The
University
Of
Sheffield.

Glycoengineering *Escherichia coli*: Identifying and characterizing increased mannose availability within the cell for N-glycoprotein production

Olufikayo Adepoju

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield
Faculty of Engineering

The Department of Chemical and Biological Engineering

Submitted December 2021

Acknowledgements

I acknowledge God – my source, without whom none of this would have been possible. For life, health and every blessing I say thank you.

I would also like to thank my amazing supervisor Dr Jagroop Pandhal for his patience, guidance and profound support throughout this PhD journey. He had a huge impact not only in my research output but also helping me maintain a healthy work-life balance for metal stability during these past 4 years. To my sponsors – Petroleum Technology Development Fund Nigeria and the University of Sheffield, I am indeed grateful for the opportunity.

To the brilliant researchers I had the pleasure of working with in the Chemical and Biological Engineering Department over the years particularly Dr Stephen Jaffe, Dr Caroline Evans, Dr Joy Mukherjee, James Gringham and Kasia I say thank you for your support and advice over the years. I would also like to thank Dr Michael Trikic, Sue Clark and Alice Seleiro for their support.

To the wonderful team / research members of the Pandhal group – Josie, Alaa, Juliano, Wan, Hannah, Helen, Mengxun, Zongting, José and Ali, thanks for being the best support system one could ever need in an enjoyable lab existence.

A big thank you to my friends Ziba, Jide, Toni, Lillian and Ope. The joys and tears shared on this journey only make the experience all the more cherished as it draws to a close. To my church and fellowship family, you have made Sheffield home and I could never repay you for all you have done for me.

Special thanks to my parents for their sacrifices and unwavering support throughout the years. I am grateful for your giving of not only time, money but also self to ensure I complete this successfully. To my in-laws and sibling's thanks for all the love.

Finally, to my husband Abiodun and girls – we did this! You kept me grounded and accepted all I could spare after all had been spent trying to achieve this goal. Eri and Ari this is for you, I run so you can fly and soar. Love always.

Thesis summary

The experiments contained within this thesis are aimed at improving N-glycoprotein production potential within *Escherichia coli* by increasing the availability of the N-glycan precursor mannose within the cell factory using a flow cytometric approach, chemically induced mutagenesis, Next Generation Sequencing techniques, as well as other bioinformatics tools and a glycoproteomics strategy in identifying promising candidates to achieve this aim.

The eukaryotic type N-glycosylation pathway requires mannose as a key precursor, and its enhanced availability in the cell has been linked with increased glycosylation efficiency. In this work, the genetic diversity of a glycan surface display *E. coli* parent strain W3110 was increased via a chemical mutagen and increased mannose generating mutants were identified using flow cytometry. The isolated mutant cells showed a 2.4-fold increase in cell surface mannose display compared to the wild type strain.

A Next Generation Sequencing approach was then employed in analysing and identifying the genetic level changes within the mutant strains which led to the significant increase observed compared to the wild type strain. Using bioinformatics tools and techniques, several gene variants within the cell were identified and characterised to highlight potential targeted genetic engineering gene candidates for enhanced mannose production within the *E. coli* cell factory.

Finally, glycoproteomics strategies were employed in investigating the possibility of non-targeted N-glycosylation of native proteins in *E. coli* containing a glycosylation machinery. Glycosylation prediction tools were used to identify potential endogenous N-glycoproteins, while also investigating possible relationships and interactions within these proteins for predicted characteristics they possess that could significantly influence the N-glycosylation potential in this bacterial system.

The work in this thesis has contributed further insights into the genetic pathways and potential for enhancement within *E. coli* that make it a suitable and efficient N-glycoprotein production cell factory. The findings can be taken forward to identify specific gene combinations which can be targeted to achieve even higher mannose availability within the strain. Findings from other researchers in the field of N-glycosylation engineering within *E. coli* particularly secretion pathway research findings can be combined with the findings in this thesis to achieve more efficient protein N-glycoprotein production in *E. coli*. Ascertaining the possibility of native protein N-glycosylation within E. coli will determine its industrial applicability as an efficient N-glycoprotein production factory.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

°C	degrees Celsius
A	adenine
Bp	base pair
C	cytosine
Da	daltons
dH_2O	deionised water
DNA	deoxyribonucleic acid
DTT	dithiothreitol
E.	*Escherichia*
EDTA	ethylenediaminetetraacetic acid
EMS	ethylmethanesulphonate
FACS	fluorescence-activated cell sorting
FDR	false discovery rate
H_2O	water
His_6-tag	6x poly-histidine tag
HPLC	high performance liquid chromatography
HRP	horseradish peroxidase
kbp	kilobase pair
kDa	kilodalton
λ	wavelength
L	litre
LB	Luria broth
LC	liquid chromatography
LFQ	label-free quantification
µL	microliter
µM	micromolar
µmol	micromoles
m	metre
M	molar
mg	milligram

MgCl2	magnesium chloride
min	minute
mL	millilitre
mm	millimetre
mM	millimolar
MOPS	3-(N-morpholino)propanesulfonic acid
NaCl	sodium chloride
NCBI	National Center for Biotechnology Information
NEB	New England Biolabs
nfH2O	nuclease-free water
ng	nanogram
O2	oxygen
OD	optical density
ORF	open reading frame
PCR	polymerase chain reaction
SDS	sodium dodecyl sulphate
SDS-PAGE	sodium dodecyl sulphate – polyacrylamide gel electrophoresis
SNP(s)	single nucleotide polymorphisms
T	thymine
TAE	tris-acetate-EDTA
V	volts
v/ v	volume/ volume
w/ v	weight/ volume
WT	wild-type

Declaration

I, the author, confirm that the Thesis is my own work. I am aware of the University's Guidance on the Use of Unfair Means (www.sheffield.ac.uk/ssid/unfair-means). This work has not been previously presented for an award at this, or any other, university.

Chapter 1: Introduction

Enzymes and other useful proteins were originally obtained from cell extracts. However, with increasing demands for large amounts, they are now being recombinantly produced in heterologous cell machineries. This is based on the molecular biology central dogma (DNA to mRNA to protein). Bacteria, yeast, insect and plant systems are examples of the host cells used. To choose the appropriate host cell factory, the structural and functional requirements of the enzyme are considered. There has been extensive research into manipulating the metabolism and processing parameters for different host organisms which has led to highly specialised products and increased product yield. In selecting the right host cell for enzyme production, it is important to consider if the protein requires post translational modifications (PTMs) that can affect properties like stability, half-life and function of the protein. Normally, enzymes with specific PTMs requirements would be expected to be produced in host cells that can natively perform these modifications, (i.e., yeast, insect, plant and mammalian cells), the low processing costs in bacteria production systems means this is often mostly preferred.

The most characterised diverse and energetically demanding post translational modification for human life to date is protein glycosylation. This is mainly because of the effect of carbohydrate-protein interactions in many biological processes. Because of its wide occurrence and obvious impact, approximately 40% of recombinant therapeutic drugs approved are glycosylated (Walsh, 2014) with a large number produced in a host system possessing a natural glycosylation machinery. Asparagine (N-linked) glycosylation of proteins has been recognized in protein folding and their ability to traverse cellular secretory pathways in eukaryotes which invariably increases their stability. For example, N-glycosylation decreases the portion of aggregation prone proteins thus creating a more stable protein structure (Culyba *et al*, 2011). Also, the addition of sugar groups to non-natively glycosylated proteins can change the protein structure and function with biotechnologically useful results. An example is in the case of native Leaf branch compost cutinase (LCC) – (a bacteria enzyme that hydrolyzes polyethylene terephthalate (PET)) expressed in *Pichia pastoris* which resulted in the production of glycosylated LCC. The glycosylation made the native state aggregation of the enzyme more stable while increasing the enzyme's thermally induced aggregation temperature by 10^{o}C. This improved the enzyme activity and thermostability in the long run (Shirke *et al*, 2018). Recombinant hydroxynitrile lyase enzymes from passion fruits also showed similar properties when expressed in *E. coli* and *Pichia pastoris*. The enzyme expressed in *Pichia sp*. a glycosylation capable host cell (N-glycosylated) displayed improved catalytic properties (thermostability, pH stability and organic solvent tolerance) compared to the *E. coli* expressed aglycosylated enzyme (Nuylert *et al*, 2017).

Currently *E. coli* is the most studied and understood recombinant protein production system (Rosano & Ceccarelli, 2014). With its ever expanding synthetic biology toolbox, flexibility for genetic manipulation and recent successful transfer of the *Campylobacter jejuni* glycosylation pathway (Wacker *et al*, 2002b) a vast new area of research has opened up. Although a relatively inefficient system for glycosylated protein production (Ding *et al*, 2017), the established infrastructure for recombinant protein cultivation in *E. coli* and the "blank canvas" for addition of glycosylation related components to its genome, has provided a well-placed adaptable system for use in the production of bespoke glycoenzymes for industrial use. The production of glycosylated enzymes in *E. coli* can be explored and optimized to better understand the abilities conferred on these enzyme products.

The studies contained within this thesis are based on progressing research stemming from the successful transfer of the glycosylation machinery from *Campylobacter jejuni* into *E. coli* to produce recombinant glycoprotein in an engineered *E. coli* host (Szymanski *et al*, 1999; Wacker *et al*, 2002a). The *E. coli* host glycoprotein production capabilities were remarkably improved. A review of the glycosylation process provides understanding into this significant development. It highlights the original characterization and discovery in the 3 life domains. With focus on evolutionary significance and comparison within the three domains, an understanding into the expression/secretion yield and other industrially desirable traits in glycoenzyme/glycoprotein production in the *E. coli* cell factory is gained.

1.1 Research Objectives

Many novel glycoproteins with different sugar assortments and improved characteristics have been found to possess significant biotechnological benefits. The right set of required sugar building blocks and glycosylation enzymes coupled with practical glycoengineering knowledge of *E. coli* has led to renewed exploration on the effect N-glycans have on glycoproteins. Presently there is need for increased efficiency of this modification being achieved on recombinant proteins and ascertaining the multi-effect of this enhancement in protein function regulation, on the binding affinity, substrate specificity, thermostability, activity, and the general role of N-glycans (Skropeta, 2009). The overall aim was to improve *E. coli* as a host cell for making N-glycoproteins, with specific objectives outlined below:

- To create an *E. coli* strain that can synthesise increased amounts of GDP-mannose – a nucleotide sugar for the biosynthesis of $Man_3GlcNAc_2$, an exemplar glycan: A mutagenesis and screening methodology was developed based on mannose surface display and fluorescence-based cell sorting.

- To sequence and characterise the enhanced mannose substrate generating strain: Identify genetic modifications within the mutant strains responsible for the new phenotype. Modification of the mutant strains for transfer of glycan onto target protein and measuring N-glycoprotein production efficiency within the cell factory.
- To investigate the incidence of off-target proteins being glycosylated within the *E. coli* cell factory: A tandem mass spectrometry analysis of *E. coli* periplasmic proteins with bacterial N-glycan consensus sequences.

It is however worthy to note the impact of the COVID-19 pandemic on the experimental plans and output of this research project. Due to extended months of lockdown in which access to the laboratory was denied and subsequent periods of rotational phased return to work, a lot of the planned experiments were modified and re-designed to fit into the timescales for research completion.

Chapter 2: Literature review

Understanding the origins of enhanced protein engineering and recombinant protein production in the *E. coli* cell factory would allow the development of better strategies to engineer cell-lines with improved competencies and efficient bioprocessing abilities. This chapter will provide an overview of the genetic and process challenges associated with recombinant protein production in *E. coli*. Also to be mentioned are the post translational modifications in proteins with specific focus on N-glycosylation and the current updates in *E. coli* glycoengineering. Finally, this review will highlight the methods and techniques for glycoprotein detection and quantification.

2.1 Enzyme/protein engineering

Various steps of biochemical processing and metabolism require very important organic reagents known as biocatalysts. The 20th century discovery of microbial enzymes has led to increased interests in them particularly for utilization on a big sustainable scale in industrial applications. Research into their characteristic properties, isolation, scale-up from laboratory to pilot plant level and use in industry has led to understanding and discovery of more microbial enzymes. Several microbially sourced enzymes are in use for a wide variety of commercial reactions. Microorganisms like fungi, yeasts and bacteria have been extensively studied for their ability to biologically produce cheap and useful amounts of various enzymes for commercial applications. Many novel enzymes have been successfully designed by using techniques such as metagenomics, bio-reaction engineering and protein engineering. Various molecular methods have been used to improve the microbial enzyme quality and activity for widespread applications in industries. The global market uses many newly created valuable products in recognized bioprocess technology to purposefully engineer biological enzymes (Nigam, 2013).

Protein engineering involves altering the protein structure to improve its properties. This added understanding coupled with the evolution and function of enzymes is critical in the improvement of enzyme properties for different applications like pharmaceuticals, biofuel production and green chemistry (Romas & Uwe, 2009). Protein/biocatalyst engineering and specifically directed enzyme evolution are rapidly developing research areas focused on understanding challenges related to the use of enzymes in enabling creation of innovative products which are comparable or better enhanced to native proteins (Goldsmith & Tawfik, 2012). Biologists have been making use of the site-specific mutagenesis molecular tool to change an amino acid in a specific position within the protein sequence to one of the other available 19 amino acids. Consequently, this resulted in the development of vital methods which have been used in enzyme mechanism studies. These studies have offered better understanding into complex enzyme mechanisms through the use of site-specific mutagenesis combined with many other established experimental and computational methods developed over the past few years. So far, this has emphasized biocatalysis complexities as going further than originally predicted (Reetz, 2013).

Recent concept improvements to existing methods have increased the tools and abilities accessible to enzyme engineers. Amino acid substitutions are predominantly how protein engineering is used to improve enzyme performance as catalysts for use in synthetic organic chemistry and biotechnology. Emphasis on improving thermostability and stereo selectivity is of particularly useful interest (Goldsmith & Tawfik, 2012; Reetz, 2013). Engineered enzymes can catalyze reactions using substrates that the original native protein would not normally catalyze. An example is the use of engineered mono-oxygenase - butane mono-oxygenase or cytochrome P450 for the hydroxylation of propane. Thermostability changes, organic solvent compatibility, low pH and peroxide resistance have been discovered using this method (Garcia-Ruiz *et al*, 2012).

Protein variants have been created not only through the replacement of one amino acid for another but also the addition or removal of extra amino acids or changing the termination locations (a circular permutation). These changes can take place at specific regions or all through the protein sequence to cause substantial improvements (Hawkins-Hooker *et al.*, 2021). One of the main targets of protein engineering is increased enzyme thermostability. Structure based approaches rely on the underlying assumption that the rigidity of the enzyme corresponds to better stability at higher temperatures (Prakash and Jaiswal, 2009). Disulfide bonds or salt bridges are examples of stabilizing structures which can be designed from the X-ray structure of the enzyme. Target areas for mutagenesis can be identified from experimental B-factors to find the most flexible protein region or through the removal of glycine and introduction of proline in stabilizing the loop region of the protein. Through a bioinformatics-based approach, it is assumed that conserved amino acid regions provide the greatest contribution to stability (Greene *et al.*, 2001). With the engineering approach, related sequences are compared and the target protein is engineered to mimic the most commonly available amino acids at each position (consensus sequence). Increase in the enzyme catalytic activity or stereoselectivity is also a main protein engineering goal. Single amino acid substitution analyses revealed substitutions closer to the active site yielded massive improvements in enantioselectivity or diastereoselectivity compared to substitutions that are further from the active site. The substitutions which increase catalytic activity are nevertheless believed to be randomly located in the protein (Romas & Uwe, 2009).

2.1.1 Applications of proteins produced in microbial systems

The main attraction of microbially produced proteins lies in their substantial ability to yield better products in stressful temperature and pH conditions. An example is microbial enzymes which are categorized as being either thermophilic, alkalophilic or acidophilic. Thermostable enzymes originating from microorganisms have been known to function at higher than normal reaction temperatures and thus reduces the risk of microbial contamination during longer period large scale industrial reactions. More thermostable enzymes can easily digest and breakdown substrate raw materials as more enzyme penetration can be achieved at higher temperatures (Zhu *et al*, 2022). Due to their characteristic activity at higher temperatures and their ability to remain stable for longer processing periods at varied temperatures, hydrolytic enzymes are widely sought after. During the hydrolysis of substrates or raw materials in industrial processes, reduction and mass transfer of the substrate viscosity is increased with the existence of high temperature enzymes. Thermophilic xylanases are of high commercial interest in many industries, for example in brewing for the mashing process. Other useful applications of the thermostable plant xerophytic variants of laccase enzyme include uses in textile dyeing, pulp and paper and bioremediation (Nigam, 2013).

The process where two organic molecules (usually similar sized) are joined together through the action of an enzyme is known as coupling. This type of reaction is gradually gaining importance in some industries. Laccase enzymes are known to have the ability to mediate coupling reactions because they can basically act on any substrate that possesses characteristics related to a p-diphenol. Laccase-mediated coupling reactions are important in the textile industry, lignocellulosic material modifications, control of environmental pollution, organic product synthesis, pharmaceutical and food industry. With the use of laccases in coupling reactions, a green alternative to chemical methods (that are less specific, costly and environmentally harmful) is provided (Kudanga *et al*, 2011).

Recombinant gene technology has improved manufacturing processes and led to the production and advancement of commercialized enzymes. Increased production of industrial enzymes has been encouraged by the recent introduction of protein engineering and directed evolution in modern biotechnology. This has also led to the development of specialized enzymes that display new characteristics and modified to new conditions that have generated further industrial uses (Kirk *et al*, 2002). This is demonstrated by the amount of a highly diversified industry that continues to advance as shown in the table below:

Table 2.1: Industrial applications of microbial enzymes.

	Type of Enzyme	Industry	Application	Substrates
1	Cellulase	Textile, Food processing, Detergent, Pulp & Paper, Bioremediation, Bioethanol production, Agriculture, Wine & brewery, Waste management & Oil extraction industries	Cleaning, colour clarification, cotton softening, Denim finishing, de-inking, fibre modification, coffee bean drying, improved feed digestion, beer/wine production, extraction of carotenoids & oils.	Cellulosic substances
2	Lipase	Pharmaceutical, Leather, Cosmetic, Paper, Food, Detergent, Fine chemicals, Biodiesel & Bioremediation	Detergent production, flavour enhancement of cheeses, control of oil spills, transesterification, lipid hydrolysis, production of biodegradable polymers, textile dyeing, lipid stain removal, dough stability and conditioning & leather de-pickling.	Triglycerides, organic pollutants (oil spill)
3	Protease	Leather, Detergent, Starch and Fuel, Food, Bioremediation, Medical,	Cheese production, pre-digesting proteins in baby food, unhearing of leather, biofilm removal, protein stain removal, milk clotting.	Proteins
4	Peroxidase	Personal care, Pharmaceutical, Bioremediation, Textile, Food, Pulp	Excess dye removal, antimicrobial uses, analysis & diagnostic kits, biosensors, bio-bleaching, dye degradation, enzyme immunoassays,	Phenolic compounds, Polycyclic aromatic compounds, Methoxybenzene

		and paper, Diagnostics	removal of phenolic contaminants, fruit growth and ripening, hair dyeing and determination of lipid peroxidation extent in meat food products.	
5	α-Amylase	Detergent, Pulp and paper, Food, Biofuel	Baking (softness and bread volume), starch stain removal, de-inking and drainage improvement, juice treatments, glucose & fructose syrup production	Carbohydrates
6	Xylanase	Food, Agriculture, Pulp & paper, Starch and fuel	Animal feed digestibility, dough conditioner, bleach boosting, starch viscosity reduction	Xylan
7	Phytase	Agriculture	Animal feed (composition, phytate digestibility – phosphorus release)	Phytic acid
8	Laccase	Bioremediation, Medical, Pharmaceutical, Biosensors, Personal care/Cosmetics	Biofuel cells, medical diagnostic tools, biosensors, cleaning agents in water purification	Phenols, Amines, Aromatic compounds, highly resistant environmental pollutants and lignin related compounds

(Andualema & Gessesse, 2012; Ramesh Chander *et al*, 2011; Rigoldi *et al*, 2018)

2.1.2 Challenges of enzyme use in industry

Properties like cost-effectiveness, low energy consumption, low environmental impact and stability in mild temperature and pH conditions are highlights of the usefulness of enzymes in industry. However, it also comes with its own share of challenges including controlled production requirements in terms of appropriate hosts, well established transformation techniques and acceptable expression vectors which cannot be guaranteed (Rigoldi *et al*, 2018). Efficient enzyme design can be used to demonstrate adequate grasp of enzyme catalysis. This is mainly due to codon variation in frequently used expression systems such as *E. coli* or *Bacillus sp*., as few systems have been able to successfully achieve this (Frushicheva *et al*, 2011; Rigoldi *et al*, 2018).

2.2 Enhanced protein discovery

Due to the significant environmental problems being created by chemically produced catalysts and the ever-expanding market for commercial enzymes in industries, there has been an increased demand for new biocatalysts and the production capacity has also improved. The common aims of technology development in enzyme bio-manufacturing include use of new enzymes, enzyme property and production process improvement. Systematic techniques in enzyme engineering has given rise to the creation of new enzymes through engineering existing enzymes using genetic engineering methods, the screening of natural samples that have better characteristics and refining enzyme manipulation methods to combat catalyst limitations like downstream manufacturing processing, enzyme formulation and immobilization (Li *et al*, 2012).

Biotechnology uses a huge number of commercially manufactured enzymes by using purposefully screened microorganisms. These microorganisms have been specifically designed, characterised and enhanced to produce superior enzymes in large quantities for industrial applications. Microbial enzymes are studied for the exclusive characteristics which make them appropriate for various industrial bioprocesses. Specific microorganisms have recently been modified to produce high yield enzymes, enzymes with desirable features like thermostability and acid or alkaline stable enzymes. They can also retain their activity in reaction conditions such as in the presence of heavy metals and compounds (Nigam, 2013).

2.2.1 Post-Translational Modifications

Several proteins cannot function as non-modified folded polypeptides because they require either permanent or temporary molecular alterations in most cases to function appropriately. Post-translational modification (PTM) of proteins usually occur as covalent modifications at particular amino acids or proteolytic cleavage actions (Blom *et al*, 2004). Protein diversity can be created by the control of PTMs that increases the possible protein applications through the addition of small chemical molecules to specific amino acids or alternative splicing of mRNA. Cellular functions like metabolism, signal transduction and protein stability have been found to be affected by different types of PTMs (Figure 2.1) (Yonathan Lissanu *et al*, 2010).

The modifications affect specific amino acids. For example, in phosphorylation the variation occurs mostly on serine, threonine and tyrosine amino acids, while for covalent glycosylation, it affects asparagine, serine and threonine residues. Not all these amino acids in a protein undergo modification. The transferase involved in enzymatic post translational modification in most cases, only recognizes acceptor motifs (sequence patterns) around the specific amino acid to make the PTM (Blom *et al*, 2004).

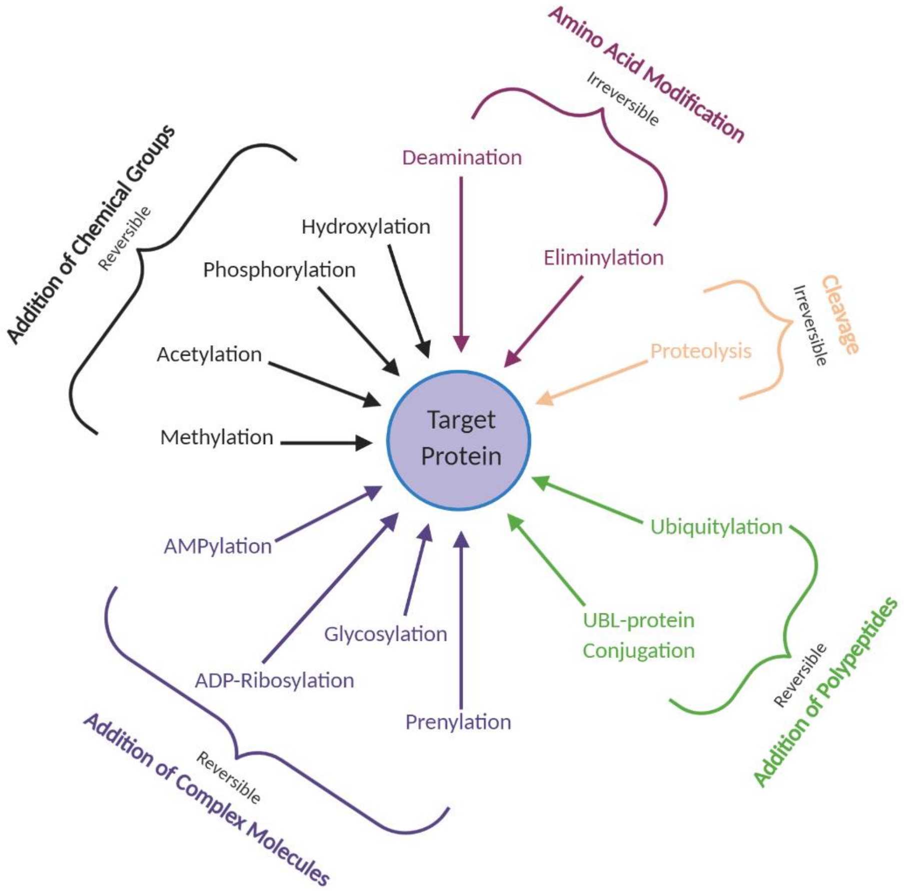


Figure 2.1 Types of post-translational modifications (PTMs)

Recent technological advances have meant post translational modifications can now be detected at an increasing rate and with great quality and precision. An example is with the use of mass spectrometry (MS) based methods (Kim *et al*, 2006).

2.2.2 Glycosylation

Glycosylation is an important post translational modification in proteins. It involves the attachment of sugars to amino acid side chains which can endow proteins with a wide variety of properties of great interest to the engineering biology community (Kightlinger *et al*, 2020). The monosaccharides are linked by glycosydic bonds to form a glycan covalently attached to the biological molecule (Baker *et al*, 2013). Glycosylation takes place through chemoenzymatic activity in biological systems and is consequently referred to as an enzyme-catalysed reaction involving the covalent addition of carbohydrate units to polypeptides, lipids, polynucleotides, carbohydrates or other organic compounds. Glycosyltransferases are the enzymes required for the catalysis reaction and they make use of specific sugar molecules as donor substrates (Lin *et al*, 2020; Varki, 2017).

Johansen and colleagues' 1961 research discovered glycosylation by detecting the sugar residue - GlcNAc linked to asparagine residue within a polypeptide chain of an ovalbumin protein in a GlcNAc-β-Asparagine linkage. Later a few other polypeptide monosaccharide linkages were identified. Glycosylation was thus defined as the attachment of sugar molecules called glycans to protein (Johansen *et al*, 1961).

Protein glycosylation was originally thought to occur strictly in eukaryotes until the 1970s. The S-layer (surface layer) glycoprotein of the archaeobacterium - *Halobacterium salinarium* was the first bacterial glycoprotein characterized in detail. This subsequently led to more research into S-layer eubacteria and archaeobacteria glycoproteins (Messner, 1997). It later became obvious that protein glycosylation occurs in all the three domains of life. It is now recognized that about 70% of eukaryotic and 50% prokaryotic proteins are post-translationally glycosylated (Dell *et al.*, 2010). Due to the fact that they carry oligosaccharide chains which are covalently bound to some amino acids, many eukaryotic proteins are glycoproteins. Protein glycosylation, of all naturally occurring post-translational modification processes, is probably the most important and most common. It impacts protein expression, folding, cell localisation and half-life, solubility, biological activity, antigenicity and cell-cell interactions. These in turn are relevant for downstream biological processes like cell immune behaviour and protein function (Baker *et al*, 2013; Blom *et al*, 2004; Schäffer *et al*, 2017).

2.2.2.1 Types of Glycosylation

A vast number of glycosylation types can occur in proteins because there are generally multiple sites within a protein sequence with various glycosidic linkages., This is however subject to several factors including:

1. The availability of required enzyme (glycan processing step can be regulated by varying the enzyme concentration)
2. The amino acid sequence (consensus sequences are required for glycosidic bond formation) and

3. The accessibility of target amino acids required for glycosylation to occur largely depends on the protein conformation (ability of the synthesized protein to fold into the developed secondary structure) (Spiro, 2002).

The specific group glycopeptide bonds fall into is based on the nature of the peptide-sugar bond and the oligosaccharide attached to in the formation of any of either N-, O-, C-linked glycosylation, phosphoglycosylation or glypiation.

Table 2.2: A description of the protein-sugar group linkage within the five types of glycosylation.

Types of Glycosylation	
Linkage	**Sugar Attachment Position**
N-glycosylation	The glycan forms a covalent bond with the amino group of asparagine residues in the endoplasmic reticulum.
O-glycosylation	The sugar group forms a covalent bond with the hydroxyl group of either serine or threonine in the nucleus, endoplasmic reticulum or cytosol.
C-glycosylation	Mannose is covalently attached to a carbon atom in the indole ring of tryptophan.
Glypiation	This involves the covalent attachment of phospholipid and a polypeptide chain.
Phosphoglycosylation	A phosphodiester bond is covalently attached to a glycan carrying a phosphor group to serine.

2.2.2.1.1 N-glycosylation

Due to the sugar group in N-glycosylation being linked to a protein before its subsequent transportation into the endoplasmic reticulum (ER) it is said to occur co-translationally. The translation and processing site of most membrane-bound and secreted proteins is the ER which makes most to be classified as N-linked glycoproteins. A common carbohydrate–peptide bond is the β-glycosylamine linkage of GlcNAc to asparagine. Other complex and polymannose oligosaccharides are also site specifically attached to biologically significant proteins such as antibodies (Spiro, 2002). A substantial amount of the enzymes and processes involved in N-glycosylation are conserved across different species and also with nearly 90% of glycoproteins being N-glycosylated, it is consequently the most common glycosylation (Trombetta, 2003).

The *N*-linked glycosylation process within eukaryotes has been extensively characterised. Glycans in the *N*-type are attached to asparagine residues within a defined potential glycosylation consensus sequence of N-X-S/T, where X can be any amino acid except proline (Marshall 1973). The *N*-type glycans added have very different structure and compositions when compared to the other types of glycosylation. Within eukaryotes, the core first five sugars form the basis for all the *N*-type glycans, see Figure 2.2 (Lyons *et al*, 2015). This consists of two GlcNAc residues known as the chitobiose core, which is followed by one mannose residue and two additional mannose residues in a branched formation. It is written chemically as, Manα1–6(Manα1–3)Manβ1–4GlcNAcβ1–4GlcNAcβ1-Asn-X-Ser/Thr.

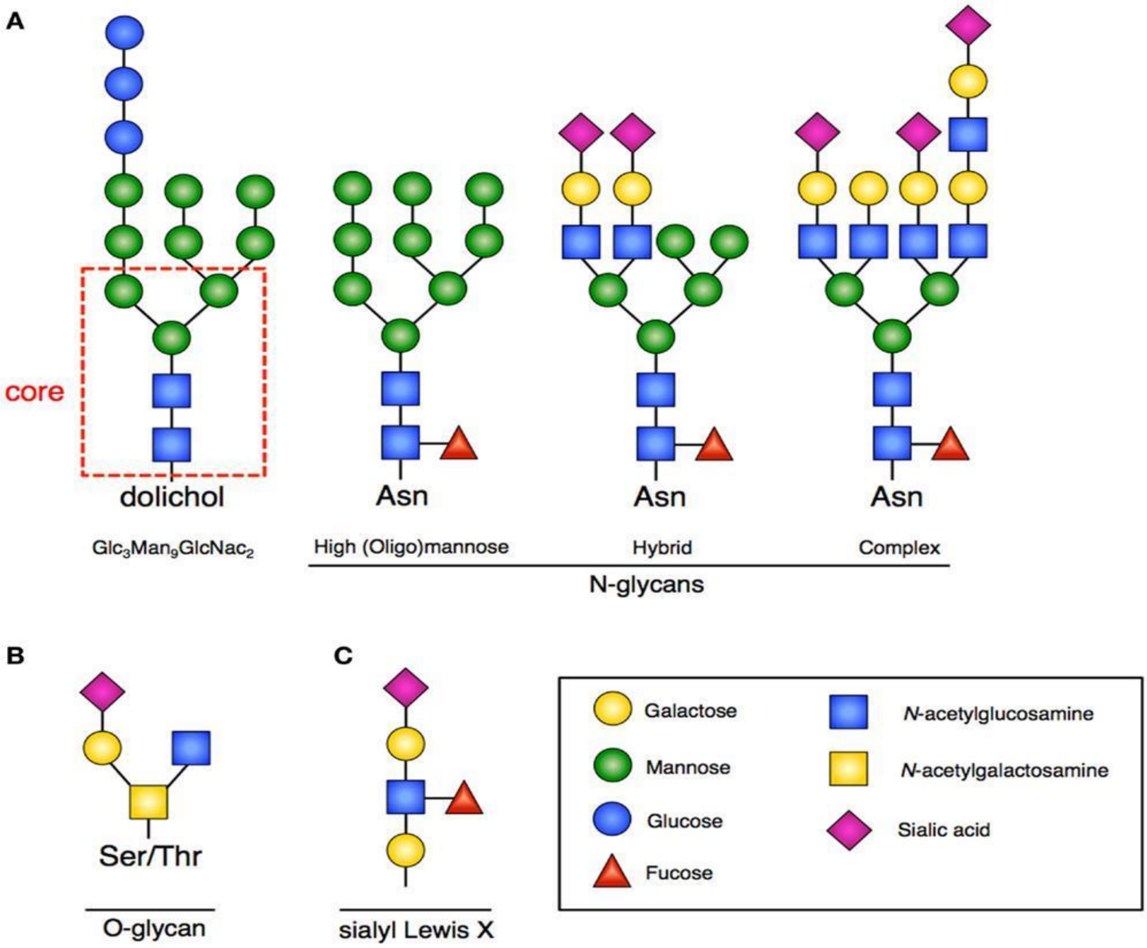


Figure 2.2 Glycan structures. (A) N-glycans with increasing complexity from left to right. (B) O-glycan (C) Sialylated glycan (Lyons *et al*, 2015)

The production and attachment of glycans to asparagine residues within the consensus sequence of the proteins is a highly complex process that has been extensively studied using *Saccharomyces* *cerevisiae* as the model organism for understanding how the process happens in eukaryotic cells (Kukuruzinska, Bergh and Jackson, 1987)

2.2.2.1.2 O-glycosylation

O-glycosylation is a very common post-translational modification and takes place on the oxygen atoms of the side chains of serine, threonine or tyrosine amino acids. It is observed in different pathological and biological processes and functions distinctively in the biosynthesis of mucins, cell–cell adhesion and communication, protein–protein interaction and immunization. O-Glycosylation take place post-translationally and originates with the addition of anyone of six different monosaccharides - α-GalNAc, β-GlcNAc, α-Fuc, α-Man, β-Xyl, β-Gal, and β-Glc to serine or threonine side chains in the Golgi apparatus. In glycoproteomics and glycomics, *O-N*-acetylgalactosamine (O-GalNAc) and *O-N*-acetylglucosamin (O-GlcNAc) glycosylations are more widely studied because of their crucial biological roles. Mucin-type O-glycosylation – O-GalNAc, is present across many species including fungi, insects, worms and mammals (You *et al*, 2018). N-glycosylation does not however deter the occurrence of O-glycosylation because O-glycosylation usually takes place on glycoproteins that have mainly been N-glycosylated in the endoplasmic reticulum.

It has been widely accepted lately that protein *N*- and *O*-glycosylation systems exist in both eukaryotic and prokaryotic organisms. As over 70% of the eukaryotic proteins are believed to be glycosylated, the magnitude of prokaryotic glycosylation will be tougher to guess. Based on the variety of recently discovered prokaryotic glycosylated proteins, it is apparent that glycosylation in these organisms is the norm rather than an exception (Dell *et al*, 2010). With the information available, over two-thirds of all eukaryotic proteins are predicted to be glycosylated (Apweiler *et al*, 1999), but similar estimation is not offered for prokaryotic glycoproteins due to limited information in this area. This is mainly due to the large variety of glycan structures and crucial glycosylation processes that in most cases are accompanied by the absence of genetic manipulation tools.

2.2.2.1.3 C-glycosylation

C-glycosylation comprises a different process of glycosylation because the reaction produces carbon-carbon linkage rather than the carbon-nitrogen or carbon oxygen interactions observed in the others. C-mannosylytransferase enzymes link the C1 of mannose to the C2 of the indole ring of tryptophan (De Beer *et al*, 1995). Mammalian proteins such as RNase2, interleukin-12 and properdin have been found to possess this type of linkage (Hess & Hofsteenge, 1999; Spiro, 2002). While generally little is known about the biological role of C-glycosylation, current research focus is on the production of C-glycosylated molecules by bacteria, plants and insects in drug discovery, because of their resistance to metabolic hydrolysis (Beilen & Li, 2002; Brazier-Hicks *et al*, 2009; Li *et al*, 2013; Zeng *et al*, 2011).

2.2.2.1.4 Glypiation

The covalent attachment of glycosylphosphatidylinositol (GPI) anchor to proteins in the cell membrane is known as glypiation. It occurs post-translationally and is mostly found on the surface of archaea and eukaryotic glycoproteins (Kobayashi *et al*, 1997). The GPI consists of a phosphoethanolamine linker which binds to the target protein C-terminus, a glycan core and a phospholipid tail that attaches the structure to the membrane. Differences in the lipid moiety and sugar residues of the tail confers a distinctive modification which leads to signal transduction, immune recognition and cell adhesion (Vainauskas & Menon, 2006).

2.2.2.1.5 Phosphoglycosylation

The attachment GlcNAc, Man, Xyl, and Fuc linked sugar to serine or threonine residues through phosphodiester bonds is referred to as phosphoglycosylation (Spiro, 2002). It is a post-translational modification restricted to parasites such as *Trypanosoma* and slime molds (Haynes, 1998). Being the most abundant PTM to be used in making proteophosphoglycans (PPGs) in some parasitic species like *Leishmania*, phosphoglycosylation is important for promoting parasite aggregation in the host and also protection against host complement reactions (Sacks *et al*, 2000). As with the case of N-glycosylation, the enzyme phosphoglycosyltransferase is responsible for the transfer of assembled phosphoglycans from a membrane-bound molecule. The structure and enzyme however differs across species (Haynes, 1998).

A comprehensive understanding of the effects of glycosylation on structure and function in protein is often lacking due to the absence of specific homogeneous glycopeptides/glycoproteins for analysis. These are mainly hard to derive from natural sources in adequate quantities and to resolve this, the glycobiology research community should make the development of various bioengineering, enzymatic, chemical and chemoenzymatic methods for the production of homogeneous samples a main research goal (Schäffer *et al*, 2017).

2.2.3 Post-glycosylation modifications

Apart from the different types of glycosylation that can take place on the same protein, glycans can be altered to produce additional variants of glycoproteins. Some of these alterations could include acetylation, sulfation and phosphorylation. For example, some glycoproteins, proteoglycans and glycolipids contain sulphated carbohydrates. They are specifically useful in molecular recognition processes (Yu & Chen, 2007).

The direct interaction of glycan structures with binding proteins is essential in protein stabilization or for masking core glyco-conjugate and carrying out biological functions. The main intermediaries of this process are the different terminal capping residues on *N*-glycans, *O*-glycans and glycosphingolipids and negatively charged sialic acid residues. It has been widely studied and well established that these terminal residues influence the action of glycans (Meng *et al*, 2013). Sialic acid-containing structures perform vital functions in different physiological processes like cellular recognition and communication in vertebrates. To avoid attack or recognition by mechanisms of the hosts' immune system, it mimics sialylated host cell surface carbohydrate structures which is believed to be a crucial virulence factor in bacteria (Yu & Chen, 2007).

Studies on recombinant human sialylated Erythropoietin (EPO) produced in Chinese Hamster Ovary (CHO) cells suggest it has a half-life of around 3 hours compared to the 2 minutes observed for the recombinant glycosylated erythropoietin counterpart produced within the same host (Fukuda *et al*, 1989). Recombinant proteins in CHO cells show a substantial difference in glycosylation that was detected with higher sialylation, which could lead to a reduction in antibody-dependent cell-mediated cytotoxicity (ADCC) activity of an antibody (Croset *et al*, 2012; Scallon *et al*, 2007).

2.2.4 Role of Glycosylation in nature

Protein molecules can be effectively diversified by modifying their properties and glycosylation due to glycans' characteristic structural differences. Some of the most important roles of glycans include their ability to modulate immune responses, take part in pathogenic interactions, function in the regulation of protein turnover and also function as recognition markers (Lis & Sharon, 1993). Majority of the knowledge and information on microbial protein glycosylation is based on S-layers studies of archaea and bacteria. In the past few decades, glycosylated surface additions like flagella and pili were considered in other organisms except in bacteria. Many of the defined bacterial glycoproteins are surface exposed which gives the modified proteins important roles in pathogenicity. As demonstrated in a recent study, protein glycosylation plays vital roles in protein assembly, adhesion, solubility, antigenic variation, protective immunity and in protection against proteolytic cleavage (Christine & Brendan, 2005).

The use of three-dimensional structure analysis has been able to elucidate the role of *N*-glycosylation on protein structure and function and has showed that the core glycosylation molecular mechanisms can be explained in various ways. Through positive interactions, *N*-glycans can serve as molecular binders for amino acids around glycosylation sites thus creating a stable protein structure (Hui Sun *et al*, 2015). *In silico* folding studies of engineered SH3 domain types glycosylated at different sites on the protein's surface resulted in observed thermal stabilization arising from the addition of the polysaccharide chains at different sites and glycosylation positions (Dalit & Yaakov, 2008).

2.3 Comparing and contrasting glycosylation in the 3 life domains

Protein glycosylation is common in all life domains. The earliest discovery of a general N-glycosylation system in *Campylobacter jejuni* gave rise to the rapid progress that has been made in understanding prokaryotic glycosylation (Christine & Brendan, 2005). S-layers, pilins, flagellins and a selection of cell surface and secreted proteins that have been identified in adhesion and biofilm formation are the most understood prokaryotic glycoproteins. Based on the type of glycan linkage to the modified protein, protein glycosylation has been divided into five classes. N-, O-glycosylation and glycosylphosphatidylinositiol have especially been well studied biochemically while little information is available on C-glycosylation and phosphoglycosylation. N-glycosylation is the only glycosylation type that has been lengthily studied in archaea (Jarrell *et al*, 2014) while in both pathogenic and symbiotic bacteria, new general O-glycosylation systems were recently discovered (Anne *et al*, 2010). With the high conservation of the process across all three domains, it is easier differentiating the basic principles of each glycosylation pathway and establish general theories of N-linked protein glycosylation (Aebi, 2013) with a table briefly highlighting this below:

Table 2.3: Comparison of N-glycosylation in the three domains.

Characteristic			Domain		
			Archaea	Bacteria	Eukarya
1	Oligosaccharyltransferase -mediated N-glycosylation	Cytoplasmic assembly of sugars to form an oligosaccharide precursor	Attached to phosphate	Attached to pyrophosphate	
		Fate of Lipid-linked oligosaccharide (LLO) after assembly	Flipped across the cytoplasmic membrane to position the LLO on the exterior surface of the cell.	Flipped from cytoplasm to face the plasma membrane in Gram negative bacteria.	Flipped from cytoplasm to face the lumen of the ER.
		Mode of oligosaccharide transfer to protein acceptor	"*En bloc*" from the lipid carrier onto the acceptor protein in a step catalysed by the enzyme oligosaccharyltransferase		
		Catalytic subunit	AglB (Stt3 homologue)	PglB (Stt3 homologue)	Stt3
		Recognized Sequon(s)	N-X-S/T (X not P), N-X-N/L/V (X not P)	D/E-Z-N-X-S/T (Z and X not P), (*Campylobacter*), N-X-S/T (X not P), (others)	N-X-S/T (X not P)
		Composition	Single subunit		Multimeric complex, single subunit

2	Lipid-linked Oligosaccharide	Lipid carrier	Dolichol phosphate	Undecaprenol phosphate	Dolichol phosphate
		Isoprene units	Variable (8-12)	Typically 11 but varies between 9-12	Variable (14-21)
3	Flippases	Transfer protein involved	AglR; other(s) likely	PglK	Rft1; multiple flippases likely
		Mechanism	Unknown	ATP dependent	ATP independent
4	N-glycans	Linking Sugar	GlcNAc, GalNAc, glucose, other hexoses	HexNAc, Diacetyl-bacillosamine	GlcNac
		Diversity	Extensive	Limited	Conserved 14-sugar glycan in higher species.
		Multi-branched and possibility of modification	Yes	No	Yes

2.3.1 Glycosylation process/defined modification (co- or post)

One of the most complicated processes in protein engineering is the co- or post-translational addition of sugar residues to protein (Spiro, 2002). Bacterial glycoproteins are altered predominantly on the asparagine amino acid (*N*-glycosylation) or the Serine/Threonine residues (*O*-glycosylation). However, these contrasts eukaryotic glycosylation in which *N*-glycans are arranged onto a lipid carrier before it is transported to the acceptor protein and *O*-glycans are assembled on the acceptor protein. Bacterial glycosylation has different processes and carbohydrate structures present. Glycosylation in prokaryotes occurs post-translationally while glycosylation in eukaryotes occurs both post-translationally and co-translationally, (Latousakis & Juge, 2018).

2.3.2 Recombinant glycoprotein production

Although generally occurring in eukaryotes and archaea, N-linked glycosylation hardly occurs in bacteria. The glycan linkage to glycoprotein depends on the protein amino acid sequence and the host organism used for protein expression. N-glycosylation differs between species and depending on the species, the N-linked glycan variants synthesized are also different. In eukaryotes, the core glycan can be rearranged to produce various N-glycan structures like: fucose, mannose, galactose, N-acetylgalactosamine, neuraminic acid, N-acetylglucosamine and other monosaccharides (Mizukami *et al*, 2018). In eubacteria and archaea N-glycosylation, the entire glycan structure produced is similar to the basic structure produced in eukaryotes without the variations peculiar to eukaryotic type glycans (Taylor, 2006).

2.3.2.1 Recombinant glycoprotein production factories

Glycosylation capacity varies significantly in eukaryotes and different mammals. Choosing the most compatible host for recombinant glycoprotein production is key to efficient protein production. Numerous factors must be considered such as the hosts' main characteristics, production costs, product efficacy, safety, stability, biochemical composition, and the hosts' capacity for processing and translating the RNA transcript (Doran, 2000; Çelik & Çalık, 2012). The different production platforms and their focal characteristics are highlighted below:

2.3.2.1.1 Bacteria

Recombinant insulin production in the bacterial expression system *E. coli* was first approved in 1982 (Ghaderi *et al*, 2012). Subsequently, the technology has progressed to be used in the production of numerous commercially approved non-glycosylated proteins such as enzymes, monoclonal antibodies and cytokines. Conventionally, recombinant glycoproteins were not produced in bacterial expression systems because of the lack of the enzyme machinery required to produce mammalian-like glycosylation in bacteria. This view is however changing with recent advancements and the successful engineering of the N-glycosylation machinery in *Campylobacter jejuni* into *E. coli* (Jaffé *et al*, 2014; Schwarz *et al*, 2010; Strutton *et al.*, 2018). With recent substantial developments in *E. coli* glycoengineering, further research outcomes will give rise to a cost-effective production process which solves the issue of current yields being insufficient for commercial uses (Jaffé *et al*, 2014).

Desirable characteristics of bacteria as a production platform include very fast cell growth, minimal and inexpensive media, an established and easy genetic modification template as well as a high recombinant protein yield. However, the post translational modifications in this system are limited and the high risk of contamination (endotoxins) are some of the downsides to using bacteria as a production host. (Mizukami *et al*, 2018).

2.3.2.1.2 Yeast

Over the years, the expression of recombinant proteins in yeast has been widely explored. The ease of culturing, rapid growth and ability to achieve higher densities, widely understood glycosylation pathways and ability for fermentation scale up for use in industrial processes (Ghaderi *et al*, 2012). Minimal and inexpensive media, low risk of contamination, easy and well established genetic modification protocol and high recombinant protein yield are also some desirable characteristic traits for glycosylation in yeast (Mizukami *et al*, 2018). Yeast has been used for recombinant protein expression because of its ability to properly fold protein and the ease of purifying the secreted protein in extracellular medium (Nielsen, 2013). The most commonly used yeast species for protein expression are *Saccharomyces cerevisiae* and *Pichia pastoris* in vaccine, insulin, human serum albumin and recombinant human platelet derived growth factor production (Meehl & Stadheim, 2014).

The glycosylation potential of yeasts makes it yield hypermannosylated N-glycans that affects protein half-life negatively when administered and can cause immune reactions in human beings (Ghaderi *et al*, 2012). There are cases of recombinant protein produced in yeasts by genetically modifying and removing the native genes responsible for the production of hypermannosylated glycans (Hamilton & Gerngross, 2007). A glycoengineered strain of *Pichia pastoris* for N-glycosylation (GlycoFi) was acquired by Merck in 2006 to merge with its expertise in *Saccharomyces cerevisiae* biopharmaceutical production (Gardasil®) to develop improved low cost products compared to what can be obtained from mammalian cell lines (Beck & Reichert, 2012).

2.3.2.1.3 Plant cells

Plant cells can manufacture complex protein and glycoproteins and produce healthy cell cultures. Like with yeast and bacteria, they can also be cultured in basic media and easily scaled up as well. The N-glycan structure in plants is very similar to that of human-like glycans thereby creating increased interest in its use for biopharmaceuticals production and use in the biotechnology industry (Paul &

Ma, 2011). The production of recombinant antibodies in plant cells and recombinant glycoprotein yields is however a limitation of this system as levels obtained are smaller compared to yields in mammalian (CHO) cell systems.

Characteristics of plant cells as a production factory include inexpensive and minimal media, low risk of contamination, moderate cell growth, easy genetic modifications resulting in low to moderate yield of recombinant protein products. This system can also perform glycosylation (with the likelihood of producing plant-specific glycoforms). However, achieving protein sialylation in this system is difficult (Mizukami *et al*, 2018).

2.3.2.1.4 Insect cells

The use of insect cells as a production factory often yields moderate cell growth. Media requirement is however complex and expensive as well as being prone to viral contamination. It however possesses moderate genetic modification potential and a high recombinant protein yield. In this system, post translational modification is possible with the production of insect cell specific glycoproducts and highly mannosylated products. Protein sialylation is also difficult in this system (Mizukami *et al*, 2018).

2.3.2.1.5 Mammalian cells

The most common mammalian cell lines for recombinant protein production are the Chinese hamster ovary (CHO), baby hamster kidney (BHK-21) and murine myeloma cells (NS0 and Sp2/0). Some of the desirable traits/characteristics of CHO cells include that it facilitates scale up, allows for gene amplification, improved selection and has stronger expression units (Mizukami *et al*, 2018).

Recombinant protein production in mammalian cells is characterized by the use of complex and expensive media, high risk of viral contamination and slow cell growth. It however possesses a moderate and well-established genetic modification pathway. The system can post translationally modify recombinant proteins with human-like glycosylation and moderate to high range of recombinant protein yield. It can also be used in the production of immunogenic non-human epitopes (Mizukami *et al*, 2018).

An assortment of reaction factors such as the host cell line, culture process, protein structure and the extracellular environment affects the recombinant protein glycoforms yields (Butler, 2006; Hossler *et al*, 2009). Hence, glycoprotein modification must be monitored and analysed to certify the quality and product acceptability (Zhang *et al*, 2016).

Table 2.4: Advantages and disadvantages of different host systems for recombinant protein expression

Host system	Advantages	Disadvantages
Bacteria (*E. coli, Bacillus sp.*)	Rapid growth rate, easily transformed, direct secretion of protein into culture medium, economical, ability of continuous fermentation.	Codon bias, lack of post translational modification, protein degradation, endotoxin accumulation, reduced/non-expression of target protein, plasmid instability, production of extracellular proteases which could degrade heterologous protein, protein production in insoluble form or as inclusion bodies.
Yeast	Fast growth, low-cost medium, high level of expression, no endotoxins produced, appropriate post translational modification.	Codon bias, hyperglycosylation, inefficient protein secretion into growth medium,
Plant	Easily scaled-up at low cost, high protein yield, localization of protein in different organs at different growth stages.	Target dependent expression levels, undeveloped functional assays.
Insect	High expression levels, post translational modification, appropriate tool for recombinant glycoprotein production.	Demanding culture conditions and lack of continuous expression.
Mammalian	Proper glycosylation and protein folding, appropriate post translational modification and product assembly.	Complex technology, high cost and susceptibility to contamination.

2.4 Glycans

In the three life domains, the sugars added into the glycans and their combination significantly varies. Eukaryotic *N*-glycan core structure is made up of β1,4GlcNAc-β1,4GlcNAc-β1,4Man-α1,3Man-α1,6Man (Van Patten *et al.* 2007, Stanley *et al.* 2009). The three main *N*-glycan types in eukaryotes are created by varying this core structure and subsequently building upon it (see Figure 2.2 A). (1) When more mannose sugars are added the branching mannose residues of the core glycan, Oligomannose glycans are formed (2), when an assortment of sugars like GlcNAc, Gal and sialic acid are built on mannose residues, with, Complex type glycans are formed (3) and when one of the two branching mannose residues on the core has a complex type "antenna" and the other one has only mannose residues attached making an oligomannose branch, Hybrid type glycans are formed (Stanley *et al.* 2009, Corfield and Berry 2015).

In both the archaeal and prokaryotic domains of life, a wider variety of glycans has been identified. There is a longer list of saccharides being included and no core structure between these species. With its core structure present, eukaryotic glycans are considered homologous in nature compared to archaeal and prokaryotic domain glycans which are heterologous (Schwarz and Aebi 2011). The archaeal domain contains a wider variety of glycans with the addition of amino acids into the glycan structure (Chaban *et al.* 2006, Kelly *et al.* 2009). This diversity of glycans is emphasized by the presence of sugars not seen in the eukaryotic domain being present in the archaeal and bacterial domains, like the bacillosamine residue seen in the *N*-glycans of *C. jejuni* (Young *et al.* 2002) and the presence of a 6-sulfoquinovose subunit in the archaeal *Sulfolobus acidocaldarius* species isolated from the Yellowstone National Park that has optimal growth conditions of acidic pH 2-3 and a temperature of 75°C (Hettmann *et al.* 1998, Zähringer *et al.* 2000). The existence of this archaeal species at such extreme conditions could explain the diversity in the structure and content of their glycans (Figure 2.3).

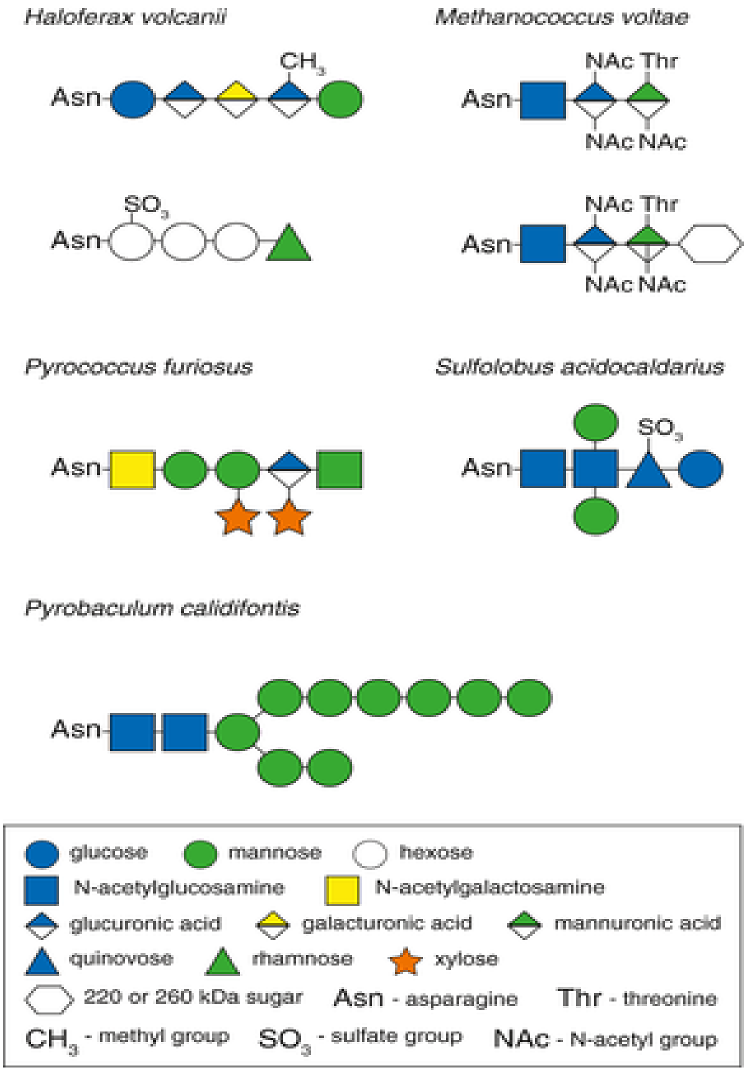


Figure 2.3 Examples of archaeal-assembled N-glycans. Structure of N-glycans linked to target proteins *Haloferax volcanii*, *Pyrococcus furiosus* and *Pyrobaculum calidifontis* shown (Eichler, 2020).

2.4.1 Production of uniform glycoforms

Glycoproteins produced in living cells usually contain a complex mixture of glycoforms, with differences in both the oligosaccharide structures and glycosylation sites. A key challenge in studies that aim to understand the activity and properties of structurally and site-specifically defined glycoforms and therefore, the development and optimization of glycoproteins for biotechnological applications is the lack of control. While significant breakthroughs have been recorded in glycoengineering bacterial, yeast and mammalian cells, a universal technique for preparing user-defined glycoforms from cells remains a problem and the possibility for understanding or exploiting synergistic interactions between multiple, distinct glycans on a single protein remains largely uncharted (Lin *et al.*, 2020).

New developments in chemical and chemoenzymatic methods for in vitro construction of homogeneous glycoproteins have enabled the synthesis and study of diverse glycoproteins with rigorously defined glycan structures. For example, total chemical synthesis has been used to produce human erythropoietin and test the function of each glycan by assembling constituent peptides and glycopeptides. However, total chemical synthesis has only been successfully applied to a few proteins and is particularly inefficient for larger proteins. The chemoenzymatic method is now important for remodelling glycans or installing defined glycans on proteins that are first modified with monosaccharides (Wang and Amin, 2014).

Glycoproteins often contain multiple glycosylation sites, each with distinct glycosylation structures that can synergistically interact to affect protein functions (Schriebl *et al.*, 2006). There is a significant need for methods that can site-specifically control glycosylation at multiple sites so that glycoproteins with defined combinations of glycans and the interactions between them can be studied and optimized to engineer precise or multifunctional glycoprotein. Lin & colleagues developed a strategy to site-specifically control the glycosylation of four sites within a single protein based on the conditionally orthogonal specificities of N-glycosyltransferase (NGT) variants to install monosaccharides at unique acceptor sites (Figure 2.4). NGTs are a class of enzymes that post-translationally modify an asparagine residue (at the canonical N-X-S/T acceptor site) with an N-linked glucose from uracil-diphosphate-glucose (UDP-Glc) sugar donor (Schwarz *et al.*, 2011; Naegeli *et al.*, 2014; Cuccui *et al.*, 2017).

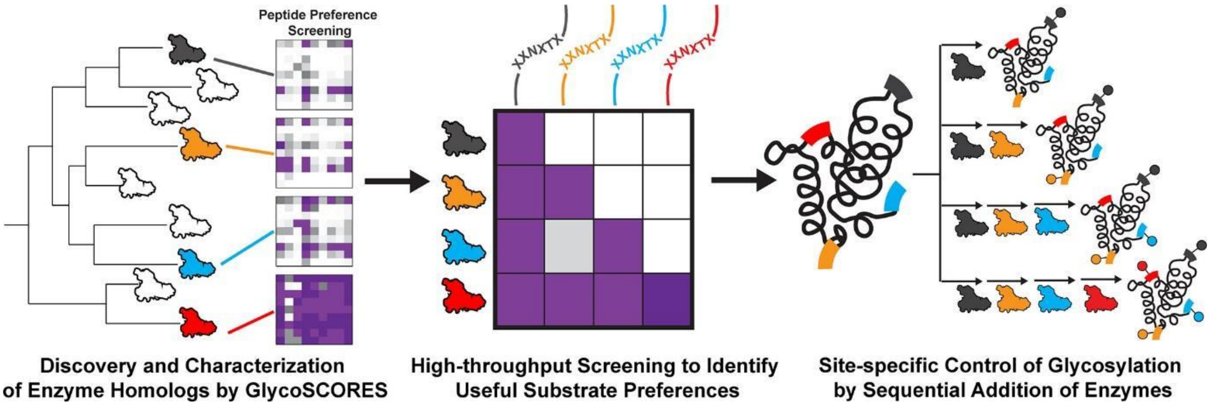

Peptide Preference Screening
xxNXTX
xxNXTX
xxNXTX
xxNXTX
Discovery and Characterization of Enzyme Homologs by GlycoSCORES
High-throughput Screening to Identify Useful Substrate Preferences
Site-specific Control of Glycosylation by Sequential Addition of Enzymes

Figure 2.4 Site-specific control of glycosylation by sequential enzymatic addition of glycans (Lin *et al.*, 2020).

2.4.2 Production of eukaryotic glycan moieties in prokaryotes

2.4.2.1 Overview of heterologous protein expression in *E. coli*

The basis for the earliest structural and functional investigations into proteins was the amount/abundance of the protein. Proteins with distinct stability and solubility profiles such as casein, albumin and haemoglobin were the focus (Gileadi, 2017). Due to the inadequate yields obtained from natural sources, the main aim/goal of biotechnology has shifted to the development of different methods for synthesis in heterologous systems (Ferrer-Miralles *et al*, 2015). In the past few years, clearer understanding into the mechanisms involved in the production of recombinant proteins has grown significantly through intense research (Roslyn, 2014). Recombinant protein production in microbial systems transformed biotechnology and from the first human protein obtained from *E. coli* in 1976 (Itakura *et al*, 1977), the research progresses with the development of various tools and techniques to further increase the capacity to achieve more with this system (Rosano & Ceccarelli, 2014). *E. coli* has thus become the most popular recombinant protein expression system because of the large amount of protein expression tools available. It is used regularly with about 30 % of all recombinant therapeutic proteins presently accepted and almost 60 % of recombinant proteins synthesized in this system. Asides the fact that *E. coli* is a suitable host to produce correctly folded, globular proteins from eukaryotes and prokaryotes (Rosano & Ceccarelli, 2014), the low cost, fast growth, ease of handling and target protein high yield coupled with the general understanding of *E. coli* genetics is also a significant reason it is widely used (Correa & Oppezzo, 2015). Flexibility within the system has been proven in large-scale protein expression trials which shows less than 50 % of bacterial proteins and less than 15% of non-bacterial proteins can be expressed in *E. coli* in a soluble form (Braun & Labaer, 2003).

Some of the disadvantages of expressing heterologous proteins in *E. coli* include misfolding and aggregation. This leads to large deposits of biologically inactive inclusion bodies in the cytoplasm (Singha *et al*, 2017). Other disadvantages include the inability for disulphide bond formation, protein degradation from insufficient expression or insufficient mRNA translation or codon bias in *E. coli* (Fakruddin *et al*, 2013).

Table 2.5: Recombinant gene expression in most commonly used host cells (Roslyn, 2014)

Year	All host cells	*E. coli*	Yeast (*S. cerevisiae* & *P. pastoris*)	Insect cells	Mammalian cells
1980	0	0	0	0	0
1985	0	0	0	0	0
1990	12	9	1	2	0
1995	37	26	2	2	3
2000	50	35	2	6	6
2005	121	103	6	7	2
2010	172	131	15	6	9
2013	128	94	16	5	5

E. coli studies show that the TAT, SRP, or SecB-dependent pathways yield fruitful results for targeted recombinant protein production. It is usually however impossible to ascertain that all recombinant protein molecules will be translocated through a single target pathway. The SRP and SecB-dependent pathways have been reported to be simultaneously involved in single protein targeting, which more or less indicates a level of overlap within the systems. Competition has also been suggested between the Sec- and TAT-dependent protein translocation pathways and it has also been reported that the Sec pathway substrates can be exported by the TAT system in Sec-deficient conditions. Due to the comprehensive studies on the SecB-dependent pathway compared to the others, secretion of most recombinant proteins has been mainly through this system (Mergulhão *et al*, 2005). The table below highlights some examples of the SecB-dependent pathway targeted recombinant proteins.

Table 2.6: Recombinant proteins secreted in *E. coli*

Protein	Signal sequence	Promoter	Secretory amount	Site of Secretion	Scale	Reference
PhoA	Enx	trc	5.2 g/l	Periplasm	Fermenter, 6 l, OD 150	(Choi *et al*, 2000)
PhosD	PelB	*T7*	1.3 mg/l	Medium	Shake flask, 500 ml, OD 3	(Zambonelli *et al*, 2003)
scFv antibody	PelB	*lambda*	160 mg/l	Medium	Fermenter, 4 l, OD 50	(Mukherjee *et al*, 2004)

Staphylokinase	OmpA	*tac*	15 µg/ml	Periplasm	Shake flask, 250 ml, CDW 0.5 g/l	(Lee *et al*, 1998)
scFv multimers	PelB	*lambda*	1 mg/l	Periplasm	Fermenter, 10 l, OD 14	(Bayly *et al*, 2002)
TPA derivatives	PelB	*lac*	29.6 µg/l	Medium	Shake flask, 100 ml, OD (n/a)	(Tayapiwatana *et al*, 2001)
Immunotoxins	PelB	T7	0.6 g/l	Periplasm	Shake flask, 1 l, CDW 4 g/l	(Barth *et al*, 2000)
Antifreeze peptide	OmpA	*tac*	16 mg/l	Medium	Shake flask, 200 ml, OD 2	(Tong *et al*, 2000)
hGCSF	Exl	*trc*	3.2 g/l	Periplasm	Shake flask, 50 ml, OD (n/a)	(Jeong & Lee, 2000)
Hirudin	Asparaginase	*tac*	60 mg/l	Medium	Shake flask, 1 l, OD 6	(Tan *et al*, 2002)
Human proinsulin	SpA	spA	1.2 mg/l	Periplasm	Shake flask, 25 ml, OD 2	(Mergulhao *et al*, 2000)

CDW—cell dry weight; Enx—endoxylanase from *Bacillus* sp.; Exl—*Bacillus* sp. signal peptide; hGCSF—human granulocyte colony-stimulating factor; OmpA—outer membrane protease A; PelB—pectate lysate from *Erwinia carotovora*; scFv—single chain variable fragment; Spa—protein A frfom *Staphylococcus aureus*;

Note: Scale information is included when it is available from the cited reference. When it is not available, it is indicated as (n/a).

2.4.2.2 Glycosylation in *E. coli*

Modifying protein with carbohydrates is common in all life domains and offers tools and methods for the regulation of different cellular processes which includes protein folding, signal transduction, targeting, stability, cell-cell and virus-cell interactions and host immune responses. Initially, glycosylation was thought to exclusively be of eukaryotic origin. It however has been well established that protein glycosylation also occurs in Bacteria and Archaea. Although the discovery of bacterial protein modification systems has increased significantly, the pathway in *Campylobacter jejuni* still remains the most widely studied one in which over a decade ago, a general *N*-linked protein glycosylation (*pgl*) system was first described. More than 60% of periplasmic and membrane-bound proteins are known to be *N*-glycosylated with a conserved heptasaccharide. The non-specificity of the PglB OTase in *C. jejuni* combined with its ability to function in *Escherichia coli* when used with enzymes from the lipopolysaccharide O-antigen biosynthetic pathways and different GTases of bacterial and eukaryotic origin provides avenues of producing defined sugar structures on specific acceptor proteins (Nothaft & Szymanski, 2013).

With the bacterial N-linked glycosylation machinery discovered in *Campylobacter jejuni,* a new area of research opened up for producing glycoproteins in prokaryotes. Wacker and colleagues successfully transferred a functional glycosylation machinery from *C. jejuni* into *E. coli (Wacker et al, 2002b)*. Thus a highly adaptable "plug and play" system was developed for the creation of choice glycans. Genes and enzymes needed for sugar biosynthesis were identified and the use of bacteria became a practicable option in the production of recombinant glycoproteins (Jaffé *et al*, 2014).

The first eukaryotic sugar – GlcNAc was added to an *E. coli* strain with the putative glycosyltransferease pgl2 cluster by Schwartz and colleagues. The resulting glycoprotein was then purified and partially digested to remove the glycan from the protein leaving the sugar residue still attached. This combined *in vivo* and *in vitro* processing steps led to the creation of the eukaryotic glycan core which has consequently become the basis of eukaryotic type glycan production in *E. coli* (Schwarz *et al*, 2010).

To achieve the core eukaryotic glycan in a completely in vivo system, extensive engineering was used to remove some existing genes and add eukaryotic glycosyltransferases. Attaching sugars in the right formation was possible by expressing yeast (*S. cerevisiae*) glycosyltransferase in *E. coli* to produce $Man_3GlcNAc_2$. The strain engineering could increase the cell stress already experienced when foreign proteins are being produced which could in turn affect the process glycosylation efficiency (Valderrama-Rincon *et al*, 2012b).

Naegeli and colleagues characterized cytoplasmic N-glycosylation in detail through the functional transfer of alternative N-glycosylation machinery from *Actinobacillus pleuropneumoniae* into *E. coli*. The pathway which is facilitated by a soluble N-glycosyltransferase (NGT) takes place in the cytoplasm. From the experiment, autotransporter adhesins were identified as the preferred protein substrate of NGT *in vivo*. A relaxed peptide substrate specificity was discovered in analysis of the modified sites in *E. coli* although the preferred acceptor sequon was still N-X-(S/T). The ability to glycosylate heterologous proteins which led to a novel route for engineering of N-glycoproteins in bacteria was validated by NGT (Naegeli *et al*, 2014).

2.4.2.3 Current technologies

E. coli became fit for recombinant protein production with the successful transfer of N-glycosylation machinery into it and have thus been explored for the manufacture of new glycosylated enzymes/proteins for industry. Current challenges of *E. coli* glycosylation include strain engineering restrictions such as the requirement for knocking-out the *waaL* gene to achieve effective glycosylation, glycosylation sequence specificity which could alter amino acid sequences and ultimately have negative effects in therapeutic protein production.

Some recombinant proteins are expressed in either the periplasmic or extracellular compartment of *E. coli* cells. The translocation of these in *E. coli* compartments require their designs to incorporate both expression and translocation with N-terminal signal peptides so that an expressed protein can be translocated to these environments through the host translocon (Schlegel *et al*, 2013). Secretory signal peptides (SPs) which are sequence motifs for target protein translocation to the endoplasmic reticulum and golgi membranes in eukaryotes. Most proteins after translocation are secreted into the culture supernatant. Absence of secretion machinery means the proteins are mostly accumulated in the cells as aggregates or inclusion bodies (IBs). However, through the use of signal sequences or cell engineering, protein expression can now be directed to either the periplasmic space or the extracellular environment of *E. coli*. Table 2.7 highlights examples of *in vivo* glycoengineered protein studies in *E. coli*.

Table 2.7: Example of *in vivo* glycoengineering protein studies in *E. coli*

Protein	*E. coli* host cell	Conditions	Quantification method	Glycosylation Efficiency (%)	Reference
F8	SCM3	Shake flask (batch), L-arabinose (2 g/L), Induction OD_{600} 0.5 for 4 h or overnight	Western blot	40	(Schwarz *et al*, 2010)
AcrA	SCM3	Shake flask (batch and fed batch), L-arabinose (2 g/L), Induction OD_{600} 0.5 for 4 h or overnight	Western blot, Absorbance (280 nm), anion exchange & size exclusion chromatography	-	(Schwarz & Aebi, 2015)
ScFv 3D5	SCM6	Shake flask (batch), L-arabinose (2 g/L), Induction OD_{600} 0.5 for 4 h	Western blot, absorbance (280nm), anion exchange & size exclusion chromatography	20*	(Lizak *et al*, 2011)
AcrA	CLM24	Shake flask (batch), L-arabinose (2 g/L), Induction OD_{600} 0.5 for 3 h	RC/DC Assay, western blot and pSRM	47	(Pandhal *et al*, 2011)
AcrA	CLM24	Shake flask (batch), L-arabinose (2 g/L), Induction OD_{600} 0.5 for 3 h	Western blot	25	(Pandhal *et al*, 2012)
Maltose-binding protein & ScFv 13-R4	MC4100	Shake flasks (batch), L-arabinose (2 g/L) induction	Western blot	<1	(Valderrama-Rincon *et al*, 2012a)
Fc of IgG1, RNA-seA & hGHv	MC4100	Shake flasks (batch), L-arabinose (2 g/L) induction	Western blot	<1	(Valderrama-Rincon *et al*, 2012a)

* Small volumes had a glycosylation efficiency of 40% which increased to 90% as the glycosylation consensus sequence was changed (Chen *et al.*, 2007) and a flexible region added. NB, efficiency value did not scale up to 5 L cultures.

Signal peptide optimization is essential for efficient expression and secretion of small peptides and antibody fragments in *E. coli*. Difference in the expression-secretion rates of desired protein has been observed when minor modifications have been made in the c-region of the natural signal sequence from SAP1 (*in silico* predictions) (Borrero *et al*, 2011). Overall recombinant protein yield in *E. coli* can also be enhanced by optimizing the signal sequences (Klatt & Konthur, 2012).

Considerably improved amounts of recombinant proteins have been produced in the extracellular space of *E. coli* cells via a new secretory platform developed by Wacker called ESETEC. Through this efficient and innovative system, a cost-effective downstream process was developed as a result of the secretion of the native conformation and properly folded recombinant products into the extracellular space of the cultured cells (Figure 2.5). The purified yield of biologically active recombinant ScFv and Fab was recorded at 3.5 g/L and 4.0 g/L respectively with Wacker's ESETEC secretion technology compared to 0.5 mg/L to 400 mg/L obtained using other cloning strategies, strains and culture conditions (André *et al*, 2013).

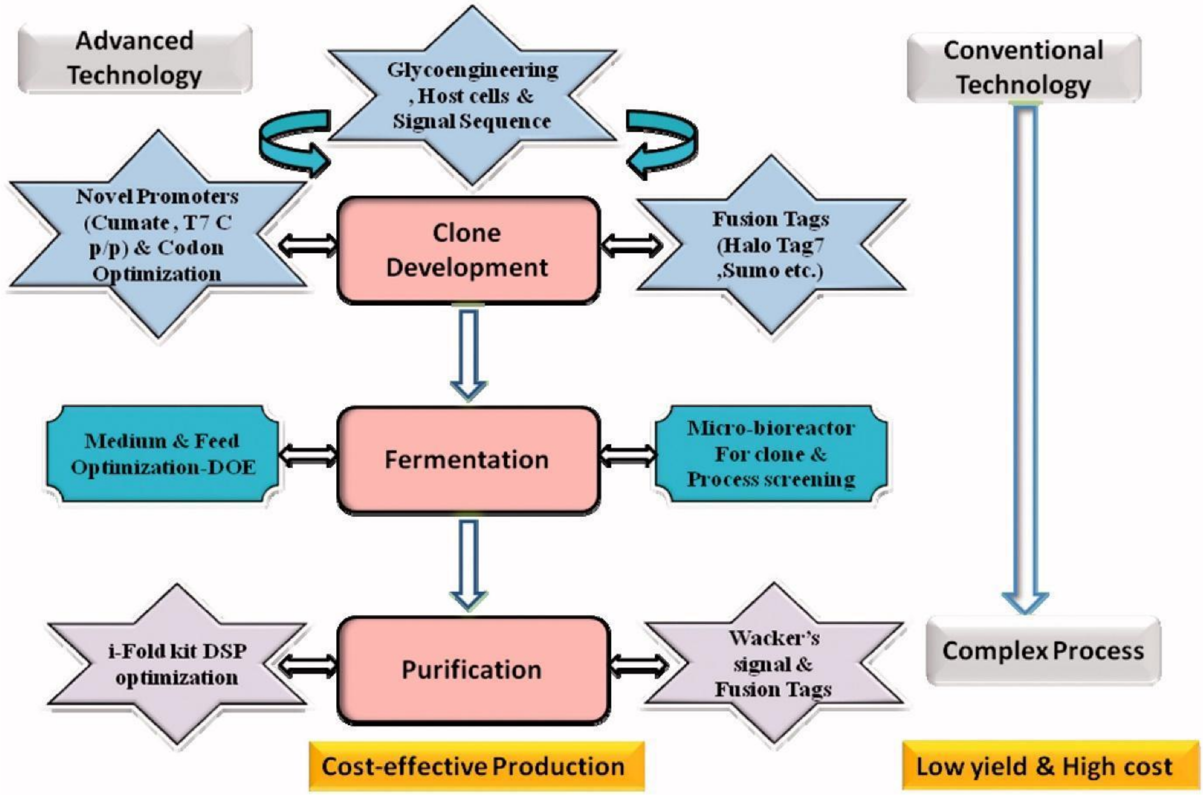

Advanced Technology
Glycoengineering ,Host cells & Signal Sequence
Conventional Technology
Novel Promoters (Cumate , T7 C p/p) & Codon Optimization
Clone Development
Fusion Tags (Halo Tag7 ,Sumo etc.)
Medium & Feed Optimization-DOE
Fermentation
Micro-bioreactor For clone & Process screening
i-Fold kit DSP optimization
Purification
Wacker's signal & Fusion Tags
Complex Process
Cost-effective Production
Low yield & High cost

Figure 2.5 Outline of technology innovation in *E. coli* for process and yield improvement (Gupta & Shukla, 2016).

2.4.2.4 Competing function of WaaL in *E. coli*

E. coli CLM24 strain, a W3110 single gene knockout variant is primarily used for bacterial glycosylation studies. The O-antigen ligase *waaL* gene which is involved in the lipopolysaccharide (LPS) of gram negative bacteria (an essential part of its outer membrane vital to maintaining its structural integrity) is knocked out (Nikaido and Vaara 1985). The three components that make up the LPS are lipid A, an O-antigen and a core oligosaccharide. The O-antigen on the periplasmic side of the cytoplasmic membrane is recognised by the waaL protein ligase and is attached to the lipid A core before being exported to the cell surface (McGrath and Osborn 1991, Han *et al.* 2014). The *C. jejuni* glycosylation machinery produced glycan is also recognised as a substrate and will be exported to the cell surface if still functional within the system. Exporting this glycan to the cell surface will deplete the glycan pool available for PglB present to attach to the protein of interest. To this end, it is advantageous to knock this gene out.

In chapter four, a strain of *E. coli* W3110 with the *waaL* gene still functional is used in order to exploit the O-antigen presenting pathway and test the relative glycan production capabilities of different engineered strains through the assessment of glycans displayed on the cell surface.

2.4.2.5 Periplasmic expression of target protein

Glycans in the bacterial system are assembled on the cytoplasmic face of the cytoplasmic membrane and flipped to the periplasmic face by PglK in an ATP-dependent process (Lehrman 2015). The protein of interest must be available in the periplasm alongside the glycan and PglB for glycosylation to take place.

There are different export systems present in bacteria (Papanikou *et al.* 2007). Three of these have been highly characterised and widely used to target the protein of interest into the periplasm. They are the Sec transport system, the SRP (signal recognition particle) pathway, and the TAT (twin-arginine translocation) export system. The Sec system will be used in these studies due to its ease of use as it only requires the placement of a leader sequence at the C-terminus of the target protein for translocation.

The most characterised system within bacteria which is also present in archaea and the endoplasmic reticulum of eukaryotic cells is believed to be the Sec system (Kudva *et al.* 2013). Its role across the three domains of life is to transport secretory proteins across the inner membrane of the cell and position membrane proteins within the inner membrane. Cytosolic proteins like SecA or SRP that help in the recognition of the signal sequences located at the C-terminus of the polypeptide chain and initiate the target protein's translocation are key for the pathway to function (Koch *et al.* 2003). The most frequently utilised leader peptide for periplasmic translocation for targeting the protein of interest to the periplasm within this system is the *Erwinia carotovora* pectate lyase B leader peptide - pelB (Thie *et al.* 2008). This pelB sequence is made up of a 22 amino acid sequence (MKYLLPTAAAGLLLLAAQPAMA) and is cut after translocation by signal peptidases contained in the membrane (Lei *et al.* 1987, Paetzel *et al.* 2002).

The disadvantage of using the SEC export system over the TAT export system is that the protein of interest is delivered in an unfolded state into the periplasm (Nilsson *et al.* 1991). Because bacterial glycosylation is believed to be a completely post-translational modification, it would be preferential in theory to use the TAT pathway as it delivers the protein in its folded state (DeLisa *et al.* 2003).

2.4.2.6 Glycosylation machinery

The experiments within this thesis were carried out using the glycosylation machinery pgl2, situated on the pACYC backbone. This was originally from *C. jejuni* and it produces and transfers the hexasaccharide glycan GalNAc-α1,4-GalNAc-α1,4GalNAc-α1,4-GalNAc-α1,4-GalNAc-α1,3-GlcNAcβ1 to the consensus sequences. The machinery is situated on a pACYC plasmid backbone which confers chloramphenicol resistance. The genes contain the necessary glycosyltransferases, oligosaccharyl transferase and flippase required and are placed under a constitutive promoter which therefore removes the need for an inducer molecule. Consequently, this means that the glycans are continuously being built up and flipped into the periplasm.

The pACYC(pgl2) machinery was chosen as the model system to allow for easy analysis as this plasmid yields moderately high glycosylation efficiency compared to the eukaryotic machinery. A successful transfer of the glycan produced to the consensus sequence within the target protein is feasible as it has been established as a substrate for the OST available. Extensive characterization and studies into the glycan produced provides a good model that can be analysed using various mass spectrometry methods as well as allowing easier optimisation for method development. This means efficiency values and titres can be compared as long as absolute quantification was specified.

2.4.2.7 Target protein

AcrA was used as the model protein for the studies in this thesis. Being the first glycoprotein to be transferred from *C. jejuni* into *E. coli,* it has subsequently been used in many bacterial glycosylation studies thus making it the most studied bacterial glycoprotein and an ideal model protein. AcrA has two *N*-linked glycosylation sites - present at both asparagine 105 and asparagine 255. These are naturally occurring consensus sites and are located in flexible regions of the protein and available for glycosylation. The existence of the two sites also allows a look into this protein's glycan saturation and formation of a di-glycosylated product in this system. This is significant because recombinant therapeutic glycoproteins usually have more than one glycosylation. For example, the SWISS-PROT study which showed that most glycoproteins comprise 1.9 glycans per protein (Apweiler *et al.* 1999). Recombinant expression of AcrA in the *E. coli* CLM24 strain, its genetic sequence was cloned onto a pEC plasmid backbone. An araBAD promoter which requires L-arabinose for induction of transcription was used to control the expression. At the N-terminus of the genetic sequence, is the 22 amino acid pelB leader peptide for translocation purposes and at the C-terminus is a 6 x histidine tag to help in the purification of the 39 kDa protein. The amino acid sequence of AcrA is shown in Figure 2.6, with the relevant characteristics highlighted.

```
MKYLLPTAAAGLLLLAAQPAMAMHMSKEEAPKIQMPPQPVTTMSAKSEDLPLSFTYPAKLVSDYDVIIKPQVSGV
IVNKLFKAGDKVKKGQTLFIIEQDKFKASVDSAYGQALMAKATFENASKDFNRSKALFSKSAISQKEYDSSLATFNN
SKASLASARAQLANARIDLDHTEIKAPFDGTIGDALVNIGDYVSASTTELVRVTNLNPIYADFFISDTDKLNLVRNTQ
SGKWDLDSIHANLNLNGETVQGKLYFIDSVIDANSGTVKAKAVFDNNNSTLLPGAFATITSEGFIQKNGFKVPQIG
VKQDQNDVYVLLVKNGKVEKSSVHISYQNNEYAIIDKGLQNGDKIILDNFKKIQVGSEVKEIGAQLEHHHHHH
```

Figure 2.6 AcrA protein amino acid sequence. The PelB sequence required for export with the Sec export system is highlighted in red. The two glycosylation consensus sequences recognized by PglB are highlighted in yellow with the asparagine residues for glycan attachment underlined. The 6 x histidine residue tag for purification purposes is highlighted in green.

AcrA in this thesis is simply a model glycoprotein for study purposes. In nature however, it is part of an *E. coli* multi-drug efflux complex with AcrB and TolC (Zgurskaya and Nikaido 1999), that is partly responsible for pumping antibiotics out of the cell into the medium. Its role in *C. jejuni* is also similar although the two homologues have a 29.01% sequence identity as the native *E. coli* variant does not contain any consensus bacterial glycosylation sites. The absence of glycosylation sites on the *E. coli* form is of importance because proteins competing for the glycan supply which the machinery is capable of building are not desired.

Glycosylated AcrA can be produced using the pACYC(pgl) machinery within the outlined system.

2.4.3 Current industrial examples

The N-linked glycan modification of protein structure and function has been extensively studied and its effect on the physicochemical properties, folding, secondary structure, stability and recognition events of proteins cannot be overemphasized. Recent researches have revealed the effect of N-glycosylation on enzyme activity, protein targeting and substrate specificity. Some of these studies removed recognised N-glycosylation sites of various glycosylated enzymes to examine and understand the exact effect each N-glycan has on regulating enzyme secretion, activity and substrate specificity

(Skropeta, 2009). Besides stability and protection against proteolysis, many protease enzymes have confirmed the significant effect of glycosylation on catalytic activity. Enzyme turnover rates, specificity and binding affinity as well as substrate recognition have been known to change from the addition of glycans (Goettig, 2016). In the case of unglycosylated RNase B variants of RNase A for example, the addition of carbohydrate chains has contributed to enzyme thermostability and eventually sterically inhibited the oligomerization process to significantly affect the enzyme activity (Gotte *et al*, 2003).

Enhanced heterologous protein expression in *Pichia* species is affected by glycosylation. Research has shown that N-glycan addition to recombinant elastase through the insertion of an N-glycosylation sequon at the right locations can stimulate expression. N-glycosylation effects on protein folding and secretion was also confirmed to be site specific (Han & Yu, 2015). With the engineering of *Bacillus subtilis* xylanase A glycosylation pattern for expression in *P. pastoris*, the enzymes thermostability increased compared to that of the unglycosylated enzyme expressed in *E. coli* (Fonseca-Maldonado *et al*, 2013). Stimulating protein glycosylation at its native or underglycosylated sites has also often led to discovery of enhanced features. N-glycans positioned within the loop region or near aromatic amino acids in proteins have been known to confer stability (Culyba *et al*, 2011; Greene *et al*, 2015).

From the thermal inactivation studies on glycosylated *Aspergillus oryzae* Cutinase (AoC) expressed in *P. pastoris,* it was revealed that thermal aggregation inhibition was higher than in the unglycosylated AoC enzyme (Shirke *et al*, 2017). Comparable findings were recorded when the effect of glycosylation on stabilizing Leaf and Branch Compost Cutinase (LCC) for PET hydrolysis was analysed. It was proven that glycosylating the enzyme was able to slow down LCC aggregation by raising the temperature for thermally induced aggregation by 10^{o}C thereby increasing its kinetic stability and improving its catalytic ability in the PET recycling process (Shirke *et al*, 2018). The effect of glycosylation on the biocatalytic properties of Hydroxynitrile lyase (HNL) enzymes from passion fruit plant which was recombinantly expressed in both *E. coli* and *Pichia pastoris* showed when compared that the N-glycosylated HNL enzyme produced in *Pichia* exhibited better thermostability, solvent tolerance and pH stability than the aglycosylated variant produced in *E. coli* (Nuylert *et al*, 2017).

Multi-functional cutinase enzymes which breakdown different substrates such as polyesters, insoluble triglycerides and soluble esters. They also have the ability to catalyze esterification and transesterification. It is therefore used potentially in the textile, detergent, ester synthesis, and environmental protection industries (Su *et al*, 2015).

2.5 Detection methods, quantification and analysis of glycoprotein

The detection, characterization and quantification of glycans and synthesized glycoproteins is challenging. Various techniques that have been recently developed for carbohydrate and glycopeptide synthesis have sustained substantial growth in glycoprotein folding studies (Dalit & Yaakov, 2008). Confirming and quantifying protein glycosylation as well as measuring the amount of glycoprotein yield in other production hosts require the use of some of the following methods:

2.5.1 Western blot analysis

Western blotting is normally used to quantify and calculate glycosylation efficiency in most prokaryotic glycosylation studies. This is usually based on the premise that the target protein has an engineered terminal histidine tag that binds specific antibodies for identification. These antibodies bind to the protein for easy recognition during immunoblotting from a rather complex protein sample. The protein mass can increase by as much as 1.4 kDa (depending on the glycan added) when a glycan is added to an asparagine residue in a prokaryotic sequence (Scott *et al*, 2012). The expected mass variation with glycan addition should result in the appearance of multiple bands in SDS PAGE and glycosylation can be confirmed through western blotting. The quality of protein produced can be measured by the intensity of the bands formed and as such the efficiency of glycoprotein production can be calculated by comparing this to the aglycosylated form/bands.

Disadvantages of this method include the fact that band intensity can be affected by other factors like transfer and development time that can make comparison across different samples' membrane blots difficult (Aebersold *et al*, 2013).

2.5.2 Sugar specific Affinity reagent - Lectins

The molecular weight difference between glycosylated and aglycosylated protein can be used to confirm glycosylation from western blots. This result could however be challenged as antibody binding is not glycan specific. Lectin screening can better confirm protein glycosylation with a higher degree of confidence.

Unlike in western blots, the lectin protocol is not specific to the protein of interest. Lectins bind and interact with sugar molecules in the glycans (Hirabayashi, 2004). Most lectins are not very useful as analytic tools due to their low affinity levels *in vitro*. To increase lectins affinity to targeted glycans and achieve higher binding to purified glycoproteins, many biotinylated lectins are joined together by streptavidin bonds in a process known as lectin multimerization. This resulted in glycoprotein detection at lower concentrations than would have been possible with monomeric detection (Cao *et al*, 2013).

This combined with western blotting is used to detect glycosylation. The most comprehensive method of analysis is still Mass spectrometry. The main challenge however remains the difficulty in interpreting the data generated.

2.5.3 Mass spectrometry

Protein glycosylation can be studied via two main approaches either as glycopeptide analysis or glycan-based analysis. The glycopeptide analysis requires the glycan to remain covalently attached to the peptide while the glycan-based analysis requires the removal of the glycan from the protein before analysing the structure and content. To analyse glycan content via mass spectrometry, previous studies in prokaryotes have shown the glycopeptide approach to be most effective in glycosylation research.

Tandem mass spectrometry (MS/MS) is more effective for glycoprotein analysis since it can be used in analysing a mixture of glycoproteins. However, this might be difficult in cases of low glycosylation efficiency as only a low amount of glycopeptides will be available for analysis.

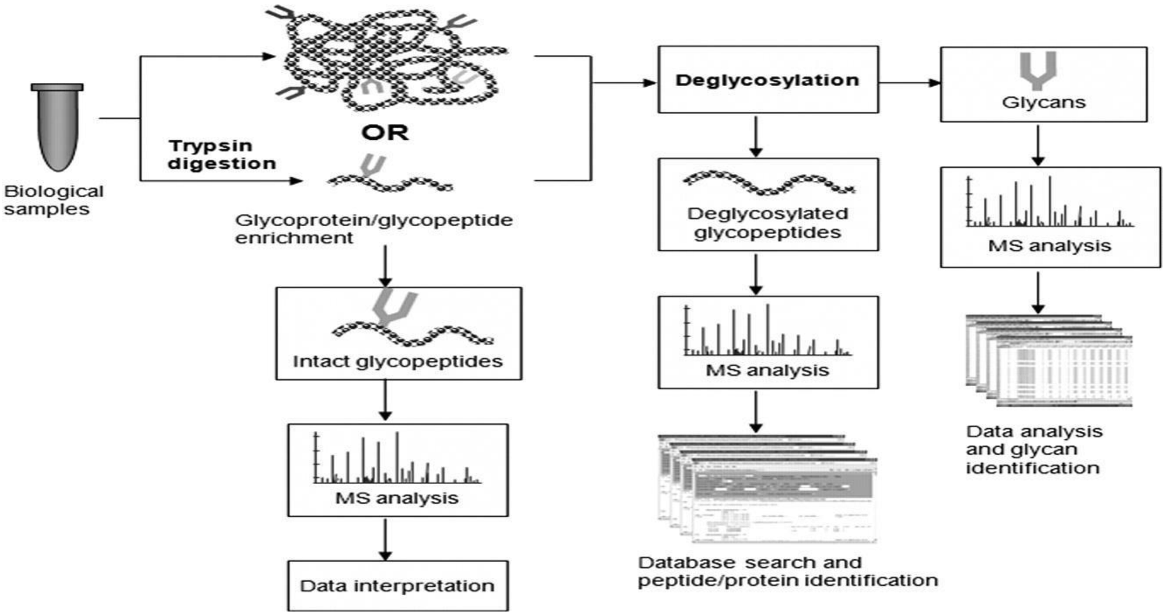

Biological samples
Trypsin digestion
OR
Glycoprotein/glycopeptide enrichment
Intact glycopeptides
MS analysis
Data interpretation
Deglycosylation
Deglycosylated glycopeptides
MS analysis
Database search and peptide/protein identification
Glycans
MS analysis
Data analysis and glycan identification

Figure 2.7 Strategies of mass spectrometry based glycoproteomic analysis (Pan *et al.*, 2011)

The systematic enzymatic or chemical release of glycan attachments has been identified as the best method for characterizing protein glycosylation. The released glycans are then analysed through reductive amination with aromatic or aliphatic amines or permethylation (Morelle & Michalski, 2007). Few studies have measured and stated the glycoprotein titres produced for western blot quantifications. For quantification purposes however, a method where the heavily labelled form of the target protein is produced has been developed (Pandhal *et al*, 2013).

Chapter 3: Materials and methods

3.1 Standard buffers, reagents and media

These stated methods have been used throughout this thesis. All buffers and media were prepared using deionised filtered water (dH_2O) and (Qiagen) nuclease-free water (nfH_2O) was used for all DNA preparations. High performance liquid chromatography (HPLC) grade solvents were used throughout the experiments. Growth media was sterilized for 20 minutes or filtered using 0.2 micrometre (µm) sterile syringe filters. Before the addition of antibiotics, all media was cooled to < 55°C. Chemicals and reagents were purchased from Sigma / Merck unless otherwise stated. For media and buffer recipes not provided in the text, see appendix. Specific information on primer design has been stated in the relevant experimental chapter. The working antibiotic concentrations used in this study were as follows: kanamycin (35µg/L), chloramphenicol (35 µg/L) and ampicillin (100 µg/L).

Table 3.1. Table showing plasmid studies carried out in previous studies, highlighting the antibiotic resistance cassette and origin of replication present.

Plasmid	Origin of replication	Antibiotic resistance cassette	Source
pEC(acrA)	ColE1	Ampicillin	(Wacker *et al.*, 2002a; Wacker *et al.*, 2002b)
pYCG	p15A	Chloramphenicol	(Valderrama-Rincon *et al.*, 2012)
pACYC(pgl2)	p15A	Chloramphenicol	(Schwarz *et al.*, 2010)

Table 3.2. Specific genotypes and strains used in this study

E. coli Strain	Genotype	Recombinant protein plasmid	Source (This work)
W3110	F^- λ^- *rph-1 INV(rrnD, rrnE)*	pYCG	Chapter 4
7HS2	EMS-Mutated variant of W3110 (above)	-	Chapter 5
2EWL7	Twice EMS-Mutated variant of W3110 (above)	-	Chapter 5
CLM24	Variant of W3110 (above) with the addition of Δ*waaL*	pEC(acrA)	Chapter 6

3.2 Molecular biology methods

3.2.1 Polymerase chain reaction (PCR)

Polymerase Chain Reaction were mostly carried out using either the Phusion$^{®}$ polymerase kit from NEB (New England Biolabs) or Dreamtaq$^{®}$ from Thermo Fisher Scientific with the primers designed using the SnapGene program and being ordered from IDT custom DNA oligos (Integrated DNA Technologies).

25 or 50 µL reactions were set according to protocol specified in the kit as shown in Table 3.3.

Table 3.3. Phusion High-Fidelity DNA Polymerase reaction mixture

Component	25 µL Reaction	50 µL Reaction	Final Concentration
5x Phusion HF or GC Buffer	5 µL	10 µL	1x
10 mM dNTPs	0.5 µL	1 µL	200 µM
10 µM Forward Primer	1.25 µL	2.5 µL	0.5 µM
10 µM Reverse Primer	1.25 µL	2.5 µL	0.5 µM
Phusion DNA Polymerase	0.25 µL	0.5 µL	1.0 units/ 50 µL PCR

Template DNA	variable	variable	1 pg – 250 ng
nfH2O	to 25 µL	to 50 µL	

The PCR reaction thermocycling condition was as follows, 98 °C for 3 minutes, 35 cycles of a 30 second denaturation step at 98 °C, following another 30 second annealing step at varying temperatures depending on the primer design, with an extension step set at 72 °C and run at 1 kb per 30 seconds which was extended depending on the size of fragment being amplified. Finally, the last extension step was run for 7 minutes at 72 °C before cooling the reaction to 4 °C. PCR reactions were subject to an agarose gel electrophoresis on a 1% agarose gel for 1 hour at 120 mA, with the DNA fragments of the correct size extracted from the gel piece using a Qiagen® kit.

If colony screen PCRs were conducted, a colony was picked and mixed in 50 µL of nuclease free water. 5 µL of this resuspended colony was used as the template DNA, and the initial heating step in the PCR reaction extended to 5 minutes.

3.2.2 Agarose gels

For analysing DNA fragments between the size range of 200-10,000 bp, 1X TAE (Tris Base, Acetic acid, EDTA) buffer was prepared from a 10X stock solution. 1% agarose gel was prepared by dissolving 1.2g Agarose in 120ml TAE buffer and heating on high heat in the microwave for 2 minutes. The molten agarose was left to cool before the addition of 5 µl of ethidium bromide. Agarose was poured into the gel casting rig to the top of the fingers of the comb containing the appropriate wells needed. The gel was left to solidify for approximately 10-20 minutes. The dams from the casting rig were removed and the chamber was filled with a 1X TAE buffer up to the fill line. Gel well-comb was carefully removed. The samples were mixed with the required volume of 5X loading buffer and loaded in the wells. 5 µl of 1kbp Bioline® hyperladder™ (Figure 3.1) was loaded in the first well for size quantification analysis. The gel was run at 120mA for 1 hour to attain separation. Finally, the DNA bands were visualized using a UV doc (GelDoc-It-Imager, UVP).

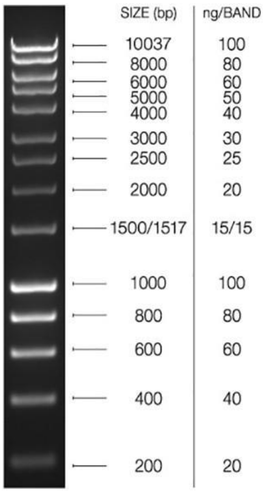


Figure 3.1: Bioline 1 kbp HyperLadder DNA ladder used for DNA agarose gel electrophoresis. Image taken from Bioline website (https://www.bioline.com/us/ on 30/09/2021)

3.2.3 Gel extraction of DNA from an agarose gel

Gel pieces were carefully cut with a sterile razor from the gel using a UV illuminator to visualise bands which was then placed in sterile centrifuge tubes (1.5 mL) before being weighed. Buffer QG (QIAquick® – Gel Extraction Kit) was added to the gel piece that equated to 3x the volume of the gel (100 mg = 100 µL) and incubated at 50 °C for 16 minutes (vortex mixed every 2 minutes) until the gel had completely dissolved. 1 gel volume of isopropanol was added after the gel pieces dissolved and was vortex mixed for 10 seconds. The solution is placed in a QIAquick® spin column before centrifugation at 13,300 x g for 1 minute. The column was then washed with 750 µL of buffer PE (QIAquick® – Gel Extraction Kit). The wash solution was centrifuged to remove any residual buffer. 30 µL of nuclease water was introduced to the column and the QIAquick® spin column placed in a clean centrifuge tube. The column was left to stand for 5 minutes and centrifuged at 13,300 x g for 1 minute. The resulting solution contained the cleaned DNA from the agarose gel.

3.2.4 Digestions

Digests were set up according to the specific enzymes used based on the type of DNA. A 50 µL reaction was set up containing, 1 µL of each restriction enzyme used to digest 1 µg of DNA, 10 µL of 5 x digestion buffer, with the reaction volume brough to a total volume of 50 µL using nuclease free water. Reactions were left at 37°C for 1 hour. Vector DNA would subsequently have 1 µL of alkaline phosphatase added to the digestion mix post digestion, and left for another hour at 37°C.

Digested DNA was cleaned up using either a Qiagen® PCR clean up kit, or was gel purified post analysis on an agarose gel. 1 µL of digested product was run on an agarose gel for quantification purposes.

3.2.5 Ligations

20 µL was used as a final reaction volume. 2 µL of the 20 µL, comprising NEB 10 x ligation buffer and 1 µL T4 DNA ligase (NEB). Insert and vector DNA were used to make up the reaction volume or 20 µL of nuclease free water. The amount and length of DNA in the reaction was used to calculate the litigation ratios of insert to vector. For most of the reactions, 20 ng was used as a vector concentration, with the insert amount varying depending on the chosen ligation ratio. An insert to vector ratio of 3:1 was used in most instances. If larger fragments were to be inserted (≥ 6 kb), more ratios would be tested ranging from 1:1 to 10:1. Negative control ligations with no insert present were routinely set up to determine the success of ligations post transformations.

Ligations were conducted at room temperature for 1 hour before the reaction mix was transformed into the chosen cell line. If challenges were experienced with the cloning, the ligation step was often optimised by working with various litigation conditions, including; 16 °C incubation overnight, and 4 °C incubation overnight.

3.2.6 Plasmid DNA extraction

The desired Plasmid DNA was extracted using the Qiagen® maxi-prep protocol. 100 ml of overnight culture was harvested at 4 °C using a centrifuge set to 6000 x g for a duration of 15 minutes. Pellets were resuspended in 10 ml of buffer P1 (Qiagen® - Plasmid extraction kit) after the supernatant had been discarded. 10 ml of buffer P2 was added and mixed severally for homogeneity. Tubes were incubated at 20°C for 5 minutes followed by the addition of 10 ml of pre-chilled buffer P3. Solution was mixed thoroughly until it turned colourless and incubated for 20 minutes on ice after which it was centrifuged at 4 °C for 30 minutes at 20,000 xg speed till supernatant was clear. A Qiagen tip 500 was equilibrated with 10 ml of buffer QBT which was allowed to flow through by gravity. Supernatant was added to the tip and allowed to flow through the resin tip by gravity. Tip was then washed twice with 30 ml of buffer QC. 15 ml of buffer QF was used to elute the DNA into clean 50 ml falcon tubes followed by the addition of 10.5ml isopropanol (room temperature) to the eluted solution for DNA precipitation. The mix was centrifuged at 15,000 x g speed for 30 minutes and the supernatant was carefully removed while the tube was allowed to sit for 2 minutes before aspirating residual supernatant. 5ml of 70% ethanol (room temperature) was used to wash the DNA pellet which was then centrifuged for 10 minutes at 15,000 x g speed. Supernatant was carefully removed and pellet was air dried for 5 minutes before re-dissolving the DNA in 400µl of TE buffer (pH 8). DNA concentration of the sample was measured using the Nanodrop.

3.2.7 Preparation of chemically competent cells

10 mL of LB (Lysogeny Broth, Tryptone 10 g, NaCl 10 g, Yeast extract 5 g) with or without the appropriate antibiotics, was inoculated with one bacterial colony of the strain of interest picked using

a sterile inoculating loop. This starter culture was left in an incubator at a shaking speed of 180 rpm and temperature of 37°C. 200 mL LB was inoculated using 2 mL of starter culture of the desired cell line. The culture was incubated at 37°C with a shaking speed of 180 rpm, until an O.D 600 nm of 0.5 was reached. The flask was then kept on ice for 10 minutes to chill while swirling every minute. The culture was then decanted into 4 x 50 mL ice cold falcon tubes and the cells were harvested at 4°C spinning at 4,000 x g for 10 minutes. Each of the 4 pellets were carefully resuspended in 20 mL ice cold 100 mM $MgCl_2$. Cells harvesting procedure was repeated. One pellet was then resuspended in 6 mL ice cold $CaCl_2$ and transferred to the next falcon tube until all 4 pellets were suspended in the 6 mL solution. Cells were left on ice for 1.5 hours to become competent. 1.8 mL of 50% (v/v) glycerol was added and gently swirled until mixed thoroughly with the cells. 50 µL competent cells were aliquoted into cold 1.5 mL centrifuge tubes (Eppendorf) and were snap frozen by immersing into liquid nitrogen before being stored at -80°C.

3.2.8 Heat shock transformations

Chemically competent *E. coli* cells were thawed on ice and 5µl of DNA or 10pg - 100ng equivalent of plasmid DNA of choice was added and kept on ice for 30 minutes. The cells were subjected to 30 second heat shock by placing in a water bath at 42 ºC temperature and then placed on ice for 2 minutes. 1 ml of pre-warmed LB medium was added and the cells were allowed to incubate at 37 ºC for 1 hour. 10 µl of the resulting cells were plated on agar plates with appropriate selective antibiotics and incubated for 12-16 hours at 37 ºC.

3.2.9 Electroporation

10 mL starter culture was achieved from a single picked *E. coli* colony inoculated into 10 mL of LB media supplemented with the required antibiotics. The culture was incubated at 37 °C for 3 hours at 180 rpm. Post growth the cells were spun at 4,500 x g for 10 minutes at 4 °C. The resulting pellet was washed with ice cold 10% glycerol and spun again with the same centrifugal conditions. This washing and spinning cycle was repeated thrice. After the last spin, the pellet was gently resuspended in 100 µL of the 10% ice cold glycerol. 1-2 µL of the desired plasmid was placed with the cells and gently mixed before incubating on ice for 1 minute. This solution was then placed in an ice cold electroporation cuvette ensuring no air bubbles were present after transfer. The solution was subjected to a 1.8 kV pulse for 5 milliseconds giving a field strength of 12.5 kV/cm in the 0.1 cm cuvette. Following that, the pulse 1 mL of SOC media warmed to 37°C was added and the solution mixed by pipetting up and down. The 1 mL culture was transferred to a sterile 5 mL centrifuge tube and incubated at 37 °C for a minimum of 1 hour at 180 rpm. After incubation, the culture was spun down at 2,000 x g for 2 minutes in a centrifuge and the majority of the supernatant removed leaving

approximately 100 µL. The pellet was resuspended in the left over supernatant and spread on an agar plate with the required antibiotics.

3.2.10 DNA sequencing

Gel extracted DNA fragments and sequences from PCR amplifications were sent to the Core Genomics sequencing facility at Sheffield University or GENEWIZ®, for confirmation of sequence construct. 10 µL of 100 ng/µL of plasmid DNA was required, along with 10 µL of 1 pmol/µL of the primer per reaction. Results were checked using the FinchTV programme to search for any faults in the screened sequence.

3.2.11 DNA quantification

DNA concentration was roughly estimated by comparing DNA gel electrophoresis bands to the DNA ladder (Figure 3.1). More accurate DNA measurements were carried out using a NanoDrop™ 1000 spectrophotometer (Thermo Fisher Scientific); 1-2 µL DNA sample was measured per run.

3.3 Bacterial growth and expression

3.3.1 Making starter cultures

10 ml of Luria-Bertani (LB) medium with or without appropriate antibiotics was dispensed in a falcon tube and inoculated with one bacterial colony picked using an inoculating loop under sterile conditions. The culture was left to shake at 180rpm overnight at an incubation temperature of 37^{o}C.

3.3.2 Preparation of glycerol stock

Bacterial starter culture of the desired strain was set to incubate overnight at 37^{o}C shaking at 180 rpm. 50% (v/v) glycerol was prepared. 0.5 ml of the starter culture was aliquoted in a sterile microcentrifuge tube and mixed with 0.5 ml of 50% glycerol. The stock was then stored at a temperature of -80^{o}C.

3.3.3 Growth measurements

In 100 ml of LB culture medium with or without appropriate antibiotics, 1 ml of fresh overnight starter culture was added. The initial optical density at 600nm (OD_{600}) was measured using a spectrophotometer. The culture was then incubated at 37^{o}C shaking at 180 rpm. 1ml aliquots were taken at 30 minute intervals and OD_{600} measured for 8 hours. OD readings were plotted on a graph (Time – X axis vs OD_{600} – Y axis).

3.3.4 Bacterial growth and protein expression

Starter cultures of the bacterial strains of interest were set up and left overnight at 37 °C, 180 rpm, to be used to inoculate 100 mL LB in triplicate the next day. Appropriate antibiotic concentrations for plasmid maintenance were added as required. Cultures were inoculated with 1 mL of the starter culture and incubated at 37 °C, shaking at 180 rpm. When an optical density (O.D) at 600 nm of 0.5 was reached, the protein expression was induced based on the promoter sequence present on the plasmid for example, the addition of 0.2% (v/v) L-arabinose for the *araBAD* promoter located on the

AcrA plasmid. Cells were left to incubate and express the protein for another 4 hours at 30°C 180 rpm. The final OD of the cultures was measured and 40 O.D units' worth was harvested through centrifugation at 4 °C, at a speed of 4,500 x g for 10 minutes. The supernatant was discarded and the pellet stored at -20 °C prior to protein extraction.

3.3.5 Periplasmic protein extraction

To extract the protein, frozen pellets were thawed on ice and resuspended in 1 ml of periplasmic lysis buffer (20% sucrose, 1g/L lysozyme, 30mM Tris-HCl pH 8.5, 1X Halt protease inhibitor complex (Thermo Fisher Scientific)) and left to roll on ice for 2 hours. The soluble protein fraction was collected through centrifugation at 4,500 x g at 4 °C for 10 minutes, with the supernatant being harvested containing the periplasmic protein sample.

3.3.6 Bradford assay

A standard curve was produced using a serial dilution of 0.5 mg/mL bovine serum albumin (BSA) down to a concentration of 10 µg/mL. 20 µL of the protein was mixed with 980 µL of Bradford assay and left to incubate at room temperature for 5 minutes before measuring the OD at 595 nm. 20 µL of the soluble periplasmic protein fraction was measured in the same way and the protein concentration determined from the standard curve.

3.3.7 Nanodrop quantification

Periplasmic protein extract from *E. coli* cells that had been purified using nickel affinity chromatography and processed through buffer exchange columns, was OD measured at 280 nm on a NanoDrop™ 1000 (Thermo Fisher Scientific). The equipment was cleaned with deionised water, and blanked at 280 nm with 2 µL of the solution that the sample protein was suspended in. Once blanked 2 µL sample was placed on the lower measurement pedestal, the sample arm was closed and the measurement taken by selecting the correct application on the associated software. Between samples, the lower and upper pedestal were cleaned by wiping with a clean blue roll. Measurements of each sample were taken twice and the average used for quantification purposes.

3.4 Gel analysis

3.4.1 SDS PAGE

For visualization and quantification of protein, secreted proteins from culture supernatant was analysed using SDS-polyacrylamide gel electrophoresis (PAGE). SDS PAGE was performed using precast NuPAGE® Novex 4-12% Bis-Tris gels (Thermo Fisher Scientific). Samples were prepared according to the composition in Table 3.4 below with appropriate volume of protein sample used.

Table 3.4. SDS PAGE reaction mixture

Component	Volume (µl)
Protein sample	X
LDS sample loading buffer	5
Sample reducing agent	2
Deionized water	13-X
Total Volume	20

The samples were heated in a water bath to 70°C temperature for 10 minutes and left to cool at room temperature. Precast gels were loaded onto the gel apparatus and both chambers filled with 1X MOPS SDS buffer. Samples were loaded into the wells and gel ran for 65 minutes at 200V. 5 µl of EZ-Run pre-stained Rec protein ladder (Fisher Scientific; Figure 3.2) was also run for size analysis of the proteins. After successful runs, gels were further analyzed by Coomassie staining and Western blot.

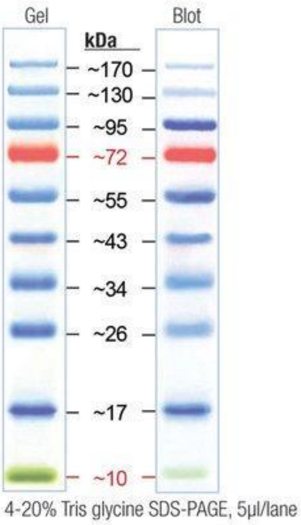


Figure 3.2. Protein ladder used as a marker for SDS-PAGE experiments. EZ-Run prestained Rec protein ladder (Fisher Scientific). Image taken from Fisher website (https://www.fishersci.com/ on 30/09/2021).

3.4.2 Coomassie stain

After running the SDS PAGE, gels were removed from the cassette and immediately placed in 25 ml of InstantBlue™ (Expedeon) for 1 hour. The gel could be left to incubate overnight if the bands were faint. To prevent further development, the gel was washed twice in deionised water and subsequently stored at 4°C in deionized water.

3.4.3 Western blot

The SDS PAGE gel was removed from the cassette and washed in deionised water. Proteins were transferred from 4-12% Bis-Tris SDS-PAGE to the PDVF nitrocellulose membrane using an iBlot® (Thermo Fisher Scientific). The membrane blot was rinsed with 25 ml TBS for 5 minutes at room temperature. The membrane was subsequently blocked in a 25 ml volume of blocking buffer (5% milk powder in Tris-Buffered Saline (TBS) with 0.1% v/v Tween20) for an hour at room temperature. 3 x 5 minute washes of the membrane in 15 ml TBS 0.1% Tween20 (TBST) was followed by overnight incubation at 4 °C in blocking buffer with a His-tag antibody (abcam® Anti-6X His tag® (HRP)) (1:10,000). Following incubation, the excess unbound antibody was washed off with 3 x 5 minute washes in 15 ml TBST. The HRP linked antibody was detected on the blot using 20 ml of ECL Reagent (Thermo Fisher Scientific) and incubated for 1 minute. Pictures were captured using the BIORAD® Imagelab software.

3.5 Purification techniques

3.5.1 Histidine nickel affinity chromatography purification

Purification was carried out with a HisTrap HP Nickel affinity column (GE® Healthcare) that actively binds proteins with a Histidine-tag. The column was washed with 5 column volumes of sterile deionized water and then equilibrated with 5 column volumes of binding buffer (20mM sodium phosphate, 500mM sodium chloride and 40mM Imidazole) pH 7.4. The periplasmic extract was loaded unto the column before 10 column volumes of binding buffer was used to wash. 5 column volumes of elution buffer (20mM sodium phosphate, 500mM sodium chloride and 500mM imidazole) pH 7.4 which contained a higher concentration of imidazole was used to elute His-tagged proteins bound to the column. Five 1ml fractions of the elusions were collected for SDS-gel analysis.

3.6 Mass spectrometry methods

3.6.1. 2-D protein clean-up

100 µL of the sample protein extract was transferred into a 1.5ml Eppendorf tube. 300 µL precipitant reagent was added and mixed thoroughly before incubating on ice for 15 minutes. 300 µL co-precipitant reagent was added and mixture was centrifuged for 10 minutes at 8000 x g. The supernatant was immediately removed carefully. 100 µL co-precipitant was added and centrifuged again at 8000 x g for 10 minutes. Supernatant was removed and 100 µL dH_2O was added and vortex mixed for several seconds. 1 ml of pre-chilled wash buffer and 5 µL of wash additive was added and vortex mixed until the pellet was fully dispersed. Mixture was incubated at -20°C for 30 minutes (vortex mixing every 10 minutes for 20-30 seconds). Centrifuging at 8000 x g for 10 minutes the supernatant was discarded and pellet air-dried for a maximum of 5 minutes before 100 µL lysis buffer was added and the mixture incubated at room temperature to dissolve pellet fully. It was further centrifuged at 8000 x g for 10 minutes to remove any insoluble material and the pellet was stored at -20°C.

3.6.2 Protein reduction, alkylation and in-solution digestion

The pellet from the 2-D protein clean-up was dissolved in 50 µL Urea Buffer (8 M urea; 100 mM Tris-HCl [pH 8.5]; 5 mM DTT) and placed in a sonication bath for 5 minutes until protein suspension became clear. Protein concentration was quantified and ~50 µg protein was transferred to a fresh 1.5 mL protein LoBind Eppendorf tube. Protein samples were reduced by diluting up to 10 µL with Urea Buffer and incubating at 37°C for 30 min. Proteins were S-alkylated by adding 1 µL 100 mM iodoacetamide and incubating in the dark at room temperature for 30 min. 2 µg trypsin endoproteinase LysC enzyme mix (Promega) was added to the protein solution and incubated at 37°C for 3 hours for LysC digestion, after which the solution was diluted with 75 µL 50 mM Tris-HCl (pH 8.5)/ 10 mM $CaCl_2$ and incubated overnight for trypsin digestion. The digestion was stopped via acidification by adding 0.05 volumes of 10% trifluoroacetic acid (TFA) to the peptide solution. The samples were dried by SpeedVac and stored at -20°C.

3.6.3 C18 clean up

All solutions were made up using mass spectrometry grade reagents and pulled through the column by centrifugation at 1,500 x g for 1 minute unless stated otherwise. Pierce® C18 Spin columns (Thermo Scientific) were used to clean Peptides for mass spectrometry analysis. Dried samples were resuspended in 20 µL of 0.5% TFA in 5% ACN, vortexed for a short period and sonicated on ice for 5 minutes. Spin columns were placed into a centrifuge tube and the resin activated with 2 x 200 µL of 50% ACN. Columns were centrifuged and the flow through discarded. 2 x 200 µL 0.5% TFA in 5% ACN were added to equilibrate the column. The sample was loaded onto the column resin and centrifuged.

Flow through was reapplied to the column and passed through the column once more. 2 column washes with 200 µL 0.5% TFA in 5% ACN were conducted. Post washing the column was placed in a clean Lo-Bind centrifuge tube (Eppendorf) and 20 µL of elution buffer (70% ACN) was applied to the resin and drawn through the column. The elution step was repeated collecting the flow through in the same centrifuge tube. Samples were dried in a vacuum centrifuge tube and stored at -20 °C before running on the mass spectrometer.

3.6.4 LC-MS/ MS for proteomics

Peptide sample pellets were thawed and resuspended in 15 µL loading Buffer (97% acetonitrile, 3% H2O, 0.1 % TFA v/ v) and sonicated in a water bath for 5 min until fully in suspension. Following 5 min centrifugation, 2 µL sample (~4 µg) was diluted 1 in 8 with a loading buffer and transferred to a vial for liquid chromatography (LC)-MS/ MS analysis. 500 ng protein sample was analysed by nanoflow LC (Dionex UltiMate 3000 RSLCnano system) coupled online to a Q Exactive HF mass spectrometer (Thermo Scientific). An automated data-dependent switch between full MS and tandem MS/MS scans through stepped collision energy was used in acquisition of peptide spectra.

Chapter 4: A flow cytometric approach to using EMS-induced mutagenesis in *Escherichia coli* for improved mannose production

4.1 Summary

In this chapter, a random mutagenic approach was applied to successfully increase the mannose availability in *E. coli*. The increasing abundance of the tri-mannosyl core structure - $Man_3GlcNAc_2$ in the N-glycan biosynthetic pathway within the glycosylation cell factory to be utilized has previously led to increased and more efficient glycosylation in several organisms. Here, a flow cytometric fluorescence based assay was used to identify *E. coli* cells with increased nucleotide sugar biosynthesis capabilities which could ultimately be available as precursors for the production of the tri-mannosyl core structure. W3110 cells carrying the plasmid pYCG for the subsequent expression and cell surface display of GDP-mannose – a precursor in the biosynthesis of $Man_3GlcNAc_2$ core glycan structure was subjected to random chemical mutagenesis using ethyl methanesulfonate (EMS) and subsequently analysed and sorted for higher fluorescence. Significant increases in the cell surface mannose displayed were observed in mutant sorted high producers regrown in the 2nd generation at 2.4-fold more per 100,000 fluorescence events recorded than the wild-type strain. This work shows the use of flow cytometry screening as a useful tool for investigating the surfaces of glyco-engineered *E. coli* and identifying significant enhancements in the N-glycan biosynthesis pathway within the strain which complements existing strain engineering techniques.

4.2 Introduction

N-glycosylation has been established as a widely homologous process in all life forms. Although bacteria and archaea glycan precursors appear to be heterogeneous, a conserved lipid-linked oligosaccharide structure is observed in eukaryotes. With a large pool of building blocks available to them, archaea display a wide variety in glycosylation pathways and they produce both dolichyl phosphate-linked and pyrophosphate-linked glycans. They can also synthesize more than one type of LLO in the same cell. This diversity is however limited in bacteria where N-glycosylation is restricted to a small number of species. In eukaryotes, the conserved Man_5-$GlcNAc_2$ core units are extended with added units to form Glc_3-Man_9-$GlcNAc_2$. Successive expansion in the late ER and in the Golgi results in the creation of a notable assortment of N-glycans found in eukaryotes, highlighting a different evolutionary origin (Schwarz and Aebi, 2011).

With the transfer of N-glycosylation machinery from *Campylobacter jejuni* into *E. coli* for the production of recombinant glycoproteins, great interest in the field of glycoengineering in bacteria has emerged. This engineering breakthrough subsequently led to the use of *E. coli* being successfully used to glycosylate recombinant proteins with both bacterial and some eukaryotic type N-glycans.

Example of bacterial type N-glycan expression in *E.coli* is the production of glycoconjugate vaccine candidates through the coupling of *Shigella* O1-antigen to either *C. jejuni* CmeA or a toxoid form of *Pseudomonas aeruginosa* exotoxin A with engineered *N*-glycosylation sites (Ihssen *et al.*, 2010) and the *in vivo* synthesis of an eukaryotic trimannosylchitobiose glycan in *E. coli* by expressing four eukaryotic glycosyltransferases and the subsequent transfer (through *C. jejuni* PglB) to acceptor proteins (Valderrama-Rincon *et al.*, 2012).

In order to harness the versatility of this system for the production of new glycoproteins of interest, glycoengineering strategies and tools have been developed to address the challenges encountered in glycoprotein production within the system. While *E. coli* often incurs translational errors, accumulates inclusion bodies, and completely lacks the eukaryotic organelles and machinery necessary to produce fundamental post-translational modifications (Barolo *et al.*, 2020), one major challenge of recombinant glycoprotein production in the system is low glycosylation efficiency. The strategies that have been developed and employed in combating this are divided into two main groups: protein engineering and cell engineering. In protein glyco-engineering, the recombinant glycoprotein is targeted through modifying its DNA sequence before translation, through modifying its subcellular location during translation, or by modifying its glycosylation pattern after translation. Cell glyco-engineering methods however introduce or modify the expression and activity of target glycosylation pathway involved enzymes by either random genetic insertion/manipulation, targeted gene knock-in or knock-out methods or inhibitor interference (Wang and Lomino, 2012; Costa *et al.*, 2014).

In earlier glycoengineering work on *E. coli*, a synthetic heterologous pathway that enables site-specific glycosylation of proteins with a eukaryotic trimannosyl chitobiose glycan - mannose$_3$-N-acetylglucosamine$_2$ ($Man_3GlcNAc_2$) a core structure of all human N-linked glycans was developed (Valderrama-Rincon *et al.*, 2012). This pathway, made up of multiple glycosyltransferases (GTases) from yeast and the oligosaccharyltransferase (OTase), PglB, from *Campylobacter jejuni*, an archetype for bacterial N-linked glycosylation is divided into three separate stages: glycan biosynthesis, membrane translocation of glycans, and glycan transfer onto polypeptide acceptor sequon (Figure 4.1). More recent research centered on this engineered pathway was able to show that sufficient availability of the substrate precursor GDP-mannose was able to solve the problem of poor accumulation of lipid-linked $Man_3GlcNAc_2$ substrate in the glycan biosynthesis stage. This in turn led to an almost 50-fold increase in the levels of $Man_3GlcNAc_2$-containing LLOs which ultimately resulted in a 14% increase in glycosylated acceptor protein (Glasscock *et al.*, 2018).

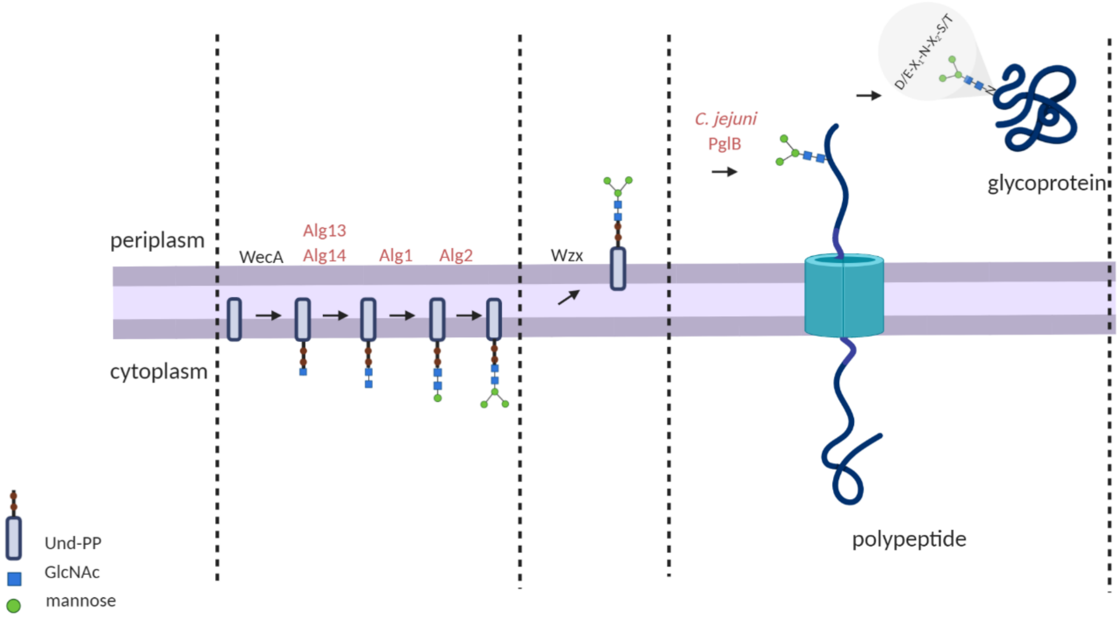


Figure 4.1. Eukaryotic N-glycosylation pathway in *E. coli*. The pathway consists of 3 separate stages; the first consisting of glycan assembly to Und-PP on the cytoplasmic face of the inner membrane through the action of endogenous *E. coli* enzyme WecA and yeast GTases Alg13, Alg14, Alg 1 flipped into the periplasm. The flippase enzyme Wzx is responsible for lipid-linked glycans being flipped unto the periplasmic face of the inner membrane in the second stage and finally the OTase PglB catalyzes the transfer of glycans from Und-PP to exported acceptor proteins' asparagine residues in the periplasm.

In this study, to improve the overall protein glycosylation levels, we sought to increase the amount of GDP-mannose, a substrate precursor for the glycan - $Man_3GlcNAc_2$ biosynthesis which can be produced by *E. coli*. W3110 cell (ancestral strain to a widely utilized glycocompetent strain) carrying the recombinant protein pYCG plasmid was used to investigate increased mannose levels within *E. coli*. Here chemical mutagenesis method was employed to generate random mutants and cell-based fluorescence assay was leveraged on to screen for increased mannose production within the cell.

Fluorescence-activated Cell Sorting (FACS) is a special type of flow cytometry where the physical or chemical characteristics of cells are measured. These measurements are performed while cells are passing in a fluid stream across an illuminated light path. This method uses a laser based technology which allows the quantitative and qualitative analysis of several properties of cell populations from any type of non-fixed tissue or fluid body. Cells are suspended in a narrow fluid system and passes in a single file in front of a detection laser for counting and sorting. Fluorescently labelled cell components are then excited by a laser to emit light at different wavelengths (Weaver, 2000). Fluorescence measured is used to determine the amount and type of cells in a sample. A beam of laser light is directed at a hydrodynamically-focused stream of fluid that carries the cells. Several detectors are carefully placed around the stream, at the point where the fluid passes through the light beam. One of these detectors is in line with the light beam and is used to measure Forward Scatter or FSC while another detector is placed perpendicular to the stream and is used to measure Side Scatter (SSC). Since fluorescent labels are used to detect the different cells or components, fluorescent detectors are also in place. The detectors therefore pick up a combination of scattered and fluorescent light. This data is then analysed and interpreted using the computer software FLOWJO (Givan, 2010).

The potential advantage of using chemical mutagenesis is that it confers mutations which improve the function or expression of a protein by base pair changes. While insertional mutagenesis is restricted to gene disruption and is more convenient for the identification of a mutation site, the main mechanism of mutagenesis by EMS involves guanine alkylation (Sega, 1984). EMS is an efficient mutagen with excellent preservation of viability (Miller, 1972). When guanine interacts with the ethyl group of EMS, O^6 -ethylguanine which is an atypical base is generated which then leads to the replacement of cytosine with thymine as the matching base pair for O^6 -ethylguanine during DNA replication. This results in a point mutation with GC pairs being replaced by AT pairs. Although another popular method for mutant generation in *E. coli* is UV mutagenesis, EMS mutagenesis has been applied to generate *E. coli* mutants (Coulondre and Miller, 1977) and used for this experiment due to equipment availability.

4.3 Specific methods

4.3.1 Reagents, strains and plasmids

The *E. coli* strain W3110 was used for all experiments. The construction of the plasmid pYCG (Figure 4.2) has been described (Valderrama-Rincon *et al.*, 2012).

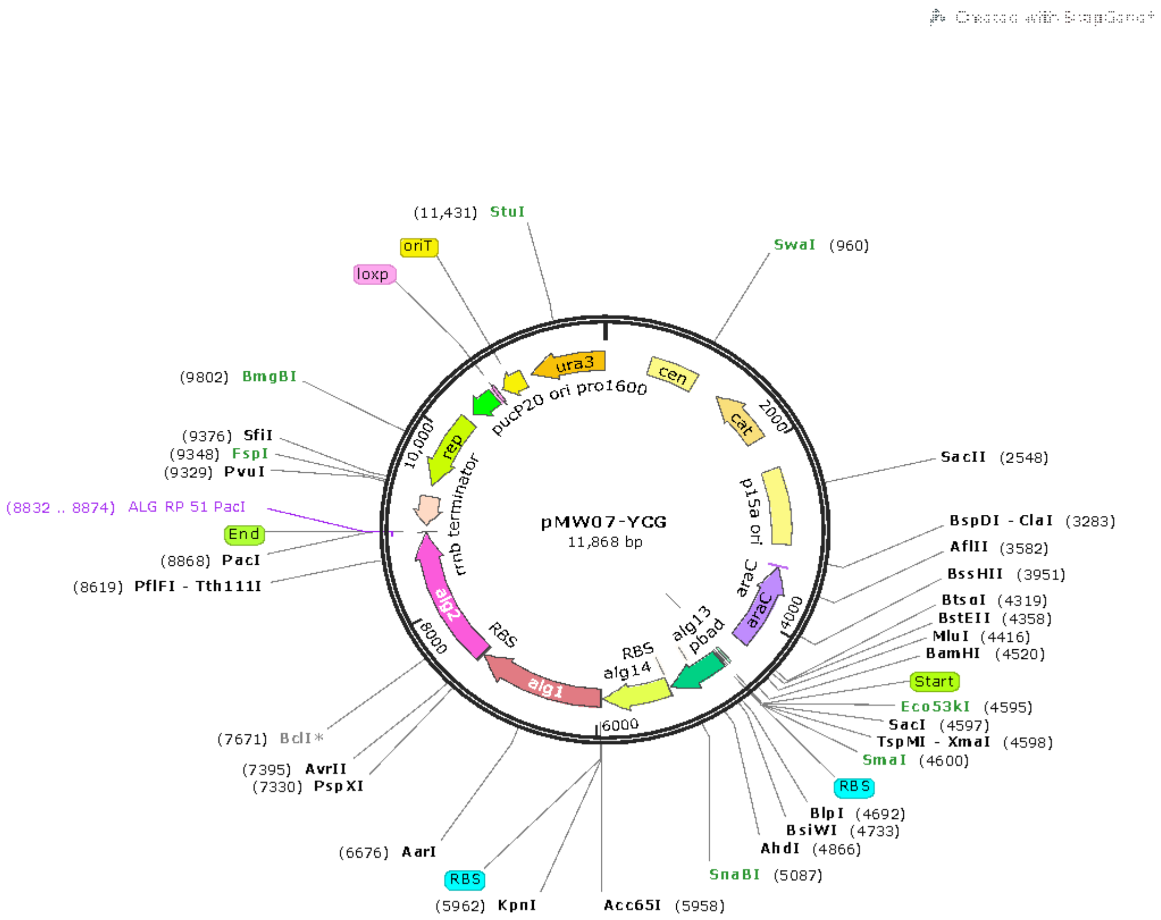


Figure 4.2 Plasmid map of pYCG. Map highlights relevant features, such as a p15a *ori*, the cat gene for Cam resistance and unique restriction enzyme digestion sites. Plasmid vector pMW07 was generated by (Valderrama-Rincon *et al.*, 2012), from pMQ70 as part of the publicly available vector suite. Map created using SnapGene software.

4.3.2 Mannose overexpression

E. coli W3110 cells freshly transformed with the pYCG expression vector were grown in Luria-Bertani (LB) medium (10 g/L tryptone, 5 g/L yeast extract and 5 g/L NaCl). Single bacterial colonies were used to inoculate liquid LB starter cultures containing appropriate antibiotics (chloramphenicol 35 µg/mL) and grown overnight at 37°C temperature in a shaker incubator at 180 rpm. The following day, fresh LB media was inoculated with overnight starter culture at a 1:100 dilutions and grown at 37°C to an optical density at 600nm (OD_{600}) of 0.5 with shaking. The temperature was then decreased to 30°C and after a temperature equilibration period of 5-10 minutes, cell surface mannose production was induced by adding 0.2 % (v/v) L-arabinose and cells were left to incubate and express protein for an additional 4 hours (Figure 4.3).

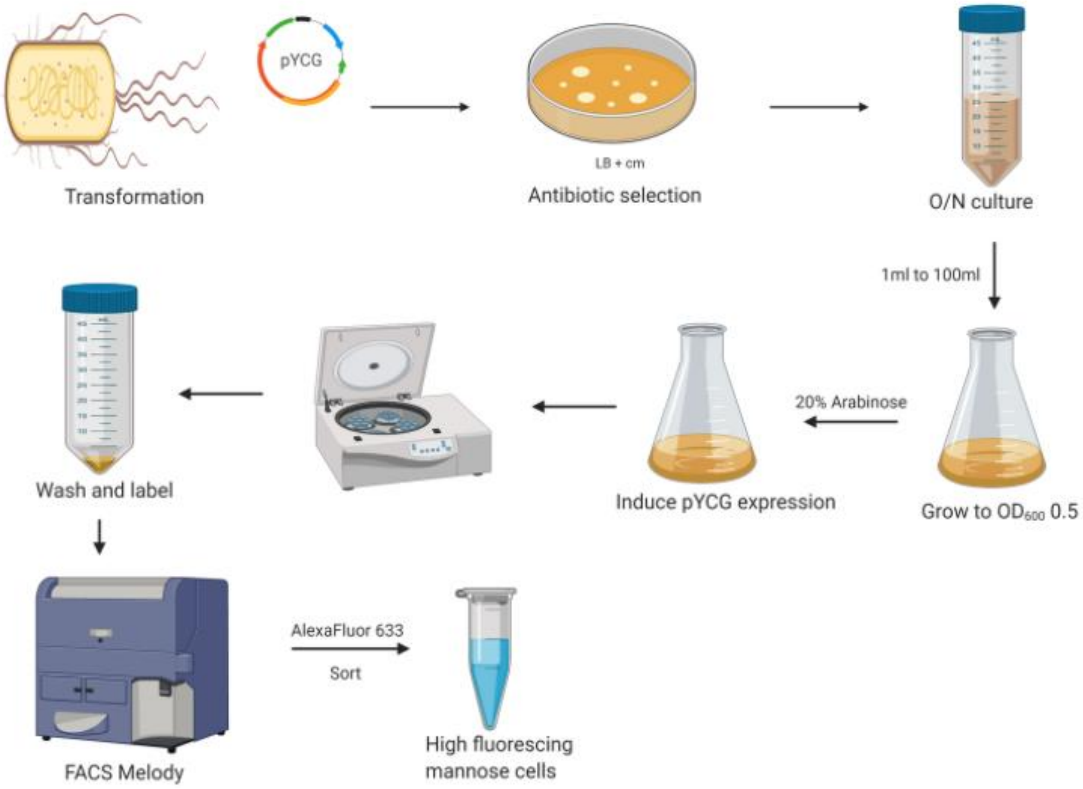


Figure 4.3. Outline of the methodology used to achieve cell surface display of $Man_3GlcNAc_2$ and the process of screening for the high representation of the cell surface glycans. O/N culture refers to inoculated overnight (starter) cultures of *E. coli* from a single colony grown in LB broth at 37 °C for between 12-16 hours shaking at 180 rpm.

4.3.3 Chemical mutagenesis

Following previous method for EMS mutagenesis in *E. coli* (Burns, Allen and Glickman, 1986), cells were grown in LB broth to OD_{600} 0.3, washed twice with PBS then resuspended in PBS to original density. To 2 mL of suspension 45 µl of EMS (10g) was added and the cells were incubated for 45 minutes at 37°C. Cells were washed twice in PBS, resuspended in 2 mL of PBS. 100 µl of resuspension was added to 2 mL of LB and the cells were grown for 3 hours at 37°C and plated on LB plates at different dilutions.

4.3.4 Fluorescent labelling and fluorescence-activated cell sorting

For labelling and sorting, cells were harvested by centrifugation and washed twice in cold PBS (Phosphate Buffered Saline, 137.93 mM NaCl, 2.67 mM KCl, 1.47mM KH_2PO_4, 8.1 mM Na_2HPO_4, pH of 7.4, Invitrogen). Cell pellets were then resuspended in cold PBS buffer to original density. The suspension was then diluted to OD_{600} 0.3. The cells were labelled with 2.5 µg/mL concentration of Concanavalin A, Alexa Fluor 633 conjugate (Thermo Fisher Scientific, Massachusetts, MA, USA) and incubated in the dark for 15 minutes at room temperature with shaking. After incubation, cells were washed twice in PBS and resuspended to original density. The fluorescence of cells expressing mannose on the cell surface was monitored using a Becton-Dickinson FACSMelody^{TM} flow cytometer (see Appendix B, page 134). 500 µl of each sample was measured with 100,000 events recorded. Data were collected for Alexa Fluor-633 fluorescence (632 nm excitation, 647 nm emission) and analyzed with the FlowJo software. For FACS screening, cells were initially gated based on brightness on a side-scatter (SSC-H) versus forward-scatter (FSC-H) plot. Subsequently, cells lying within the approximately 10^5 clones corresponding to the top 1-5% fluorescent events were isolated, grown in liquid LB media containing appropriate antibiotics with repeated rounds of FACS to sort for top mannose producing cells. All experiments have three biological replicates and the Median MEFL values which are representative of the specifically targeted event within our sample population were calculated over replicates.

4.3.5 Confocal microscopy

For fluorescence microscopy, pellet cells were washed and resuspended to original density in PBS. The cells were labeled with 2.5 µg/mL concentration of Alexa Fluor-633 ConA and left to incubate in the dark for 15 minutes with shaking. Labelled cells were washed twice in PBS and 50 µl of sample was pipetted onto a flat slide and covered with a cover slip. Slides were viewed on a Leica Microsystems-SP8 TCS confocal fluorescent microscope using different lens magnifications.

4.4 Results

4.4.1 Cell-surface glycan display for screening $Man_3GlcNAc_2$ levels in living cells

Flow cytometry was used to measure the amount of lipid-linked oligosaccharide displayed as $Man_3GlcNAc_2$ glycans cells on the *E. coli* cell surface. Due to the presence of the O-antigen ligase WaaL which is responsible for the periplasmic Und-PP-linked oligosaccharide transfer unto lipid A, the lipopolysaccharide (LPS) layer of gram-negative bacteria cell surfaces can support engineered oligosaccharides. The LPS transport system is used to transfer oligosaccharides to the cell surface where it can be labelled and readily measured via flow cytometry (Glasscock *et al.*, 2018; Fisher *et al.*, 2011).

In line with observations from previous research (Valderrama-Rincon *et al.*, 2012; Glasscock *et al.*, 2018), W3110 cells transformed with pYCG plasmid when labelled with AlexaFluor-633 ConA conjugate were fluorescent compared to the control cells without the plasmid. The fluorescence assay is based on the studies which suggests fluorophores preferentially bind to internal and non-reducing terminal α-mannose in oligosaccharides while microscopy studies have shown that the ConA binding is visibly localized on the cell surface (Glasscock *et al.*, 2018) as seen in Fig. 4.5.

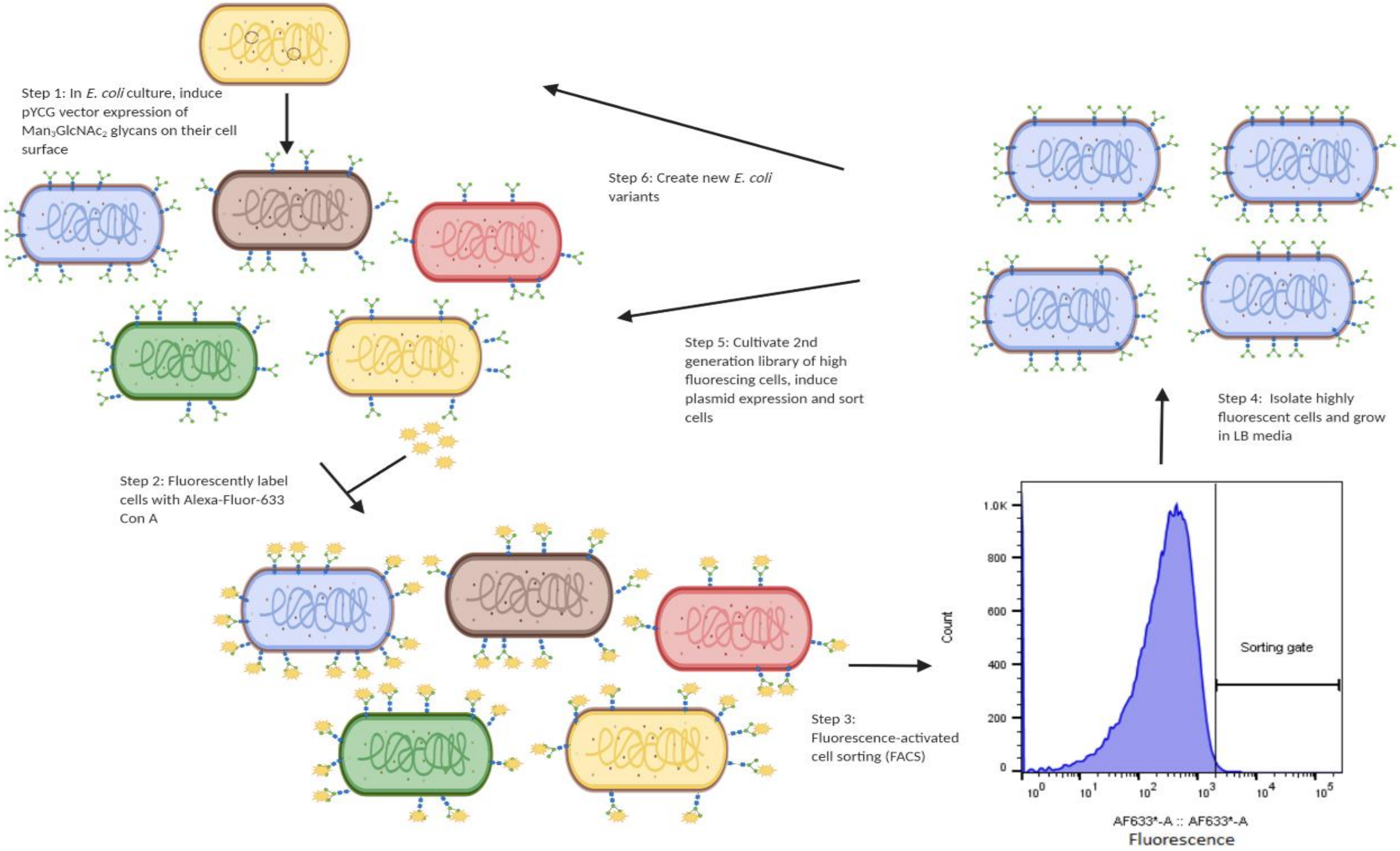

Step 1: In *E. coli* culture, induce pYCG vector expression of $Man_3GlcNAc_2$ glycans on their cell surface
Step 2: Fluorescently label cells with Alexa-Fluor-633 Con A
Step 3: Fluorescence-activated cell sorting (FACS)
Step 4: Isolate highly fluorescent cells and grow in LB media
Step 5: Cultivate 2nd generation library of high fluorescing cells, induce plasmid expression and sort cells
Step 6: Create new *E. coli* variants
1.0K
800
600
400
200
0
Count
10^0
10^1
10^2
10^3
10^4
10^5
AF633*-A :: AF633*-A
Fluorescence
Sorting gate

Figure 4.4. Fluorescent screening of pYCG expression in *E. coli* W3110. General order for the *E. coli* culture and sorting of cells displaying increased mannose expression on the cell surface.

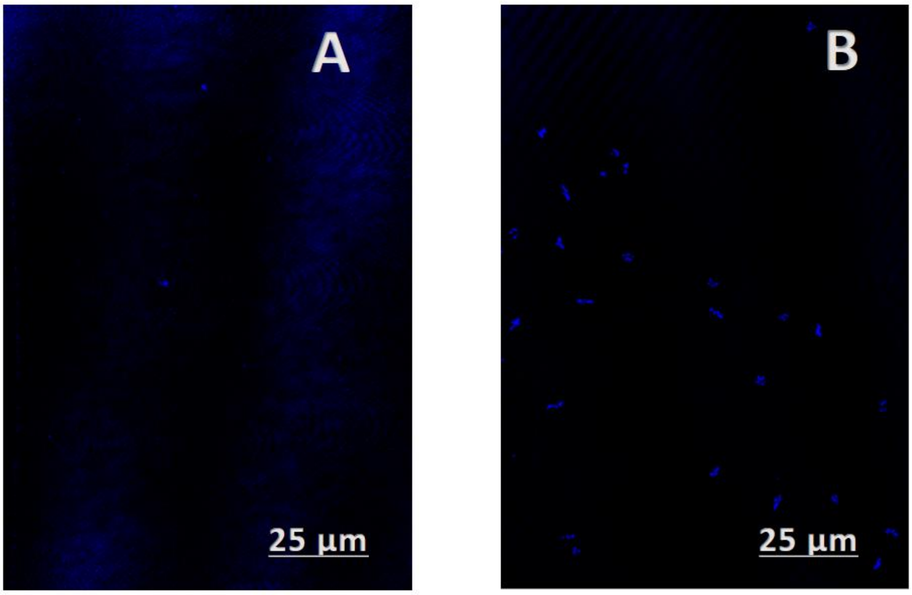


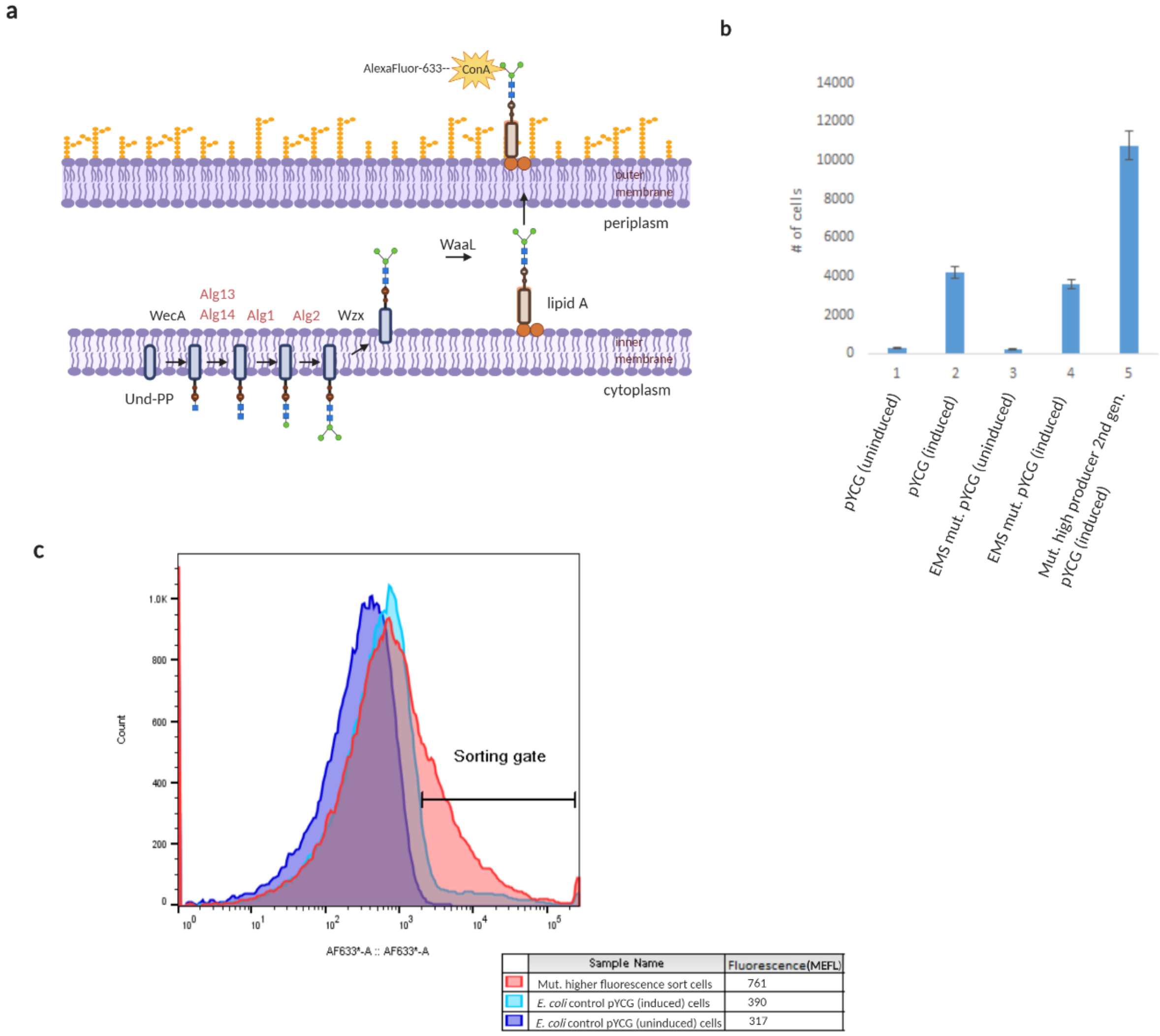
Figure 4.5. Confocal images showing $Man_3GlcNAc_2$ fluorescence of *E. coli* W3110 pYCG cells. Strains shown are A, uninduced labelled cells; B, induced labelled cells. Mannose fluorescence was imaged by laser excitation at 633 nm. Magnification X20. Cells harvested after 4 hours.

4.4.2 Increasing the $Man_3GlcNAc_2$ levels by random mutagenesis

E. coli W3110 cells carrying the pYCG vector were subjected to random mutagenesis using the chemical mutagen ethyl methanesulphonate (EMS) as described in section 4.3.3. The resulting mutants were pooled to form the pre-sort library propagated in LB medium and $Man_3GlcNAc_2$ expression was induced by the addition of arabinose as described in section 4.3.3. After labelling with 2.5 µg/mL concentration of Concanavalin A, Alexa Fluor 633 conjugate, approximately 100,000 cells were subjected to screening using FACS. Single cells from the population exhibiting the top 1-5% mannose display fluorescence were collected into liquid LB medium in 96-well plates, grown and subjected to additional rounds of FACS. An increase in the average AlexaFluor fluorescence intensity was observed in every round. The median fluorescence of the population was approximately 2.4 fold higher than the AlexaFluor fluorescence exhibited by the initial library (Fig. 4.6c).

4.4.2 Increasing the $Man_3GlcNAc_2$ levels by random mutagenesis

E. coli W3110 cells carrying the pYCG vector were subjected to random mutagenesis using the chemical mutagen ethyl methanesulphonate (EMS) as described in section 4.3.3. The resulting mutants were pooled to form the pre-sort library propagated in LB medium and $Man_3GlcNAc_2$ expression was induced by the addition of arabinose as described in section 4.3.3. After labelling with 2.5 µg/mL concentration of Concanavalin A, Alexa Fluor 633 conjugate, approximately 100,000 cells were subjected to screening using FACS. Single cells from the population exhibiting the top 1-5% mannose display fluorescence were collected into liquid LB medium in 96-well plates, grown and subjected to additional rounds of FACS. An increase in the average AlexaFluor fluorescence intensity was observed in every round. The median fluorescence of the population was approximately 2.4 fold higher than the AlexaFluor fluorescence exhibited by the initial library (Fig. 4.6c).



	Sample Name	Fluorescence(MEFL)
🟥	Mut. higher fluorescence sort cells	761
🟦	*E. coli* control pYCG (induced) cells	390
🟦	*E. coli* control pYCG (uninduced) cells	317

Figure 4.6. Detection of $Man_3GlcNAc_2$ pathway with glycan display (a) Scheme for flow cytometric analysis of glycan cell surface display. Cytoplasmic LLOs are a substrate for Wzx-mediated translocation across the inner membrane into the periplasm. Glycans are subsequently transferred to lipid A by O-antigen ligase WaaL and

transported to the cell surface where it is made available for AlexaFluor-633 ConA labeling. Labeled cells are analyzed by flow cytometry. (b) Number of cells displaying brightest fluorescence within the AlexaFluor-633 excitation gated population. All strains were grown in LB and labelled with ConA before flow cytometric analysis. (c) Fluorescence histograms of *E. coli* W3110 cells expressing mannose. MEFL: median fluorescence intensity for gated cell population of interest.

The first round of sorted cells showing higher mannose fluorescing abilities were then either subjected to growth and induction in the second generation or a second round of mutation (Figure 4.7). Control cells refer to wild type *E. coli* containing the plasmid pYCG for cell surface mannose display. Mutant refers to the first round of these wild type cells exposed to EMS for mutagenesis. Mutant high producer refers to cells grown from a single cell selected from within the high mannose displaying mutant population (7HS2). Mutant growth round 2 refers to high mannose displaying cell sort (7HS2) subjected to another round of EMS mutagenesis and single cell sorted for higher mannose display (2EWL7).

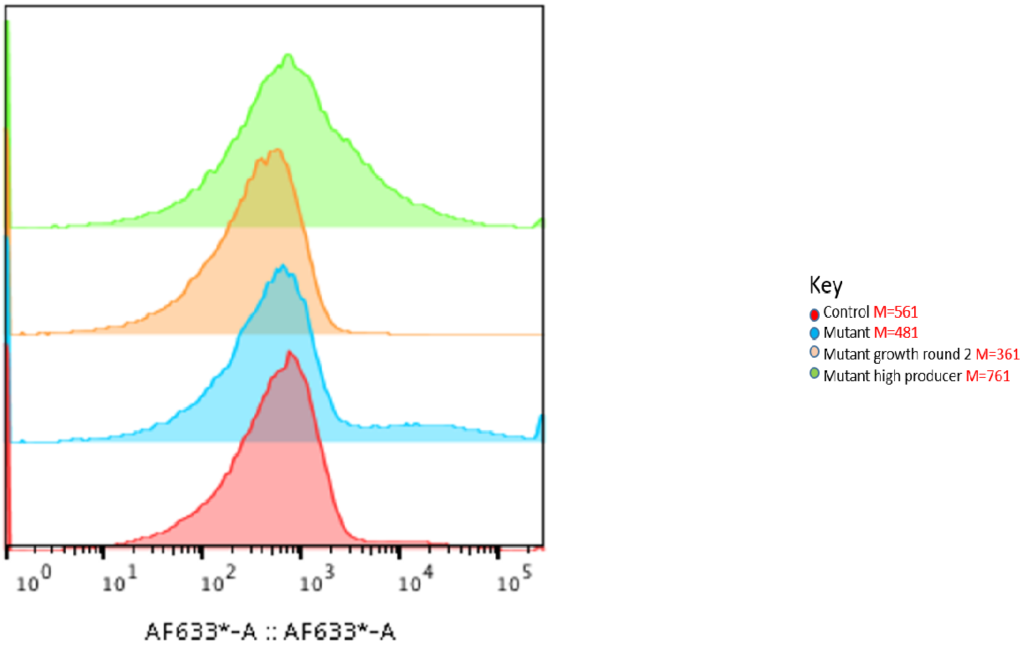


Figure 4.7. Cell surface GDP-mannose display. *E. coli* W3110 cells carrying plasmid pYCG. Cells were labeled with AlexaFluor-633 ConA prior to flow cytometry. Labeled cells are analyzed by flow cytometry and the MEFL: median fluorescence intensity for each histogram is given based on every 100,000 events recorded (see appendix B).

4.5 Discussion

The need for simple and cost-effective methods of producing various glycomolecules for use in the ever growing world of glycobiology cannot be overemphasized. Due to the widely variable array of glycans that can be produced in nature, the uniformity in the production materials and precursors will go a long way in facilitating the study and characterization of effects of this post translational protein modification. While reprogramming the glycosylation pathway in the microbial strain to achieve glycan uniformity has been considered (Anyaogu *et al.*, 2021), earlier research has also explored the idea of efficiently converting microbially derived precursor oligonucleotides in the formation of uniform N-type glycans (Hamilton *et al.*, 2017; Valderrama-Rincon *et al.*, 2012). N-glycan structures are generally classified into the high mannose, complex and hybrid type categories. All three are composed of a common tri-mannosyl ($Man_3GlcNAc_2$) core structure. The high mannose glycans contain 5 to 9 mannose ($Man_{5–9}GlcNAc_2$) sugars. The complex type has 2 GlcNAc's attached to the tri-mannosyl core while the hybrid type has a combination of both high mannose and complex glycans with at least three mannose sugars, but only one GlcNAc on one non-reducing mannose (Hossler, Mulukutla and Hu, 2007).

The tri-mannosyl $Man_3GlcNAc_2$ glycan core which is the basis for a wide variety of complex glycans has been assembled *in vitro* using specific glycosyltransferases and sugar-nucleotide donors. The increased availability of these components required in the N-glycosylation pathway within the endoplasmic reticulum of cell for instance, led to the achievement of unprecedented homogeneity levels of over 85% from its engineered synthesis in the yeast species *Yarrowia lipolytica (De Pourcq et al., 2012)*. In mammalian studies, an inadequate supply of mannose in N-linked oligosaccharide synthesis has also been linked to reduced protein glycosylation (Sharma, Ichikawa and Freeze, 2014; Li *et al.*, 2019; Zalai *et al.*, 2016). It is suspected that inefficient glycosylation (relative to < 50% glycosylation efficiency) often observed in other prokaryotic N-linked glyco-systems was due in part to relatively poor accumulation of the lipid-linked Man3GlcNAc2 substrate during the glycan biosynthesis (Parsaie Nasab *et al.*, 2013). With this in mind, the existence of increased N-glycosylation pathway components within a glycocompetent *E. coli* strain for example should yield more uniformly glycosylated proteins leading to higher glycosylation efficiency levels.

Mannose is an important metabolite in glycosylation reactions (Elizabeth *et al.*, 2021; Sharma, Ichikawa and Freeze, 2014) and increased amounts that can be channelled into the glycosylation pathway should lead to more efficient glycosylation reactions within the cell. Nasab and colleagues used the combined approach of engineering glycosylation efficiency and glycan structure in *Saccharomyces cerevisiae* to produce recombinant proteins with human-like N-glycans (Nasab *et al.*, 2013) By employing random mutagenesis and targeted selection methods, strains of *E. coli* derived from the widely known and studied K-12 variety have been identified as predisposed to higher mannose production. Confocal microscopy was used to test the increased cell surface display hypothesis by inducing protein production being coded for by the ALG yeast genes to direct the GDP mannose onto the cell surface of the *E. coli* strain (Figure 4.5). This was then taken a step further by subjecting the cells with induced mannose expression being displayed on the cell surface to viewing and sorting for cells showing increased fluorescence. The wild type W3110 containing the pYCG plasmid was used as baseline control for the cell surface mannose display to ensure the cells being targeted within the sort gate had more mannose available/displayed (Figure 4.6).

Apart from the increased mannose display of these sorted cells as evidenced by the MEFL values (Figure 4.7) compared to the controls, other methods for accurate quantification of the amount of mannose precursors available within the strains should be explored such as combining ion-pair assisted extraction with hydrophilic interaction chromatography-solid phase extraction (HILIC-SPE). Investigation into the amount of mannose obtained from the mutation which ends up in the N-glycosylation pathway could possibly highlight changes/adaptations within the strains. Further studies and improvements to the mutant strains in subsequent generations should also be explored.

Based on the results obtained, it is expected that the isolated *E. coli* strains identified from this study possess the ability to produce a higher amount of GDP-mannose precursors which can subsequently be channelled into the glycan production pathway. In this chapter, the mutant high producer 2nd generation (7HS2) or the twice mutated high producer strain 2EWL7 which have been demonstrated to be the *E. coli* mutants with their particularly high display of cell surface mannose, could be a useful strain to take forward for further experimentation and perhaps for industrial cultivation. With further optimisation and improvements perhaps these strains could become a widely used glycocompetent *E. coli* strain in industry to synthesise homologous glycans in larger quantities due to the abundant availability of precursors within the cell. The genetic changes within these strains will be tracked and identified in follow up experiments to validate the new phenotype displayed and measure glycosylation efficiency in the cells. However, these insights into cell surface mannose display within the model workhorse *E. coli* strain could be applied directly to other industrially relevant microorganisms, thus improving the prospects for production of precursors for eukaryotic glycan synthesis in prokaryotes.

Chapter 5: Sequencing and characterization of mannose substrate enhanced *Escherichia coli* strains for N-glycoprotein production efficiency

5.1 Summary

In this chapter, a bioinformatics based approach was used in characterizing the genetic changes that resulted in the enhanced mannose display phenotype exhibited by mutant strains compared to the parent strain. The gene variations which resulted in protein-coding changes, were identified and mapped onto the *E. coli* K-12 MG1655 genome which is parent strain to the WT W3110 strain to understand their relationships and pathways involved in the enhanced phenotype. The mutant phenotype characterization has revealed candidate metabolic engineering gene targets for metabolic engineering for better understanding and further improved mannose substrate availability within the *E. coli* cell factory for N-glycoprotein protein production.

5.2 Introduction

An organisms' complete hereditary information is enclosed inside its genome structure, organization, and function. The probability of mutation however depends largely on the mutation spectrum which relies on the fundamental understanding of mutational properties in relation to fitness distribution. In chapter 4, *E. coli* from K-12 family – W3110 strain carrying the pYCG recombinant protein plasmid was genetically modified through random mutagenesis and sorted for their ability to produce higher levels of mannose than the WT strain using flow cytometry. Understanding the mutational profile of these strains is key to matching these genomic level changes to their contributions in the cells' evolutionary process (Shibai *et al.*, 2017).

For evolutionary process variation, mutations to the genetic code are essential. To facilitate understanding of these genome level changes, Next-generation sequencing (NGS) technologies has been employed. It uses DNA, RNA, or methylation sequencing and has come to mostly replace traditional Sanger sequencing because of its low running cost and high-throughput production output of sequencing data. NGS has recently grown to become more than about how various organisms use genetic information and molecular biology to survive and reproduce with and without mutations, to understanding disease and diversity within their altering environments and population networks. The frequent development of several new public bioinformatics databases on the World Wide Web validates and shows the influence NGS has in the life sciences as well as the need to continuously create new methods to query and interpret hereditary information in and around DNA and its nucleotide sequences (Kulski, 2016).

With the emergence of NGS, genomics has been defined as the mapping of genome structure and organization to classify them as either *de novo* sequences, re-sequenced genomes, exonic or targeted sequences and metagenomic sequences. Hence, NGS expands the understanding of structural and functional genomics through the concepts of “omics” to offer new insight into the workings and meaning of genetic conservation and diversity of living things (Kulski, 2016).

The genomic differences between mutant strains developed in the previous chapter and the WT strain were mapped out using Illumina sequencing platform. The Illumina sequencer uses removable fluorescently labelled chain-terminating nucleotides in a technology known as sequencing by synthesis to produce a larger output at lower reagent costs (Metzker, 2010). PCR bridge amplification (also known as cluster generation) is used to generate clonally enriched template DNA which is sequenced into smaller colonies called polonies (Shendure and Ji, 2008). Sequencing data output per run is higher (600 Gb), the read lengths are shorter (approximately 100 bp), the cost is cheaper, and the run times are much longer (3-10 days) than most other systems (Liu *et al.*, 2012).

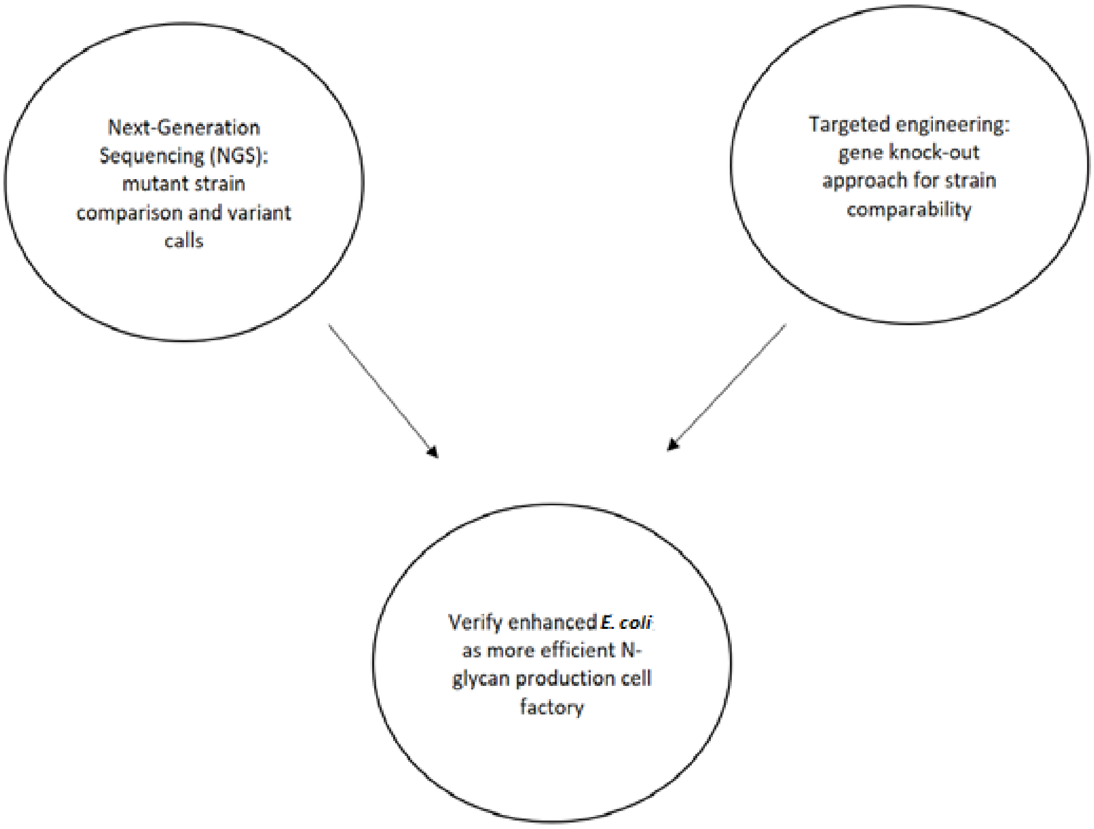

Next-Generation Sequencing (NGS): mutant strain comparison and variant calls
Targeted engineering: gene knock-out approach for strain comparability
Verify enhanced *E. coli* as more efficient N-glycan production cell factory

Figure 5.1 Summary of chapter aims and objectives. Two strategies were used to identify and characterize the enhanced *E. coli* as an N-glycoprotein production chassis. A Next-generation sequencing approach was first taken, in which the mutated strains identified as containing enhanced mannose substrate were sequenced, annotated and variants called to identify gene changes. An engineered gene knock-out approach was also applied by targeting the O-antigen ligase gene – WaaL to prevent cell surface representation of produced glycans. The strains were then subjected to protein expression/growth analysis to gain insights into the mutant phenotype and physiology, to facilitate strain optimization and highlight potential genetic engineering targets for future experiments.

The data generated from the illumina sequencing was then processed using various bioinformatics tools to check the quality of sequencing output, align the sequences against the reference genome, identify variants and annotate the variant calls to determine which genes the mutations fall within. Despite achieving a 2.4-fold increase in cell surface mannose display (per 100,000 events recorded) than what is obtained in the WT strain, further modifications to the mutant strains will be required to enable them to be competitive with other glyco-competent *E. coli* strains and create an ample reference point to quantify the effects of the modified cellular enhancement from mutagenesis. To determine the impacts of increased mannose availability within the cell on protein glycosylation efficiency and glycoprotein production in the *E. coli* strain, O-antigen ligase, WaaL, was knocked out to prevent cell surface representation of recombinant glycans (Mario *et al.*, 2005).

5.3 Specific methods

5.3.1 Strains and plasmids

The *E. coli* strain W3110 (WT) and generated mutant strains from the previous chapter were used for these experiments. The plasmid pKD4-rfaL was generated by amplification of the chromosomal O-antigen ligase (*WaaL*) *rfaL* gene from W3110 cells with *rfaL*-specific DNA primers (see appendix A, page 134).

5.3.2 Sample preparation and sequencing

Cryovial sample preparation for each strain to be sequenced was done by picking a single colony and mixing in a 100 µl volume of sterile PBS buffer. Two-third lawn of *E. coli* culture was plated on LB agar and streaked out to determine culture purity. The strains were grown at 37 °C until adequate growth was observed (Fig. 5.2). With a large sterile loop, all the *E. coli* culture was scrapped off the plate and mixed into barcoded bead tubes. The tubes were inverted 10 times to ensure adequate mixing with the cryopreservant liquid and samples shipped off per MicrobesNG packaging instructions.

The sample DNA was quantified in triplicates using Illumina sequencing method via the Whole Genome Sequencing service and the reads were trimmed using Trimmomatic. Microbes NG in-house quality assessment was carried out using scripts combined with the following software: Samtools, BedTools and bwa-mem. The reads were assembled using SPAdes and turned into contigs and the data annotated using Prokka software (MicrobesNG, 2021)

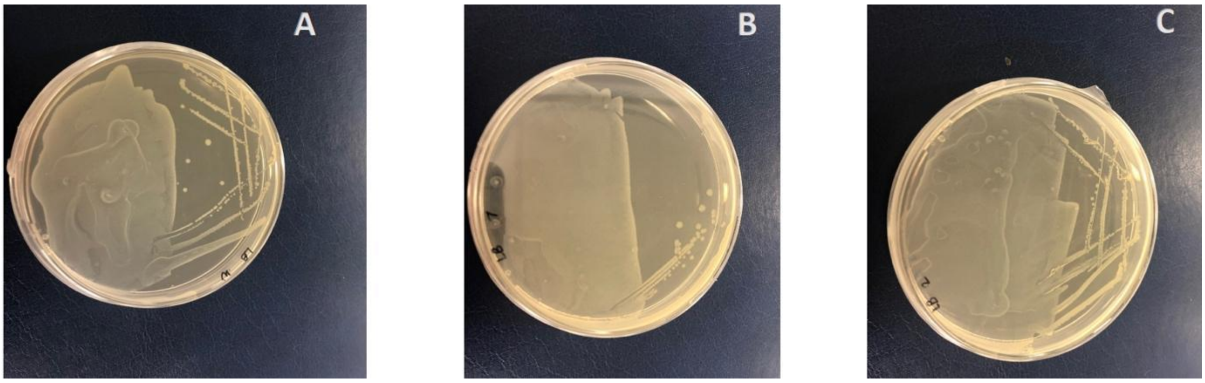


Figure 5.2 *E. coli* lawn culture plates of pure cultures grown at standard conditions for strain sequencing.

5.3.3 Sequencing read quality control using KBase/Galaxy

Sequencing data from the strains was returned in FASTQ file formats with the trimmed data reads labelled U1 and U2. These files are used to identify sequence clusters with each cluster passing through a fluorescent reader. The read files generated from the paired end run are then imported into the KBase online platform through the staging area. From the applications section, the read quality was assessed using the FastQC – v0.11.9. This application runs checks on the imputed sequences and is compatible with data generated from the NGS Illumina sequencing platform. The output is in the form of graphs presented in the visual format with data displayed using the RAG (Red, Amber and Green) rating system to classify data as normal or unusual.

The 3 FASTQ files were uploaded into the MultiQC tool (a modular tool for aggregation of multiple samples from bioinformatics analyses into one single report) on the Galaxy bioinformatics platform to compare the 3 strains using the same reference data points.

5.3.4 Variant calling and annotation

Identifying the differences between the data reads generated and a known reference genome was used to generate single nucleotide polymorphism (SNP) files and small insertions and deletions from the reads aligned to the reference genome (Syme *et al*., 2021). The sequencing contig data was assembled and aligned against the reference genome for *E. coli* K-12 MG1655 using the Burrow-Wheeler Aligner tool - BWA MEM algorithm (Li, 2013) to map the data and a ".bam" file was generated which is a binary format of the sequence alignment map for each sample. The files were then visualised using the Integrative Genomics Viewer (IGV) software (Thorvaldsdóttir, Robinson and Mesirov, 2013; Robinson *et al.*, 2017). Variants in the strains were then called using the haplotype-based variant detection tool – Freebayes (Garrison and Marth, 2012) to generate variant call format (vcf) text files which were also loaded into IGV for viewing.

The Variant Effect Predictor web tool from Ensembl Genomes (EnsemblBacteria) (Howe *et al.*, 2020) was used to annotate the sample vcf files after realigning the data to the Ensembl provided genome. The vcf files were uploaded into the web tool and annotated against genes to determine the functional consequences of the detected variations.

5.3.5 Gene variant mutation analysis

The functional consequences of the resulting variant annotated against genes in reference genome *E. coli* K-12 MG1655 was recorded as either an upstream gene variant, downstream gene variant, synonymous variant or missense variant with the biotype identified as either protein coding or pseudogene. Thus, the identified shared genes between the WT and mutant strains were mapped out for overlaps to distinguish the independent new gene mutations from those identified in the shared common ancestor (see appendix C Table 1, page 149 for further details).

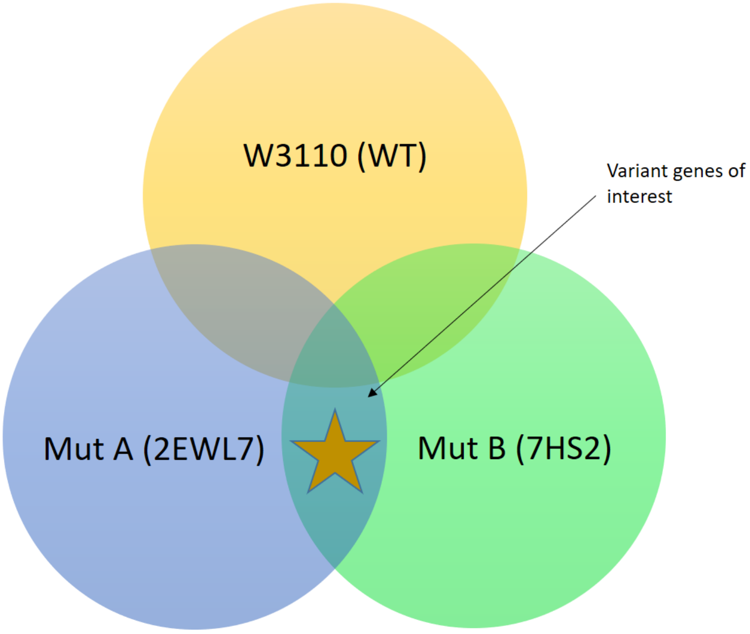


Figure 5.3 Intersects between the WT and mutant strain gene variants used to identify impact bearing genes of consequence in enhanced mannose production within the cell chassis.

5.3.6 *WaaL* knockout

Chromosomal knockout mutant of *E. coli* W3110 was generated using the gene replacement strategy as described by Datsenko and Wanner (Datsenko and Wanner, 2000). Using plasmid pKD4 as a template, the kanamycin-resistant gene flanked by homologues of *waaL* gene was amplified by PCR (Fig. 5.4) using specific primers as described (see Appendix A, page 134). The PCR products were electro-transferred into the *E. coli* strains, with the help of the Red recombinant system. *WaaL* gene was replaced by the kanamycin-resistant gene. Then the kanamycin-resistant gene was eliminated by the FLP-promoted recombination system.

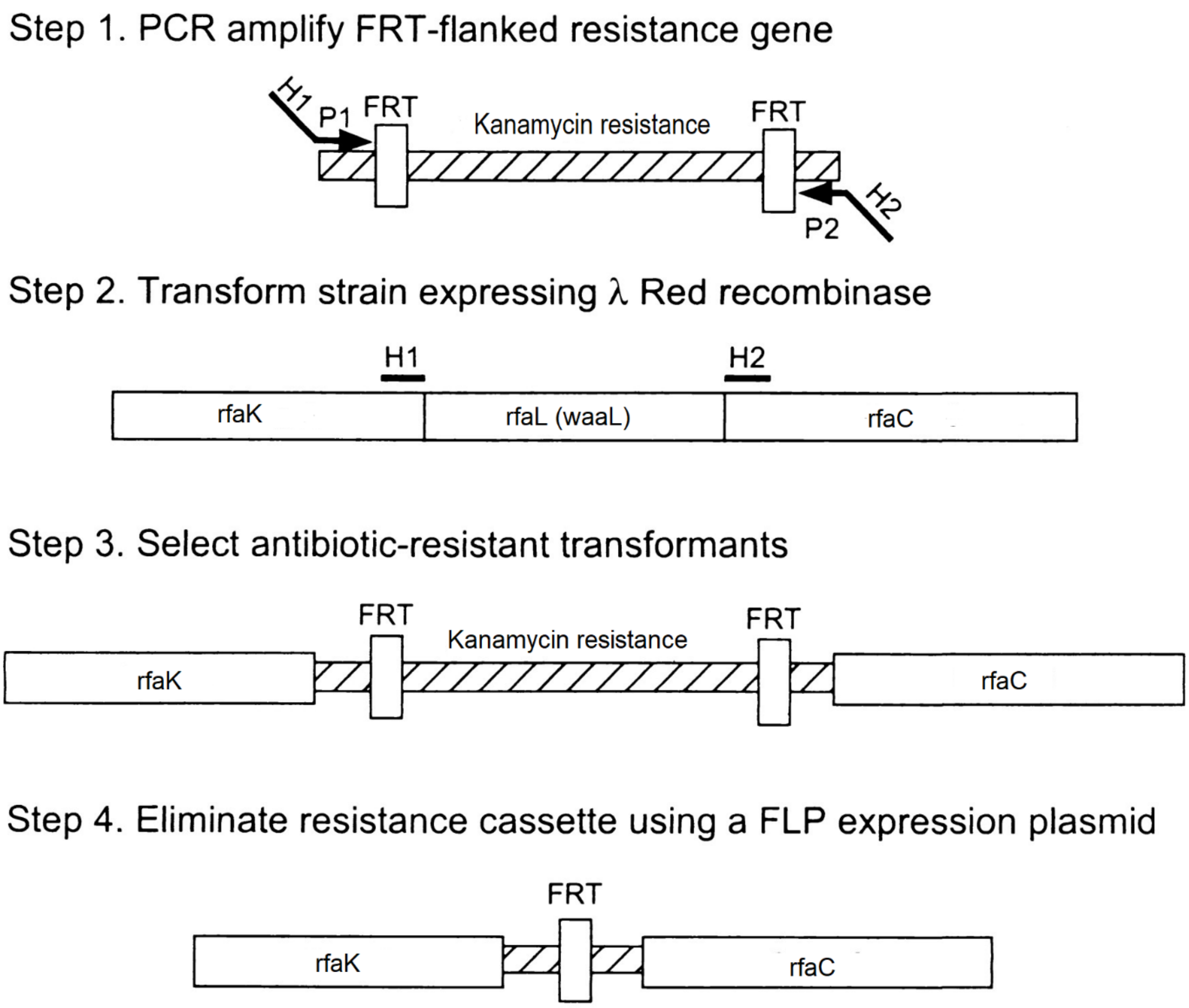

Step 1. PCR amplify FRT-flanked resistance gene
H1
P1
FRT
Kanamycin resistance
FRT
H2
P2
Step 2. Transform strain expressing λ Red recombinase
H1
H2
rfaK
rfaL (waaL)
rfaC
Step 3. Select antibiotic-resistant transformants
FRT
Kanamycin resistance
FRT
rfaK
rfaC
Step 4. Eliminate resistance cassette using a FLP expression plasmid
FRT
rfaK
rfaC

Figure 5.4 Gene disruption strategy. rfaL (waaL gene) homology regions labelled H1 and H2. P1 and P2 are sequence priming sites. FRT refers to FLP recognition target (sites).

5.3.7 Glycoprotein expression and periplasmic extraction

Overnight culture of cells carrying the glycan biosynthesis plasmids along with pEC(AcrA) were inoculated into 100ml of LB in triplicates and grown at a temperature of 37°C shaking at 180 rpm until absorbance OD_{600} was 0.5. Culture temperature was reduced to 30°C and after a temperature equilibration period of 10 minutes, AcrA protein expression was induced by adding 0.2 % (v/v) L-arabinose. Cells were left to incubate and express protein for an additional 4 hours. The cultures final OD was measured and volumes corresponding to 40 OD units were harvested through centrifugation at a speed of 4,500 x g and 4 °C temperature for 10 minutes. The supernatant was discarded and pellets collected.

Pellets were resuspended in 1 ml periplasmic lysis buffer (20% sucrose, 1 g/L lysozyme, 30mM Tris-HCl pH 8.5, 1 x Halt Protease inhibitor complex) and allowed to roll on ice for 2 hours. The soluble protein fraction was collected by spinning down the cells at speed 4,500 x g and temperature 4 °C for

10 minutes. The supernatant was collected as the protein soluble periplasmic fraction. To isolate glycoproteins, the periplasmic fractions were affinity purified using a His-Trap HP Nickel affinity column (GE Healthcare). Elutes were collected for SDS-polyacrylamide gel analysis.

5.4 Results

5.4.1 Growth kinetics and physiology of strains

Two randomly selected colonies from the mutant high mannose fluorescing *E. coli* transformants, 7HS2 and 2EWL7, were grown under standard conditions alongside the parental strain W3110. Figure 5.5 shows the growth kinetics of the strains while Table 5.1 highlights the growth rate parameters.

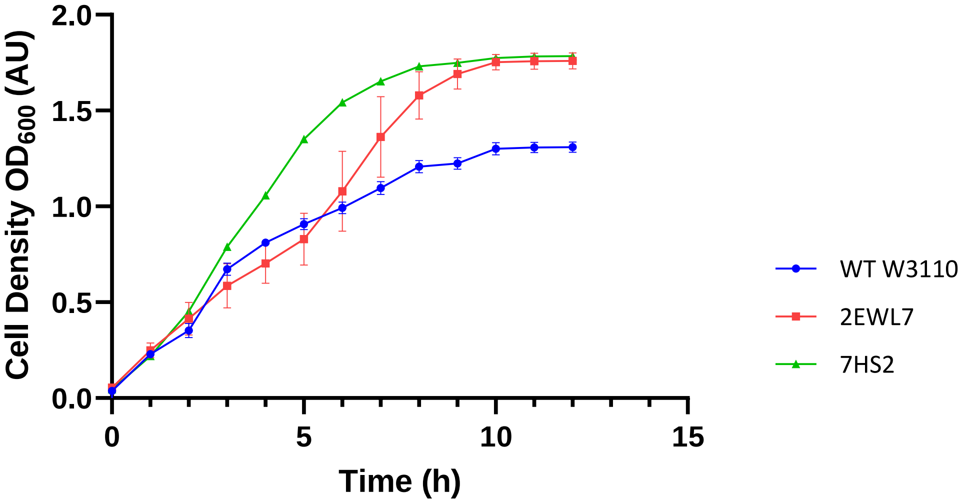


Figure 5.5 Growth kinetics of wild-type W3110 and mutant strains grown under standard conditions. Data shown are taken from three biological replicates. Error bars represent standard deviation, all of which are < 10% of the mean. Growth was measured by optical density at 600 nm. Values interpolated using Graphpad Prism.

Table 5.1: Growth kinetics of WT W3110 and mutant strains grown under standard conditions

Strain	Specific growth rate (SGR, h-1)	Doubling Time (min)
W3110 (WT)	0.32 ± 0.02	22 ± 0.08
2EWL7 (Mut-1)	0.23 ± 0.05	26 ± 0.08
7HS2 (Mut-2)	0.38 ± 0.04**	18 ± 0.02*

Growth rate and doubling time determined using cell count measurements at 2 h and 5 h time points. Data is expressed as means ± standard deviation (SD), number of replicates (n) = 3. Significant differences were calculated using a student's t-test. *P < 0.05, **P < 0.01

5.4.2 Read quality assessment of strains

The FastQC – v0.11.9 ran multiple checks on read sequences and generated quality assessment data for the individual strains and the MultiQC tool from Galaxy Europe interface produced a single report containing graphs visually showing the sequencing quality for the W3110 WT and 2 mutant strains which are presented below:

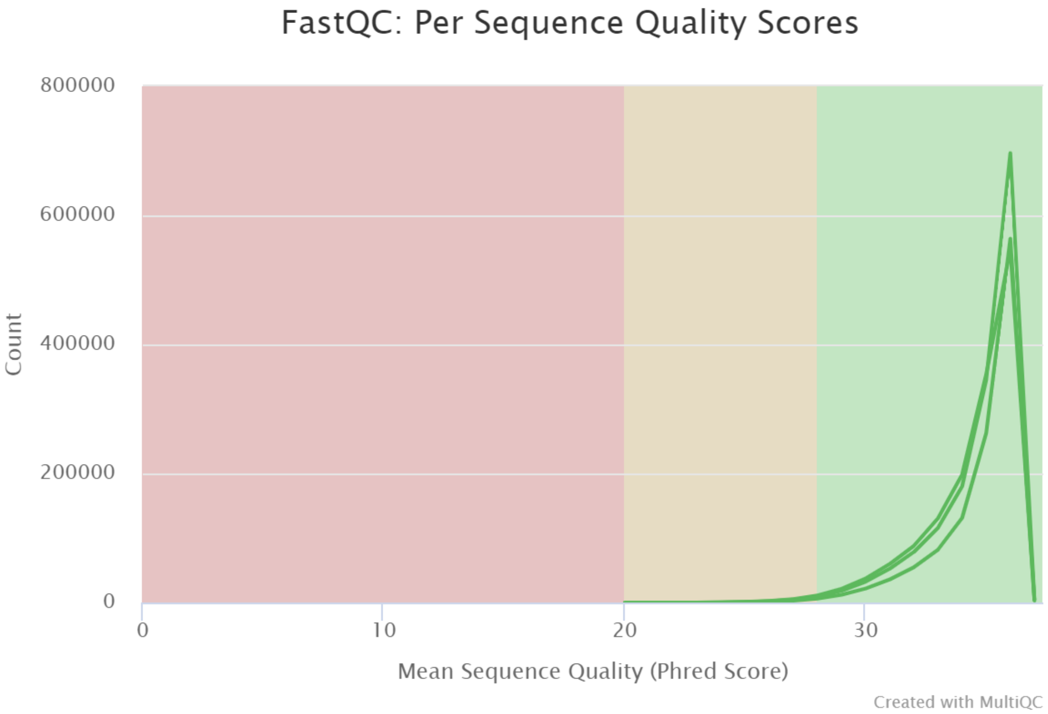


Figure 5.6 Per base sequence quality scores graph obtained for the 3 strains from MultiQC.

The 3 samples had an average sequence length of between 235-239 bps and all had mean sequence quality phred scores ranging from between 33-36. For the WT W3110, 77.9% of the sequence counts were identified as unique reads with an estimated 22.1% duplicate reads. 81.6% unique read sequences were identified in the 7HS2 sequence counts while duplicate reads were an estimated 18.4%. In the 2EWL7 samples, 83.1% of the sequence counts were unique while an estimated 16.9% were identified as duplicate reads. The average GC content for all the samples also followed a normal distribution pattern and there were no N base calls across all the samples. Less than 1% of the reads were made up of overrepresented sequences and no samples were found to be contaminated with > 0.1% adapter sequences.

5.4.3 Variant calling of strains and functional consequences

The BAM *".bam"* files generated from the data mapping tool BWA MEM were loaded onto the IGV software with corresponding index files with the same filename and extension *".bai"* to view the aligned sequences against the known reference genome. The sorted sequences were loaded and each sample BAM file creates 3 associated tracks which are: Alignment track to view individual aligned reads, a Coverage track to view the depth of coverage and a Splice junction track which shows an alternative view of reads covering the splice junctions although only the alignment and coverage tracks are displayed by default (Fig. 5.7).

The read depth at each locus is displayed as a grey bar and if a nucleotide differs from the reference sequence in more than 20% of quality weighed reads, the bar is coloured in proportion to the read count of each base. Structural variants can be detected and viewed with IGV as it displays colour and visual markers to highlight potential gene alterations in the read compared to the reference sequence/genome. Insertions and deletions with respect to the reference genome are also displayed.

Each strain read sequence was also individually aligned to the annotated reference genome and this clearly shows the variations within sequenced reads and within which exact genes these fall into to call the functional consequences of such variations.

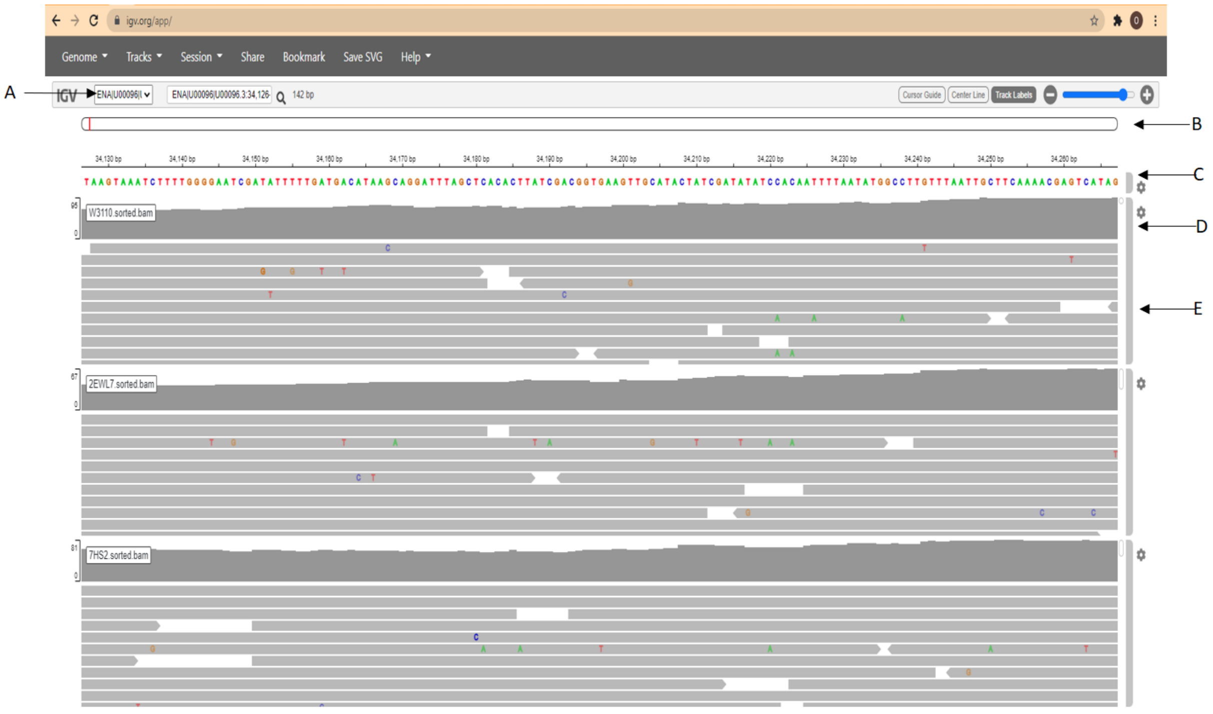


Figure 5.7 IGV display of aligned WT and mutant strains' sequences against known *E. coli* K-12 MG1655 reference genome. (A) Reference Genome (B) Chromosome locator (C) Reference genome sequence (D) Track coverage (E) Sequence alignment track

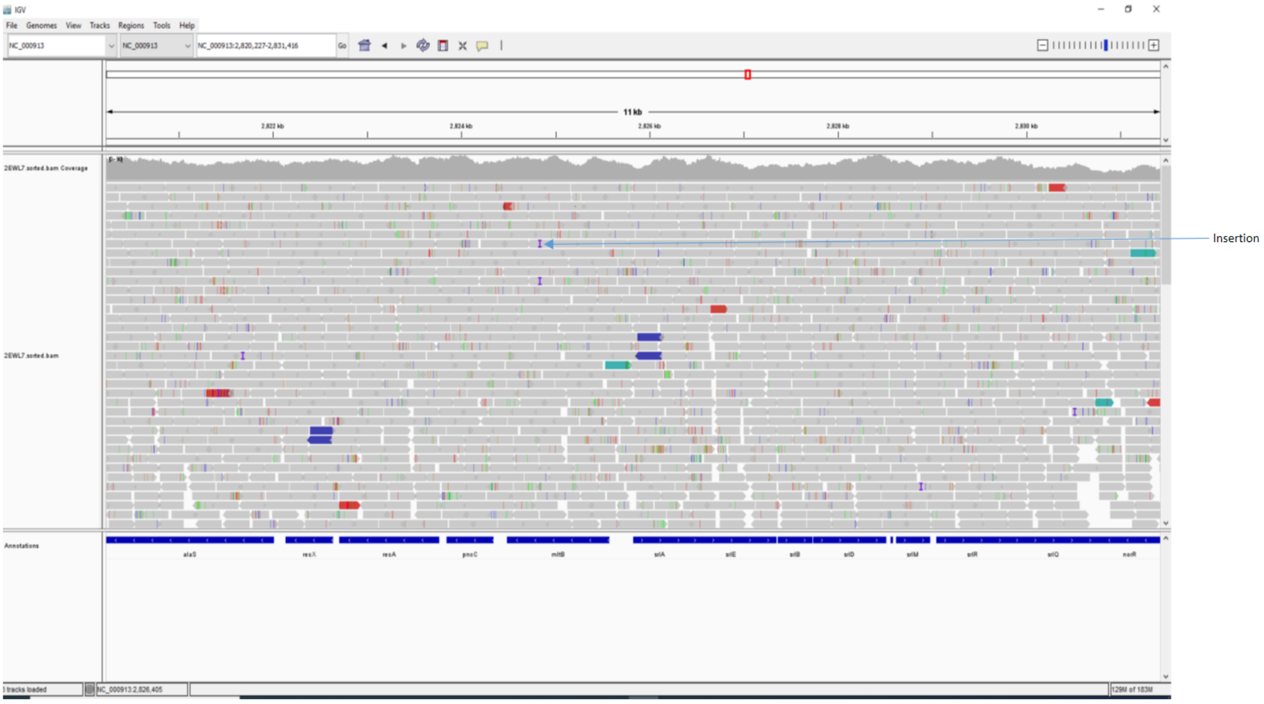
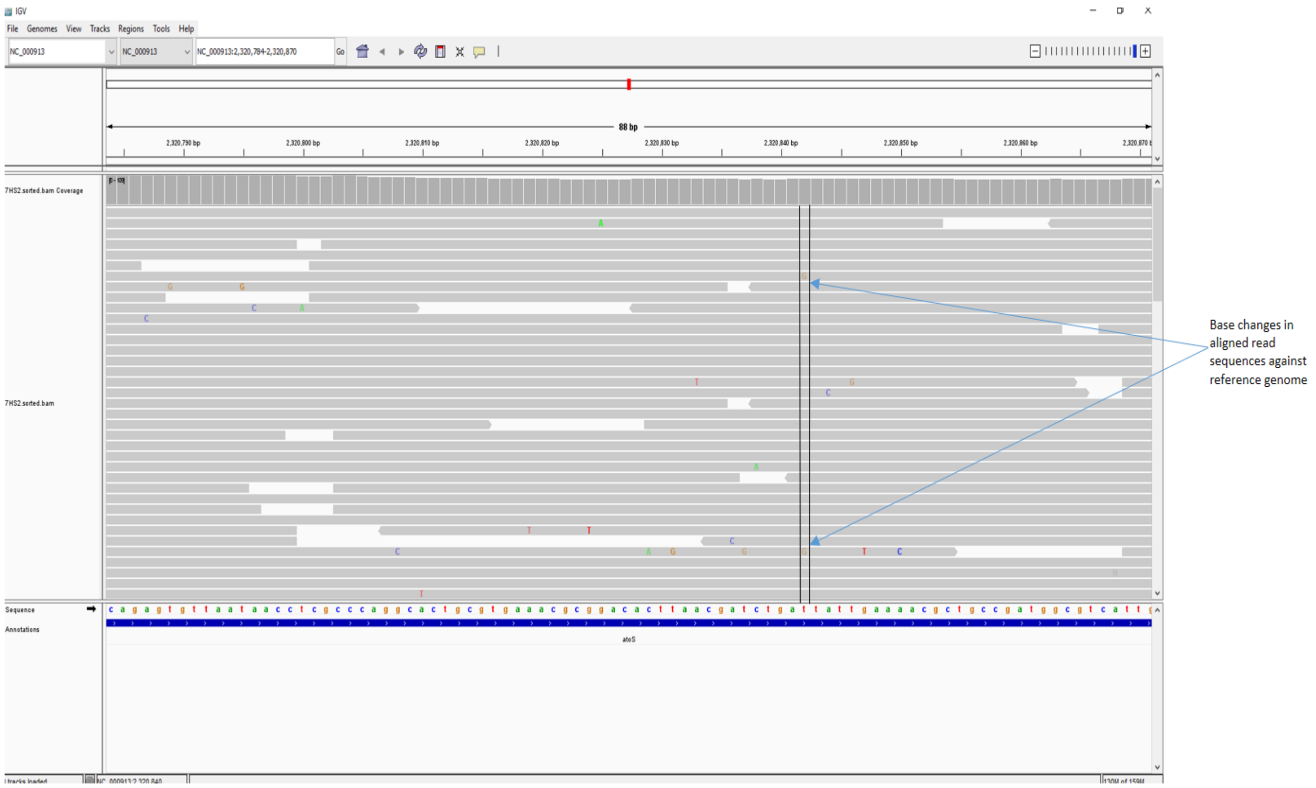


Figure 5.8 IGV display of sorted aligned mutant variant files against reference genome. "I" indicates points of insertion and "-" deletion within the sequence. Red arrow indicates


IGV
File Genomes View Tracks Regions Tools Help
NC_000913
NC_000913
NC_000913:2,320,784-2,320,870
Go
88 bp
2,320,790 bp
2,320,800 bp
2,320,810 bp
2,320,820 bp
2,320,830 bp
2,320,840 bp
2,320,850 bp
2,320,860 bp
2,320,870 bp
7HS2.sorted.bam Coverage
7HS2.sorted.bam
Sequence
Annotations
c a g a g t g t t a a t a a c c t c g c c c a g g c a c t g c g t g a a a c g c g g a c a c t t a a c g a t c t g a t t a t t g a a a a c g c t g c c g a t g g c g t c a t t
atoS

Base changes in aligned read sequences against reference genome

Figure 5.9 Zoomed-in IGV display of aligned mutant sequences against reference genome showing specific base variations to reference sequences.

5.4.4 General features and pathway mapping of gene variants

The impact bearing genes with functional consequence were assembled to identify new gene variants from the alignment of sequences of the wild type W3110 parent and mutant strains against reference genome K-12 MG1655 using the Variant Effect Predictor tool by Ensembl Bacteria. A rule based approach is used to predict the effects each allele of the variant has on each transcript. The results generated were available in either a text, VEP or VCF format. Intersect genes in the overlapping region when comparing variants from both mutant strains and WT were analysed and genes present in all 3 were excluded to focus on gene variants not duplicated in both mutant strains as targets to be mapped using Pathway Tools v25.0 (see Appendix C, page 148). Table 5.2 lists the affected pathways within the mutant strains as presented in a Gene-Reaction schematic to illustrate the relationships between the gene and the reaction of its protein product. These genes are also mapped out onto the *E. coli* K-12 MG1655 genome using the circular genome viewer in ECOCYC (Fig. 5.10).

Table 5.2: Transcript feature gene variants in identified pathways in *E. coli* K-12 MG1655.

Gene Variant (Transcript Feature)	Pathway	Summary
NadE	Cofactor, Carrier and Vitamin Biosynthesis	Biosynthesis of small molecules including cofactors, prosthetic groups, electron carriers and vitamins which participate in enzyme reactions pathway contains
HisG, HisI, HisA, HisH, HisF, HisB, HisD, GlsA, GlsB, GuaA, PyrG, LeuB, DmlA, AroA, AroL, AroK	Amino Acid Biosynthesis	Involved in pathways for the biosynthesis of the 22 amino acids normally present in proteins and other amino acids and modified amino acids incorporated into proteins.
ArgF, ArgI	Amine and Polyamine Biosynthesis	This contains pathways in the biosynthesis of amines and polyamines which play different roles in metabolism including acting as osmoprotectants and keeping DNA in a condensed state and serving as intermediates in macromolecule synthesis.
ClsB, ClsA, KdsB, IpxK, BirA	Fatty Acid and Lipid Biosynthesis	Involved in the synthesis of fatty acids and other lipids including phospholipids.
FadK, FadD, GlpQ,UgpQ, GlxK, GarK	Fatty Acid and Lipid Degradation	These contain pathways in which different fatty acids and other lipids are degraded to become sources of nutrients and energy.

WcaL, RfbA, RffH, RffG, RfbB, RfbD, Prs, CpsB, CpsG, WcaK, WzxC, WcaJ, RfbC	Carbohydrate Biosynthesis	The colanic acid biosynthesis pathway involved in sugar, polysaccharide and glycan biosynthesis.
MalP, YbiV, Agp, YidA, YigL, YihX, GarK, GlxK	Carbohydrate Degradation	This pathway contains various enzymes which enables the organism in degradation of substrates to be used as energy and nutrient sources as well as the use of exogenous sources for the production of essential metabolites.
PyrI, PyrB, PurD, PurK	Nucleoside and Nucleotide Biosynthesis	Contains pathways of synthesis of the 8 nucleoside triphosphates that are RNA and DNA building blocks.
OtsA, OtsB	Metabolic Regulator Biosynthesis	Involved with organic solute biosynthesis
LtaE, GlsB, GlsA, HisH, HisF, GuaA, PyrG, AstA, AstD, AstE	Amino Acid Degradation	Degradation of different amino acids to be utilized for energy and nutrients
Rpe, TktB, TktA	Pentose Phosphate Pathway	Part of the central metabolism pathway essential for the supply of precursor metabolites.
IspE	Secondary Metabolite Biosynthesis	Biosynthesis of organic compounds not directly involved in growth, development and reproduction.
AmiD, AmpD, AmnK, FrlB, FrlC, FrlD, MurB	Secondary Metabolite Degradation	Degradation of organic compounds not directly

		involved in growth, development and reproduction.
AllD, Allc, AllE, YbcF, CarA, CarB,	Amine and Polyamine Degradation	Involved in the degradation of different amines except amino acids.
GloA, PaaE, PaaA, PaaC, PaaB,	Aromatic Compound Degradation	The pathway provides nutrients and energy from the degradation of heterocyclic and sulphur-containing compounds.
MurI, MurB,IpxK, KdsB, PbpC, MgtA,MrcA, MrcB	Cell Structure Biosynthesis	This contains enzymes involved in biosynthesis of cellular organelles such as cell wall components and associated substances.
PrpE, Acs, DmlA	Carboxylate Degradation	Pathway utilizes aliphatic carboxylates as sources of nutrients and energy.
CheA, NarQ	Signal transduction pathways	Involved in signalling pathways within the cell.
ThrS, PheS, PheT, ValS,	Aminoacyl tRNA Charging	This contains a set of reactions involved in the bonding of amino acids to corresponding tRNA molecules.
Rnd, Rnt, TsaE, TsaB, TsaD	Macromolecule Modification	This pathway modifies pre-existing large macromolecules such as proteins and nucleic acids.

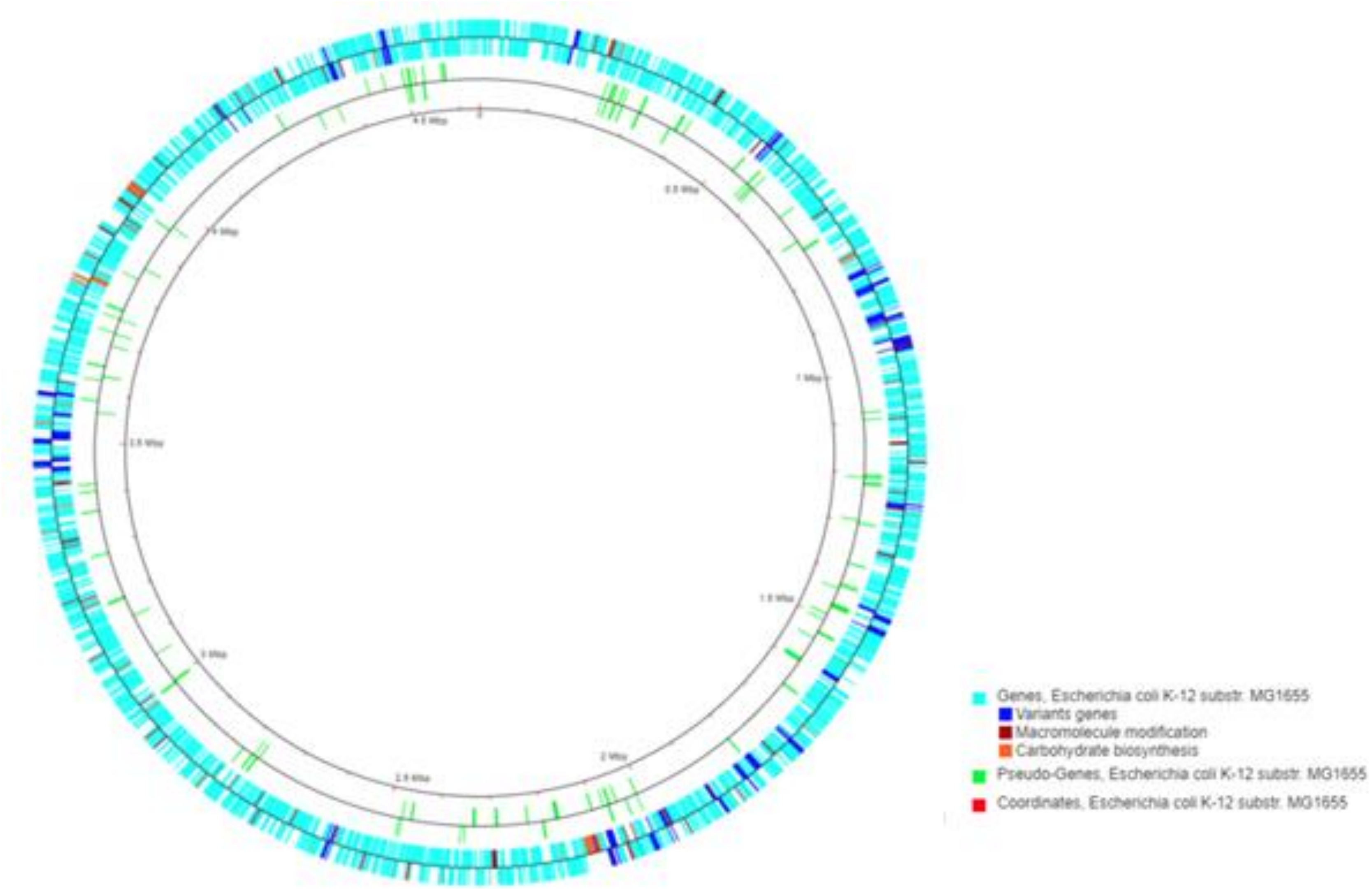

4.5 Mbp
0.5 Mbp
4 Mbp
1 Mbp
3.5 Mbp
1.5 Mbp
3 Mbp
2 Mbp
2.5 Mbp
Genes, Escherichia coli K-12 substr. MG1655
Variants genes
Macromolecule modification
Carbohydrate biosynthesis
Pseudo-Genes, Escherichia coli K-12 substr. MG1655
Coordinates, Escherichia coli K-12 substr. MG1655

Figure 5.10 Transcript feature variant genes and carbohydrate biosynthetic cluster genes mapped onto the *E. coli* K-12 MG1655 genome using Circular Genome Viewer tool from ECOCYC.

The metabolic changes which could occur in the EMS mutated strains 2EWL7 and 7HS2 strains compared to the WT W3110 strain was explored by mapping the identified transcript feature bearing protein coding gene variants using KEGG pathway analysis. 320 genes could be assigned to KEGG IDs and were mapped to *E. coli* K-12 MG1655 metabolism with the search and colour function (https://www.genome.jp/kegg/tool/map_pathway2.html?cre) as shown in Figure 5.11.

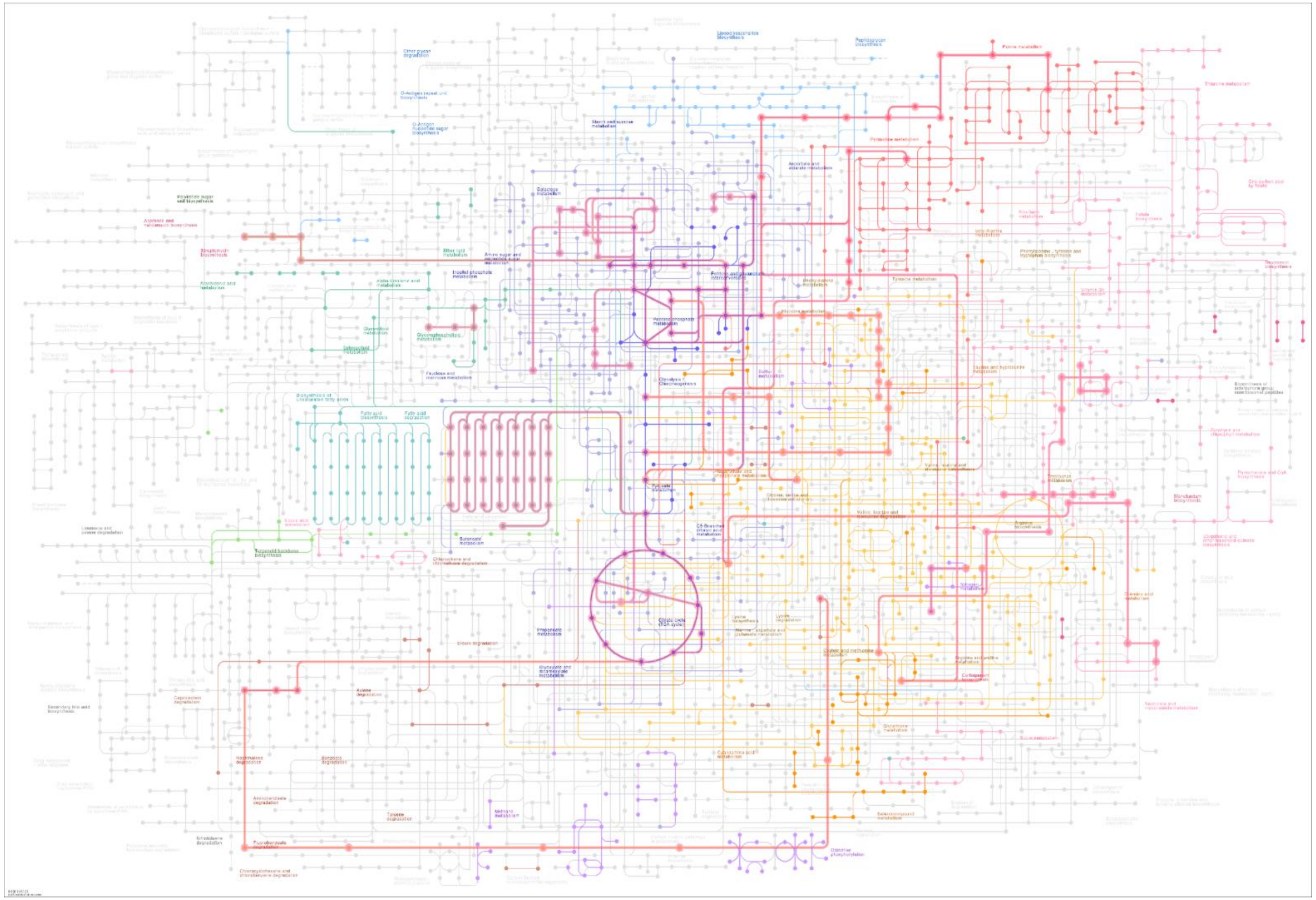


Figure 5.11 Metabolic pathway diagram from KEGG showing protein coding gene variants identified in the mutant strains which possess transcript feature bearing consequences mapped to the *E. coli* K-12 MG1655 metabolism. The pink highlight provides a general overview of *E. coli* metabolism pathways affected.

5.4.5 Analysis of gene variant consequences with amino acid association in respect to WT

The difference in the gene variants observed in the two EMS mutated strains relative to the WT W3110 with consequent attached amino acid changes showing the biological processes they are involved in are presented in Figure 5.12 while a summary of the amino acids affected is shown in table in appendix C, page 149. These genes were also mapped onto the *E. coli* K-12 MG1655 circular genome along with the carbohydrate biosynthesis cluster genes to identify any potential areas of overlap which could have an effect on the carbohydrate biosynthesis within the mutant strains.

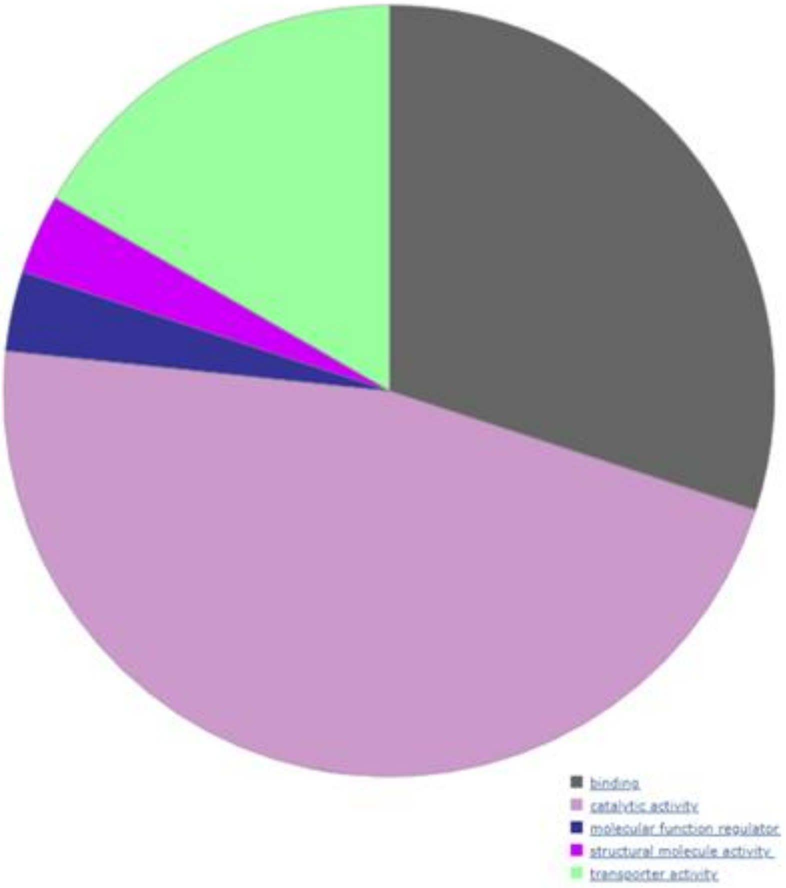


Figure 5.12 Variant genes with consequent amino acid changes showing biological processes they are involved in displayed using PANTHER gene analysis tool.

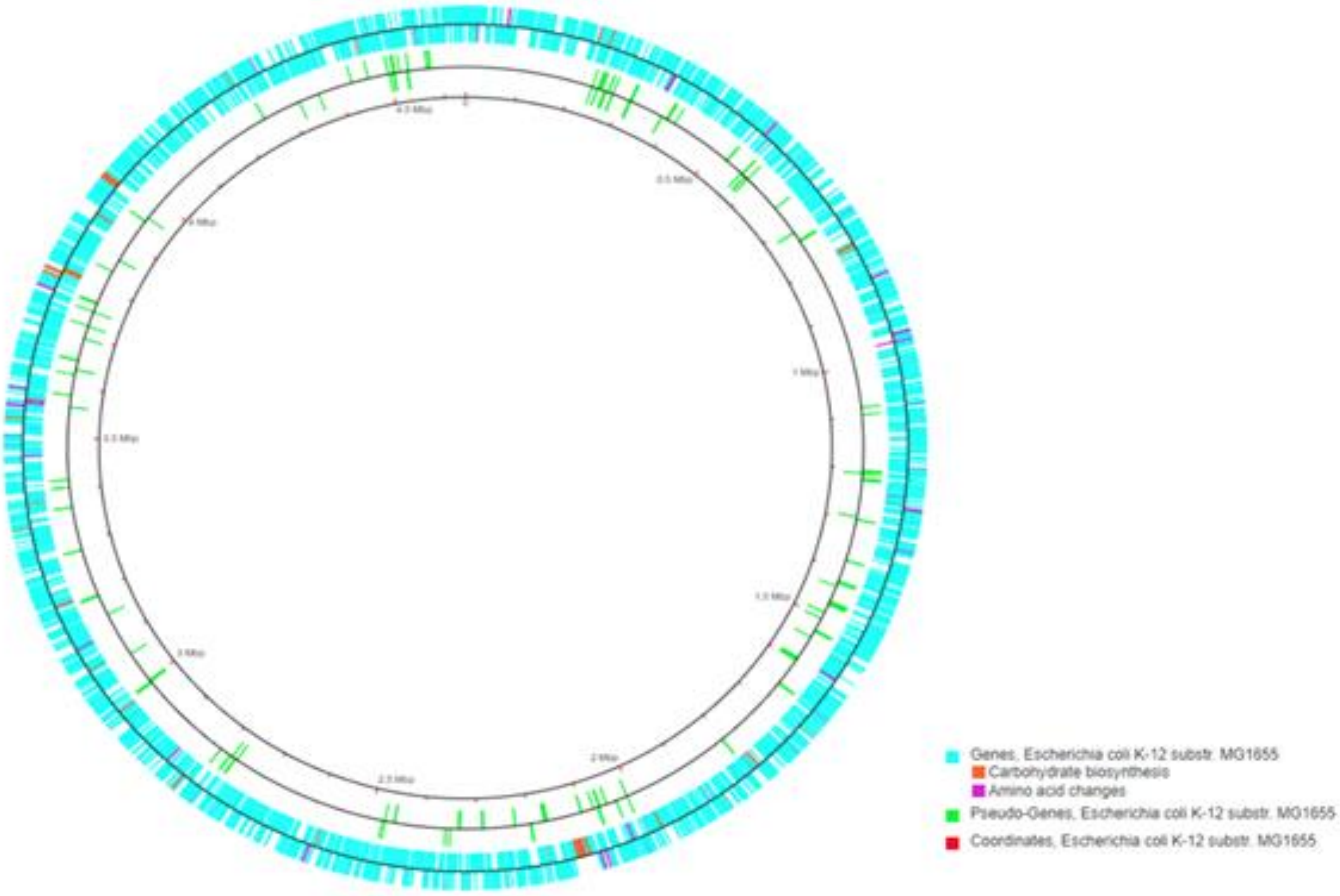

4.5 Mbp
0.5 Mbp
4 Mbp
1 Mbp
3.5 Mbp
1.5 Mbp
3 Mbp
2 Mbp
2.5 Mbp
Genes, Escherichia coli K-12 substr. MG1655
Carbohydrate biosynthesis
Amino acid changes
Pseudo-Genes, Escherichia coli K-12 substr. MG1655
Coordinates, Escherichia coli K-12 substr. MG1655

Figure 5.13 Variant genes with consequent amino acid changes and carbohydrate biosynthetic cluster genes mapped onto the *E. coli* K-12 MG1655 genome using Circular Genome Viewer tool from ECOCYC.

5.4.6 Description of the waaL disruption strategy

In order to measure the effect higher mannose availability within the mutant strains has on N-glycoprotein production efficiency, a known N-glycoprotein had to be expressed in the mutant strain comparable to established glycocompetent *E. coli* strain. The *E. coli* strain CLM24 – a characterized variant of W3110 with the waaL gene knocked out was used for the comparison. Following the standard protocol as illustrated in Fig. 5.4, PCR products were generated using several pairs of 60-70 nucleotide-long primers that included homology extensions and priming sequences for pKD4 as template (see appendix A, page 134 for primer sequences). The respective PCR products were then gel purified and transformed into *E. coli* carrying the Red helper pKD46 plasmid (ampicillin resistant) and the cells were plated unto kanamycin selective plates and left to grow overnight at 37 °C. Loss of the pKD46 plasmid in the cells were checked by streaking colonies out on ampicillin plates and kanamycin plates. The cells that grew on only kanamycin plates were taken forward.

Mutants were verified using colony PCR with primers located within upstream (rfaK) and downstream (rfaC) gene sequences of the waaL (rfaL) gene as well as with a primer downstream rfaL and one that binds to the antibiotic resistance gene (K2). For PCR verification associated gels and primer sequences, see appendix A

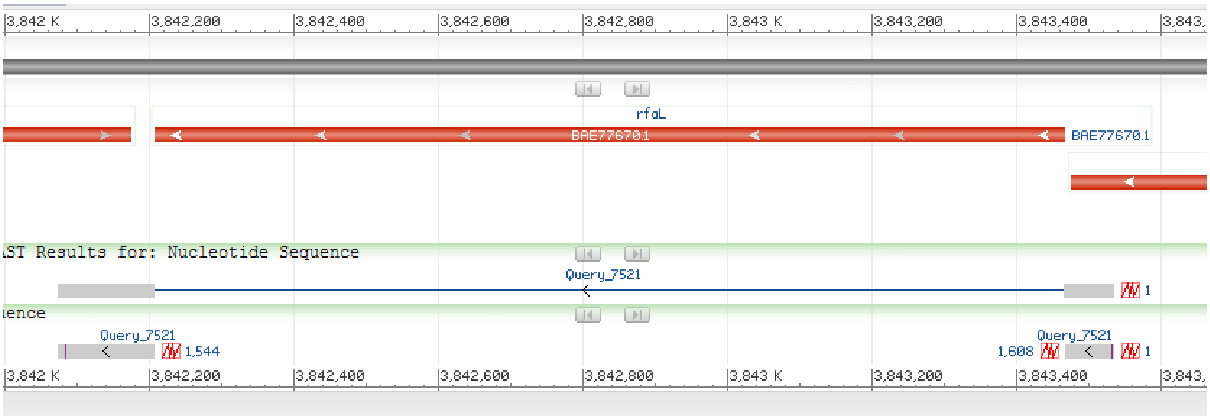


Figure 5.14 NCBI display of aligned mutant 7H2S sequence against reference genome showing query rfaL sequence not found.

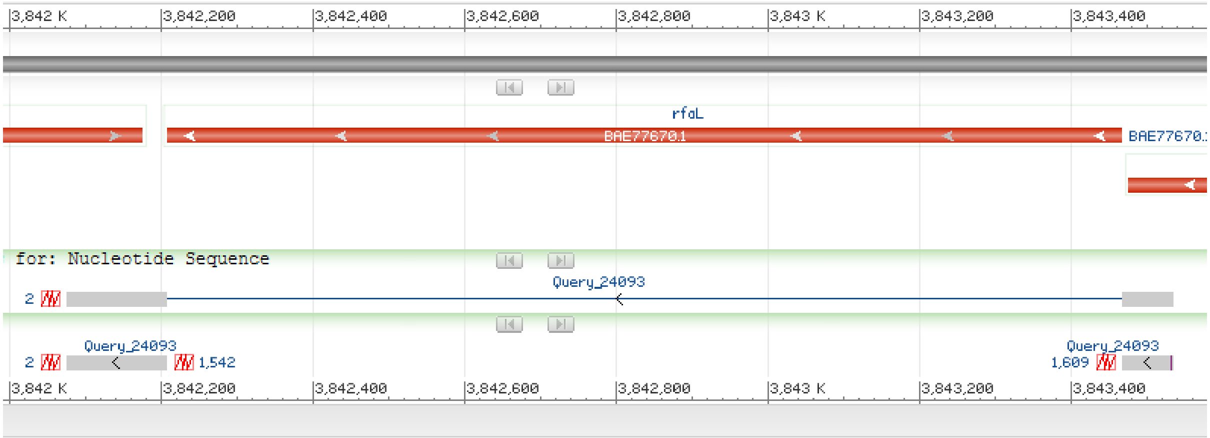


Figure 5.15 NCBI display of aligned mutant 2EWL7 sequence against reference genome showing query rfaL sequence not found.

5.5 Discussion

In this chapter, Next Generation Sequencing techniques and tools were employed for understanding the genetic changes resulting from the random mutagenesis of *E. coli* cells which were selected for their higher cell surface display of mannose compared to the WT W3110 cells. All three strains had different growth rates with the mutant strain 7HS2 displaying higher exponential growth compared to the remaining 2 strains. Mutant strain 2EWL7 exhibited biphasic and slower growth compared to the WT and other mutant strain. However, they all reached the stationary phase at similar times, after 9 hours while remaining stable for the rest of the experiment (Figure 5.5). The table (5.1) shows 7HS2 mutant had a significantly higher specific growth rate (and lower doubling time) than W3110, whereas there was no significant difference between 2EWL7 and the WT strain at the initial stages.

Following sequencing read results, quality scores were assigned based on the equivalent base call accuracy of the sequence read. From the assessment of all strains, the base calling is acceptable on the interquartile ranges falling within the adequate quality section. The mean sequence quality of the bases (Phred score) for all strains assessed was above 33 (Fig. 5.6) which indicates the base calling was above 99.9% assured from the sequencing data (see Appendix C, page 136). There was slight quality deterioration with increased read position to the end which is consistent with observed outputs with Illumina technology due to a cycle in the sequencing process. There were no overrepresented sequences in the read data for all the strains sequenced.

FastQC quality analysis was essential for the sequences to make sure the calls are correct from the samples and not artefacts. While it is possible to identify sequencing variants by viewing the bam files on IGV, Freebayes rules out background noise and sequencing errors coupled with statistical testing to ensure each variant call is properly defined. The Variant Effect Predictor tool from Ensembl Bacteria was able to allocate correct SO terms and consequences to the positional variants that have been identified.

Sequence Onthology (SO) refers to terms used to define the consequence of the effect of a specific variant identified in a gene transcript. The variant call process is based on processing the sequencing information against what is held in the database such as the cDNA, CDS coordinates, amino acid coding transcripts and codons affected. The severity assigned to the variations is based on the impact the identified set of consequences will have on the allele being considered. Although not all the variant genes affected resulted in amino acid changes, these variants have also been called based on the genes being identified as affecting a transcript feature. These transcript feature variants are seen mapped within metabolic and biosynthetic gene clusters (Fig. 5.10) indicating these identified features are nestled within these clusters and could have effects on the up or down regulation of the genes in the clusters. 46.7% of the genes are involved in molecular catalytic activity, followed by 30% in binding, 16.7% in transport activity and 3.3% each in structural and regulator activity (Fig. 5.12). This distribution also cuts across 12 different pathways predominantly affecting biosynthetic pathways within *E. coli*.

The folD, acnA, missense variants with moderate impact protein coding biotype and the non-coding transcript exon variants intQ and gatC pseudogene biotype with modifier impact was found to be conserved in the 3 (WT W3110, 2EWL7 and 7HS2 mutant) strains with the same codons and amino acid identifiers. While yedJ was identified as both a synonymous and missense variant in both mutant strains, it was called as only a synonymous variant with low impact in the WT strain. The amino acid change associated with this variation from a V to S/N suggests this point change in the protein coded for might be responsible for some of the phenotypic changes observed in the mutant strains.

In all strains, there were certain variants with amino acid changes at multiple locations within the same gene. In the araC gene of the wild type strain, different amino acid changing synonymous variants were identified at 7 different locations within the gene. While synonymous variants are classified as low impact, multiple amino acid changes within the gene could have a cumulative downstream effect on the protein structure. All 3 strains had the rpoS Sigma S factor gene identified as a high impact stop gained variant. This gene plays a central role in the cells adaptation mechanism to suboptimal growth conditions by controlling the expression of other genes (Schellhorn, 2020) and as such may be located in a highly conserved region of the genome of the WT W3110 to have remained unchanged and passed down to the mutants.

The low impact protein coding biotype gene – araC and missense variants protein coding genes yabI and lacZ with moderate impact were only found in the WT W3110 but absent in the mutant strains.

rRNA genes were also called as variants in all 3 strains, however only the rrlE gene was conserved in all 3 strains while the mutant strains had other rRNA genes identified as non-coding transcript exon variants with modifier impacts. Investigations into the specific effects these genes have on strain characteristics is key in understanding the increased mannose availability within the cells.

The increase in mannose availability in the EMS mutated strains can be likely attributed to the upregulation of several enzymes involved in nutrient and energy producing pathways as well as the carbohydrate biosynthetic pathway (Table 5.2). A synergistic or co-dependent relationship between the biosynthetic and degradation pathways influenced by these variant genes could account for the increased mannose quantities attributed to these cells. The interlinked pathways highlighted in the KEGG Metabolic Pathway map (Figure 5.11) are also mainly energy intensive/generating pathways suggesting these cells are involved in higher energy requiring processes than the WT strains.

Trehalose-6-phosphate synthase (*otsA*) identified as a transcript feature variant in the 2 mutant strain~~ss~~ catalyzes the first step in the biosynthetic reaction involving the conversion of UDP-α-D-glucose + D-glucopyranose 6-phosphate → UDP + α, α-trehalose 6-phosphate + H^+. In addition to being nonreducing, trehalose possesses several unique properties such as high hydrophilicity, chemical stability, nonhygroscopic glass formation and no internal hydrogen bond formation. The combination of these features explains the principal role of trehalose as a stress metabolite. Coupled with the fact that this gene along with the amino acid – stop gained high impact Sigma S factor gene rpoS were both identified as variants in the mutant strain, the expression in an rpoS mutant strain could account for the increased fitness displayed (Stoebel *et al.*, 2009) by both mutants compared to the WT W3110 strain as observed in Fig. 5.5.

Lewis and colleagues confirmed oxidative stress causes an upregulation in mannose biosynthesis and glycoslysis while down regulating hexosamine biosynthesis and acetyl-CoA formation (Lewis *et al.*, 2016). It is likely that the mutations in the corresponding pathways within the *E. coli* strains have emerged from this same factor.

With the amino acid changes identified and locations within the gene known, targeted metabolic engineering can be used to understand the enhanced mannose production in these cells through proteomics study. Other isolates identified from the FACS mutant selection experiments in the previous chapter could also be sequenced and analysed for their precise genetically enhanced properties.

Chapter 6: Identification and characterization of non-targeted N-glycosylation in native *E. coli* proteins using shotgun glycoproteomics

6.1 Summary

In this chapter, a bottom-up strategy was used to investigate the presence of non-targeted N-glycosylation of endogenous *E. coli* proteins in a glyco-competent strain containing a glycosylation machinery. The mass spectrometry approach was employed to investigate the presence of N-glycoproteins from the LC-MS/MS run of periplasmic protein samples. Various glycoproteomics tools were used to analyse the samples and predict N-glycopeptide candidates from the *E. coli* proteome. Following initial data processing in MaxQuant, the data was further analysed using glycosylation directed bioinformatics tools to enable prediction of potential N-glycoprotein candidates while comparing data across multiple platforms to validate the identified candidates. A list of potential endogenous proteins with predicted presence of required bacterial N-glycosylation sequon was curated and presence of HexNAc oxonium ions at an additional retention time region in the spectrum suggests the presence of a unique peptide with N-glycan attached. Strategies to further validate these results are discussed as well as potential direction for future experiments highlighted.

6.2 Introduction

Glycosylation was widely believed to take place exclusively in eukaryotes until a few decades ago, further research has shown that not only can bacteria glycosylate protein, a few prokaryotic native glycoproteins have been identified over time. With the first isolation of the cell surface protein in *Halobacterium halobium*, over 70 other bacterial glycoproteins have been found to exist as either surface or secreted proteins (Wang *et al.*, 2012; Szymanski and Wren, 2005). In recent years however, only 3 (O-glycosylated proteins) have been identified in *E. coli* with 2 of these – the adhesion involved in diffuse adherence I (AIDA) and adhesion-invasion protein (TibA) found only in pathogenic *E. coli* strains. The autoaggregation factor antigen 43 (Ag43) has been found to exist in both pathogenic and non-pathogenic *E. coli* strains (Wang *et al.*, 2012).

With the minimal understanding of the biochemical and cellular functions within prokaryotes for glycosylation, there is a possibility of discovering more glycoprotein existence in *E. coli* as prokaryotic glycoprotein diversity could very well mean glycosylation is more common in them than initially predicted. Recent research alludes to the existence of even more widespread protein glycosylation in prokaryotes (Schäffer *et al.*, 2017) and prompts this research chapter to investigate the existence/evidence of more glycosylation particularly non-targeted N-glycosylation within *E. coli*. This chapter will focus on using the pgl2 glycosylation machinery from *Campylobacter* sp. to test for non-targeted glycosylation of native *E. coli* proteins within the cell factory.

This study is aimed at identifying these N-glycoprotein candidates because of the attendant effects and amplified metabolic stress (Rosano and Ceccarelli, 2014) non-targeted glycosylation within the cell will have on recombinant protein N-glycosylation efficiency within the same cell factory.

Large-scale characterization of peptides and proteins and the identification of their structure and function is referred to as proteomics. This study generates information on protein abundance, variations and polymorphisms, modifications, and their interactions and networks in cellular processes. In proteomics, a variety of hardware and software tools are used to construct the protein and peptide profiles in an organism. These include tools for the detection and analysis of protein functions from 2D polyacrylamide gels, liquid chromatography combined with tandem mass spectrometry, affinity-tagged proteins as well as two-hybrid assays. All the information generated are curated in a number of public databases and are available on internet sites such as ExPASy and PRIDE for understanding proteomics and protein-protein interactions (Kulski, 2016).

The use of mass spectrometry (MS) methods for proteomics analysis has advanced and with optimization, better understanding into the glycan structures and localization within the glycoprotein profile is achievable (Ohyama *et al*., 2020). A bottom–up approach consisting of proteolytic digestion of the sample glycoprotein and LC-MS/MS analyses is useful in probing glycopeptide suspects (Schirm *et al.*, 2005). Most analytical approaches have used the bottom-up strategy in glycoeptide analysis using both chemical and enzymatic methods (Liu *et al.*, 2014) and this experiment will be employing the same strategy by using trypsin enzymes to digest intact peptides for MS analysis.

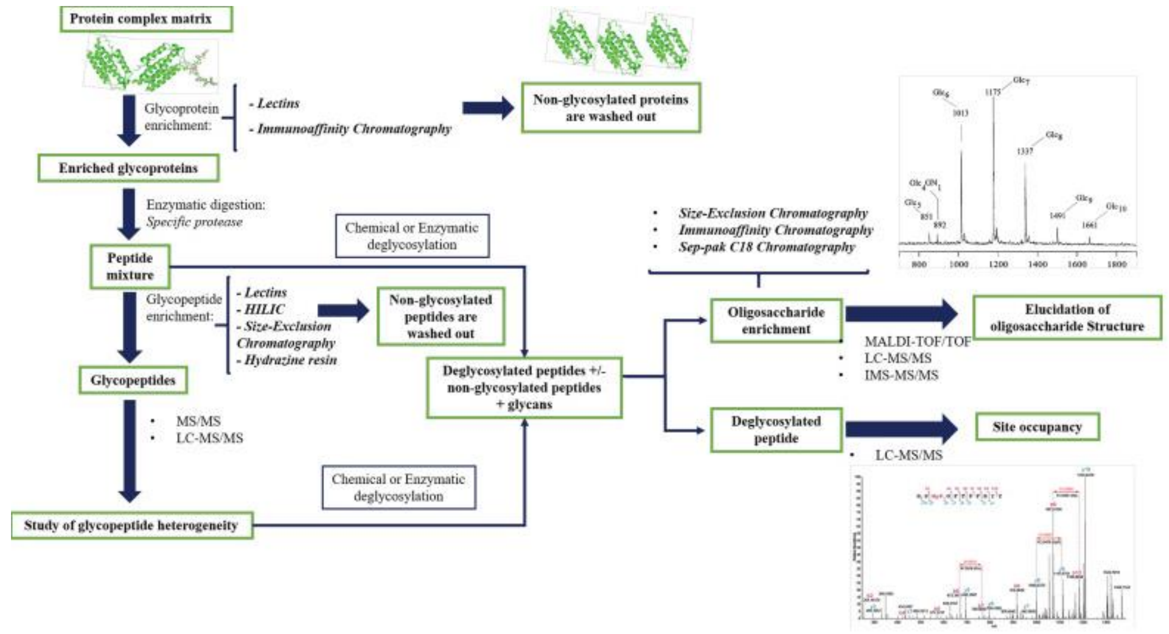

Protein complex matrix
Glycoprotein enrichment:
- Lectins
- Immunoaffinity Chromatography
Non-glycosylated proteins are washed out
Enriched glycoproteins
Enzymatic digestion: Specific protease
Chemical or Enzymatic deglycosylation
Peptide mixture
Glycopeptide enrichment:
- Lectins
- HILIC
- Size-Exclusion Chromatography
- Hydrazine resin
Non-glycosylated peptides are washed out
Glycopeptides
Deglycosylated peptides +/- non-glycosylated peptides + glycans
MS/MS
LC-MS/MS
Chemical or Enzymatic deglycosylation
Study of glycopeptide heterogeneity
Size-Exclusion Chromatography
Immunoaffinity Chromatography
Sep-pak C18 Chromatography
Oligosaccharide enrichment
MALDI-TOF/TOF
LC-MS/MS
IMS-MS/MS
Elucidation of oligosaccharide Structure
Deglycosylated peptide
LC-MS/MS
Site occupancy
Glc6
1013
1175
Glc7
Glc8
1337
Glc4GN1
Glc5
851
892
Glc9
1499
Glc10
1661
800
1000
1200
1400
1600
1800

Figure 6.1 Current workflow for MS-based glycoproteomics (Illiano *et al.*, 2020).

This chapter sought to explore the possibility of non-targeted native protein N-glycosylation within the *E. coli* cell factory and specifically the use of bioinformatics resources for the identification of N-glycosylation sites as well as the characterization of the glycopeptides and glycan structures identified. A widely characterized glycocompetent *E. coli* strain CLM24 with a known recombinant protein AcrA and the glycosylation machinery pgl2 was used to test this hypothesis. Experimental LC-MS/MS data obtained from *E. coli* protein samples were processed and analysed for N-glycan and glycopeptide profiling. MaxQuant software was used for the initial proteomics analysis of the data to identify the protein profiles contained within the sample. While it is not specifically designed for glycan analysis, the information generated from it was useful in probing the data for glycopeptide identification.

6.3 Specific methods

Glycosylation machinery pgl2 was transformed into *E. coli* CLM24 with a known recombinant glycoprotein AcrA and the total uninduced protein extract was harvested according to standard protocol. Control data was generated by extracting total periplasmic protein expressed and extracted using the methods described in section 3.3 from *E. coli* CLM24 carrying either the recombinant glycoprotein AcrA plasmid or the glycosylation machinery pgl2 alone. Samples were prepared for LC-MS/MS analysis following the methods as described in section 3.6 and run with two different MS settings - one directed to protein identification and the other one directed towards glycopeptide analysis (Yang *et al.,* 2018).

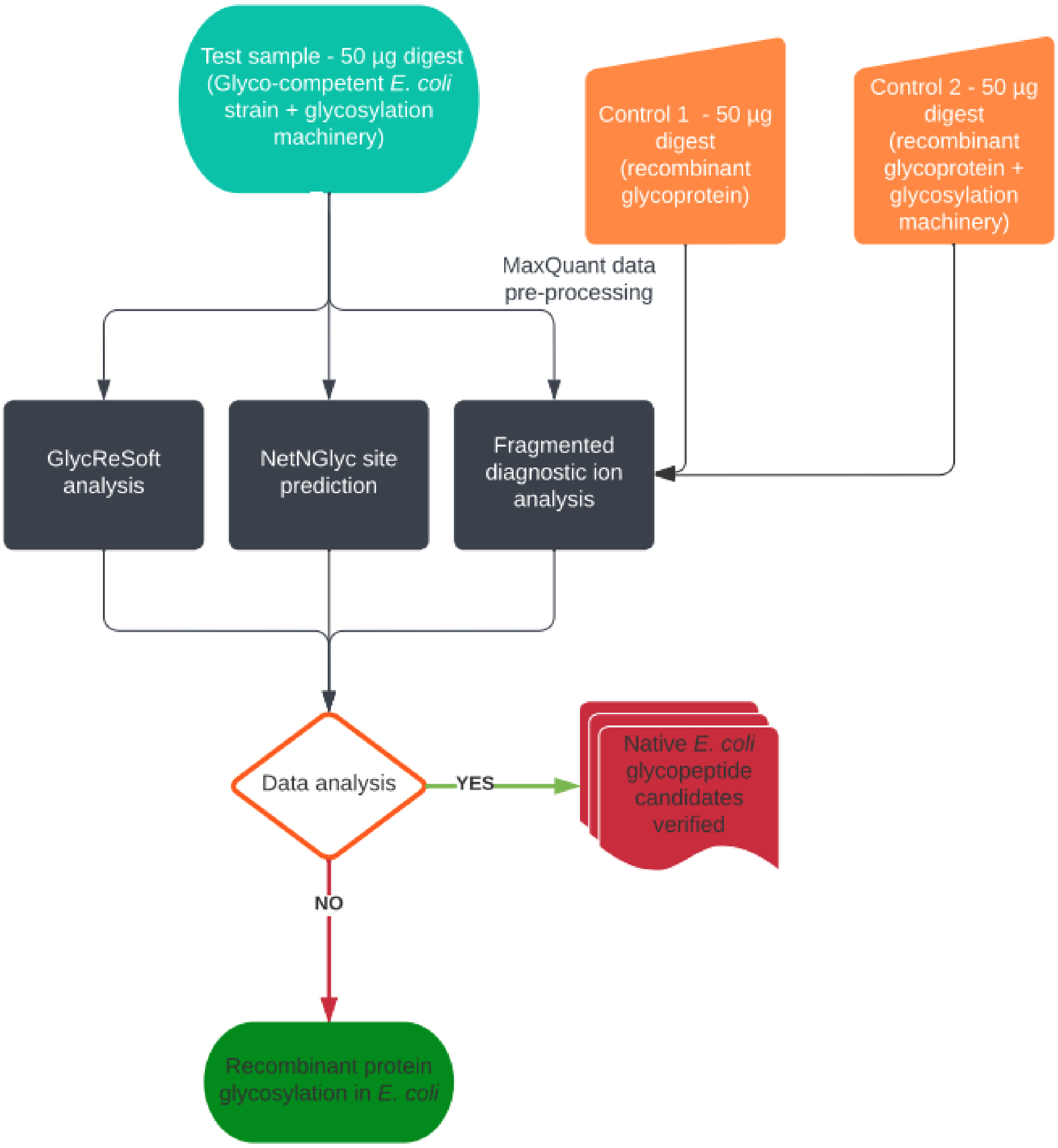

Test sample - 50 µg digest (Glyco-competent *E. coli* strain + glycosylation machinery)
Control 1 - 50 µg digest (recombinant glycoprotein)
Control 2 - 50 µg digest (recombinant glycoprotein + glycosylation machinery)
MaxQuant data pre-processing
GlycReSoft analysis
NetNGlyc site prediction
Fragmented diagnostic ion analysis
Data analysis
YES
Native *E. coli* glycopeptide candidates verified
NO
Recombinant protein glycosylation in *E. coli*

Figure 6.2 Flowchart of specific methods

All 3 protein samples for this experiment were uninduced extracts from the glycocompetent *E. coli* strain CLM24. ~50 µg concentration of each was prepared for in-solution digestion and subsequent analysis. The specific plasmids contained and expected functional outcomes from analysis of each sample is indicated in Table 6.1 below:

Table 6.1: Sample description and expected functional outcome

Sample	Recombinant protein present?	Glycosylation machinery present ?	Functional outcome
Control 1	Yes	No	No glycosylation
Control 2	Yes	Yes	Possibility of AcrA glycosylation in sample
Test	No	Yes	Native protein N - glycosylation

6.3.1 LC-MS/MS data analysis

6.3.1.1 MaxQuant pre-processing

Raw file data generated (from Thermo instruments) was loaded unto the pre-processing software tool MaxQuant (version 2.0.3.0) (Cox and Mann, 2008) for initial analysis. Maintaining default global parameters, carbamidomethyl (C) was selected as fixed modification in the protein quantification and acetyl (protein N-term) and oxidation (M) as variable modifications. Trypsin/P enzyme was selected as the specific proteolytic digestion with up to a maximum of two missed cleavages. The first and main digestion were set to 20 ppm and 4.5 ppm mass tolerances respectively and amino acid minimum length of seven was set for peptides. A false discovery rate (FDR) limit of 1% was used in filtering the result while the remaining settings were left as standard.

6.3.1.2 GlycReSoft glycopeptide analysis

The GlycReSoft software (version 0.4.3) was used to analyse the raw data generated from LC-MS/MS runs. Conversion of the .raw file to .mzml index input format required by GlycReSoft was done using the Thermo RAW file converter tool on Galaxy Europe interface. A combinatorial GalNAc N-glycan search space was built with the following parameters for monosaccharides: HexNAc lower bound 1 and higher bound 6 based on the glycan structure from plasmid pgl2 (Schwarz *et al.*, 2010) with both reduction and derivatization parameters left as default native to generate 5 different compositions of HexNAc and their theoretical masses. Custom glycopeptide search space was also built using the output from the N-glycan search space. For specific proteolytic digestion, trypsin enzyme was selected and the missed cleavages allowed parameter was set to 2 and a maximum of 1 glycosylation per peptide. Fixed modification selected was carbamidomethyl (57.021464 Da), a grouping tolerance of 15ppm was set while the minimum fit isotopic score was set to 20. Minimum MS^1 score filter of 3 was used and a 10ppm error tolerance was fixed. A protein list for the *E. coli* K-12 reference proteome from UniProt (ID UP000000625 downloaded 13th November 2021 containing 4,438 proteins) was used for the search.

6.3.1.3 N-glycosylation site prediction with NetNGlyc server

NetNGlyc – 1.0 online server tool (Gupta and Brunak, 2002) was used to predict potential N-glycosylation sites for comparison with data generated from MaxQuant. With a confidence threshold of 0.5, only the Asn-Xaa-Ser/Thr consensus sequon match was used for generating predictions. The FASTA sequences of each protein from the raw file data was downloaded from UniProt and submitted to NetNGlyc site. The number of predicted N-glycosylation sites and sequences were entered in a table against peptide sequences from MaxQuant.

6.3.2 Data analysis using strings interaction tool

Cross-referencing glycopeptide prediction data from MaxQuant against the NetNGlyc data for predicted N-glycosylation sites that match the bacterial sequon requirement of D/E-Z-N-X-S/T (see appendix D Table 2, page 195), a STRINGS interaction map was generated (Szklarczyk *et al.*, 2021).

6.3.3 Data analysis using fragmented peptide data

To assess the presence of all glycoproteins within the samples, the glycan oxonium ion was linked to the corresponding precursor sugar residue based on the observed retention time. The oxonium ions represent different sugars based on the *m/z* values. Hex^+ sugars give an *m/z* value of 163, $HexNAc^+$ an *m/z* of 204 and $HexNAc\text{-}Hex^+$ an *m/z* of 366. The presence of the 204.0867 diagnostic ions in the .mgf data can confirm the occurrence of HexNAc glycosylation with a high degree of confidence, as this is the value of the monosaccharide content of the glycan being investigated on the glycopeptide spectra.

6.4 Results

Peptide mapping is widely used to verify primary sequences and determine the location and type of post translational modification present within the sequence. Data from MS^1 level were used for quantification of identified proteins in the processed raw files. Intensity column (protein groups table of the MaxQuant output) represents the summed up extracted ion current of all isotopic clusters associated with the identified amino acid sequence. Oxonium ion data from the peptides identified in extracted protein from *E. coli* cells with only the glycosylation machinery (test sample) was compared against controls (protein extracts from a strain carrying the known recombinant glycoprotein AcrA without the pgl2 machinery and extracts from a strain with the glycosylation machinery and recombinant glycoprotein AcrA present). The experimental control of protein digest from cells carrying a known recombinant N-glycorotein with no glycosylation machinery (control 1) was built into the design to verify the protein content in the sample data while the digest from cells carrying the glycosylation machinery and known glycoprotein was included as a second control to confirm the glycosylation observed in the test sample can be attributed to the presence of the acceptor sequon and glycosylation machinery. While the protein expression was uninduced, leaky expression of protein has been observed within the *E. coli* strain.

6.4.1 MaxQuant data analysis

A list of 91 glycoprotein candidates with a minimum of 2 unique identified peptides was generated from the MaxQuant combined output folder and was extracted in an uncompressed format into an excel spreadsheet. A full list is contained in Appendix D Table 1, page 167. The data was trimmed by deleting contaminants and sorted into the biological processes that will be influenced as shown below in Fig. 6.3.

The glycoprotein candidates identified within the data were classed based on the specific cellular processes they influence. A high percentage of these are proteins that affect cellular processes while the other majority are classed under proteins which influence metabolic activities.

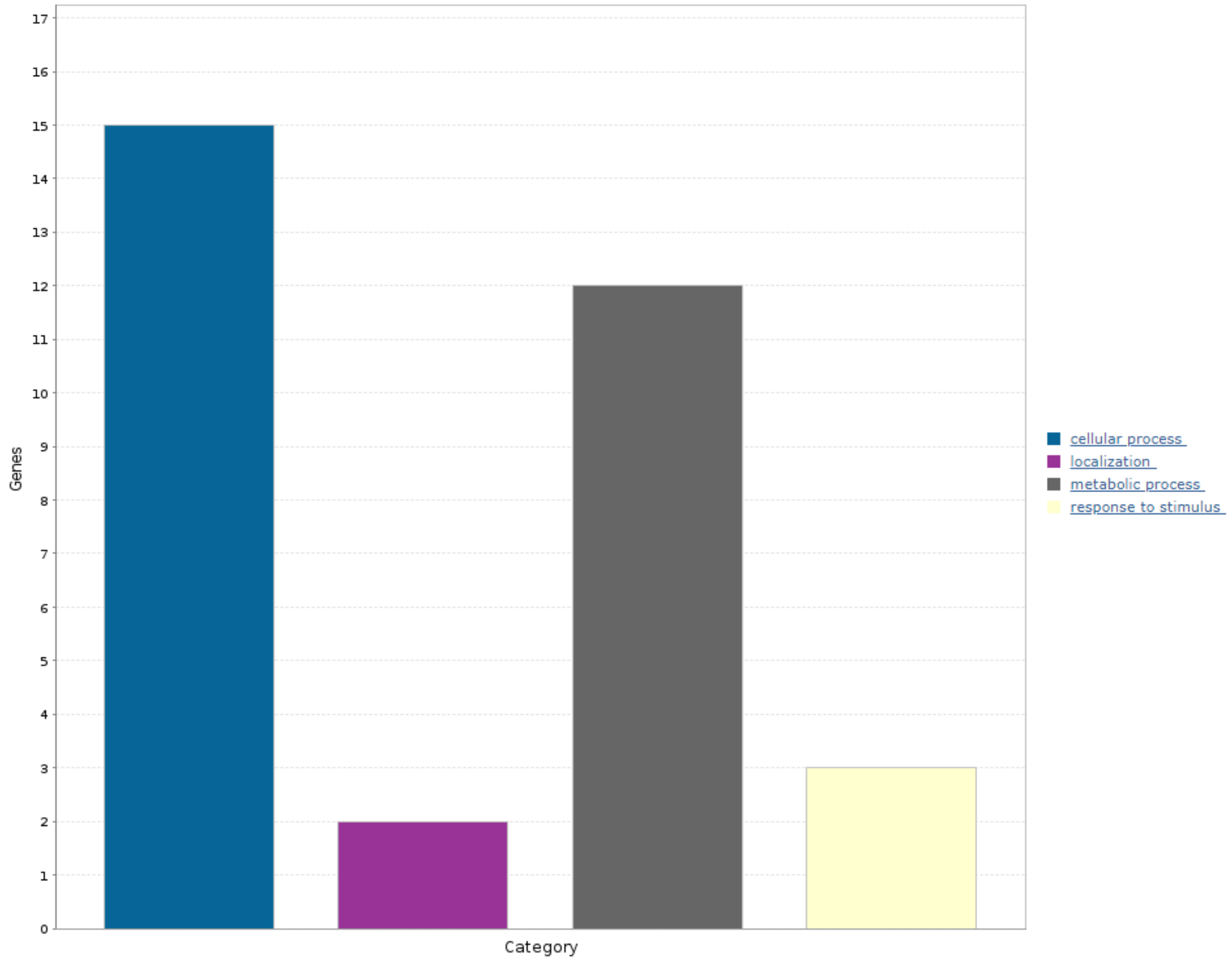


Figure 6.3 Proteins containing at least 2 unique peptides as identified from MaxQuant raw data. Distribution showing various processes within the cell that they influence and data is displayed using PANTHER gene analysis tool.

Mascot Generic Format (.mgf) files generated from the MaxQuant raw file data was analysed using the oxonium ion technique to rapidly detect the presence or absence of glycosylation (Madsen *et al.*, 2018). The presence of the 204.xxx HexNAc oxonium ion fingerprint was investigated from the data and while the same retention time region was identified in both the control and glycosylated sample, an additional region was detected in the glycosylated sample.

6.4.2 Predicted bacterial N-glycosylation sites from NetNglyc

Following the predicted N-glycoprotein candidates with unique peptide sequences generated from MaxQuant, the NetNGlyc server was also used to predict N-glycopeptide candidates were analysed for the consensus bacterial sequon and table 6.2 below highlights the differences between the predicted sites from MaxQuant data and the sites identified with the bacterial N-glycosylation consensus sequence.

Table 6.2: Top 30 predicted N-glycoprotein candidates

Rank	UniProt Entry name	Protein name	MaxQuant unique peptides	NetNGlyc glycosylation site number	Bacterial sequon match*
1	P10384	Long-chain fatty acid transport protein (Outer membrane FadL protein) (Outer membrane flp protein)	1	7	2
2	P0AFK9	Spermidine/putrescine-binding periplasmic protein (SPBP)	1	6	2
3	P23843	Periplasmic oligopeptide-binding protein	2	6	1
4	P0A8V2	DNA-directed RNA polymerase subunit beta (RNAP subunit beta) (EC 2.7.7.6) (RNA polymerase subunit beta) (Transcriptase subunit beta)	1	6	1
5	P23538	Phosphoenolpyruvate synthase (PEP synthase) (EC 2.7.9.2) (Pyruvate, water dikinase)	3	5	1
6	P0A6Y8	Chaperone protein DnaK (HSP70) (Heat shock 70 kDa protein) (Heat shock protein 70)	2	5	1
7	P52697	6-phosphogluconolactonase (6-P-gluconolactonase) (Pgl) (EC 3.1.1.31)	1	5	1
8	P37636	Multidrug resistance protein MdtE	2	3	1
9	P0AFG6	Dihydrolipoyllysine-residue succinyltransferase component of 2-oxoglutarate dehydrogenase complex (EC 2.3.1.61) (2-oxoglutarate dehydrogenase complex component E2) (OGDC-E2) (Dihydrolipoamide succinyltransferase component of 2-oxoglutarate dehydrogenase complex)	1	3	1

10	P07102	Periplasmic AppA protein [Includes: Phosphoanhydride phosphohydrolase (EC 3.1.3.2) (pH 2.5 acid phosphatase) (AP); 4-phytase (EC 3.1.3.26)]	1	3	1
11	P26616	NAD-dependent malic enzyme (NAD-ME) (EC 1.1.1.38)	1	3	1
12	P0C8J8	D-tagatose-1,6-bisphosphate aldolase subunit GatZ	1	3	1
13	P0AAI3	ATP-dependent zinc metalloprotease FtsH (EC 3.4.24.-) (Cell division protease FtsH)	1	3	1
14	P0AG80	sn-glycerol-3-phosphate-binding periplasmic protein UgpB	1	3	1
15	P0A6M8	Elongation factor G (EF-G)	2	2	1
16	P0ABB0	ATP synthase subunit alpha (EC 7.1.2.2) (ATP synthase F1 sector subunit alpha) (F-ATPase subunit alpha)	2	2	1
17	P69797	PTS system mannose-specific EIIAB component (EC 2.7.1.191) (EIIAB-Man) (EIII-Man) [Includes: Mannose-specific phosphotransferase enzyme IIA component (PTS system mannose-specific EIIA component); Mannose-specific phosphotransferase enzyme IIB component (PTS system mannose-specific EIIB component)]	1	2	1
18	P05055	Polyribonucleotide nucleotidyltransferase (EC 2.7.7.8) (Polynucleotide phosphorylase) (PNPase)	1	2	1
19	P0A7V3	30S ribosomal protein S3 (Small ribosomal subunit protein uS3)	1	2	1
20	P0AG86	Protein-export protein SecB (Chaperone SecB)	1	1	1
21	P00509	Aspartate aminotransferase (AspAT) (EC 2.6.1.1) (Transaminase A)	1	6	0

22	P02931	Outer membrane porin F (Outer membrane protein 1A) (Outer membrane protein B) (Outer membrane protein F) (Outer membrane protein IA) (Porin OmpF)	1	6	0
23	P00452	Ribonucleoside-diphosphate reductase 1 subunit alpha (EC 1.17.4.1) (Protein B1) (Ribonucleoside-diphosphate reductase 1 R1 subunit) (Ribonucleotide reductase 1)	1	6	0
24	P25516	Aconitate hydratase A (ACN) (Aconitase) (EC 4.2.1.3) (Iron-responsive protein-like) (IRP-like) (RNA-binding protein) (Stationary phase enzyme)	3	5	0
25	P06996	Outer membrane porin C (Outer membrane protein 1B) (Outer membrane protein C) (Porin OmpC)	2	4	0
26	P13029	Catalase-peroxidase (CP) (EC 1.11.1.21) (Hydroperoxidase I) (HPI) (Peroxidase/catalase)	2	4	0
27	P0C0V0	Periplasmic serine endoprotease DegP (EC 3.4.21.107) (Heat shock protein DegP) (Protease Do)	1	4	0
28	P13482	Periplasmic trehalase (EC 3.2.1.28) (Alpha,alpha-trehalase) (Alpha,alpha-trehalose glucohydrolase) (Tre37A)	1	4	0
29	P16700	Thiosulfate-binding protein	1	4	0
30	P77717	Uncharacterized lipoprotein YbaY	3	3	0

*Bacterial sequon: D/E-X-N-X-S/T (X can be any amino acid except P)

All 109 predicted candidates are listed in Appendix D, page 195 Table 2.

The top predicted N-glycopeptide candidates like the long-chain fatty acid transport protein, the periplasmic oligopeptide-binding protein and the DNA-directed RNA-polymerase subunit beta are all proteins involved in the periplasmic translocation pathway.

6.4.3 Fragmented oxonium ions search for N-glycopeptide candidates

Following the predicted N-glycoprotein candidates with identified unique peptides greater than 2, the fragmented ions detected within the spectra data for the test sample and both controls were investigated for the presence of candidates containing the $HexNAc^+$ glycan oxonium ion *m/z* value of 204.0867. Of the 3 data sets investigated, 204.0867 glyconium ions were only found in the control 2 data. All the identified peptide masses along with their corresponding retention times within the spectra of Control 2 sample containing AcrA and glycosylation machinery is shown in Table 6.3 below. The presence of AcrA protein production was detected in control 2 protein fractions as evidenced in Appendix D Fig. 1, page 205. The spectra chromatograph of the data at the retention time of the 818.4422 glycopeptide candidate with both the glycan oxonium ions 1 and 2 present in the .mgf file is also shown in Appendix D Fig. 2 page 205. No glycan ions were detected in the test and control 1 samples.

Table 6.3: Glycan oxonium ion glycopeptide candidates

S/N	Glycopeptide candidate m/z charge state	Ret. time	Glycan oxonium ion 1 204.0867 Present in the glyco .mgf file Y/N	Glycan oxonium ion 2 186.0761 present in the glyco .mgf file Y/N
1	1068.999889851907 2+	1399.80726	Y	N
2	928.4260341436 2+	1468.50138	Y	N
3	1212.048181284059 2+	1519.68666	Y	N
4	818.442215604665 2+	1748.898	Y	Y
5	809.901316495702 2+	1762.26432	Y	N
6	824.445806799401 2+	1767.6384	Y	N

6.4.4 Protein strings interaction

The proteins containing potential bacterial N-glycosylation sites were mapped using the STRINGS tool to view the protein-protein relationship that have been identified from previous experiments and this is shown in the figure below and data contained in appendix D Table 2, page 195.

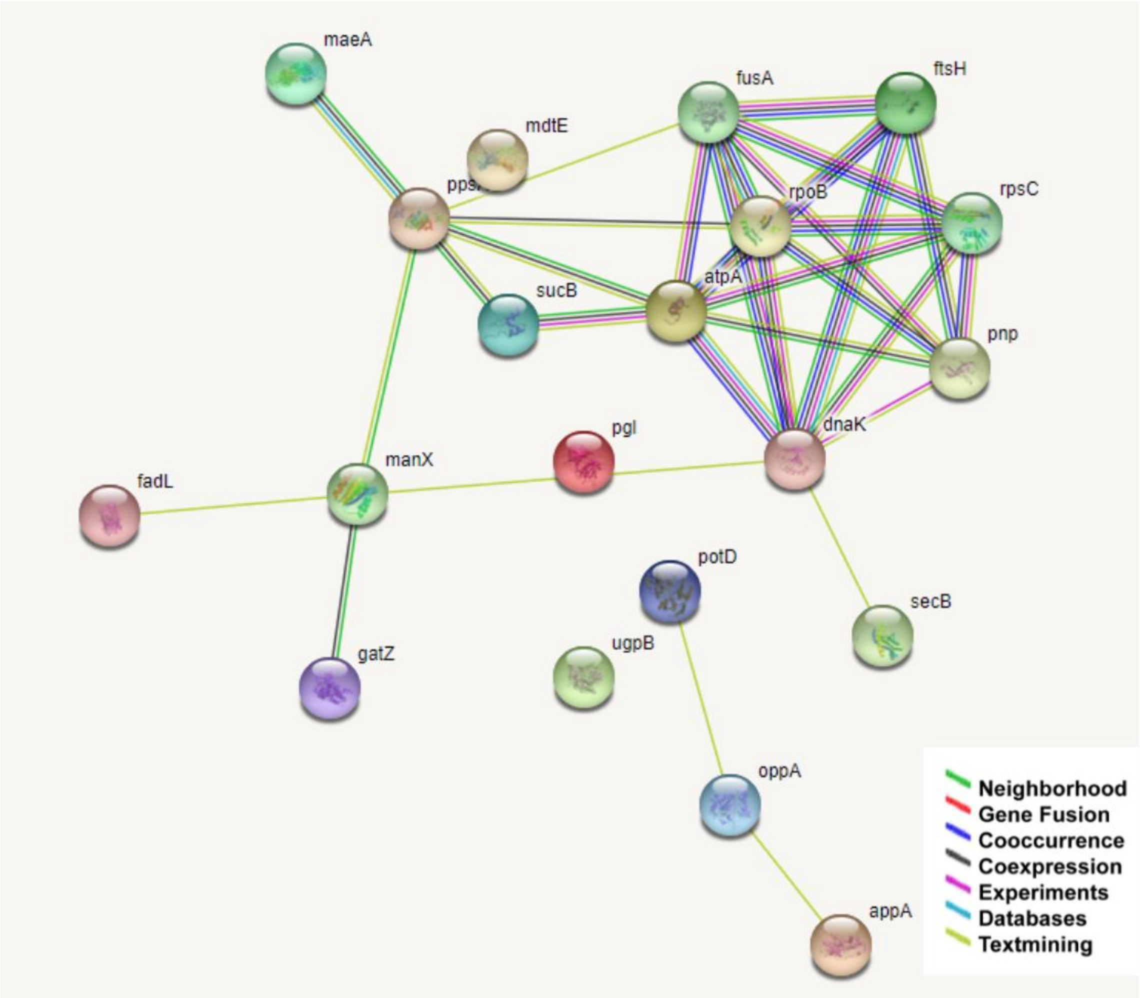


Figure 6.4 The protein-protein interaction of the 20 predicted glycoprotein candidates from NetNGlyc with bacterial acceptor sequon D/E-Z-N-X-S/T match. STRING was used in network mapping with the confidence parameter set to 0.4. The color-coded lines between proteins stand for possible interactions with each color representative of a type of interaction.

The identified predicted glycopeptide candidates have also been shown to have different existing relationships and interactions between them. Some of these proteins are co-located within the same gene region. DNAK protein for instance has multiple layers of interaction with several proteins like atpA, pnp, rpoB, ftsH and fusA suggesting this gene is a key and central component of non-targeted N-glycoprotein production in *E. coli* and these protein candidates have a common property which makes them susceptible to non-targeted N-glycosylation.

6.5 Discussion

The *E. coli* strain CLM24 used for this study is strategically set up to constitutively express the glycosylation machinery as the glycans are built up in the periplasmic membrane while awaiting protein expression before it can be exported. Based on this, non-targeted N-glycosylation of native *E. coli* proteins was checked by harvesting the proteins within the periplasm of the CLM24 strain containing the glycosylation machinery pgl2. Using a high-throughput mass spectrometry/ proteomics strategy, this study was unable to identify non-targeted native N-glycoproteins within the *E. coli* proteome. By employing the stepped approach in the MS run of samples which would have ensured that both the peptide backbones and glycan chains of the glycopeptides were fragmented while still being recorded in a single MS/MS spectrum (Yang, Yang and Sun, 2018) the protein extract was analysed with data interpretation by standard bioinformatics tools and techniques.

As the LC-MS/MS run was set up using parameters skewed towards identifying glycopeptides, the .raw output file generated was pre-processed using MaxQuant for protein and unique peptide identification. From the .mgf data file peptide masses with charges ≥2+ were investigated and a peptide with a mass of 818.4422 was found to contain both the glycan oxonium ion fragment of 204.XXX and also the additional second glycan ion *m/z* value of 186.0761 in the control extract from the recombinant protein AcrA and glycosylation machinery sample demonstrating to a high degree of confidence the ability to identify the presence of glycopeptides within test result parameters. This was however not present in spectra data for the test extract from strain containing glycosylation machinery only or the second control of extracts from recombinant protein (no glycosylation machinery). Glycopeptide candidates corresponding to this could not be validated by a parallel prediction in GlycReSoft using the HexNAc targeted search space designed to identify glycosylation in sample data. The identity of the glycoprotein candidate could thus not be confirmed in this approach. Perhaps sample protein enrichment of the glycopeptides should be employed before MS-glycoproteomics analysis (Yang *et al.*, 2020). With the wide range of enrichment approaches available (Riley, Bertozzi and Pitteri, 2021), exploring this could lead to identification of more glycopeptide peaks. The use of lectin affinity probe (Wu *et al.*, 2019) could be considered in identifying the possibility of non-targeted N-glycosylation within the cell. Also worthy of note is adopting strategies similar to BEMAP mass spectrometry for O-glycosylation in enterotoxigenic *E. coli* (Boysen *et al.*, 2016) to probe N-glycopeptides from our sample.

Another method to be considered for N-glycan native protein identification is the use of the PNGase F enzyme. This enzyme from *Flavobacterium meningosepticum*, catalyzes the hydrolysis of N-linked high mannose residues and other hybrid or complex oligosaccharides from glycoproteins (Elder and Alexander, 1982). A PNGase F cleavage converts the asparagine residue to an aspartic residue which can be identified by mass spectrometry (Aebersold *et al.*, 2003) as it cleaves the β-aspartylglucosamine bond between the innermost GlcNAc of N-glycans and asparagine residues of the glycoproteins.

While the NetNGlyc server is set up for prediction of N-glycosylation sites in human proteins by examining sequence matches to the Asn-Xaa-Ser/Thr sequons, the results generated were further analysed for the presence of the recognized bacterial sequon requirement (D/E-Z-N-X-S/T) where Z and X are not P. The biological processes within these identified glycoprotein candidates (Figure 6.3) points to these predicted N-glycopeptides existing within protein regions involved in cellular and metabolic activities which could ultimately place added stress on the cellular N-glycosylation process and reduce N-glycosylation efficiency of recombinant proteins in the cell.

Although this experiment has been unable to validate the possibility of non-targeted N-glycosylation of native proteins within *E. coli*, it has however generated leads into focus areas and other methods to be considered in releasing glycopeptide molecules from sample protein while utilizing existing glycoproteomics workflow and bioinformatics tools in identifying N-glycopeptides of interest.

The protein interaction information generated from the STRING mapping (Fig. 6.4) could be useful in understanding commonalities between these predicted glycopeptide containing candidates which make them susceptible to non-targeted N-glycosylation particularly the proteins with multiple layers of established interactions.

Chapter 7: Final discussion and future work

The main aim of this research work was to identify an *E. coli* cell chassis with increased capacity and more natural glycosylation substrate availability for efficient recombinant glycoprotein production. The focus was not only identifying this chassis but also gaining a better understanding into the mechanism and unique genetic differences this cell possesses over the existing glyco-competent *E. coli* strains. This was achieved by using a combination of genetic cell engineering techniques and bioinformatics tools as described in chapters 4, 5 and 6. The main conclusions of each chapter will be summarized while discussing future directions for the research findings.

7.1 A flow cytometric approach to using EMS-induced mutagenesis in *Escherichia coli* for improved mannose production

In chapter 4, the pYCG plasmid which consists of a synthetic pathway for the site specific glycosylation of proteins with the eukaryotic trimannosyl chitobiose glycan – mannose$_3$-N-acetylglucosamine$_2$ (Man$_3$GlcNAc$_2$) was transformed into the *E. coli* W3110 strain with the O-antigen ligase WaaL still intact thus ensuring cell surface display of lipid-linked oligosaccharides Man$_3$GlcNAc$_2$ glycans through the actions of the LPS transport system in the cells. Subsequent chemical mutagenesis and fluorescence based screening through flow cytometry was used to target cells with higher surface display of mannose for further characterization.

This experiment has contributed by not only validating flow cytometry as a useful tool in cell assay and isolation, but also generating 2 mutant strains with significantly higher levels of cell surface mannose display. With these new strains however, it would be of interest to analyze the effect of further random mutagenesis on the mannose availability within the cells. A comparison of the effects across multiple generations could further enhance our understanding of the mechanism and cellular changes occurring within the *E. coli* chassis. Investigating ways to label the natural mannose availability within the cells to quantify amounts that are eventually channelled into the glycan synthesis pathway would be useful in correlating mannose availability with glycosylation efficiency.

Also of interest would be investigating the application of this new method in identifying cells with improved substrate availability for other cellular process requirements such as the cell surface display of CMP-Neu5Ac in screening for higher sialylation competent *E. coli* cells (Zhu *et al.*, 2020).

While some studies have linked increased mannose levels to oxidative stress in CHO cells (Lewis *et al.*, 2016), it would be of interest to investigate the effect of oxidative stress on the mannose availability within the mutant cells.

7.2 Sequencing and characterization of mannose substrate enhanced *Escherichia coli* strains for N-glycoprotein production efficiency

In chapter 5, the isolated higher mannose fluorescing mutant cells were subjected to Next Generation Sequencing methods and the resulting output was compared to known *E. coli* K-12 MG1655 reference genome sequence to identify variations in the mutant strains and assign biological consequences to these variants. The bioinformatics tools and methods used in this experiment provided a better insight into affected pathways within the mutant strains and specific genes that could be further targeted or engineered to better understand or optimize mannose substrate availability within the strains.

The next steps in verifying the capabilities of these mutant strains to enable further engineering and characterization would also involve removing the kanamycin antibiotic resistance gene from the new mutant (2EWL7 ΔrfaL and 7HS2 ΔrfaL) strains (Figs. 5.14 and 5.15) and testing glycosylation of a recombinantly expressed glycoprotein in them against the same glycoprotein expressed in a well characterized glyco-competent *E. coli* CLM24 strain which is derived from the common ancestral strain – *E. coli* W3110.

Targeted upregulation or downregulation of gene combinations could be used to further understanding into the mannose biosynthesis pathway. For example, downregulation of genes involved in hexosamine biosynthesis and Acetyl-CoA formation pathway to measure effects on mannose availability. It is worth exploring if these identified genes are under the influence of strong promoters (Glasscock *et al.*, 2018) as engineering or testing the promoters to characterize their effect on mannose availability within the strain could lead to better understanding of the genetic changes that are desirable in the strain.

It would also be of interest if the mutant strains are evolved using a culture media that has been optimized to enhance specific characteristics in microbial growth to check for the effect on the mutant strains' phenotype. Determination of whether displayed phenotypes in the mutant strains depend on the presence of initial variations identified between the WT W3110 and its parent reference genome – K-12 MG1655 or the new variant genes identified between the WT and 2 chemically mutated strains mutations are dependent or independent occurrences.

Mutations upstream of certain genes have been known to increase protein abundance (Morgenthaler *et al.*, 2019). Investigating the verified variants with amino acid changes and their regulatory effects on downstream genes would be useful information. Checking to see if these variations lie in promoter regions of genes that feed directly into carbohydrate biosynthesis would be key to understanding cellular processes.

7.3 Identification and characterization of non-targeted n-glycosylation in native *E. coli* proteins using shotgun glycoproteomics

In chapter 6, an established glycosylation machinery was transformed into a glycocompetent *E. coli* strain – CLM24 with the aim of analysing the native periplasmic proteins for non-targeted N-glycosylation using a bottom up glycoproteomics approach. The LC-MS/MS generated raw data was processed using a series of mass spectrometry analysis tools to predict the glycoprotein/peptide composition within the sample.

The goal of this experiment was to investigate the possibility of *E. coli* N-glycosylating its own native proteins in the presence of the right machinery to support this process. Validating this would have ultimately required further investigation into what this might mean for the efficiency of recombinant protein N-glycosylation within the same cell. While the methods used in chapter 6 have been unable to clearly validate this, other methods as discussed in the chapter would be worth considering. Paramount in this would be exploration of various glycopeptide enrichment approaches to establish the predictions that have been generated from the experiment so far. Further to this would be checking if more mannose availability within the mutant strains in chapter 4 would translate into discovery of novel native protein glycosylation within the cellular proteome. Another direction to be considered would be the effect a different glycosylation machinery plasmid would have on non-targeted native protein glycosylation in *E. coli*.

7.4 Final remarks

The central aim of this thesis was to identify an enhanced glyco-competent *E. coli* strain using cellular engineering methods while developing a process for selection and getting a deeper understanding of the underlying genetic changes that are responsible for the phenotype. Random mutagenesis and flow cytometric methods were used to achieve this while sequence characterization using bioinformatics tools was useful in understanding the genetic changes that have occurred within the strains. With the potentially predicted non-targeted N-glycosylation native *E. coli* protein candidates, it begs to postulate that this would subsequently affect N-glycosylation efficiency of recombinant proteins within the *E. coli* genus.

These strains with the increased mannose substrate available in the cells could be the key to achieving higher glycosylated protein titres within the *E. coli* cell factory while making it competitive with other glycoprotein production cell factories. The continued understanding of gene interactions could yield an optimized candidate suitable for wider industrial use. The experiments in this thesis aims to contribute to the development of an enhanced glyco-competent strain for use in various bioprocessing applications.

References

Aebersold, R., Zhang, H., Li, X.-j. and Martin, D. B. (2003) 'Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry', *Nat Biotechnol,* 21(6), pp. 660-666.

André, E., Michael, E. & Thomas, E. (2013) Expression of recombinant Antibodies. *Frontiers in Immunology*, 4.

Andualema, B. & Gessesse, A. (2012) Microbial Lipases and Their Industrial Applications: Review. *Biotechnology*, 11(3), 100-118.

Anyaogu, D. C., Hansen, A. H., Hoof, J. B., Majewska, N. I., Contesini, F. J., Paul, J. T., Nielsen, K. F., Hobley, T. J., Yang, S., Zhang, H., Betenbaugh, M. and Mortensen, U. H. (2021) 'Glycoengineering of Aspergillus nidulans to produce precursors for humanized N-glycan structures', *Metabolic engineering,* 67, pp. 153-163.

Apweiler, R., Hermjakob, H. and Sharon, N. (1999) 'On the frequency of protein glycosylation, as deduced from analysis of the SWISS- PROT database 1 1 Dedicated to Prof. Akira Kobata and Prof. Harry Schachter on the occasion of their 65th birthdays', *BBA - General Subjects,* 1473(1), pp. 4-8.
Baker, J. L., Çelik, E. & DeLisa, M. P. (2013) Expanding the glycoengineering toolbox: The rise of bacterial N-linked protein glycosylation. Trends in Biotechnology, 31(5), 313-323.

Barolo, L., Abbriano, R. M., Commault, A. S., George, J., Kahlke, T., Fabris, M., Padula, M. P., Lopez, A., Ralph, P. J. and Pernice, M. (2020) 'Perspectives for Glyco-Engineering of Recombinant Biopharmaceuticals from Microalgae', *Cells,* 9(3), pp. 633.

Barth, S., Huhn, M., Matthey, B., Klimka, A., Galinski, E. A. & Engert, A. (2000) Compatible- solute-supported periplasmic expression of functional recombinant proteins under stress conditions. *Applied and environmental microbiology*, 66(4), 1572-1579.

Bayly, A. M., Kortt, A. A., Hudson, P. J. & Power, B. E. (2002) Large- scale bacterial fermentation and isolation of scFv multimers using a heat- inducible bacterial expression vector. *Journal of immunological methods*, 262(1-2), 217-227.

Beck, A. & Reichert, J. M. (2012) Marketing approval of mogamulizumab: A triumph for glyco-engineering. *mAbs*, 4(4), 419-425.

Beilen, J. B. v. & Li, Z. (2002) Enzyme technology: an overview. *Current Opinion in Biotechnology*, 13(4), 338-344.

Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S. & Brunak, S. (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *In:* Appel, R. D. & Bairoch, A. (eds.). Weinheim.

Borrero, J., Jiménez, J. J., Gútiez, L., Herranz, C., Cintas, L. M. & Hernández, P. E. (2011) Protein expression vector and secretion signal peptide optimization to drive the production, secretion, and functional expression of the bacteriocin enterocin A in lactic acid bacteria. *Journal of Biotechnology*, 156(1), 76-86.

Braun, P. & Labaer, J. (2003) High throughput protein production for functional proteomics. *Trends in Biotechnology*, 21(9), 383-388.

Brazier-Hicks, M., Evans, K. M., Gershater, M. C., Puschmann, H., Steel, P. G. & Edwards, R. (2009) The C- glycosylation of flavonoids in cereals. *Journal of Biological Chemistry*, 284(27), 17926-17934.

Butler, M. (2006) Optimisation of the Cellular Metabolism of Glycosylation for Recombinant Proteins Produced by Mammalian Cell Systems. *International Journal of Cell Culture and Biotechnology*, 50(1), 57-76.

Boysen, A., Palmisano, G., Krogh, T. J., Duggin, I. G., Larsen, M. R. and Møller-Jensen, J. (2016) 'A novel mass spectrometric strategy "bEMAP" reveals Extensive O-linked protein glycosylation in Enterotoxigenic Escherichia coli', *Sci Rep,* 6(1), pp. 32016-32016.

Burns, P. A., Allen, F. L. and Glickman, B. W. (1986) 'DNA SEQUENCE ANALYSIS OF MUTAGENICITY AND SITE SPECIFICITY OF ETHYL METHANESULFONATE IN Uvr+ AND UvrB - STRAINS OF ESCHERICHIA COLI', *Genetics,* 113(4), pp. 811-819.

Cao, Z., Partyka, K., McDonald, M., Brouhard, E., Hincapie, M., Brand, R., Hancock, W. & Haab, B. (2013) Modulation of Glycan Detection on Specific Glycoproteins by Lectin Multimerization. *Analytical Chemistry*, 85(3), 1689.

Choi, J. H., Jeong, K. J., Kim, S. C. & Lee, S. Y. (2000) Efficient secretory production of alkaline phosphatase by high cell density culture of recombinant Escherichia coli using the Bacillus sp. endoxylanase signal sequence. *Applied Microbiology and Biotechnology*, 53(6), 640-645.

Christine, M. S. & Brendan, W. W. (2005) Protein glycosylation in bacterial mucosal pathogens. *Nature Reviews Microbiology*, 3(3), 225.

Correa, A. and Oppezzo, P. (2015) 'Overcoming the solubility problem in E. coli: Available approaches for recombinant protein production', *Methods in Molecular Biology,* 1258, pp. 27-44.

Costa, A. R., Rodrigues, M. E., Henriques, M., Oliveira, R. and Azeredo, J. (2014) 'Glycosylation: impact, control and improvement during therapeutic protein production', *Crit Rev Biotechnol,* 34(4), pp. 281-299.

Coulondre, C. and Miller, J. H. (1977) 'Genetic studies of the lac repressor : IV. Mutagenic specificity in the lacI gene of Escherichia coli', *J Mol Biol,* 117(3), pp. 577-606.

Cox, J. and Mann, M. (2008) 'MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification', *Nat Biotechnol,* 26(12), pp. 1367-1372.

Croset, A., Delafosse, L., Gaudry, J.-P., Arod, C., Glez, L., Losberger, C., Begue, D., Krstanovic, A., Robert, F., Vilbois, F., Chevalet, L. & Antonsson, B. (2012) Differences in the glycosylation of recombinant proteins expressed in HEK and CHO cells. *Journal of Biotechnology*, 161(3), 336-348.

Cuccui, J., Terra, V. S., Bossé, J. T., Naegeli, A., Abouelhadid, S., Li, Y., Lin, C.-W., Vohra, P., Tucker, A. W., Rycroft, A. N., Maskell, D. J., Aebi, M., Langford, P. R. and Wren, B. W. (2017) 'The N-linking glycosylation system from Actinobacillus pleuropneumoniae is required for adhesion and has potential use in glycoengineering', *Open Biol,* 7(1), pp. 160212.

Culyba, E., Price, J., Hanson, S., Dhar, A., Wong, C.-H., Gruebele, M., Powers, E. and Kelly, J. (2011) 'Protein Native-State Stabilization by Placing Aromatic Side Chains in N-Glycosylated Reverse Turns', *Science,* 331(6017), pp. 571-575.

Dalit, S.-B. & Yaakov, L. (2008) Effect of glycosylation on protein folding: A close look at thermodynamic stabilization. *Proceedings of the National Academy of Sciences*, 105(24), 8256.

Datsenko, K., A. and Wanner, B., L. (2000) 'One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products', *Proceedings of the National Academy of Sciences of the United States of America,* 97(12), pp. 6640.

David, L. S., Govind, M., Edgar, R., Gerald, S., Linda, E., Salvatore, J. T. & Stephen, M. B. (2000) The role of phosphoglycans in Leishmania– sand fly interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 97(1), 406.

De Beer, T., Vliegenthart, J. F. G., Loffler, A. & Hofsteenge, J. (1995) The hexopyranosyl residue that is C- glycosidically linked to the side chain of tryptophan- 7 in human RNase U(s) is α- mannopyranose. *Biochemistry*, 34(37), 11785-11789.

De Pourcq, K., Tiels, P., Van Hecke, A., Geysens, S., Vervecken, W. and Callewaert, N. (2012) 'Engineering Yarrowia lipolytica to produce glycoproteins homogeneously modified with the universal Man3GlcNAc2 N-glycan core', *PLoS One,* 7(6), pp. e39976-e39976.

Dell, A., Galadari, A., Sastre, F. and Hitchen, P. (2010) 'Similarities and differences in the glycosylation mechanisms in prokaryotes and eukaryotes', *International journal of microbiology,* 2010(2010), pp. 148178-148178.

Ding, N., Yang, C., Sun, S., Han, L., Ruan, Y., Guo, L., Hu, X. & Zhang, J. (2017) Increased glycosylation efficiency of recombinant proteins in Escherichia coli by auto- induction. *Biochemical and Biophysical Research Communications*, 485(1), 138-143.

Doran, P. M. (2000) Foreign protein production in plant tissue cultures. *Current Opinion in Biotechnology*, 11(2), 199-204.

Elder, J. H. and Alexander, S. (1982) 'endo-β -N-acetylglucosaminidase F: Endoglycosidase from Flavobacterium meningosepticum That Cleaves Both High-Mannose and Complex Glycoproteins', *Proc Natl Acad Sci U S A,* 79(15), pp. 4540-4544.

Elizabeth, L. L., Neil, K., Pratibha, B. and Jiyeon, K. (2021) 'Fructose and Mannose in Inborn Errors of Metabolism and Cancer', *Metabolites,* 11(8), pp. 479.

Fakruddin, M., Mohammad Mazumdar, R., Bin Mannan, K. S., Chowdhury, A. & Hossain, M. N. (2013) Critical Factors Affecting the Success of Cloning, Expression, and Mass Production of Enzymes by Recombinant E. coli. *ISRN biotechnology*, 2013, 590587-590587.

Ferrer-Miralles, N., Saccardo, P., Corchero, J. L., Xu, Z. & García-Fruitós, E. (2015) General introduction: recombinant protein production and purification of insoluble proteins. *Methods in molecular biology (Clifton, N.J.)*, 1258, 1-24.

Fisher, A. C., Haitjema, C. H., Guarino, C., Çelik, E., Endicott, C. E., Reading, C. A., Merritt, J. H., Ptak, A. C., Zhang, S. and Delisa, M. P. (2011) 'Production of secretory and extracellular N- linked glycoproteins in Escherichia coli', *Applied and environmental microbiology,* 77(3), pp. 871-881.

Fonseca-Maldonado, R., Vieira, D., Alponti, J., Bonneil, E., Thibault, P. & Ward, R. (2013) Engineering the Pattern of Protein Glycosylation Modulates the Thermostability of a GH11 Xylanase. *J. Biol. Chem.*, 288(35), 25522-25534.

Frushicheva, M. P., Cao, J. & Warshel, A. (2011) Challenges and advances in validating enzyme design proposals: The case of kemp eliminase catalysis. *Biochemistry*, 50(18), 3849-3858.

Fukuda, M., Sasaki, H. & Lopez, L. (1989) Survival of recombinant erythropoietin in the circulation: The role of carbohydrates. *Blood*, 73(1), 84-89.

Garcia-Ruiz, E., Gonzalez-Perez, D., Ruiz-Dueñas, F. J., Martínez, A. T. & Alcalde, M. (2012) Directed evolution of a temperature-, peroxide- and alkaline pH- tolerant versatile peroxidase. *Biochemical Journal*, 441(1), 487-498.

Garrison, E. and Marth, G. (2012) 'Haplotype-based variant detection from short-read sequencing'.

Ghaderi, D., Zhang, M., Hurtado-Ziola, N. & Varki, A. (2012) Production platforms for biotherapeutic glycoproteins. Occurrence, impact, and challenges of non- human sialylation. *Biotechnology and Genetic Engineering Reviews*, 28(1), 147-176.

Gileadi, O. (2017) *Recombinant protein expression in E. coli: A historical perspective*, *1586*.

Givan, A. L. (2010) 'Flow Cytometry: An Introduction', *Methods Mol Biol,* 699, pp. 1-29.

Glasscock, C. J., Yates, L. E., Jaroentomeechai, T., Wilson, J. D., Merritt, J. H., Lucks, J. B. and Delisa, M. P. (2018) 'A flow cytometric approach to engineering Escherichia coli for improved eukaryotic protein glycosylation', *Metabolic Engineering,* 47, pp. 488-495.

Goettig, P. (2016) Effects of Glycosylation on the Enzymatic Activity and Mechanisms of Proteases. *International Journal of Molecular Sciences*, 17(12).

Goldsmith, M. & Tawfik, D. S. (2012) Directed enzyme evolution: beyond the low- hanging fruit. *Current Opinion in Structural Biology*, 22(4), 406-412.

Gotte, G., Libonati, M. & Laurents, D. V. (2003) Glycosylation and Specific Deamidation of Ribonuclease B Affect the Formation of Three- dimensional Domain- swapped Oligomers. *Journal of Biological Chemistry*, 278(47), 46241-46251.

Greene, L. H., Chrysina, E. D., Irons, L. I., Papageorgiou, A. C., Acharya, K. R. and Brew, K. (2001) 'Role of conserved residues in structure and stability: Tryptophans of human serum retinol-binding protein, a model for the lipocalin superfamily', *Protein Sci,* 10(11), pp. 2301-2316.

Gupta, R. and Brunak, S. 2002. Prediction of glycosylation across the human proteome and the correlation to protein function. Pac Symp Biocomput.

Hamilton, S. R. & Gerngross, T. U. (2007) Glycosylation engineering in yeast: the advent of fully humanized yeast. *Current Opinion in Biotechnology*, 18(5), 387-392.

Hamilton, B. S., Wilson, J. D., Shumakovich, M. A., Fisher, A. C., Brooks, J. C., Pontes, A., Naran, R., Heiss, C., Gao, C., Kardish, R., Heimburg-Molinaro, J., Azadi, P., Cummings, R. D., Merritt, J. H. and Delisa, M. P. (2017) 'A library of chemically defined human N-glycans synthesized from microbial oligosaccharide precursors', *Sci Rep,* 7(1), pp. 15907-12.

Han, M. & Yu, X. (2015) Enhanced expression of heterologous proteins in yeast cells via the modification of N -glycosylation sites. *Bioengineered*, 6(2), 00-00.

Hawkins-Hooker, A., Depardieu, F., Baur, S., Couairon, G., Chen, A. and Bikard, D. (2021) 'Generating functional protein variants with variational autoencoders', *PLoS Comput Biol,* 17(2), pp. e1008736-e1008736.

Haynes, P. A. (1998) Phosphoglycosylation: a new structural class of glycosylation? *Glycobiology*, 8(1), 1-5.

Hess, D. & Hofsteenge, J. (1999) Recombinant human interleukin- 12 is the second example of a C-mannosylated protein. *Glycobiology*, 9(5), 435.

Hirabayashi, J. (2004) Lectin- based structural glycomics: Glycoproteomics and glycan profiling. *Glycoconjugate Journal*, 21(1-2), 35-40.

Hossler, P., Mulukutla, B. C. and Hu, W.-S. (2007) 'Systems analysis of N-glycan processing in mammalian cells', *PLoS One,* 2(8), pp. e713-e713.

Hossler, P., Khattak, S. F. & Li, Z. J. (2009) Optimal and consistent protein glycosylation in mammalian cell culture. *Glycobiology*, 19(9), 936-949.

Howe, K. L., Contreras-Moreira, B., De Silva, N., Maslen, G., Akanni, W., Allen, J., Alvarez-Jarreta, J., Barba, M., Bolser, D. M., Cambell, L., Carbajo, M., Chakiachvili, M., Christensen, M., Cummins, C., Cuzick, A., Davis, P., Fexova, S., Gall, A., George, N., Gil, L., Gupta, P., Hammond-Kosack, K. E., Haskell, E., Hunt, S. E., Jaiswal, P., Janacek, S. H., Kersey, P. J., Langridge, N., Maheswari, U., Maurel, T., McDowall, M. D., Moore, B., Muffato, M., Naamati, G., Naithani, S., Olson, A., Papatheodorou, I., Patricio, M., Paulini, M., Pedro, H., Perry, E., Preece, J., Rosello, M., Russell, M., Sitnik, V., Staines, D. M., Stein, J., Tello-Ruiz, M. K., Trevanion, S. J., Urban, M., Wei, S., Ware, D., Williams, G., Yates, A. D. and Flicek, P. (2020) 'Ensembl Genomes 2020—enabling non-vertebrate genomic research', *Nucleic Acids Res,* 48(D1), pp. D689-D695.

Hui Sun, L., Yifei, Q. & Wonpil, I. (2015) Effects of N- glycosylation on protein conformation and dynamics: Protein Data Bank analysis and molecular dynamics simulation study. *Scientific Reports*, 5(1).

Ihssen, J., Kowarik, M., Dilettoso, S., Tanner, C., Wacker, M. and Thöny-Meyer, L. (2010) 'Production of glycoprotein vaccines in Escherichia coli', *Microb Cell Fact,* 9(1), pp. 61-61.

Illiano, A., Pinto, G., Melchiorre, C., Carpentieri, A., Faraco, V. and Amoresano, A. (2020) 'Protein Glycosylation Investigated by Mass Spectrometry: An Overview', *Cells,* 9(9), pp. 1986.

Itakura, K., Hirose, T., Crea, R., Riggs, A. D., Heyneker, H. L., Bolivar, F. & Boyer, H. W. (1977) Expression in Escherichia coli of a chemically synthesized gene for the hormone somatostatin. *Science (New York, N.Y.)*, 198(4321), 1056-1063.

Jaffé, S. R. P., Strutton, B., Levarski, Z., Pandhal, J. and Wright, P. C. (2014) 'Escherichia coli as a glycoprotein production host: Recent developments and challenges', *Current Opinion in Biotechnology,* 30, pp. 205-210.

Jarrell, K., Ding, Y., Meyer, B., Albers, S.-V., Kaminski, L. & Eichler, J. (2014) N- Linked Glycosylation in Archaea: A Structural, Functional, and Genetic Analysis. Washington: American Society for Microbiology.

Jeong, K. J. & Lee, S. Y. (2000) Secretory production of human leptin in Escherichia coli. *Biotechnology and bioengineering*, 67(4), 398-407.

Johansen, P. G., Marshall, R. D. & Neuberger, A. (1961) Carbohydrates in protein. 3 The preparation and some of the properties of a glycopeptide from hen's-egg albumin. *The Biochemical journal*, 78(3), 518-527.

Kightlinger, W., Warfel, K. F., Delisa, M. P. & Jewett, M. C. (2020) Synthetic Glycobiology: Parts, Systems, and Applications. *ACS synthetic biology*.

Kim, S. C., Sprung, R., Chen, Y., Xu, Y., Ball, H., Pei, J., Cheng, T., Kho, Y., Xiao, H., Xiao, L., Grishin, N. V., White, M., Yang, X.-J. & Zhao, Y. (2006) Substrate and Functional Diversity of Lysine Acetylation Revealed by a Proteomics Survey. *Molecular Cell*, 23(4), 607-618.

Kirk, O., Borchert, T. V. & Fuglsang, C. C. (2002) Industrial enzyme applications. *Current Opinion in Biotechnology*, 13(4), 345-351.

Klatt, S. & Konthur, Z. (2012) Secretory signal peptide modification for optimized antibody- fragment expression- secretion in Leishmania tarentolae. *Microbial Cell Factories*, 11, 97.

Kobayashi, T., Nishizaki, R. & Ikezawa, H. (1997) The presence of GPI- linked protein (s) in an archaeobacterium, Sulfolobus acidocaldarius, closely related to eukaryotes. *BBA - General Subjects*, 1334(1), 1-4.

Kudanga, T., Nyanhongo, G. S., Guebitz, G. M. & Burton, S. (2011) Potential applications of laccase-mediated coupling and grafting reactions: A review. *Enzyme and Microbial Technology*, 48(3), 195-208.

Kukuruzinska, M. A., Bergh, M. L. E. and Jackson, B. J. (1987) 'Protein Glycosylation in Yeast', *Annu. Rev. Biochem.,* 56(1), pp. 915-944.

Kulski, J. K. 2016. Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications. IntechOpen.

Latousakis, D. & Juge, N. (2018) How Sweet Are Our Gut Beneficial Bacteria? A Focus on Protein Glycosylation in Lactobacillus. *Int. J. Mol. Sci.*

Lee, S., Kim, I., Kim, D., Bae, K. & Byun, S. (1998) High level secretion of recombinant staphylokinase into periplasm of Escherichia coli. *Biotechnology Letters*, 20(2), 113-116.

Lewis, A. M., Croughan, W. D., Aranibar, N., Lee, A. G., Warrack, B., Abu-Absi, N. R., Patel, R., Drew, B., Borys, M. C., Reily, M. D. and Li, Z. J. (2016) 'Understanding and Controlling Sialylation in a CHO Fc-Fusion Process', *PLoS One,* 11(6), pp. e0157111-e0157111.

Li, L., Peng, W. & Yi, T. (2013) C- glycosylation of anhydrotetracycline scaffold with SsfS6 from the SF2575 biosynthetic pathway. *The Journal of Antibiotics*, 67(1), 65.

Li, S., Yang, X., Yang, S., Zhu, M. & Wang, X. (2012) TECHNOLOGY PROSPECTING ON ENZYMES: APPLICATION, MARKETING AND ENGINEERING. *Computational and Structural Biotechnology Journal*, 2(3), e201209017.

Li, H. (2013) 'Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM'.

Li, S.-T., Lu, T.-T., Xu, X.-X., Ding, Y., Li, Z., Kitajima, T., Dean, N., Wang, N. and Gao, X.-D. (2019) 'Reconstitution of the lipid-linked oligosaccharide pathway for assembly of high-mannose N-glycans', *Nat Commun,* 10(1), pp. 1813-1813.

Lin, L., Kightlinger, W., Prabhu, S. K., Hockenberry, A. J., Li, C., Wang, L.-X., Jewett, M. C. and Mrksich, M. (2020) 'Sequential Glycosylation of Proteins with Substrate-Specific N-Glycosyltransferases', *ACS central science,* 6(2), pp. 144-154.

Lis, H. & Sharon, N. (1993) PROTEIN GLYCOSYLATION - STRUCTURAL AND FUNCTIONAL- ASPECTS. *Eur. J. Biochem.*

Liu, H., Zhang, N., Wan, D., Cui, M., Liu, Z. and Liu, S. (2014) 'Mass spectrometry-based analysis of glycoproteins and its clinical applications in cancer biomarker discovery', *Clin Proteomics,* 11(1), pp. 14-14.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. and Law, M. (2012) 'Comparison of Next-Generation Sequencing Systems', *J Biomed Biotechnol,* 2012, pp. 251364-11.

Lizak, C., Fan, Y.-Y., Weber, T. C. & Aebi, M. (2011) N-Linked Glycosylation of Antibody Fragments in Escherichia coli. *Bioconjugate Chemistry*, 22(3), 488-496.

Lyons, J., Milner, J. D. & Rosenzweig, S. D. (2015) Glycans instructing immunity: the emerging role of altered glycosylation in clinical immunology. *Front. Pediatr.*

Madsen, J. A., Farutin, V., Lin, Y. Y., Smith, S. and Capila, I. (2018) 'Data-independent oxonium ion profiling of multi-glycosylated biotherapeutics', *MAbs,* 10(7), pp. 968-978.

Mario, F. F., Michael, W., Marcela, H., Paul, G. H., Cristina, L. M., Michael, K., Howard, R. M., Anne, D., Miguel, A. V. and Markus, A. (2005) 'Engineering N-linked protein glycosylation with diverse O antigen lipopolysaccharide structures in Escherichia coli', *Proceedings of the National Academy of Sciences of the United States of America,* 102(8), pp. 3016.

Meehl, M. A. & Stadheim, T. A. (2014) Biopharmaceutical discovery and production in yeast. *Current Opinion in Biotechnology*, 30, 120-127.

Meng, L., Forouhar, F., Thieker, D., Gao, Z., Ramiah, A., Moniz, H., Xiang, Y., Seetharaman, J., Milaninia, S., Su, M., Bridger, R., Veillon, L., Azadi, P., Kornhaber, G., Wells, L., Montelione, G., Woods, R., Tong, L. & Moremen, K. (2013) Enzymatic Basis for N- Glycan Sialylation STRUCTURE OF RAT alpha 2,6-SIALYLTRANSFERASE (ST6GAL1) REVEALS CONSERVED AND UNIQUE FEATURES FOR GLYCAN SIALYLATION. *J. Biol. Chem.*, 288(48), 34680-34698.

Mergulhao, F., Monteiro, G., Kelly, A. G., Taipa, M. A. & Joaquim, M. (2000) Recombinant human proinsulin: A new approach in gene assembly and protein expression. *J. Microbiol. Biotechnol.*, 10(5), 690-693.

Mergulhão, F. J. M., Summers, D. K. & Monteiro, G. A. (2005) Recombinant protein secretion in Escherichia coli. *Biotechnology Advances*, 23(3), 177-202.

Messner, P. (1997) Bacterial glycoproteins. *Official Journal of the International Glycoconjugate Organization*, 14(1), 3-11.

Metzker, M. L. (2010) 'Sequencing technologies - the next generation', *Nat Rev Genet,* 11(1), pp. 31-46.

Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. and Gurevich, A. (2018) 'Versatile genome assembly evaluation with QUAST-LG', *Bioinformatics,* 34(13), pp. i142-i150.

Miller, J. H. (1972) *Experiments in molecular genetics.* Cold Spring Harbor, N.Y.: Cold Spring Harbor, N.Y. : Cold Spring Harbor Laboratory, 1972.

Mizukami, A., Caron, A. L., Picanço-Castro, V. & Swiech, K. (2018) *Platforms for recombinant therapeutic glycoprotein production*, *1674*.

Morelle, W. & Michalski, J.-C. (2007) Analysis of protein glycosylation by mass spectrometry. *Nature Protocols*, 2, 1585.

Morgenthaler, A. B., Kinney, W. R., Ebmeier, C. C., Walsh, C. M., Snyder, D. J., Cooper, V. S., Old, W. M. and Copley, S. D. (2019) 'Mutations that improve efficiency of a weak-link enzyme are rare compared to adaptive mutations elsewhere in the genome', *Elife,* 8.

Mukherjee, K. J., Rowe, D. C. D., Watkins, N. A. & Summers, D. K. (2004) Studies of single- chain antibody expression in quiescent Escherichia coli. *Applied and environmental microbiology*, 70(5), 3005-3012.

Mulagapati, S., Koppolu, V. & Raju, T. (2017) Decoding of O- Linked Glycosylation by Mass Spectrometry. *Biochemistry*, 56(9), 1218-1226.

Naegeli, A., Neupert, C., Fan, Y.-Y., Lin, C.-W., Poljak, K., Papini, A. M., Schwarz, F. & Aebi, M. (2014) Molecular analysis of an alternative N- glycosylation machinery by functional transfer from

Actinobacillus pleuropneumoniae to Escherichia coli. *The Journal of biological chemistry*, 289(4), 2170-2179.

Naegeli, A., Michaud, G., Schubert, M., Lin, C.-W., Lizak, C., Darbre, T., Reymond, J.-L. and Aebi, M. (2014) 'Substrate Specificity of Cytoplasmic N-Glycosyltransferase', *J Biol Chem,* 289(35), pp. 24521-24532.

Nasab, F. P., Aebi, M., Bernhard, G. and Frey, A. D. (2013) 'A Combined System for Engineering Glycosylation Efficiency and Glycan Structure in Saccharomyces cerevisiae', *Appl Environ Microbiol,* 79(3), pp. 997-1007.

Nielsen, J. (2013) Production of biopharmaceutical proteins by yeast Advances through metabolic engineering. *Bioengineered.*

Nigam, P. (2013) Microbial Enzymes with Special Characteristics for Biotechnological Applications. Basel: MDPI AG.

Nothaft, H. and Szymanski, C. M. (2013) 'Bacterial protein n-glycosylation: New perspectives and applications', *Journal of Biological Chemistry,* 288(10), pp. 6912-6920.

Nuylert, A., Ishida, Y. & Asano, Y. (2017) Effect of Glycosylation on the Biocatalytic Properties of Hydroxynitrile Lyase from the Passion Fruit, Passiflora edulis: A Comparison of Natural and Recombinant Enzymes. *ChemBioChem*, 18(3), 257-265.

Ohyama, Y., Nakajima, K., Renfrow, M.B., Novak, J. & Takahashi, K. (2020) Mass spectrometry for the identification and analysis of highly complex glycosylation of therapeutic or pathogenic proteins, Expert Review of Proteomics, 17(4), 275-296.

Pan, S., Chen, R., Aebersold, R. and Brentnall, T. A. (2011) 'Mass Spectrometry Based Glycoproteomics—From a Proteomics Perspective', *Mol Cell Proteomics,* 10(1), pp. R110.003251-R110.003251.

Pandhal, J., Desai, P., Walpole, C., Doroudi, L., Malyshev, D. & Wright, P. C. (2012) Systematic metabolic engineering for improvement of glycosylation efficiency in Escherichia coli. *Biochemical and Biophysical Research Communications*, 419(3), 472-476.

Pandhal, J., Ow, S. Y., Noirel, J. & Wright, P. C. (2011) Improving N-glycosylation efficiency in Escherichia coli using shotgun proteomics, metabolic network analysis, and selective reaction monitoring. *Biotechnology and Bioengineering*, 108(4), 902-912.

Pandhal, J., Woodruff, L. B. A., Jaffe, S., Desai, P., Ow, S. Y., Noirel, J., Gill, R. T. & Wright, P. C. (2013) Inverse metabolic engineering to improve Escherichia coli as an N-glycosylation host. *Biotechnology and Bioengineering*, 110(9), 2482-2493.

Paul, M. & Ma, J. (2011) Plant- made pharmaceuticals: Leading products and production platforms. *Biotechnol. Appl. Biochem.*

Prakash, O. and Jaiswal, N. (2009) 'α-Amylase: An Ideal Representative of Thermostable Enzymes', *Appl Biochem Biotechnol,* 160(8), pp. 2401-2414.

Ramesh Chander, K., Rishi, G. & Ajay, S. (2011) Microbial Cellulases and Their Industrial Applications. *Enzyme Research*, 2011(1).

Reetz, M. T. (2013) The Importance of Additive and Non- Additive Mutational Effects in Protein Engineering. *Angewandte Chemie International Edition*, 52(10), 2658-2666.

Rigoldi, F., Donini, S., Redaelli, A., Parisini, E. & Gautieri, A. (2018) Review: Engineering of thermostable enzymes for industrial applications. *APL Bioengineering*, 2(1).

Riley, N. M., Bertozzi, C. R. and Pitteri, S. J. (2021) 'A Pragmatic Guide to Enrichment Strategies for Mass Spectrometry–Based Glycoproteomics', *Mol Cell Proteomics,* 20, pp. 100029-100029.

Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A. and Mesirov, J. P. (2017) 'Variant Review with the Integrative Genomics Viewer', *Cancer Res,* 77(21), pp. e31-e34.

Romas, J. K. & Uwe, T. B. (2009) Finding better protein engineering strategies. *Nature Chemical Biology*, 5(8), 526.

Rosano, G. and Ceccarelli, E. 2014. Recombinant protein expression in Escherichia coli: advances and challenges. *Front. Microbiol.*

Roslyn Mary, B. (2014) 'Playing catch- up with Escherichia coli: Using yeast to increase success rates in recombinant protein production experiments', *Frontiers in Microbiology,* 5.

Scallon, B. J., Tam, S. H., McCarthy, S. G., Cai, A. N. & Raju, T. S. (2007) Higher levels of sialylated Fc glycans in immunoglobulin G molecules can adversely impact functionality. *Molecular Immunology*, 44(7), 1524-1534.

Schlegel, S., Rujas, E., Ytterberg, A. J., Zubarev, R., Luirink, J. & de Gier, J.-W. (2013) Optimizing heterologous protein production in the periplasm of E. coli by regulating gene expression levels. *Microbial Cell Factories*, 12(1), 24.

Schellhorn, H. E. (2020) 'Function, Evolution, and Composition of the RpoS Regulon in Escherichia coli', *Frontiers in microbiology,* 11, pp. 560099-560099.

Schirm, M., Schoenhofen, I. C., Logan, S. M., Waldron, K. C. and Thibault, P. (2005) 'Identification of Unusual Bacterial Glycosylation by Tandem Mass Spectrometry Analyses of Intact Proteins', *Anal. Chem,* 77(23), pp. 7774-7782.

Schriebl, K., Trummer, E., Lattenmayer, C., Weik, R., Kunert, R., Müller, D., Katinger, H. and Vorauer-Uhl, K. (2006) 'Biochemical characterization of rhEpo-Fc fusion protein expressed in CHO cells', *Protein Expr Purif,* 49(2), pp. 265-275.

Schwarz, F. & Aebi, M. (2015) Production of Glycoproteins with Asparagine-Linked N-Acetylglucosamine in Escherichia coli, in Castilho, A. (ed), *Glyco-Engineering: Methods and Protocols*. Methods in Molecular Biology, 49-56.

Schwarz, F. and Aebi, M. (2011) 'Mechanisms and principles of N-linked protein glycosylation', *Current Opinion in Structural Biology,* 21(5), pp. 576-582.

Schwarz, F., Fan, Y.-Y., Schubert, M. and Aebi, M. (2011) 'Cytoplasmic N-Glycosyltransferase of Actinobacillus pleuropneumoniae Is an Inverting Enzyme and Recognizes the NX(S/T) Consensus Sequence', *J Biol Chem,* 286(40), pp. 35267-35274.

Schwarz, F., Huang, W., Li, C., Schulz, B. L., Lizak, C., Palumbo, A., Numao, S., Neri, D., Aebi, M. and Wang, L. X. (2010) 'A combined method for producing homogeneous glycoproteins with eukaryotic *N*-glycosylation', *Nature Chemical Biology,* 6(4), pp. 264-266.

Schäffer, C., Messner, P. and Pohlschroder, M. (2017) 'Emerging facets of prokaryotic glycosylation', *FEMS Microbiology Reviews,* 41(1), pp. 49-91.

Scott, N. E., Cordwell, S. J., Nothaft, H., Szymanski, C. M., Edwards, A. V. G., Larsen, M. R., Labbate, M. & Djordjevic, S. P. (2012) Modification of the Campylobacter jejuni N- linked glycan by EptC protein-mediated addition of phosphoethanolamine. *Journal of Biological Chemistry*, 287(35), 29384-29396.

Sega, G. A. (1984) 'A review of the genetic effects of ethyl methanesulfonate', *Mutation research. Reviews in genetic toxicology,* 134(2), pp. 113-142.

Sharma, V., Ichikawa, M. and Freeze, H. H. (2014) 'Mannose metabolism: More than meets the eye', *Biochemical and Biophysical Research Communications,* 453(2), pp. 220-228.

Shendure, J. and Ji, H. (2008) 'Next-generation DNA sequencing', *Nat Biotechnol,* 26(10), pp. 1135-1145.

Shibai, A., Takahashi, Y., Ishizawa, Y., Motooka, D., Nakamura, S., Ying, B.-W. and Tsuru, S. (2017) 'Mutation accumulation under UV radiation in Escherichia coli', *Sci Rep,* 7(1), pp. 14531-12.

Shirke, A. N., Su, A., Jones, J. A., Butterfoss, G. L., Koffas, M. A. G., Kim, J. R. & Gross, R. A. (2017) Comparative thermal inactivation analysis of Aspergillus oryzae and Thiellavia terrestris cutinase: Role of glycosylation. *Biotechnology and Bioengineering*, 114(1), 63-73.

Shirke, A. N., White, C., Englaender, J. A., Zwarycz, A., Butterfoss, G. L., Linhardt, R. J. & Gross, R. A. (2018) Stabilizing Leaf and Branch Compost Cutinase (LCC) with Glycosylation: Mechanism and Effect on PET Hydrolysis. *Biochemistry*, 57(7), 1190-1200.

Singha, T. K., Gulati, P., Mohanty, A., Khasa, Y. P., Kapoor, R. K. & Kumar, S. (2017) Efficient genetic approaches for improvement of plasmid based expression of recombinant protein in Escherichia coli: A review. *Process Biochemistry*, 55, 17-31.

Skropeta, D. (2009) The effect of individual N- glycans on enzyme activity. *Bioorganic & Medicinal Chemistry*, 17(7), 2645-2653.

Spiro, R. G. (2002) Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology*, 12(4), 43R-56R.

Stoebel, D. M., Hokamp, K., Last, M. S. and Dorman, C. J. (2009) 'Compensatory evolution of gene regulation in response to stress by Escherichia coli lacking RpoS', *PLoS Genet,* 5(10), pp. e1000671-e1000671.

Strutton, B., Jaffé, S. R. P., Pandhal, J. and Wright, P. C. (2018) 'Producing a glycosylating Escherichia coli cell factory: The placement of the bacterial oligosaccharyl transferase pglB onto the genome', *Biochemical and Biophysical Research Communications,* 495(1), pp. 686-692.

Su, L., Yu, L., Xu, C. & Wu, J. (2015) Extracellular expression of Thermobifida fusca cutinase with pelB signal peptide depends on more than type II secretion pathway in Escherichia coli. *Journal of Biotechnology*, 204, 47-52.

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N. T., Legeay, M., Fang, T., Bork, P., Jensen, L. J. and von Mering, C. (2021) 'The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets', *Nucleic Acids Res,* 49(D1), pp. D605-D612.

Szymanski, C. M. and Wren, B. W. (2005) 'Protein glycosylation in bacterial mucosal pathogens', *Nat Rev Microbiol,* 3(3), pp. 225-237.

Szymanski, C. M., Yao, R., Ewing, C. P., Trust, T. J. & Guerry, P. (1999) Evidence for a system of general protein glycosylation in *Campylobacter jejuni*. *Molecular Microbiology*, 32(5), 1022-1030.

Tan, S., Wu, W., Liu, J., Kong, Y., Pu, Y. & Yuan, R. (2002) Efficient expression and secretion of recombinant hirudin III in *E. coli* using the l- asparaginase II signal sequence. *Protein Expression and Purification*, 25(3), 430-436.

Tayapiwatana, C., Gotz, F., Werner, R. & Manosroi, A. (2001) Secretion of active recombinant human tissue plasminogen activator derivatives in Escherichia coli. *Applied and Environmental Microbiology*, 67(6), 2657-2664.

Taylor, M. E. (2006) *Introduction to glycobiology*, 2nd ed. edition. Oxford: Oxford: Oxford University Press, 2006.

Thorvaldsdóttir, H., Robinson, J. T. and Mesirov, J. P. (2013) 'Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration', *Brief Bioinform,* 14(2), pp. 178-192.

Tong, L., Lin, Q., Wong, W. K. R., Ali, A., Lim, D., Sung, W. L., Hew, C. L. & Yang, D. S. C. (2000) Extracellular Expression, Purification, and Characterization of a Winter Flounder Antifreeze Polypeptide from Escherichia coli. *Protein Expression and Purification*, 18(2), 175-181.

Trombetta, E. S. (2003) The contribution of N- glycans and their processing in the endoplasmic reticulum to glycoprotein biosynthesis. *Glycobiology*, 13(9), 77R-91R.

Vainauskas, S. & Menon, A. K. (2006) Ethanolamine phosphate linked to the first mannose residue of glycosylphosphatidylinositol (GPI) lipids is a major feature of the GPI structure that is recognized by human GPI transamidase. *Journal of Biological Chemistry*, 281(50), 38358-38364.

Valderrama-Rincon, J. D., Fisher, A. C., Merritt, J. H., Fan, Y. Y., Reading, C. A., Chhiba, K., Heiss, C., Azadi, P., Aebi, M. and DeLisa, M. P. (2012) 'An engineered eukaryotic protein glycosylation pathway in Escherichia coli', *Nature Chemical Biology,* 8(5), pp. 434-436.

Wacker, M., Linton, D., Hitchen, P. G., Nita-Lazar, M., Haslam, S. M., North, S. J., Panico, M., Morris, H. R., Dell, A. and Wren, B. W. (2002) '*N*-linked glycosylation in *Campylobacter jejuni* and its functional transfer into *E. coli*', *Science,* 298(5599), pp. 1790-1793.

Wang, L.-X. and Amin, Mohammed N. (2014) 'Chemical and Chemoenzymatic Synthesis of Glycoproteins for Deciphering Functions', *Chem Biol,* 21(1), pp. 51-66.

Wang, L.-X. and Lomino, J. V. (2012) 'Emerging Technologies for Making Glycan-Defined Glycoproteins', *ACS Chem. Biol,* 7(1), pp. 110-122.

Wang, Z.-x., Deng, R.-p., Jiang, H.-W., Guo, S.-J., Le, H.-y., Zhao, X.-d., Chen, C.-S., Zhang, J.-b. and Tao, S.-c. (2012) 'Global Identification of Prokaryotic Glycoproteins Based on an Escherichia coli Proteome Microarray', *PLoS One,* 7(11), pp. e49080-e49080.

Weaver, J. L. (2000) 'Introduction to Flow Cytometry', *Methods,* 21(3), pp. 199-201.

Wu, D., Li, J., Struwe, W. B. and Robinson, C. V. (2019) 'Probing N -glycoprotein microheterogeneity by lectin affinity purification-mass spectrometry analysis', *Chem Sci,* 10(19), pp. 5146-5155.

Yang, G., Höti, N., Chen, S.-Y., Zhou, Y., Wang, Q., Betenbaugh, M. and Zhang, H. (2020) 'One-Step Enrichment of Intact Glycopeptides From Glycoengineered Chinese Hamster Ovary Cells', *Front Chem,* 8, pp. 240-240.

Yang, H., Yang, C. and Sun, T. (2018) 'Characterization of glycopeptides using a stepped higher-energy C-trap dissociation approach on a hybrid quadrupole orbitrap', *Rapid Commun Mass Spectrom,* 32(16), pp. 1353-1362.

Zalai, D., Hevér, H., Lovász, K., Molnár, D., Wechselberger, P., Hofer, A., Párta, L., Putics, Á. and Herwig, C. (2016) 'A control strategy to investigate the relationship between specific productivity and high-mannose glycoforms in CHO cells', *Appl Microbiol Biotechnol,* 100(16), pp. 7011-7024.

Zhu, B., Wang, D. and Wei, N. (2022) 'Enzyme discovery and engineering for sustainable plastic recycling', Trends Biotechnol, 40(1), pp. 22-37.

Zhu, J., Ruan, Y., Fu, X., Zhang, L., Ge, G., Wall, J. G., Zou, T., Zheng, Y., Ding, N. and Hu, X. (2020) 'An Engineered Pathway for Production of Terminally Sialylated N -glycoproteins in the Periplasm of Escherichia coli', Front Bioeng Biotechnol, 8, pp. 313-313.

Zhang, L., Luo, S. and Zhang, B. 2016. Glycan analysis of therapeutic glycoproteins. Taylor & Francis.

Appendix A

This appendix contains a list of primer sequences and recipes used throughout this thesis.

Appendix A Table1: Gene knockout primer sequences for plasmid amplification and verification

Primer	**Sequence 5' -> 3'**	**Properties**
H1P1	CATTGAAACCTTACACTCTGAAATCATCGTGTAGGCTGGAGCTGCTTC	Tm, 67°C
rfaL check	GAGATTAAGTTGTATAGATAAGAAG	Tm, 47°C
K1	CAGTCATAGCCGAATAGCCT	Tm, 54°C

Appendix A Table 2: 10X Tris-acetate EDTA buffer

Component	**Formula**	**Weight or Volume**
Tris base	$C_4H_{11}NO_3$	96.8 g
Glacial acetic acid	CH3COOH	22.8 mL
Ethylenediaminetetraacetic acid (EDTA)	$C_{10}H_{16}N_2O_8$	7.4 g

Add components to 2L dH_20 and dilute 10X to make a 1X DNA gel electrophoresis running buffer.

Appendix B: Supplementary material for Chapter 4

Appendix B Table 1: Mannose fluorescence baseline tests for Alexafluor633 Con A label concentrations. AF 633+ represents target higher mannose fluorescing *E. coli* cell populations and AF 633- represents *E. coli* cells within standard fluorescent gated population of the sample.

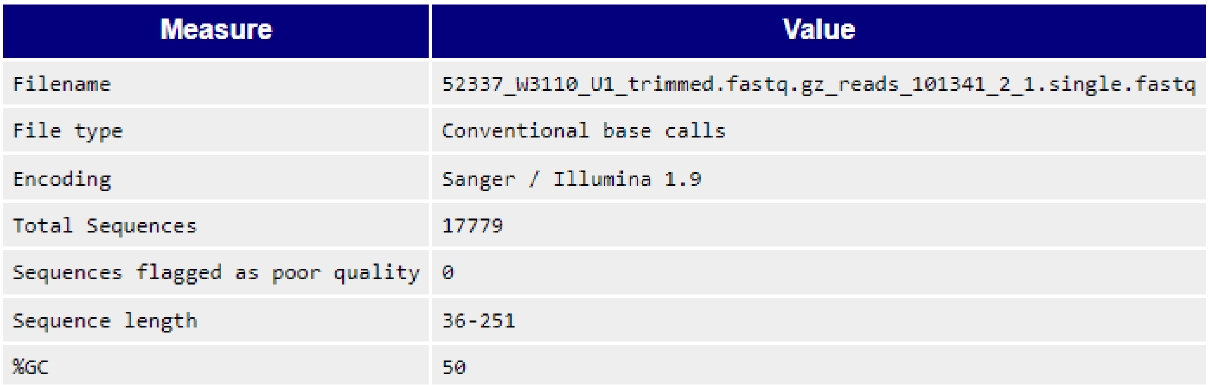
Sample	**Population**	**Events recorded**	**% Parent**	**% Total**
PBS	E. coli	6	0.01	0.01
	AF 633+	0	0	0
	AF633-	1	16.67	0

LB Media	E. coli	23	0.02	0.02
	AF 633+	2	8.7	0
	AF633-	11	47.83	0.01
Ctrl 1 (No pathway) labelled	E. coli	95845	95.85	95.85
	AF 633+	3	0	0
	AF633-	40140	41.88	40.14
Ctrl 2 (with plasmid) unlabelled	E. coli	65106	65.11	65.11
	AF 633+	9	0.02	0.01
	AF633-	27891	42.83	27.89
Uninduced 1µg/µl conc	E. coli	61867	61.87	61.87
	AF 633+	1852	2.99	1.85
	AF633-	18176	29.38	18.18
Induced 1µg/µl conc	E. coli	83172	83.17	83.17
	AF 633+	47555	57.18	47.56
	AF633-	3356	4.01	3.36
Uninduced 2.5µg/µl conc	E. coli	89415	89.42	89.42
	AF 633+	1154	1.29	1.15
	AF633-	31270	34.97	31.27
Induced 2.5µg/µl conc	E. coli	75235	75.23	75.23
	AF 633+	50819	67.76	50.82
	AF633-	4883	6.40	4.89
Uninduced 5µg/µl conc	E. coli	83994	83.99	83.99
	AF 633+	1124	1.34	1.12
	AF633-	32093	38.21	32.09
Induced 5µg/µl conc	E. coli	44898	44.90	44.90
	AF 633+	30024	66.84	30.02
	AF633-	4282	9.55	4.28
Uninduced 10µg/µl conc	E. coli	56382	56.38	56.38
	AF 633+	1995	3.54	2
	AF633-	16772	29.75	16.77
Induced 10µg/µl conc	E. coli	30465	30.47	30.47

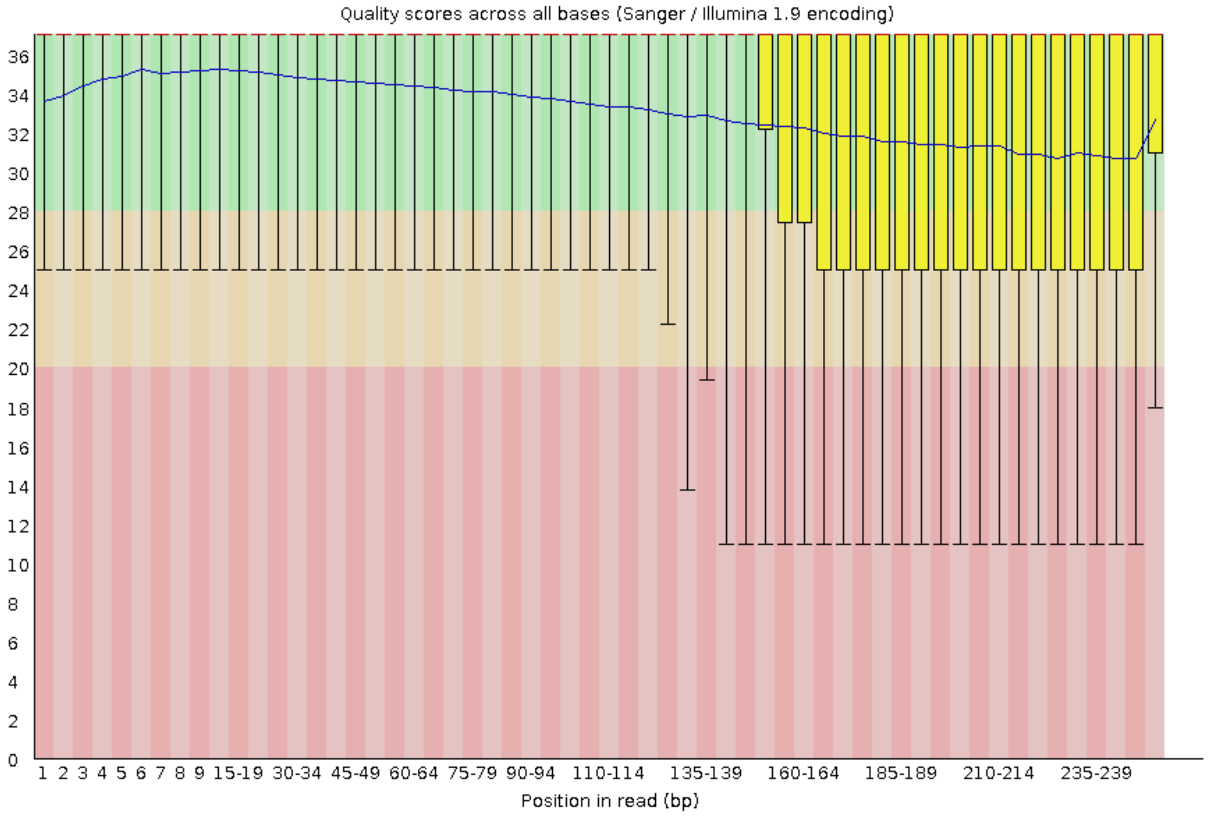
AF 633+	20643	68.49	20.65
AF633-	2886	9.22	2.88

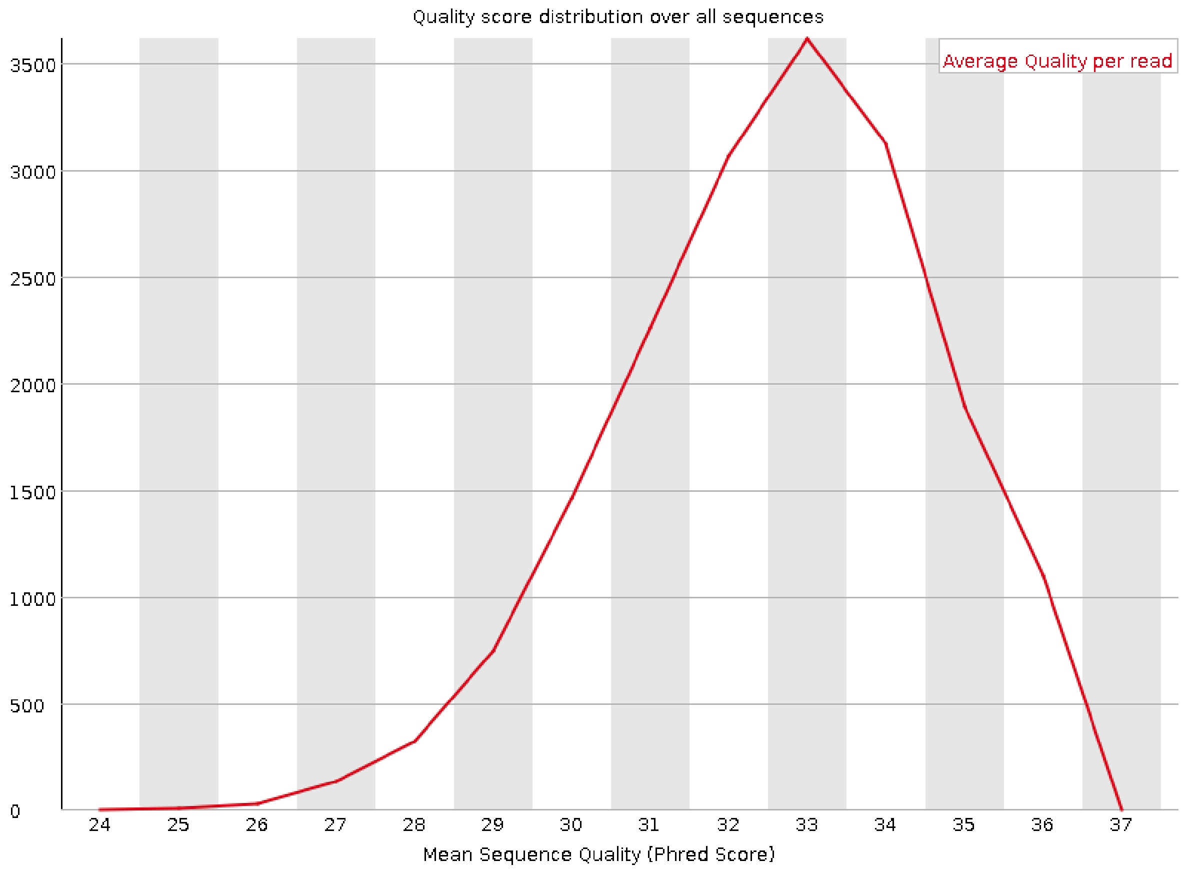
Appendix C: Supplementary material for Chapter 5

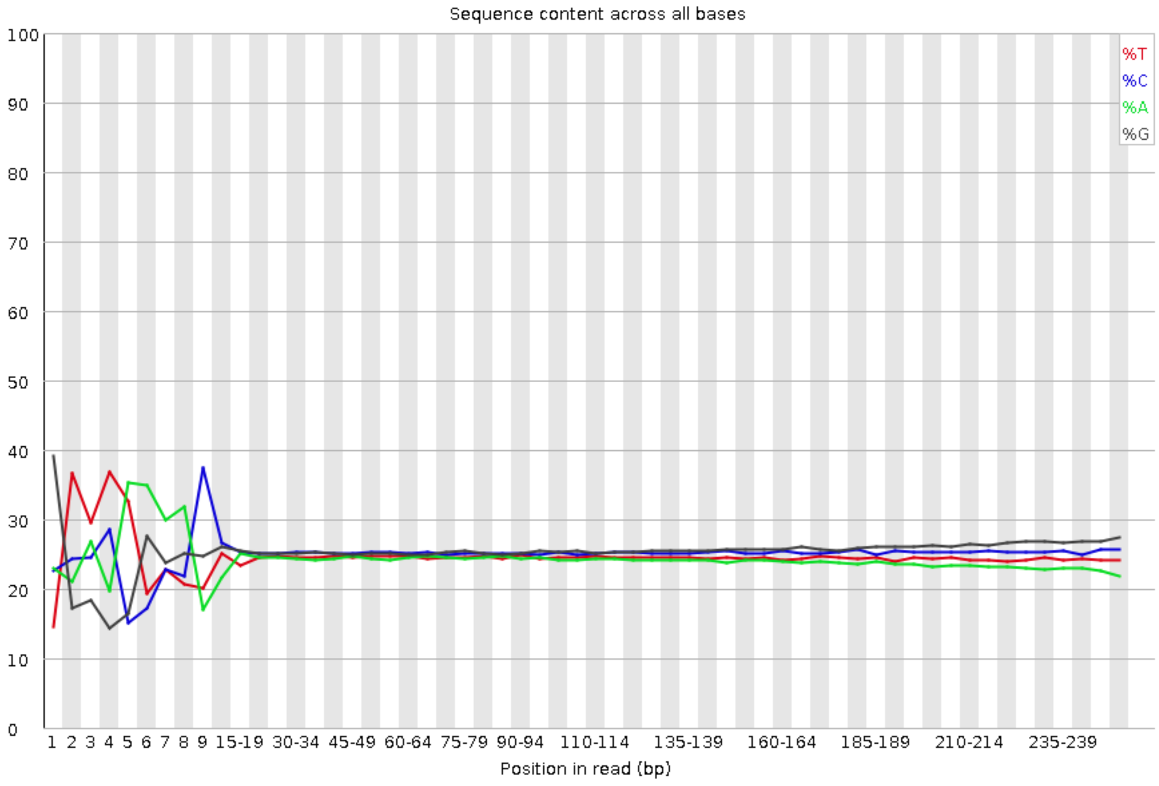
Basic Statistics for W3110 cell Next Generation Sequencing. Data was used to verify strain identity for subsequent mutant strain data analyses.

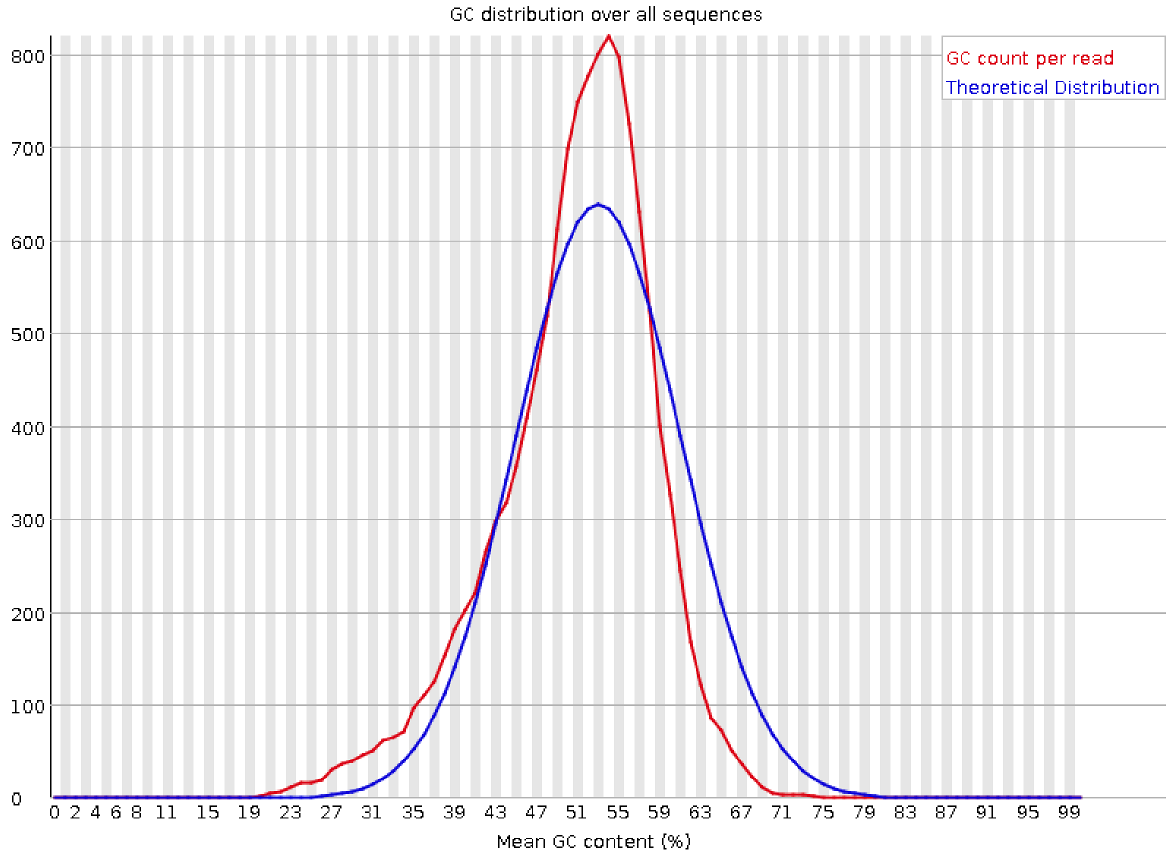


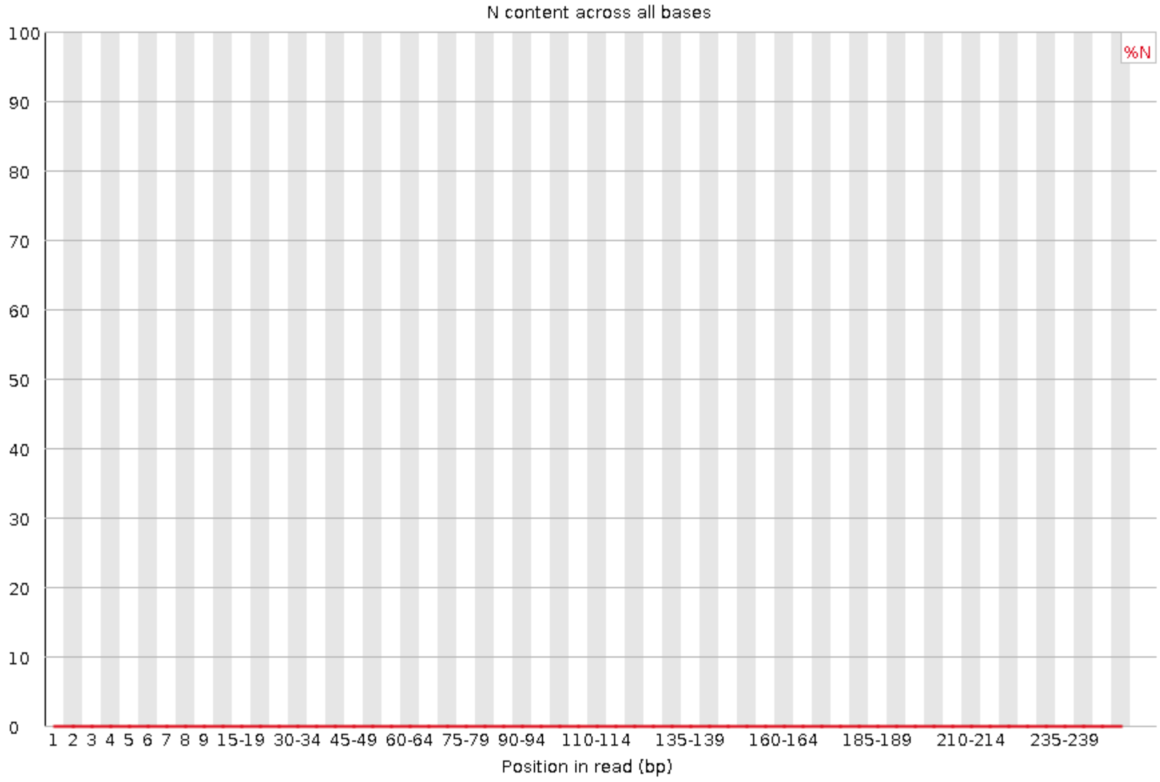
Measure	Value
Filename	52337_W3110_U1_trimmed.fastq.gz_reads_101341_2_1.single.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	17779
Sequences flagged as poor quality	0
Sequence length	36-251
%GC	50

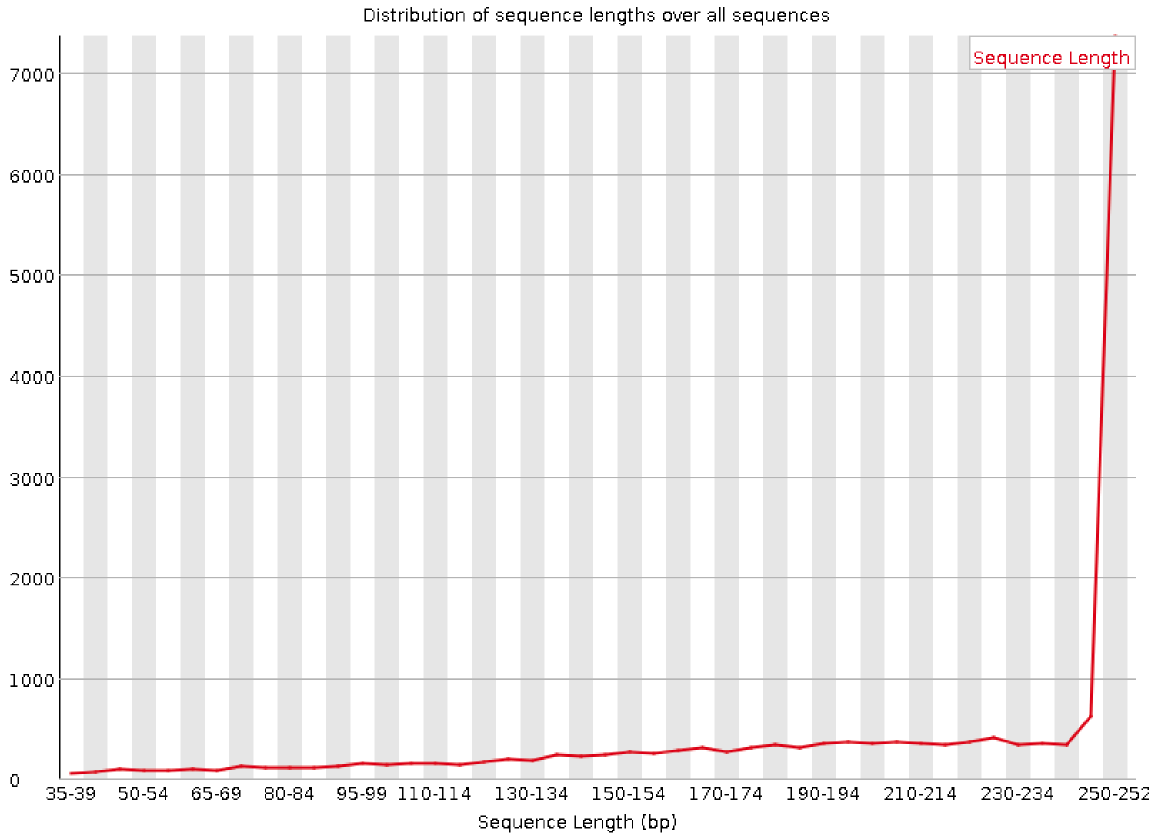



Quality score distribution over all sequences
Average Quality per read
3500
3000
2500
2000
1500
1000
500
0
24
25
26
27
28
29
30
31
32
33
34
35
36
37
Mean Sequence Quality (Phred Score)

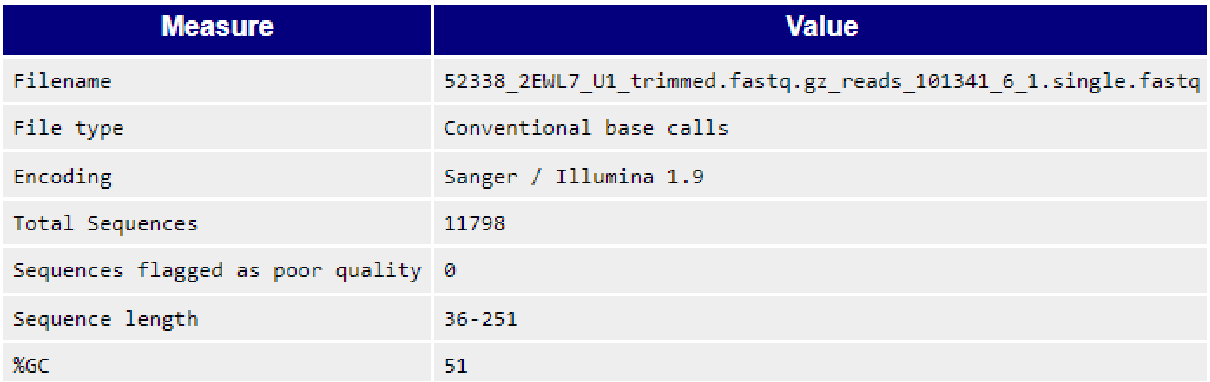

Sequence content across all bases
%T
%C
%A
%G
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 15-19 30-34 45-49 60-64 75-79 90-94 110-114 135-139 160-164 185-189 210-214 235-239
Position in read (bp)


GC distribution over all sequences
GC count per read
Theoretical Distribution
800
700
600
500
400
300
200
100
0
0 2 4 6 8 11 15 19 23 27 31 35 39 43 47 51 55 59 63 67 71 75 79 83 87 91 95 99
Mean GC content (%)

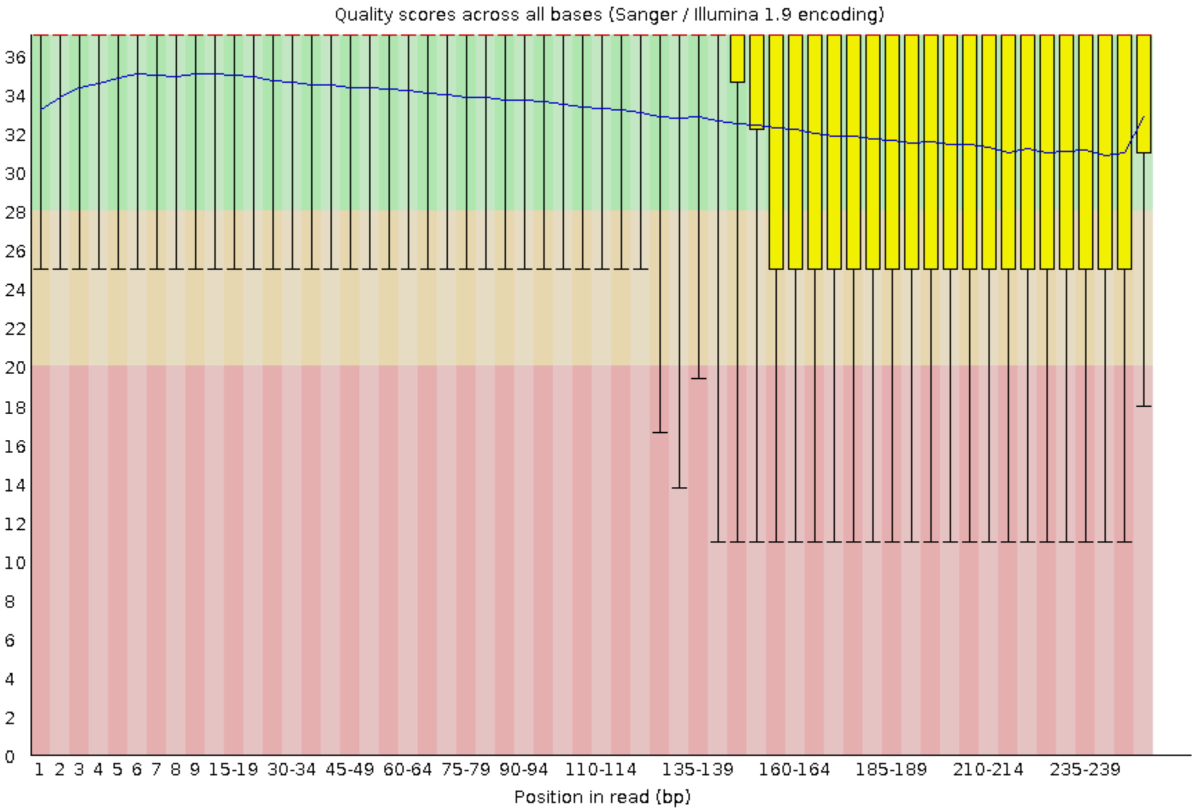

N content across all bases
%N
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 15-19 30-34 45-49 60-64 75-79 90-94 110-114 135-139 160-164 185-189 210-214 235-239
Position in read (bp)

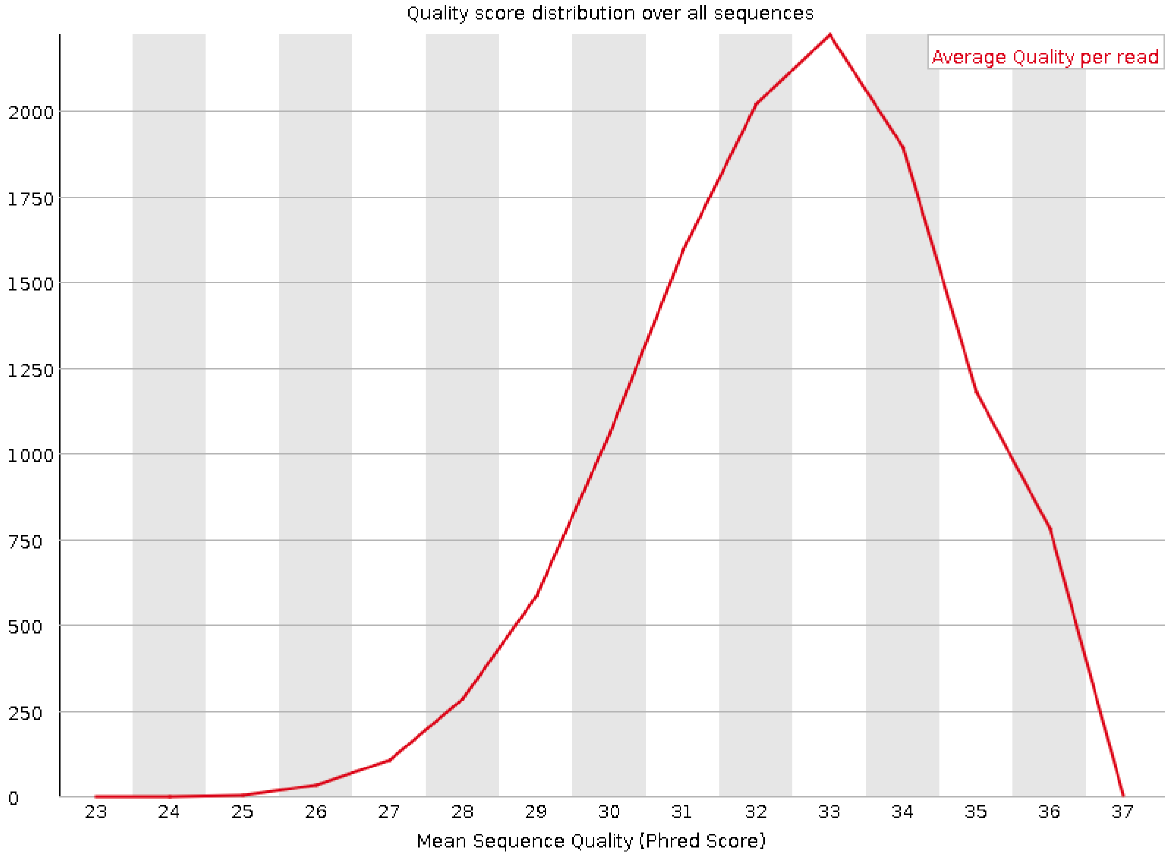


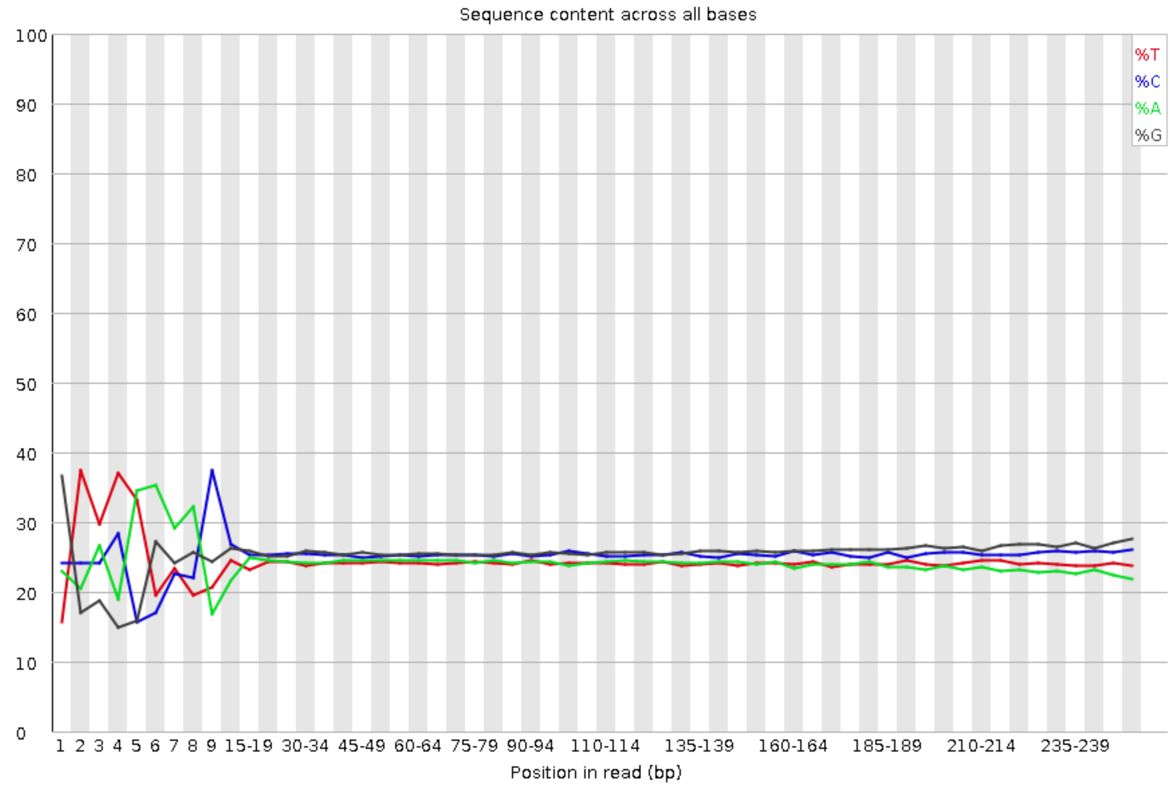
Basic Statistics for 2EWL7

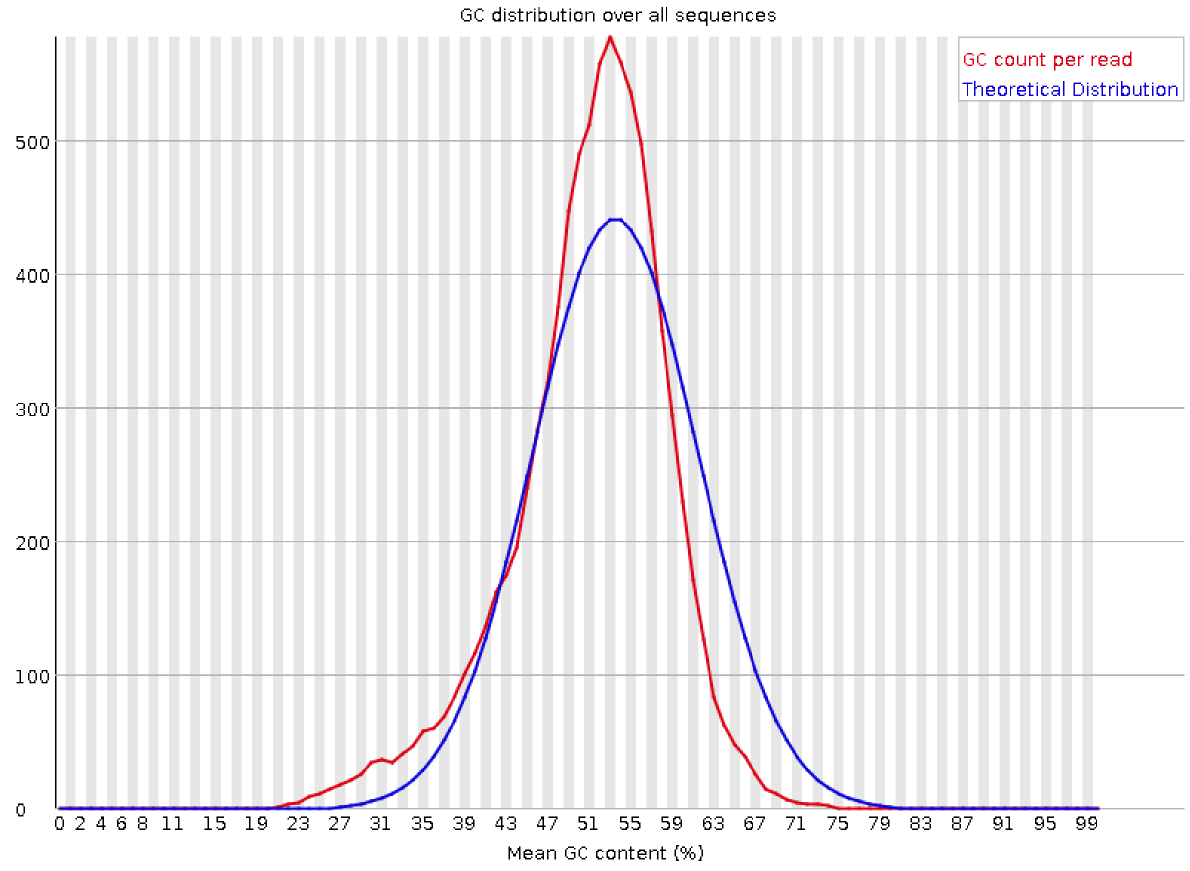


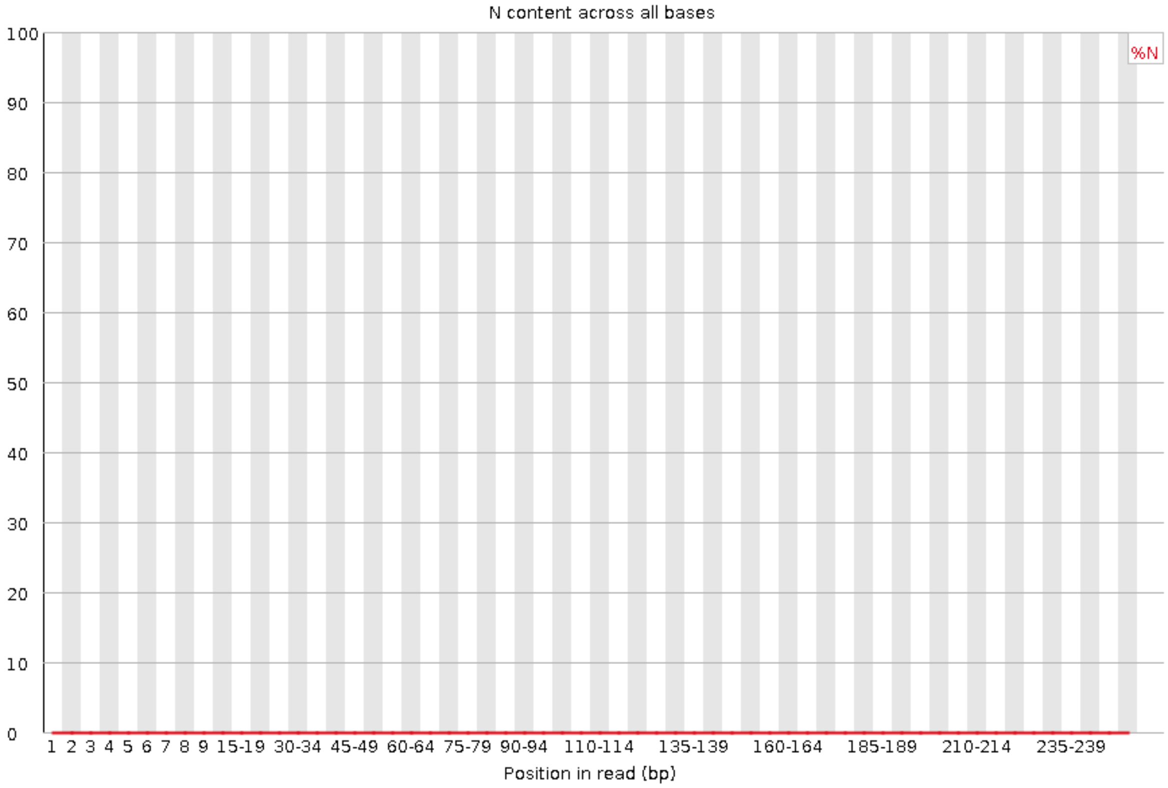
Measure	Value
Filename	52338_2EWL7_U1_trimmed.fastq.gz_reads_101341_6_1.single.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	11798
Sequences flagged as poor quality	0
Sequence length	36-251
%GC	51

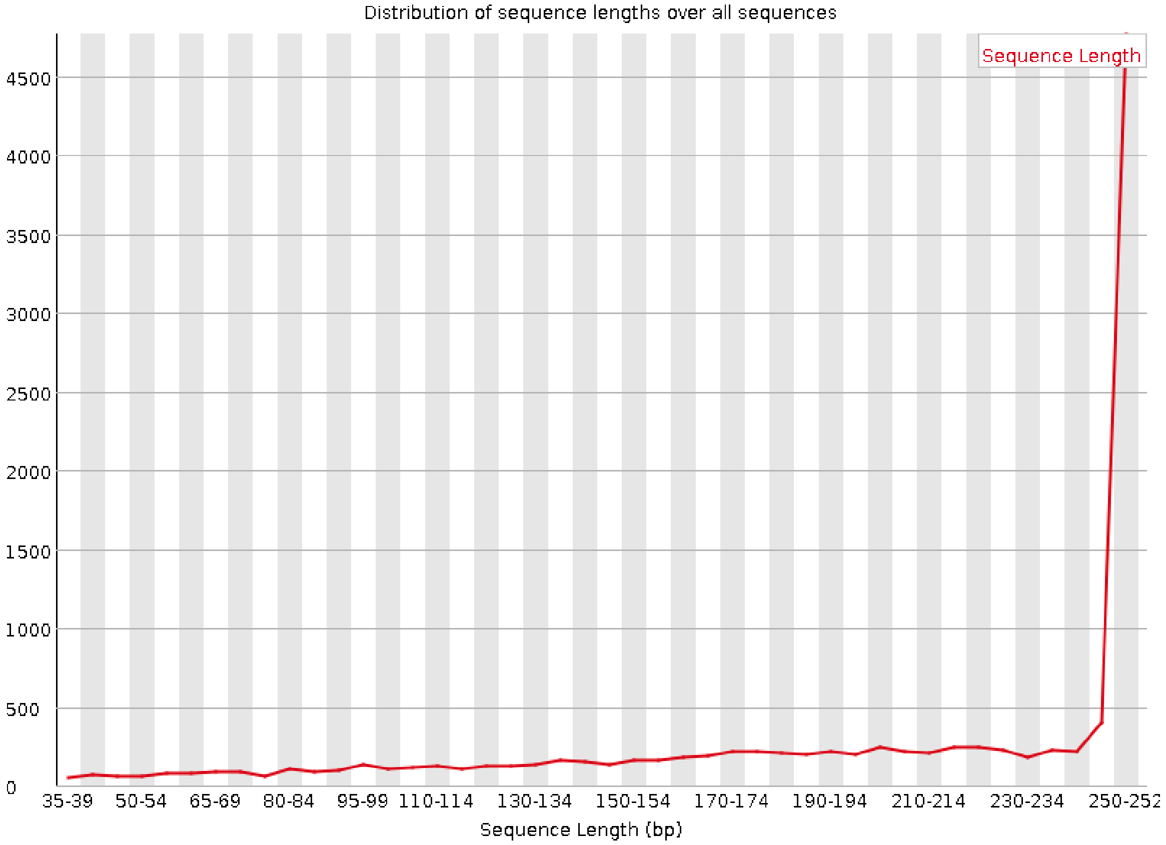

Quality scores across all bases (Sanger / Illumina 1.9 encoding)
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 15-19 30-34 45-49 60-64 75-79 90-94 110-114 135-139 160-164 185-189 210-214 235-239
Position in read (bp)


Quality score distribution over all sequences
Average Quality per read
2000
1750
1500
1250
1000
750
500
250
0
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
Mean Sequence Quality (Phred Score)

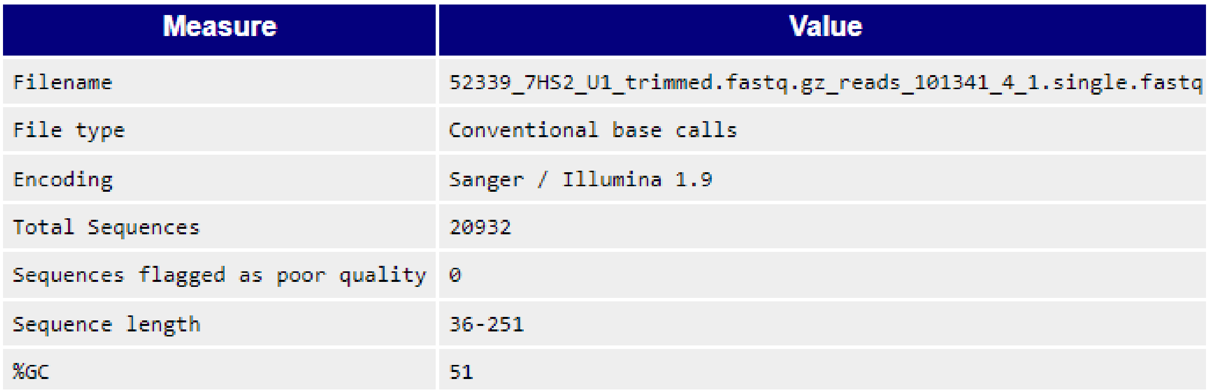

Sequence content across all bases
100
90
80
70
60
50
40
30
20
10
0
%T
%C
%A
%G
1 2 3 4 5 6 7 8 9 15-19 30-34 45-49 60-64 75-79 90-94 110-114 135-139 160-164 185-189 210-214 235-239
Position in read (bp)


GC distribution over all sequences
GC count per read
Theoretical Distribution
500
400
300
200
100
0
0 2 4 6 8 11 15 19 23 27 31 35 39 43 47 51 55 59 63 67 71 75 79 83 87 91 95 99
Mean GC content (%)

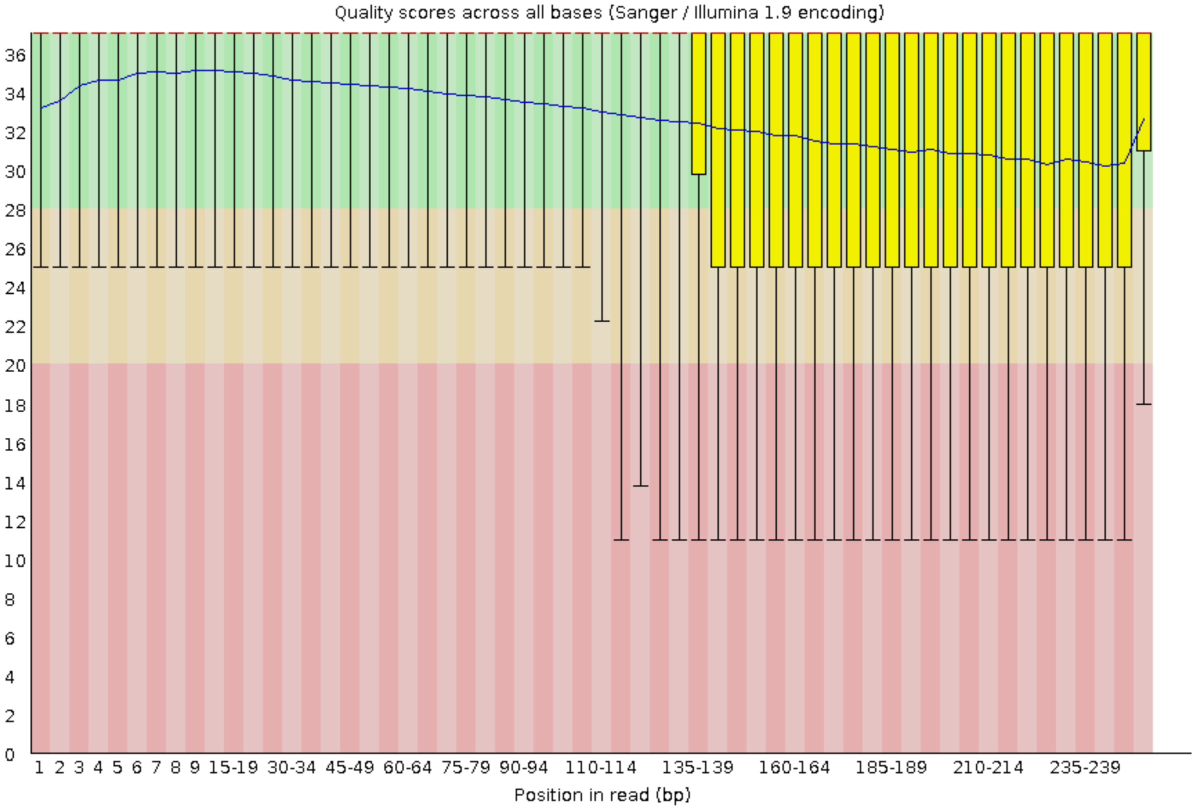

N content across all bases
%N
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 15-19 30-34 45-49 60-64 75-79 90-94 110-114 135-139 160-164 185-189 210-214 235-239
Position in read (bp)

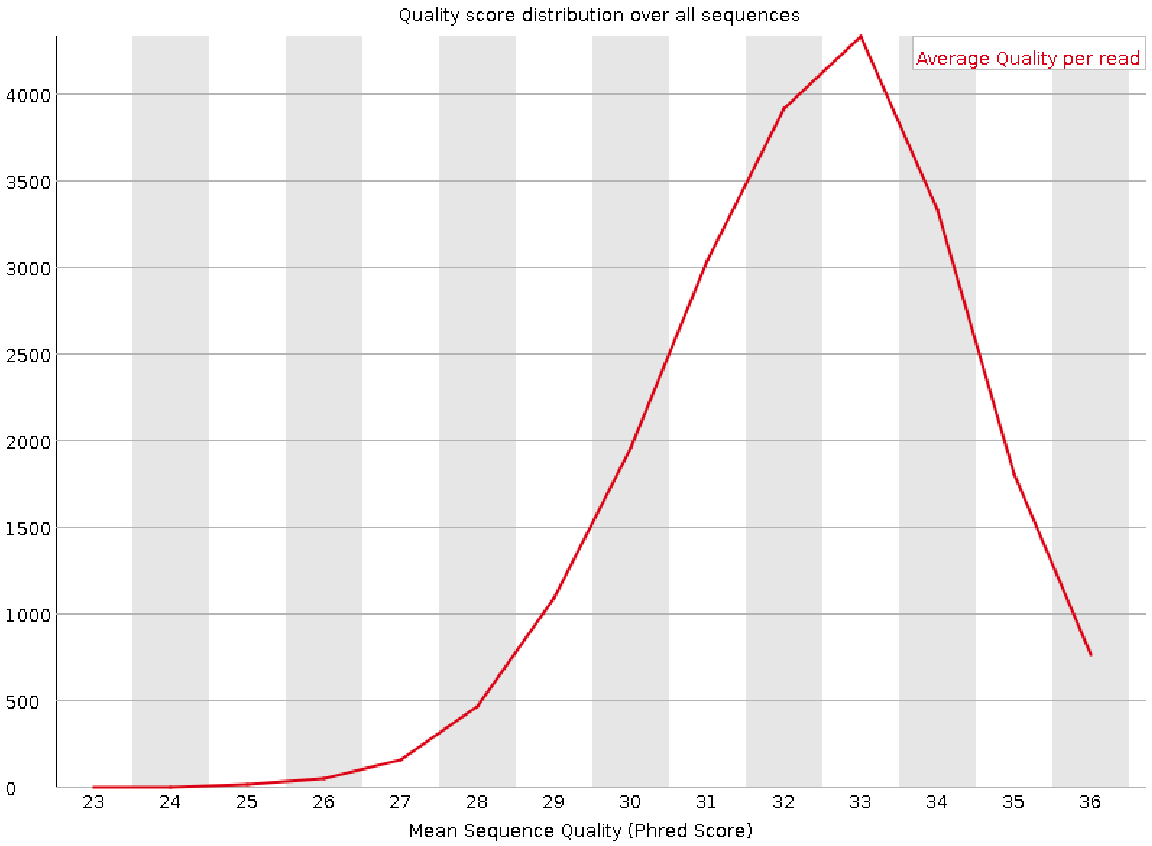

Distribution of sequence lengths over all sequences
Sequence Length
4500
4000
3500
3000
2500
2000
1500
1000
500
0
35-39 50-54 65-69 80-84 95-99 110-114 130-134 150-154 170-174 190-194 210-214 230-234 250-252
Sequence Length (bp)

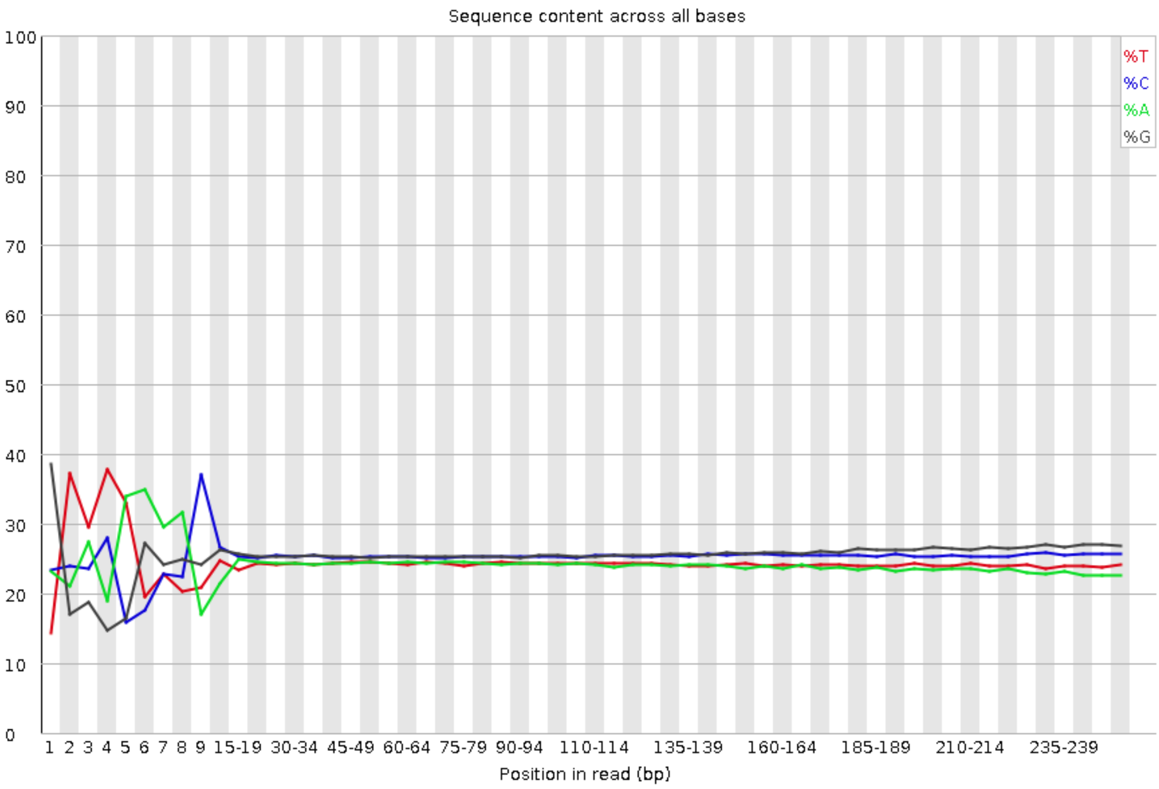
Basic Statistics for 7HS2

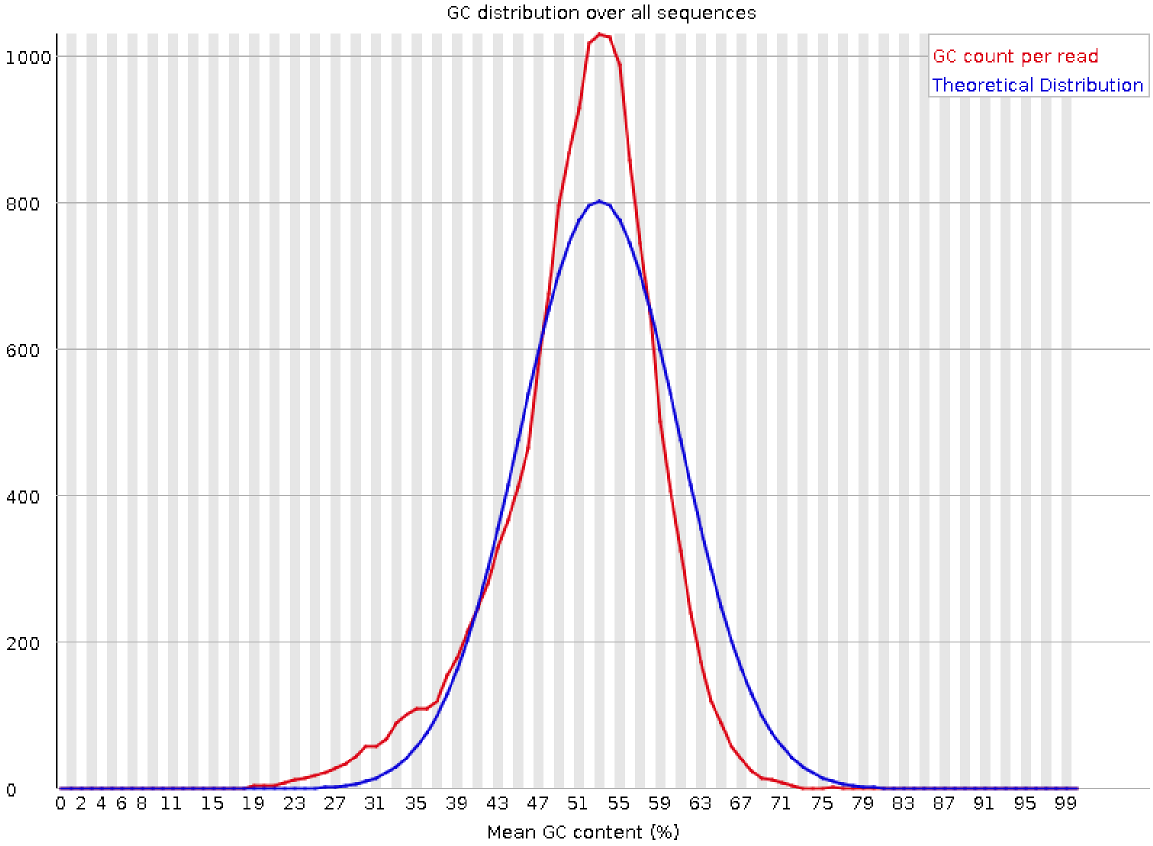


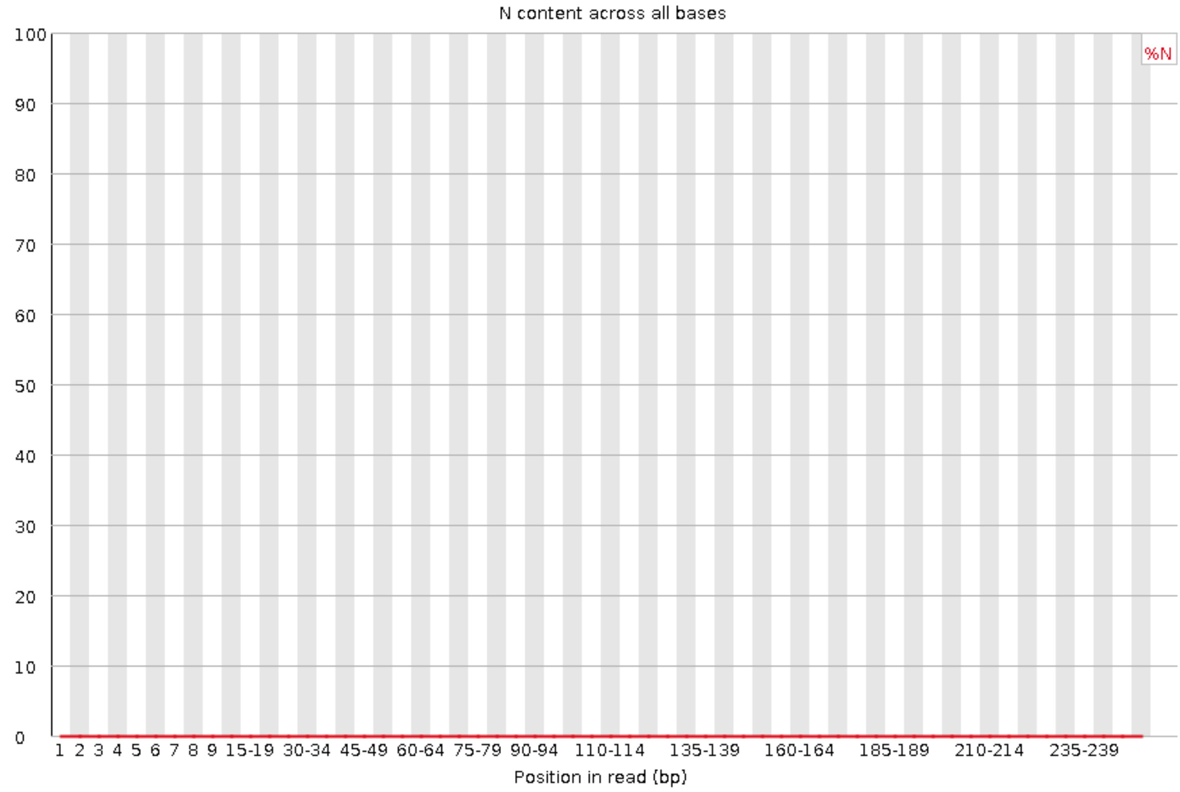
Measure	Value
Filename	52339_7HS2_U1_trimmed.fastq.gz_reads_101341_4_1.single.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	20932
Sequences flagged as poor quality	0
Sequence length	36-251
%GC	51

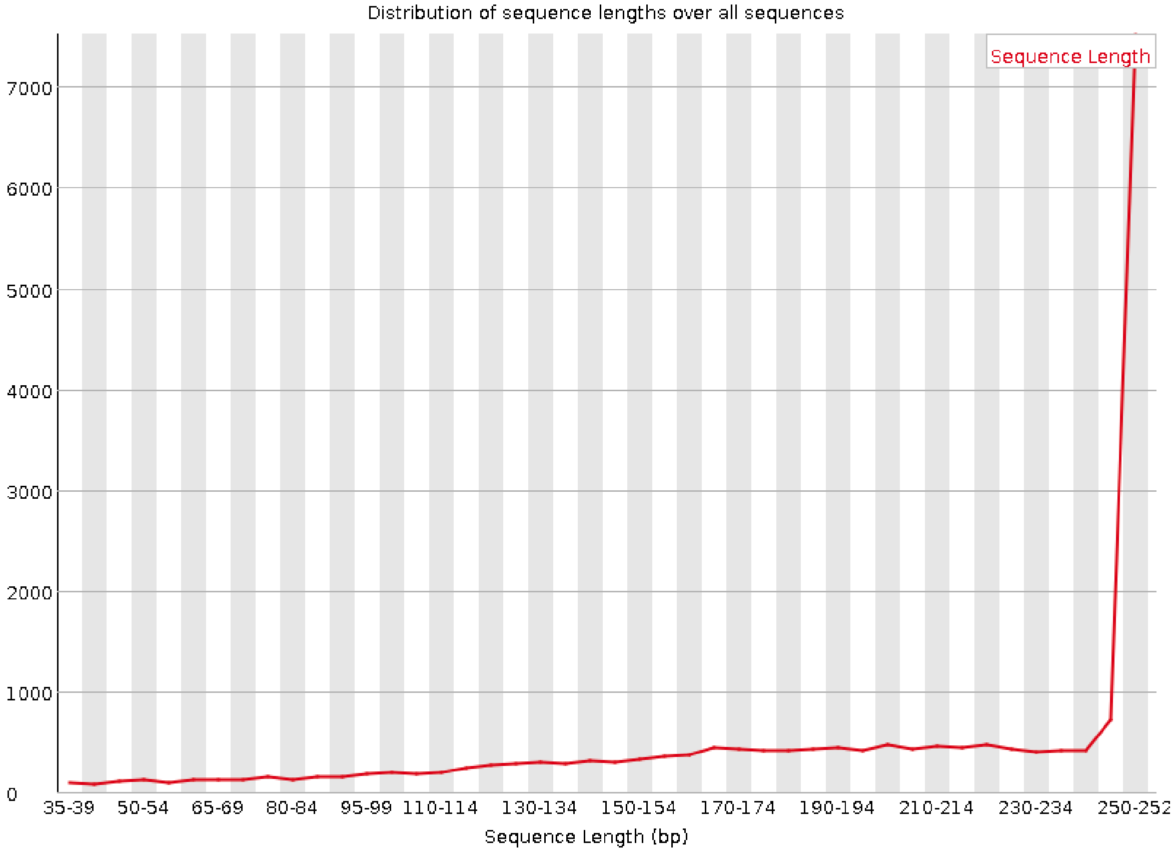



Quality score distribution over all sequences
Average Quality per read
4000
3500
3000
2500
2000
1500
1000
500
0
23
24
25
26
27
28
29
30
31
32
33
34
35
36
Mean Sequence Quality (Phred Score)

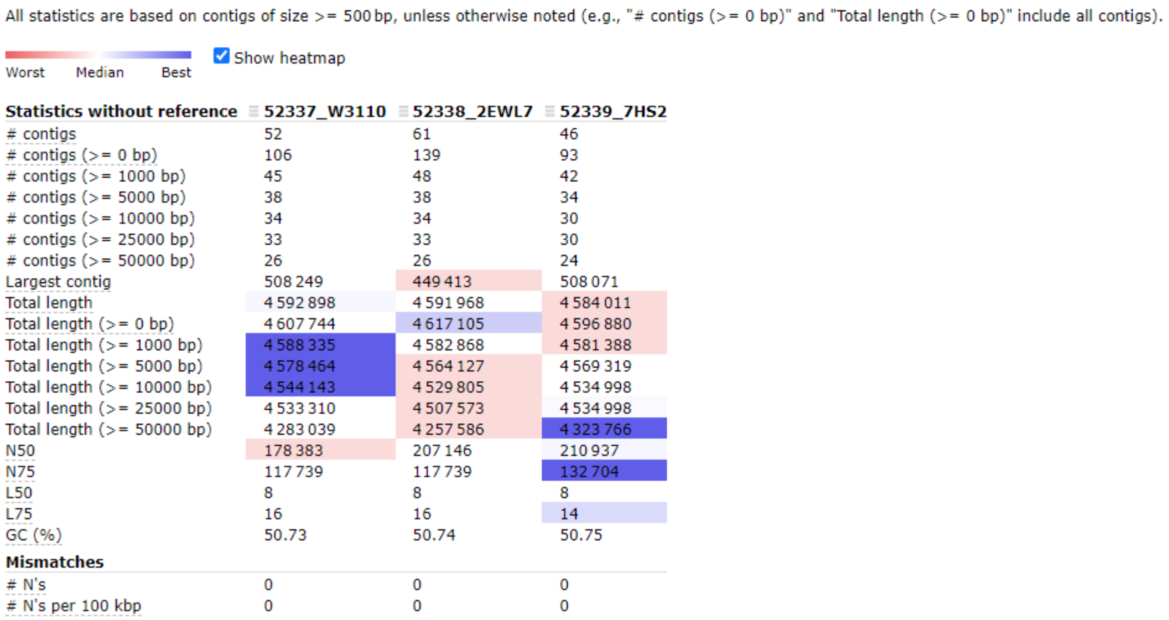

Sequence content across all bases
100
90
80
70
60
50
40
30
20
10
0
%T
%C
%A
%G
1 2 3 4 5 6 7 8 9 15-19 30-34 45-49 60-64 75-79 90-94 110-114 135-139 160-164 185-189 210-214 235-239
Position in read (bp)


GC distribution over all sequences
1000
800
600
400
200
0
GC count per read
Theoretical Distribution
0 2 4 6 8 11 15 19 23 27 31 35 39 43 47 51 55 59 63 67 71 75 79 83 87 91 95 99
Mean GC content (%)


N content across all bases
%N
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 15-19 30-34 45-49 60-64 75-79 90-94 110-114 135-139 160-164 185-189 210-214 235-239
Position in read (bp)


Distribution of sequence lengths over all sequences
Sequence Length
7000
6000
5000
4000
3000
2000
1000
0
35-39 50-54 65-69 80-84 95-99 110-114 130-134 150-154 170-174 190-194 210-214 230-234 250-252
Sequence Length (bp)

QUAST Summary (Mikheenko *et al.*, 2018)



All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

☑ Show heatmap

Statistics without reference	52337_W3110	52338_2EWL7	52339_7HS2
# contigs	52	61	46
# contigs (>= 0 bp)	106	139	93
# contigs (>= 1000 bp)	45	48	42
# contigs (>= 5000 bp)	38	38	34
# contigs (>= 10000 bp)	34	34	30
# contigs (>= 25000 bp)	33	33	30
# contigs (>= 50000 bp)	26	26	24
Largest contig	508 249	449 413	508 071
Total length	4 592 898	4 591 968	4 584 011
Total length (>= 0 bp)	4 607 744	4 617 105	4 596 880
Total length (>= 1000 bp)	4 588 335	4 582 868	4 581 388
Total length (>= 5000 bp)	4 578 464	4 564 127	4 569 319
Total length (>= 10000 bp)	4 544 143	4 529 805	4 534 998
Total length (>= 25000 bp)	4 533 310	4 507 573	4 534 998
Total length (>= 50000 bp)	4 283 039	4 257 586	4 323 766
N50	178 383	207 146	210 937
N75	117 739	117 739	132 704
L50	8	8	8
L75	16	16	14
GC (%)	50.73	50.74	50.75
Mismatches			
# N's	0	0	0
# N's per 100 kbp	0	0	0

Metrics description

# contigs is the total number of contigs in the assembly.

Largest contig is the length of the longest contig in the assembly.

Total length is the total number of bases in the assembly.

Reference length is the total number of bases in the reference genome.

GC (%) is the total number of G and C nucleotides in the assembly, divided by the total length of the assembly.

Reference GC (%) is the percentage of G and C nucleotides in the reference genome.

N50 is the length for which the collection of all contigs of that length or longer covers at least half an assembly.

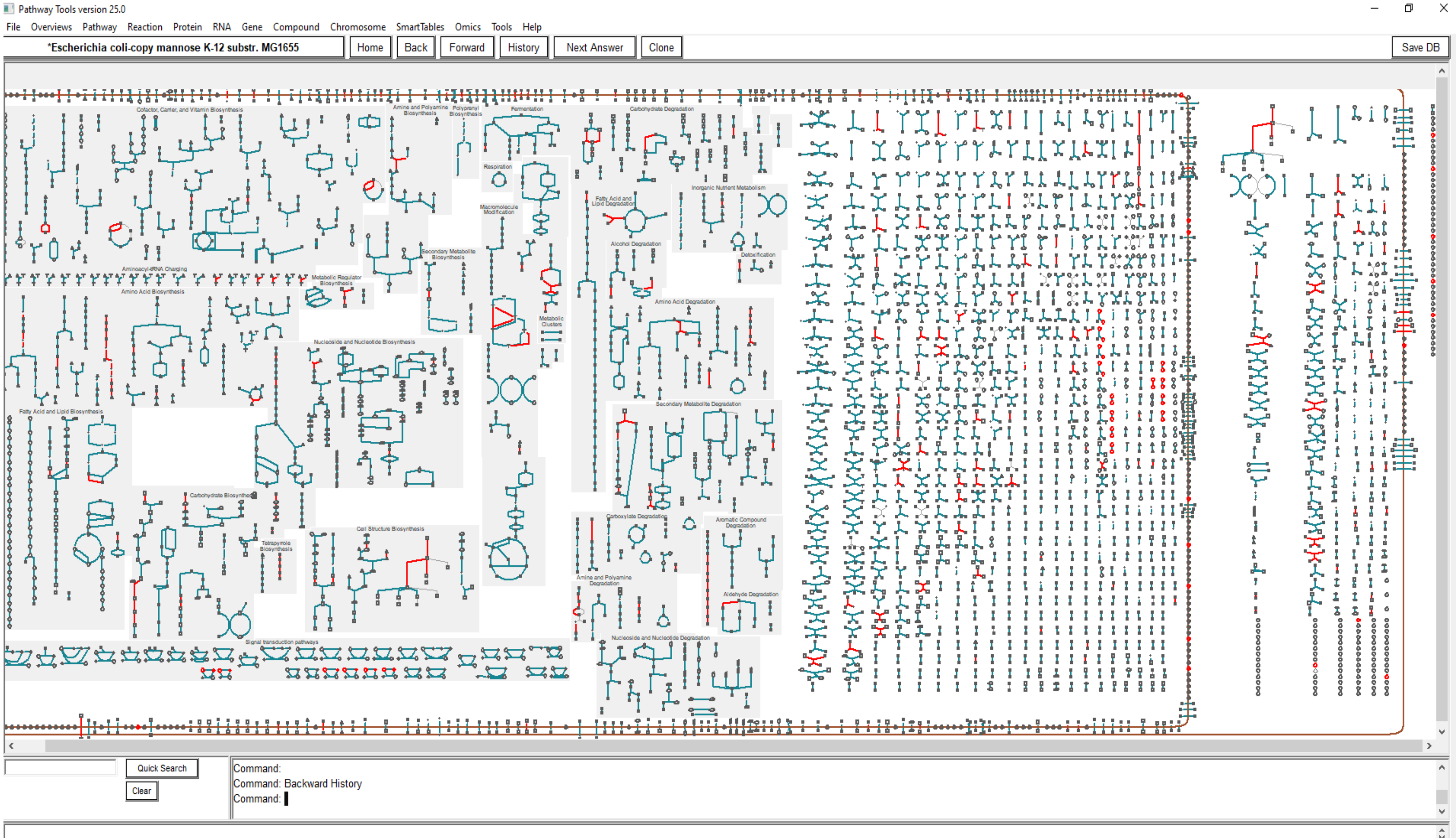
NG50 is the length for which the collection of all contigs of that length or longer covers at least half the reference genome.

This metric is computed only if the reference genome is provided.

N75 and NG75 are defined similarly to N50 but with 75 % instead of 50 %.

L50 (L75, LG50, LG75) is the number of contigs equal to or longer than N50 (N75, NG50, NG75)

In other words, L50, for example, is the minimal number of contigs that cover half the assembly.


Pathway Tools version 25.0
File
Overviews
Pathway
Reaction
Protein
RNA
Gene
Compound
Chromosome
SmartTables
Omics
Tools
Help
*Escherichia coli-copy mannose K-12 substr. MG1655
Home
Back
Forward
History
Next Answer
Clone
Save DB
Cofactor, Carrier, and Vitamin Biosynthesis
Amine and Polyamine Biosynthesis
Polyprenyl Biosynthesis
Fermentation
Carbohydrate Degradation
Respiration
Macromolecule Modification
Fatty Acid and Lipid Degradation
Inorganic Nutrient Metabolism
Alcohol Degradation
Detoxification
Secondary Metabolite Biosynthesis
Aminoacyl-tRNA Charging
Metabolic Regulator Biosynthesis
Amino Acid Biosynthesis
Amino Acid Degradation
Metabolic Clusters
Nucleoside and Nucleotide Biosynthesis
Fatty Acid and Lipid Biosynthesis
Secondary Metabolite Degradation
Carbohydrate Biosynthesis
Cell Structure Biosynthesis
Carboxylate Degradation
Aromatic Compound Degradation
Tetrapyrrole Biosynthesis
Amine and Polyamine Degradation
Aldehyde Degradation
Signal transduction pathways
Nucleoside and Nucleotide Degradation
Quick Search
Clear

```
Command:
Command: Backward History
Command:
```

Appendix C Table 1: Full list of amino acid changes in each variant strain

Variant	Location	Allele	Consequence	IMPACT	SYMBOL	BIOTYPE	cDNA_position	Protein_position	Amino_acids	Codons
7HS2	:547393-547393	T	synonymous_variant	LOW	fdrA	protein_coding	714	238	I	atC/atT
.	:557635-557635	T	missense_variant	MODERATE	folD	protein_coding	107	36	L/Q	cTg/cAg
.	:736663-736663	A	synonymous_variant	LOW	ybfQ	protein_coding	219	73	T	acC/acA
.	:826907-826907	A	missense_variant	MODERATE	ybhS	protein_coding	346	116	R/C	Cgc/Tgc
.	:856214-856214	A	synonymous_variant	LOW	ybiT	protein_coding	252	84	T	acG/acA
.	:910156-910156	A	synonymous_variant	LOW	poxB	protein_coding	894	298	I	atC/atT
.	:920907-920907	A	synonymous_variant	LOW	macB	protein_coding	561	187	L	ctG/ctA

.	:958937-958937	T	synonymous_variant	LOW	aroA	protein_coding	126	42	T	acC/acT
.	:964071-964071	T	missense_variant	MODERATE	ihfB	protein_coding	244	82	P/S	Cct/Tct
.	:972786-972786	A	missense_variant	MODERATE	elyC	protein_coding	616	206	P/S	Cca/Tca
.	:1286831-1286831	A	synonymous_variant	LOW	ychS	protein_coding	123	41	A	gcT/gcA
.	:1337394-1337394	G	missense_variant	MODERATE	acnA	protein_coding	1564	522	S/G	Agc/Ggc
.	:1456323-1456323	T	synonymous_variant	LOW	paaD	protein_coding	399	133	V	gtC/gtT
.	:1481793-1481793	T	synonymous_variant	LOW	ynbD	protein_coding	885	295	R	cgC/cgT
.	:1589769-1589769	A	missense_variant	MODERATE	ydeT	protein_coding	233	78	T/M	aCg/aTg

.	:1652331-1652331	C	non_coding_transcript_exon_variant	MODIFIER	intQ	pseudogene	820	-	-	-
.	:1723830-1723830	T	synonymous_variant	LOW	ydhK	protein_coding	1710	570	R	cgC/cgT
.	:1797018-1797018	T	synonymous_variant	LOW	pheT	protein_coding	927	309	K	aaG/aaA
.	:1825325-1825325	T	missense_variant	MODERATE	spy	protein_coding	301	101	A/T	Gct/Act
.	:1887576-1887576	A	missense_variant	MODERATE	rnd	protein_coding	416	139	W/L	tGg/tTg
.	:1983124-1983124	T	missense_variant	MODERATE	araH	protein_coding	417	139	M/I	atG/atA
.	:2016289-2016289	T	missense_variant	MODERATE	fliH	protein_coding	422	141	S/L	tCg/tTg
.	:2032423-2032423	A	synonymous_variant	LOW	yedJ	protein_coding	657	219	V	gtC/gtT

.	:2032454-2032454	T	missense_variant	MODERATE	yedJ	protein_coding	626	209	S/N	aGt/aAt
.	:2084042-2084042	A	synonymous_variant	LOW	sbcB	protein_coding	1287	429	E	gaG/gaA
.	:2091844-2091844	A	missense_variant	MODERATE	hisD	protein_coding	748	250	D/N	Gat/Aat
.	:2115845-2115845	A	missense_variant	MODERATE	wcaM	protein_coding	52	18	L/F	Ctt/Ttt
.	:2118265-2118265	A	missense_variant	MODERATE	wcaK	protein_coding	140	47	S/F	tCc/tTc
.	:2173360-2173363	C	non_coding_transcript_exon_variant	MODIFIER	gatC	pseudogene	916-918	-	-	-
.	:2586057-2586057	A	missense_variant	MODERATE	narQ	protein_coding	327	109	M/I	atG/atA
.	:2867455-2867455	A	stop_gained	HIGH	rpoS	protein_coding	97	33	Q/*	Cag/Tag

.	:3424235-3424237	CAT	non_coding_transcript_exon_variant	MODIFIER	-	rRNA	2547-2549	-	-	-
.	:3452110-3452110	T	synonymous_variant	LOW	rplD	protein_coding	177	59	P	ccG/ccA
.	:3466340-3466340	T	missense_variant	MODERATE	bfr	protein_coding	386	129	G/D	gGc/gAc
.	:3486205-3486205	A	missense_variant	MODERATE	crp	protein_coding	86	29	T/K	aCg/aAg
.	:3503729-3503729	T	missense_variant	MODERATE	frlD	protein_coding	563	188	T/I	aCa/aTa
.	:3515802-3515802	T	missense_variant	MODERATE	dam	protein_coding	112	38	A/T	Gcc/Acc
.	:3553567-3553567	T	synonymous_variant	LOW	malT	protein_coding	483	161	N	aaC/aaT
.	:3560455-3560456	GC	non_coding_transcript_exon_variant	MODIFIER	glpR	pseudogene	150	-	-	-

.	:3584919-3584919	T	missense_variant	MODERATE	yrhB	protein_coding	161	54	A/V	gCt/gTt
.	:3973249-3973249	T	missense_variant	MODERATE	rffG	protein_coding	728	243	G/V	gGg/gTg
.	:4167241-4167241	T	non_coding_transcript_exon_variant	MODIFIER	-	rRNA	583	-	-	-
.	:4170711-4170711	C	non_coding_transcript_exon_variant	MODIFIER	-	rRNA	2071	-	-	-
.	:4170806-4170806	A	non_coding_transcript_exon_variant	MODIFIER	-	rRNA	2166	-	-	-
.	:4209062-4209062	A	non_coding_transcript_exon_variant	MODIFIER	-	rRNA	916	-	-	-
.	:4210718-4210718	C	non_coding_transcript_exon_variant	MODIFIER	-	rRNA	676	-	-	-
.	:4212298-4212298	A	non_coding_transcript_exon_variant	MODIFIER	-	rRNA	2256	-	-	-

.	:4281063-4281063	T	synonymous_variant	LOW	yjcE	protein_coding	1084	362	L	Ctg/Ttg
.	:4375623-4375623	T	missense_variant	MODERATE	epmB	protein_coding	35	12	R/K	aGa/aAa
.	:4476252-4476252	A	synonymous_variant	LOW	yjgL	protein_coding	816	272	T	acG/acA
2EWL7										
.	:19901-19901	T	synonymous_variant	LOW	insB-1	protein_coding	414	138	R	cgG/cgA
.	:547393-547393	T	synonymous_variant	LOW	fdrA	protein_coding	714	238	I	atC/atT
.	:557635-557635	T	missense_variant	MODERATE	folD	protein_coding	107	36	L/Q	cTg/cAg
.	:826907-826907	A	missense_variant	MODERATE	ybhS	protein_coding	346	116	R/C	Cgc/Tgc
.	:856214-856214	A	synonymous_variant	LOW	ybiT	protein_coding	252	84	T	acG/acA
.	:910156-	A	synonymous_variant	LOW	poxB	protein_coding	894	298	I	atC/atT

	910156									
.	:920907-920907	A	synonymous_variant	LOW	macB	protein_coding	561	187	L	ctG/ctA
.	:958937-958937	T	synonymous_variant	LOW	aroA	protein_coding	126	42	T	acC/acT
.	:964071-964071	T	missense_variant	MODERATE	ihfB	protein_coding	244	82	P/S	Cct/Tct
.	:972786-972786	A	missense_variant	MODERATE	elyC	protein_coding	616	206	P/S	Cca/Tca
.	:973642-973642	T	missense_variant	MODERATE	cmoM	protein_coding	106	36	R/C	Cgc/Tgc
.	:1083915-1083915	A	synonymous_variant	LOW	efeB	protein_coding	540	180	Q	caG/caA
.	:1196220-1196220	T	synonymous_variant	LOW	icd	protein_coding	1098	366	H	caC/caT
.	:1196232-	T	synonymous_variant	LOW	icd	protein_coding	1110	370	T	acC/acT

	1196232									
.	:1196245-1196247	CTG	synonymous_variant	LOW	icd	protein_coding	1123-1125	375	L	TTA/CTG
.	:1196277-1196283	TGCCAAG	synonymous_variant	LOW	icd	protein_coding	1155-1161	385-387	NAK	aaCGCGAAA/aaTGCCAAG
.	:1196292-1196292	T	synonymous_variant	LOW	icd	protein_coding	1170	390	T	acC/acT
.	:1248239-1248239	T	synonymous_variant	LOW	dhaM	protein_coding	876	292	T	acG/acA
.	:1267312-1267312	A	missense_variant	MODERATE	ychQ	protein_coding	389	130	G/E	gGg/gAg
.	:1267884-1267884	A	missense_variant	MODERATE	ychA	protein_coding	565	189	A/T	Gcc/Acc
.	:1337394-1337394	G	missense_variant	MODERATE	acnA	protein_coding	1564	522	S/G	Agc/Ggc
.	:1456323-	T	synonymous_variant	LOW	paaD	protein_coding	399	133	V	gtC/gtT

	1456323									
.	:1466210-1466210	A	non_coding_transcript_exon_variant	MODIFIER	ydbA	pseudogene	819	-	-	-
.	:1466258-1466258	C	non_coding_transcript_exon_variant	MODIFIER	ydbA	pseudogene	867	-	-	-
.	:1481793-1481793	T	synonymous_variant	LOW	ynbD	protein_coding	885	295	R	cgC/cgT
.	:1589769-1589769	A	missense_variant	MODERATE	ydeT	protein_coding	233	78	T/M	aCg/aTg
.	:1652331-1652331	C	non_coding_transcript_exon_variant	MODIFIER	intQ	pseudogene	820	-	-	-
.	:1723830-1723830	T	synonymous_variant	LOW	ydhK	protein_coding	1710	570	R	cgC/cgT
.	:1797018-1797018	T	synonymous_variant	LOW	pheT	protein_coding	927	309	K	aaG/aaA
.	:1825325-	T	missense_variant	MODERATE	spy	protein_coding	301	101	A/T	Gct/Act

	1825325									
.	:1887576-1887576	A	missense_variant	MODERATE	rnd	protein_coding	416	139	W/L	tGg/tTg
.	:1983124-1983124	T	missense_variant	MODERATE	araH	protein_coding	417	139	M/I	atG/atA
.	:2001248-2001248	T	missense_variant	MODERATE	fliA	protein_coding	542	181	R/Q	cGg/cAg
.	:2016289-2016289	T	missense_variant	MODERATE	fliH	protein_coding	422	141	S/L	tCg/tTg
.	:2032423-2032423	A	synonymous_variant	LOW	yedJ	protein_coding	657	219	V	gtC/gtT
.	:2032454-2032454	T	missense_variant	MODERATE	yedJ	protein_coding	626	209	S/N	aGt/aAt
.	:2062323-2062323	A	non_coding_transcript_exon_variant	MODIFIER	-	tRNA	64	-	-	-
.	:2084042-	A	synonymous_variant	LOW	sbcB	protein_coding	1287	429	E	gaG/gaA

	2084042									
.	:2091844-2091844	A	missense_variant	MODERATE	hisD	protein_coding	748	250	D/N	Gat/Aat
.	:2115845-2115845	A	missense_variant	MODERATE	wcaM	protein_coding	52	18	L/F	Ctt/Ttt
.	:2118265-2118265	A	missense_variant	MODERATE	wcaK	protein_coding	140	47	S/F	tCc/tTc
.	:2173360-2173363	C	non_coding_transcript_exon_variant	MODIFIER	gatC	pseudogene	916-918	-	-	-
.	:2586057-2586057	A	missense_variant	MODERATE	narQ	protein_coding	327	109	M/I	atG/atA
.	:2793544-2793544	T	synonymous_variant	LOW	gabT	protein_coding	810	270	I	atC/atT
.	:2793560-2793560	T	synonymous_variant	LOW	gabT	protein_coding	826	276	L	Ctg/Ttg
.	:2867455-	A	stop_gained	HIGH	rpoS	protein_coding	97	33	Q/*	Cag/Tag

	2867455									
.	:2965105-2965105	A	missense_variant	MODERATE	thyA	protein_coding	51	17	Q/H	caG/caT
.	:3113408-3113408	T	missense_variant	MODERATE	yghG	protein_coding	70	24	G/S	Ggc/Agc
.	:3203725-3203725	A	missense_variant	MODERATE	bacA	protein_coding	407	136	A/V	gCc/gTc
.	:3203745-3203745	A	synonymous_variant	LOW	bacA	protein_coding	387	129	G	ggC/ggT
.	:3204055-3204055	A	missense_variant	MODERATE	bacA	protein_coding	77	26	S/F	tCc/tTc
.	:3424235-3424237	CAT	non_coding_transcript_exon_variant	MODIFIER	-	rRNA	2547-2549	-	-	-
.	:3426170-3426170	G	non_coding_transcript_exon_variant	MODIFIER	-	rRNA	614	-	-	-
.	:3452110-	T	synonymous_variant	LOW	rplD	protein_coding	177	59	P	ccG/ccA

	3452110									
.	:3466340-3466340	T	missense_variant	MODERATE	bfr	protein_coding	386	129	G/D	gGc/gAc
.	:3486205-3486205	A	missense_variant	MODERATE	crp	protein_coding	86	29	T/K	aCg/aAg
.	:3503729-3503729	T	missense_variant	MODERATE	frlD	protein_coding	563	188	T/I	aCa/aTa
.	:3515802-3515802	T	missense_variant	MODERATE	dam	protein_coding	112	38	A/T	Gcc/Acc
.	:3553567-3553567	T	synonymous_variant	LOW	malT	protein_coding	483	161	N	aaC/aaT
.	:3560455-3560456	GC	non_coding_transcript_exon_variant	MODIFIER	glpR	pseudogene	150	-	-	-
.	:3565072-3565072	A	stop_gained	HIGH	glgP	protein_coding	1510	504	Q/*	Caa/Taa
.	:3584900-	A	missense_variant	MODERATE	yrhB	protein_coding	142	48	D/N	Gat/Aat

	3584900									
.	:3584919-3584919	T	missense_variant	MODERATE	yrhB	protein_coding	161	54	A/V	gCt/gTt
.	:3763232-3763232	A	synonymous_variant	LOW	rhsA	protein_coding	1050	350	R	cgT/cgA
.	:4212298-4212298	A	non_coding_transcript_exon_variant	MODIFIER	-	rRNA	2256	-	-	-
.	:4281063-4281063	T	synonymous_variant	LOW	yjcE	protein_coding	1084	362	L	Ctg/Ttg
.	:4375623-4375623	T	missense_variant	MODERATE	epmB	protein_coding	35	12	R/K	aGa/aAa
.	:4412363-4412363	G	missense_variant	MODERATE	yjfL	protein_coding	385	129	C/G	Tgt/Ggt
.	:4476252-4476252	A	synonymous_variant	LOW	yjgL	protein_coding	816	272	T	acG/acA
.	:4555970-	T	synonymous_variant	LOW	yjiC	protein_coding	351	117	A	gcG/gcA

	4555970									
.	:4621059-4621059	T	synonymous_variant	LOW	deoD	protein_coding	177	59	S	tcC/tcT
W3110										
.	:70434-70434	T	synonymous_variant	LOW	araC	protein_coding	48	16	N	aaC/aaT
.	:70581-70581	T	synonymous_variant	LOW	araC	protein_coding	195	65	V	gtC/gtT
.	:70740-70740	G	synonymous_variant	LOW	araC	protein_coding	354	118	G	ggT/ggG
.	:71079-71085	TATCAGC	synonymous_variant	LOW	araC	protein_coding	693-699	231-233	RIS	cgCATTAGT/cgTATCAGC
.	:71109-71109	C	synonymous_variant	LOW	araC	protein_coding	723	241	T	acT/acC
.	:71175-71175	G	synonymous_variant	LOW	araC	protein_coding	789	263	R	cgA/cgG
.	:71214-71214	C	synonymous_variant	LOW	araC	protein_coding	828	276	F	ttT/ttC
.	:71356-71356	T	missense_variant	MODERATE	yabI	protein_coding	6	2	Q/H	caA/caT
.	:366138-366138	T	missense_variant	MODERATE	lacZ	protein_coding	168	56	N/K	aaT/aaA
.	:366226-366226	A	missense_variant	MODERATE	lacZ	protein_coding	80	27	R/L	cGc/cTc

.	:557635-557635	T	missense_variant	MODERATE	folD	protein_coding	107	36	L/Q	cTg/cAg
.	:1270419-1270419	C	non_coding_transcript_exon_variant	MODIFIER	-	ncRNA	27	-	-	-
.	:1337394-1337394	G	missense_variant	MODERATE	acnA	protein_coding	1564	522	S/G	Agc/Ggc
.	:1652331-1652331	C	non_coding_transcript_exon_variant	MODIFIER	intQ	pseudogene	820	-	-	-
.	:2032423-2032423	A	synonymous_variant	LOW	yedJ	protein_coding	657	219	V	gtC/gtT
.	:2173360-2173363	C	non_coding_transcript_exon_variant	MODIFIER	gatC	pseudogene	916-918	-	-	-
.	:2867455-2867455	A	stop_gained	HIGH	rpoS	protein_coding	97	33	Q/*	Cag/Tag
.	:3424235-3424237	CAT	non_coding_transcript_exon_variant	MODIFIER	-	rRNA	2547-2549	-	-	-

.	:3486205-3486205	A	missense_variant	MODERATE	crp	protein_coding	86	29	T/K	aCg/aAg
.	:3560455-3560456	GC	non_coding_transcript_exon_variant	MODIFIER	glpR	pseudogene	150	-	-	-
.	:4212298-4212298	A	non_coding_transcript_exon_variant	MODIFIER	-	rRNA	2256	-	-	-
.	:4412779-4412779	G	synonymous_variant	LOW	yjfM	protein_coding	393	131	V	gtT/gtG

Appendix D: Supplementary material for Chapter 6

Appendix D Table 1: Full list of proteins with unique peptide

Protein names	Gene names	Unique peptides	Mol. weight [kDa]
Chaperone protein DnaK (HSP70) (Heat shock 70 kDa protein) (Heat shock protein 70)	dnaK groP grpF seg b0014 JW0013	36	69.114
Aconitate hydratase B (ACN) (Aconitase) (EC 4.2.1.3) ((2R,3S)-2-methylisocitrate dehydratase) ((2S,3R)-3-hydroxybutane-1,2,3-tricarboxylate dehydratase) (2-methyl-cis-aconitate hydratase) (EC 4.2.1.99) (Iron-responsive protein-like) (IRP-like) (RNA-binding protein)	acnB yacI yacJ b0118 JW0114	33	93.497
Catalase HPII (EC 1.11.1.6) (Hydroxyperoxidase II)	katE b1732 JW1721	29	84.162
Glutamate decarboxylase beta (GAD-beta) (EC 4.1.1.15)	gadB b1493 JW1488	26	52.668
Glutamate decarboxylase alpha (GAD-alpha) (EC 4.1.1.15)	gadA gadS b3517 JW3485	26	51.481
6-phosphogluconate dehydrogenase, decarboxylating (EC 1.1.1.44)	gnd b2029 JW2011	26	77.58
Elongation factor G (EF-G)	fusA far fus b3340 JW3302	25	87.377
Phenylalanine--tRNA ligase beta subunit (EC 6.1.1.20) (Phenylalanyl-tRNA synthetase beta subunit) (PheRS)	pheT b1713 JW1703	25	63.561
Phosphoenolpyruvate-protein phosphotransferase (EC 2.7.3.9) (Phosphotransferase system, enzyme I)	ptsI b2416 JW2409	25	80.023
Catalase-peroxidase (CP) (EC 1.11.1.21) (Hydroperoxidase I) (HPI) (Peroxidase/catalase)	katG b3942 JW3914	24	85.356
Formate acetyltransferase 1 (EC 2.3.1.54) (Pyruvate formate-lyase 1)	pflB pfl b0903 JW0886	24	73.042
PFL-like enzyme TdcE (Keto-acid formate acetyltransferase) (Keto-acid formate-lyase) (Ketobutyrate formate-lyase) (KFL) (EC 2.3.1.-) (Pyruvate formate-lyase) (PFL) (EC 2.3.1.54)	tdcE yhaS b3114 JW5522	24	95.584
Transketolase 2 (TK 2) (EC 2.2.1.1)	tktB b2465 JW2449	24	96.031
Chaperone protein ClpB (Heat shock protein F84.1)	clpB htpM b2592 JW2573	23	52.272

Alanine--tRNA ligase (EC 6.1.1.7) (Alanyl-tRNA synthetase) (AlaRS)	alaS lovB b2697 JW2667	23	60.273
Lactaldehyde dehydrogenase (EC 1.2.1.22) (Aldehyde dehydrogenase A) (Glycolaldehyde dehydrogenase) (EC 1.2.1.21)	aldA ald b1415 JW1412	22	57.328
Malate synthase A (MSA) (EC 2.3.3.9)	aceB mas b4014 JW3974	21	52.773
60 kDa chaperonin (GroEL protein) (Protein Cpn60)	groL groEL mopA b4143 JW4103	20	60.898
Tryptophanase (EC 4.1.99.1) (L-tryptophan indole-lyase) (TNase)	tnaA ind b3708 JW3686	20	97.676
Periplasmic oligopeptide-binding protein	oppA b1243 JW1235	20	35.532
Aconitate hydratase A (ACN) (Aconitase) (EC 4.2.1.3) (Iron-responsive protein-like) (IRP-like) (RNA-binding protein) (Stationary phase enzyme)	acnA acn b1276 JW1268	20	41.118
Glyceraldehyde-3-phosphate dehydrogenase A (GAPDH-A) (EC 1.2.1.12) (NAD-dependent glyceraldehyde-3-phosphate dehydrogenase)	gapA b1779 JW1768	20	56.306
Phosphoglycerate kinase (EC 2.7.2.3)	pgk b2926 JW2893	19	99.667
Aldehyde dehydrogenase B (EC 1.2.1.4) (Acetaldehyde dehydrogenase)	aldB yiaX b3588 JW3561	19	71.422
Pyruvate dehydrogenase E1 component (PDH E1 component) (EC 1.2.4.1)	aceE b0114 JW0110	19	98.918
Chaperone protein HtpG (Heat shock protein C62.5) (Heat shock protein HtpG) (High temperature protein G)	htpG b0473 JW0462	19	50.729
Aminopeptidase N (EC 3.4.11.2) (Alpha-aminoacylpeptide hydrolase)	pepN b0932 JW0915	19	45.774
Pyruvate kinase I (EC 2.7.1.40) (PK-1)	pykF b1676 JW1666	19	55.221
4-aminobutyrate aminotransferase GabT (EC 2.6.1.19) (5-aminovalerate transaminase) (EC 2.6.1.48) (GABA aminotransferase) (GABA-AT) (Gamma-amino-N-butyrate transaminase) (GABA transaminase) (Glutamate:succinic semialdehyde transaminase) (L-AIBAT)	gabT b2662 JW2637	19	47.521
ATP synthase subunit alpha (EC 7.1.2.2) (ATP synthase F1 sector subunit alpha) (F-ATPase subunit alpha)	atpA papA uncA b3734 JW3712	18	41.392

Isocitrate lyase (ICL) (EC 4.1.3.1) (Isocitrase) (Isocitratase)	aceA icl b4015 JW3975	18	77.1
Succinate--CoA ligase [ADP-forming] subunit beta (EC 6.2.1.5) (Succinyl-CoA synthetase subunit beta) (SCS-beta)	sucC b0728 JW0717	18	32.337
Polyribonucleotide nucleotidyltransferase (EC 2.7.7.8) (Polynucleotide phosphorylase) (PNPase)	pnp b3164 JW5851	18	61.529
Malate dehydrogenase (EC 1.1.1.37)	mdh b3236 JW3205	17	51.719
Glucose-6-phosphate isomerase (GPI) (EC 5.3.1.9) (Phosphoglucose isomerase) (PGI) (Phosphohexose isomerase) (PHI)	pgi b4025 JW3985	17	60.293
Succinate-semialdehyde dehydrogenase [NADP(+)] GabD (SSDH) (EC 1.2.1.79) (Glutarate-semialdehyde dehydrogenase) (EC 1.2.1.-)	gabD b2661 JW2636	16	61.157
Dipeptide-binding protein (DBP) (Periplasmic dipeptide transport protein)	dppA b3544 JW3513	16	27.991
30S ribosomal protein S1 (Bacteriophage Q beta RNA-directed RNA polymerase subunit I) (Small ribosomal subunit protein bS1)	rpsA ssyF b0911 JW0894	16	58.679
Lysine/arginine/ornithine-binding periplasmic protein (LAO-binding protein)	argT b2310 JW2307	16	104.38
GMP synthase [glutamine-hydrolyzing] (EC 6.3.5.2) (GMP synthetase) (GMPS) (Glutamine amidotransferase)	guaA b2507 JW2491	16	43.313
Glycine dehydrogenase (decarboxylating) (EC 1.4.4.2) (Glycine cleavage system P-protein) (Glycine decarboxylase) (Glycine dehydrogenase (aminomethyl-transferring))	gcvP b2903 JW2871	16	108.19
Elongation factor Tu 2 (EF-Tu 2) (Bacteriophage Q beta RNA-directed RNA polymerase subunit III) (P-43)	tufB b3980 JW3943	16	72.211
Elongation factor Tu 1 (EF-Tu 1) (Bacteriophage Q beta RNA-directed RNA polymerase subunit III) (P-43)	tufA b3339 JW3301	15	30.423
Valine--tRNA ligase (EC 6.1.1.9) (Valyl-tRNA synthetase) (ValRS)	valS b4258 JW4215	15	29.777
Transketolase 1 (TK 1) (EC 2.2.1.1)	tktA tkt b2935 JW5478	15	55.704

Elongation factor Ts (EF-Ts) (Bacteriophage Q beta RNA-directed RNA polymerase subunit IV)	tsf b0170 JW0165	15	29.039
Succinate--CoA ligase [ADP-forming] subunit alpha (EC 6.2.1.5) (Succinyl-CoA synthetase subunit alpha) (SCS-alpha)	sucD b0729 JW0718	15	38.109
Glucose-6-phosphate 1-dehydrogenase (G6PD) (EC 1.1.1.49) [Cleaved into: Extracellular death factor (EDF)]	zwf b1852 JW1841	15	76.254
L-cystine-binding protein TcyJ (CBP) (Protein FliY) (Sulfate starvation-induced protein 7) (SSI7)	tcyJ fliY yzzR b1920 JW1905	15	44.369
Fructose-bisphosphate aldolase class 1 (EC 4.1.2.13) (Fructose-bisphosphate aldolase class I) (FBP aldolase)	fbaB dhnA b2097 JW5344	14	104.3
Methionine--tRNA ligase (EC 6.1.1.10) (Methionyl-tRNA synthetase) (MetRS)	metG b2114 JW2101	14	52.915
Phosphopentomutase (EC 5.4.2.7) (Phosphodeoxyribomutase)	deoB drm thyR b4383 JW4346	14	48.192
Isoleucine--tRNA ligase (EC 6.1.1.5) (Isoleucyl-tRNA synthetase) (IleRS)	ileS ilvS b0026 JW0024	14	48.014
Cytosol non-specific dipeptidase (EC 3.4.13.18) (Aminoacyl-histidine dipeptidase) (Beta-alanyl-histidine dipeptidase) (Carnosinase) (Cysteinylglycinase) (Peptidase D) (Xaa-His dipeptidase) (X-His dipeptidase)	pepD pepH b0237 JW0227	14	38.867
Trigger factor (TF) (EC 5.2.1.8) (PPIase)	tig b0436 JW0426	14	96.126
Citrate synthase (EC 2.3.3.16)	gltA gluT icdB b0720 JW0710	14	50.829
Spermidine/putrescine-binding periplasmic protein (SPBP)	potD b1123 JW1109	14	87.434
Aldehyde-alcohol dehydrogenase [Includes: Alcohol dehydrogenase (ADH) (EC 1.1.1.1); Acetaldehyde dehydrogenase [acetylating] (ACDH) (EC 1.2.1.10); Pyruvate-formate-lyase deactivase (PFL deactivase)]	adhE ana b1241 JW1228	14	65.913
Gamma-aminobutyraldehyde dehydrogenase (ABALDH) (EC 1.2.1.19) (1-pyrroline dehydrogenase) (4-aminobutanal dehydrogenase) (5-aminopentanal dehydrogenase) (EC 1.2.1.-)	patD prr ydcW b1444 JW1439	14	35.712

Phosphoenolpyruvate synthase (PEP synthase) (EC 2.7.9.2) (Pyruvate, water dikinase)	ppsA pps b1702 JW1692	14	77.171
Aspartate--tRNA ligase (EC 6.1.1.12) (Aspartyl-tRNA synthetase) (AspRS)	aspS tls b1866 JW1855	14	46.18
D-galactose-binding periplasmic protein (GBP) (D-galactose/ D-glucose-binding protein) (GGBP)	mglB b2150 JW2137	14	45.654
Phosphate acetyltransferase (EC 2.3.1.8) (Phosphotransacetylase)	pta b2297 JW2294	14	77.166
Peptidase B (EC 3.4.11.23) (Aminopeptidase B)	pepB yfhI b2523 JW2507	14	30.95
Enolase (EC 4.2.1.11) (2-phospho-D-glycerate hydro-lyase) (2-phosphoglycerate dehydratase)	eno b2779 JW2750	14	49.593
Oligopeptidase A (EC 3.4.24.70)	prlC opdA b3498 JW3465	14	99.061
Ribose import binding protein RbsB	rbsB prlB rbsP b3751 JW3730	13	35.219
ATP-dependent protease ATPase subunit HslU (Heat shock protein HslU) (Unfoldase HslU)	hslU htpI b3931 JW3902	13	28.556
Phosphoenolpyruvate carboxylase (PEPC) (PEPCase) (EC 4.1.1.31)	ppc glu b3956 JW3928	13	43.573
Transaldolase B (EC 2.2.1.2)	talB yaaK b0008 JW0007	13	45.756
2,3-bisphosphoglycerate-dependent phosphoglycerate mutase (BPG-dependent PGAM) (PGAM) (Phosphoglyceromutase) (dPGM) (EC 5.4.2.11)	gpmA gpm pgm pgmA b0755 JW0738	13	47.344
Aspartate aminotransferase (AspAT) (EC 2.6.1.1) (Transaminase A)	aspC b0928 JW0911	12	62.011
Isocitrate dehydrogenase [NADP] (IDH) (EC 1.1.1.42) (IDP) (NADP(+)-specific ICDH) (Oxalosuccinate decarboxylase)	icd icdA icdE b1136 JW1122	12	37.2
Adenylosuccinate synthetase (AMPSase) (AdSS) (EC 6.3.4.4) (IMP--aspartate ligase)	purA adeK b4177 JW4135	12	63.197
Pyruvate dehydrogenase [ubiquinone] (EC 1.2.5.1) (Pyruvate oxidase) (POX) (Pyruvate:ubiquinone-8 oxidoreductase)	poxB b0871 JW0855	12	32.609
Outer membrane protein A (OmpA) (Outer membrane porin A) (Outer membrane protein 3A) (Outer membrane protein B) (Outer membrane protein II*) (Outer membrane protein d)	ompA con tolG tut b0957 JW0940	12	34.489

NAD-dependent malic enzyme (NAD-ME) (EC 1.1.1.38)	maeA sfcA b1479 JW5238	12	37.614
Glycine betaine-binding protein YehZ	yehZ osmF b2131 JW2119	12	39.147
Cysteine synthase A (CSase A) (EC 2.5.1.47) (O-acetylserine (thiol)-lyase A) (OAS-TL A) (O-acetylserine sulfhydrylase A) (S-carboxymethylcysteine synthase) (EC 4.5.1.5) (Sulfate starvation-induced protein 5) (SSI5)	cysK cysZ b2414 JW2407	12	76.812
Thiosulfate-binding protein	cysP b2425 JW2418	12	66.894
Fructose-bisphosphate aldolase class 2 (FBP aldolase) (FBPA) (EC 4.1.2.13) (Fructose-1,6-bisphosphate aldolase) (Fructose-bisphosphate aldolase class II)	fbaA fba fda b2925 JW2892	12	50.325
Glycine--tRNA ligase beta subunit (EC 6.1.1.14) (Glycyl-tRNA synthetase beta subunit) (GlyRS)	glyS glyS(B) b3559 JW3530	11	47.283
Glutamine--fructose-6-phosphate aminotransferase [isomerizing] (EC 2.6.1.16) (D-fructose-6-phosphate amidotransferase) (GFAT) (Glucosamine-6-phosphate synthase) (Hexosephosphate aminotransferase) (L-glutamine--D-fructose-6-phosphate amidotransferase)	glmS b3729 JW3707	11	56.073
ATP synthase subunit beta (EC 7.1.2.2) (ATP synthase F1 sector subunit beta) (F-ATPase subunit beta)	atpD papB uncD b3732 JW3710	11	50.688
Chaperone SurA (Peptidyl-prolyl cis-trans isomerase SurA) (PPIase SurA) (EC 5.2.1.8) (Rotamase SurA) (Survival protein A)	surA b0053 JW0052	11	49.354
L-arabinose isomerase (EC 5.3.1.4)	araA b0062 JW0061	11	29.892
Dihydrolipoyl dehydrogenase (EC 1.8.1.4) (Dihydrolipoamide dehydrogenase) (E3 component of pyruvate and 2-oxoglutarate dehydrogenases complexes) (Glycine cleavage system L protein)	lpdA lpd b0116 JW0112	11	27.19
Periplasmic serine endoprotease DegP (EC 3.4.21.107) (Heat shock protein DegP) (Protease Do)	degP htrA ptd b0161 JW0157	11	21.265
2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase (EC 2.3.1.117) (Succinyl-CoA: tetrahydrodipicolinate N-	dapD b0166 JW0161	11	36.831

succinyltransferase) (Tetrahydrodipicolinate N-succinyltransferase) (THDP succinyltransferase) (THP succinyltransferase) (Tetrahydropicolinate succinylase)			
Glutamine-binding periplasmic protein (GlnBP)	glnH b0811 JW0796	11	82.416
Superoxide dismutase [Fe] (EC 1.15.1.1)	sodB b1656 JW1648	11	16.063
Phenylalanine--tRNA ligase alpha subunit (EC 6.1.1.20) (Phenylalanyl-tRNA synthetase alpha subunit) (PheRS)	pheS b1714 JW5277	11	31.488
NADP-dependent malic enzyme (NADP-ME) (EC 1.1.1.40)	maeB ypfF b2463 JW2447	11	28.882
Potassium binding protein Kbp (K(+) binding protein Kbp)	kbp ygaU yzzM b2665 JW2640	11	57.826
Uncharacterized oxidoreductase YghA (EC 1.-.-.-)	yghA b3003 JW2972	11	52.356
FKBP-type peptidyl-prolyl cis-trans isomerase FkpA (PPIase) (EC 5.2.1.8) (Rotamase)	fkpA yzzS b3347 JW3309	11	19.703
Lysine--tRNA ligase, heat inducible (EC 6.1.1.6) (Lysyl-tRNA synthetase) (LysRS)	lysU b4129 JW4090	11	25.95
Aspartate ammonia-lyase (Aspartase) (EC 4.3.1.1)	aspA b4139 JW4099	10	44.63
Inorganic pyrophosphatase (EC 3.6.1.1) (Pyrophosphate phospho-hydrolase) (PPase)	ppa b4226 JW4185	10	105.06
Purine nucleoside phosphorylase DeoD-type (PNP) (EC 2.4.2.1)	deoD pup b4384 JW4347	10	26.929
Gamma-glutamyl phosphate reductase (GPR) (EC 1.2.1.41) (Glutamate-5-semialdehyde dehydrogenase) (Glutamyl-gamma-semialdehyde dehydrogenase) (GSA dehydrogenase)	proA b0243 JW0233	10	23.1
2-oxoglutarate dehydrogenase E1 component (EC 1.2.4.2) (Alpha-ketoglutarate dehydrogenase)	sucA b0726 JW0715	10	52.57
Putative ABC transporter arginine-binding protein 2	artI b0863 JW0847	10	17.835
Probable hydrolase YcaC (EC 4.-.-.-)	ycaC b0897 JW0880	10	42.295
Asparagine--tRNA ligase (EC 6.1.1.22) (Asparaginyl-tRNA synthetase) (AsnRS)	asnS tss b0930 JW0913	10	35.379
Thiol peroxidase (Tpx) (EC 1.11.1.24) (Peroxiredoxin tpx) (Prx) (Scavengase	tpx yzzJ b1324 JW1317	10	51.357

p20) (Thioredoxin peroxidase) (Thioredoxin-dependent peroxiredoxin)			
Bifunctional polyhydroxybutyrate synthase / ABC transporter periplasmic binding protein (Poly-3-hydroxybutyrate synthase) (PHB synthase) (EC 2.3.1.-) (cPHB synthase)	ydcS b1440 JW1435	10	31.19
Alcohol dehydrogenase, propanol-preferring (EC 1.1.1.1)	adhP yddN b1478 JW1474	10	47.108
Pyruvate kinase II (EC 2.7.1.40) (PK-2)	pykA b1854 JW1843	10	35.658
Protein/nucleic acid deglycase 1 (EC 3.1.2.-) (EC 3.5.1.-) (EC 3.5.1.124) (Glyoxalase III) (EC 4.2.1.130) (Holding molecular chaperone) (Hsp31) (Maillard deglycase)	hchA yedU yzzC b1967 JW1950	10	37.973
D-tagatose-1,6-bisphosphate aldolase subunit GatZ	gatZ b2095 JW2082	10	51.903
D-tagatose-1,6-bisphosphate aldolase subunit KbaZ	kbaZ agaZ yhaX b3132 JW3101	10	72.093
Transaldolase A (EC 2.2.1.2)	talA b2464 JW2448	10	62.442
Protein RecA (Recombinase A)	recA lexB recH rnmB tif umuB zab b2699 JW2669	10	106.82
Glutamine synthetase (GS) (EC 6.3.1.2) (Glutamate--ammonia ligase) (Glutamine synthetase I beta) (GSI beta)	glnA b3870 JW3841	9	41.438
Acetyl-coenzyme A synthetase (AcCoA synthetase) (Acs) (EC 6.2.1.1) (Acetate--CoA ligase) (Acyl-activating enzyme)	acs yfaC b4069 JW4030	9	48.413
Energy-dependent translational throttle protein EttA (EC 3.6.1.-) (Translational regulatory factor EttA)	ettA yjjK b4391 JW4354	9	45.682
Antigen 43 (AG43) (Fluffing protein) [Cleaved into: Antigen 43 alpha chain; Antigen 43 beta chain]	flu yeeQ yzzX b2000 JW1982	9	40.368
Serine--tRNA ligase (EC 6.1.1.11) (Seryl-tRNA synthetase) (SerRS) (Seryl-tRNA(Ser/Sec) synthetase)	serS b0893 JW0876	9	40.146
Glucose-1-phosphatase (G1Pase) (EC 3.1.3.10)	agp b1002 JW0987	9	31.109
Outer membrane porin C (Outer membrane protein 1B) (Outer membrane protein C) (Porin OmpC)	ompC meoA par b2215 JW2203	9	18.858
Aminomethyltransferase (EC 2.1.2.10) (Glycine cleavage system T protein)	gcvT b2905 JW2873	9	59.643

2,5-diketo-D-gluconic acid reductase A (2,5-DKG reductase A) (2,5-DKGR A) (25DKGR-A) (EC 1.1.1.274) (AKR5C)	dkgA yqhE b3012 JW5499	9	34.842
Protein/nucleic acid deglycase 2 (EC 3.1.2.-) (EC 3.5.1.-) (EC 3.5.1.124) (Maillard deglycase)	yhbO b3153 JW5529	8	60.823
Phosphoenolpyruvate carboxykinase (ATP) (PCK) (PEP carboxykinase) (PEPCK) (EC 4.1.1.49)	pckA pck b3403 JW3366	8	97.233
ATP-dependent 6-phosphofructokinase isozyme 1 (ATP-PFK 1) (Phosphofructokinase 1) (EC 2.7.1.11) (6-phosphofructokinase isozyme I) (Phosphohexokinase 1)	pfkA b3916 JW3887	8	33.42
Protein UshA [Includes: UDP-sugar hydrolase (EC 3.6.1.45) (UDP-sugar diphosphatase) (UDP-sugar pyrophosphatase); 5'-nucleotidase (5'-NT) (EC 3.1.3.5)]	ushA b0480 JW0469	8	58.36
Leucine--tRNA ligase (EC 6.1.1.4) (Leucyl-tRNA synthetase) (LeuRS)	leuS b0642 JW0637	8	27.864
Glutamate/aspartate import solute-binding protein	gltI ybeJ yzzK b0655 JW5092	8	15.088
Phosphoglucomutase (PGM) (EC 5.4.2.2) (Glucose phosphomutase)	pgm b0688 JW0675	8	33.903
Enoyl-[acyl-carrier-protein] reductase [NADH] FabI (ENR) (EC 1.3.1.9) (NADH-dependent enoyl-ACP reductase)	fabI envM b1288 JW1281	8	17.681
Peroxiredoxin OsmC (EC 1.11.1.-) (Osmotically-inducible protein C)	osmC b1482 JW1477	8	26.384
Protein YdgH	ydgH b1604 JW1596	8	49.815
Superoxide dismutase [Cu-Zn] (EC 1.15.1.1) (Bacteriocuprein)	sodC b1646 JW1638	8	26.635
Pyridoxine 5'-phosphate synthase (PNP synthase) (EC 2.6.99.2)	pdxJ b2564 JW2548	8	33.557
Xaa-Pro aminopeptidase (EC 3.4.11.9) (Aminoacylproline aminopeptidase) (Aminopeptidase P II) (APP-II) (X-Pro aminopeptidase)	pepP b2908 JW2876	8	80.488
Uncharacterized protein YggE	yggE b2922 JW2889	8	14.011
Agmatinase (EC 3.5.3.11) (Agmatine ureohydrolase) (AUH)	speB b2937 JW2904	8	18.904
Malate synthase G (MSG) (EC 2.3.3.9)	glcB glc b2976 JW2943	8	18.495
Protein YgiW	ygiW b3024 JW2992	8	93.171
50S ribosomal protein L6 (Large ribosomal subunit protein uL6)	rplF b3305 JW3267	8	48.448

Bacterioferritin (BFR) (EC 1.16.3.1) (Cytochrome b-1) (Cytochrome b-557)	bfr b3336 JW3298	8	27.159
Glycogen phosphorylase (EC 2.4.1.1)	glgP glgY b3428 JW3391	8	50.176
sn-glycerol-3-phosphate-binding periplasmic protein UgpB	ugpB b3453 JW3418	8	48.369
Uridine phosphorylase (UPase) (UrdPase) (EC 2.4.2.3)	udp b3831 JW3808	8	21.073
Xaa-Pro dipeptidase (X-Pro dipeptidase) (EC 3.4.13.9) (Imidodipeptidase) (Proline dipeptidase) (Prolidase)	pepQ b3847 JW3823	7	63.692
Metalloprotease PmbA (EC 3.4.-.-) (Protein TldE)	pmbA tldE b4235 JW4194	7	52.911
Osmotically-inducible protein Y	osmY b4376 JW4338	7	37.978
Proline--tRNA ligase (EC 6.1.1.15) (Global RNA synthesis factor) (Prolyl-tRNA synthetase) (ProRS)	proS drpA b0194 JW0190	7	28.145
Betaine aldehyde dehydrogenase (BADH) (EC 1.2.1.8)	betB b0312 JW0304	7	20.761
Aldehyde reductase YahK (EC 1.1.1.2) (Zinc-dependent alcohol dehydrogenase YahK)	yahK b0325 JW0317	7	40.949
Pyrroline-5-carboxylate reductase (P5C reductase) (P5CR) (EC 1.5.1.2) (PCA reductase)	proC b0386 JW0377	7	44.011
Alkyl hydroperoxide reductase C (EC 1.11.1.26) (Alkyl hydroperoxide reductase protein C22) (Peroxiredoxin) (SCRP-23) (Sulfate starvation-induced protein 8) (SSI8) (Thioredoxin peroxidase)	ahpC b0605 JW0598	7	36.307
N-acetylglucosamine-6-phosphate deacetylase (GlcNAc 6-P deacetylase) (EC 3.5.1.25)	nagA b0677 JW0663	7	18.695
Dihydrolipoyllysine-residue succinyltransferase component of 2-oxoglutarate dehydrogenase complex (EC 2.3.1.61) (2-oxoglutarate dehydrogenase complex component E2) (OGDC-E2) (Dihydrolipoamide succinyltransferase component of 2-oxoglutarate dehydrogenase complex)	sucB b0727 JW0716	7	22.497
6-phosphogluconolactonase (6-P-gluconolactonase) (Pgl) (EC 3.1.1.31)	pgl ybhE b0767 JW0750	7	20.912
DNA protection during starvation protein (EC 1.16.-.-)	dps pexB vtm b0812 JW0797	7	24.35

Outer-membrane lipoprotein carrier protein (P20)	lolA lplA yzzV b0891 JW0874	7	51.542
Protein YceI	yceI b1056 JW1043	7	38.494
Glutaredoxin 2 (Grx2)	grxB b1064 JW1051	7	38.612
Adenylosuccinate lyase (ASL) (EC 4.3.2.2) (Adenylosuccinase) (ASase)	purB b1131 JW1117	7	14.121
Uncharacterized protein YncE	yncE b1452 JW1447	7	36.082
Uncharacterized protein YdeI	ydeI b1536 JW1529	7	20.452
Probable L,D-transpeptidase YnhG (EC 2.-.-.-)	ynhG b1678 JW1668	7	57.603
Isochorismatase family protein YecD (EC 3.-.-.-)	yecD b1867 JW5307	7	49.898
Lysine--tRNA ligase (EC 6.1.1.6) (Lysyl-tRNA synthetase) (LysRS)	lysS asuD herC b2890 JW2858	7	51.363
Argininosuccinate synthase (EC 6.3.4.5) (Citrulline--aspartate ligase)	argG b3172 JW3140	7	36.511
Metalloprotease TldD (EC 3.4.-.-)	tldD yhdO b3244 JW3213	7	55.527
DNA-directed RNA polymerase subunit alpha (RNAP subunit alpha) (EC 2.7.7.6) (RNA polymerase subunit alpha) (Transcriptase subunit alpha)	rpoA pez phs sez b3295 JW3257	7	26.972
Protein YhjJ	yhjJ b3527 JW3495	6	66.095
Triosephosphate isomerase (TIM) (TPI) (EC 5.3.1.1) (Triose-phosphate isomerase)	tpiA tpi b3919 JW3890	6	90.552
Dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex (EC 2.3.1.12) (Dihydrolipoamide acetyltransferase component of pyruvate dehydrogenase complex) (E2)	aceF b0115 JW0111	6	20.815
Outer membrane protein assembly factor BamA (Omp85)	bamA yaeT yzzN yzzY b0177 JW0172	6	35.624
Phosphoheptose isomerase (EC 5.3.1.28) (Sedoheptulose 7-phosphate isomerase)	gmhA lpcA tfrA yafI b0222 JW0212	6	16.156
Delta-aminolevulinic acid dehydratase (ALAD) (ALADH) (EC 4.2.1.24) (Porphobilinogen synthase)	hemB ncf b0369 JW0361	6	63.477
6,7-dimethyl-8-ribityllumazine synthase (DMRL synthase) (LS) (Lumazine synthase) (EC 2.5.1.78)	ribE ribH ybaF b0415 JW0405	6	25.56
Glutamine--tRNA ligase (EC 6.1.1.18) (Glutaminyl-tRNA synthetase) (GlnRS)	glnS b0680 JW0666	6	63.636

3-oxoacyl-[acyl-carrier-protein] reductase FabG (EC 1.1.1.100) (3-ketoacyl-acyl carrier protein reductase) (Beta-Ketoacyl-acyl carrier protein reductase) (Beta-ketoacyl-ACP reductase)	fabG b1093 JW1079	6	30.832
Periplasmic trehalase (EC 3.2.1.28) (Alpha,alpha-trehalase) (Alpha,alpha-trehalose glucohydrolase) (Tre37A)	treA osmA b1197 JW1186	6	36.534
2-dehydro-3-deoxyphosphooctonate aldolase (EC 2.5.1.55) (3-deoxy-D-manno-octulosonic acid 8-phosphate synthase) (KDO-8-phosphate synthase) (KDO 8-P synthase) (KDOPS) (Phospho-2-dehydro-3-deoxyoctonate aldolase)	kdsA b1215 JW1206	6	36.684
D-lactate dehydrogenase (D-LDH) (EC 1.1.1.28) (Fermentative lactate dehydrogenase)	ldhA hslI htpH b1380 JW1375	6	26.778
Autoinducer 2-binding protein LsrB (AI-2-binding protein LsrB)	lsrB yneA b1516 JW1509	6	20.469
7alpha-hydroxysteroid dehydrogenase (7alpha-HSDH) (EC 1.1.1.159) (NAD-dependent 7alpha-hydroxysteroid dehydrogenase)	hdhA hsdH b1619 JW1611	6	32.456
Thioredoxin/glutathione peroxidase BtuE (EC 1.11.1.24) (EC 1.11.1.9)	btuE b1710 JW1700	6	74.479
ATP-dependent 6-phosphofructokinase isozyme 2 (ATP-PFK 2) (Phosphofructokinase 2) (EC 2.7.1.11) (6-phosphofructokinase isozyme II) (Phosphohexokinase 2)	pfkB b1723 JW5280	6	12.962
Uncharacterized protein YeaG	yeaG b1783 JW1772	6	22.284
Protein YebF	yebF b1847 JW1836	6	28.483
KHG/KDPG aldolase [Includes: 4-hydroxy-2-oxoglutarate aldolase (EC 4.1.3.16) (2-keto-4-hydroxyglutarate aldolase) (KHG-aldolase); 2-dehydro-3-deoxy-phosphogluconate aldolase (EC 4.1.2.14) (2-keto-3-deoxy-6-phosphogluconate aldolase) (KDPG-aldolase) (Phospho-2-dehydro-3-deoxygluconate aldolase) (Phospho-2-keto-3-deoxygluconate aldolase)]	eda hga kdgA b1850 JW1839	6	14.284
Histidine-binding periplasmic protein (HBP)	hisJ b2309 JW2306	6	21.798
Autonomous glycyl radical cofactor	grcA yfiD b2579 JW2563	6	37.36

Protein GrpE (HSP-70 cofactor) (HSP24) (Heat shock protein B25.3)	grpE b2614 JW2594	6	36.022
Glutarate 2-hydroxylase (G-2-H) (EC 1.14.11.64) (Carbon starvation induced protein D)	glaH csiD gab ygaT b2659 JW5427	6	36.094
Glycine betaine/proline betaine-binding periplasmic protein (GBBP)	proX proU b2679 JW2654	6	55.36
tRNA-modifying protein YgfZ	ygfZ yzzW b2898 JW2866	6	44.175
6-phospho-beta-glucosidase BglA (EC 3.2.1.86) (Phospho-beta-glucosidase A)	bglA bglD yqfC b2901 JW2869	6	26.429
D-3-phosphoglycerate dehydrogenase (PGDH) (EC 1.1.1.95) (2-oxoglutarate reductase) (EC 1.1.1.399)	serA b2913 JW2880	6	32.391
Uncharacterized protein YggN	yggN b2958 JW2925	6	97.349
Disulfide-bond oxidoreductase YghU (EC 1.8.4.-) (GSH-dependent disulfide-bond oxidoreductase YghU) (GST N2-2) (Organic hydroperoxidase) (EC 1.11.1.-)	yghU b2989 JW5492	6	47.543
Translation initiation factor IF-2	infB gicD ssyG b3168 JW3137	6	17.641
Phosphoglucosamine mutase (EC 5.4.2.10)	glmM mrsA yhbF b3176 JW3143	6	23.962
Transcription elongation factor GreA (Transcript cleavage factor GreA)	greA b3181 JW3148	6	16.066
Intermembrane phospholipid transport system binding protein MlaC	mlaC yrbC b3192 JW3159	6	48.772
Universal stress protein A	uspA b3495 JW3462	6	43.117
Glutathione reductase (GR) (GRase) (EC 1.8.1.7)	gor b3500 JW3467	6	23.104
2-amino-3-ketobutyrate coenzyme A ligase (AKB ligase) (EC 2.3.1.29) (Glycine acetyltransferase)	kbl b3617 JW3592	6	56.23
Thiol:disulfide interchange protein DsbA	dsbA dsf ppfA b3860 JW3832	6	35.172
Glycerol kinase (EC 2.7.1.30) (ATP:glycerol 3-phosphotransferase) (Glycerokinase) (GK)	glpK b3926 JW3897	6	27.733
Quinone oxidoreductase 1 (EC 1.6.5.5) (NADPH:quinone reductase 1) (Zeta-crystallin homolog protein)	qorA hcz qor qor1 b4051 JW4011	5	89.119
Deoxyribose-phosphate aldolase (DERA) (EC 4.1.2.4) (2-deoxy-D-ribose 5-phosphate aldolase)	deoC dra thyR b4381 JW4344	5	61.089

(Phosphodeoxyriboaldolase) (Deoxyriboaldolase)			
Bifunctional aspartokinase/homoserine dehydrogenase 1 (Aspartokinase I/homoserine dehydrogenase I) (AKI-HDI) [Includes: Aspartokinase (EC 2.7.2.4); Homoserine dehydrogenase (EC 1.1.1.3)]	thrA thrA1 thrA2 b0002 JW0001	5	9.8953
Ribulokinase (EC 2.7.1.16)	araB b0063 JW0062	5	53.951
Uncharacterized protein YahO	yahO b0329 JW0321	5	19.859
2-methylcitrate dehydratase (2-MC dehydratase) (EC 4.2.1.79) ((2S,3S)-2-methylcitrate dehydratase) (Aconitate hydratase) (ACN) (Aconitase) (EC 4.2.1.3)	prpD yahT b0334 JW0325	5	32.903
Adenine phosphoribosyltransferase (APRT) (EC 2.4.2.7)	apt b0469 JW0458	5	19.737
Glutaminase 1 (EC 3.5.1.2)	glsA1 ybaS b0485 JW0474	5	26.892
Flavodoxin 1 (Flavodoxin A)	fldA b0684 JW0671	5	38.009
GTP cyclohydrolase 1 type 2 homolog (Radiation resistance protein YbgI)	ybgI b0710 JW0700	5	40.839
Phospho-2-dehydro-3-deoxyheptonate aldolase, Phe-sensitive (EC 2.5.1.54) (3-deoxy-D-arabino-heptulosonate 7-phosphate synthase) (DAHP synthase) (Phospho-2-keto-3-deoxyheptonate aldolase)	aroG b0754 JW0737	5	33.515
Putrescine-binding periplasmic protein PotF	potF b0854 JW0838	5	34.218
3-oxoacyl-[acyl-carrier-protein] synthase 3 (EC 2.3.1.180) (3-oxoacyl-[acyl-carrier-protein] synthase III) (Beta-ketoacyl-ACP synthase III) (KAS III) (EcFabH)	fabH b1091 JW1077	5	18.597
Ribose-phosphate pyrophosphokinase (RPPK) (EC 2.7.6.1) (5-phospho-D-ribosyl alpha-1-diphosphate) (Phosphoribosyl diphosphate synthase) (Phosphoribosyl pyrophosphate synthase) (P-Rib-PP synthase) (PRPP synthase) (PRPPase)	prs prsA b1207 JW1198	5	59.899
Protein YciF	yciF b1258 JW1250	5	31.892
Periplasmic murein peptide-binding protein	mppA ynaH b1329 JW1322	5	42.849

3-hydroxy-5-phosphonooxypentane-2,4-dione thiolase (EC 2.3.1.245)	lsrF yneB b1517 JW1510	5	47.526
Mannose-6-phosphate isomerase (EC 5.3.1.8) (Phosphohexomutase) (Phosphomannose isomerase) (PMI)	manA pmi b1613 JW1605	5	30.636
Tyrosine--tRNA ligase (EC 6.1.1.1) (Tyrosyl-tRNA synthetase) (TyrRS)	tyrS b1637 JW1629	5	18.199
NH(3)-dependent NAD(+) synthetase (EC 6.3.1.5) (Nicotinamide adenine dinucleotide synthetase) (NADS) (Nitrogen regulatory protein)	nadE efg ntrL b1740 JW1729	5	43.665
Periplasmic chaperone Spy (Spheroplast protein Y)	spy b1743 JW1732	5	20.059
Succinylornithine transaminase (SOAT) (EC 2.6.1.81) (Carbon starvation protein C) (Succinylornithine aminotransferase)	astC argM cstC ydjW b1748 JW1737	5	35.54
Putative NAD(P)H nitroreductase YdjA (EC 1.-.-.-)	ydjA b1765 JW1754	5	42.965
L-arabinose-binding periplasmic protein (ABP)	araF b1901 JW1889	5	26.951
UDP-galactopyranose mutase (UGM) (EC 5.4.99.9) (UDP-GALP mutase) (Uridine 5-diphosphate galactopyranose mutase)	glf yefE b2036 JW2021	5	18.192
Uncharacterized oxidoreductase YohF (EC 1.-.-.-)	yohF yohE b2137 JW2125	5	42.613
Ecotin	eco eti b2209 JW2197	5	53.815
3-oxoacyl-[acyl-carrier-protein] synthase 1 (EC 2.3.1.41) (3-oxoacyl-[acyl-carrier-protein] synthase I) (Beta-ketoacyl-ACP synthase I) (KAS I)	fabB fabC b2323 JW2320	5	18.251
Glutamate--tRNA ligase (EC 6.1.1.17) (Glutamyl-tRNA synthetase) (GluRS)	gltX b2400 JW2395	5	60.373
PTS system glucose-specific EIIA component (EIIA-Glc) (EIII-Glc) (Glucose-specific phosphotransferase enzyme IIA component)	crr gsr iex tgs treD b2417 JW2410	5	35.56
CTP synthase (EC 6.3.4.2) (Cytidine 5'-triphosphate synthase) (Cytidine triphosphate synthetase) (CTP synthetase) (CTPS) (UTP--ammonia ligase)	pyrG b2780 JW2751	5	47.204
Glutathione synthetase (EC 6.3.2.3) (GSH synthetase) (GSH-S) (GSHase) (Glutathione synthase)	gshB gsh-II b2947 JW2914	5	40.017
Periplasmic pH-dependent serine endoprotease DegQ (EC 3.4.21.107) (Protease Do)	degQ hhoA b3234 JW3203	5	37.239

Aspartate-semialdehyde dehydrogenase (ASA dehydrogenase) (ASADH) (EC 1.2.1.11) (Aspartate-beta-semialdehyde dehydrogenase)	asd hom b3433 JW3396	5	11.806
L-threonine 3-dehydrogenase (TDH) (EC 1.1.1.103) (L-threonine dehydrogenase)	tdh b3616 JW3591	5	23.097
Thioredoxin 1 (Trx-1)	trxA fipA tsnC b3781 JW5856	5	150.63
Superoxide dismutase [Mn] (EC 1.15.1.1) (MnSOD)	sodA b3908 JW3879	5	81.259
DNA-directed RNA polymerase subunit beta (RNAP subunit beta) (EC 2.7.7.6) (RNA polymerase subunit beta) (Transcriptase subunit beta)	rpoB groN nitB rif ron stl stv tabD b3987 JW3950	5	22.216
Inducible lysine decarboxylase (LDCI) (EC 4.1.1.18)	cadA ldcI b4131 JW4092	5	20.42
Constitutive lysine decarboxylase (LDCC) (EC 4.1.1.18)	ldcC ldc ldcH b0186 JW0181	5	33.675
FKBP-type 22 kDa peptidyl-prolyl cis-trans isomerase (FKBP22) (PPIase) (EC 5.2.1.8) (Rotamase)	fklB ytfC b4207 JW5746	5	67.355
Uncharacterized protein YtfJ	ytfJ b4216 JW4175	4	28.756
Oxidoreductase YdhF (EC 1.-.-.-)	ydhF b1647 JW1639	4	31.597
50S ribosomal subunit assembly factor BipA (EC 3.6.5.-) (GTP-binding protein BipA/TypA) (Ribosome assembly factor BipA) (Ribosome-dependent GTPase BipA) (Tyrosine phosphorylated protein A)	bipA o591 typA yihK b3871 JW5571	4	20.638
4-hydroxy-tetrahydrodipicolinate reductase (HTPA reductase) (EC 1.17.1.8)	dapB b0031 JW0029	4	36.42
Pantothenate synthetase (PS) (EC 6.3.2.1) (Pantoate--beta-alanine ligase) (Pantoate-activating enzyme)	panC b0133 JW0129	4	23.586
Ribosome-recycling factor (RRF) (Ribosome-releasing factor)	frr rrf b0172 JW0167	4	23.622
1-deoxyxylulose-5-phosphate synthase YajO (EC 1.1.-.-)	yajO b0419 JW0409	4	56.176
Adenylate kinase (AK) (EC 2.7.4.3) (ATP-AMP transphosphorylase) (ATP:AMP phosphotransferase) (Adenylate monophosphate kinase)	adk dnaW plsA b0474 JW0463	4	7.4634
Thioesterase 1/protease 1/lysophospholipase L1 (TAP) (Acyl-CoA thioesterase 1) (TESA) (EC 3.1.2.2) (Acyl-CoA thioesterase I) (Arylesterase) (EC 3.1.1.2)	tesA apeA pldC b0494 JW0483	4	45.955

(Lysophospholipase L1) (EC 3.1.1.5) (Oleoyl-[acyl-carrier-protein] hydrolase) (EC 3.1.2.14) (Phospholipid degradation C) (Pldc) (Protease 1) (EC 3.4.21.-) (Protease I) (Thioesterase I/protease I) (TEP-I)			
Alkyl hydroperoxide reductase subunit F (EC 1.8.1.-) (Alkyl hydroperoxide reductase F52A protein)	ahpF b0606 JW0599	4	28.231
Cold shock-like protein CspE (CSP-E)	cspE gicA msmC b0623 JW0618	4	36.494
Tol-Pal system protein TolB	tolB b0740 JW5100	4	34.623
Cell division coordinator CpoB	cpoB ybgF b0742 JW0732	4	39.783
Low specificity L-threonine aldolase (Low specificity L-TA) (EC 4.1.2.48)	ltaE ybjU b0870 JW0854	4	32.942
Thioredoxin reductase (TRXR) (EC 1.8.1.9)	trxB b0888 JW0871	4	18.961
Phosphoserine aminotransferase (EC 2.6.1.52) (Phosphohydroxythreonine aminotransferase) (PSAT)	serC pdxC pdxF b0907 JW0890	4	77.515
UTP--glucose-1-phosphate uridylyltransferase (EC 2.7.7.9) (Alpha-D-glucosyl-1-phosphate uridylyltransferase) (UDP-glucose pyrophosphorylase) (UDPGP) (Uridine diphosphoglucose pyrophosphorylase)	galU ychD b1236 JW1224	4	22.868
Protein YciE	yciE b1257 JW1249	4	48.906
Dipeptidyl carboxypeptidase (EC 3.4.15.5) (Peptidyl-dipeptidase Dcp)	dcp b1538 JW1531	4	7.4023
Glutathione S-transferase GstA (EC 2.5.1.18) (GST B1-1)	gstA gst b1635 JW1627	4	64.682
L-serine dehydratase 1 (SDH 1) (EC 4.3.1.17) (L-serine deaminase 1) (L-SD1)	sdaA b1814 JW1803	4	19.424
Cold shock-like protein CspC (CSP-C)	cspC msmB b1823 JW1812	4	32.693
Arginine--tRNA ligase (EC 6.1.1.19) (Arginyl-tRNA synthetase) (ArgRS)	argS b1876 JW1865	4	32.829
Bacterial non-heme ferritin (EC 1.16.3.2) (Ferritin-1)	ftnA ftn gen-165 rsgA b1905 JW1893	4	43.29
Glucose-1-phosphate thymidylyltransferase 1 (G1P-TT 1) (EC 2.7.7.24) (dTDP-glucose pyrophosphorylase 1) (dTDP-glucose synthase 1)	rfbA rmlA rmlA1 b2039 JW2024	4	17.634

UTP--glucose-1-phosphate uridylyltransferase (EC 2.7.7.9) (Alpha-D-glucosyl-1-phosphate uridylyltransferase) (UDP-glucose pyrophosphorylase) (UDPGP) (Uridine diphosphoglucose pyrophosphorylase)	galF wcaN b2042 JW2027	4	15.463
Acetate kinase (EC 2.7.2.1) (Acetokinase)	ackA ack b2296 JW2293	4	40.149
Peroxiredoxin Bcp (EC 1.11.1.24) (Bacterioferritin comigratory protein) (Thioredoxin peroxidase) (Thioredoxin-dependent peroxiredoxin Bcp)	bcp b2480 JW2465	4	38.499
Nucleoside diphosphate kinase (NDK) (NDP kinase) (EC 2.7.4.6) (Nucleoside-2-P kinase)	ndk b2518 JW2502	4	51.05
Murein hydrolase activator NlpD	nlpD b2742 JW2712	4	166.71
Protein tas	tas ygdS b2834 JW2802	4	24.305
Bifunctional protein HldE [Includes: D-beta-D-heptose 7-phosphate kinase (EC 2.7.1.167) (D-beta-D-heptose 7-phosphotransferase) (D-glycero-beta-D-manno-heptose-7-phosphate kinase); D-beta-D-heptose 1-phosphate adenylyltransferase (EC 2.7.7.70) (D-glycero-beta-D-manno-heptose 1-phosphate adenylyltransferase)]	hldE rfaE waaE yqiF b3052 JW3024	4	34.723
Stringent starvation protein A	sspA pog ssp b3229 JW3198	4	20.431
Probable acrylyl-CoA reductase AcuI (EC 1.3.1.84) (Acryloyl-coenzyme A reductase AcuI)	acuI yhdH b3253 JW3222	4	24.554
Peptidyl-prolyl cis-trans isomerase A (PPIase A) (EC 5.2.1.8) (Cyclophilin A) (Rotamase A)	ppiA rot rotA b3363 JW3326	4	11.857
Ribulose-phosphate 3-epimerase (EC 5.1.3.1) (Pentose-5-phosphate 3-epimerase) (PPE) (R5P3E)	rpe dod yhfD b3386 JW3349	4	17.277
Acid stress chaperone HdeA (10K-S protein)	hdeA yhhC yhiB b3510 JW3478	4	9.1374
Protein-export protein SecB (Chaperone SecB)	secB b3609 JW3584	4	33.175
Glutaredoxin 3 (Grx3)	grxC yibM b3610 JW3585	4	20.375
UPF0701 protein YicC	yicC b3644 JW3619	4	19.093
Quinone reductase (EC 1.6.5.2) (Chromate reductase) (CHRR) (EC	chrR yieF b3713 JW3691	4	24.729

1.6.-.-) (NAD(P)H dehydrogenase (quinone))			
ATP-dependent protease subunit HslV (EC 3.4.25.2) (Heat shock protein HslV)	hslV htpO yiiC b3932 JW3903	4	17.711
50S ribosomal protein L1 (Large ribosomal subunit protein uL1)	rplA b3984 JW3947	4	9.5349
50S ribosomal protein L10 (50S ribosomal protein L8) (Large ribosomal subunit protein uL10)	rplJ b3985 JW3948	4	10.387
DNA-binding protein HU-alpha (HU-2) (NS2)	hupA b4000 JW3964	4	27.176
10 kDa chaperonin (GroES protein) (Protein Cpn10)	groS groES mopB b4142 JW4102	4	34.344
3'(2'),5'-bisphosphate nucleotidase CysQ (EC 3.1.3.7) (3'(2'),5-bisphosphonucleoside 3'(2')-phosphohydrolase) (3'-phosphoadenosine 5'-phosphate phosphatase) (PAP phosphatase) (DPNPase)	cysQ amtA b4214 JW4172	4	18.111
Galactofuranose-binding protein YtfQ	ytfQ b4227 JW4186	3	21.222
Type-1 fimbrial protein, A chain (Type-1A pilin)	fimA pilA b4314 JW4277	3	40.323
Molybdopterin adenylyltransferase (MPT adenylyltransferase) (EC 2.7.7.75)	mog chlG mogA yaaG b0009 JW0008	3	24.354
Cell division protein FtsZ	ftsZ sfiB sulB b0095 JW0093	3	18.344
5'-methylthioadenosine/S-adenosylhomocysteine nucleosidase (MTA/SAH nucleosidase) (MTAN) (EC 3.2.2.9) (5'-deoxyadenosine nucleosidase) (DOA nucleosidase) (dAdo nucleosidase) (5'-methylthioadenosine nucleosidase) (MTA nucleosidase) (S-adenosylhomocysteine nucleosidase) (AdoHcy nucleosidase) (SAH nucleosidase) (SRH nucleosidase)	mtnN mtn pfs yadA b0159 JW0155	3	23.186
UPF0234 protein YajQ	yajQ b0426 JW5058	3	87.437
ATP-dependent Clp protease proteolytic subunit (EC 3.4.21.92) (Caseinolytic protease) (Endopeptidase Clp) (Heat shock protein F21.5) (Protease Ti)	clpP lopP b0437 JW0427	3	12.015
Lon protease (EC 3.4.21.53) (ATP-dependent protease La)	lon capR deg lopA muc b0439 JW0429	3	31.791
Nucleoid-associated protein YbaB	ybaB b0471 JW0460	3	29.774

Chaperedoxin (Heat shock protein CnoX) (Trxsc)	cnoX ybbN b0492 JW5067	3	64.421
Glucosamine-6-phosphate deaminase (EC 3.5.99.6) (GlcN6P deaminase) (GNPDA) (Glucosamine-6-phosphate isomerase)	nagB glmD b0678 JW0664	3	12.872
Succinate dehydrogenase flavoprotein subunit (EC 1.3.5.1)	sdhA b0723 JW0713	3	76.225
Uncharacterized protein YbgS	ybgS b0753 JW0736	3	18.602
UvrABC system protein B (Protein UvrB) (Excinuclease ABC subunit B)	uvrB b0779 JW0762	3	33.325
Outer membrane protein X	ompX ybiG b0814 JW0799	3	44.067
Probable L,D-transpeptidase YbiS (EC 2.-.-.-)	ybiS b0819 JW0803	3	56.47
Molybdopterin molybdenumtransferase (MPT Mo-transferase) (EC 2.10.1.1)	moeA bisB chlE narE b0827 JW0811	3	23.713
Glutathione-binding protein GsiB	gsiB yliB b0830 JW5111	3	26.829
Glutathione S-transferase GstB (EC 2.5.1.18)	gstB yliJ b0838 JW0822	3	14.701
ABC transporter arginine-binding protein 1	artJ b0860 JW0844	3	8.5244
Uncharacterized protein YccU	yccU b0965 JW5130	3	20.845
Uncharacterized protein YccJ	yccJ b1003 JW0988	3	8.6394
NAD(P)H dehydrogenase (quinone) (EC 1.6.5.2) (Flavoprotein WrbA) (NAD(P)H:quinone oxidoreductase) (NQO)	wrbA b1004 JW0989	3	10.235
Acyl carrier protein (ACP) (Cytosolic-activating factor) (CAF) (Fatty acid synthase acyl carrier protein)	acpP b1094 JW1080	3	29.614
Cell division topological specificity factor	minE b1174 JW1163	3	57.641
Septum site-determining protein MinD (Cell division inhibitor MinD)	minD b1175 JW1164	3	18.321
Probable D,D-dipeptide-binding periplasmic protein DdpA	ddpA yddS b1487 JW5240	3	27.249
Protein YdeJ	ydeJ b1537 JW1530	3	60.298
NADP-dependent 3-hydroxy acid dehydrogenase YdfG (L-allo-threonine dehydrogenase) (EC 1.1.1.381) (Malonic semialdehyde reductase) (EC 1.1.1.298)	ydfG b1539 JW1532	3	11.354
Fumarate hydratase class I, aerobic (EC 4.2.1.2) (Fumarase A) (Oxaloacetate keto--enol-isomerase)	fumA b1612 JW1604	3	53.026

(OAAKE isomerase) (Oxaloacetate tautomerase) (EC 5.3.2.2)			
Fumarate hydratase class I, anaerobic (EC 4.2.1.2) (D-tartrate dehydratase) (EC 4.2.1.81) (Fumarase B)	fumB b4122 JW4083	3	32.666
Integration host factor subunit alpha (IHF-alpha)	ihfA hid himA b1712 JW1702	3	18.121
N-succinylglutamate 5-semialdehyde dehydrogenase (EC 1.2.1.71) (Succinylglutamic semialdehyde dehydrogenase) (SGSD)	astD ydjU b1746 JW5282	3	12.378
Putative glucose-6-phosphate 1-epimerase (EC 5.1.3.15) (Putative D-hexose-6-phosphate mutarotase) (Unknown protein from 2D-page spots T26/PR37)	yeaD yzzQ b1780 JW1769	3	35.153
Free methionine-R-sulfoxide reductase (fRMsr) (EC 1.8.4.14)	msrC yebR b1832 JW1821	3	18.081
Uncharacterized protein YebY	yebY b1839 JW1828	3	29.68
D-cysteine desulfhydrase (EC 4.4.1.15)	dcyD yedO b1919 JW5313	3	40.558
DNA gyrase inhibitor	sbmC gyrI yeeB b2009 JW1991	3	85.774
Protein YeeZ	yeeZ b2016 JW1998	3	11.306
dTDP-glucose 4,6-dehydratase 1 (EC 4.2.1.46)	rfbB rmlB b2041 JW2026	3	41.367
Ribonucleoside-diphosphate reductase 1 subunit alpha (EC 1.17.4.1) (Protein B1) (Ribonucleoside-diphosphate reductase 1 R1 subunit) (Ribonucleotide reductase 1)	nrdA dnaF b2234 JW2228	3	9.1193
Protein ElaB	elaB yfbD b2266 JW2261	3	31.27
Erythronate-4-phosphate dehydrogenase (EC 1.1.1.290)	pdxB b2320 JW2317	3	52.022
Phosphocarrier protein HPr (Histidine-containing protein)	ptsH hpr b2415 JW2408	3	12.425
4-hydroxy-tetrahydrodipicolinate synthase (HTPA synthase) (EC 4.3.3.7)	dapA b2478 JW2463	3	22.86
Inosine-5'-monophosphate dehydrogenase (IMP dehydrogenase) (IMPD) (IMPDH) (EC 1.1.1.205)	guaB guaR b2508 JW5401	3	36.85
Nitrogen regulatory protein P-II 1	glnB b2553 JW2537	3	70.531
Ribose-5-phosphate isomerase A (EC 5.3.1.6) (Phosphoriboisomerase A) (PRI)	rpiA ygfC b2914 JW5475	3	42.097

L-asparaginase 2 (EC 3.5.1.1) (L-asparaginase II) (L-ASNase II) (L-asparagine amidohydrolase II) (Colaspase)	ansB b2957 JW2924	3	37.386
Bifunctional glutathionylspermidine synthetase/amidase (GspSA) [Includes: Glutathionylspermidine amidase (Gsp amidase) (EC 3.5.1.78) (Glutathionylspermidine amidohydrolase [spermidine-forming]); Glutathionylspermidine synthetase (Gsp synthetase) (EC 6.3.1.8) (Glutathione:spermidine ligase [ADP-forming]) (Gsp synthase)]	gss gsp b2988 JW2956	3	10.75
Alcohol dehydrogenase YqhD (EC 1.1.1.-)	yqhD b3011 JW2978	3	49.32
Glutathionyl-hydroquinone reductase YqjG (GS-HQR) (EC 1.8.5.7)	yqjG b3102 JW3073	3	13.099
Ribosome hibernation promoting factor (HPF) (Hibernation factor HPF)	hpf yhbH b3203 JW3170	3	20.301
Biotin carboxylase (EC 6.3.4.14) (Acetyl-CoA carboxylase subunit A) (ACC) (EC 6.4.1.2)	accC fabG b3256 JW3224	3	20.853
30S ribosomal protein S13 (Small ribosomal subunit protein uS13)	rpsM b3298 JW3260	3	20.997
50S ribosomal protein L5 (Large ribosomal subunit protein uL5)	rplE b3308 JW3270	3	12.043
FKBP-type peptidyl-prolyl cis-trans isomerase SlyD (PPIase) (EC 5.2.1.8) (Histidine-rich protein) (Metallochaperone SlyD) (Rotamase) (Sensitivity to lysis protein D) (WHP)	slyD b3349 JW3311	3	22.545
Fe/S biogenesis protein NfuA	nfuA gntY yhgI b3414 JW3377	3	37.023
Acid stress chaperone HdeB (10K-L protein)	hdeB yhhD yhiC b3509 JW5669	3	9.6347
Uncharacterized GST-like protein YibF	yibF b3592 JW3565	3	38.712
Phosphate-binding protein PstS (PBP)	pstS phoS b3728 JW3706	3	12.295
Cell division protein ZapB	zapB yiiU b3928 JW3899	3	155.16
Glycerol dehydrogenase (GDH) (GLDH) (EC 1.1.1.6)	gldA b3945 JW5556	3	43.387
50S ribosomal protein L7/L12 (L8) (Large ribosomal subunit protein bL12)	rplL b3986 JW3949	3	13.519
DNA-directed RNA polymerase subunit beta' (RNAP subunit beta') (EC 2.7.7.6) (RNA polymerase subunit beta') (Transcriptase subunit beta')	rpoC tabB b3988 JW3951	3	15.769

Maltose/maltodextrin-binding periplasmic protein (MMBP) (Maltodextrin-binding protein) (Maltose-binding protein) (MBP)	malE b4034 JW3994	3	36.833
Uncharacterized protein YjbR	yjbR b4057 JW4018	3	13.611
50S ribosomal protein L9 (Large ribosomal subunit protein bL9)	rplI b4203 JW4161	3	36.501
Fructose-1,6-bisphosphatase class 1 (FBPase class 1) (EC 3.1.3.11) (D-fructose-1,6-bisphosphate 1-phosphohydrolase class 1)	fbp fdp b4232 JW4191	3	27.292
2-iminobutanoate/2-iminopropanoate deaminase (EC 3.5.99.10) (Enamine/imine deaminase)	ridA yjgF b4243 JW5755	3	34.093
Aldehyde reductase Ahr (EC 1.1.1.2) (Zinc-dependent alcohol dehydrogenase Ahr)	ahr yjgB b4269 JW5761	2	109.77
Aerobic respiration control protein ArcA (Dye resistance protein)	arcA cpxC dye fexA msp seg sfrA b4401 JW4364	2	47.447
Branched-chain-amino-acid aminotransferase (BCAT) (EC 2.6.1.42) (Transaminase B)	ilvE b3770 JW5606	2	32.134
RNA polymerase-associated protein RapA (EC 3.6.4.-) (ATP-dependent helicase HepA)	rapA hepA yabA b0059 JW0058	2	39.315
UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligase (EC 6.3.2.10) (D-alanyl-D-alanine-adding enzyme) (UDP-MurNAc-pentapeptide synthetase)	murF mra b0086 JW0084	2	7.2811
2-methylisocitrate lyase (2-MIC) (MICL) (EC 4.1.3.30) ((2R,3S)-2-methylisocitrate lyase)	prpB yahQ b0331 JW0323	2	46.355
D-alanine--D-alanine ligase A (EC 6.3.2.4) (D-Ala-D-Ala ligase A) (D-alanylalanine synthetase A)	ddlA b0381 JW0372	2	19.431
Uncharacterized protein YaiA	yaiA b0389 JW0380	2	87.872
ATP-dependent Clp protease ATP-binding subunit ClpX (ATP-dependent unfoldase ClpX)	clpX lopC b0438 JW0428	2	18.153
Uncharacterized lipoprotein YbaY	ybaY b0453 JW0443	2	19.476
Copper-exporting P-type ATPase (EC 7.2.2.8) (Copper-exporting P-type ATPase A) (Cu(+)-exporting ATPase) (Soluble copper chaperone CopA(Z))	copA atcU f834 ybaR b0484 JW0473	2	18.797

Peptidyl-prolyl cis-trans isomerase B (PPIase B) (EC 5.2.1.8) (Rotamase B)	ppiB b0525 JW0514	2	33.823
UPF0098 protein YbcL	ybcL b0545 JW0533	2	27.163
Uncharacterized protein YbeL	ybeL b0643 JW0638	2	16.795
Pyrimidine-specific ribonucleoside hydrolase RihA (EC 3.2.-.-) (Cytidine/uridine-specific hydrolase)	rihA ybeK b0651 JW0646	2	26.77
Ribonucleotide monophosphatase NagD (EC 3.1.3.5)	nagD b0675 JW0661	2	17.085
Ferric uptake regulation protein (Ferric uptake regulator)	fur b0683 JW0669	2	59.857
Succinate dehydrogenase iron-sulfur subunit (EC 1.3.5.1)	sdhB b0724 JW0714	2	18.969
UPF0098 protein YbhB	ybhB b0773 JW0756	2	7.7806
Probable ATP-binding protein YbiT	ybiT b0820 JW0804	2	35.343
3-hydroxydecanoyl-[acyl-carrier-protein] dehydratase (EC 4.2.1.59) (3-hydroxyacyl-[acyl-carrier-protein] dehydratase FabA) (Beta-hydroxydecanoyl thioester dehydrase) (Trans-2-decenoyl-[acyl-carrier-protein] isomerase) (EC 5.3.3.14)	fabA b0954 JW0937	2	32.417
Cold shock-like protein CspG (CPS-G)	cspG cspI b0990 JW0974	2	43.045
Glyoxylate/hydroxypyruvate reductase A (EC 1.1.1.79) (EC 1.1.1.81) (2-ketoacid reductase)	ghrA ycdW b1033 JW5146	2	21.226
Malonyl CoA-acyl carrier protein transacylase (MCT) (EC 2.3.1.39)	fabD tfpA b1092 JW1078	2	38.844
3-oxoacyl-[acyl-carrier-protein] synthase 2 (EC 2.3.1.179) (3-oxoacyl-[acyl-carrier-protein] synthase II) (Beta-ketoacyl-ACP synthase II) (KAS II)	fabF fabJ b1095 JW1081	2	15.539
UPF0227 protein YcfP	ycfP b1108 JW5158	2	25.493
Alanine racemase, catabolic (EC 5.1.1.1)	dadX alnB dadB b1190 JW1179	2	11.475
DNA-binding protein H-NS (Heat-stable nucleoid-structuring protein) (Histone-like protein HLP-II) (Protein B1) (Protein H1)	hns bglY cur drdX hnsA msyA osmZ pilG topS b1237 JW1225	2	10.136
Phage shock protein A	pspA b1304 JW1297	2	16.016
Thiosulfate sulfurtransferase PspE (TST) (EC 2.8.1.1) (Phage shock protein E)	pspE b1308 JW1301	2	29.705

Universal stress protein F	uspF ynaF yzzL b1376 JW1370	2	29.006
Protein YdcF	ydcF b1414 JW1411	2	10.477
Trans-aconitate 2-methyltransferase (EC 2.1.1.144)	tam yneD b1519 JW1512	2	12.879
Uncharacterized protein YnfD	ynfD b1586 JW5259	2	23.445
Glutaredoxin 4 (Grx4) (Monothiol glutaredoxin)	grxD ydhD b1654 JW1646	2	27.582
Riboflavin synthase (RS) (EC 2.5.1.9)	ribC ribE b1662 JW1654	2	54.745
Probable ATP-dependent transporter SufC	sufC ynhD b1682 JW1672	2	10.865
FeS cluster assembly protein SufB	sufB ynhE b1683 JW5273	2	32.458
Uncharacterized protein YdiZ	ydiZ b1724 JW1713	2	12.021
Probable ketoamine kinase YniA (EC 2.7.1.-)	yniA b1725 JW1714	2	51.294
Osmotically-inducible putative lipoprotein OsmE (Activator of ntr-like gene protein)	osmE anr b1739 JW1728	2	8.5791
Flagellin	fliC flaF hag b1923 JW1908	2	30.439
Uncharacterized protein YodD	yodD b1953 JW5317	2	53.994
Mannosyl-3-phosphoglycerate phosphatase (MPGP) (EC 3.1.3.70)	yedP b1955 JW1938	2	12.466
AMP nucleosidase (EC 3.2.2.4)	amn b1982 JW1963	2	31.54
Nickel/cobalt homeostasis protein RcnB	rcnB yohN b2107 JW5346	2	43.517
Cytidine deaminase (EC 3.5.4.5) (Cytidine aminohydrolase) (CDA)	cdd b2143 JW2131	2	58.958
Ribonucleoside-diphosphate reductase 1 subunit beta (EC 1.17.4.1) (Protein B2) (Protein R2) (Ribonucleotide reductase 1)	nrdB ftsB b2235 JW2229	2	44.225
Anaerobic glycerol-3-phosphate dehydrogenase subunit A (G-3-P dehydrogenase) (EC 1.1.5.3)	glpA b2241 JW2235	2	31.633
NMN amidohydrolase-like protein YfaY	yfaY b2249 JW2243	2	26.995
1,4-dihydroxy-2-naphthoyl-CoA synthase (DHNA-CoA synthase) (EC 4.1.3.36)	menB b2262 JW2257	2	22.533
Phosphoribosylaminoimidazole-succinocarboxamide synthase (EC 6.3.2.6) (SAICAR synthetase)	purC b2476 JW2461	2	35.749

Uracil phosphoribosyltransferase (EC 2.4.2.9) (UMP pyrophosphorylase) (UPRTase)	upp uraP b2498 JW2483	2	12.288
Sigma-E factor regulatory protein RseB	rseB b2571 JW2555	2	19.416
S-ribosylhomocysteine lyase (EC 4.4.1.21) (AI-2 synthesis protein) (Autoinducer-2 production protein LuxS)	luxS ygaG b2687 JW2662	2	66.269
Sulfite reductase [NADPH] flavoprotein alpha-component (SiR-FP) (EC 1.8.1.2)	cysJ b2764 JW2734	2	50.971
Pyrimidine/purine nucleotide 5'-monophosphate nucleosidase (EC 3.2.2.-) (EC 3.2.2.10) (AMP nucleosidase) (EC 3.2.2.4) (CMP nucleosidase) (GMP nucleosidase) (IMP nucleosidase) (UMP nucleosidase) (dTMP nucleosidase)	ppnN ygdH b2795 JW2766	2	41.25
Peptide chain release factor RF2 (RF-2)	prfB supK b2891 JW5847	2	13.811
Glycine cleavage system H protein	gcvH b2904 JW2872	2	73.898
Biosynthetic arginine decarboxylase (ADC) (EC 4.1.1.19)	speA b2938 JW2905	2	41.951
S-adenosylmethionine synthase (AdoMet synthase) (EC 2.5.1.6) (MAT) (Methionine adenosyltransferase)	metK metX b2942 JW2909	2	13.737
Protein GlcG	glcG yghC b2977 JW2944	2	11.051
Uncharacterized protein YqjD	yqjD b3098 JW3069	2	54.87
Transcription termination/antitermination protein NusA (N utilization substance protein A) (Transcription termination/antitermination L factor)	nusA b3169 JW3138	2	44.817
UDP-N-acetylglucosamine 1-carboxyvinyltransferase (EC 2.5.1.7) (Enoylpyruvate transferase) (UDP-N-acetylglucosamine enolpyruvyl transferase) (EPT)	murA murZ b3189 JW3156	2	35.196
Arabinose 5-phosphate isomerase KdsD (API) (L-API) (EC 5.3.1.13)	kdsD yrbH b3197 JW3164	2	20.127
Lipopolysaccharide export system protein LptA	lptA yhbN b3200 JW3167	2	22.981
Glyoxalase ElbB (EC 4.2.1.-) (Sigma cross-reacting protein 27A) (SCRP-27A)	elbB elb2 yzzB b3209 JW3176	2	52.015
Glutamate synthase [NADPH] small chain (EC 1.4.1.13) (Glutamate	gltD aspB b3213 JW3180	2	9.1963

synthase subunit beta) (GLTS beta chain) (NADPH-GOGAT)			
Uncharacterized protein YhcN	yhcN b3238 JW5540	2	36.952
Cell shape-determining protein MreB (Actin-like MreB protein) (Rod shape-determining protein MreB)	mreB envB rodY b3251 JW3220	2	17.603
30S ribosomal protein S5 (Small ribosomal subunit protein uS5)	rpsE spc b3303 JW3265	2	25.983
30S ribosomal protein S3 (Small ribosomal subunit protein uS3)	rpsC b3314 JW3276	2	29.86
50S ribosomal protein L2 (Large ribosomal subunit protein uL2)	rplB b3317 JW3279	2	27.353
DNA-binding dual transcriptional regulator OmpR (Transcriptional regulatory protein OmpR)	ompR kmt ompB b3405 JW3368	2	48.697
Glucose-1-phosphate adenylyltransferase (EC 2.7.7.27) (ADP-glucose pyrophosphorylase) (ADPGlc PPase) (ADP-glucose synthase)	glgC b3430 JW3393	2	38.765
Uncharacterized oxidoreductase YhhX (EC 1.-.-.-)	yhhX b3440 JW3403	2	61.767
Glutathione hydrolase proenzyme (EC 3.4.19.13) (Gamma-glutamyltranspeptidase proenzyme) (GGT) (EC 2.3.2.2) [Cleaved into: Glutathione hydrolase large chain; Glutathione hydrolase small chain]	ggt b3447 JW3412	2	16.624
Uncharacterized protein YhhA (ORFQ)	yhhA b3448 JW3413	2	35.395
Glyoxylate/hydroxypyruvate reductase B (EC 1.1.1.79) (EC 1.1.1.81) (2-ketoaldonate reductase) (2-ketogluconate reductase) (2KR) (EC 1.1.1.215)	ghrB tkrA yiaE b3553 JW5656	2	36.361
Glycerol-3-phosphate dehydrogenase [NAD(P)+] (EC 1.1.1.94) (NAD(P)H-dependent glycerol-3-phosphate dehydrogenase)	gpsA b3608 JW3583	2	56.193
2,3-bisphosphoglycerate-independent phosphoglycerate mutase (BPG-independent PGAM) (Phosphoglyceromutase) (iPGM) (EC 5.4.2.12)	gpmI pgmI yibO b3612 JW3587	2	16.155
Deoxyuridine 5'-triphosphate nucleotidohydrolase (dUTPase) (EC 3.6.1.23) (dUTP pyrophosphatase)	dut dnaS sof b3640 JW3615	2	31.577
ATP synthase gamma chain (ATP synthase F1 sector gamma subunit) (F-ATPase gamma subunit)	atpG papC uncG b3733 JW3711	2	84.672

5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase (EC 2.1.1.14) (Cobalamin-independent methionine synthase) (Methionine synthase, vitamin-B12 independent isozyme)	metE b3829 JW3805	2	79.593
Fatty acid oxidation complex subunit alpha [Includes: Enoyl-CoA hydratase/Delta(3)-cis-Delta(2)-trans-enoyl-CoA isomerase/3-hydroxybutyryl-CoA epimerase (EC 4.2.1.17) (EC 5.1.2.3) (EC 5.3.3.8); 3-hydroxyacyl-CoA dehydrogenase (EC 1.1.1.35)]	fadB oldB b3846 JW3822	2	10.273
Protein YihD	yihD b3858 JW3830	2	16.293
Universal stress protein D	uspD yiiT b3923 JW3894	2	8.3252
UPF0337 protein YjbJ	yjbJ b4045 JW4005	2	18.975
Single-stranded DNA-binding protein (SSB) (Helix-destabilizing protein)	ssb exrB lexC b4059 JW4020	2	16.171
Protein YjdN	yjdN phnB b4107 JW4068	2	20.591
Elongation factor P (EF-P)	efp b4147 JW4107	2	54.332
Uncharacterized protein YjgR	yjgR b4263 JW4220	2	17.352
Uncharacterized protein YjjA (Protein P-18)	yjjA b4360 JW5795	2	73.352
Soluble lytic murein transglycosylase (EC 4.2.2.n1) (Exomuramidase) (Peptidoglycan lytic exotransglycosylase) (Slt70)	slt sltY b4392 JW4355	2	77.1
Periplasmic protein CpxP (ORF_o167) (Periplasmic accessory protein CpxP)	cpxP yiiO b4484 JW5558	2	18.965

Appendix D Table 2: predicted N-glycoprotein candidates

Rank	UniProt Entry name	Protein name	MaxQuant unique peptides	NetNGlyc glycosylation site number	Bacterial sequon match
1	P10384	Long-chain fatty acid transport protein (Outer membrane FadL protein) (Outer membrane flp protein)	1	7	2
2	P0AFK9	Spermidine/putrescine-binding periplasmic protein (SPBP)	1	6	2
3	P23843	Periplasmic oligopeptide-binding protein	2	6	1
4	P0A8V2	DNA-directed RNA polymerase subunit beta (RNAP subunit beta) (EC 2.7.7.6) (RNA polymerase subunit beta) (Transcriptase subunit beta)	1	6	1
5	P23538	Phosphoenolpyruvate synthase (PEP synthase) (EC 2.7.9.2) (Pyruvate, water dikinase)	3	5	1
6	P0A6Y8	Chaperone protein DnaK (HSP70) (Heat shock 70 kDa protein) (Heat shock protein 70)	2	5	1
7	P52697	6-phosphogluconolactonase (6-P-gluconolactonase) (Pgl) (EC 3.1.1.31)	1	5	1
8	P37636	Multidrug resistance protein MdtE	2	3	1
9	P0AFG6	Dihydrolipoyllysine-residue succinyltransferase component of 2-oxoglutarate dehydrogenase complex (EC 2.3.1.61) (2-oxoglutarate dehydrogenase complex component E2) (OGDC-E2) (Dihydrolipoamide succinyltransferase component of 2-oxoglutarate dehydrogenase complex)	1	3	1

10	P07102	Periplasmic AppA protein [Includes: Phosphoanhydride phosphohydrolase (EC 3.1.3.2) (pH 2.5 acid phosphatase) (AP); 4-phytase (EC 3.1.3.26)]	1	3	1
11	P26616	NAD-dependent malic enzyme (NAD-ME) (EC 1.1.1.38)	1	3	1
12	P0C8J8	D-tagatose-1,6-bisphosphate aldolase subunit GatZ	1	3	1
13	P0AAI3	ATP-dependent zinc metalloprotease FtsH (EC 3.4.24.-) (Cell division protease FtsH)	1	3	1
14	P0AG80	sn-glycerol-3-phosphate-binding periplasmic protein UgpB	1	3	1
15	P0A6M8	Elongation factor G (EF-G)	2	2	1
16	P0ABB0	ATP synthase subunit alpha (EC 7.1.2.2) (ATP synthase F1 sector subunit alpha) (F-ATPase subunit alpha)	2	2	1
17	P69797	PTS system mannose-specific EIIAB component (EC 2.7.1.191) (EIIAB-Man) (EIII-Man) [Includes: Mannose-specific phosphotransferase enzyme IIA component (PTS system mannose-specific EIIA component); Mannose-specific phosphotransferase enzyme IIB component (PTS system mannose-specific EIIB component)]	1	2	1
18	P05055	Polyribonucleotide nucleotidyltransferase (EC 2.7.7.8) (Polynucleotide phosphorylase) (PNPase)	1	2	1
19	P0A7V3	30S ribosomal protein S3 (Small ribosomal subunit protein uS3)	1	2	1
20	P0AG86	Protein-export protein SecB (Chaperone SecB)	1	1	1
21	P00509	Aspartate aminotransferase (AspAT) (EC 2.6.1.1) (Transaminase A)	1	6	0

22	P02931	Outer membrane porin F (Outer membrane protein 1A) (Outer membrane protein B) (Outer membrane protein F) (Outer membrane protein IA) (Porin OmpF)	1	6	0
23	P00452	Ribonucleoside-diphosphate reductase 1 subunit alpha (EC 1.17.4.1) (Protein B1) (Ribonucleoside-diphosphate reductase 1 R1 subunit) (Ribonucleotide reductase 1)	1	6	0
24	P25516	Aconitate hydratase A (ACN) (Aconitase) (EC 4.2.1.3) (Iron-responsive protein-like) (IRP-like) (RNA-binding protein) (Stationary phase enzyme)	3	5	0
25	P06996	Outer membrane porin C (Outer membrane protein 1B) (Outer membrane protein C) (Porin OmpC)	2	4	0
26	P13029	Catalase-peroxidase (CP) (EC 1.11.1.21) (Hydroperoxidase I) (HPI) (Peroxidase/catalase)	2	4	0
27	P0C0V0	Periplasmic serine endoprotease DegP (EC 3.4.21.107) (Heat shock protein DegP) (Protease Do)	1	4	0
28	P13482	Periplasmic trehalase (EC 3.2.1.28) (Alpha,alpha-trehalase) (Alpha,alpha-trehalose glucohydrolase) (Tre37A)	1	4	0
29	P16700	Thiosulfate-binding protein	1	4	0
30	P77717	Uncharacterized lipoprotein YbaY	3	3	0
31	P09373	Formate acetyltransferase 1 (EC 2.3.1.54) (Pyruvate formate-lyase 1)	2	3	0
32	P76193	Probable L,D-transpeptidase YnhG (EC 2.-.-.-)	2	3	0
33	P0A9B2	Glyceraldehyde-3-phosphate dehydrogenase A (GAPDH-A) (EC 1.2.1.12) (NAD-dependent glyceraldehyde-3-phosphate dehydrogenase)	2	3	0

34	P33570	Transketolase 2 (TK 2) (EC 2.2.1.1)	2	3	0
35	P0A6Z3	Chaperone protein HtpG (Heat shock protein C62.5) (Heat shock protein HtpG) (High temperature protein G)	1	3	0
36	P0ABJ9	Cytochrome bd-I ubiquinol oxidase subunit 1 (EC 7.1.1.7) (Cytochrome bd-I oxidase subunit I) (Cytochrome d ubiquinol oxidase subunit I)	1	3	0
37	P0A953	3-oxoacyl-[acyl-carrier-protein] synthase 1 (EC 2.3.1.41) (3-oxoacyl-[acyl-carrier-protein] synthase I) (Beta-ketoacyl-ACP synthase I) (KAS I)	1	3	0
38	P02930	Outer membrane protein TolC (Multidrug efflux pump subunit TolC) (Outer membrane factor TolC)	1	3	0
39	P27302	Transketolase 1 (TK 1) (EC 2.2.1.1)	1	3	0
40	P69908	Glutamate decarboxylase alpha (GAD-alpha) (EC 4.1.1.15)	6	2	0
41	P08997	Malate synthase A (MSA) (EC 2.3.3.9)	6	2	0
42	P0A9G6	Isocitrate lyase (ICL) (EC 4.1.3.1) (Isocitrase) (Isocitratase)	5	2	0
43	P0A9Q7	Aldehyde-alcohol dehydrogenase [Includes: Alcohol dehydrogenase (ADH) (EC 1.1.1.1); Acetaldehyde dehydrogenase [acetylating] (ACDH) (EC 1.2.1.10); Pyruvate-formate-lyase deactivase (PFL deactivase)]	2	2	0
44	P25553	Lactaldehyde dehydrogenase (EC 1.2.1.22) (Aldehyde dehydrogenase A) (Glycolaldehyde dehydrogenase) (EC 1.2.1.21)	2	2	0

45	P76177	Protein YdgH	2	2	0
46	P04079	GMP synthase [glutamine-hydrolyzing] (EC 6.3.5.2) (GMP synthetase) (GMPS) (Glutamine amidotransferase)	2	2	0
47	P39099	Periplasmic pH-dependent serine endoprotease DegQ (EC 3.4.21.107) (Protease Do)	2	2	0
48	P07024	Protein UshA [Includes: UDP-sugar hydrolase (EC 3.6.1.45) (UDP-sugar diphosphatase) (UDP-sugar pyrophosphatase); 5'-nucleotidase (5'-NT) (EC 3.1.3.5)]	1	2	0
49	P19926	Glucose-1-phosphatase (G1Pase) (EC 3.1.3.10)	1	2	0
50	P0AFL3	Peptidyl-prolyl cis-trans isomerase A (PPIase A) (EC 5.2.1.8) (Cyclophilin A) (Rotamase A)	1	2	0
51	P37194	Outer membrane protein Slp	1	2	0
52	P23847	Dipeptide-binding protein (DBP) (Periplasmic dipeptide transport protein)	1	2	0
53	P0CE48	Elongation factor Tu 2 (EF-Tu 2) (Bacteriophage Q beta RNA-directed RNA polymerase subunit III) (P-43)	4	1	0
54	P63284	Chaperone protein ClpB (Heat shock protein F84.1)	6	1	0
55	P04128	Type-1 fimbrial protein, A chain (Type-1A pilin)	3	1	0
56	P0A9D8	2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase (EC 2.3.1.117) (Succinyl-CoA: tetrahydrodipicolinate N-succinyltransferase) (Tetrahydrodipicolinate N-succinyltransferase) (THDP succinyltransferase) (THP succinyltransferase) (Tetrahydropicolinate succinylase)	1	1	0
57	P45955	Cell division coordinator CpoB	1	1	0

58	P61316	Outer-membrane lipoprotein carrier protein (P20)	1	1	0
59	P0A6A8	Acyl carrier protein (ACP) (Cytosolic-activating factor) (CAF) (Fatty acid synthase acyl carrier protein)	1	1	0
60	P37903	Universal stress protein F	1	1	0
61	P0AD61	Pyruvate kinase I (EC 2.7.1.40) (PK-1)	1	1	0
62	P09551	Lysine/arginine/ornithine-binding periplasmic protein (LAO-binding protein)	1	1	0
63	P0ADG7	Inosine-5'-monophosphate dehydrogenase (IMP dehydrogenase) (IMPD) (IMPDH) (EC 1.1.1.205)	1	1	0
64	P0A763	Nucleoside diphosphate kinase (NDK) (NDP kinase) (EC 2.7.4.6) (Nucleoside-2-P kinase)	1	1	0
65	P0AFF6	Transcription termination/antitermination protein NusA (N utilization substance protein A) (Transcription termination/antitermination L factor)	1	1	0
66	P0AFX0	Ribosome hibernation promoting factor (HPF) (Hibernation factor HPF)	1	1	0
67	P61889	Malate dehydrogenase (EC 1.1.1.37)	1	1	0
68	P0AG44	50S ribosomal protein L17 (Large ribosomal subunit protein bL17)	1	1	0
69	P0A7W1	30S ribosomal protein S5 (Small ribosomal subunit protein uS5)	1	1	0
70	P0ADY3	50S ribosomal protein L14 (Large ribosomal subunit protein uL14)	1	1	0
71	P60438	50S ribosomal protein L3 (Large ribosomal subunit protein uL3)	1	1	0
72	P45523	FKBP-type peptidyl-prolyl cis-trans isomerase FkpA (PPIase) (EC 5.2.1.8) (Rotamase)	1	1	0

73	P0A6Y5	33 kDa chaperonin (Heat shock protein 33) (HSP33)	1	1	0
74	P0ADX7	Uncharacterized protein YhhA (ORFQ)	1	1	0
75	P12758	Uridine phosphorylase (UPase) (UrdPase) (EC 2.4.2.3)	1	1	0
76	P00448	Superoxide dismutase [Mn] (EC 1.15.1.1) (MnSOD)	1	1	0
77	P0A9L3	FKBP-type 22 kDa peptidyl-prolyl cis-trans isomerase (FKBP22) (PPIase) (EC 5.2.1.8) (Rotamase)	1	1	0
78	P36683	Aconitate hydratase B (ACN) (Aconitase) (EC 4.2.1.3) ((2R,3S)-2-methylisocitrate dehydratase) ((2S,3R)-3-hydroxybutane-1,2,3-tricarboxylate dehydratase) (2-methyl-cis-aconitate hydratase) (EC 4.2.1.99) (Iron-responsive protein-like) (IRP-like) (RNA-binding protein)	2	0	2
79	P0A7V0	30S ribosomal protein S2 (Small ribosomal subunit protein uS2)	2	0	2
80	P0A836	Succinate--CoA ligase [ADP-forming] subunit beta (EC 6.2.1.5) (Succinyl-CoA synthetase subunit beta) (SCS-beta)	2	0	1
81	P0ABT2	DNA protection during starvation protein (EC 1.16.-.-)	2	0	1
82	P08200	Isocitrate dehydrogenase [NADP] (IDH) (EC 1.1.1.42) (IDP) (NADP(+)-specific ICDH) (Oxalosuccinate decarboxylase)	2	0	1
83	P31130	Uncharacterized protein YdeI	2	0	1
84	P0ABB4	ATP synthase subunit beta (EC 7.1.2.2) (ATP synthase F1 sector subunit beta) (F-ATPase subunit beta)	2	0	1
85	P0A6F5	Chaperonin GroEL (EC 5.6.1.7) (60 kDa chaperonin) (Chaperonin-60) (Cpn60) (GroEL protein)	2	0	1

86	P0A9P0	Dihydrolipoyl dehydrogenase (EC 1.8.1.4) (Dihydrolipoamide dehydrogenase) (E3 component of pyruvate and 2-oxoglutarate dehydrogenases complexes) (Glycine cleavage system L protein)	1	0	1
87	P0A6P1	Elongation factor Ts (EF-Ts) (Bacteriophage Q beta RNA-directed RNA polymerase subunit IV)	1	0	1
88	P0A850	Trigger factor (TF) (EC 5.2.1.8) (PPIase)	1	0	1
89	P69503	Adenine phosphoribosyltransferase (APRT) (EC 2.4.2.7)	1	0	1
90	P0AE08	Alkyl hydroperoxide reductase C (EC 1.11.1.26) (Alkyl hydroperoxide reductase protein C22) (Peroxiredoxin) (SCRP-23) (Sulfate starvation-induced protein 8) (SSI8) (Thioredoxin peroxidase)	1	0	1
91	P09323	PTS system N-acetylglucosamine-specific EIICBA component (EIICBA-Nag) (EII-Nag) [Includes: N-acetylglucosamine permease IIC component (PTS system N-acetylglucosamine-specific EIIC component); N-acetylglucosamine-specific phosphotransferase enzyme IIB component (EC 2.7.1.193) (PTS system N-acetylglucosamine-specific EIIB component); N-acetylglucosamine-specific phosphotransferase enzyme IIA component (PTS system N-acetylglucosamine-specific EIIA component)]	1	0	1
92	P0A910	Outer membrane protein A (OmpA) (Outer membrane porin A) (Outer membrane protein 3A) (Outer membrane protein B) (Outer membrane protein II*) (Outer membrane protein d)	1	0	1
93	P33136	Glucans biosynthesis protein G	1	0	1
94	P0ACE7	Purine nucleoside phosphoramidase (EC 3.9.1.-) (Histidine triad nucleotide binding protein HinT) (HIT protein)	1	0	1

95	P0A862	Thiol peroxidase (Tpx) (EC 1.11.1.24) (Peroxiredoxin tpx) (Prx) (Scavengase p20) (Thioredoxin peroxidase) (Thioredoxin-dependent peroxiredoxin)	1	0	1
96	P77674	Gamma-aminobutyraldehyde dehydrogenase (ABALDH) (EC 1.2.1.19) (1-pyrroline dehydrogenase) (4-aminobutanal dehydrogenase) (5-aminopentanal dehydrogenase) (EC 1.2.1.-)	1	0	1
97	P0C0L2	Peroxiredoxin OsmC (EC 1.11.1.-) (Osmotically-inducible protein C)	1	0	1
98	P06610	Thioredoxin/glutathione peroxidase BtuE (EC 1.11.1.24) (EC 1.11.1.9)	1	0	0
99	P0A908	MltA-interacting protein	1	0	0
100	P0A6A3	Acetate kinase (EC 2.7.2.1) (Acetokinase)	1	0	0
101	P04805	Glutamate--tRNA ligase (EC 6.1.1.17) (Glutamyl-tRNA synthetase) (GluRS)	1	0	0
102	P45578	S-ribosylhomocysteine lyase (EC 4.4.1.21) (AI-2 synthesis protein) (Autoinducer-2 production protein LuxS)	1	0	0
103	P0AA10	50S ribosomal protein L13 (Large ribosomal subunit protein uL13)	1	0	0
104	P0ABD8	Biotin carboxyl carrier protein of acetyl-CoA carboxylase (BCCP)	1	0	0
105	P0A7M6	50S ribosomal protein L29 (Large ribosomal subunit protein uL29)	1	0	0
106	P02359	30S ribosomal protein S7 (Small ribosomal subunit protein uS7)	1	0	0
107	P0AC62	Glutaredoxin 3 (Grx3)	1	0	0

108	P0A7J3	50S ribosomal protein L10 (50S ribosomal protein L8) (Large ribosomal subunit protein uL10)	1	0	0
109	P0A6L0	Deoxyribose-phosphate aldolase (DERA) (EC 4.1.2.4) (2-deoxy-D-ribose 5-phosphate aldolase) (Phosphodeoxyriboaldolase) (Deoxyriboaldolase)	1	0	0

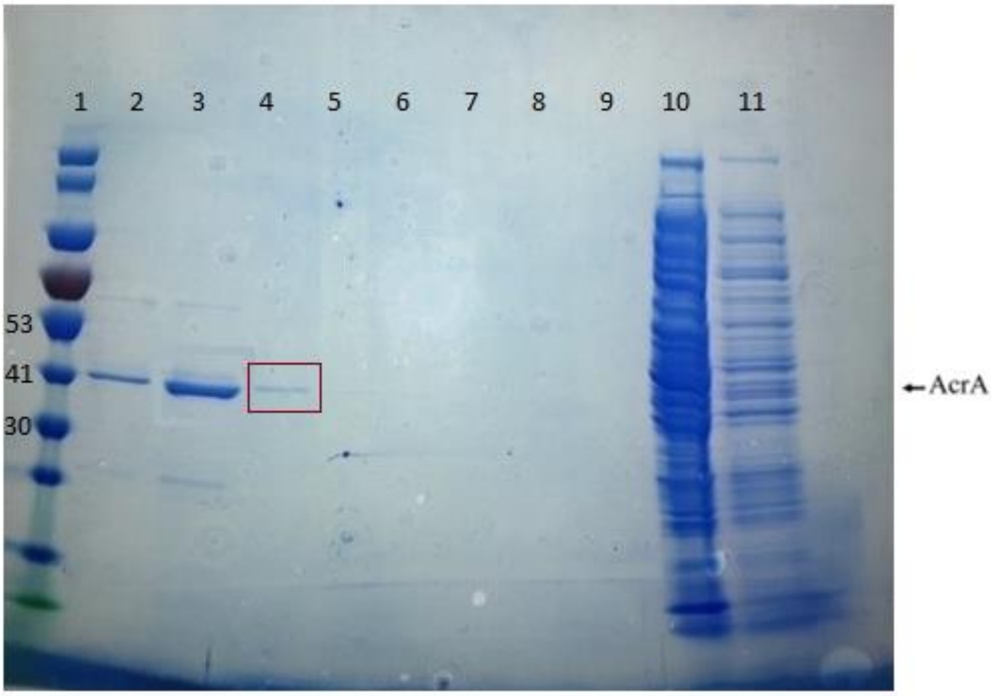
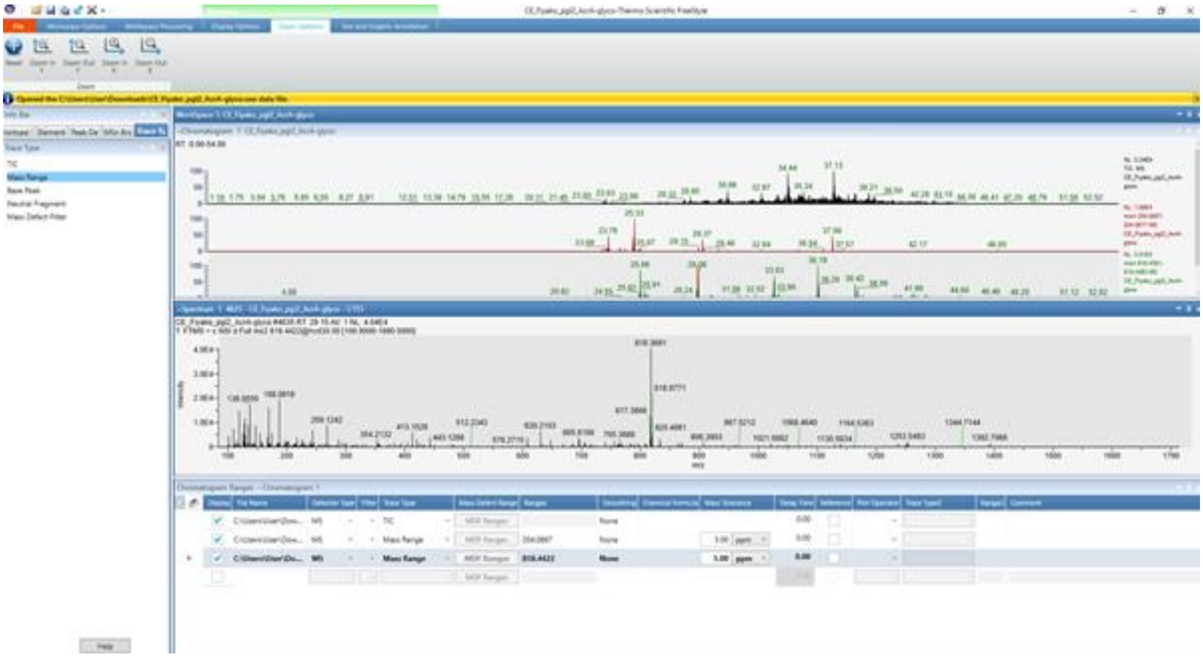


Fig 1: InstantBlue (coomassie) stain of a 4-12% Bis-Tris acrylamide gel containing the His-tag purified AcrA protein from CLM24 pEC(acrA) pACYC(pgl2). Lane 1 - EZ Pre-stained protein marker, Lane 2 (Control induced pEC(AcrA), 3 (His-trap column elution) induced pEC(AcrA) pACYC(pgl2) & 4 (His-trap column elution) uninduced pEC(AcrA) pACYC – Control 2. Lanes 5 & 6 were also elution fractions from His-trap column while 7 – 9 were wash fractions. Lane 10 (total protein fraction) and Lane 11 (flow through).



Display	File Name	Detector Type	Trace Type	Ranges	Smoothing	Mass Tolerance	Delay Time
✓	C:\Users\User\Dow...	MS	TIC		None		0.00
✓	C:\Users\User\Dow...	MS	Mass Range	204.0867	None	5.00 ppm	0.00
✓	C:\Users\User\Do...	MS	Mass Range	818.4422	None	5.00 ppm	0.00

Fig 2: Fragmented ion chromatograph of Control 2 (Acra with glycosylation machinery) data.