). The third position of each triplicate bases (i.e. a codon) could only encode 'G' or 'T', as shown in the logo. Nucleotide bases at other positions were diverse, demonstrating the success of the degenerate primer and HiFi cloning method to capture a multiplicity of putative signal peptide sequences for further study. Error bars indicate an approximate Bayesian 95% confidence interval.

Signal Peptide	Nucleotide sequence	G/C content (%)	Translated amino acid sequence	Hydrophobic residues (%)	SignalP secretion prediction
1.4	ATGTGTGCTTTTAGGGTGAGTCGTGTTGGTGTGCGTCATGATTTTCGTAGGATTGGT	45.6	MCAFRVSRVGVRHDFRRIG	42	No
1.11	ATGGGGTTGGTTCTGCTGTGTGGGTCTTTGGCTACTGAGGGGCGTAGGCGGATGGGG	61.4	MGLVLLCGSLATEGRRRMG	42	No
1.15	ATGGTGGAGGGTGTTTGGCTTGGGCTTATTTCGGTTGCGAAGGGGTCTAGTGATCTT	50.9	MVEGVWLGLISVAKGSSDL	53	No
1.17	ATGGCGGCTACGCGTCCGCGGCCGGGTTTGTTTAATATTAAGTCGGGTTATTATTCT	49.1	MAATRPRPGLFNIKSGYYS	37	No
1.19	ATGTCTCAGATGTATAATATGCATCGGCATTGTAGGGTGACTACGTTGTTGCTTGGG	43.9	MSQMYNMHRHCRVTTLLLG	37	No
2.1	ATGGGGATGCCTTTGCGGTGTATTGCGTGGTTTTCTGATGGGTGTGCGTCGAGGGGG	57.9	MGMPLRCIAWFSDGCASRG	47	No
2.2	ATGCAGTTGGCGGTTCGTCCGTGTACGTATTGTTATCATGTGTGGTTGTGGCGTATG	49.1	MQLAVRPCTYCYHVWLWRM	47	No
2.5	ATGCCGCAGAGGCTTAGTTCGGTTTGTTTGTTGTCGTTTTATCCGAATGCGGAGTCG	49.1	MPQRLSSVCLLSFYPNAES	47	No

2.6	ATGGGTTTTGCGTTTTCGTATCGGATTCGTCGGCATCGGCGTGGTACGACTGGGGTC	56.1	MGFAFSYRIRRHRRGTTGV	26	No
2.8	ATGTCGTTGTCGTTGAGGGGTTATTGGCCGAGGGGGAGTGATTGGTCGATGGGTCGT	56.1	MSLSLRGYWPRGSDWSMGR	37	No
2.11	ATGAAGCCTGGGGATCTGGCGAATGATGTGTTTAAGGCTACTAGTGCTTTTAGGGAT	45.6	MKPGDLANDVFKATSAFRD	47	No
2.12	ATGAGGAATGTTTTTATTATGGATCAGGGTGGTACTCAGCTGTTTCTTTCTTCGATT	36.8	MRNVFIMDQGGTQLFLSSI	47	No
2.15	ATGTCTTTTATGGTGCATCTTGAGCGGACTAGTGTGAGTGTGATTAGGTGTTGTTGG	43.9	MSFMVHLERTSVSVIRCCW	47	No
2.16	ATGGGGATTCCGAATTTGCTTTTTCCGTGGACGGTTGCGAGGTTGGTTAGTACTAGT	47.4	MGIPNLLFPWTVARLVSTS	63	*Yes – cleavage site predicted after 'A' in the protein sequence (+1) 20-21
2.17	ATGAGTATGTCTTGGCAGACTGCGAGGGTGTTGTTTAATAGGGGGCAGCATTCGTCT	49.1	MSMSWQTARVLFNRGQHSS	37	No
3.2	ATGAAGCTTCGGCGGGCTAGTCTGCATGTGCATAGGACTACGGAGGTGGCGCTGTGT	57.9	MKLRRASLHVHRTTEVALC	42	No
3.5	ATGCAGTTGGGTATGTGTGGGATTGAGTCGCTGGTGGGGAGTACTTCGTCGACTTAT	50.9	MQLGMCGIESLVGSTSSTY	32	No

3.6	ATGTGTTTGTTGGAGTGGGCTGTGGATATGGAGTGTCCTCGTTGTAGGTTTAGGGTT	47.4	MCLLEWAVDMECPRCRFRV	53	No
3.10	ATGCGTGTTAATACGGTGATGTCGTCTGGTGCGTGTTATTTTTCTTATCGGACGACG	45.6	MRVNTVMSSGACYFSYRTT	32	No
3.11	ATGGTGTATACGGAGTGGGGGCATATGCGGTTGCGTCTTATGCCGTTTCTGGTGGAT	52.6	MVYTEWGHMRLRLMPFLVD	58	No
3.15	ATGATGGAGCAGTTTGAGGTGGTTTGGTTTTGTGAGAGGGTGTGGCCGTTTCTGTCG	50.9	MMEQFEVVWFCERVWPFLS	63	No
3.16	ATGGGGGGGTGTGCGATGATGGTGCCGAGGTATGTGGCTGTTTGGGTGGGTGAGAGG	61.4	MGGCAMMVPRYVAVWVGER	58	No
3.17	ATGACGTGGACGAATGGGGCGTGGCTGTGTTGTGCGCCGCCTGAGGATAGTTTTCAG	57.9	MTWTNGAWLCCAPPEDSFQ	47	No
3.18	ATGCATGTTTCTGTTATTAATTATTGGATTAATAATGCTCTGCTGCTGTTTTTTGTT	21.8	MHVSVINYWINNALLLFFV	74	*Yes – cleavage site predicted after ‘A’ in protein sequence (+10) 29-30

Table 5.1. 50 independent *E. coli* transformants were selected for plasmid extraction and Sanger Sequencing to identify putative signal peptide sequences. Sequences were unique and 48% were expected 57 bp length with no stop codons - these sequences were analysed for G/C content, translated protein sequences were analysed for % hydrophobicity, and SignalP 3.0 (Bendtsen et al., 2004) with default eukaryote parameters was used to identify putative N-terminal cleavage sites (sequences pasted upstream of *P. sojae*_482953 missing its native signal peptide). Amino acid residues were coloured by their side chain properties: red (hydrophobic), blue (acidic), magenta (basic), green (hydroxyl, sulfhydryl, amine). Grey boxes indicate sequences 2.16 and 3.18, with predicted cleavage sites in the downstream protein sequence (+1 and +10 amino acids after the signal peptide sequence, respectively).

5.5.2 Pilot study of the N-terminal signal peptide library in *S. cerevisiae* (halo screening method)

HiFi reactions were transformed directly into *S. cerevisiae* BY4742 and transformants were selected on SCM-URA agar (+ 0.2% (w/v) xyloglucan). Transformation plates were stained with Congo red to reveal extracellular enzyme activity (indicated by a clearing or 'halo' around colonies of enzyme-secreting strains) – shown in Figure 5.6 (A).

Transformants were stocked and a total of 31 *S. cerevisiae* strains were cultured individually to confirm secreted enzymatic activity. OD_{600} –matched cells (0.5-1) were spotted onto SCM-URA agar containing 0.2% (w/v) xyloglucan (with 2% (w/v) glucose as an additional carbon source), and following incubation, plates were stained with Congo red to visualise halos (Table 5.2). The N-terminal signal peptide sequences encoded by the 31 transformants were identified by colony PCR (using GPD pro-F and 482953_NT_ClaI_Rv primers), followed by PCR purification and Sanger Sequencing (using a GPD pro-F primer) (Figure 5.6 (B and C)). Table 5.2 shows the nucleotide sequences derived from the 31 recombinant *S. cerevisiae* strains, as well as the translated amino acid sequences, and whether those yeast transformants produced a halo (i.e. if there was secretion of the xylo-glucanase by the N-terminal sequence encoded in that plasmid). For N-terminal signal peptide sequences that did not include a stop codon, SignalP 3.0 (Bendtsen et al., 2004) with default eukaryote parameters was used to confirm prediction of the signal peptide sequence when pasted upstream of *P. sojae*_482953 (missing its native signal peptide) (Table 5.2). Interestingly, sequences P2.C5 and P1.A4 (the same signal peptide sequence) were predicted to encode a cleavage site +1 amino acid after the signal peptide sequence; secreted enzyme activities were detected for these clones (Table 5.2).

5.5.2 Pilot study of the N-terminal signal peptide library in *S. cerevisiae* (halo screening method)

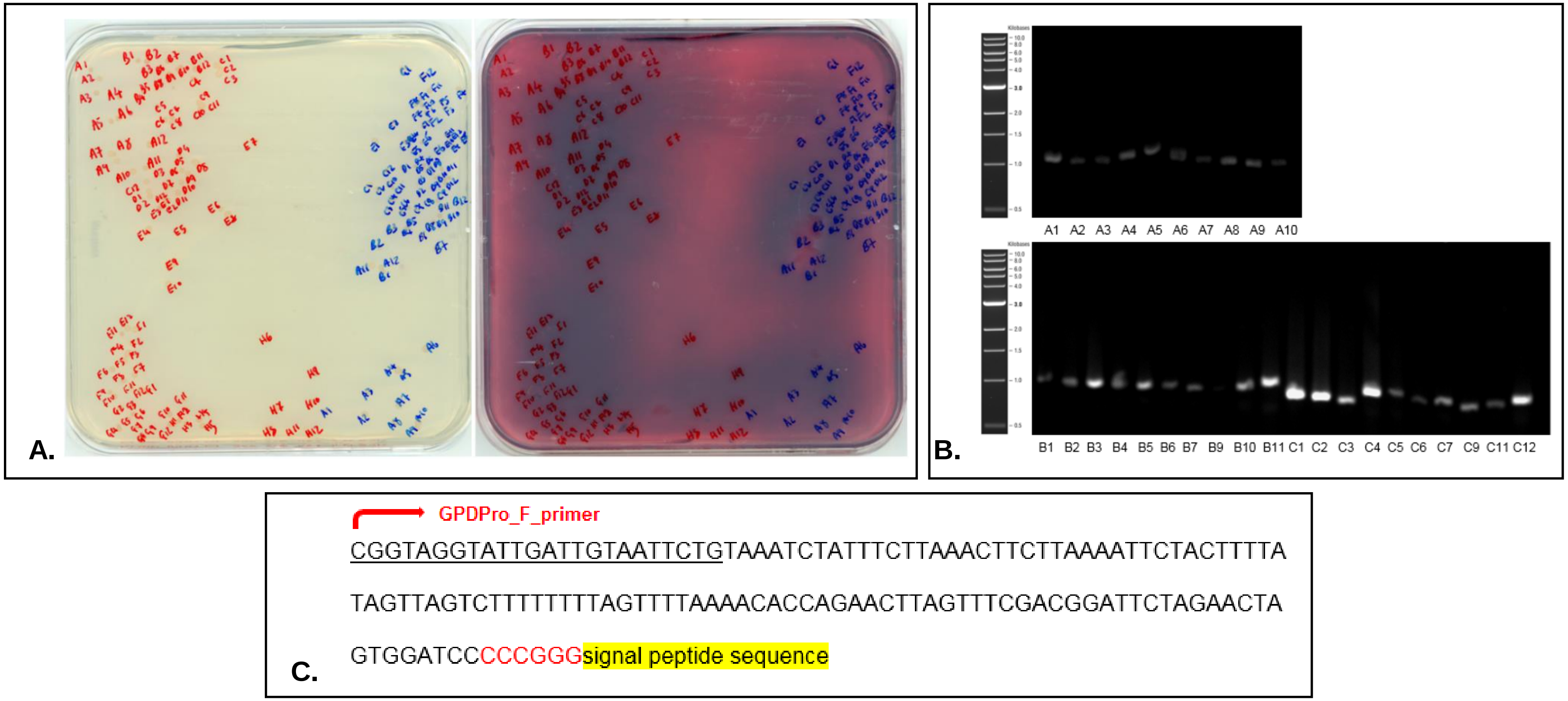
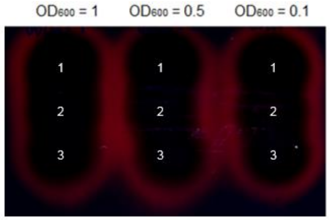
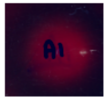
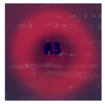
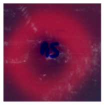
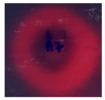
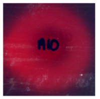
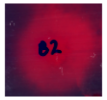
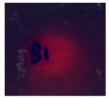
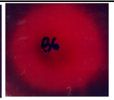
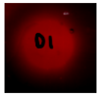
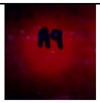
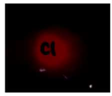
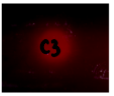
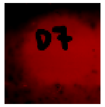
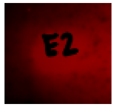

A.
B.
Kilobases
10.0
8.0
6.0
5.0
4.0
3.0
2.0
1.5
1.0
0.5
A1 A2 A3 A4 A5 A6 A7 A8 A9 A10
B1 B2 B3 B4 B5 B6 B7 B9 B10 B11 C1 C2 C3 C4 C5 C6 C7 C9 C11 C12
C.
GPDPro_F_primer
CGGTAGGTATTGATTGTAATTCTGTAAATCTATTTCTTAAACTTCTTAAAATTCTACTTTTA
TAGTTAGTCTTTTTTTTAGTTTTAAAACACCAGAACTTAGTTTCGACGGATTCTAGAACTA
GTGGATCCCCCGGGsignal peptide sequence

Figure 5.6. A. HiFi reactions were transformed into *S. cerevisiae* and selected on SCM-URA (0.2% (w/v) xyloglucan); mutants were stocked and plates were stained with Congo red to reveal extracellular enzyme activity. 31 independent *S. cerevisiae* transformants were selected for Sanger Sequencing to identify putative signal peptide sequences. **B.** The approach to identify the N-terminal signal peptides involved PCR (GPD_pro-F and 482953_NT_ClaI_Rv), followed by **(C)** Sanger Sequencing of purified PCR products using GPD_pro-F primer to retrieve N-terminal sequences.

Signal Peptide	Nucleotide sequence	G/C content (%)	Translated amino acid sequence	Hydrophobic residues (%)	SignalP secretion prediction	Halo	
MFA +ve control	ATGAGATTTCCTTCAATTTTTACTGCTGTTT TATTCGCAGCATCCTCCGCATTAGCTGCTC CAGTCAACACTACAACAGAAGATGAAACG GCACAAATTCCGGCTGAAGCTGTCATCGG TTACTCAGATTTAGAAGGGGATTTCGATGT TGCTGTTTTGCCATTTTCCAACAGCACAAA TAACGGGTTATTGTTTATAAATACTACTATT GCCAGCATTGCTGCTAAAGAAGAAGGGGT ATCTCTCGAGAAAAGAGAGGCTGAAGCT	42	MRFPSIFTAVLFAASSALAAPV NTTTEDETAQIPAEAVIGYSDL EGDFDVAVLPFSNSTNNGLLFI NTTIASIAAKEEGVSLEKREAE A	49	Yes		
P2.A1 P2.A8 P2.B11	ATGACGCGTATGACTTAGGATCCGGTTCTT GTGGGGTATATGTTGTGTCTTGGTTAT	43	MTRMTStopDPVLVGYMLCLGY	47	No		P2.A1 – NO P2.A8 – NO
P2.A2 P2.A4 P2.B4	ATGGGTATGCGTGTGGTTATTGTGGGTGAT AAGCGTGCTGAGTGTGGTGTTGGGTTG	50	MGMRVVIVGDKRAECGVGL	47	No		P2.A4 - YES

P2.A3 P2.A5 P2.A7 P2.A10 P2.B2 P2.B5 P2.B7	ATGGTGCCTGCGGCGTTTTTTTTTCAGGGG TTTAAGTATGCGGTTCGTCTGCAGAGT	48	MVPAAFFFQGFKYAVRLQS	63	No	    	P2.A3 - YES P2.A5 - YES P2.A7 - YES P2.A10 - YES P2.B2 - YES
P2.A6 P2.B1 P2.B6 P2.C12 P2.D1	ATGTATCAGGGGTCTCGGAGTCGGTAGGC GTCTGGTAAGATGCGGTTTCGGGGTAAT	54	MYQGSRSRStopASGKMRFRG N	21	No	  	P2.B1 - NO P2.B6 - YES P2.D1 – NO
P2.A9 P2.C1 P2.C3 P2.D7 P2.E2	CTGCAGGAATTCGATATC	45	LQEFDI	33	No	    	P2.A9 - NO P2.C1 - NO P2.C3 – NO P2.D7 – NO P2.E2 – NO

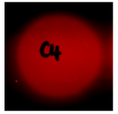
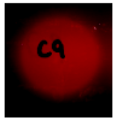
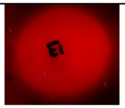
P2.C2 P2.C7 P1.A5	ATGGTGTGTTTTATGATTAATTCTGTTCATC ATCGTGGGCTTTATTAGAGGCGGTTG	38	MVCFMINSVHHRGLYStopRRL	42	No		P2.C2 – NO
P2.C4	ATGTCGGAGTTTCATGAGTGGGGGGGTGC	59	MSEFHEWGG	33	No		P2.C4 - YES
P2.C5 P1.A4	ATGTGGTTTCGGGCGAGGGTGTATACGTG GTCGTGGTTTGCGTGTAGTGAGATTTTG	51	MWFRARVYTWSWFACSEIL	58	*Yes – cleavage site predicted after 'A' in protein sequence (+1) 20-21		P2.C5 – YES
P2.C9	ATGGATGCGCAGCTTCGGGTTATGTATGTT CTTATT	42	MDAQLRVMYVLI	58	No		P2.C9 - YES
P2.E1	ATGCTGTTTAGGGCTAGTGAGTGTATTAGG GTTGCTGGGCGTGAGCCTGAGACGCGT	55	MLFRASECIRVAGREPETR	37	No		P2.E1 - YES

Table 5.2. 31 independent S. cerevisiae transformants were selected for PCR and Sanger Sequencing to identify putative signal peptide sequences, alongside individual culture on SCM-URA (+0.2% (w/v) xyloglucan) with Congo red staining, to confirm extracellular enzyme activity. Nucleotide sequences highlighted in grey indicate those that were less than the expected 57bp length. Sequences P2.A9, P2.C1, P2.C3, P2.D7 and P2.E2 (the same sequence) encoded a signal sequence of 6 amino acids missing a start codon – consistently, none of these clones showed extracellular enzyme activity. Sequences P2.A1, P2.A8, P2.B11, P2.A6, P2.B1, P2.B6, P2.C12, P2.D1 encoded a stop codon followed by a subsequent start codon (sequences underlined); sequences P2.C2, P2.C7, P1.A5 encoded a stop codon and consistently, clones did not produce a halo after staining (suggesting the protein was not translated). Sequences P2.C4 and P2.C9 encoded putative signal peptide sequences of 9 and 12 amino acids, respectively – interestingly, both transformants demonstrated secreted enzyme activity. SignalP 3.0 (Bendtsen et al., 2004) with default eukaryote parameters was used to identify putative N-terminal cleavage sites (sequences pasted upstream of *P. sojae*_482953 missing its native signal peptide). Sequences P2.C5 and P1.A4 (the same nucleotide sequence) predicted a cleavage site +1 amino acid after the signal peptide sequence. Amino acid residues were coloured by their side chain properties: red (hydrophobic), blue (acidic), magenta (basic), (hydroxyl, sulfhydryl, amine). 'MFA +ve control' refers to the *S. cerevisiae* MFA signal peptide sequence that successfully directs enzyme secretion (see previous chapter).

5.5.3 Microfluidics approach to screen the N-terminal signal peptide library in *S. cerevisiae*

Carbohydrate degradation can be inferred by a shift in Congo red absorbance – previously demonstrated by a microplate assay to quantify cellulase activities (Haft et al., 2012). The method was adapted in order to develop a high-throughput microdroplet technique for simultaneously detecting xylo-glucanase activities in large libraries of N-terminal signal peptide mutants.

The starting concentration of xyloglucan and the concentration of Congo red for detection were identified by a microplate assay: SCM-URA media containing 1%, 0.5%, 0.25% and 0% (w/v) xyloglucan were mixed with 0.4%, 0.3%, 0.2% and 0.1% (w/v) Congo red (+ 1 M NaCl for colour development), and the absorbance measured at 550 nm (Figure 5.7 (A)). At 0.3% and 0.4% Congo red, the upper limit for accurate absorbance detection was reached by the plate reader and it was therefore impossible to distinguish between samples with and without xyloglucan. The optimal Congo red concentration (where a maximum difference in absorbance was achieved between 1% xyloglucan and 0% xyloglucan (i.e. SCM-URA only – representing complete xyloglucan breakdown)) was found to be 0.1% (w/v) Congo red (Figure 5.7 (B)).

In order to begin to develop a high-throughput micro-droplet assay for rapid detection of secreted xylo-glucanase activities by recombinant *S. cerevisiae* expressing a number of randomised N-terminal signal peptides, it was important to establish that ‘positive’ droplets (i.e. those containing *S. cerevisiae* single cells secreting an active cognate enzyme) could be distinguished from ‘negative’ droplets. Two samples were prepared - (i) *P. sojae*_482953 expressed in plasmid p426-GPD downstream of the *S. cerevisiae* MFα N-terminal signal peptide sequence (positive control), and (ii) a p426-GPD vector-only (negative control). Both strains were cultured overnight, the cells were washed and resuspended in SCM-URA (+ 1% (w/v) xyloglucan) at an OD_{600} of (0.1). Cells were encapsulated within water-in-oil droplets using a microfluidic chip designed and fabricated in the group of Dr Gielen (University of Exeter) (Figure 5.8 (A, B)). Droplets were collected in 0.2 mL tubes and incubated at 30°C for 24 hours. Following incubation, a small volume of droplets were extracted to check for cell growth using a light microscope – several cells per droplet were visualised, suggesting encapsulation did not affect *S. cerevisiae* growth and division (Figure 5.8 (C)). Both positive and negative *S. cerevisiae* droplets were separately reinjected into the microfluidic chip. Equal volumes of 0.1% (w/v) Congo red and 1 M NaCl were premixed immediately before pico-injection into each droplet (Figure 5.8 (D)) - following injection, the voltage signal for each droplet was recorded using a custom Labview program. Raw data (time and voltage) (Figure 5.9 (A)) was processed using a custom MatLab script (to exclude droplet edges and correct for the background of oil); probability distribution for positive and negative samples are shown in (Figure 5.9 (B)).

5.5.3 Microfluidics approach to screen the N-terminal signal peptide library in *S. cerevisiae*

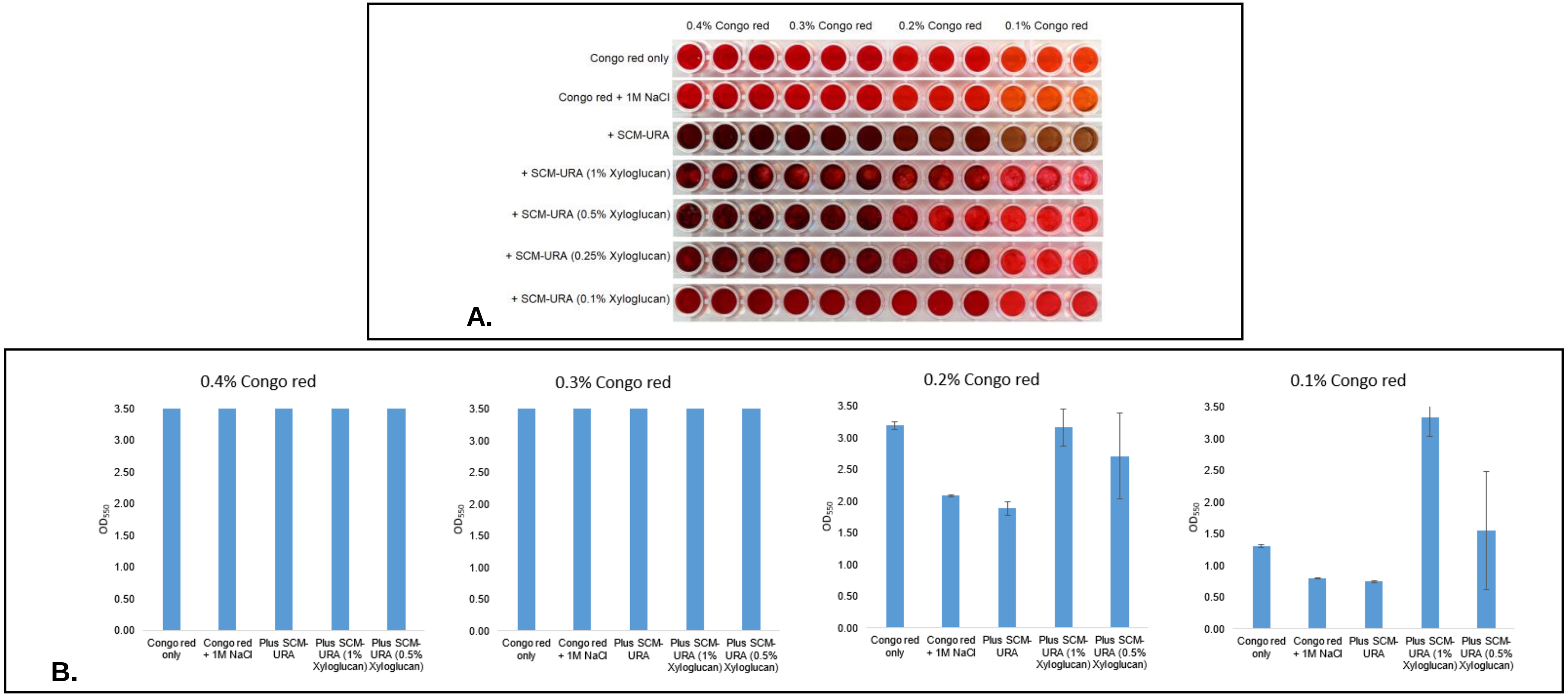


Figure 5.7. The starting concentration of xyloglucan to encapsulate with yeast in the droplets, and the concentration of Congo red for detection were identified by a microplate assay: **A.** SCM-URA media containing 1%, 0.5%, 0.25% and 0% (w/v) xyloglucan were mixed with 0.4%, 0.3%, 0.2% and 0.1% (w/v) Congo red (+ 1 M NaCl for colour development), and the absorbance measured at 550 nm. **B.** At 0.3% and 0.4% Congo red, the upper limit for accurate absorbance detection was reached by the plate reader; the optimal Congo red concentration (where a maximum difference in absorbance was achieved between 1% xyloglucan and 0% xyloglucan (i.e. SCM-URA only – representing complete xyloglucan breakdown)) was found to be 0.1% (w/v) Congo red. N=3, +/- SD.

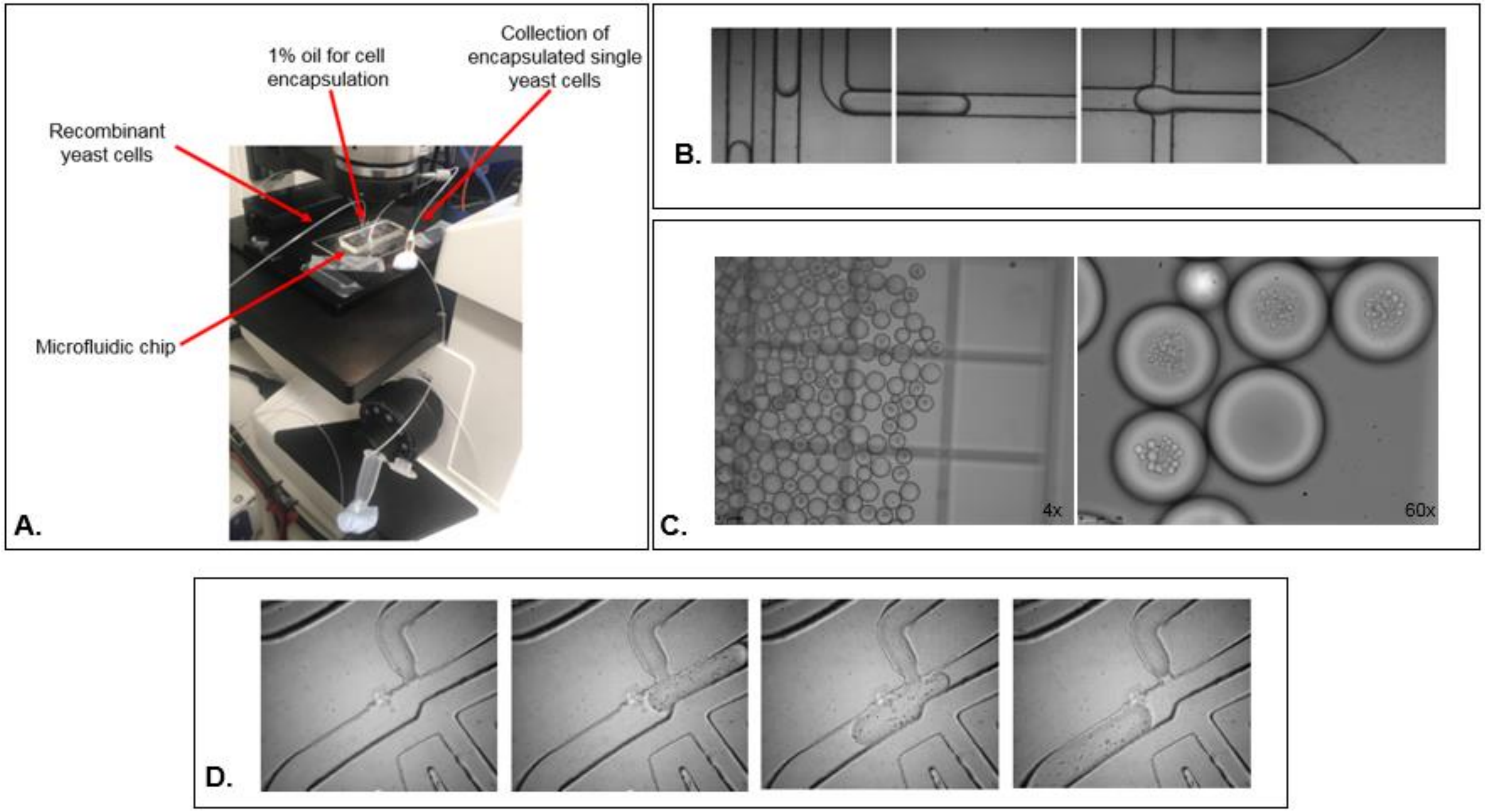
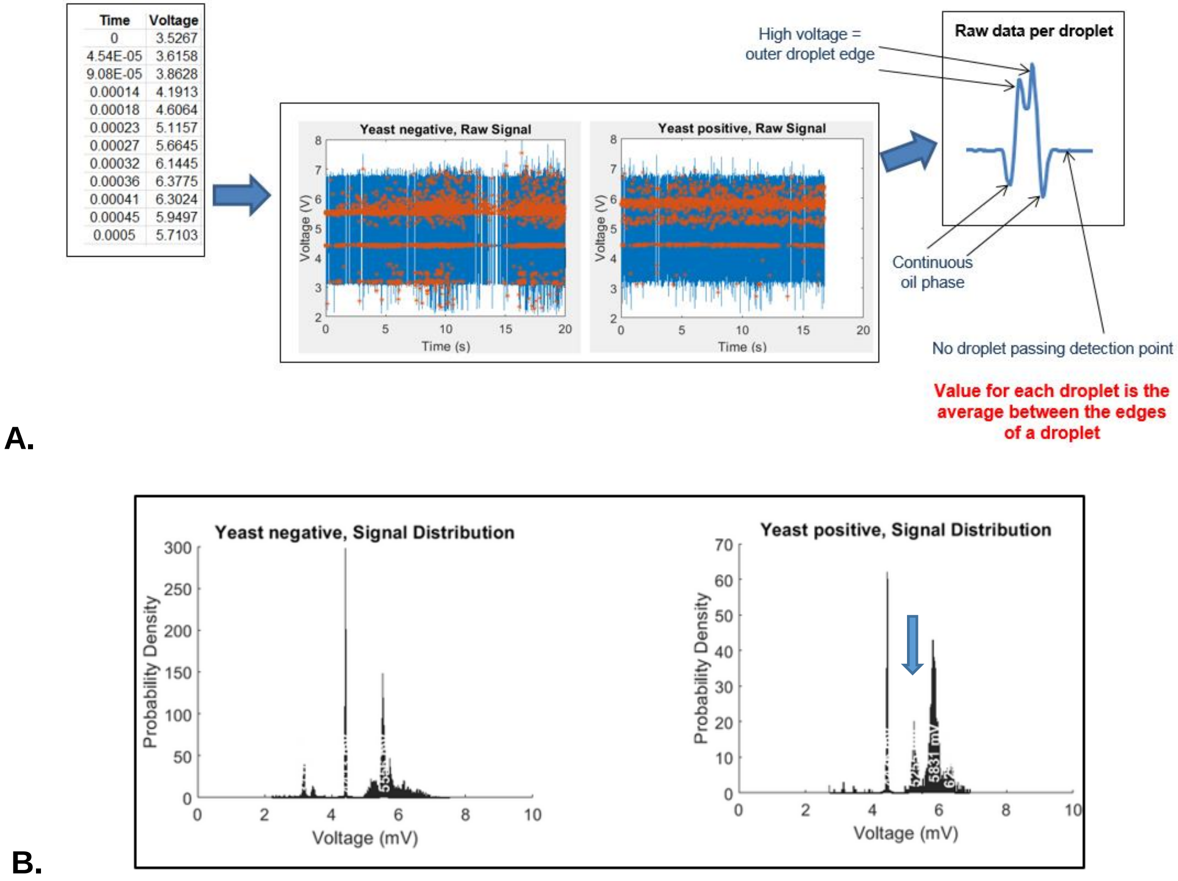

Collection of encapsulated single yeast cells
1% oil for cell encapsulation
Recombinant yeast cells
Microfluidic chip
A.
B.
C.
4x
60x
D.

Figure 5.8. A. Micro-droplet set up; yeast cells were encapsulated within water-in-oil droplets, and collected in 0.2 mL tubes. **B.** Example of droplet formation, **C.** encapsulated cells were visualised after incubation for 24 hours at 30°C - several cells per droplet were visualised, suggesting encapsulation did not affect *S. cerevisiae* growth and division. **D.** Following incubation, droplets were reinjected into the microfluidic chip; 0.1% Congo red (+ 1 M NaCl) were premixed before pico-injection (upper channel in the figure) into each droplet (passing along the horizontal channel in the figure).



Time	Voltage
0	3.5267
4.54E-05	3.6158
9.08E-05	3.8628
0.00014	4.1913
0.00018	4.6064
0.00023	5.1157
0.00027	5.6645
0.00032	6.1445
0.00036	6.3775
0.00041	6.3024
0.00045	5.9497
0.0005	5.7103

Figure 5.9. A. Preliminary micro-droplet data for positive and negative S. cerevisiae samples. Time and voltage of droplets were recorded in series as they passed through the detector; the raw data for each droplet is illustrated in the image on the right - the value for each droplet is the average between the edges of the droplet. **B.** Raw data for a ‘yeast positive’ (i.e. *P. sojae*_482953 expressed downstream of an MFα signal peptide), and a ‘yeast negative’ (i.e. *S. cerevisiae* expressing a p426-GPD vector only) were processed using a custom MatLab script (probability distributions shown in the figure). Compared to the negative control, a small peak (population of droplets) was observed at ~4.5 mV in the positive sample (blue arrow) – putatively indicating droplets in which there has been xyloglucan degradation.

5.6 Discussion

Previously reported HGT events from fungi to oomycetes include a suite of secreted enzymes predicted to degrade plant cell wall-specific substrates (Torto et al., 2002; Belbahri et al., 2008; Richards et al., 2011; Misner et al., 2015; Savory et al., 2015). Whilst mechanistic costs to HGT have been described previously (e.g. differences in codon usage (Sorensen et al., 1989; Buchan., 2006; Hershberg and Petrov., 2008; Qian et al., 2012; Callens et al., 2020)), what influences the transferability of genes encoding secreted proteins is less well understood. Secreted proteins are targeted to the secretory pathway through the recognition of a short N-terminal amino acid sequence by translocation machinery (Blobel and Sabatini., 1971; Blobel and Dobberstein., 1975; von Heijne., 1983; 1985; 1990) by an evolutionarily-conserved mechanism; such sequences therefore play a fundamental role in controlling the rate of protein secretion, as well as determining downstream protein folding and modifications. Yet, there are no consensus N-terminal sequences, and they are largely heterogeneous (even amongst paralogs of the same gene family (e.g. see Chapter 3, Figure 3.4 *P. sojae* GH12 proteins).

The aim of the present chapter was to develop methods that will be used to investigate the diversity of signal peptide-encoding sequences, and therefore understand how such peptides can evolve, and address the hypothesis that N-terminal signal peptide sequences represent a porous barrier to the acquisition of foreign genes by HGT – i.e. if a donor signal peptide is not recognised by translocation machinery in the host, it could be selected against in favour of genes that more closely match the host cellular system requirements. The putative degeneracy of protein targeting sequences has been described previously (e.g.

by Cory and Dunn., 2019), and such liberal recognition of N-terminal signal peptides could potentially buffer against the initial incompatibilities with translocation machinery (with subsequent gene duplication events fostering beneficial mutations in favour of increasing the efficiency of secretion). The methods described in this chapter will ultimately be used for a large mutational perturbation study, which will lead to the generation of a wide population of potential mutant signal peptides for the synthetic 'acquisition' of a secreted xylo-glucanase by *S. cerevisiae*. Due to the heterogeneity of N-terminal sequences, it is hypothesised that the resulting sequence space would have many peaks and valleys (Romero and Arnold., 2009; Carneiro and Hartl., 2010), representing the many possible sequences that still retain functionality (i.e. secretion), although putatively at varying efficiencies. The sequence landscape will be fundamental for understanding how a protein acquired through HGT could theoretically traverse mutational sequence space in order to evolve, improve, or lose a functional signal peptide.

5.6.1 Pilot study of the N-terminal signal peptide library in *E. coli*

The experimental approach to generate exhaustive libraries of randomised signal peptide sequences used a degenerate oligonucleotide (ATG followed by NNK^{18}), cloned upstream of a xylo-glucanase missing its native N-terminal signal peptide sequence (expressed in high-copy vector, p426-GPD). This combination allowed for all 20 amino acids to be coded, but reduced the probability of stop codons to TAG only (therefore, a stop codon has a 3% probability of being encoded by each set of randomised triplicate nucleotides) (e.g. Reetz et al., 2008). It is important to note that some amino acids are coded for by more than one set of codons (resulting in bias caused by uneven degeneracy) – for example, assuming a NNK library, tryptophan is encoded by 1 codon, whereas leucine is encoded by 3 - putatively reducing diversity and the amount of sequence space sampled (Makowski and Soares., 2003; Krumpe et al., 2007). Nevertheless, large randomised libraries can be generated by sufficient replication of the cloning method, in order to expand the genomic diversity captured.

NEBuilder HiFi DNA cloning enabled synthesis of the library of variants in a single reaction - this allowed for many randomised N-terminal sequences to be assembled. Transformation of the HiFi cloning reactions into *E. coli* enabled initial analysis of the efficiency of the cloning method; 50 individual *E. coli* transformants across three independent HiFi reactions were selected for plasmid isolation and Sanger Sequencing; the data indicated that all 50 N-terminal sequences were unique, and that 48% (24 of the sequences) were the expected length (57 bp) and contained no stop codons (i.e. these were putatively viable 19 amino acid N-terminal signal peptides). GC content within these 24 sequences was in the range 21.8-61.4% (Table 5.1, and see Figure 5.5 for WebLogo of the frequency of nucleotides observed at each position) (Crooks et al., 2004), and percentage hydrophobicity was in the range of 26-74% (Table 5.1). Hydrophobicity is a useful parameter to include because the hydrophobic core amino acids play an important role in determining the efficiency of secretion across of the tree of life (e.g. in bacteria (Gennity et al., 1990; Keenan et al., 1998), mammals (Bird et al., 1990; Nilsson et al., 2015) and yeast (Rothe and Lehle., 1998)). Amino acid residues were coloured by their side chain properties, confirming that a diverse set of library variants could be generated by this cloning method (Table 5.1).

Of the 24 unique sequences, 22 did not result in prediction of an N-terminal signal peptide when pasted upstream of the *P. sojae*_482953 protein sequence (missing its native signal peptide) (Table 5.1) – although, it is important to be cautious in the absence of direct experimental evidence for secretion, because characterised sequences in protein training datasets may be distantly related. Interestingly, two of the sequences did elicit the prediction of putative cleavage sites by SignalP 3.0 (using default eukaryote parameters; Bendtsen et al., 2004) – suggesting the recognition of appropriately composed N-terminal signal peptides (Table 5.1). Notably, the sequences share two of the highest hydrophobicity scores (63% and 74%), which as mentioned previously, is an important determinant for recognition by translocation machinery. However, the predicted signal peptide cleavage sites for both sequences was not following the final amino acid of the signal peptide sequence - in sample 2.16, a cleavage site was predicted +1 amino acid following the signal peptide sequence, and in sample 3.18, a cleavage site was predicted +10 amino acids following the signal peptide sequence (Table 5.1), suggesting putative N-terminal signal peptides of 20 and 29 amino acids, respectively. It would be interesting to further confirm signal peptide recognition and cleavage at these sites, and (particularly for sample 3.18) – the impact of the putative cleavage site in the downstream protein sequence to the folding and function of the mature protein.

5.6.2 Pilot study of the N-terminal signal peptide library in *S. cerevisiae* (halo screening method)

After establishing an appropriate method to generate large libraries of randomised N-terminal signal peptide sequences, the method was used to transform the HiFi cloning reactions directly into heterologous host, *S. cerevisiae*

- enabling direct selection of transformants expressing one of a number of N-terminal sequences. Colonies were stocked and transformation plates were subjected to Congo red staining to identify recombinant *S. cerevisiae* with secreted enzyme activity as a proxy for signal peptide function (Figure 5.6 (A)). It was difficult to confidently distinguish between positive and negative yeast colonies by direct staining of the transformation plates - this was due to (i) an unknown reaction between Congo red and the transformation mixture background on the agar plates resulting in 'hazy' patches of staining, (ii) the size of the individual transformants (i.e. the diameter of each colony) meant that halo diffusion was not obvious in some cases, and (iii) some colonies were too close together and difficult to distinguish. As stocks of the *S. cerevisiae* mutants were maintained following the transformations, 31 strains were individually cultured, spotted on fresh media containing xyloglucan, and the plates were stained as before. This allowed clear identification of the *S. cerevisiae* mutants expressing either a 'positive' or a 'negative' N-terminal signal peptide sequence (i.e. by detection of the secreted enzyme activity). N-terminal sequences were identified using a PCR and Sanger Sequencing approach; interestingly (and unlike the previous result in *E. coli*), many recombinant *S. cerevisiae* clones expressed the same signal peptide sequences (Table 5.2). Whilst the redundancy suggests that a reduced diversity of sequences was captured and implies that much larger libraries would need to be screened to capture more complete diversity, it is also particularly useful for replication if multiple clones harbour the same N-terminal signal peptide. Many *S. cerevisiae* transformants expressing the same N-terminal signal peptide sequence gave the same result in the halo screen (Table 5.2), improving the reliability of the screening method.

Wangen and Green (2020) recently demonstrated genome-wide stop codon readthrough (SCR) in mammalian cells treated with aminoglycosides for both premature stop codons and normal stop codons, suggesting that the presence of a termination codon within a gene sequence is not necessarily indicative of a halt in translation (Wangen and Green., 2020). Interestingly, in this study there were examples of N-terminal signal peptide sequences that contained one or more premature stop codons (i.e. found within the 19 amino acid signal sequence preceding the gene of interest) (Table 5.2). Sequences P2.C2, P2.C7 and P1.A5 all contained a single stop codon, and did not appear to secrete an active enzyme, suggesting that the protein was not translated (as expected). Clones P2.A1, P2.A8, P2.B11 and P2.A6, P2.B1, P2.B6, P2.C12 contained a stop codon followed by another start codon within the signal peptide sequence, so for these clones it is possible that the second methionine signalled the beginning of the nascent polypeptide chain. As the N-terminal signal peptide sequence preceding the protein would then be only 6-7 amino acids in length, no protein secretion would be expected (based on previous studies concerning the length and basic structure of signal peptide sequences (e.g. von Heijne., 1984, 1990)). Interestingly, all except one of these *S. cerevisiae* clones did not demonstrate secreted enzyme activity as expected – however, P2.B6 (sequence consisting of a stop codon, followed by a terminal ‘M-R-F-R-G-N’) did appear to secrete an active enzyme as suggested by a halo (Table 5.2). SCR could explain this result – although it also implies that the 6 amino acid ‘signal peptide’ sequence is also sufficient to drive translocation of this peptide (or otherwise due to background protein secretion resulting from the expression vector – discussed below).

It was not possible to retrieve an N-terminal signal peptide of 57 bp in length, with no stop codons, which did not result in extracellular enzyme activity (i.e. most sequences with a start codon appeared to act as functional signal peptides), suggesting that N-terminal sequences may be largely degenerate. Whilst a negative control (*P. sojae*_482953 expressed in p426-GPD without a signal peptide) was included in the study, it is possible that the choice of a high-copy plasmid and constitutive promoter for gene expression also resulted in background secretion of the protein in *S. cerevisiae* (i.e. due to excess protein accumulation inside the cell, which may then be removed by a mechanism independent to the classical secretory pathway involving signal peptide recognition). Park and Zhang (2012) hypothesise that high expression of HGT genes presents a fitness cost due to increased energy expenditure, protein misfolding and reductions in translation efficiency (Park and Zhang., 2012), but it is possible that gene expression level could play a role in the HGT of secreted proteins with 'incompatible' signal peptide sequences – for example, high gene expression and non-efficient recognition of proteins by translocation machinery could allow use of the extracellular space as a putative 'dumping ground' for novel proteins. Multiple organelle targeting of novel proteins has been previously suggested to buffer against the putative disruption of recently introduced sequences (e.g. Llorente et al., 2016), and has been demonstrated in the removal of (for example) putatively toxic proteins from the cytosol to mitochondria in yeast and human cells (Ruan et al., 2017). Subsequent gene duplications and mutation then provide the opportunity to evolve a signal peptide sequence that more closely matches the host's translocation requirements. However, it has also been shown to be deleterious for pre-proteins to accumulate in the cell without signal peptide cleavage (e.g. Dalbey and Wickner., 1985; Nesmeyanova et al., 1991;

Nesmeyanova et al., 1997), so it would be interesting to further explore gene expression level in the context of secreted protein acquisition by HGT events. It is also interesting to consider paralog evolution for secreted proteins and whether these events could provide a means to optimise signal peptide sequences (i.e. increase or decrease the efficiency of secretion through gene duplication and mutation); unfortunately it was beyond the scope of this work, but it would be informative to compare the native secretion efficiency of GH12 and GH10 paralogs (for example, by *in vivo* fluorescent tagging of the proteins in *P. sojae*, and monitoring the secreted protein concentration by Western blotting or mass spectrometry).

5.6.3 Microfluidics approach to screen the N-terminal signal peptide library in *S. cerevisiae*

As a large number of *S. cerevisiae* transformants would need to be screened in order to capture the full diversity of N-terminal signal peptide sequences for analysis, a high-throughput microdroplet technique was sought for the detection of secreted enzyme activities, based on Congo red absorbance shift (Haft et al., 2012). Two samples ((i) *P. sojae*_482953 expressed in plasmid p426-GPD downstream of the *S. cerevisiae* MFα N-terminal signal peptide sequence (positive control), and (ii) a p426-GPD vector-only (negative control)) were initially trialled by this method – single cells were encapsulated within water-in-oil droplets in SCM-URA (+ 1% (w/v) xyloglucan), and the droplets were incubated overnight at 30°C. Following incubation, droplets were visualised under a light microscope to confirm cell viability (Figure 5.8C)). Empty droplets were also visualised (i.e. droplets that did not harbour a *S. cerevisiae* cell during generation); as previously shown, cell encapsulation into droplets generally follows the Poisson distribution, therefore an appropriate cell dilution is needed to achieve mostly single cells in droplets - for this study, a starting cell density of OD_{600} 0.1 (0.34 X 10^{7} cells per mL) (Beneyton et al., 2017)) was used. Following pico-injection of Congo red (+ NaCl) (premixed immediately before injection because longer exposure of the Congo red to NaCl resulted in precipitation within the microfluidic channels), the voltage signal for each droplet was recorded and raw data (time and voltage) was processed using a custom MatLab script, visualised as probability distributions (Figure 5.9). The preliminary data indicates that the positive control sample (i.e. *S. cerevisiae* secreting an active xylo-glucanase) contained a population of droplets after 24 hours (small peak at ~4.5 mV), which putatively represents the droplets in which there is a shift in Congo red absorbance (i.e. putative xyloglucan degradation). This is a promising result, although further work will be necessary to confirm accurate distinction between known populations of droplets (i.e. generating spectra for droplets containing varying concentrations of xyloglucan to optimise detection of subtle degradation). Unfortunately (due to time constraints), it was not possible to continue optimisation of this method, however further work should also involve mixing the positive and negative *S. cerevisiae* cultures before encapsulation with the enzyme substrate – this will allow optimisation of the next stage, the droplet sorting (i.e. the ability to detect and collect positive droplets based on xyloglucan product absorbance). After sorting, plasmid DNA can be recovered from the droplets to confirm the droplet identities (also enabling the enrichment ratio to be calculated for positive and negative variants).

5.7 General Conclusion

This chapter aimed to develop a method that will ultimately be used to investigate signal peptide evolution, specifically exploring if N-terminal signal peptide sequences represent a putative barrier to cross-phylum HGT of secreted proteins - or if such sequences are largely degenerate (indicating lowered requirements for protein targeting in eukaryotes that putatively allows increased mobility across theoretical sequence space). With sufficient replication, large libraries of randomised signal peptide sequences can be generated using a degenerate oligonucleotide approach (ATG-NNK^{18}), with sequences cloned upstream of a secreted xylo-glucanase missing its native signal peptide, and expressed in *S. cerevisiae*. Secreted enzyme activity can then be used a proxy for signal peptide function – Congo red staining and halo screening of agar plates is a convenient approach to identify positive and negative variants, and this chapter has demonstrated how to link a variants' phenotype to its genotype using this method (by subsequent PCR and Sanger Sequencing). Although laborious, this method can be used to screen high numbers of variants. However, to reduce the time demands, screening of enzyme mutants compartmentalised in monodisperse droplets in microfluidic devices was also explored based on the shift in Congo red absorbance when no longer bound to soluble substrate. Initial data suggests it is possible to distinguish between positive and negative secreted enzyme activities within water-in-oil droplets, and with further optimisation, this could be an invaluable tool for high-throughput data generation for the N-terminal signal peptide sequence landscape.

Chapter 6

General Discussion and Conclusions

The oomycetes are a group of heterotrophic protists that form part of the stramenopiles (heterokonts) lineage, within the SAR supergroup (Riisberg et al., 2009). They resemble fungi in their filamentous growth and osmotrophic feeding habits, which resulted in their initial placement within the same kingdom (Ainsworth., 1961) - however, it is now known that the oomycetes and fungi have very different evolutionary histories (Förster et al., 1990; Leclerc et al., 2000; Hudspeth et al., 2000; Hudspeth et al., 2003; Thines et al., 2007; McCarthy and Fitzpatrick., 2017), and so some of their similarities may be due to convergent evolution (Latijnhouwers et al., 2003; Money et al., 2004). Whilst many species have not yet been sampled, we know the oomycetes are diverse and include ecologically-destructive parasites of plants, fungi and animals, as well as saprotrophic species (Beakes et al., 2012).

Plant pathogenicity is thought to have evolved independently in multiple oomycete lineages (Thines and Kamoun., 2010). Some of the best-known species of phytopathogenic oomycetes include those of the hemibiotrophic genus, *Phytophthora* - late blight of potato (caused by *P. infestans*), sudden oak death (caused by *P. ramorum*), and root rot of soybean (caused by *P. sojae*) are some of the ecologically-important diseases triggered by oomycetes that occupy a hemibiotrophic lifestyle, involving an initial biotrophic association with a plant host that causes little damage, before switching to a necrotrophic phase (involving maceration of host tissues as the infection spreads). Through comparative phylogenetic analyses, it has previously been proposed that multiple HGTs from fungi have contributed to the evolution of phytopathogenicity in the oomycetes, and interestingly, many of the identified genes encode secreted proteins that display a strong functional trend towards plant cell wall degradation (Torto et al., 2002; Richards et al., 2006; Belbahri et al., 2008; Richards et al., 2011; Misner et al., 2015; Savory et al., 2015). Oomycete-plant interactions are complex, involving the secretion of diverse proteins involved in feeding and virulence - of particular interest to this work are the digestive enzymes secreted to break down plant cell wall-specific carbohydrates - providing a means of entry into plant tissues, as well as an abundant source of fixed carbon (Latijnhouwers et al., 2003; Oliver and Ipcho., 2004; Kabbage et al., 2015). Most of the HGTs have undergone subsequent gene duplications (e.g. Richards et al., 2011., Savory et al., 2015) – an important mechanism that can influence both the expression of the acquired DNA (i.e. by putatively increasing the transcriptional dosage or fidelity), and the resulting gene function. As a result, paralogs evolved from ancestral HGT events could encode unexplored functional differences important for plant cell wall digestion (Ohno., 1970; Stoltzfus., 1999; Force et al., 1999; Long et al., 2003).

The aim of this thesis was to explore the functional significance of protein paralogs (which form part of multi-gene enzymes families) acquired by HGT into the oomycetes. It was hypothesised that whilst HGT may have facilitated a novel gain of function, the subsequent expansion (by iterative gene duplication events) resulted in a further enhancement of function (i.e. by putatively widening the encoded protein activities). Gene duplication can be a source of phenotypic diversity, therefore it was hypothesised that some duplicates could be functional, some could be non-functional, and some may have putatively gained additional or novel functions over the course of their evolution (e.g. neofunctionalization (Ohno., 1970); subfunctionalization (Stoltzfus., 1999; Force et al., 1999)). By considering both evolutionary phenomena together (HGT and gene duplication), we can better understand their contribution to oomycete evolution - paralog functions post-HGT have not been broadly studied previously, although the phenomenon is evident (e.g. Richards et al., 2009; Richards et al., 2011; Savory et al., 2015). Exploring functional evidence of the phenotypic fate post-acquisition of horizontally-transferred, expanded enzyme families also allows us to better understand how phytopathogenic oomycetes are able to digest host-derived substrates with greater efficiency, as well as how they respond to the host immune system (e.g. Ma et al., 2017).

Chapter Three re-confirmed 11 previously-identified HGT events associated with plant cell wall digestion (Torto et al., 2002; Richards et al., 2006; Belbahri et al., 2008; Richards et al., 2011; Misner et al., 2015; Savory et al., 2015) with updated taxon sampling, and confirmed the total paralog numbers of HGT genes across oomycetes spanning different ecological lifestyles. All of the HGTs are absent in the sister group to the oomycetes, the hyphochytridiomycetes (specifically *H. catenoides*) - consistent with previous work that demonstrates many gene families associated with phytopathogenicity are absent in this organism (Richards et al., 2011; Leonard et al., 2018). Interestingly, many of the HGTs have been largely expanded in hemibiotrophic oomycetes (*Phytophthora* spp.), suggesting a lifestyle-specific influence of HGT (of these gene families) to oomycete evolution. Most of the HGTs encode secreted CAZymes (GHs), and the results of this work are consistant with previous studies that demonstrate higher numbers of GH genes in *Phytophthora* genomes compared to *Pythium* genomes (e.g. Zerillo et al., 2013), as well as fewer CAZymes encoded in the genomes of obligate biotrophs within the *Phytophthora* lineage (e.g. Baxter et al., 2010; Kemen et al., 2011). Biotrophs must maintain a close association with their hosts without stimulating cell death, so it is unsurprising that higher numbers of plant cell wall digestive enzymes would be encoded by hemibiotrophs associated with a later necrotrophic phase.

P. sojae (a hemibiotrophic pathogen of soybean) was used as a model species for analysis in this work, because it is a well-studied organism within the *Phytophthora* genus, and an economically-significant plant parasite (Kamoun et al., 2015). The *P. sojae* genome is 95 Mb (Tyler et al., 2006), and the putative secretome is predicted to include 1464-1756 proteins (Tyler et al., 2006; Richards et al., 2011; Adhikari et al., 2013; McGowan and Fitzpatrick., 2017). The selective benefit of maintaining the fungal-oomycete HGTs is suggested by widespread gene duplication, putatively contributing to >6% of the *P. sojae* predicted secretome (Tyler et al., 2006; Richards et al., 2011; Adhikari et al., 2013; McGowan and Fitzpatrick., 2017). Rapid evolution of the laterally-acquired genes (through selection pressures and subsequent gene duplication events) reflects their putative importance for phytopathogenicity for hemibiotrophic oomycetes. Two HGT gene families were selected for further analysis in *P. sojae* - a GH12 enzyme family (endo-1,4-β-glucanase (EC 3.2.1.4), xyloglucan endo-hydrolase (EC 3.2.1.151) and endo-1,3-1,4-β-glucanase (EC 3.2.1.73) activities), and a GH10 enzyme family (endo-1,4- β-xylanase (EC 3.2.1.8), endo-1,3- β-xylanase (EC 3.2.1.32) and xylan endotransglycosylase (EC 2.4.2.-) activities). The presence of alternative cellulose- and xylan-degrading capabilities across the oomycetes suggests there is greater diversity and expansion of such activities among hemibiotrophs (consistent with the distribution of the 11 original HGT events) – again, higher numbers of CAZymes associated with GH12- or GH10-like activities were identified in *Phytophthora* spp., suggesting their presence and expansion by subsequent gene duplication events correlates with the importance of the specific functions for a hemibiotrophic lifestyle.

Publically-available transcriptome (RNA-sequencing) data (FungiDB; Stajich et al., 2012; Basenko et al., 2018)) indicated that all GH12 and GH10 paralogs are uniquely expressed across *P. sojae* life stages, and most were shown to be upregulated during infection of its soybean host (i.e. when host-derived substrates are likely to be digested at increased capacity). It was hypothesised that differences in amino acid sequences and three-dimensional protein structures between the paralogs could result in differences in enzyme activity or in novel ways to interact with carbohydrate substrates (or plant hosts) in addition to increasing the transcriptional load - therefore, putative functional differences between *P. sojae* GH12 and GH10 paralogs (11 and 4 proteins, respectively) were investigated using bioinformatics tools. Ten of the *P. sojae* GH12 paralogs possess both glutamic acid residues theoretically required for enzymatic activity (Okada et al., 2000) - this is consistent with other *in vivo* work involving the eleventh paralog (*P. sojae*_360375), which is missing a significant portion of its C-terminus and does not appear to possess hydrolytic activity (Ma et al., 2017). Interestingly, this non-active paralog has a strong binding affinity to host immune protein, GmGIP1 (Ma et al., 2017), suggesting important roles of paralogs outside of the functions they are thought to encode (i.e. in this case, subverting or exploiting host defences).

GH12 paralogs, *P. sojae*_559651 and *P. sojae*_482953 were found to possess distinctive structural features and unique genomic locations with tandem-repeat paralogs. Using available carbohydrate-binding site-prediction tools (3DLigandSite: Wass et al., 2010), GH12 paralog, *P. sojae*_559651, was predicted to encode a putative 'second' carbohydrate binding site, involving the residues Gly55, Ala56, Ala57, Thr58, Val97, Phe205, Val206 (residue position numbers given for the protein sequence in the absence of its N-terminal signal peptide), and the binding site prediction was also conserved amongst orthologous proteins in *P. cactorum* and *P. nicotiniae*. Interestingly, two putative indels coding for alanine and serine are conserved amongst the orthologs, and are important for the site prediction (their removal abolished the prediction of the 'second' binding site (whilst the primary binding site common amongst all GH12 paralogs was left intact)).

The full-length protein was expressed and secreted into *S. cerevisiae* culture supernatants – released reducing sugars from DNS assays (Miller., 1959) demonstrated the protein was active towards xyloglucan (β-1,4 linkages), but not towards CMC (β-1,4 linkages), Laminarin (β-1,3 with some β-1,6 linkages), or Avicel (typically used to demonstrate exo-glucanase activity). Higher reducing sugars were released at 30°C compared to 20°C at pH5, 7 and 10, consistent with the optimum temperature for disease development by *P. sojae* (25-30°C; warm soil (Dorrance and Mills., 2012)), and enzymatic activity was also observed by spotting culture supernatants onto SCM-URA agar plates containing xyloglucan – halos around enzyme spots were visible after staining with Congo red. Whilst it was not possible to confirm putative differences in carbohydrate-binding affinities between *P. sojae*_559651 and the remaining GH12 paralogs due to time constraints, the results of the computational analysis suggest that unique sequence and structural features of the paralogs could enhance their enzyme activity and/or interactions with substrates. Future experimental work involving *P. sojae*_559651 will include further characterisation of the putative differences to enzymatic activity resulting from the removal of the two indels responsible for the second binding site prediction - it is hypothesised that the additional carbohydrate-binding site could enable more efficient substrate digestion through increased binding of the xyloglucan backbone, and therefore removal of the additional binding site would be expected to reduce enzymatic activity (but not abolish it as the primary binding site remains intact).

GH12 paralog *P. sojae*_482953, and orthologs in *P. cactorum* and *P. nicotiniae* all possess a long, disordered, significantly phosphorylated C-terminus 'tail', and the C-terminal sequences were unable to be modelled to available protein structures using Phyre2 (Kelly and Sternberg., 2009) – therefore, using computational tools alone, it was not possible to identify a putative functional significance of the 'tail' sequence. The native protein and a truncated version (missing the 186 amino acid C-terminal 'tail') were both expressed and secreted into *S. cerevisiae* culture supernatants – interestingly, the full-length protein demonstrated higher activity towards xyloglucan (by DNS reducing sugar detection (Miller., 1959)), whilst the truncated protein released a significantly lower concentration of reducing sugars during the same incubation time. The results demonstrate that HGT followed by subsequent gene duplication can foster the expansion of paralog function (in this case, the evolved C-terminus tail, which appears to enhance the rate of enzymatic function). This also highlights the importance of investigating the functions of multiple paralogs of the same gene families of CAZymes to better-understand their complex functional roles within the milieu of plant cell walls. Future work will involve characterisation of the orthologous proteins (to confirm the observed reduction in enzymatic activity) - however, it would also be interesting to further explore the significance of the tail sequence to carbohydrate-binding affinity and identify its final conformation within the folded protein (i.e. using crystallography methods) to better understand how the tail sequence enhances the proteins' interaction with xyloglucan.

The gene encoding *P. sojae*_482953 was also knocked out *in vivo*, using CRISPR/Cas9 methods published by Fang et al. (2017)[11]. Whilst no significant effect to *P. sojae* growth was observed using xyloglucan as the sole carbon source, the result indicates that the remaining GH12 paralogs (as well as other genes with overlapping activities (as previously discussed)) likely provide functional compensation for the loss of the paralog and therefore maintain the encoded function. In nature, this would be particularly important for a parasitic microbe under attack from host immune proteins – secretion of multiple paralogs with subtle sequence and structural differences could affect detection by the host (as previously explored with a catalytically-inactive GH12 paralog (Ma et al., 2017)). Of particular interest would be future work involving multiplexed CRISPR/Cas methods to knockout the entire GH12 family in *P. sojae* – enabling us to better understand the impact of HGT and subsequent gene duplication events to carbon utilisation *in vivo*.

[11] Generation of the *P. sojae* mutants was carried out in the research group of Professor Joseph Heitman (Duke University, North Carolina), using methods implemented in this organism (Fang and Tyler., 2016); Fang et al., 2017), Thanks to Professor Heitman and Dr Fang for their support during this part of the project.

The oligosaccharides released from xyloglucan breakdown by GH12 paralogs, *P. sojae*_482953 and *P. sojae*_559651, were investigated using mass spectrometry (MALDI analysis carried out at the NMSF at Swansea University). After 90 hours incubation with xyloglucan (pH7, 30°C), three peaks were observed (ions with *m/z* 1085, 1247 and 1409), correlating to the released oligosaccharides XXXG, XXLG (or XLXG), and XLLG (Fry et al., 1993). Comparison of the relative intensities of the oligosaccharide species released by both paralogs indicated that *P. sojae*_482953 and *P. sojae*_559651 putatively bind to different parts of the xyloglucan backbone - whilst *P. sojae*_482953 released increased XXLG/XLXG oligosaacharides following incubation with xyloglucan, *P. sojae*_559651 released increased XLLG, so it is possible that the paralogs display preferences for the carbohydrate backbone (by, for example, binding to different side chains (Fry et al., 1993)). Again, with an optimised carbohydrate-binding assay, it would be interesting to further investigate these putative binding preferences, but it would be feasible to hypothesise that a greater efficiency of *in vivo* xyloglucan digestion could be achieved through the secretion of two catalytically-active GH12 paralogs with preferences in binding to different parts of the xyloglucan chain. Unfortunately, it was not possible to accurately detect the release of single sugars from the MS spectra generated – therefore, it is also possible that the enzymes are also further processing the larger oligosaccharides released, which would be useful to investigate further (potentially with alternative spectrometry techniques). It would be interesting to better understand the release of smaller mono- and di-saccharides from xyloglucan breakdown by the proteins, to gain a clearer picture of putative differences in digestive activity between the paralogs.

All *P. sojae* GH10 paralogs possess both glutamic acid residues theoretically required for enzymatic activity. Interestingly, *P. sojae*_527497 was also found to encode a phosphorylated ‘tail’ sequence of ~63 amino acids – when expressed and secreted into *S. cerevisiae* culture supernatants, this paralog displayed activity towards xylan at pH5, pH7 and pH10 (30°C) (by DNS assay (Miller., 1959). *P. sojae*_519234 was the only other paralog expressed in *S. cerevisiae* that displayed activity towards xylan; using available carbohydrate-binding site prediction tools (3DLigandSite: Wass et al., 2010), all GH10 paralogs were found to putatively encode varying amino acid residues within their predicted carbohydrate-binding sites – suggesting there could be exposure of different amino acid side chains to the xylan substrate between the paralogs secreted *in vivo*, putatively interacting with different parts of the xylan chain for more efficient digestion (as mentioned for GH12). This further emphasises that it is crucial to investigate the functions of multiple paralogs of HGT enzyme families and their interactions with substrates, to improve understanding of the functional consequences of HGT and gene duplication events.

As formerly discussed, previously identified fungal-oomycete HGTs encode secreted proteins – these are targeted extracellularly by N-terminal signal peptide sequences. Although biological and physical barriers to HGT have been previously described (e.g. Thomas and Nielsen., 2005), less is known about the transferability of genes encoding secreted proteins. Therefore, this work also aimed to develop methods that will be used for an on-going mutational perturbation study, to investigate the putative degeneracy of signal peptide sequences (e.g. Dunn and Paavilainen., 2019), alongside the hypothesis that N-terminal signal peptide sequences represent a putative barrier to the HGT of genes encoding secreted proteins. Eukaryotic signal peptides are heterogeneous (even amongst paralogs of the same gene family), despite sharing a basic tripartite architecture; co-translational translocation is complex and involves recognition of an appropriate N-terminal sequence by cellular machinery, therefore it is interesting to consider how cross-phylum HGT could be restricted by signal peptide differences (i.e. if a donor sequence is incompatible with the recipient's secretory pathway, or results in inefficient secretion).

As a proof-of-principle experiment, randomised N-terminal signal peptide sequences were generated using a degenerate oligonucleotide (ATG-NNK[18]) cloned upstream of a *P. sojae* GH12 xylo-glucanase missing it's native signal peptide sequence. Transformation of the library into *S. cerevisiae* resulted in yeast transformants expressing one of a number of possible signal peptide sequences upstream of the enzyme, and secreted enzyme activity in the recombinant host was used as a proxy for signal peptide function and efficacy. Congo red staining and halo screening of transformants allowed successful identification of positive clones, providing a useful method for testing large libraries of randomised N-terminal signal peptide sequences. PCR and Sanger sequencing of representative clones suggested redundancy of the cloning method (i.e. several clones harboured the same signal peptide sequence) - whilst this reduces the diversity of variants sampled and requires more exhaustive screening, the increased replication gives further confidence in the method as variants harbouring the same N-terminal sequence generally exhibited the same phenotype.

Ultimately, it would be useful to screen large libraries of signal peptide sequences rapidly, therefore a micro-droplet method was also trialled, which aimed to encapsulate single recombinant *S. cerevisiae* cells (expressing one of all possible signal peptide sequences), and infer substrate breakdown through a shift in Congo red absorbance when it is no longer bound to soluble carbohydrate (Haft et al., 2012)[12]. Based on initial analysis involving (i) *P. sojae*_482953 expressed in plasmid p426-GPD downstream of the *S. cerevisiae* MFα N-terminal signal peptide sequence (positive control), and (ii) a p426-GPD vector-only (negative control)), it was possible to culture the yeast cells in droplets supplemented with SCM-URA (+ 1% (w/v) xyloglucan). Pico-injection of Congo red (+ NaCl) followed by signal capture of a series of droplets (each conceptually representing an experiment) enabled a huge amount of data to be generated in a relatively short amount of time. The preliminary data is promising because it highlights a population of droplets in the positive sample (i.e. *S. cerevisiae* secreting the active enzyme) which putatively represents droplets in which there is a shift in Congo red absorbance (and therefore may represent xyloglucan degradation). It will be exciting to further optimise this method in order to screen several orders of magnitude higher numbers of N-terminal sequences for signal peptide function.

Assessment of signal peptide variants will ultimately allow us to construct a mutational landscape – a theoretical sequence space containing all possible N-terminal signal peptides (19 amino acids in length), linked to their function (as inferred by detection of the secreted enzyme activity). The sequence space will

[12] Thanks to Dr Fabrice Gielen and Vasilis Anagnostidis (University of Exeter) for their help and support during the micro-droplet component of the project.

contain both positive and negative sequences for the secretion of a heterologously-expressed (*'acquired'*) xylo-glucanase in *S. cerevisiae* - informative for exploring how a protein acquired laterally could theoretically traverse sequence space in order to evolve, improve, or lose a functional signal peptide, as well as improving insights into the evolution of N-terminal secretion sequences. By inferring how costly different mutations may be, this will improve our understanding of the evolutionary dynamics relating to successful acquisition and processing of foreign secreted proteins gained through HGT events, as well as provide a better understanding of the relationship between sequence variation and protein coding.

In conclusion, strong phylogenetic evidence has demonstrated that the oomycetes branch separately to the "true" fungi (Förster et al., 1990; Leclerc et al., 2000; Hudspeth et al., 2000; Hudspeth et al., 2003; Thines et al., 2007; McCarthy and Fitzpatrick., 2017). Their diversity in forms includes hemibiotrophic plant parasites, whose genomes are abundant in genes encoding secreted digestive enzymes, including previously identified HGTs from fungi (Torto et al., 2002; Richards et al., 2006; Belbahri et al., 2008; Richards et al., 2011; Misner et al., 2015; Savory et al., 2015). Such gene transfers between organisms with very divergent evolutionary histories can enable organisms to take great leaps across evolutionary sequence space and 'bypass' some of the constraints associated with vertical evolution. Genes can also be acquired through duplication events, and this work has demonstrated that for an acquired GH12, subsequent paralog evolution has given rise to at least two proteins with unique sequence and structural features important for (plant-specific) substrate digestion. Confident functional annotation is limited if proteins are only distantly related to characterised ones (or are paralogous, or have promiscuous functions), therefore it is crucial to experimentally investigate multiple paralogs of protein families, including those acquired by HGT. By coupling gene-duplicated horizontal transfers with improved knowledge of function, we can better understand the selective benefit of the transfer and maintenance, which is important for understanding how HGT genes are maintained (or fixed) within recipient genomes. Of further interest is how N-terminal signal peptide sequences affect the transferability of secreted proteins by HGT – this work has optimised a functional agar plate screen (and demonstrated a promising micro-droplet approach) in order to explore signal peptide evolution in future work.

Summary

- Previously identified HGTs from fungi to oomycetes encode secreted proteins with putative functions for plant cell wall degradation - subsequent gene duplications have given rise to unexplored paralogs that could encode important functional differences.
- *P. sojae* GH12 paralog _559651 and orthologs in *P. cactorum* and *P. nicotiniae* encode a putative ‘second’ carbohydrate binding site, involving the residues Gly55, Ala56, Ala57, Thr58, Val97, Phe205, Val206; two putative indels coding for alanine and serine are important for the site prediction.
- *P. sojae* GH12 paralog _482953 and orthologs in *P. cactorum* and *P. nicotiniae* encode a disordered, significantly phosphorylated C-terminal ‘tail’, which increases the proteins enzymatic activity towards xyloglucan.
- Knockout of the gene encoding *P. sojae*_482953 *in vivo* demonstrated no significant effect to *P. sojae* growth with xyloglucan as a sole carbon

source, suggesting functional compensation for the loss of the paralog by other genes.

- MALDI-MS analysis suggests that GH12 paralogs *P. sojae*_482953 and *P. sojae*_559651 putatively bind to different parts of the xyloglucan backbone.
- The results demonstrate that HGT followed by subsequent gene duplication can foster the expansion of paralog function.
- There are biological and physical barriers to HGT, but less is known about the transferability of genes encoding secreted proteins; this work has optimised a functional agar plate screen (and demonstrated a promising micro-droplet approach) in order to explore signal peptide evolution in future work.

Appendix I

List of HGTs with putative protein functions.

HGT (reference)	Annotation based on sequence similarity	Putative function	Evidence of function in oomycetes
CO-esterase (Richards et al., 2011, Savory et al., 2015)	Carboxylesterase EC 3.1.1.1	Breakdown of plant waxy cuticle	
Cutinase (Belbahri et al 2008, Savory et al., 2015)	Cutinase EC 3.1.1.74	Breakdown of plant waxy cuticle	
GH12 (Richards et al., 2011, Savory et al., 2015)	Endoglucanase Glycosyl hydrolase 12 family	Breakdown of **cellulose**	*P. sojae* XEG1 (_559651) (Ma et al., 2015); silencing and overexpression reduced virulence. XLP1 (_360375) (Ma et al., 2017); no reducing sugars released in apoplastic fluid of *Nicotiana benthamiana* leaves transiently expressing protein (catalytically inactive); deletion reduced

			virulence, overexpression increased virulence.
GH10 (Richards et al., 2011, Savory et al., 2015)	Xylanase Glycosyl hydrolase 10 family	Breakdown of **xylan** (component of **hemicellulose**)	*P. parasitica* (pp) xyn1-4 (Lai and Liou., 2018); all genes upregulated during infection; silencing xyn1 and xyn2 reduced virulence towards *N. benthamiana.*
GH43 (Richards et al., 2011, Savory et al., 2015)	Endo alpha-L-arabinosidase Glycosyl hydrolase 43 family	Breakdown of **arabinan** (component of **hemicellulose**)	
GH78 (Richards et al., 2011, Savory et al., 2015)	Alpha-L-rhamnosidase Glycosyl hydrolase 78 family	Breakdown of **pectin**	
GH53 (Richards et al., 2011, Savory et al., 2015)	Arabinogalactan endo-1,4-beta-galactosidase Glycosyl hydrolase 53 family	Breakdown of arabinogalactans (component of **pectin)**	
Pectolyase (Richards et al., 2011, Savory et al., 2015)	Pectolyase EC 4.2.2.2	Breakdown of **pectin**	
GH88 (Richards et al., 2011, Savory et al., 2015)	d-4,5-unsaturated beta-glucuronyl hydrolase Glycosyl hydrolase 88 family	Breakdown of **pectin**	

FAD-binding (Richards et al., 2011, Savory et al., 2015)	FAD-binding Broad specificity	Broad specificity	
GH28 (Torto et al 2002, Savory, 2015)	Polygalacturonase Glycosyl hydrolase 28 family	Breakdown of **pectin**	*P. parasitica* pppg1 (Lan and Yiou., 2005); strong induction during infection, pppg1-10 (Wu and Liou., 2008); pppg4, 6 and 7 were upregulated post-inoculation on tomato leaves (ppp3, 5, 8 and 9 not detectable)

Bibliography

Adhikari, B. N. *et al.* (2013) ‘Comparative Genomics Reveals Insight into Virulence Strategies of Plant Pathogenic Oomycetes’, *PLOS ONE*, 8(10), p. e75072. doi: 10.1371/journal.pone.0075072.

Agresti, J. J. *et al.* (2010) ‘Ultrahigh-throughput screening in drop-based microfluidics for directed evolution’, *Proceedings of the National Academy of Sciences*, 107(9), pp. 4004–4009. doi: 10.1073/pnas.0910781107.

Ahn, D. H., Kim, H. and Young Pack, M. (1997) ‘Imobilization of β-glucosidase using the cellulose-binding domain of Bacillus subtilis endo-β-1,4-glucanase’, *Biotechnology Letters*, 19(5), p. 483. doi: 10.1023/A:1018360530691.

Akiba, T. *et al.* (1960) ‘On the Mechanism of the Development of Multiple Drug-Resistant Clones of Shigella’, *Japanese Journal of Microbiology*, 4(2), pp. 219–227. doi: 10.1111/j.1348-0421.1960.tb00170.x.

Akopian, D. *et al.* (2013) ‘Signal recognition particle: an essential protein-targeting machine’, *Annual Review of Biochemistry*, 82, pp. 693–721. doi: 10.1146/annurev-biochem-072711-164732.

Alsmark, C. *et al.* (2013) ‘Patterns of prokaryotic lateral gene transfers affecting parasitic microbial eukaryotes’, *Genome Biology*, 14(2), p. R19. doi: 10.1186/gb-2013-14-2-r19.

Altschul, S. F. *et al.* (1990) ‘Basic local alignment search tool’, *Journal of Molecular Biology*, 215(3), pp. 403–410. doi: 10.1016/S0022-2836(05)80360-2.

Álvarez, C., Reyes-Sosa, F. M. and Díez, B. (2016) ‘Enzymatic hydrolysis of biomass from wood’, *Microbial Biotechnology*, 9(2), pp. 149–156. doi: 10.1111/1751-7915.12346.

Amorós-Moya, D. *et al.* (2010) ‘Evolution in Regulatory Regions Rapidly Compensates the Cost of Nonoptimal Codon Usage’, *Molecular Biology and Evolution*, 27(9), pp. 2141–2151. doi: 10.1093/molbev/msq103.

Anderson, C. T. *et al.* (2010) ‘Real-Time Imaging of Cellulose Reorientation during Cell Wall Expansion in Arabidopsis Roots’, *Plant Physiology*, 152(2), pp. 787–796. doi: 10.1104/pp.109.150128.

Andersson, J. O. *et al.* (2003) ‘Phylogenetic Analyses of Diplomonad Genes Reveal Frequent Lateral Gene Transfers Affecting Eukaryotes’, *Current Biology*, 13(2), pp. 94–104. doi: 10.1016/S0960-9822(03)00003-4.

Andersson, J. O. (2005) ‘Lateral gene transfer in eukaryotes’, *Cellular and Molecular Life Sciences*, 62(11), pp. 1182–1197. doi: 10.1007/s00018-005-4539-z.

Andersson, J. O. (2009) ‘Gene Transfer and Diversification of Microbial Eukaryotes’, *Annual Review of Microbiology*, 63(1), pp. 177–193. doi: 10.1146/annurev.micro.091208.073203.

Andersson, J. O. and Roger, A. J. (2003) ‘Evolution of glutamate dehydrogenase genes: evidence for lateral gene transfer within and between prokaryotes and eukaryotes’, *BMC Evolutionary Biology*, 3(1), p. 14. doi: 10.1186/1471-2148-3-14.

Arakawa, K. (2016) ‘No evidence for extensive horizontal gene transfer from the draft genome of a tardigrade’, *Proceedings of the National Academy of Sciences*, 113(22), pp. E3057–E3057. doi: 10.1073/pnas.1602711113.

Archibald, J. M. *et al.* (2003) ‘Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga Bigelowiella natans’, p. 6.

Atmodjo, M. A., Hao, Z. and Mohnen, D. (2013) ‘Evolving Views of Pectin Biosynthesis’, *Annual Review of Plant Biology*, 64(1), pp. 747–779. doi: 10.1146/annurev-arplant-042811-105534.

Bae, J. *et al.* (2013) ‘Cellulosome complexes: natural biocatalysts as arming microcompartments of enzymes’, *Journal of Molecular Microbiology and Biotechnology*, 23(4–5), pp. 370–378. doi: 10.1159/000351358.

Baldauf, S. L. and Palmer, J. D. (1993) ‘Animals and fungi are each other’s closest relatives: congruent evidence from multiple proteins’, *Proceedings of the National Academy of Sciences*, 90(24), pp. 11558–11562. doi: 10.1073/pnas.90.24.11558.

Baltrus, D. A. (2013) ‘Exploring the costs of horizontal gene transfer’, *Trends in Ecology & Evolution*, 28(8), pp. 489–495. doi: 10.1016/j.tree.2013.04.002.

Barbas, C. F. *et al.* (1994) ‘In vitro evolution of a neutralizing human antibody to human immunodeficiency virus type 1 to enhance affinity and broaden strain cross-reactivity’, *Proceedings of the National Academy of Sciences*, 91(9), pp. 3809–3813. doi: 10.1073/pnas.91.9.3809.

Baret, J.-C. *et al.* (2009) ‘Fluorescence-activated droplet sorting (FADS): efficient microfluidic cell sorting based on enzymatic activity’, *Lab on a Chip*, 9(13), pp. 1850–1858. doi: 10.1039/B902504A.

Barrangou, R. *et al.* (2007) ‘CRISPR provides acquired resistance against viruses in prokaryotes’, *Science (New York, N.Y.)*, 315(5819), pp. 1709–1712. doi: 10.1126/science.1138140.

BARTNICKI-GARCIA, S. (1966) ‘Chemistry of Hyphal Walls of Phytophthora’, *Microbiology,* 42(1), pp. 57–69. doi: 10.1099/00221287-42-1-57.

Bartnicki-Garcia, S. (1968) ‘Cell Wall Chemistry, Morphogenesis, and Taxonomy of Fungi’, *Annual Review of Microbiology*, 22(1), pp. 87–108. doi: 10.1146/annurev.mi.22.100168.000511.

Bartnicki-Garcia, S. and Lippman, E. (1969) ‘Fungal Morphogenesis: Cell Wall Construction in Mucor rouxii’, *Science*, 165(3890), pp. 302–304. doi: 10.1126/science.165.3890.302.

Basenko, E. Y. *et al.* (2018) ‘FungiDB: An Integrated Bioinformatic Resource for Fungi and Oomycetes’, *Journal of Fungi*, 4(1). doi: 10.3390/jof4010039.

Bauer, W. D. *et al.* (1973) ‘The Structure of Plant Cell Walls: II. The Hemicellulose of the Walls of Suspension-cultured Sycamore Cells’, *Plant Physiology*, 51(1), pp. 174–187. doi: 10.1104/pp.51.1.174.

Baxter, L. *et al.* (2010) ‘Signatures of Adaptation to Obligate Biotrophy in the Hyaloperonospora arabidopsidis Genome’, *Science*, 330(6010), pp. 1549–1551. doi: 10.1126/science.1195203.

Beakes, G. W., Glockling, S. L. and Sekimoto, S. (2012) ‘The evolutionary phylogeny of the oomycete “fungi”’, *Protoplasma*, 249(1), pp. 3–19. doi: 10.1007/s00709-011-0269-2.

Behura, S. K. and Severson, D. W. (2012) ‘Comparative analysis of codon usage bias and codon context patterns between dipteran and hymenopteran sequenced genomes’, *PloS one*. 2012/08/17 edn, 7(8), pp. e43111–e43111. doi: 10.1371/journal.pone.0043111.

Belbahri, L. *et al.* (2008) ‘Evolution of the cutinase gene family: Evidence for lateral gene transfer of a candidate Phytophthora virulence factor’, *Gene*, 408(1), pp. 1–8. doi: 10.1016/j.gene.2007.10.019.

Bemm, F. *et al.* (2016) ‘Genome of a tardigrade: Horizontal gene transfer or bacterial contamination?’, *Proceedings of the National Academy of Sciences*, 113(22), pp. E3054–E3056. doi: 10.1073/pnas.1525116113.

Beneyton, T. *et al.* (2014) ‘CotA laccase: high-throughput manipulation and analysis of recombinant enzyme libraries expressed in E. coli using droplet-based microfluidics’, *Analyst*, 139(13), pp. 3314–3323. doi: 10.1039/C4AN00228H.

Beneyton, T. *et al.* (2017) 'Droplet-based microfluidic high-throughput screening of heterologous enzymes secreted by the yeast Yarrowia lipolytica', *Microbial Cell Factories*, 16(1), p. 18. doi: 10.1186/s12934-017-0629-5.

Benhamou, N. *et al.* (2012) 'Pythium oligandrum: an example of opportunistic success', *Microbiology,* 158(11), pp. 2679–2694. doi: 10.1099/mic.0.061457-0.

Bergthorsson, U. *et al.* (2003) 'Widespread horizontal transfer of mitochondrial genes in flowering plants', *Nature*, 424(6945), pp. 197–201. doi: 10.1038/nature01743.

Bershtein, S. *et al.* (2015) 'Protein Homeostasis Imposes a Barrier on Functional Integration of Horizontally Transferred Genes in Bacteria', *PLOS Genetics*, 11(10), p. e1005612. doi: 10.1371/journal.pgen.1005612.

Bershtein, S., Goldin, K. and Tawfik, D. S. (2008) 'Intense neutral drifts yield robust and evolvable consensus proteins', *Journal of Molecular Biology*, 379(5), pp. 1029–1044. doi: 10.1016/j.jmb.2008.04.024.

Bhattacharjee, S. *et al.* (2006) 'The Malarial Host-Targeting Signal Is Conserved in the Irish Potato Famine Pathogen', *PLOS Pathogens*, 2(5), p. e50. doi: 10.1371/journal.ppat.0020050.

Biely, P. *et al.* (1997) 'Endo-β-1,4-xylanase families: differences in catalytic properties', *Journal of Biotechnology*, 57(1), pp. 151–166. doi: 10.1016/S0168-1656(97)00096-5.

Biely, P., Krátký, Z. and Vršanská, M. (1981) 'Substrate-Binding Site of Endo-1,4-β-Xylanase of the Yeast Cryptococcus albidus', *European Journal of Biochemistry*, 119(3), pp. 559–564. doi: 10.1111/j.1432-1033.1981.tb05644.x.

Biely, P., Singh, S. and Puchart, V. (2016) 'Towards enzymatic breakdown of complex plant xylan structures: State of the art', *Biotechnology Advances*, 34(7), pp. 1260–1274. doi: 10.1016/j.biotechadv.2016.09.001.

Bird, P., Sambrook, J. and Gething, M.-J. (1990) ‘The Functional Efficiency of a Mammalian Signal Peptide Is Directly Related to Its Hydrophobicity’, p. 7.

Bischof, R. H., Ramoni, J. and Seiboth, B. (2016) ‘Cellulases and beyond: the first 70 years of the enzyme producer Trichoderma reesei’, *Microbial Cell Factories*, 15(1), p. 106. doi: 10.1186/s12934-016-0507-6.

Blobel, G. and Dobberstein, B. (1975) ‘Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma’, *The Journal of Cell Biology*, 67(3), pp. 835–851. doi: 10.1083/jcb.67.3.835.

Blobel, G. and Sabatini, D. D. (1971) ‘Ribosome-Membrane Interaction in Eukaryotic Cells’, in Manson, L. A. (ed.) *Biomembranes: Volume 2*. Boston, MA: Springer US, pp. 193–195. doi: 10.1007/978-1-4684-3330-2_16.

Blom, N., Gammeltoft, S. and Brunak, S. (1999) ‘Sequence and structure-based prediction of eukaryotic protein phosphorylation sites1 1Edited by F. E. Cohen’, *Journal of Molecular Biology*, 294(5), pp. 1351–1362. doi: 10.1006/jmbi.1999.3310.

Bogorad, L. (2008) ‘Evolution of early eukaryotic cells: genomes, proteomes, and compartments’, *Photosynthesis Research*, 95(1), pp. 11–21. doi: 10.1007/s11120-007-9236-3.

Boraston, A. B. *et al.* (2004) ‘Carbohydrate-binding modules: fine-tuning polysaccharide recognition’, *Biochemical Journal*, 382(3), pp. 769–781. doi: 10.1042/BJ20040892.

Boucher, Y. *et al.* (2003) ‘Lateral Gene Transfer and the Origins of Prokaryotic Groups’, *Annual Review of Genetics*, 37(1), pp. 283–328. doi: 10.1146/annurev.genet.37.050503.084247.

Brachmann, C. B. *et al.* (1998) 'Designer deletion strains derived from Saccharomyces cerevisiae S288C: A useful set of strains and plasmids for PCR-mediated gene disruption and other applications', *Yeast*, 14(2), pp. 115–132. doi: 10.1002/(SICI)1097-0061(19980130)14:2<115::AID-YEA204>3.0.CO;2-2.

Brown, D. M. *et al.* (2007) 'Comparison of five xylan synthesis mutants reveals new insight into the mechanisms of xylan synthesis', *The Plant Journal*, 52(6), pp. 1154–1168. doi: 10.1111/j.1365-313X.2007.03307.x.

Brown, J. D. *et al.* (1994) 'Subunits of the Saccharomyces cerevisiae signal recognition particle required for its functional expression.', *The EMBO Journal*, 13(18), pp. 4390–4400. doi: 10.1002/j.1460-2075.1994.tb06759.x.

Buchan, J. R., Aucott, L. S. and Stansfield, I. (2006) 'tRNA properties help shape codon pair preferences in open reading frames', *Nucleic acids research*, 34(3), pp. 1015–1027. doi: 10.1093/nar/gkj488.

Buckeridge, M. S. *et al.* (1992) 'Xyloglucan structure and post-germinative metabolism in seeds of Copaifera langsdorfii from savanna and forest populations', *Physiologia Plantarum*, 86(1), pp. 145–151. doi: 10.1111/j.1399-3054.1992.tb01323.x.

Busse-Wicher, M. *et al.* (2014) 'The pattern of xylan acetylation suggests xylan may interact with cellulose microfibrils as a twofold helical screw in the secondary plant cell wall of Arabidopsis thaliana', *The Plant Journal*, 79(3), pp. 492–506. doi: 10.1111/tpj.12575.

Buszard, B. J. *et al.* (2013) 'The Nucleus- and Endoplasmic Reticulum-Targeted Forms of Protein Tyrosine Phosphatase 61F Regulate Drosophila Growth, Life Span, and Fecundity', *Molecular and Cellular Biology*, 33(7), pp. 1345–1356. doi: 10.1128/MCB.01411-12.

Byrne, K. A. *et al.* (1999) 'Isolation of a cDNA encoding a putative cellulase in the red claw crayfish Cherax quadricarinatus', *Gene*, 239(2), pp. 317–324. doi: 10.1016/S0378-1119(99)00396-0.

Callens, M., Scornavacca, C. and Bedhomme, S. (2020) ‘Evolutionary responses to codon usage of horizontally transferred genes in Pseudomonas aeruginosa’, *bioRxiv*, p. 2020.07.11.198432. doi: 10.1101/2020.07.11.198432.

Carneiro, M. and Hartl, D. L. (2010) ‘Adaptive landscapes and protein evolution’, *Proceedings of the National Academy of Sciences*, 107(suppl 1), pp. 1747–1751. doi: 10.1073/pnas.0906192106.

Carter, P. J. (2006) ‘Potent antibody therapeutics by design’, *Nature Reviews Immunology*, 6(5), pp. 343–357. doi: 10.1038/nri1837.

Castresana, J. (2000) ‘Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis’, *Molecular Biology and Evolution*, 17(4), pp. 540–552. doi: 10.1093/oxfordjournals.molbev.a026334.

Cavalier, D. M. *et al.* (2008) ‘Disrupting Two Arabidopsis thaliana Xylosyltransferase Genes Results in Plants Deficient in Xyloglucan, a Major Primary Cell Wall Component’, *The Plant Cell*, 20(6), pp. 1519–1537. doi: 10.1105/tpc.108.059873.

Cavalier-Smith, T. (1981) ‘Eukaryote kingdoms: Seven or nine?’, *Biosystems*, 14(3), pp. 461–481. doi: 10.1016/0303-2647(81)90050-2.

CAVALIER-SMITH, T. (1986) ‘The kingdom Chromista : Origin and systematics’, *Progress in phycological research*, 4, pp. 309–347.

Cavalier-Smith, T. (1999) ‘Principles of Protein and Lipid Targeting in Secondary Symbiogenesis: Euglenoid, Dinoflagellate, and Sporozoan Plastid Origins and the Eukaryote Family Tree1,2’, *Journal of Eukaryotic Microbiology*, 46(4), pp. 347–366. doi: 10.1111/j.1550-7408.1999.tb04614.x.

Cavalier-Smith, T. and Chao, E. E.-Y. (2006) ‘Phylogeny and Megasystematics of Phagotrophic Heterokonts (Kingdom Chromista)’, *Journal of Molecular Evolution*, 62(4), pp. 388–420. doi: 10.1007/s00239-004-0353-8.

Choi, E.-Y. *et al.* (2002) 'Construction of an industrial polyploid strain of Saccharomyces cerevisiae containing Saprolegnia ferax β-amylase gene and secreting β-amylase', *Biotechnology Letters*, 24(21), pp. 1785–1790. doi: 10.1023/A:1020613306127.

Choi, J. Y., Bubnell, J. E. and Aquadro, C. F. (2015) 'Population Genomics of Infectious and Integrated Wolbachia pipientis Genomes in Drosophila ananassae', *Genome Biology and Evolution*, 7(8), pp. 2362–2382. doi: 10.1093/gbe/evv158.

Chothia, C. *et al.* (2003) 'Evolution of the Protein Repertoire', *Science*, 300(5626), pp. 1701–1703. doi: 10.1126/science.1085371.

Chu, Y. *et al.* (2017) 'Insights into the roles of non-catalytic residues in the active site of a GH10 xylanase with activity on cellulose', *Journal of Biological Chemistry*, 292(47), pp. 19315–19327. doi: 10.1074/jbc.M117.807768.

de Cock, A. W. A. M. *et al.* (2015) 'Phytopythium: molecular phylogeny and systematics', *Persoonia : Molecular Phylogeny and Evolution of Fungi*, 34, pp. 25–39. doi: 10.3767/003158515X685382.

Cohen, O., Gophna, U. and Pupko, T. (2011) 'The Complexity Hypothesis Revisited: Connectivity Rather Than Function Constitutes a Barrier to Horizontal Gene Transfer', *Molecular Biology and Evolution*, 28(4), pp. 1481–1489. doi: 10.1093/molbev/msq333.

Cohen, R., Suzuki, M. R. and Hammel, K. E. (2005) 'Processive Endoglucanase Active in Crystalline Cellulose Hydrolysis by the Brown Rot Basidiomycete Gloeophyllum trabeum', *Applied and Environmental Microbiology*, 71(5), pp. 2412–2417. doi: 10.1128/AEM.71.5.2412-2417.2005.

Colin, P.-Y., Zinchenko, A. and Hollfelder, F. (2015) 'Enzyme engineering in biomimetic compartments', *Current Opinion in Structural Biology*, 33, pp. 42–51. doi: 10.1016/j.sbi.2015.06.001.

Collins, T., Gerday, C. and Feller, G. (2005) ‘Xylanases, xylanase families and extremophilic xylanases’, *FEMS Microbiology Reviews*, 29(1), pp. 3–23. doi: 10.1016/j.femsre.2004.06.005.

Cong, L. *et al.* (2013) ‘Multiplex Genome Engineering Using CRISPR/Cas Systems’, *Science*, 339(6121), pp. 819–823. doi: 10.1126/science.1231143.

Connolly, M. S. *et al.* (2005) ‘Heterologous expression of a pleiotropic drug resistance transporter from Phytophthora sojae in yeast transporter mutants’, *Current Genetics*, 48(6), pp. 356–365. doi: 10.1007/s00294-005-0015-4.

Cooney, E. W., Barr, D. J. S. and Barstow, W. E. (1985) ‘The ultrastructure of the zoospore of Hyphochytrium catenoides’, *Canadian Journal of Botany*, 63(3), pp. 497–505. doi: 10.1139/b85-062.

Costanzo, S. *et al.* (2007) ‘Alternate intron processing of family 5 endoglucanase transcripts from the genus Phytophthora’, *Current Genetics*, 52(3), pp. 115–123. doi: 10.1007/s00294-007-0144-z.

Crooks, G.E. *et al*. (2004) ‘WebLogo: A sequence logo generator’, *Genome Research*, 14, pp. 1188-1190. https://doi.org/10.1101/gr.849004.

Crowley, K. S. *et al.* (1994) ‘Secretory proteins move through the endoplasmic reticulum membrane via an aqueous, gated pore’, *Cell*, 78(3), pp. 461–471. doi: 10.1016/0092-8674(94)90424-3.

Cumming, C. M. *et al.* (2005) ‘Biosynthesis and cell-wall deposition of a pectin–xyloglucan complex in pea’, *Planta*, 222(3), pp. 546–555. doi: 10.1007/s00425-005-1560-2.

Dagan, T., Artzy-Randrup, Y. and Martin, W. (2008) ‘Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution’, *Proceedings of the National Academy of Sciences*, 105(29), pp. 10039–10044. doi: 10.1073/pnas.0800679105.

Dalbey, R. E. and Wickner, W. (1985) ‘Leader peptidase catalyzes the release of exported proteins from the outer surface of the Escherichia coli plasma membrane.’, *Journal of Biological Chemistry*, 260(29), pp. 15925–15931.

Danchin, E. G. J. *et al.* (2010) ‘Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes’, *Proceedings of the National Academy of Sciences*, 107(41), pp. 17651–17656. doi: 10.1073/pnas.1008486107.

Daniels, S. B. *et al.* (1990) ‘Evidence for horizontal transmission of the P transposable element between Drosophila species’, *Genetics*, 124(2), pp. 339–355.

Dantzig, A. H., Zuckerman, S. H. and Andonov-Roland, M. M. (1986) ‘Isolation of a Fusarium solani mutant reduced in cutinase activity and virulence.’, *Journal of Bacteriology*, 168(2), pp. 911–916. doi: 10.1128/jb.168.2.911-916.1986.

Davies, G. and Henrissat, B. (1995) ‘Structures and mechanisms of glycosyl hydrolases’, *Structure*, 3(9), pp. 853–859. doi: 10.1016/S0969-2126(01)00220-9.

Davis, C. C. and Wurdack, K. J. (2004) ‘Host-to-Parasite Gene Transfer in Flowering Plants: Phylogenetic Evidence from Malpighiales’, *Science*, 305(5684), pp. 676–678. doi: 10.1126/science.1100671.

Den Haan, R. *et al.* (2007) ‘Hydrolysis and fermentation of amorphous cellulose by recombinant Saccharomyces cerevisiae’, *Metabolic Engineering*, 9(1), pp. 87–94. doi: 10.1016/j.ymben.2006.08.005.

Derelle, R. *et al.* (2016) ‘A Phylogenomic Framework to Study the Diversity and Evolution of Stramenopiles (=Heterokonts)’, *Molecular Biology and Evolution*, 33(11), pp. 2890–2898. doi: 10.1093/molbev/msw168.

Diao, X., Freeling, M. and Lisch, D. (2005) ‘Horizontal Transfer of a Plant Transposon’, *PLOS Biology*, 4(1), p. e5. doi: 10.1371/journal.pbio.0040005.

DiCarlo, J. E. *et al.* (2013) 'Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems', *Nucleic Acids Research*, 41(7), pp. 4336–4343. doi: 10.1093/nar/gkt135.

Dong, S. *et al.* (2011) 'Sequence Variants of the Phytophthora sojae RXLR Effector Avr3a/5 Are Differentially Recognized by Rps3a and Rps5 in Soybean', *PLOS ONE*, 6(7), p. e20172. doi: 10.1371/journal.pone.0020172.

Dong, S., Raffaele, S. and Kamoun, S. (2015) 'The two-speed genomes of filamentous pathogens: waltz with plants', *Current Opinion in Genetics & Development*, 35, pp. 57–65. doi: 10.1016/j.gde.2015.09.001.

Doolittle, W. F. (1999) 'Phylogenetic Classification and the Universal Tree', *Science*, 284(5423), pp. 2124–2128. doi: 10.1126/science.284.5423.2124.

Doolittle, W. F. and Bapteste, E. (2007) 'Pattern pluralism and the Tree of Life hypothesis', *Proceedings of the National Academy of Sciences*, 104(7), pp. 2043–2049. doi: 10.1073/pnas.0610699104.

Dorrance, A. E. (2018) 'Management of Phytophthora sojae of soybean: a review and future perspectives', *Canadian Journal of Plant Pathology*, 40(2), pp. 210–219. doi: 10.1080/07060661.2018.1445127.

Dorrance, A.E., D. Mills, A.E. Robertson, M.A. Draper, L. Giesler, and A.Tenuta (2007) *Phytophthora root and stem rot of soybean*. Available at: https://www.apsnet.org/edcenter/disandpath/oomycete/pdlessons/Pages/PhytophthoraSojae.aspx

Driouich, A. *et al.* (2012) 'Golgi-Mediated Synthesis and Secretion of Matrix Polysaccharides of the Primary Cell Wall of Higher Plants', *Frontiers in Plant Science*, 3. doi: 10.3389/fpls.2012.00079.

Driouich, A., Faye, L. and Staehelin, A. (1993) 'The plant Golgi apparatus: a factory for complex polysaccharides and glycoproteins', *Trends in Biochemical Sciences*, 18(6), pp. 210–214. doi: 10.1016/0968-0004(93)90191-O.

Drummond, D. A. and Wilke, C. O. (2008) 'Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution', *Cell*, 134(2), pp. 341–352. doi: 10.1016/j.cell.2008.05.042.

Duffy, J., Patham, B. and Mensa-Wilmot, K. (2010) 'Discovery of functional motifs in h-regions of trypanosome signal sequences', *The Biochemical Journal*, 426(2), pp. 135–145. doi: 10.1042/BJ20091277.

Dufresne, M. *et al.* (2000) 'A GAL4-like Protein Is Involved in the Switch between Biotrophic and Necrotrophic Phases of the Infection Process of Colletotrichum lindemuthianum on Common Bean', *The Plant Cell*, 12(9), pp. 1579–1590.

Dunn, C. D. and Paavilainen, V. O. (2019) 'Wherever I may roam: organellar protein targeting and evolvability', *Current Opinion in Genetics & Development*, 58–59, pp. 9–16. doi: 10.1016/j.gde.2019.07.012.

Dunning, L. T. *et al.* (2019) 'Lateral transfers of large DNA fragments spread functional genes among grasses', *Proceedings of the National Academy of Sciences*, 116(10), pp. 4416–4425. doi: 10.1073/pnas.1810031116.

Dunning Hotopp, J. C. and Estes, A. M. (2014) 'Biology Wars: The Eukaryotes Strike Back', *Cell Host & Microbe*, 16(6), pp. 701–703. doi: 10.1016/j.chom.2014.11.014.

Dybdahl, M. F. and Storfer, A. (2003) 'Parasite local adaptation: Red Queen versus Suicide King', *Trends in Ecology & Evolution*, 18(10), pp. 523–530. doi: 10.1016/S0169-5347(03)00223-4.

Dyrløv Bendtsen, J. *et al.* (2004) 'Improved Prediction of Signal Peptides: SignalP 3.0', *Journal of Molecular Biology*, 340(4), pp. 783–795. doi: 10.1016/j.jmb.2004.05.028.

Edgar, R. C. (2004) 'MUSCLE: multiple sequence alignment with high accuracy and high throughput', *Nucleic Acids Research*, 32(5), pp. 1792–1797. doi: 10.1093/nar/gkh340.

Efron, B., Halloran, E. and Holmes, S. (1996) ‘Bootstrap confidence levels for phylogenetic trees’, *Proceedings of the National Academy of Sciences*, 93(23), pp. 13429–13429. doi: 10.1073/pnas.93.23.13429.

Egea, P. F., Stroud, R. M. and Walter, P. (2005) ‘Targeting proteins to membranes: structure of the signal recognition particle’, *Current Opinion in Structural Biology*, 15(2), pp. 213–220. doi: 10.1016/j.sbi.2005.03.007.

Eigen, M. and Schuster, P. (1977) ‘A principle of natural self-organization: Part A: Emergence of the hypercycle’, *Naturwissenschaften*, 64(11), pp. 541–565. doi: 10.1007/BF00450633.

Elvekrog, M. M. and Walter, P. (2015) ‘Dynamics of co-translational protein targeting’, *Current Opinion in Chemical Biology*, 29, pp. 79–86. doi: 10.1016/j.cbpa.2015.09.016.

EMERSON, R. (1941) ‘An experimental study on the life cycles and taxonomy of Allomyces’, *Lloydia*, 4, pp. 77–144.

Evans, E. A., Gilmore, R. and Blobel, G. (1986) ‘Purification of microsomal signal peptidase as a complex’, *Proceedings of the National Academy of Sciences of the United States of America*, 83(3), pp. 581–585. doi: 10.1073/pnas.83.3.581.

Fabritius, A.-L., Cvitanich, C. and Judelson, H. S. (2002) ‘Stage-specific gene expression during sexual development in Phytophthora infestans’, *Molecular Microbiology*, 45(4), pp. 1057–1066. doi: 10.1046/j.1365-2958.2002.03073.x.

Fang, C. *et al.* (2017) ‘High copy and stable expression of the xylanase XynHB in Saccharomyces cerevisiae by rDNA-mediated integration’, *Scientific Reports*, 7(1), p. 8747. doi: 10.1038/s41598-017-08647-x.

Fang, Y. *et al.* (2017) ‘Efficient Genome Editing in the Oomycete Phytophthora sojae Using CRISPR/Cas9’, *Current Protocols in Microbiology*, 44(1), p. 21A.1.1-21A.1.26. doi: 10.1002/cpmc.25.

Fang, Y. and Tyler, B. M. (2016) ‘Efficient disruption and replacement of an effector gene in the oomycete Phytophthora sojae using CRISPR/Cas9’, *Molecular Plant Pathology*, 17(1), pp. 127–139. doi: 10.1111/mpp.12318.

Felsenstein, J. (1985) ‘Confidence Limits on Phylogenies: An Approach Using the Bootstrap’, *Evolution*, 39(4), pp. 783–791. doi: 10.1111/j.1558-5646.1985.tb00420.x.

Ferrari, S. *et al.* (2013) ‘Oligogalacturonides: plant damage-associated molecular patterns and regulators of growth and development’, *Frontiers in Plant Science*, 4. doi: 10.3389/fpls.2013.00049.

Finn, R. D. *et al.* (2015) ‘HMMER web server: 2015 update’, *Nucleic Acids Research*, 43(W1), pp. W30–W38. doi: 10.1093/nar/gkv397.

Finn, R. D. *et al.* (2016) ‘The Pfam protein families database: towards a more sustainable future’, *Nucleic Acids Research*, 44(D1), pp. D279–D285. doi: 10.1093/nar/gkv1344.

Finn, R. D. *et al.* (2017) ‘InterPro in 2017—beyond protein family and domain annotations’, *Nucleic Acids Research*, 45(D1), pp. D190–D199. doi: 10.1093/nar/gkw1107.

Fischlechner, M. *et al.* (2014) ‘Evolution of enzyme catalysts caged in biomimetic gel-shell beads’, *Nature Chemistry*, 6(9), pp. 791–796. doi: 10.1038/nchem.1996.

Fitch, W. M. (1970) ‘Distinguishing Homologous from Analogous Proteins’, *Systematic Biology*, 19(2), pp. 99–113. doi: 10.2307/2412448.

Flanagan, J. J. *et al.* (2003) ‘Signal recognition particle binds to ribosome-bound signal sequences with fluorescence-detected subnanomolar affinity that does not diminish as the nascent chain lengthens’, *The Journal of Biological Chemistry*, 278(20), pp. 18628–18637. doi: 10.1074/jbc.M300173200.

Force, A. *et al.* (1999) ‘Preservation of Duplicate Genes by Complementary, Degenerative Mutations’, *Genetics*, 151(4), pp. 1531–1545.

Ford Doolittle, W. (1998) ‘You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes’, *Trends in Genetics*, 14(8), pp. 307–311. doi: 10.1016/S0168-9525(98)01494-2.

Förster, H. *et al.* (1990) ‘Sequence Analysis of the Small Subunit Ribosomal Rnas of Three Zoosporic Fungi and Implications for Fungal Evolution’, *Mycologia*, 82(3), pp. 306–312. doi: 10.1080/00275514.1990.12025885.

Fortune, P. M., Roulin, A. and Panaud, O. (2008) ‘Horizontal transfer of transposable elements in plants’, *Communicative & integrative biology*, 1(1), pp. 74–77. doi: 10.4161/cib.1.1.6328.

Francino, M. P. (2012) ‘The Ecology of Bacterial Genes and the Survival of the New’, *International Journal of Evolutionary Biology*, 2012, pp. 1–14. doi: 10.1155/2012/394026.

Friesen, T. L. *et al.* (2006) ‘Emergence of a new disease as a result of interspecific virulence gene transfer’, *Nature Genetics*, 38(8), pp. 953–956. doi: 10.1038/ng1839.

Frumkin, I. *et al.* (2018) ‘Codon usage of highly expressed genes affects proteome-wide translation efficiency’, *Proceedings of the National Academy of Sciences*, 115(21), pp. E4940–E4949. doi: 10.1073/pnas.1719375115.

Fry, S. C. *et al.* (1993) ‘An unambiguous nomenclature for xyloglucan-derived oligosaccharides’, *Physiologia Plantarum*, 89(1), pp. 1–3. doi: 10.1111/j.1399-3054.1993.tb01778.x.

Futatsumori-Sugai, M. and Tsumoto, K. (2010) ‘Signal peptide design for improving recombinant protein secretion in the baculovirus expression vector system’, *Biochemical and Biophysical Research Communications*, 391(1), pp. 931–935. doi: 10.1016/j.bbrc.2009.11.167.

Gabaldón, T. and Pittis, A. A. (2015) 'Origin and evolution of metabolic sub-cellular compartmentalization in eukaryotes', *Biochimie*, 119, pp. 262–268. doi: 10.1016/j.biochi.2015.03.021.

Galtier, N., Gouy, M. and Gautier, C. (1996) 'SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny', *Bioinformatics*, 12(6), pp. 543–548. doi: 10.1093/bioinformatics/12.6.543.

Garcia-Vallve, S. *et al.* (2003) 'HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes', *Nucleic Acids Research*, 31(1), pp. 187–189. doi: 10.1093/nar/gkg004.

Gardiner, D. M. *et al.* (2012) 'Comparative Pathogenomics Reveals Horizontally Acquired Novel Virulence Genes in Fungi Infecting Cereal Hosts', *PLOS Pathogens*, 8(9), p. e1002952. doi: 10.1371/journal.ppat.1002952.

Gaston, D. and Roger, A. J. (2013) 'Functional Divergence and Convergent Evolution in the Plastid-Targeted Glyceraldehyde-3-Phosphate Dehydrogenases of Diverse Eukaryotic Algae', *PLOS ONE*, 8(7), p. e70396. doi: 10.1371/journal.pone.0070396.

Gennity, J., Goldstein, J. and Inouye, M. (1990) 'Signal peptide mutants of Escherichia coli', *Journal of Bioenergetics and Biomembranes*, 22(3), pp. 233–269. doi: 10.1007/BF00763167.

Gielen, F. *et al.* (2013) 'A Fully Unsupervised Compartment-on-Demand Platform for Precise Nanoliter Assays of Time-Dependent Steady-State Enzyme Kinetics and Inhibition', *Analytical Chemistry*, 85(9), pp. 4761–4769. doi: 10.1021/ac400480z.

Gielen, F. *et al.* (2016) 'Ultrahigh-throughput–directed enzyme evolution by absorbance-activated droplet sorting (AADS)', *Proceedings of the National Academy of Sciences*, 113(47), pp. E7383–E7389. doi: 10.1073/pnas.1606927113.

Gingold, H. and Pilpel, Y. (2011) ‘Determinants of translation efficiency and accuracy’, *Molecular Systems Biology*, 7(1), p. 481. doi: 10.1038/msb.2011.14.

Grange, D. C. la, Pretorius, I. S. and Zyl, W. H. van (1996) ‘Expression of a Trichoderma reesei beta-xylanase gene (XYN2) in Saccharomyces cerevisiae.’, *Applied and Environmental Microbiology*, 62(3), pp. 1036–1044.

Green, R., Kramer, R. A. and Shields, D. (1989) ‘Misplacement of the amino-terminal positive charge in the prepro-alpha-factor signal peptide disrupts membrane translocation in vivo.’, *Journal of Biological Chemistry*, 264(5), pp. 2963–2968.

Grenville-Briggs, L. J. *et al.* (2008) ‘Cellulose Synthesis in Phytophthora infestans Is Required for Normal Appressorium Formation and Successful Infection of Potato’, *The Plant Cell*, 20(3), pp. 720–738. doi: 10.1105/tpc.107.052043.

Grishutin, S. G. *et al.* (2004) ‘Specific xyloglucanases as a new class of polysaccharide-degrading enzymes’, *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1674(3), pp. 268–281. doi: 10.1016/j.bbagen.2004.07.001.

Guerriero, G. *et al.* (2010) ‘Chitin Synthases from Saprolegnia Are Involved in Tip Growth and Represent a Potential Target for Anti-Oomycete Drugs’, *PLOS Pathogens*, 6(8), p. e1001070. doi: 10.1371/journal.ppat.1001070.

Guindon, S. and Gascuel, O. (2003) ‘A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood’, *Systematic Biology*, 52(5), pp. 696–704. doi: 10.1080/10635150390235520.

Gupta, V. K. *et al.* (2016) ‘Fungal Enzymes for Bio-Products from Sustainable and Waste Biomass’, *Trends in Biochemical Sciences*, 41(7), pp. 633–645. doi: 10.1016/j.tibs.2016.04.006.

Haas, B. J. *et al.* (2009) ‘Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans’, *Nature*, 461(7262), pp. 393–398. doi: 10.1038/nature08358.

Haft, R. J. F., Gardner, J. G. and Keating, D. H. (2012) ‘Quantitative colorimetric measurement of cellulose degradation under microbial culture conditions’, *Applied Microbiology and Biotechnology*, 94(1), pp. 223–229. doi: 10.1007/s00253-012-3968-5.

Haguenauer-Tsapis, R. and Hinnen, A. (1984) ‘A deletion that includes the signal peptidase cleavage site impairs processing, glycosylation, and secretion of cell surface yeast acid phosphatase.’, *Molecular and Cellular Biology*, 4(12), pp. 2668–2675. doi: 10.1128/MCB.4.12.2668.

Hahn, M. G., Darvill, A. G. and Albersheim, P. (1981) ‘Host-Pathogen Interactions: XIX. THE ENDOGENOUS ELICITOR, A FRAGMENT OF A PLANT CELL WALL POLYSACCHARIDE THAT ELICITS PHYTOALEXIN ACCUMULATION IN SOYBEANS’, *Plant Physiology*, 68(5), pp. 1161–1169. doi: 10.1104/pp.68.5.1161.

Hakariya, M., Hirose, D. and Tokumasu, S. (2007) ‘A molecular phylogeny of Haptoglossa species, terrestrial peronosporomycetes (oomycetes) endoparasitic on nematodes’, *Mycoscience*, 48(3), pp. 169–175. doi: 10.1007/s10267-007-0355-7.

Hall, C. and Dietrich, F. S. (2007) ‘The Reacquisition of Biotin Prototrophy in Saccharomyces cerevisiae Involved Horizontal Gene Transfer, Gene Duplication and Gene Clustering’, *Genetics*, 177(4), pp. 2293–2307. doi: 10.1534/genetics.107.074963.

Hamann, T. (2012) ‘Plant cell wall integrity maintenance as an essential component of biotic stress response mechanisms’, *Frontiers in Plant Science*, 3. doi: 10.3389/fpls.2012.00077.

Hann, B. C. and Walter, P. (1991) ‘The signal recognition particle in S. cerevisiae’, *Cell*, 67(1), pp. 131–144. doi: 10.1016/0092-8674(91)90577-L.

Hansen, E. M. (no date) ‘Alien forest pathogens: Phytophthora species are changing world forests’, 13, p. 9.

Haring, E., Hagemann, S. and Pinsker, W. (2000) ‘Ancient and Recent Horizontal Invasions of Drosophilids by P Elements’, *Journal of Molecular Evolution*, 51(6), pp. 577–586. doi: 10.1007/s002390010121.

Harman, G. E. and Kubicek, C. P. (1998) *Trichoderma And Gliocladium, Volume 2: Enzymes, Biological Control and commercial applications*. CRC Press.

Harper, J. T., Waanders, E. and Keeling, P. J. (2005) ‘On the monophyly of chromalveolates using a six-protein phylogeny of eukaryotes’, *International Journal of Systematic and Evolutionary Microbiology,* 55(1), pp. 487–496. doi: 10.1099/ijs.0.63216-0.

Harris, G. W. *et al.* (1994) ‘Structure of the catalytic core of the family F xylanase from Pseudomonas fluorescens and identification of the xylopentaose-binding sites’, *Structure*, 2(11), pp. 1107–1116. doi: 10.1016/S0969-2126(94)00112-X.

Hasper, A. A. *et al.* (2002) ‘EglC, a New Endoglucanase from Aspergillus niger with Major Activity towards Xyloglucan’, *Applied and Environmental Microbiology*, 68(4), pp. 1556–1560. doi: 10.1128/AEM.68.4.1556-1560.2002.

Hayashi, T. and Kaida, R. (2011) ‘Functions of xyloglucan in plant cells’, *Molecular Plant*, 4(1), pp. 17–24. doi: 10.1093/mp/ssq063.

von Heijne, G. (1983) ‘Patterns of amino acids near signal-sequence cleavage sites’, *European Journal of Biochemistry*, 133(1), pp. 17–21. doi: 10.1111/j.1432-1033.1983.tb07424.x.

von Heijne, G. (1985) ‘Signal sequences. The limits of variation’, *Journal of Molecular Biology*, 184(1), pp. 99–105. doi: 10.1016/0022-2836(85)90046-4.

von Heijne, G. (1990) ‘The signal peptide’, *The Journal of Membrane Biology*, 115(3), pp. 195–201. doi: 10.1007/BF01868635.

von Heijne, G. and Abrahmsén, L. (1989) ‘Species-specific variation in signal peptide design. Implications for protein secretion in foreign hosts’, *FEBS letters*, 244(2), pp. 439–446. doi: 10.1016/0014-5793(89)80579-4.

Henrissat, B. *et al.* (1985) ‘Synergism of Cellulases from Trichoderma reesei in the Degradation of Cellulose’, *Bio/Technology*, 3(8), pp. 722–726. doi: 10.1038/nbt0885-722.

Hershberg, R. and Petrov, D. A. (2008) ‘Selection on Codon Bias’, *Annual Review of Genetics*, 42(1), pp. 287–299. doi: 10.1146/annurev.genet.42.110807.091442.

Hirt, R. P., Alsmark, C. and Embley, T. M. (2015) ‘Lateral gene transfers and the origins of the eukaryote proteome: a view from microbial parasites’, *Current Opinion in Microbiology*, 23, pp. 155–162. doi: 10.1016/j.mib.2014.11.018.

Hittinger, C. T. and Carroll, S. B. (2007) ‘Gene duplication and the adaptive evolution of a classic genetic switch’, *Nature*, 449(7163), pp. 677–681. doi: 10.1038/nature06151.

Ho, H. H. and Hickman, C. J. (1967) ‘Asexual Reproduction and Behavior of Zoospores of Phytophthora Megasperma Var. Sojae’, *Canadian Journal of Botany*, 45(11), pp. 1963–1981. doi: 10.1139/b67-215.

Hoffman, M. *et al.* (2005) ‘Structural analysis of xyloglucans in the primary cell walls of plants in the subclass Asteridae’, *Carbohydrate Research*, 340(11), pp. 1826–1840. doi: 10.1016/j.carres.2005.04.016.

Hönigschmid, P. *et al.* (2018) ‘Evolutionary Interplay between Symbiotic Relationships and Patterns of Signal Peptide Gain and Loss’, *Genome Biology and Evolution*, 10(3), pp. 928–938. doi: 10.1093/gbe/evy049.

van den Hoogen, J. and Govers, F. (2018) ‘Attempts to implement CRISPR/Cas9 for genome editing in the oomycete <em>Phytophthora infestans</em>’, *bioRxiv*, p. 274829. doi: 10.1101/274829.

Horton, P. *et al.* (2007) 'WoLF PSORT: protein localization predictor', *Nucleic Acids Research*, 35(suppl_2), pp. W585–W587. doi: 10.1093/nar/gkm259.

Hosokawa, M. *et al.* (2015) 'Droplet-based microfluidics for high-throughput screening of a metagenomic library for isolation of microbial enzymes', *Biosensors and Bioelectronics*, 67, pp. 379–385. doi: 10.1016/j.bios.2014.08.059.

Hsieh, Y. S. Y. and Harris, P. J. (2009) 'Xyloglucans of Monocotyledons Have Diverse Structures', *Molecular Plant*, 2(5), pp. 943–965. doi: 10.1093/mp/ssp061.

Huber, D. *et al.* (2005) 'Use of Thioredoxin as a Reporter To Identify a Subset of Escherichia coli Signal Sequences That Promote Signal Recognition Particle-Dependent Translocation', *Journal of Bacteriology*, 187(9), pp. 2983–2991. doi: 10.1128/JB.187.9.2983-2991.2005.

Hudson, A. O. *et al.* (2006) 'An ll-Diaminopimelate Aminotransferase Defines a Novel Variant of the Lysine Biosynthesis Pathway in Plants', *Plant Physiology*, 140(1), pp. 292–301. doi: 10.1104/pp.105.072629.

Hudspeth, D. S. S., Nadler, S. A. and Hudspeth, M. E. S. (2000) 'A COX2 molecular phylogeny of the Peronosporomycetes', *Mycologia*, 92(4), pp. 674–684. doi: 10.1080/00275514.2000.12061208.

Hughes, A. L. (1994) 'The evolution of functionally novel proteins after gene duplication', *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 256(1346), pp. 119–124. doi: 10.1098/rspb.1994.0058.

Hughes, M. K. and Hughes, A. L. (1993) 'Evolution of duplicate genes in a tetraploid animal, Xenopus laevis.', *Molecular Biology and Evolution*, 10(6), pp. 1360–1369. doi: 10.1093/oxfordjournals.molbev.a040080.

Husnik, F. and McCutcheon, J. P. (2016) 'Repeated replacement of an intrabacterial symbiont in the tripartite nested mealybug symbiosis', *Proceedings* *of the National Academy of Sciences*, 113(37), pp. E5416–E5424. doi: 10.1073/pnas.1603910113.

Husnik, F. and McCutcheon, J. P. (2018) 'Functional horizontal gene transfer from bacteria to eukaryotes', *Nature Reviews Microbiology*, 16(2), pp. 67–79. doi: 10.1038/nrmicro.2017.137.

Hyams, G. *et al.* (2018) 'CRISPys: Optimal sgRNA Design for Editing Multiple Members of a Gene Family Using the CRISPR System', *Journal of Molecular Biology*, 430(15), pp. 2184–2195. doi: 10.1016/j.jmb.2018.03.019.

Idiris, A. *et al.* (2010) 'Engineering of protein secretion in yeast: strategies and impact on protein production', *Applied Microbiology and Biotechnology*, 86(2), pp. 403–417. doi: 10.1007/s00253-010-2447-0.

Ilmen, M. (1997) 'Molecular mechanisms of glucose repression in the filamentous fungus Trichoderma reesei: Dissertation'. Available at: https://cris.vtt.fi/en/publications/molecular-mechanisms-of-glucose-repression-in-the-filamentous-fun

Iwai, H. *et al.* (2002) 'A pectin glucuronyltransferase gene is essential for intercellular attachment in the plant meristem', *Proceedings of the National Academy of Sciences*, 99(25), pp. 16319–16324. doi: 10.1073/pnas.252530499.

Jacquier, H. *et al.* (2013) 'Capturing the mutational landscape of the beta-lactamase TEM-1', *Proceedings of the National Academy of Sciences*, 110(32), pp. 13067–13072. doi: 10.1073/pnas.1215206110.

Jain, R. *et al.* (2003) 'Horizontal Gene Transfer Accelerates Genome Innovation and Evolution', *Molecular Biology and Evolution*, 20(10), pp. 1598–1602. doi: 10.1093/molbev/msg154.

Jain, R., Rivera, M. C. and Lake, J. A. (1999) 'Horizontal gene transfer among genomes: The complexity hypothesis', *Proceedings of the National Academy of Sciences*, 96(7), pp. 3801–3806. doi: 10.1073/pnas.96.7.3801.

James, T. Y. and Berbee, M. L. (2012) ‘No jacket required – new fungal lineage defies dress code’, *BioEssays*, 34(2), pp. 94–102. doi: 10.1002/bies.201100110.

Janouškovec, J. *et al.* (2010) ‘A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids’, *Proceedings of the National Academy of Sciences*, 107(24), pp. 10949–10954. doi: 10.1073/pnas.1003335107.

Jensen, R. A. (1976) ‘Enzyme recruitment in evolution of new function’, *Annual Review of Microbiology*, 30, pp. 409–425. doi: 10.1146/annurev.mi.30.100176.002205.

Jiang, R. H. Y. and Tyler, B. M. (2011) ‘Mechanisms and Evolution of Virulence in Oomycetes’, *Annual Review of Phytopathology*, 50(1), pp. 295–318. doi: 10.1146/annurev-phyto-081211-172912.

Jones, J. D. G. and Dangl, J. L. (2006) ‘The plant immune system’, *Nature*, 444(7117), pp. 323–329. doi: 10.1038/nature05286.

Jones, M. D. M. *et al.* (2011) ‘Validation and justification of the phylum name Cryptomycota phyl. nov.’, *IMA Fungus*, 2(2), pp. 173–175. doi: 10.5598/imafungus.2011.02.02.08.

Kabbage, M., Yarden, O. and Dickman, M. B. (2015) ‘Pathogenic attributes of Sclerotinia sclerotiorum: Switching from a biotrophic to necrotrophic lifestyle’, *Plant Science*, 233, pp. 53–60. doi: 10.1016/j.plantsci.2014.12.018.

Kaiser, C. A. *et al.* (1987) ‘Many random sequences functionally replace the secretion signal sequence of yeast invertase’, *Science*, 235(4786), pp. 312–317. doi: 10.1126/science.3541205.

Kalyaanamoorthy, S. *et al*. (2017) ‘ModelFinder: Fast model selection for accurate phylogenetic estimates’, *Nature Methods*, 14, pp. 587-589. doi: 10.1038/nmeth.4285.

Kamoun, S. (2003) ‘Molecular Genetics of Pathogenic Oomycetes’, *Eukaryotic Cell*, 2(2), p. 191. doi: 10.1128/EC.2.2.191-199.2003.

Kamoun, S. (2006) ‘A Catalogue of the Effector Secretome of Plant Pathogenic Oomycetes’, *Annual Review of Phytopathology*, 44(1), pp. 41–60. doi: 10.1146/annurev.phyto.44.070505.143436.

Kamoun, S. *et al.* (2015) ‘The Top 10 oomycete pathogens in molecular plant pathology’, *Molecular Plant Pathology*, 16(4), pp. 413–434. doi: 10.1111/mpp.12190.

Karathia, H. *et al.* (2011) ‘Saccharomyces cerevisiae as a model organism: a comparative study’, *PloS one*, 6(2), pp. e16015–e16015. doi: 10.1371/journal.pone.0016015.

Keasling, J. D. (2008) ‘Synthetic Biology for Synthetic Chemistry’, *ACS Chemical Biology*, 3(1), pp. 64–76. doi: 10.1021/cb7002434.

Keeling, P. J. (2009) ‘Chromalveolates and the Evolution of Plastids by Secondary Endosymbiosis1’, *Journal of Eukaryotic Microbiology*, 56(1), pp. 1–8. doi: 10.1111/j.1550-7408.2008.00371.x.

Keeling, P. J. and Palmer, J. D. (2008) ‘Horizontal gene transfer in eukaryotic evolution’, *Nature Reviews Genetics*, 9(8), pp. 605–618. doi: 10.1038/nrg2386.

Keenan, R. J. *et al.* (1998) ‘Crystal Structure of the Signal Sequence Binding Subunit of the Signal Recognition Particle’, *Cell*, 94(2), pp. 181–191. doi: 10.1016/S0092-8674(00)81418-X.

Keenan, R. J. *et al.* (2001) ‘The signal recognition particle’, *Annual Review of Biochemistry*, 70, pp. 755–775. doi: 10.1146/annurev.biochem.70.1.755.

Kelley, B. S. *et al.* (2010) ‘A secreted effector protein (SNE1) from Phytophthora infestans is a broadly acting suppressor of programmed cell death’, *The Plant Journal*, 62(3), pp. 357–366. doi: 10.1111/j.1365-313X.2010.04160.x.

Kemen, E. *et al.* (2011) ‘Gene Gain and Loss during Evolution of Obligate Parasitism in the White Rust Pathogen of Arabidopsis thaliana’, *PLOS Biology*, 9(7), p. e1001094. doi: 10.1371/journal.pbio.1001094.

Kemen, E. and Jones, J. D. G. (2012) ‘Obligate biotroph parasitism: can we link genomes to lifestyles?’, *Trends in Plant Science*, 17(8), pp. 448–457. doi: 10.1016/j.tplants.2012.04.005.

Kidwell, M. G. and Lisch, D. R. (2002) ‘Transposable Elements as Sources of Genomic Variation’, *Mobile DNA II*, pp. 59–90. doi: 10.1128/9781555817954.ch5.

Koshland, D. E. (1953) ‘Stereochemistry and the Mechanism of Enzymatic Reactions’, *Biological Reviews*, 28(4), pp. 416–436. doi: 10.1111/j.1469-185X.1953.tb01386.x.

Krings, M., Taylor, T. N. and Dotzler, N. (2011) ‘The fossil record of the Peronosporomycetes (Oomycota)’, *Mycologia*, 103(3), pp. 445–457. doi: 10.3852/10-278.

Krumpe, L. R. *et al.* (2007) ‘Trinucleotide cassettes increase diversity of T7 phage-displayed peptide library’, *BMC Biotechnology*, 7(1), p. 65. doi: 10.1186/1472-6750-7-65.

Kubicek, C. P. (2013) ‘Systems biological approaches towards understanding cellulase production by Trichoderma reesei’, *Journal of Biotechnology*, 163(2), pp. 133–142. doi: 10.1016/j.jbiotec.2012.05.020.

Kumar, N. *et al.* (2012) ‘Efficient subtraction of insect rRNA prior to transcriptome analysis of Wolbachia-Drosophila lateral gene transfer’, *BMC Research Notes*, 5(1), p. 230. doi: 10.1186/1756-0500-5-230.

Lai, M.-W. and Liou, R.-F. (2018) ‘Two genes encoding GH10 xylanases are essential for the virulence of the oomycete plant pathogen Phytophthora parasitica’, *Current Genetics*, 64(4), pp. 931–943. doi: 10.1007/s00294-018-0814-z.

Lamour, K. H. *et al.* (2012) 'Genome Sequencing and Mapping Reveal Loss of Heterozygosity as a Mechanism for Rapid Adaptation in the Vegetable Pathogen Phytophthora capsici', *Molecular Plant-Microbe Interactions®*, 25(10), pp. 1350–1360. doi: 10.1094/MPMI-02-12-0028-R.

Lamour, K. and Kamoun, S. (2009) *Oomycete Genetics and Genomics: Diversity, Interactions and Research Tools*. John Wiley & Sons.

Langston, J. A. *et al.* (2011) 'Oxidoreductive Cellulose Depolymerization by the Enzymes Cellobiose Dehydrogenase and Glycoside Hydrolase 61', *Applied and Environmental Microbiology*, 77(19), pp. 7007–7015. doi: 10.1128/AEM.05815-11.

Lantz, S. E. *et al.* (2010) 'Hypocrea jecorina CEL6A protein engineering', *Biotechnology for Biofuels*, 3, p. 20. doi: 10.1186/1754-6834-3-20.

Larkin, M. A. *et al.* (2007) 'Clustal W and Clustal X version 2.0', *Bioinformatics*, 23(21), pp. 2947–2948. doi: 10.1093/bioinformatics/btm404.

Latijnhouwers, M. *et al.* (2004) 'A Gα subunit controls zoospore motility and virulence in the potato late blight pathogen Phytophthora infestans', *Molecular Microbiology*, 51(4), pp. 925–936. doi: 10.1046/j.1365-2958.2003.03893.x.

Latijnhouwers, M., de Wit, P. J. G. M. and Govers, F. (2003) 'Oomycetes and fungi: similar weaponry to attack plants', *Trends in Microbiology*, 11(10), pp. 462–469. doi: 10.1016/j.tim.2003.08.002.

Lawrence, J. G. and Ochman, H. (1997) 'Amelioration of Bacterial Genomes: Rates of Change and Exchange', *Journal of Molecular Evolution*, 44(4), pp. 383–397. doi: 10.1007/PL00006158.

Lawrence, J. G. and Ochman, H. (1998) 'Molecular archaeology of the Escherichia coli genome', *Proceedings of the National Academy of Sciences*, 95(16), pp. 9413–9417. doi: 10.1073/pnas.95.16.9413.

Le Hir, H., Nott, A. and Moore, M. J. (2003) ‘How introns influence and enhance eukaryotic gene expression’, *Trends in Biochemical Sciences*, 28(4), pp. 215–220. doi: 10.1016/S0968-0004(03)00052-5.

Leclerc, M. C., Guillot, J. and Deville, M. (2000) ‘Taxonomic and phylogenetic analysis of Saprolegniaceae (Oomycetes) inferred from LSU rDNA and ITS sequence comparisons’, *Antonie van Leeuwenhoek*, 77(4), pp. 369–377. doi: 10.1023/A:1002601211295.

Leclercq, S. *et al.* (2016) ‘Birth of a W sex chromosome by horizontal transfer of Wolbachia bacterial symbiont genome’, *Proceedings of the National Academy of Sciences*, 113(52), pp. 15036–15041. doi: 10.1073/pnas.1608979113.

Leonard, G. *et al.* (2018) ‘Comparative genomic analysis of the “pseudofungus” Hyphochytrium catenoides’, *Open Biology*, 8(1), p. 170184. doi: 10.1098/rsob.170184.

Lerouxel, O. *et al.* (2002) ‘Rapid Structural Phenotyping of Plant Cell Wall Mutants by Enzymatic Oligosaccharide Fingerprinting’, *Plant Physiology*, 130(4), pp. 1754–1763. doi: 10.1104/pp.011965.

Lerouxel, O. *et al.* (2006) ‘Biosynthesis of plant cell wall polysaccharides — a complex process’, *Current Opinion in Plant Biology*, 9(6), pp. 621–630. doi: 10.1016/j.pbi.2006.09.009.

Lévesque, C. A. *et al.* (2010) ‘Genome sequence of the necrotrophic plant pathogen Pythium ultimum reveals original pathogenicity mechanisms and effector repertoire’, *Genome Biology*, 11(7), p. R73. doi: 10.1186/gb-2010-11-7-r73.

Li, Z. *et al.* (2014) ‘A C-Terminal Proline-Rich Sequence Simultaneously Broadens the Optimal Temperature and pH Ranges and Improves the Catalytic Efficiency of Glycosyl Hydrolase Family 10 Ruminal Xylanases’, *Applied and Environmental Microbiology*, 80(11), pp. 3426–3432. doi: 10.1128/AEM.00016-14.

Lind, P. A. *et al.* (2010) 'Compensatory gene amplification restores fitness after inter-species gene replacements', *Molecular Microbiology*, 75(5), pp. 1078–1089. doi: 10.1111/j.1365-2958.2009.07030.x.

Links, M. G. *et al.* (2011) 'De novo sequence assembly of Albugo candida reveals a small genome relative to other biotrophic oomycetes', *BMC Genomics*, 12(1), p. 503. doi: 10.1186/1471-2164-12-503.

Liu, T. *et al.* (2014) 'Unconventionally secreted effectors of two filamentous pathogens target plant salicylate biosynthesis', *Nature Communications*, 5(1), p. 4686. doi: 10.1038/ncomms5686.

Llorente, B. *et al.* (2016) 'Selective pressure against horizontally acquired prokaryotic genes as a driving force of plastid evolution', *Scientific Reports*, 6(1), p. 19036. doi: 10.1038/srep19036.

Lombard, V. *et al.* (2014) 'The carbohydrate-active enzymes database (CAZy) in 2013', *Nucleic Acids Research*, 42(D1), pp. D490–D495. doi: 10.1093/nar/gkt1178.

Long, M. *et al.* (2003) 'The origin of new genes: glimpses from the young and old', *Nature Reviews Genetics*, 4(11), pp. 865–875. doi: 10.1038/nrg1204.

Lopez, J. V. and Yuhki, N. (1994) 'Numt, a Recent Transfer and Tandem Amplification of Mitochondrial DNA to the Nuclear Genome of the Domestic Cat', p. 17.

Ma, Z. *et al.* (2015) 'A Phytophthora sojae Glycoside Hydrolase 12 Protein Is a Major Virulence Factor during Soybean Infection and Is Recognized as a PAMP', *The Plant Cell*, 27(7), pp. 2057–2072. doi: 10.1105/tpc.15.00390.

Ma, Z. *et al.* (2017) 'A paralogous decoy protects Phytophthora sojae apoplastic effector PsXEG1 from a host inhibitor', *Science*, 355(6326), pp. 710–714. doi: 10.1126/science.aai7919.

Madeira, F. *et al.* (2019) ‘The EMBL-EBI search and sequence analysis tools APIs in 2019’, *Nucleic Acids Research*, 47(W1), pp. W636–W641. doi: 10.1093/nar/gkz268.

Madson, M. *et al.* (2003) ‘The MUR3 gene of Arabidopsis encodes a xyloglucan galactosyltransferase that is evolutionarily related to animal exostosins’, *The Plant cell*, 15(7), pp. 1662–1670. doi: 10.1105/tpc.009837.

Mahelka, V. *et al.* (2017) ‘Multiple horizontal transfers of nuclear ribosomal genes between phylogenetically distinct grass lineages’, *Proceedings of the National Academy of Sciences*, 114(7), pp. 1726–1731. doi: 10.1073/pnas.1613375114.

Makarova, K. S. *et al.* (2006) ‘A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action’, *Biology Direct*, 1(1), p. 7. doi: 10.1186/1745-6150-1-7.

Makowski, L. and Soares, A. (2003) ‘Estimating the diversity of peptide populations from limited sequence data’, *Bioinformatics*, 19(4), pp. 483–489. doi: 10.1093/bioinformatics/btg013.

Marcet-Houben, M. and Gabaldón, T. (2010) ‘Acquisition of prokaryotic genes by fungal genomes’, *Trends in Genetics*, 26(1), pp. 5–8. doi: 10.1016/j.tig.2009.11.007.

Marry, M. *et al.* (2003) ‘Structural characterization of chemically and enzymatically derived standard oligosaccharides isolated from partially purified tamarind xyloglucan’, *Carbohydrate Polymers*, 51(3), pp. 347–356. doi: 10.1016/S0144-8617(02)00189-3.

Martinez-Fleites, C. *et al.* (2006) ‘Crystal Structures of Clostridium thermocellum Xyloglucanase, XGH74A, Reveal the Structural Basis for Xyloglucan Recognition and Degradation’, *Journal of Biological Chemistry*, 281(34), pp. 24922–24933. doi: 10.1074/jbc.M603583200.

Matari, N. H. and Blair, J. E. (2014) ‘A multilocus timescale for oomycete evolution estimated under three distinct molecular clock models’, *BMC Evolutionary Biology*, 14(1), p. 101. doi: 10.1186/1471-2148-14-101.

MAYNARD SMITH, J. (1970) ‘Natural Selection and the Concept of a Protein Space’, *Nature*, 225(5232), pp. 563–564. doi: 10.1038/225563a0.

McCarter, J. D. and Stephen Withers, G. (1994) ‘Mechanisms of enzymatic glycoside hydrolysis’, *Current Opinion in Structural Biology*, 4(6), pp. 885–892. doi: 10.1016/0959-440X(94)90271-2.

McCarthy, C. G. P. and Fitzpatrick, D. A. (2016) ‘Systematic Search for Evidence of Interdomain Horizontal Gene Transfer from Prokaryotes to Oomycete Lineages’, *mSphere*, 1(5). doi: 10.1128/mSphere.00195-16.

McCarthy, C. G. P. and Fitzpatrick, D. A. (2017) ‘Phylogenomic Reconstruction of the Oomycete Phylogeny Derived from 37 Genomes’, *mSphere*, 2(2). doi: 10.1128/mSphere.00095-17.

McGowan, J. and Fitzpatrick, D. A. (2017) ‘Genomic, Network, and Phylogenetic Analysis of the Oomycete Effector Arsenal’, *mSphere*, 2(6). doi: 10.1128/mSphere.00408-17.

McLeod, A., Smart, C. D. and Fry, W. E. (2003) ‘Characterization of 1,3-β-glucanase and 1,3;1,4-β-glucanase genes from Phytophthora infestans’, *Fungal Genetics and Biology*, 38(2), pp. 250–263. doi: 10.1016/S1087-1845(02)00523-6.

McNeil, M. *et al.* (1984) ‘Structure and Function of the Primary Cell Walls of Plants’, *Annual Review of Biochemistry*, 53(1), pp. 625–663. doi: 10.1146/annurev.bi.53.070184.003205.

Meadows, R. (2011) ‘Why Biotrophs Can’t Live Alone’, *PLOS Biology*, 9(7), p. e1001097. doi: 10.1371/journal.pbio.1001097.

Meini, M.-R. *et al.* (2015) 'Quantitative Description of a Protein Fitness Landscape Based on Molecular Features', *Molecular Biology and Evolution*, 32(7), pp. 1774–1787. doi: 10.1093/molbev/msv059.

Mellerowicz, E. J., Immerzeel, P. and Hayashi, T. (2008) 'Xyloglucan: The Molecular Muscle of Trees', *Annals of Botany*, 102(5), pp. 659–665. doi: 10.1093/aob/mcn170.

Mendgen, K. and Hahn, M. (2002) 'Plant infection and the establishment of fungal biotrophy', *Trends in Plant Science*, 7(8), pp. 352–356. doi: 10.1016/S1360-1385(02)02297-5.

Metcalf, J. A. *et al.* (2014) 'Antibacterial gene transfer across the tree of life', *eLife*. Edited by W. van der Donk, 3, p. e04266. doi: 10.7554/eLife.04266.

Micali, C. *et al.* (2008) 'The Powdery Mildew Disease of Arabidopsis: A Paradigm for the Interaction between Plants and Biotrophic Fungi', *The arabidopsis book*. 2008/10/02 edn, 6, pp. e0115–e0115. doi: 10.1199/tab.0115.

Miller, G. L. (1959) 'Use of Dinitrosalicylic Acid Reagent for Determination of Reducing Sugar', *Analytical Chemistry*, 31(3), pp. 426–428. doi: 10.1021/ac60147a030.

Milstein, C. *et al.* (1972) 'A Possible Precursor of Immunoglobulin Light Chains', *Nature New Biology*, 239(91), pp. 117–120. doi: 10.1038/newbio239117a0.

Minh, B. Q., Schmidt, Heiko A, *et al.* (2020) 'Corrigendum to: IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era', *Molecular Biology and Evolution*, 37(8), p. 2461. doi: 10.1093/molbev/msaa131.

Misner, I. *et al.* (2015) 'The Secreted Proteins of Achlya hypogyna and Thraustotheca clavata Identify the Ancestral Oomycete Secretome and Reveal Gene Acquisitions by Horizontal Gene Transfer', *Genome Biology and Evolution*, 7(1), pp. 120–135. doi: 10.1093/gbe/evu276.

Miyazaki, T. *et al.* (2004) 'α-Aminoadipate aminotransferase from an extremely thermophilic bacterium, Thermus thermophilus', *Microbiology,* 150(7), pp. 2327–2334. doi: 10.1099/mic.0.27037-0.

Mock, T. *et al.* (2017) 'Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus', *Nature*, 541(7638), pp. 536–540. doi: 10.1038/nature20803.

Mojica, F. J. M. *et al.* (2005) 'Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements', *Journal of Molecular Evolution*, 60(2), pp. 174–182. doi: 10.1007/s00239-004-0046-3.

Mojica, F. J. M. *et al.* (2009) 'Short motif sequences determine the targets of the prokaryotic CRISPR defence system', *Microbiology,* 155(3), pp. 733–740. doi: 10.1099/mic.0.023960-0.

Money, N. P., Davis, C. M. and Ravishankar, J. P. (2004) 'Biomechanical evidence for convergent evolution of the invasive growth process among fungi and oomycete water molds', *Fungal Genetics and Biology*, 41(9), pp. 872–876. doi: 10.1016/j.fgb.2004.06.001.

Monier, A. *et al.* (2017) 'Host-derived viral transporter protein for nitrogen uptake in infected marine phytoplankton', *Proceedings of the National Academy of Sciences*, 114(36), pp. E7489–E7498. doi: 10.1073/pnas.1708097114.

Moreira, D. and López-García, P. (2002) 'The molecular ecology of microbial eukaryotes unveils a hidden world', *Trends in Microbiology*, 10(1), pp. 31–38. doi: 10.1016/S0966-842X(01)02257-0.

Morris, P. F., Bone, E. and Tyler, B. M. (1998) 'Chemotropic and Contact Responses of Phytophthora sojae Hyphae to Soybean Isoflavonoids and Artificial Substrates', *Plant Physiology*, 117(4), pp. 1171–1178. doi: 10.1104/pp.117.4.1171.

Morris, P. F. and Ward, E. W. B. (1992) ‘Chemoattraction of zoospores of the soybean pathogen, Phytophthora sojae, by isoflavones’, *Physiological and Molecular Plant Pathology*, 40(1), pp. 17–22. doi: 10.1016/0885-5765(92)90067-6.

Mower, J. P. *et al.* (2004) ‘Gene transfer from parasitic to host plants’, *Nature*, 432(7014), pp. 165–166. doi: 10.1038/432165b.

Mueller, O. *et al.* (2008) ‘The secretome of the maize pathogen Ustilago maydis’, *Fungal Genetics and Biology*, 45, pp. S63–S70. doi: 10.1016/j.fgb.2008.03.012.

Mueller, S. C. and Brown, R. M. (1980) ‘Evidence for an intramembrane component associated with a cellulose microfibril-synthesizing complex in higher plants.’, *Journal of Cell Biology*, 84(2), pp. 315–326. doi: 10.1083/jcb.84.2.315.

Mueller, S. C., Brown, R. M. and Scott, T. K. (1976) ‘Cellulosic Microfibrils: Nascent Stages of Synthesis in a Higher Plant Cell’, *Science*, 194(4268), pp. 949–951. doi: 10.1126/science.194.4268.949.

Näsvall, J. *et al.* (2012) ‘Real-time evolution of new genes by innovation, amplification, and divergence’, *Science (New York, N.Y.)*, 338(6105), pp. 384–387. doi: 10.1126/science.1226521.

Nesmeyanova, M. A. *et al.* (1991) ‘Secretion of the overproduced periplasmic PhoA protein into the medium and accumulation of its precursor in phoA-transformed Escherichia coli strains: involvement of outer membrane vesicles’, *World Journal of Microbiology and Biotechnology*, 7(3), pp. 394–406. doi: 10.1007/BF00329408.

Nesmeyanova, M. A. *et al.* (1997) ‘Positively charged lysine at the N-terminus of the signal peptide of the Escherichia coli alkaline phosphatase provides the secretion efficiency and is involved in the interaction with anionic phospholipids’, *FEBS Letters*, 403(2), pp. 203–207. doi: 10.1016/S0014-5793(97)00052-5.

Newell, S. Y., Cefalu, R. and Fell, J. W. (1977) ‘Myzocytium, Haptoglossa, and Gonimochaete (Fungi) in Littoral Marine Nematodes’, *Bulletin of Marine Science*, 27(2), pp. 177–207.

Nidetzky, B. and Claeyssens, M. (1994) ‘Specific quantification of trichoderma reesei cellulases in reconstituted mixtures and its application to cellulase–cellulose binding studies’, *Biotechnology and Bioengineering*, 44(8), pp. 961–966. doi: 10.1002/bit.260440812.

Nielsen, H. and Krogh, A. (1998) ‘Prediction of Signal Peptides and Signal Anchors by a Hidden Markov model’, p. 9.

Nielsen, R. I. (2002) ‘Microbial xyloglucan endotransglycosylase (XET)’. Available at: https://patents.google.com/patent/US6448056B1/en

Nilsson, I. *et al.* (2015) ‘The code for directing proteins for translocation across ER membrane: SRP cotranslationally recognizes specific features of a signal sequence’, *Journal of molecular biology*, 427(6 0 0), pp. 1191–1201. doi: 10.1016/j.jmb.2014.06.014.

Nishikubo, N. *et al.* (2007) ‘Xyloglucan Endo-transglycosylase (XET) Functions in Gelatinous Layers of Tension Wood Fibers in Poplar—A Glimpse into the Mechanism of the Balancing Act of Trees’, *Plant and Cell Physiology*, 48(6), pp. 843–855. doi: 10.1093/pcp/pcm055.

Nyathi, Y., Wilkinson, B. M. and Pool, M. R. (2013) ‘Co-translational targeting and translocation of proteins to the endoplasmic reticulum’, *Biochimica Et Biophysica Acta*, 1833(11), pp. 2392–2402. doi: 10.1016/j.bbamcr.2013.02.021.

Obel, N. *et al.* (2009) ‘Microanalysis of plant cell wall polysaccharides’, *Molecular Plant*, 2(5), pp. 922–932. doi: 10.1093/mp/ssp046.

Ochman, H., Lawrence, J. G. and Groisman, E. A. (2000) ‘Lateral gene transfer and the nature of bacterial innovation’, *Nature*, 405(6784), pp. 299–304. doi: 10.1038/35012500.

Ohno, S. (2013) *Evolution by Gene Duplication*. Springer Science & Business Media.

Okada, H. *et al.* (2000) ‘Identification of active site carboxylic residues in Trichoderma reesei endoglucanase Cel12A by site-directed mutagenesis’, *Journal of Molecular Catalysis B: Enzymatic*, 10(1), pp. 249–255. doi: 10.1016/S1381-1177(00)00137-5.

Oliver, R. P. and Ipcho, S. V. S. (2004) ‘Arabidopsis pathology breathes new life into the necrotrophs-vs.-biotrophs classification of fungal pathogens’, *Molecular Plant Pathology*, 5(4), pp. 347–352. doi: 10.1111/j.1364-3703.2004.00228.x.

Park, C. and Zhang, J. (2012) ‘High expression hampers horizontal gene transfer’, *Genome Biology and Evolution*, 4(4), pp. 523–532. doi: 10.1093/gbe/evs030.

Parker, B. C., Preston, R. D. and Fogg, G. E. (1963) ‘Studies of the structure and chemical composition of the cell walls of Vaucheriaceae and Saprolegniaceae’, *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 158(973), pp. 435–445. doi: 10.1098/rspb.1963.0056.

Peisajovich, S. G. and Tawfik, D. S. (2007) ‘Protein engineers turned evolutionists’, *Nature Methods*, 4(12), pp. 991–994. doi: 10.1038/nmeth1207-991.

Peña, M. J. *et al.* (2008) ‘Moss and liverwort xyloglucans contain galacturonic acid and are structurally distinct from the xyloglucans synthesized by hornworts and vascular plants’, *Glycobiology*, 18(11), pp. 891–904. doi: 10.1093/glycob/cwn078.

Peng, D. *et al.* (2015) ‘CRISPR-Cas9-Mediated Single-Gene and Gene Family Disruption in Trypanosoma cruzi’, *mBio*, 6(1). doi: 10.1128/mBio.02097-14.

Piskurek, O. and Okada, N. (2007) ‘Poxviruses as possible vectors for horizontal transfer of retroposons from reptiles to mammals’, *Proceedings of the National*

Academy of Sciences, 104(29), pp. 12046–12051. doi: 10.1073/pnas.0700531104.

Plath, K. *et al.* (1998) ‘Signal sequence recognition in posttranslational protein transport across the yeast ER membrane’, *Cell*, 94(6), pp. 795–807. doi: 10.1016/s0092-8674(00)81738-9.

Poon, A. and Chao, L. (2005) ‘The Rate of Compensatory Mutation in the DNA Bacteriophage φX174’, *Genetics*, 170(3), pp. 989–999. doi: 10.1534/genetics.104.039438.

Porter, B. W., Yuen, C. Y. L. and Christopher, D. A. (2015) ‘Dual protein trafficking to secretory and non-secretory cell compartments: Clear or double vision?’, *Plant Science*, 234, pp. 174–179. doi: 10.1016/j.plantsci.2015.02.013.

Puziss, J. W., Harvey, R. J. and Bassford, P. J. (1992) ‘Alterations in the hydrophilic segment of the maltose-binding protein (MBP) signal peptide that affect either export or translation of MBP.’, *Journal of Bacteriology*, 174(20), pp. 6488–6497. doi: 10.1128/jb.174.20.6488-6497.1992.

Qian, W. *et al.* (2012) ‘Balanced Codon Usage Optimizes Eukaryotic Translational Efficiency’, *PLoS Genetics*. Edited by H. S. Malik, 8(3), p. e1002603. doi: 10.1371/journal.pgen.1002603.

Quinlan, R. J. *et al.* (2011) ‘Insights into the oxidative degradation of cellulose by a copper metalloenzyme that exploits biomass components’, *Proceedings of the National Academy of Sciences*, 108(37), pp. 15079–15084. doi: 10.1073/pnas.1105776108.

Quinn, L. *et al.* (2013) ‘Genome-wide sequencing of Phytophthora lateralis reveals genetic variation among isolates from Lawson cypress (Chamaecyparis lawsoniana) in Northern Ireland’, *FEMS Microbiology Letters*, 344(2), pp. 179–185. doi: 10.1111/1574-6968.12179.

Raffaele, S., Win, J., *et al.* (2010) ‘Analyses of genome architecture and gene expression reveal novel candidate virulence factors in the secretome of Phytophthora infestans’, *BMC Genomics*, 11(1), p. 637. doi: 10.1186/1471-2164-11-637.

Raffaele, S., Farrer, R. A., *et al.* (2010) ‘Genome Evolution Following Host Jumps in the Irish Potato Famine Pathogen Lineage’, *Science*, 330(6010), pp. 1540–1543. doi: 10.1126/science.1193070.

Ramirez-Garcés, D. *et al.* (2016) ‘CRN13 candidate effectors from plant and animal eukaryotic pathogens are DNA-binding proteins which trigger host DNA damage response’, *New Phytologist*, 210(2), pp. 602–617. doi: 10.1111/nph.13774.

Raymond, J. A. and Kim, H. J. (2012) ‘Possible Role of Horizontal Gene Transfer in the Colonization of Sea Ice by Algae’, *PLOS ONE*, 7(5), p. e35968. doi: 10.1371/journal.pone.0035968.

Raymond, J. and Blankenship, R. E. (2003) ‘Horizontal gene transfer in eukaryotic algal evolution’, *Proceedings of the National Academy of Sciences*, 100(13), pp. 7419–7420. doi: 10.1073/pnas.1533212100.

Reetz, M. T., Carballeira, J. D. and Vogel, A. (2006) ‘Iterative Saturation Mutagenesis on the Basis of B Factors as a Strategy for Increasing Protein Thermostability’, *Angewandte Chemie International Edition*, 45(46), pp. 7745–7751. doi: 10.1002/anie.200602795.

Reetz, M. T., Kahakeaw, D. and Lohmer, R. (2008) ‘Addressing the Numbers Problem in Directed Evolution’, *ChemBioChem*, 9(11), pp. 1797–1804. doi: 10.1002/cbic.200800298.

Rice, D. W. *et al.* (2013) ‘Horizontal Transfer of Entire Genomes via Mitochondrial Fusion in the Angiosperm Amborella’, *Science*, 342(6165), pp. 1468–1473. doi: 10.1126/science.1246275.

Rice, D. W. and Palmer, J. D. (2006) ‘An exceptional horizontal gene transfer in plastids: gene replacement by a distant bacterial paralog and evidence that haptophyte and cryptophyte plastids are sisters’, *BMC Biology*, 4(1), p. 31. doi: 10.1186/1741-7007-4-31.

Richards, T. A. *et al.* (2006) ‘Evolution of Filamentous Plant Pathogens: Gene Exchange across Eukaryotic Kingdoms’, *Current Biology*, 16(18), pp. 1857–1864. doi: 10.1016/j.cub.2006.07.052.

Richards, T. A. *et al.* (2009) ‘Phylogenomic Analysis Demonstrates a Pattern of Rare and Ancient Horizontal Gene Transfer between Plants and Fungi’, *The Plant Cell*, 21(7), pp. 1897–1911. doi: 10.1105/tpc.109.065805.

Richards, T. A. *et al.* (2011) ‘Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes’, *Proceedings of the National Academy of Sciences*, 108(37), pp. 15258–15263. doi: 10.1073/pnas.1105100108.

Richards, T. A. and Talbot, N. J. (2013) ‘Horizontal gene transfer in osmotrophs: playing with public goods’, *Nature Reviews Microbiology*, 11(10), pp. 720–727. doi: 10.1038/nrmicro3108.

Riederer, M. and Muller, C. (2008) *Annual Plant Reviews, Biology of the Plant Cuticle*. John Wiley & Sons.

Riisberg, I. *et al.* (2009) ‘Seven Gene Phylogeny of Heterokonts’, *Protist*, 160(2), pp. 191–204. doi: 10.1016/j.protis.2008.11.004.

Rizk, S. E. *et al.* (2000) ‘Protein- and pH-dependent binding of nascent pectin and glucuronoarabinoxylan to xyloglucan in pea’, *Planta*, 211(3), pp. 423–429. doi: 10.1007/s004250000303.

Rodenburg, S. Y. A. *et al.* (2019) ‘Metabolic Model of the Phytophthora infestans-Tomato Interaction Reveals Metabolic Switches during Host Colonization’, *mBio*, 10(4). doi: 10.1128/mBio.00454-19.

Rogers, L. M. *et al.* (2000) ‘Requirement for either a host- or pectin-induced pectate lyase for infection of Pisum sativum by Nectria hematococca’, *Proceedings of the National Academy of Sciences*, 97(17), pp. 9813–9818. doi: 10.1073/pnas.160271497.

Romero, P. A. and Arnold, F. H. (2009) ‘Exploring protein fitness landscapes by directed evolution’, *Nature Reviews Molecular Cell Biology*, 10(12), pp. 866–876. doi: 10.1038/nrm2805.

Romero, P. A., Tran, T. M. and Abate, A. R. (2015) ‘Dissecting enzyme function with microfluidic-based deep mutational scanning’, *Proceedings of the National Academy of Sciences*, 112(23), pp. 7159–7164. doi: 10.1073/pnas.1422285112.

Rothe, C. and Lehle, L. (1998) ‘Sorting of invertase signal peptide mutants in yeast dependent and independent on the signal-recognition particle’, *European Journal of Biochemistry*, 252(1), pp. 16–24. doi: 10.1046/j.1432-1327.1998.2520016.x.

Rouvinen, J. *et al.* (1990) ‘Three-dimensional structure of cellobiohydrolase II from Trichoderma reesei’, *Science*, 249(4967), pp. 380–386. doi: 10.1126/science.2377893.

Ruan, L. *et al.* (2017) ‘Cytosolic proteostasis through importing of misfolded proteins into mitochondria’, *Nature*, 543(7645), pp. 443–446. doi: 10.1038/nature21695.

Rubartelli, A. and Sitia, R. (1997) ‘Secretion of Mammalian Proteins that Lack a Signal Sequence’, in Kuchler, K., Rubartelli, A., and Holland, B. (eds) *Unusual Secretory Pathways: From Bacteria to Man*. Berlin, Heidelberg: Springer (Molecular Biology Intelligence Unit), pp. 87–114. doi: 10.1007/978-3-662-22581-3_3.

Ryan, P. and Edwards, C. O. (1995) ‘Systematic Introduction of Proline in a Eukaryotic Signal Sequence Suggests Asymmetry within the Hydrophobic Core’, *Journal of Biological Chemistry*, 270(46), pp. 27876–27879. doi: 10.1074/jbc.270.46.27876.

Sahoo, D. K. *et al.* (2017) ‘A Novel Phytophthora sojae Resistance Rps12 Gene Mapped to a Genomic Region That Contains Several Rps Genes’, *PLOS ONE*, 12(1), p. e0169950. doi: 10.1371/journal.pone.0169950.

Salzberg, S. L. *et al.* (2001) ‘Microbial Genes in the Human Genome: Lateral Transfer or Gene Loss?’, *Science*, 292(5523), pp. 1903–1906. doi: 10.1126/science.1061036.

Salzberg, S. L. (2017) ‘Horizontal gene transfer is not a hallmark of the human genome’, *Genome Biology*, 18(1), p. 85. doi: 10.1186/s13059-017-1214-2.

Sandgren, M. *et al.* (2001) ‘The X-ray crystal structure of the Trichoderma reesei family 12 endoglucanase 3, Cel12A, at 1.9 Å resolution1 1Edited by A. R. Fersht’, *Journal of Molecular Biology*, 308(2), pp. 295–310. doi: 10.1006/jmbi.2001.4583.

Sandgren, M. *et al.* (2003) ‘Comparison of family 12 glycoside hydrolases and recruited substitutions important for thermal stability’, *Protein Science*, 12(4), pp. 848–860. doi: 10.1110/ps.0237703.

Sandgren, M., Ståhlberg, J. and Mitchinson, C. (2005) ‘Structural and biochemical studies of GH family 12 cellulases: improved thermal stability, and ligand complexes’, *Progress in Biophysics and Molecular Biology*, 89(3), pp. 246–291. doi: 10.1016/j.pbiomolbio.2004.11.002.

Sarkisyan, K. S. *et al.* (2016) ‘Local fitness landscape of the green fluorescent protein’, *Nature*, 533(7603), pp. 397–401. doi: 10.1038/nature17995.

Sasaki, A. and Nowak, M. A. (2003) ‘Mutation landscapes’, *Journal of Theoretical Biology*, 224(2), pp. 241–247. doi: 10.1016/S0022-5193(03)00161-9.

Savory, F., Leonard, G. and Richards, T. A. (2015) ‘The Role of Horizontal Gene Transfer in the Evolution of the Oomycetes’, *PLOS Pathogens*, 11(5), p. e1004805. doi: 10.1371/journal.ppat.1004805.

Savory, Fiona R *et al.* (2018) ‘Ancestral Function and Diversification of a Horizontally Acquired Oomycete Carboxylic Acid Transporter’, *Molecular Biology and Evolution*. Edited by K. Crandall, 35(8), pp. 1887–1900. doi: 10.1093/molbev/msy082.

Schindler, T. M. (1998) ‘The new view of the primary cell wall’, *Zeitschrift für Pflanzenernährung und Bodenkunde*, 161(5), pp. 499–508. doi: 10.1002/jpln.1998.3581610503.

Schou, C. *et al.* (1993) ‘Stereochemistry, specificity and kinetics of the hydrolysis of reduced cellodextrins by nine cellulases’, *European Journal of Biochemistry*, 217(3), pp. 947–953. doi: 10.1111/j.1432-1033.1993.tb18325.x.

Schröder, M. and Friedl, P. (1997) ‘Overexpression of recombinant human antithrombin III in Chinese hamster ovary cells results in malformation and decreased secretion of recombinant protein’, *Biotechnology and Bioengineering*, 53(6), pp. 547–559. doi: 10.1002/(SICI)1097-0290(19970320)53:6<547::AID-BIT2>3.0.CO;2-M.

Schwessinger, B. and Ronald, P. C. (2012) ‘Plant Innate Immunity: Perception of Conserved Microbial Signatures’, *Annual Review of Plant Biology*, 63(1), pp. 451–482. doi: 10.1146/annurev-arplant-042811-105518.

Secq, M.-P. O.-L. *et al.* (2006) ‘Complete mitochondrial genomes of the three brown algae (Heterokonta: Phaeophyceae) Dictyota dichotoma, Fucus vesiculosus and Desmarestia viridis’, *Current Genetics*, 49(1), pp. 47–58. doi: 10.1007/s00294-005-0031-4.

Seidl, M. F. *et al.* (2012) ‘Reconstruction of Oomycete Genome Evolution Identifies Differences in Evolutionary Trajectories Leading to Present-Day Large

Gene Families', *Genome Biology and Evolution*, 4(3), pp. 199–211. doi: 10.1093/gbe/evs003.

Sharp, P. M., Emery, L. R. and Zeng, K. (2010) 'Forces that influence the evolution of codon bias', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1544), pp. 1203–1212. doi: 10.1098/rstb.2009.0305.

Shimodaira, H. (2002) 'An Approximately Unbiased Test of Phylogenetic Tree Selection', *Systematic Biology*, 51(3), pp. 492–508. doi: 10.1080/10635150290069913.

Silva, J. C. and Kidwell, M. G. (2000) 'Horizontal Transfer and Selection in the Evolution of P Elements', *Molecular Biology and Evolution*, 17(10), pp. 1542–1557. doi: 10.1093/oxfordjournals.molbev.a026253.

Silve, S. *et al.* (1987) 'The yeast acid phosphatase can enter the secretory pathway without its N-terminal signal sequence.', *Molecular and Cellular Biology*, 7(9), pp. 3306–3314. doi: 10.1128/MCB.7.9.3306.

Simão, F. A. *et al.* (2015) 'BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs', *Bioinformatics*, 31(19), pp. 3210–3212. doi: 10.1093/bioinformatics/btv351.

Simmons, T. J. *et al.* (2016) 'Folding of xylan onto cellulose fibrils in plant cell walls revealed by solid-state NMR', *Nature Communications*, 7(1), p. 13902. doi: 10.1038/ncomms13902.

Sinnott, M. L. (1990) 'Catalytic mechanism of enzymic glycosyl transfer', *Chemical Reviews*, 90(7), pp. 1171–1202. doi: 10.1021/cr00105a006.

Soanes, D. and Richards, T. A. (2014) 'Horizontal Gene Transfer in Eukaryotic Plant Pathogens', *Annual Review of Phytopathology*, 52(1), pp. 583–614. doi: 10.1146/annurev-phyto-102313-050127.

Sørensen, M. A., Kurland, C. G. and Pedersen, S. (1989) ‘Codon usage determines translation rate in Escherichia coli’, *Journal of Molecular Biology*, 207(2), pp. 365–377. doi: 10.1016/0022-2836(89)90260-X.

Sperschneider, J. *et al.* (2015) ‘Advances and Challenges in Computational Prediction of Effectors from Plant Pathogenic Fungi’, *PLOS Pathogens*, 11(5), p. e1004806. doi: 10.1371/journal.ppat.1004806.

Srinivasan, V., Pamula, V. K. and Fair, R. B. (2004) ‘Droplet-based microfluidic lab-on-a-chip for glucose detection’, *Analytica Chimica Acta*, 507(1), pp. 145–150. doi: 10.1016/j.aca.2003.12.030.

Sriprang, R. *et al.* (2006) ‘Improvement of thermostability of fungal xylanase by using site-directed mutagenesis’, *Journal of Biotechnology*, 126(4), pp. 454–462. doi: 10.1016/j.jbiotec.2006.04.031.

Stajich, J. E. *et al.* (2012) ‘FungiDB: an integrated functional genomics database for fungi’, *Nucleic Acids Research*, 40(D1), pp. D675–D681. doi: 10.1093/nar/gkr918.

Sternberg, D. and Mandels, G. R. (1980) ‘Regulation of the cellulolytic system in Trichoderma reesei by sophorose: induction of cellulase and repression of beta-glucosidase.’, *Journal of Bacteriology*, 144(3), pp. 1197–1199.

Stiller, J. W. *et al.* (2009) ‘Are algal genes in nonphotosynthetic protists evidence of historical plastid endosymbioses?’, *BMC Genomics*, 10(1), p. 484. doi: 10.1186/1471-2164-10-484.

Stoltzfus, A. (1999) ‘On the Possibility of Constructive Neutral Evolution’, *Journal of Molecular Evolution*, 49(2), pp. 169–181. doi: 10.1007/PL00006540.

Strullu-Derrien, C. *et al.* (2011) ‘Evidence of parasitic Oomycetes (Peronosporomycetes) infecting the stem cortex of the Carboniferous seed fern Lyginopteris oldhamia’, *Proceedings of the Royal Society B: Biological Sciences*, 278(1706), pp. 675–680. doi: 10.1098/rspb.2010.1603.

Sulzenbacher, G. *et al.* (1999) ‘The Crystal Structure of a 2-Fluorocellotriosyl Complex of the Streptomyces lividans Endoglucanase CelB2 at 1.2 Å Resolution’, *Biochemistry*, 38(15), pp. 4826–4833. doi: 10.1021/bi982648i.

Sumathi, J. C. *et al.* (2006) ‘Molecular Evidence of Fungal Signatures in the Marine Protist Corallochytrium limacisporum and its Implications in the Evolution of Animals and Fungi’, *Protist*, 157(4), pp. 363–376. doi: 10.1016/j.protis.2006.05.003.

Syvanen, M. (1985) ‘Cross-species gene transfer; implications for a new theory of evolution’, *Journal of Theoretical Biology*, 112(2), pp. 333–343. doi: 10.1016/S0022-5193(85)80291-5.

Takahara, M. *et al.* (1985) ‘The ompA signal peptide directed secretion of Staphylococcal nuclease A by Escherichia coli’, *The Journal of Biological Chemistry*, 260(5), pp. 2670–2674.

Talavera, G. and Castresana, J. (2007) ‘Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments’, *Systematic Biology*, 56(4), pp. 564–577. doi: 10.1080/10635150701472164.

Taylor, F. J. R. (1978) ‘Problems in the development of an explicit hypothetical phylogeny of the lower eukaryotes’, *Biosystems*, 10(1), pp. 67–89. doi: 10.1016/0303-2647(78)90031-X.

Theberge, A. B. *et al.* (2010) ‘Microdroplets in Microfluidics: An Evolving Platform for Discoveries in Chemistry and Biology’, *Angewandte Chemie International Edition*, 49(34), pp. 5846–5868. doi: 10.1002/anie.200906653.

Thines, M. (2007) ‘Characterisation and phylogeny of repeated elements giving rise to exceptional length of ITS2 in several downy mildew genera (Peronosporaceae)’, *Fungal Genetics and Biology*, 44(3), pp. 199–207. doi: 10.1016/j.fgb.2006.08.002.

Thines, M. and Kamoun, S. (2010) ‘Oomycete–plant coevolution: recent advances and future prospects’, *Current Opinion in Plant Biology*, 13(4), pp. 427–433. doi: 10.1016/j.pbi.2010.04.001.

Thomas A. Richards Guy Leonard Jeremy G. Wideman (2017) *What Defines the “Kingdom” Fungi? - The Fungal Kingdom - Wiley Online Library*. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1128/9781555819583.ch3

Thomas, C. M. and Nielsen, K. M. (2005) ‘Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria’, *Nature Reviews Microbiology*, 3(9), pp. 711–721. doi: 10.1038/nrmicro1234.

Tomme, P. *et al.* (1996) ‘Cellulose-Binding Domains: Classification and Properties’, in *Enzymatic Degradation of Insoluble Carbohydrates*. American Chemical Society (ACS Symposium Series, 618), pp. 142–163. doi: 10.1021/bk-1995-0618.ch010.

Tonkin, C. J. *et al.* (2008) ‘Evolution of malaria parasite plastid targeting sequences’, *Proceedings of the National Academy of Sciences*, 105(12), pp. 4781–4785. doi: 10.1073/pnas.0707827105.

Torruella, G. *et al.* (2009) ‘The Evolutionary History of Lysine Biosynthesis Pathways Within Eukaryotes’, *Journal of Molecular Evolution*, 69(3), pp. 240–248. doi: 10.1007/s00239-009-9266-x.

Torto, T. A., Rauser, L. and Kamoun, S. (2002) ‘The pipg1 gene of the oomycete Phytophthora infestans encodes a fungal-like endopolygalacturonase’, *Current Genetics*, 40(6), pp. 385–390. doi: 10.1007/s00294-002-0272-4.

Trivedi, V. *et al.* (2010) ‘A modular approach for the generation, storage, mixing, and detection of droplet libraries for high throughput screening’, *Lab on a Chip*, 10(18), pp. 2433–2442. doi: 10.1039/C004768F.

Tsui, C. K. M. *et al.* (2009) ‘Labyrinthulomycetes phylogeny and its implications for the evolutionary loss of chloroplasts and gain of ectoplasmic gliding’, *Molecular Phylogenetics and Evolution*, 50(1), pp. 129–140. doi: 10.1016/j.ympev.2008.09.027.

Tyler, B. *et al.* (2006) ‘Genome Sequence of Phytophthora ramorum: Implications for Management’, p. 2.

Tyler, B. M. *et al.* (2006) ‘Phytophthora Genome Sequences Uncover Evolutionary Origins and Mechanisms of Pathogenesis’, *Science*, 313(5791), pp. 1261–1266. doi: 10.1126/science.1128796.

Tyler, B. M. (2007) ‘Phytophthora sojae: root rot pathogen of soybean and model oomycete’, *Molecular Plant Pathology*, 8(1), pp. 1–8. doi: 10.1111/j.1364-3703.2006.00373.x.

Tyler, B. M. and Gijzen, M. (2014) ‘The Phytophthora sojae Genome Sequence: Foundation for a Revolution’, in Dean, R. A., Lichens-Park, A., and Kole, C. (eds) *Genomics of Plant-Associated Fungi and Oomycetes: Dicot Pathogens*. Berlin, Heidelberg: Springer, pp. 133–157. doi: 10.1007/978-3-662-44056-8_7.

Van den Berg, B. *et al.* (2004) ‘X-ray structure of a protein-conducting channel’, *Nature*, 427(6969), pp. 36–44. doi: 10.1038/nature02218.

Van der Auwera, G. *et al.* (1995) ‘The phylogeny of the Hyphochytriomycota as deduced from ribosomal RNA sequences of Hyphochytrium catenoides.’, *Molecular Biology and Evolution*, 12(4), pp. 671–678. doi: 10.1093/oxfordjournals.molbev.a040245.

Velasco, A. M., Leguina, J. I. and Lazcano, A. (2002) ‘Molecular Evolution of the Lysine Biosynthetic Pathways’, *Journal of Molecular Evolution*, 55(4), pp. 445–449. doi: 10.1007/s00239-002-2340-2.

Vincken, J. P. *et al.* (1997) ‘Two general branching patterns of xyloglucan, XXXG and XXGG.’, *Plant Physiology*, 114(1), pp. 9–13.

de Visser, J. A. G. M., Cooper, T. F. and Elena, S. F. (2011) 'The causes of epistasis', *Proceedings of the Royal Society B: Biological Sciences*, 278(1725), pp. 3617–3624. doi: 10.1098/rspb.2011.1537.

VOGEL, H. (1960) 'Two modes of lysine synthesis among lower fungi : evolutionary significance', *Biochim. Biophys. Acta*, 41, pp. 172–173.

Vogel, H. J. (1961) 'Lysine Synthesis and Phytogeny of Lower Fungi : Some Chytrids versus Hyphochytrium', *Nature*, 189(4769), pp. 1026–1027. doi: 10.1038/1891026a0.

Vries, R. P. de and Visser, J. (2001) 'Aspergillus Enzymes Involved in Degradation of Plant Cell Wall Polysaccharides', *Microbiology and Molecular Biology Reviews*, 65(4), pp. 497–522. doi: 10.1128/MMBR.65.4.497-522.2001.

Vyas, V. K., Barrasa, M. I. and Fink, G. R. (2015) 'A *Candida albicans* CRISPR system permits genetic engineering of essential genes and gene families', *Science Advances*, 1(3), p. e1500248. doi: 10.1126/sciadv.1500248.

Walter, P. and Blobel, G. (1980) 'Purification of a membrane-associated protein complex required for protein translocation across the endoplasmic reticulum', *Proceedings of the National Academy of Sciences of the United States of America*, 77(12), pp. 7112–7116. doi: 10.1073/pnas.77.12.7112.

Walter, P. and Johnson, A. E. (1994) 'Signal sequence recognition and protein targeting to the endoplasmic reticulum membrane', *Annual Review of Cell Biology*, 10, pp. 87–119. doi: 10.1146/annurev.cb.10.110194.000511.

Wang, Y.-H. *et al.* (2018) 'Plastid Genome Evolution in the Early-Diverging Legume Subfamily Cercidoideae (Fabaceae)', *Frontiers in Plant Science*, 9. doi: 10.3389/fpls.2018.00138.

Wang, Z., Tyler, B. M. and Liu, X. (2018) 'Protocol of Phytophthora capsici Transformation Using the CRISPR-Cas9 System', in Ma, W. and Wolpert, T. (eds) *Plant Pathogenic Fungi and Oomycetes: Methods and Protocols*. New

York, NY: Springer (Methods in Molecular Biology), pp. 265–274. doi: 10.1007/978-1-4939-8724-5_17.

Wangen, J. R. and Green, R. (2020) ‘Stop codon context influences genome-wide stimulation of termination codon readthrough by aminoglycosides’, *eLife*. Edited by N. Sonenberg et al., 9, p. e52611. doi: 10.7554/eLife.52611.

Wass, M. N., Kelley, L. A. and Sternberg, M. J. E. (2010) ‘3DLigandSite: predicting ligand-binding sites using similar structures’, *Nucleic Acids Research*, 38(suppl_2), pp. W469–W473. doi: 10.1093/nar/gkq406.

Wass, M. N. and Sternberg, M. J. E. (2009) ‘Prediction of ligand binding sites using homologous structures and conservation at CASP8’, *Proteins: Structure, Function, and Bioinformatics*, 77(S9), pp. 147–151. doi: 10.1002/prot.22513.

Wen, T.-N. *et al.* (2005) ‘A Truncated Fibrobacter succinogenes 1,3−1,4-β-d-Glucanase with Improved Enzymatic Activity and Thermotolerance’, *Biochemistry*, 44(25), pp. 9197–9205. doi: 10.1021/bi0500630.

West, S. A. *et al.* (2007) ‘The Social Lives of Microbes’, *Annual Review of Ecology, Evolution, and Systematics*, 38(1), pp. 53–77. doi: 10.1146/annurev.ecolsys.38.091206.095740.

Westphal, Y. *et al.* (2010) ‘MALDI-TOF MS and CE-LIF Fingerprinting of Plant Cell Wall Polysaccharide Digests as a Screening Tool for Arabidopsis Cell Wall Mutants’, *Journal of Agricultural and Food Chemistry*, 58(8), pp. 4644–4652. doi: 10.1021/jf100283b.

Whelan, S. and Goldman, N. (2001) ‘A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach’, Molecular Biology and Evolution, 18(5), pp. 651-9. doi 10.1093/oxfordjournals.molbev.a003851.

Whisson, S. C. *et al.* (2007) ‘A translocation signal for delivery of oomycete effector proteins into host plant cells’, *Nature*, 450(7166), pp. 115–118. doi: 10.1038/nature06203.

Whitaker, J. W., McConkey, G. A. and Westhead, D. R. (2009) ‘The transferome of metabolic genes explored: analysis of the horizontal transfer of enzyme encoding genes in unicellular eukaryotes’, *Genome Biology*, 10(4), p. R36. doi: 10.1186/gb-2009-10-4-r36.

Williams, E. J. B., Pal, C. and Hurst, L. D. (2000) ‘The molecular evolution of signal peptides’, *Gene*, 253(2), pp. 313–322. doi: 10.1016/S0378-1119(00)00233-X.

Wilson, D. B. (2009) ‘The first evidence that a single cellulase can be essential for cellulose degradation in a cellulolytic microorganism’, *Molecular Microbiology*, 74(6), pp. 1287–1288. doi: 10.1111/j.1365-2958.2009.06889.x.

Wilson, R. A. and Talbot, N. J. (2009) ‘Under pressure: investigating the biology of plant infection by Magnaporthe oryzae’, *Nature Reviews Microbiology*, 7(3), pp. 185–195. doi: 10.1038/nrmicro2032.

Won, H. and Renner, S. S. (2003) ‘Horizontal gene transfer from flowering plants to Gnetum’, *Proceedings of the National Academy of Sciences*, 100(19), pp. 10824–10829. doi: 10.1073/pnas.1833775100.

WOOD, P. J. and WEISZ, J. (1987) ‘Detection and assay of (1-4)-β-D-glucanase,(1-3)-β-D-glucanase, (1-3)(1-4)-β-D-glucanase, and xylanase based on complex formation of substrate with Congo red’, *Detection and assay of (1-4)-β-D-glucanase,(1-3)-β-D-glucanase, (1-3)(1-4)-β-D-glucanase, and xylanase based on complex formation of substrate with Congo red*, 64(1), pp. 8–15.

WRIGHT, S. (1932) ‘The roles of mutation, inbreeding, crossbreeding and selection in evolution’, *Proceedings of the sixth international congress of Genetics*, 1, pp. 356–366.

Wu, A.-M. *et al.* (2009) ‘The Arabidopsis IRX10 and IRX10-LIKE glycosyltransferases are critical for glucuronoxylan biosynthesis during secondary cell wall formation’, *The Plant Journal*, 57(4), pp. 718–731. doi: 10.1111/j.1365-313X.2008.03724.x.

Yamagaki, T., Mitsuishi, Y. and Nakanishi, H. (1998) ‘Determination of structural isomers of xyloglucan octasaccharides using post-source decay fragment analysis in MALDI-TOF mass spectrometry’, *Tetrahedron Letters*, 39(23), pp. 4051–4054. doi: 10.1016/S0040-4039(98)00655-8.

Yaoi, K. *et al.* (2005) ‘Cloning and Characterization of Two Xyloglucanases from Paenibacillus sp. Strain KM21’, *Applied and Environmental Microbiology*, 71(12), pp. 7670–7678. doi: 10.1128/AEM.71.12.7670-7678.2005.

Yaoi, K. and Mitsuishi, Y. (2004) ‘Purification, characterization, cDNA cloning, and expression of a xyloglucan endoglucanase from Geotrichum sp. M12811Nucleotide sequence data reported in this paper are available in the DDBJ/EMBL/GenBank database under accession number AB116528.’, *FEBS Letters*, 560(1), pp. 45–50. doi: 10.1016/S0014-5793(04)00068-7.

Yčas, M. (1974) ‘On earlier states of the biochemical system’, *Journal of Theoretical Biology*, 44(1), pp. 145–160. doi: 10.1016/S0022-5193(74)80035-4.

Yeung, N. *et al.* (2009) ‘Rational Design of a Structural and Functional Nitric Oxide Reductase’, *Nature*, 462(7276), pp. 1079–1082. doi: 10.1038/nature08620.

Yoon, H. S. *et al.* (2002) ‘The Single, Ancient Origin of Chromist Plastids’, *Journal of Phycology*, 38(s1), pp. 40–40. doi: 10.1046/j.1529-8817.38.s1.8.x.

York, W. S. *et al.* (1990) ‘Structural analysis of xyloglucan oligosaccharides by 1H-n.m.r. spectroscopy and fast-atom-bombardment mass spectrometry’, *Carbohydrate Research*, 200, pp. 9–31. doi: 10.1016/0008-6215(90)84179-X.

York, W. S. *et al.* (1996) 'The structures of arabinoxyloglucans produced by solanaceous plants', *Carbohydrate Research*, 285, pp. 99–128. doi: 10.1016/S0008-6215(96)90176-7.

Zallot, R. *et al.* (2016) 'Functional Annotations of Paralogs: A Blessing and a Curse', *Life*, 6(3), p. 39. doi: 10.3390/life6030039.

Zerillo, M. M. *et al.* (2013) 'Carbohydrate-Active Enzymes in Pythium and Their Role in Plant Cell Wall and Storage Polysaccharide Degradation', *PLOS ONE*, 8(9), p. e72572. doi: 10.1371/journal.pone.0072572.

Zhaxybayeva, O. and Doolittle, W. F. (2011) 'Lateral gene transfer', *Current Biology*, 21(7), pp. R242–R246. doi: 10.1016/j.cub.2011.01.045.

Zinchenko, A. *et al.* (2014) 'One in a Million: Flow Cytometric Sorting of Single Cell-Lysate Assays in Monodisperse Picolitre Double Emulsion Droplets for Directed Evolution', *Analytical Chemistry*, 86(5), pp. 2526–2533. doi: 10.1021/ac403585p.