Differential gene expression analysis of aclacinomycin producing *Streptomyces galilaeus* ATCC 31615 and its mutant

Master's thesis

MD A B M Sharifuzzaman

Department of Life Technologies

Molecular Systems Biology

University of Turku

Supervisor:

Professor Mikko Metsä-Ketelä, Ph.D.

Department of Life Technologies

Antibiotic Biosynthesis Engineering Lab

University of Turku

Co-supervisor:

Keith Yamada, M.Sc.

PhD candidate

Department of Life Technologies

Antibiotic Biosynthesis Engineering Lab

University of Turku

Master's thesis 2021

The originality of this thesis has been checked in accordance with the University of Turku Quality assurance system using the Turnitin Originality check service

Summary

Streptomyces from the genus Actinomycetales are soil bacteria known to have a complex secondary metabolism that is extensively regulated by environmental and genetic factors. Consequently they produce antibiotics that are unnecessary for their growth but are used as a defense mechanism to dispel cohabiting microorganisms. In addition to other isoforms *Streptomyces galilaeus* ATCC 31615 (WT) produces aclacinomycin A (Acl A), whereas its mutant strain HO42 (MT) is an overproducer of Acl B. Acl A is an anthracycline clinically approved for cancer chemotherapy and used in Japan and China. A better understanding of the how the different isoforms of Acl are made and investigations into a possible Acl recycling system would allow us to use metabolic engineering for the generation of a strain that produces higher quantities of Acl A with a clean production profile.

In this study, RNA-Seq data from the WT, and MT strains on the 1st (D1), 2nd (D2), 3rd (D3), and 4th (D4) day of their growth was used and differentially expressed genes (DEGs) were identified. Differential gene expression (DGE) analyses were carried out using Bioconductor R-packages *i.e.,* Bowtie2, HTSeq, and edgeR embedded in Chipster. DEGs were further analyzed for gene ontology (GO) enrichment and mapped to Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. .

The number of significantly ($P < 0.05$) DEGs from the first analysis were 891 (D1), 1573 (D2), 1638 (D3), and 1392 (D4). However, DGE analysis by comparing RNA-Seq data within the strain gives two sets of gene list (WT and MT). The number of DEGs from WT were 915 (D1-D2), 810 (D2-D3), and 363 (D3-D4) and from the MT, there were 844 (D1-D2), 145 (D2-D3) and 170 (D3-D4). GO enrichment analysis of the DEGs between the strains showed significant enrichment in polysaccharide catabolic process, carbohydrate transport, cellular process, organic substance metabolic and biosynthetic process, catalytic activity, hydrolase, and oxidoreductase activity. Additionally, these DEGs were enriched from membrane origin. KEGG pathway analysis of the DEGs from both data sets showed that mutant strain had a lower number of mapped DEGs in carbon and fatty acid metabolism, metabolism of different amino acids. There were variations in some primary metabolic cycles, such as the TCA cycle, oxidative phosphorylation, glycolysis. ABC transporter, two-component system, nucleotide metabolism was also varied between the strain.

Keywords: Aclacinomycin, RNA-Seq, differential gene expression, gene ontology and enrichment, KEGG pathway

Acknowledgments

I want to thank Professor Mikko Metsä-Ketelä for allowing me to work with the ABE group and for the brilliant guidance that provides me with the right direction and supervision of this thesis. I am also thankful to my co-supervisor, Keith Yamada M.Sc. (Ph.D. student of Professor Mikko Metsä-Ketelä), for his continuous support throughout my thesis. His guidance has great importance to me; otherwise, it would not be this easy to analyze RNA-Seq data.

I am indebted to Bikash Baral M.Sc. (Ph.D. student of Professor Mikko Metsä-Ketelä) for your valuable time. Thanks to all the members of the ABE group.

I am also thankful to the Finnish Functional Genomics Center (FFGC) to provide the facilities for RNA sequencing and the IT center for science Ltd. (CSC) to provide IT support.

Finally, I want to thank my wife for her countless support that I needed during my thesis, and the simile of my little baby gave me positive energy.

Contents

List of tables

List of figures

Abbreviations

ABC	ATP binding cassette
ABG	Additional biosynthetic gene
ACCase	acetyl CoA carboxylage
Acl A/B/N/Y/X	Aclacinomycin A/B/N/Y/X
Acl/Acm	Aclacinomycin
ACP	acyl carrier protein
ACT	actinorhodin
Akn	Aklavinone
AknOx	Aclacinomycin oxidoreductase
antiSMASH	antibiotics & Secondary Metabolite Analysis Shell
AT	acyltransferase
ATP	adenosine tri phosphate
BGC	biosynthetic gene cluster
bp	base pair
BP	biological process
CBG	core biosynthetic gene
CC	cellular component
cDNA	complementary DNA
CP	cellular processing
CPM	counts per millions
CV	controlled vocabularies
D1/D2/D3/D4	day 1/day 2/ day 3/ day 4
D1-D2	day1-day2
D2-D3	day2-day3
D3-D4	day3-day4

DE	differential expression
DEGs	differentially expressed genes
dF	deoxyfucose
DGE	differential gene expression
DH	dehydratase
DNA	deoxyriobonucleic acid
dTTP	deoxythymidie triphosphate
dUTP	deoxyuradin triphosphate
EIP	environmental information processing
EM	energy metabolism
ER	enoylreductase
FASTA	Fast-all or FastA
FC	fold change
FDR	False discovery rate
FPKM	Fragments Per Kilobase Million
GBM	glycan biosynthesis and metabolism
GC	guanine-cytosine
GFF	general feature format
GIP	genetic information processing
GO	gene ontology
GOC	gene ontology consortium
GT	glycosyl transferase
GTF	gene transfer file
KAAS	KEGG Automatic Annotation Server
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG orthology
KR	ketoreductase

KS	ketosynthase
LP	lipid metabolism
MCV	metabolism of cofactors and vitamins
MET	Methyl transferase
MF	molecular functions
MIBiG	Minimum Information about a Biosynthetic Gene Cluster
minPKS	minimal PKS
mRNA	messenger RNA
MT	*Streptomyces galilaeus* ATCC 31615 HO42
MTP	metabolism of terpenoids and polyketides
NM	nucleotide metabolism
NRPSs	non-ribosomal peptide synthetases
ORFs	open reading frames
PANTHER	Protein ANalysis THrough Evolutionary Relationships
PCR	polymerase chain reaction
PGM	Phosphoglucose mutase.
PKSs	polyketide synthases
RG	Regulatory gene
Rhn	rhodosamine
Rho	rhodinose
RNA	ribonucleic acid
RNA-Seq	RNA sequencing
SARPs	*Streptomyces* antibiotic regulatory proteins
SBH	single-directional best hit
SCoA	succinyl-CoA
SD	standard deviation
SM	secondary metabolite

sp.	species
TCS	two-component system
TG	transport gene
TMM	Trimmed mean of values
UniProt	Universal Protein Resource
WT	*Streptomyces galilaeus* ATCC 31615

1. Literature reviews

1.1 Introduction

Streptomyces are Gram-positive, mycelial, filamentous, soil-dwelling bacteria belonging to the genus *Actinomyces* whose genetic material (i.e. DNA) is GC-rich (70%) (de Lima *et al.,* 2012) that is best known for producing a diverse class of biologically active compounds (Barka *et al.,* 2016). The morphological differentiation process of *Streptomyces* is unique among the Gram-positives such that after forming hyphae they differentiate into chain of spores and that require specialized and coordinated metabolism (de Lima *et al.,* 2012). They produce secondary metabolites (SMs) to get a competitive advantage over surrounding species (Netzker *et al.,* 2018) even within the same genus (Bosso *et al.,* 2010). Production of these SMs typically starts during the early stationary phase (Nieselt *et al.,* 2010). They are capable of producing a wide array of bioactive secondary metabolites (SM) such as antifungals, antivirals, antitumoral, anti-hypertensives, antibiotics and immunosuppressives (Omura et al., 2001a).

SMs produced by *Streptomyces* are not required for cell growth and reproduction, and their primary purpose is to improve the survivability of *Streptomyces* (Scherlach *et al.,* 2013). Many of these SMs are further developed to medically useful products such as antibiotic drugs, anti-cancer agents, antifungal compounds, anthelmintic drugs, and many more (Medema *et al.,* 2014). Among all the naturally derived antibiotics which are in use, two-third of them are derived from *Streptomyces* (Barka *et al.,* 2016). Apart from the fact of the stress response and defense mechanism, antibiotics can be produced as a result of the symbiosis between *Streptomyces* sp. and a host such as insects, marine animals and plants. *Streptomyces* sp. exert antibiotics as a result of this symbiotic relationships that protect the host and host provide the nutrients of *Streptomyces* sp. *Streptomyces* can also develop a parasitic relationship with the host *e.g.* plant and human, and create pathogenicity in the host (Seipke *et al.,* 2012). These SMs are usually encoded by a group of genes that cluster together and referred as biosynthetic gene cluster (BGC) (Tran *et al.,* 2019).

1.2 Structure and regulation of BGCs

BGCs located in a particular area of a genome and together they encode a specific biosynthetic pathway for the production of specific metabolites and their variants (Kjærbølling *et al.,* 2019 and Osbourn, 2010). In addition to the pathway-specific regulatory genes, a BGC encodes all enzymes required for SM synthesis. They typically contain one or more enzyme encoding

genes, that synthesize the core structure of the compound, genes encode tailoring enzymes which are involved in modifying the core structure. Additionally, there are enzymes those have regulatory functions such as transcription factors and resistance genes (Osbourn, 2010). Pathway-specific regulators residing within the BGC can act as activators or repressors (Van Der Heul *et al.,* 2018). By responding to diverse signals, the distantly localized global regulators control the expression of secondary metabolism. These regulators are pleiotropic. Therefore, they can control the expression of several BGCs simultaneously. Events of nutrient starvation, chemical stressors, environmental stressors, and physical damage are the stimulating factors for the expression of these regulatory genes (Bibb & Hesketh, 2009). Besides these core biosynthetic enzymes, many BGCs also accommodate enzymes to synthesize modified carbohydrates (*e.g.* deoxy sugars in the erythromycin gene cluster) that are appended to the core SM structures (Oliynyk *et al.,* 2007). Polyketide synthases (PKSs) and non-ribosomal peptide synthetases (NRPSs) are the most common backbone synthesizing enzymes (Kjærbølling *et al.,* 2019). Beside these there are some other notable types of BGCs such as post-translationally modified peptides (RiPPs) (Arnison *et al.,* 2013), terpenes, saccharides, alkaloids (Blin *et al.,* 2019). The size of BGCs dramatically varies on the complexity of the synthesized metabolites (Baral *et al.,* 2018). However, there might be other types of BGCs that are still unknown and have potentiality for SM production.

1.3 Biosynthesis of polyketides in *Streptomyces*

A large group of SMs made from polyketides, encompassing molecules such as macrolides, aromatics and polyenes (Cummings *et al.,* 2014). These SMs possess a wide variety of structures and functions. Some of the notable functions are antibacterial (*e.g.* streptomycin), antifungal (*e.g.* amphotericin B), anti-cancer (*aclacinomycin* A), immune-suppressing (*e.g.* rapamycin), anti-inflammatory (*e.g.* flavonoids) (Risdian *et al.,* 2019; Rokem *et al.,* 2007 and Schwecke *et al.,* 1995) and antiviral (*e.g.* antimycin A1a) (Raveh *et al.,* 2013). Polyketide biosynthesis is well studied in *Streptomyces* (Risdian *et al.,* 2019) and species from this genus produce three distinct types of polyketides (type I, type II, and type III) (Staunton & Weissman, 2001). Biosynthesis of polyketides is a complicated process that involves multiple enzymatic reactions (Table 1.1, Figure 1.1 and 1.3). These enzymes are called polyketide synthases (PKSs).

1.3.1 Type I PKS

Type I PKSs are multifunctional proteins that are organized into modules made up of semi-repetitive domains, each of which executes a particular enzymatic reaction (Risdian *et al.,* 2019). This type of PKS is also called a modular PKS (Shen, 2003). The essential domains present in the modules are acyltransferase (AT), ketosynthase (KS), and acyl carrier protein (ACP) (Table 1.1). Each module performs a set of distinct activities responsible for the catalysis of polyketide chain elongation cycle in a non-iterative way (Risdian *et al.,* 2019).

Table 1. 1: Some notable enzymes involve in Type I PKS synthesis and their functions.

Enzyme	**Function**
Acyltransferase (AT)	Acts upon acyl group and catalyze the binding of the substrate (*e.g.* acetyl or malonyl-CoA) to the acyl carrier protein (ACP)
Ketosynthase (KS)	Catalyze the condensation of ACP bound substrate
Ketoreductase (KR)	Reduce keto ester in the substrate
Dehydratase (DH)	Dehydrate the compound
Enoylreductase (ER)	Reduces the number of >C=C< in the molecule.

The essential domains collaborate to produce β-keto ester intermediates (Staunton & Weissman 2001). These keto groups are modified by other enzymes such as β-ketoreductase (KR), dehydratase (DH), and enoyl reductase (ER) (Figure 1.1). Before separating from the final polyketide product from the completed module by a specific enzyme, the growing polyketide chain is transferred from one module to another (Risdian *et al.,* 2019). Type I PKSs are generally responsible for producing macrolides, polyethers, and polyene polyketides (Shen, 2003).

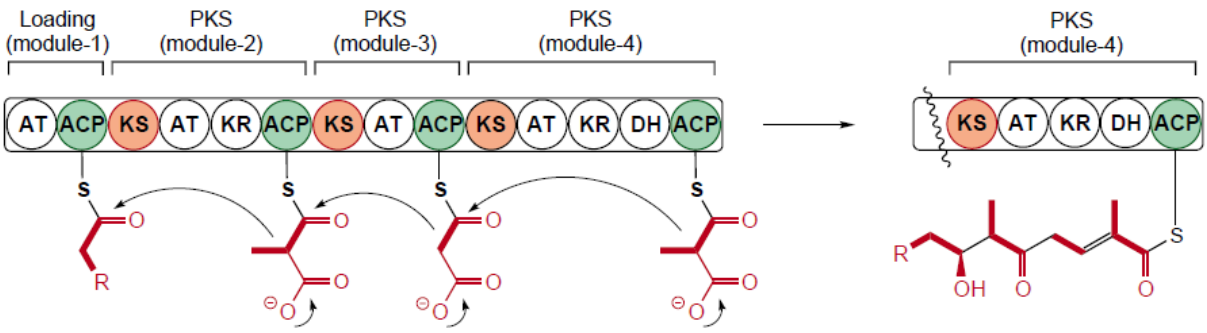
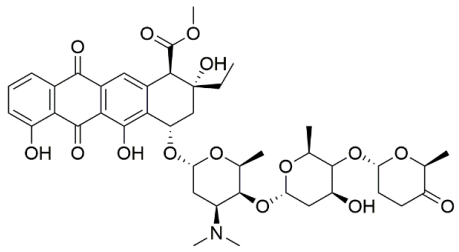


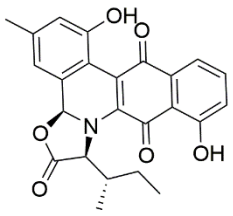
Figure 1. 1: Typical structure of a type I PKS (erythromycin A). It has 4 modules and 19 domains. ACP, acyl carrier protein; AT, acyltransferase; KS, ketosynthase; KR, ketoreductase; DH, dehydratase; ER, enoylreductase (Shen, 2003).

1.3.2 Type II PKS

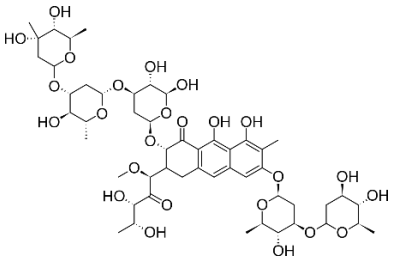
Type II PKSs are responsible for the synthesis of aromatic polyketides. They are classified into seven groups *i.e.,* anthracyclines, angucyclines, aureolic acids, tetracyclines, tetracenomycins, pradimicins and benzoisochromanequinones; based on their polyphenolic ring system and biosynthetic pathways (Risdian *et al.,* 2019). Amino sugar linked tetracyclic aglycone forms the basic structure of anthracyclines (Beretta & Zunino, 2007). Angucyclines feature a tetracyclic benz[α]anthracene skeleton and forms the largest group of aromatic PKSs (Matulova *et al.,* 2019). Members of the aureolic acid family are tricyclic polyketides. Tetracyclines have a linearly oriented tetracyclic ring system without quinone–hydroquinone groups in rings B and C, but tetracenomycins have a similer ring system with quinone group in ring B. Pradimicins are considered as modified angucyclines. Benzoisochromanequinones contain a quinone derivative from an isochromane structure (Risdian *et al.,* 2019) (Figure 1.2).



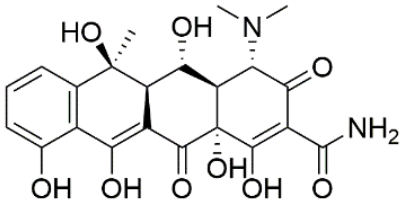
Aclacinomycin A



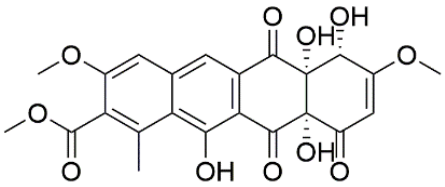
Jadomycin



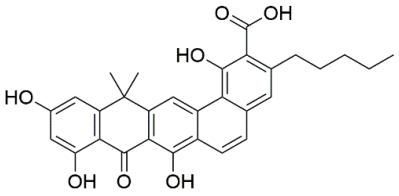
Mithramycin



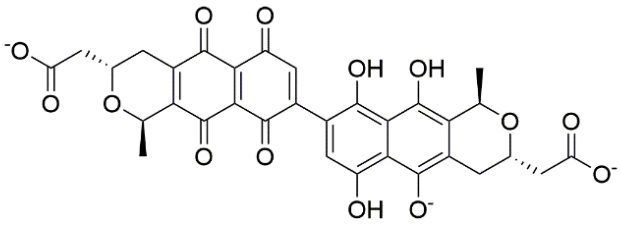
Oxytetracycline



Tetracenomycin C



Benastatin A



Actinorhodin

Figure 1. 2: Different groups of type II aromatic polyketide compounds. Aclacinomycin (anthracycline), jadomycin B (angucycline), mithramycin A (aureolic acid), oxytetracycline (tetracycline), tetracenomycin C (tetracenomycin), benastatin A (pradimicin) , and actinorhodin (benzoisochromanequinone).

Type II PKSs have a cascade of monofunctional proteins, and they work iteratively unlike type I PKSs (Risdian *et al.,* 2019). The synthesis of type II PKS starts from acetate or less commonly propionate. At first, a liner polyketide chain consist of 16-26 carbon is assembled by repeated Claisen condensation reactions from a starter unit and 2-carbon acetate extender unit (Ridley *et al.,* 2008). These reactions carried out by minimal PKS (minPKS) complex system, which comprises ACP and two KS unites (KS_α and KS_β). The minPKS cooperatively produces the poly-β-keto chain. The α unit of the KS catalyzes the condensation of the precursor molecule, and the β unit in type II PKS system acts as a length determining factor. The poly-β-keto chain is converted to an aromatic compound by some additional enzymes *e.g.* KR, cyclase and aromatase. Afterwards, the system employs other enzymes *i.e.,* oxygenases and glycosyl and methyl transferase (GT and MET) for post tailoring processes (Hertweck *et al.,* 2007, Tang *et al.,* 2017 and Staunton *et al.,* 2001). A general illustration of Type II PKS synthesis is presented in Figure 1.3.

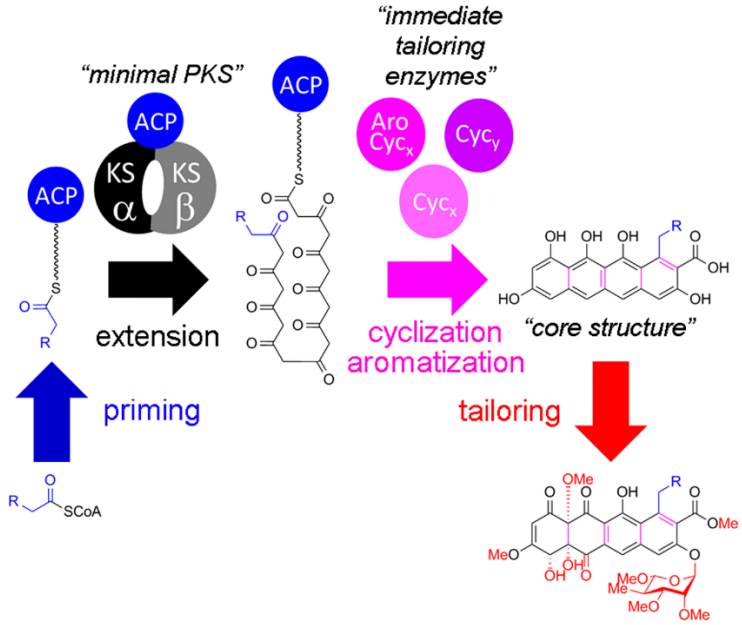


Figure 1. 3: Schematic representation of the steps and reactions occurring in elloramycin synthesis. A Type II PKSs encoded by a typical polyketide BGC (Ogasawara et al., 2015). The key steps involve here are priming of the minimal PKS, formation of poly-β-keto intermediate from the polyketide chain by ACP and ketosynthase α/β heterodimer, followed by cyclization and aromatization to form the cyclized core structure by the tailoring enzymes. The intermediates, final products and respective catalyzing enzymes are coded with similar color.

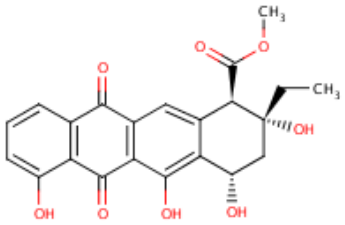
1.3.3 Type III PKS

Type III PKSs do not utilize ACP and they are comparatively simpler than the other two types of PKS systems (Risdian *et al.,* 2019) and can produce a wide array of compounds such as chalcones, pyrones, acridones, phloroglucinols, stilbenes, and resorcinolic lipids (Yu *et al.,* 2012). This system utilizes acyl-CoA as a substrate (Shen, 2003) and can be found in plants, bacteria, and fungus (Yu *et al.,* 2012).

1.4 Acl biosynthesis

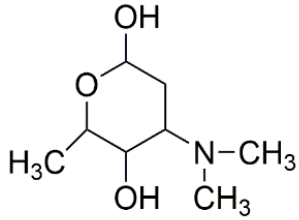
1.4.1 Structural variations of acl

Acl are anthracycline antibiotics, produced as a complex by *Streptomyces galilaeus* (Oki *et al.,* 1975) with Acl A, B, and Y forms reported. In addition, Fujii & Ebizuka (1997) reported that *S. galilaeus* could produce Acl X. Acl A is used to treat leukemias and non-Hodgkin's lymphomas in Japan and China (Räty *et al.,* 2002). Aklavinone (Akn) (Figure 1.4a) is a common precursor for many anthracyclines such as daunorubicins, rhodomycins (Ylihonko *et al.,* 1994) and ε-rhodomycinone (Chung *et al.,* 2002). This molecule (Akn) forms the aglycone skeleton of all Acl variants. The sugar components of different Acl are rhodosamine, 2-deoxyfucose, cinerulose A, L-aculose, and cinerulose B. Among these sugar molecules, the first two are common in the Acl variants (Oki *et al.,* 1979) (Figure 1.4b). Nippon Roche group of Japan isolated *S. galilaeus* ATCC 31615 (WT), which is capable of producing Acl A and Acl B (Fujii *et al.,* 1997). Random mutagenization of WT has produced several mutant strains for the purpose of overproducing Acl A (Figure 1.5). A key issue in the industrial manufacturing of Acl A is the production of a mixture of different aclacinomycins, which are very difficult to separate from each other chromatographically.

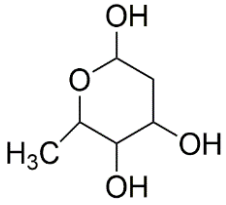


(a)

First two common sugar molecules in Acl variants

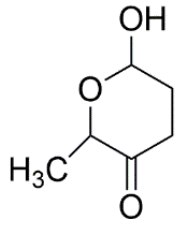


Rhodosamine (i)

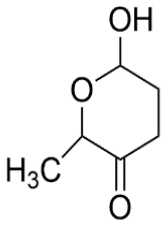


2-Deoxyfucose (ii)

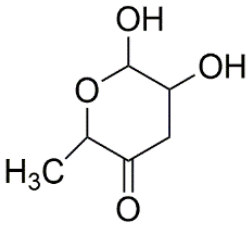
Variant specific third sugar molecule



Cinerulose A (iii)



Aculose (iv)



Cinerulose B (v)

(b)

Figure 1. 4: Acl precursor molecules and its isoform specific sugar molecules. (a) Chemical structure of Aklavinone (Akn). Suger molecules bind to C7 of this aglycone and form different variants of Acl. (b) Chemical structure of the suger components present in Acl variants. (i) and (ii) first two common sugar molecules in Acl isoforms and (iii-v)isoform specific third sugar molecules: (iii) L-cinerulose gives Acl A, (iv) aculose gives Acl Y and (v) cinerulose B gives Acl B.

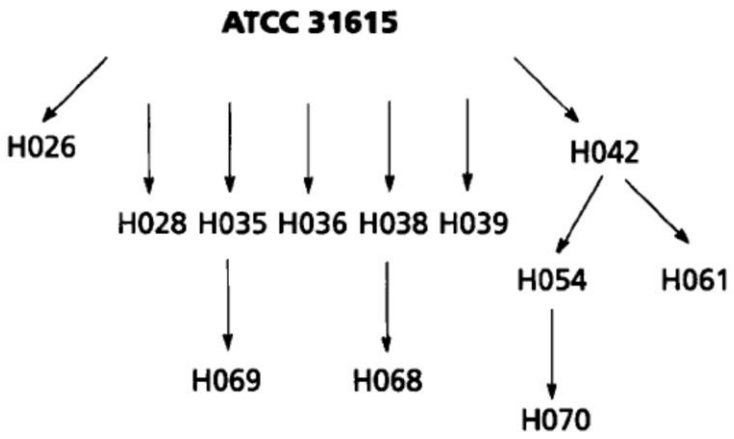


Figure 1. 5: The pedigree of mutant strains from WT. Each of the mutant strains has a characteristic production profile. HO42 is an overproducer of Acl B (Ylihonko et al., 1994).

1.4.2 Acl recycling system

Anthracyclines are a type II polyketide produced, especially by *Streptomyces* species. The structure of anthracycline antibiotics is divided into two parts *i.e.,* aglycone carbon skeleton and sugar moiety (Ylihonko *et al.,* 1994). Akn biosynthetic pathway starts by the condensation of a propionate and nine acetate molecules by a series of catalytic reactions of the minimal PKS, which result in a decaketide. A more stable aklanonic acid is formed after subsequent ketoreduction, cyclization, and oxygenation reactions of the initially formed decaketide (Räty *et al.,* 2002). Different post-polyketide reactions, such as methylation, cyclization, and reduction, results in the formation of the principle intermediate (Akn).

An interesting target to study on Acl biosynthesis could be their trisaccharide moiety. The principal intermediate (Akn) is glycosylated; therefore, a chain of three sugar molecule attached with this intermediate and form Acl N. This sugar chain attached to position C7 of Akn as follows: rhodosamine (Rhn)-2-deoxyfucose (dF)- rhodinose (Rho) (Räty *et al.,* 2002). Rho is further converted to cinerulose by extracellular oxidoreductase (AknOx) and form Acl A. Furthermore, this cinerulose is converted to aculose to form Acl Y by the same enzyme (Alexeev *et al.,* 2007). The final variant Acl B is formed by a non-enzymatic oxidation reaction, which converts aculose to cinerulose B (Gräfe *et al.,* 1988) (Figure 1.6 and 1.7).

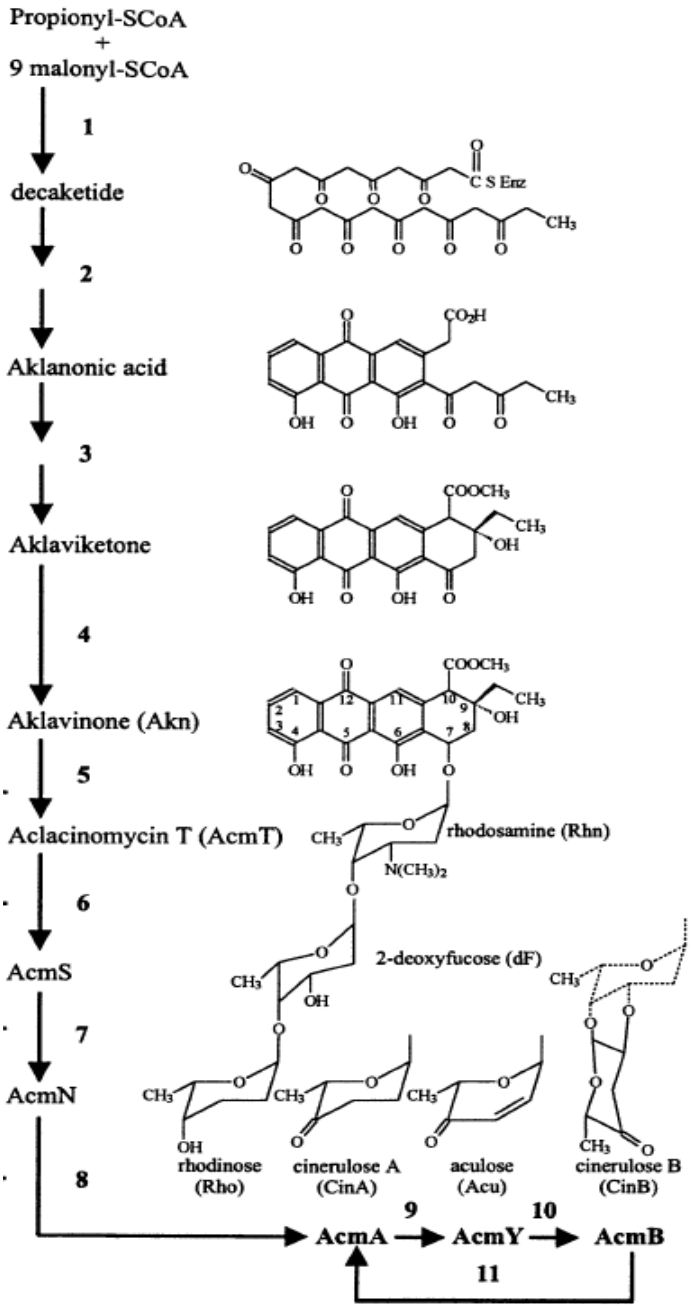


Figure 1. 6: Biosynthetic pathway of Acl A, B, and Y. Acm; aclacinomycin (Räty et al., 2002).

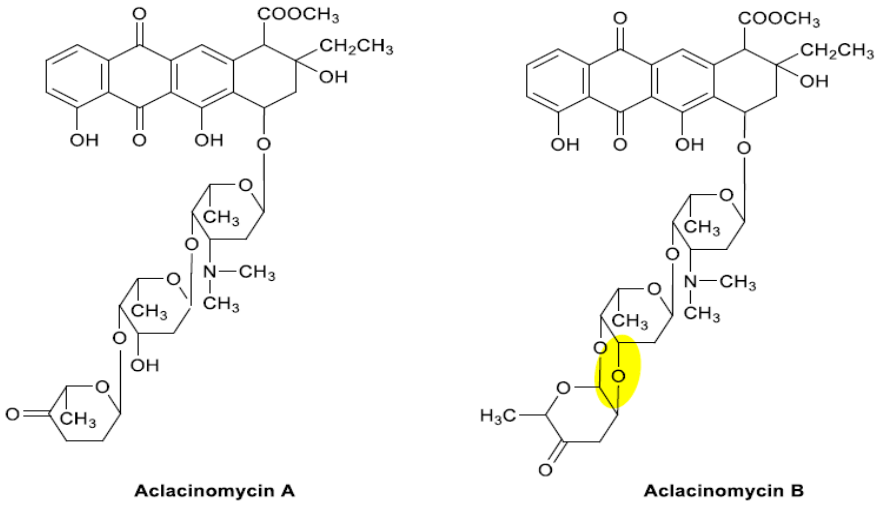


Figure 1. 7: Chemical structure of Acl A and B. They differ in the bond between the second and third sugar molecules (Oki et al., 1975).

However, the WT strain appears to be able to harvest Acl B by transporting the molecule back inside the cell and re-convert it to Acl A. The mechanisms of this Acl recycling system are still unknown. The *S. galilaeus* ATCC 31615 mutant HO42 (MT) accumulates Acl B and may have an impaired recycling system. Genome sequencing revealed that the strain does not have any mutations inside the predicted BGC, which depicts the fact that this recycling system is partially dependent on one or more genes outside the cluster system (Figure 1.8).

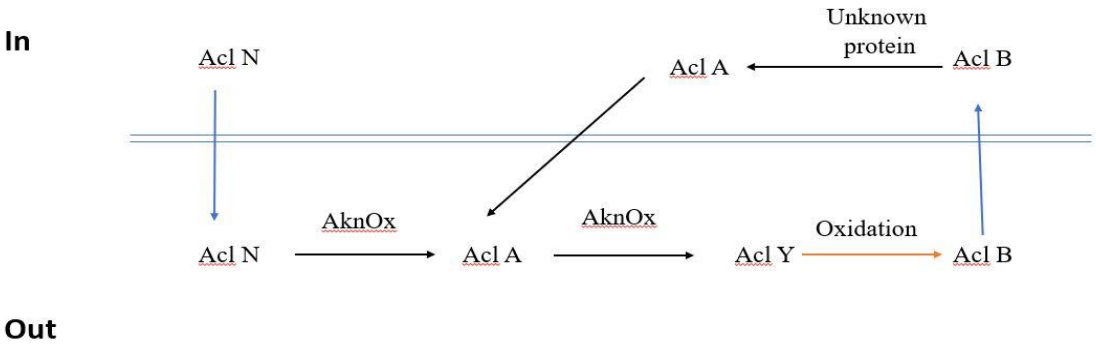


Figure 1. 8: Hypothesized recycling of Acl. (Empirically deduced by Docent Jarmo Niemi from University of Turku).

1.4.3 Characterized genes in Acl gene cluster

The MIBiG (Minimum Information about a Biosynthetic Gene Cluster) Data Standard and Repository was established in 2015 to enable curation and storage of known BGCs, and during the last five years, 851 new BGCs have been added (Kautsar *et al.,* 2020). Among these stored BGCs, this repository has information about three different publications (accession no: BGC0000191, BGC0000191, and BGC0000193) on Acl producing gene cluster. According to this repository, the BGC information for Acl biosynthesis is not complete. All the genes have not been characterized yet. However, researchers have characterized some of the genes from several experiments (Räty *et al.,* 2000, Räty *et al.,* 2002 and Chung *et al.,* 2002).

Bioinformatic analysis reveals that similar to all type II PKS BGCs, expression of the KS_{α} and KS_{β} ketosynthase genes *aknB* and *aknC*, respectively, is translationally coupled. Adjacent to them, the *aknD* gene closely resides (103 bp) with *aknC* and encodes a 91 amino acid long ACP. Another gene in the minimal PKS system is *aknE2*, which encodes a 368 amino acid long peptide, responsible for serving propionate as a starter unit. The genes *aknD* and *aknE2* are thought to be translationally coupled, and they are 4 bp overlapped. The gene *aknF* encodes a 347 amino acid peptide, which is an AT (Räty *et al.,* 2002) (figure 1.9).

After formation of the polyketide chain, AknA reduces the keto group at C2 to the -OH group, which is subsequently removed during closing and aromatization of the first ring by AknE1 (450 amino acid peptide). The *aknE1* gene is much longer than other homolog genes found in other type II PKS clusters. The *aknX* gene is assumed to be involved in oxygen addition at C12 to make aklanonic acid at step 2 (Figure 1.6). The probable function of the gene product of AknV in this cluster is the glycosylation of Acl. AknG and AknH are post-polyketide enzymes. The 286 amino acid AknG is a methyltransferase (MET) and esterify the -COOH of aklanonic acid. AknH encodes a 144 amino acid peptide, which is a cyclase that closes the fourth ring of aklanonic acid (Räty *et al.,* 2002) (Figure 1.9).

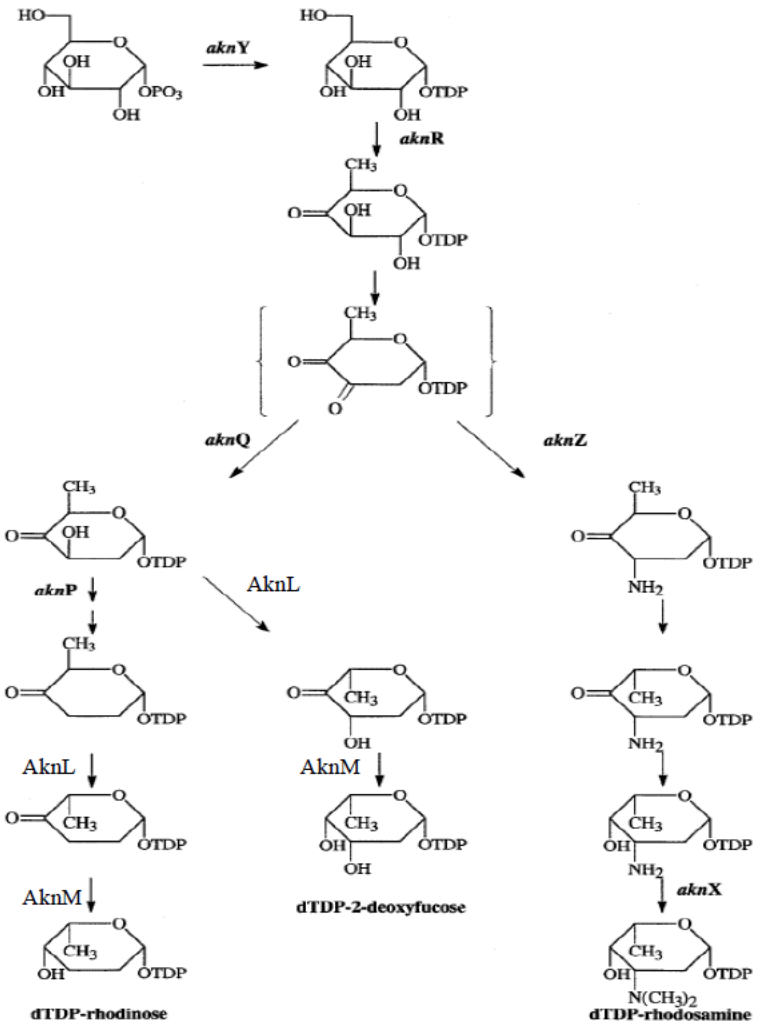


Figure 1. 9: Hypothesized biosynthesis of Acl sugar molecules. Extracted from Räty et al., (2000) and modified according to Metsä-Ketelä et al., (2007).

A 280 amino acid long peptide AknI has a regulatory function that belongs to *Streptomyces* antibiotic regulatory proteins (SARPs). It acts as a pathway specific activator to promote the expression of genes in BGCs. The product (440 amino acid) of the *aknK* gene has glycosyl transferase (GT) activity (Räty *et al.,* 2002) and involved in the biosynthesis of the trisaccharide moiety. The AknK catalyzes the transfer of a sugar molecule to the mono-glycosylated aclacinomycin T and converts to Acl S (Figure 1.6) (Lu *et al.,* 2005). The cluster has another glycosyl transferase AknS (Räty *et al.,* 2000 and Lu *et al.,* 2005). AknL is thought to convert the dTDP-4-keto-6-deoxy-D-glucose to dTDP-4-keto-L-rhamnose (Räty *et al.,* 2002). AknM is responsible for the stereospecific reduction of C4 using NADPH as a reducing agent (Räty *et al.,* 2000). The pathway for dTDP-L-rhodosamine formation involves the transaminase encoding gene *aknZ*. Finally, the demethylation of dTDP-L-rhodosamine carried out by *aknX2*

gene (aknX in figure 1.9) (Räty *et al.,* 2000 and Metsä-Ketelä *et al.,* 2007). The cluster contains a large protein (>661 amino acid) coding gene *ankN*, which doesn't show significant homology with other proteins with known functions. Next to this gene, there is another regulatory protein-encoding gene *aknO*. On the complementary strand of *aknO*, there is *aknP* which encodes a 3-dehydratase. There is a 3-ketoreductase encoding gene *aknQ* on the reverse direction. AknR is assumed to be dTDP-glucose-4,6-dehydratase based on homology with other enzymes. The *aknS* encodes a GT and shares 13 bp overlapping sequence with *aknR* (Räty *et al.,* 2000). AknS has deficient activity in the absence of AknT, which means AknT is an activation protein for AknS. It stimulates the transfer of L-2-deoxyfucose to the aglycone aklavinone (Lu *et al.,* 2005). The gene *aknU* resides in the same orientation as *aknT*, which encode aklaviketone reductase. The 259 amino acid long peptide AknW is presumably a cyclase that closes the aromatic ring of the aklavinone. Evidence suggested that *aknY* may encode glucose-1-phosphate thymidylytransferase (Räty *et al.,* 2000). Experiment from Alexeev *et al.,* (2007) showed that AknOx (Acl oxydoreductase) catalyzes two consecutive steps of this pathway. It oxidizes terminal sugar rhodinose to cinerulose A by oxidizing the -OH group at C4 to a keto group. In the next step, it culminates two hydrogen atoms from cinerulose A and converts it to L-aculose. The characterized genes of Acl BGC has listed in Table 1.2.

Table 1. 2: Deduced functions of Akn genes from Acl BGC based on homology.

Gene product	**Length (amino acids)**	**Probable function**	**Reference**
AknA	261	Polyketide ketoreductase (KR)	(Räty *et al.,* 2002)
AknB	423	KS I	
AknC	407	KS II	
AknD	91	ACP	
AknE1	450	Aromatase	
AknE2	368	Starter unit determinant	

AknF	347	AT	
AknG	286	Methyltransferase (MET)	
AknH	144	Cyclase	
AknI	280	Activator	
AknK	440	GT	
AknL	201	Epimerase	
AknM	>205	KR	
AknX	122	Monooxygenase	
AknN	>661	Unknown	(Räty *et al.,* 2000)
AknO	272	Activator	
AknP	434	dTDP-hexose-3-dehydratase	
AknQ	329	dTDP-hexose-3-ketoreductase	
AknR	323	dTDP-hexose 4,6-dehydratase	
AknS	443	GT	
AknU	267	Aklaviketone reductase	
AknV	144	Glycosylation	
AknW	259	Cyclase	
AknX2	238	Aminomethylase	
AknY	291	dTDP-glucose 1-synthase	
AknZ	>340	Aminotransferase	

AknOx	545	Aclacinomycin oxydoreductase	(Alexeev *et al.,* 2007)
AknT	443	Activating protein of AknS	(Lu *et al.,* 2005)

1.5 DEG analysis in prokaryotes

A gene is defined as differentially expressed (DE) if the abundance of its transcript level is different per cell under the two conditions (Li & Li, 2018). High-throughput transcriptome sequencing (RNA-Seq) has become the key method to study DEGs (Costa-Silva *et al.,* 2017). The DEGs between the treatment groups have been identified by software packages that have been developed based on RNA-Seq data (Zhang *et al.,* 2014). The whole process can be divided into two categories: biochemistry and bioinformatics, where the generation of RNA-Seq data in a high-throughput platform is the biochemistry part. The generated RNA-Seq data is now analyzed by using various bioinformatic tools, which is a bioinformatic part of the RNA-Seq experiment.

1.5.1 Biochemistry of RNA-Seq

The biochemistry of RNA-Seq pipeline commonly involves three steps, such as RNA extraction and processing of RNA material, cDNA library preparation, and sequencing in a next-generation sequencing (NGS) platform. Initially, the extracted RNA materials are fragmented, and small complementary DNA (cDNA) sequences (reads) are generated from these fragments (Stark *et al.,* 2019). Ribosomal RNA (rRNA) comprises >80% of the total RNA material in bacteria. However, getting a meaningful information from the RNA-Seq data, it is important to get the reads predominantly from the mRNA (Culviner *et al.,* 2020; Westermann *et al.,* 2012). Based on the objectives and protocols, depletion of rRNA and enrichment of mRNA is common for prokaryotic RNA-Seq. The fragmentation method is a crucial aspect of sequencing library construction. Later these cDNAs are sequenced using a high-throughput platform. Based on the objectives, these reads can be sequenced from one direction (single-end), both directions (paired-end) or a specific strand. However, single-end sequencing is quicker, cheaper over paired-end reads, and enough for gene expression analysis. PCR amplification of the cDNA would be necessary before sequencing to enrich for fragments

that contain the expected 5′ and 3′ adapter sequences; although this enrichment of fragments may cause bias in the results (Christodoulou *et al.,* 2011).

The word "transcriptomics" was first used in the '90s (Nelson, 2001), and serial analysis of gene expression (SAGE) is the first sequencing-based transcriptomic method, which was based on the Sanger sequencing platform (Velculescu *et al.,* 1995). Microarray, a dominant technique developed in the '90s, was used to measure the abundance of a set of transcripts (Nelson, 2001). Although it was very popular at that time, this technique had notable limitations. This was a hybridization-based method, where user-defined probes were attached to plates for hybridization with sample RNA. Therefore, users could only measure defined sets of transcripts, which makes the experiments costly. However, the emergence of next-generation sequencing (NGS) technology in the 2000s (Cloonan *et al.,* 2008) allowed researchers to overcome the limitations of microarrays through the development of RNA-Seq. It has been nearly ten years since RNA-Seq technology emerged and become a ubiquitous tool for DGE studies for molecular research at the transcriptomic level. The principle of RNA-Seq has not changed substantially since then; however, each step of RNA-Seq has gone through rigorous improvement (Stark *et al.,* 2019). Selecting a proper NGS platform is always essential and depends upon many factors such as experimental design, read length, and cost. The majority of this high throughput sequencing technology uses sequencing by synthesis method, which can generate tens of millions of sequences in parallel. Sequencing in the NGS platform can either done by sequencing many identical copies of DNA molecule (ensemble-based) or a single DNA molecule (single molecule-based) (Kukurba & Montgomery, 2015) *e.g.* PacBio enable single-molecule real-time (SMRT) sequencing (Eid *et al.,* 2009). Illumina is currently dominating the sequencing industry, uses an ensemble-based approach (Bentley *et al.,* 2008). The adaptor-ligated DNA molecules are immobilized and clonally amplified on a glass flow cell using fluorescently labeled reversible terminator nucleotides. The PCR amplification bias is removed during downstream computational analysis. This approach has much less than a 1% error (Kukurba *et al.,* 2015) (Figure 1.10).

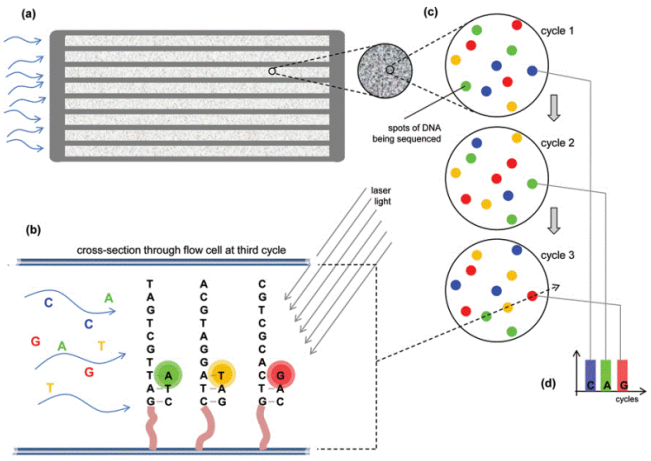


Figure 1. 10: Principle of Illumina Sequencing (Sequencing by Synthesis). (a) a simplified view of a flow cell; (b) fluorescently labeled nucleotides binds with the complementary bases, and a specialized camera captures the released fluorescent band. (c)incorporated bases in each cycle. d) different nucleotides with their specific fluorescents color (TUFTS, 2019).

1.5.2 Bioinformatics of RNA-Seq

From generating the data, the pipeline enters the bioinformatics part, which is also a multi-step process (Figure 1.11). Data analysis requires a combination of bioinformatics software tools based on experimental design and goals. RNA-Seq usually produces a large volume of raw sequence reads.

The initial raw RNA-Seq data goes through quality control step to check multiple properties of these reads *e.g.* read length, GC content and contamination, by using one or more tools such as FastQC (Andrews, 2010 and Zhao *et al.,* 2016). The small general sequences (reads) that are passed through the QC filter are mapped to a reference genome or transcriptome, to link millions of short reads into the quantification of expression.

Mapping can be defined as aligning a short read to a specific position in the reference genome, where an identical sequence exists (Zhao *et al.,* 2016). However, these short reads tend to be aligned with multiple locations of the genome, which makes RNA-Seq data analysis challenging and despite tremendous progress in RNA-Seq technologies; alignment of the reads is still considered as the most computationally intensive step of the entire bioinformatics RNA-Seq pipeline (Dao *et al.,* 2014).

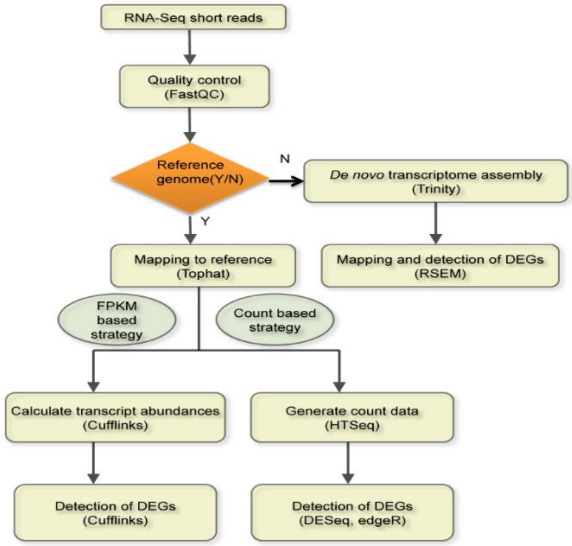


Figure 1. 11: General workflow of a typical DE analysis of RNA-Seq data. Image curtesy: Zhang et al., 2014.

De novo assembly can be used to align reads to one another to construct full-length transcript sequences without the use of a reference genome (Hölzer & Marz, 2019). There are some challenges in *de novo* alignment, such as it requires more significant computational requirements than reference-based DEGs analysis, additional validation, sequencing depth and annotation of aligned transcripts (Oshlack *et al.,* 2010).

Several file formats exist to store information about the location of transcription start sites, exons-introns (eukaryotes) *etc.* They differ in several aspects; however, all the formats agree on having one line per genomic feature. The General Feature Format (GFF) has nine required fields. There are two versions of the GFF (*e.g.* GFF2 and GFF3) format in use, which are similar. GTF (Gene Transfer File) is tab limited text file, like GFF2 format, but differs in its 9th field *i.e.* TYPE VALUE pairs of GTF files are separated by one space and must end with a semi-colon (Dündar *et al.,* 2015). It is also possible to visualize the aligned reads against the reference genome by using some genomic viewer *e.g.* Artemis (Carver *et al.,* 2012).

Quantification of sequence alignments may be performed either at the gene or transcript level (Dündar *et al.,* 2015). In gene-level quantification, one can directly count the reads overlapping the gene loci in the genome or aggregating the transcript level quantification per gene. The principle of counting reads aligning to genomic features is quite simple; however, there are some considerations needed to take care of depending upon the experiment and desired result (Figure 10). HTSeq is one of the most popular tools for gene quantification (Anders *et al.,* 2015). It has three read overlapping counting systems, *i.e.,* union, intersection_strict, and

intersection_nonempty (Figure 1.12), where union is the recommended setting. It is essential to know how a program is sensitive to different features, such as overlapping, reads mapped to more than one location, and reads overlapping multiple genomic features of the same kind, while choosing a program. The output summary file gives the overall statistics of the mapped and unmapped reads, which is very useful (Anders *et al.,* 2015 and Dündar *et al.,* 2015).

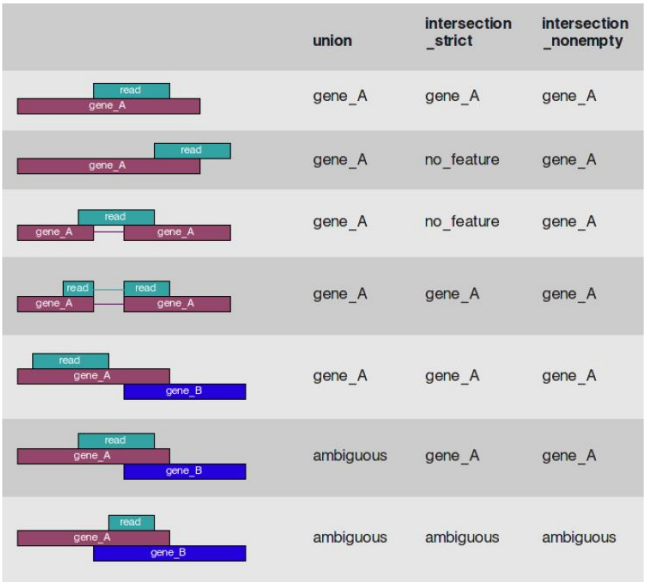


Figure 1. 12: Three different modes of read counting system in HTSeq package. Based on certain alignment conditions (left panel) the system decides how to address the read. For example, in the second case, the read was partially overlapped from the annotated region of gene_A. In "union" and "intersection_empty" mode of read overlapping, the read belongs to gene_A, but in "intersection_strict" mode, the read is not counted for a gene. (Image courtesy: Dündar et al., 2015)

The number of mapped reads to a gene depends upon multiple factors such as gene expression level itself, and other genes within the sample, the length of the gene/transcript, sequencing depth. Therefore, it is necessary to adjust the RNA-Seq data before making any comparative study (Costa-Silva *et al.,* 2017). RNA-Seq data normalization means the raw data are adjusted according to the factors that prevent direct comparison of the gene expression within (*e.g.* length and GC-content) and between (*e.g.* sequencing depth) the samples (Zyprych-Walczak *et al.,* 2015). Optimum normalization of the RNA-Seq data is one of the most important steps to avoid false-positive results (Dündar *et al.,* 2015). A number of normalization methods have adopted for RNA-Seq data analysis packages, for example, Median (DES) adopted in DESeq (Oshlack *et al.,* 2010), Trimmed Mean of -values (TMM), Upper Quartile (UQ) have adopted for edgeR (Marguerat & Bähler, 2010), Quantile (EBS) employed in the EBSeq package

(Anders & Huber, 2010) and Fragments Per Kilobase Million (FPKM) normalization (Loraine *et al.,* 2015).

Mappers have minimal impact than DEG identification on expression analysis (Costa-Silva *et al.,* 2017). Additional efforts such as replicates, avoiding bias, normalization has to be taken into account to optimize the statistical test to determine the DEGs from the normalized RNA-Seq data (Dündar *et al.,* 2015). Some of the best performing tools are edgeR (Robinson *et al.,* 2009), DESeq/DESeq2 (Anders *et al.,* 2015 and, and limma-voom (Ritchie *et al.,* 2015). Although edgeR has less control over false positives than DESeq and limma-voom (Ritchie *et al.,* 2015 and Zhang *et al.,* 2014); edgeR and DESeq2 are recommended if experiments has less than 12 replicates to capture the maximum number of DEGs even if the change is too small (Schurch *et al.,* 2015).

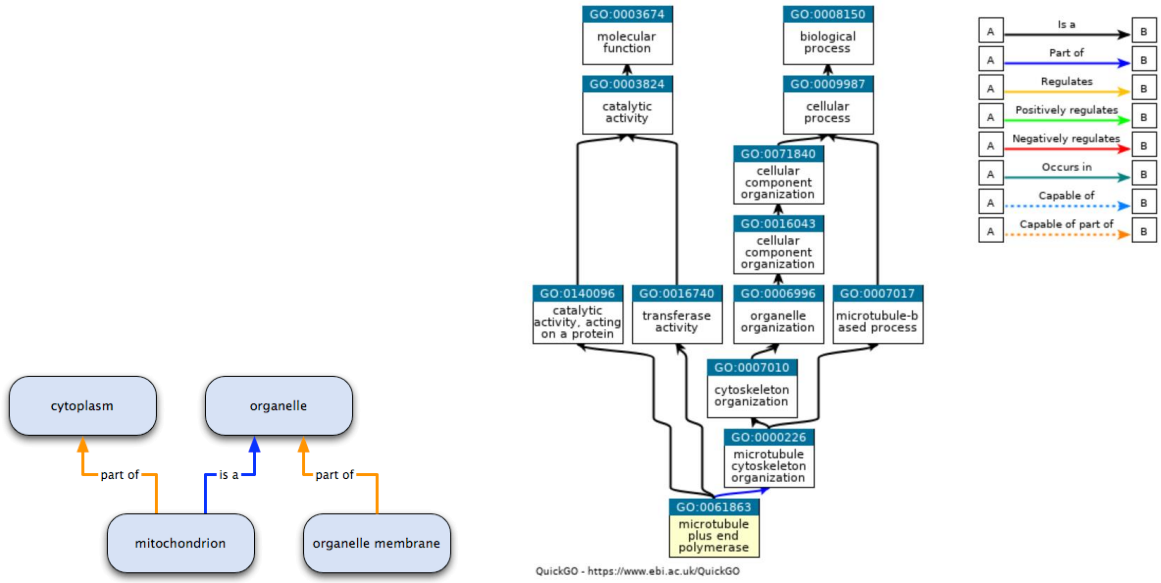
1.6 Gene ontology (GO) enrichment and pathway mapping

1.6.1 GO consortium (GOC)

GOC is a significant bioinformatics project that integrates information obtained from various sources to develop a structured controlled vocabulary (CV) to classify gene product function and location *i.e.* annotation of gene products (Huntley *et al.,* 2014). These annotations can be either experimental or computational, done by expert curators. the computationally annotated vocabularies constitutes >98% of the total annotations (Škunca *et al.,* 2012). At the highest level, these explicitly defined and structured vocabularies describe three principle information such as 1) the function of the gene product (molecular functions or MF), 2) their involvement into any biological processes (BP) and 3) their location *i.e.* cellular components (CC) (Hinderer *et al.,* 2019). The objective is to provide an extensive, publicly available resource of functional annotation of genes (Huntley *et al.,* 2014) and the functional impact of gene expression in the form of gene enrichment analyses (Hinderer *et al.,* 2019). These CV of these GO terms are assigned with a unique alphanumeric code (Dessimoz & Škunca, 2017). These unique codes are used to annotate genes and gene products in many other databases, including UniProt (Dimmer *et al.,* 2012) and Ensembl (Hubbard *et al.,* 2002 and Huntley *et al.,* 2014).

1.6.2 Gene ontology

GO is a hierarchical network of GO terms. Each term is a node, and these nodes are connected by edges (parent-child relationship) and describe how each term relates to one another. A child node can be connected to multiple parent nodes by edges. Some of the commonly used relationships in GO are "*is a"* (is a subtype of), "*part of", "has part", "regulates"* (*negatively and positively regulates) (Figure:1.13) ("Relations in the Gene Ontology").* GO terms provide general insight about the mechanism of expression changes due to differences in the condition. GO annotation refers to the functional annotation of genes with the respective GO terms (Zhou *et al.,* 2017).



(a) (b)

Figure 1. 13: A simple overview of parent-child relationship in GO. (a) Mitochondrion has two parents, it "is_an" organelle, and it is "part_of" the cytoplasm; organelle has two children: mitochondrion "is_an" organelle and organelle membrane is "part_of" organelle (Image courtesy: "Relations in the Gene Ontology"). (b) Acyclic graph of GO, each term has defined relationships to at least one term. This relation can disperse across the domain e.g. GO:0061863 is connected to both GO:0003674 (MF) and GO:0008150 (BP) (representation obtained from https://www.ebi.ac.uk/QuickGO/term/GO:0061863.

1.6.3 GO enrichment and DEG

Although the list of DEGs obtained from high-throughput sequencing is useful, it is not conveniently or directly insightful to recognize biology. Understanding the underlying biological process needs a functional profile of these gene sets. Thus, researchers can use enrichment analysis to get an understanding of the involved biological mechanisms that are differentially expressed due to the experimental conditions (Subramanian *et al.,* 2005). Enrichment analysis is an statistical approach that helps the researchers to determine the set of over and underrepresented genes by comparing the input gene sets with the reference gene set (Subramanian *et al.,* 2005 and Zhou *et al.,* 2017).

Enrichment analysis requires input files *i.e.* genes under experiment and a reference list. The reference list is important in enrichment analysis to define the background properly. For example, if the gene set comes from a particular strain of microbes than reference for that strain will generate more meaningful results than a reference from some other organism (Zhou *et al.,* 2017).

Gene enrichment is calculated by different statistical methods. Some of the popular methods used for this purpose are Fisher's exact test, chi-square test, hypergeometric distribution, and binomial distribution (Huang *et al.,* 2009). Except for binomial distribution, the rest of them are thought to be suitable for smaller population background (Khatri & Draghici, 2005). However, they have their limitations and weakness. Therefore, the selection of enrichment tools based on statistical methods will not be the best idea. Researchers should test different testing methods for their datasets and compare the results, which will help them make a biological conclusion with higher confidence (Zhou *et al.,* 2017).

There are several tools available such as AmiGO, OBO-Edit, and GOOSE; these are developed and supported by GO consortium. A large number of third-party tools such as Blast2GO (Conesa *et al.,* 2005) are also available that use data provided by the GO project (Zhou *et al.,* 2017). Blast2GO is now part of OmicsBox ("OmicsBox | BioBam | Bioinformatics Made Easy," 2019) project and seamlessly integrated as a Functional Analysis Module. It enables GO-based data mining on GO unannotated sequences data. This feature makes it suitable for the current research work since the FASTA sequences of DEGs are available. In addition to functional annotation it supports InterPro domains, RFAM IDs, enzyme codes (ECs), and KEGG maps. Graphical, and analytical tools are also present in Blast2GO for annotation

manipulation and datamining. A typical annotation of sequences in Blast2GO involves six steps: BLAST, InterPro scan, mapping, annotation, enrichment analysis and visualization. B2G uses BLAST (Altschul *et al.,* 1990) to find homologs to FASTA formatted input sequences (Figure 1.14) (Conesa *et al.,* 2005; GöTz *et al.,* 2008 and Zhou *et al.,* 2017).

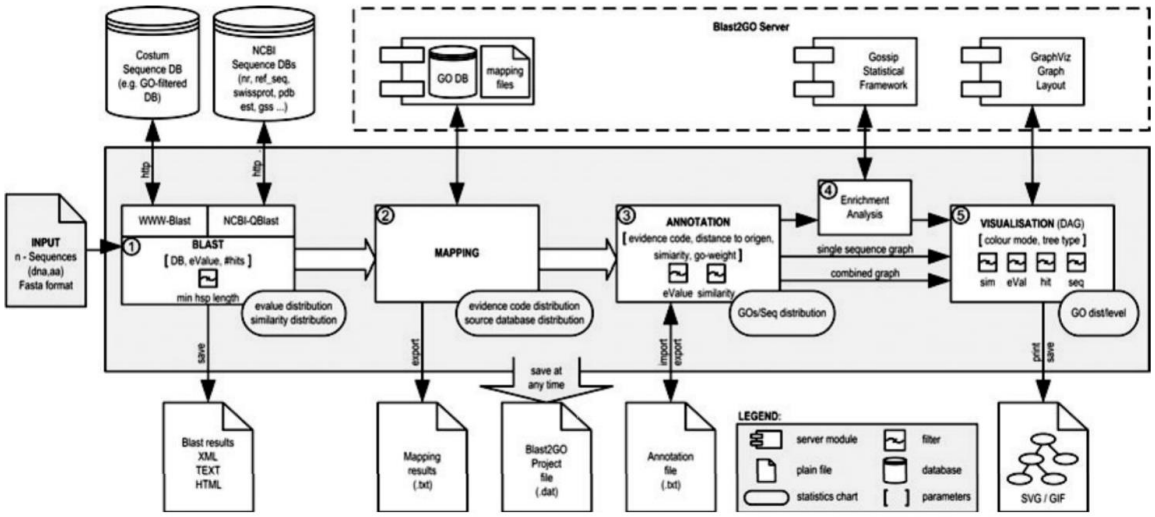


Figure 1. 14: Schematic workflow functional annotation by Blast2GO. Numbered circles denote major application steps. (1) selected sequences look for homology either in NCBI or custom database, (2) GO terms are mapped, (3) sequences are annotated according to the parameter set by the user, (4) user-defined statistical analysis (optional) e.g. GO term distribution, (5) annotation, and results can be visualized by GO DAG. Progress of the analysis can be observed and saved in each step and exported. Image courtesy: (Conesa et al., 2005)

1.7 DEG and pathway mapping

Genes function in a cooperative manner that means a gene may be one of the many genes involved in a specific pathway. From a given list of genes *e.g.* DEG, researchers may map them to known pathways. Pathway mapping of the DEGs helps to determine the biological process the enriched genes are involved in and, consequently, reduce the complexities and improve the confidence of the analysis. Pathway mapping is a good choice to understand biological phenomena that is hidden in a given gene/protein list (Zhou *et al.,* 2017). The Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, PANTHER, and BioCyc are some major metabolic pathway databases (Zhou *et al.,* 2017).

In KEGG database, genes are annotated with a KEGG orthology (KO) identifiers or K numbers that represents an orthologues group of genes and KEGG Automatic Annotation Server

(KASS) is a web-based service integrated to KEGG system that enables the reconstruction of KEGG pathways and BRITE hierarchies by assigning K numbers (based on sequence similarities) to the genes in the genome. At first, the query sequences (in FASTA format) are blasted against the reference sequence set (taken from KEGG database). The BLAST score and bi-directional best hit rate are computed and selected; those are above the threshold level. Orthologs are divided into KO groups according to the KEGG genes database. Then the KO groups are ranked based on the likelihood and heuristics. Finally, the K-number with the highest score is assigned to the query sequence (Moriya *et al.,* 2007).

2. Aims of the study

Aims of this study are:

a. Understand the unique Acl recycling system in *S. galilaeus* ATCC 31615.
b. Is there a transcriptome level reason for the overproduction of Acl B in the mutant strain HO42?

3. Materials and methods

The research work was carried out Antibiotic Biosynthetic Engineering lab located in Biocity, Turku, from September 2019 – June 2020. The RNA-Seq data was generated using the facility of The Finnish Functional Genomics Center (FFGC) located on the same premises. The mRNA library was prepared from 1000 ng total RNA, and the mRNA enrichment was done using a MICROBExpress™ Bacterial Enrichment Kit (ThermoFisher Scientific). From the fragmentation step, Illumina TruSeq® Stranded mRNA Sample Preparation Guide was followed for the library preparation. The first cDNA strand was synthesized from the fragmented mRNA, using transcriptase and random primer. dTTP was replaced by dUTP for strand specificity and used as a quencher for second strand amplification. To further improve strand specificity by allowing only RNA-dependent synthesis, Actinomycin D (Act D) was added to the reaction mixture. The synthesized cDNA was then ligated to the unique Illumina TruSeq indexing adaptors according to the protocol and enriched with PCR to create the final cDNA library. Advanced Analytical Fragment Analyzer was used to analyze the quality of the libraries and quantified by Qubit® Fluorometric Quantitation Life Technologies. The RNA-Seq library fragments were in the range of 200-700 bp, and the average size of the fragment was 250-350 bp were allowed for sequencing. Samples were sequenced with Illumina HiSeq3000 instrument with read length 1 x 50 bp. The above-mentioned experimental work was performed as a service by the FFGC.

All the analyses performed in the lab computer and the software used were either free or a trial version. Bioinformatics analysis of RNA-Seq data was divided into four categories: a) prediction of the Aclacinomycin encoding BGC by antiSMASH, b) detection of the DEGs, c) GO enrichment analysis, and d) pathway mapping of the DEGs. For DEGs identification, the web version of Chipster (Kallio *et al.,* 2011) software (version 4), which accommodates various tools required for high-throughput data analysis, was used. GO enrichment analysis of the DEGs was performed using Blast2GO version 5.2.5, which is integrated into Omicbox software (version 1.3). Finally, the DEGs were mapped to KEGG metabolic pathways by KASS webserver.

3.1 RNA-Seq dataset

The RNA-seq dataset used in this study was generated from WT and its mutant MT and in FASTQ format. Samples were taken at four-time points *i.e.* day 1, day 2, day 3 and day 4 from both strains. Data from WT was used as a control strain for the differential expression analysis.

3.2 Detection of Acl producing BGC

Acl encoding secondary metabolite was predicted and annotated using the web-based software antiSMASH 5.1.2 (Blin *et al.,* 2019). The reference genome sequence (unpublished) was uploaded in antiSMASH (https://antismash.secondarymetabolites.org/#!/start). The detection strictness was left in "relaxed" mode, and detection of all the extra features was turned ON. It uses profile Hidden Markov Model (pHMM) based on multiple sequence alignments of experimentally characterized protein or protein domains for BGC identification (Blin *et al.,* 2017). After submitting the genome sequence, the server predicts different secondary metabolite-producing clusters throughout the genome.

3.3 Identification of DEGs

Data analysis begins with the input of the raw read files and the reference files. The read files were uploaded in the Chipster system. The read files were unzipped and extracted to FASTQ format and underwent a quality control step by FastQC (Andrews, 2010). Raw data delivers immediate insight into the sequence quality for further downstream analysis. After checking the quality parameters, for example GC content which indicates any possible contamination in the raw data and eventually has great importance in transcript detection and abundance quantification, The reads from each sample were aligned with the reference genome sequence (WT) by the Bioconductor package Bowtie2 (Langmead & Salzberg, 2012). All the settings for alignment were left as default.

This tool returns a BAM file that contains the alignment and an index file for it (.bai). It also produces a log file. This log file shows the result of the alignment i.e., percentage of the aligned (unique and non-unique) and non-aligned reads on a particular setting. After completing the alignment, reads that were mapped to the feature of interest were counted. The mapped reads in the BAM files were then counted against the genes by HTSeq (Anders *et al.,* 2015). To do so, a GTF (Gene transfer format) file was provided. This tab limited text file contains the information about gene structure. This tool calculated how many reads were

mapped to the genes and makes two output; one is a tab limited file (.tsv) which contain the number of reads on each gene and a .txt file which mainly shows the how many reads were counted and not counted. Some of the parameters were changed from the default settings such as “Is the data stranded and how” changed to “reverse” which means the second read of a pair should map to the same strand of the gene, minimum alignment quality was set to 1 to increase the number of counted reads and the feature type was changed to CDS since the data came from a prokaryotic organism**.**

Before performing the DEG analysis, count files were merged using “Define NGS experiment” tools. A phenodata file was also created by this tool. For DEGs between the strains, the count files of WT and MT from the same time points were combined, and for the analysis of DEGs within the strains, count files from D1-D2, D2-D3 and D3-D4 were combined. Therefore, the read counts were ready for pairwise comparison DEG of the two groups. The pairwise DE analysis allows the identification of DE genes in a pairwise comparison of two different experimental conditions. To perform this comparison, Bioconductor package edgeR (Robinson *et al.,* 2010) was used. Here all the parameters were unchanged except the dispersion value, which was set to 0.01.

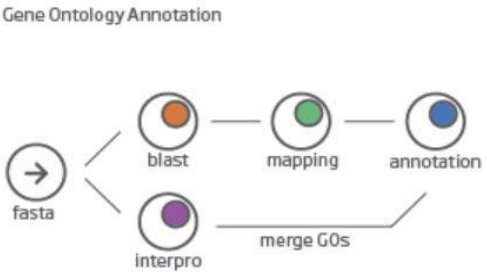
The expression differences between different conditions compared with edgeR created a list of genes that are associated with Log_2 fold change (LogFC), log- counts per million (log-CPM), p-values, and false discovery rate (FDR). All the parameters were set to default except the “dispersion value” which was set to 0.01 because the samples were pooled, and variability was counted. Fold change (FC) is often used to measure the change in gene expression level in gene expression analysis by microarray, RT-qPCR, and RNA-seq. The LogFC can be defined as the log ratio of the genes/transcript's expression values in two different conditions. LogFC is used for better scaling of fold changes. In the present study, transcripts with LogFC < -1 as down-regulated and transcripts with LogFC > 1 were characterized as up-regulated. Counts Per Million (CPM) is a filter to exclude genes with low counts across libraries. Log CPM is the log counts per million, which is measuring relative abundance of expression normalized for library size ad transcript length. The False Discovery Rate (FDR) controls the number of false discoveries in significant discovery (*i.e.* a significant result) that determines adjusted p-values for each test. In this study, the number of DE genes was detected by using the FDR < 0.05, which infers less than 5% of significant tests will result in false positives. To make the analysis simpler, I had selected 5 DEGs up and 5 DEGs downstream from the Acl producing cluster

detected by antiSMASH, including the genes residing in the cluster. In addition to these genes, ten most up-regulated and ten most down-regulated genes on different days were analyzed.

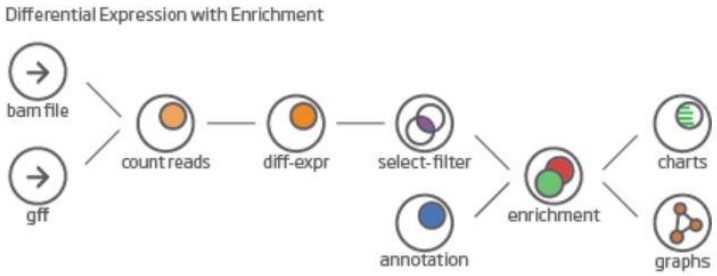
The results from this DE analysis was used to determine the genes regulations possibly associated with the Acl gene cluster. Furthermore, a tab limited files of these DEGs were sorted according to LogFC value and exported. A python script (provided by Keith Yamada M.Sc.) was used to extract the corresponding FASTA sequence of these sorted DEGs. These FASTA sequences later used for GO enrichment analysis and KEGG pathway mapping.

3.4 GO enrichment analysis:

The amino acid sequences in the FASTA file created from DEG analysis were sorted according to the log_2FC value. GO enrichment analysis of these up and down-regulated genes was done by the functional analysis module of Omicsbox software in two steps: gene ontology annotation and gene ontology enrichment (Figure 2.1). Omicsbox uses Blas2GO methodology (Conesa *et al.,* 2005) for this purpose by following steps. The FASTA sequences of the DEGs at different time points were uploaded into the module. After that NCBI QBlast service (non-redundant) was used to blast those sequences against protein database. BlastP algorithm was used since the query sequence were amino acid sequence. The taxonomic filter was set to *Streptomyces* to widen the specificity of the annotation. The purpose of this step was to find sequences similar to the query set. The sequences were mapped by retrieving blast hit associated GO terms. By applying an annotation rule ("Gene Ontology Annotation"), the GO terms were selected from the GO pool obtained in the mapping step. After that, the selected GO terms were assigned to the query sequence. InterPro annotation helps to retrieve domain/motif information of the amino acid sequences. The sequences were annotated with the corresponding GO terms merged with already existing GO terms. For InterPro annotation, EMBL-EBI InterPro web-service was used. The sequence annotation results were then exported as .annot format and used as reference set for enrichment analysis (Figure 3.1a).



(a)



(b)

Figure 3. 1: Gene enrichment analysis workflow of the DEGs. a) the annotation of the DEGs and b) gene enrichment analysis with the help of a GO annotated file. Image curtesy: *(OmicsBox | BioBam | Bioinformatics Made Easy, 2019)*

For enrichment analysis *(Fisher's Exact Test),* Blast2GO employs FatiGO package (Al-Shahrour *et al.,* 2004). In the gene enrichment step (Figure 3.1b), the tab limited files (.tsv) created in the DEG analysis were uploaded, and the previously created ".annot" file was used as a reference set. Then the overrepresentation of the up and downregulated genes was tested for gene enrichment. The filter value was set to $p < 0.05$.

3.5 KEGG pathway mapping of DEGs:

The DEGs were mapped to different pathways by using KAAS. The FASTA sequences of the up-regulated and downregulated genes were functionally annotated by BLAST (https://www.genome.jp/kaas-bin/kaas_main?mode=partial) against the manually curated KEGG GENEs database. *Streptomyces coelicolor, Streptomyces avermitilis, Streptomyces griseus, Streptomyces scabiei and Streptomyces noursei* were selected as a template data set for KO assignment to find the ortholog of the corresponding amino acid sequences by single-directional best hit (SBH).

4. Results

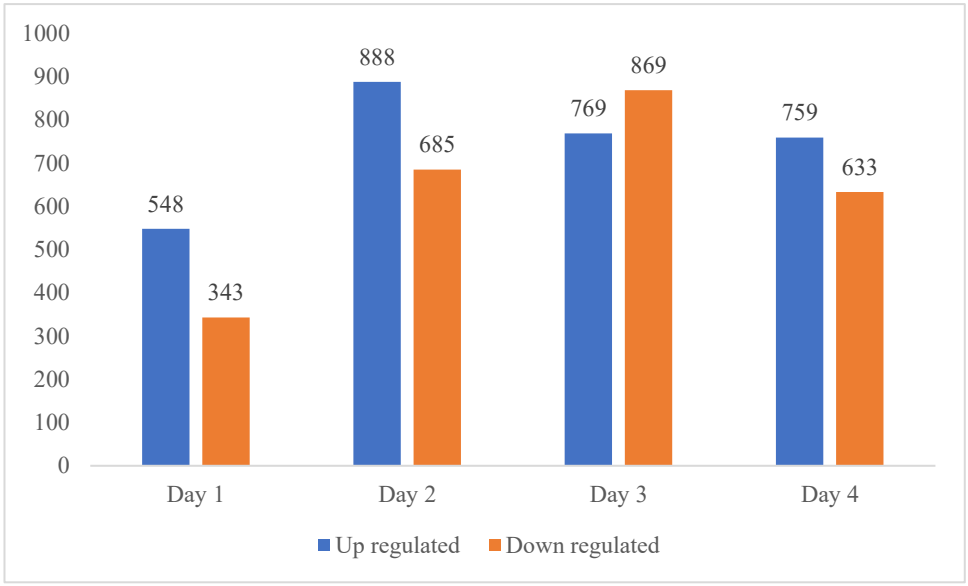
4.1. Analysis of the transcriptomic data

In this study, RNA-Seq data obtained from WT and MT was compared between the strains (WT vs. MT) and within the strains (time points). Transcriptomic data from WT and MT were compared with each other obtained at the same sampling time. Measurement of DE within the strain came from the transcriptomic data of a strain with 1-day intervals. WT considered as a control group during between the strain comparison. In contrast, there was no direct control while comparing within the strain, but a comparative analysis has made by comparing WT vs MT at the same time interval *i.e.* D1-D2, D2-D3 and D3-D4.

The number of raw reads obtained from each sample ranged from 13.5 to 18.3 million single end fragments (one read per fragment). The average number of reads from WT and MT were 15.3 and 16.4 million, respectively. Bowtie2 aligned those reads with the WT genome at the rate ranges between 96.21% to 98.14%, respectively. This result indicated the consistency of the mapping of reads, especially with regards to the mutations in the MT (Table 4.1).

There are 8807 annotated genes in the WT genome. The number of significantly ($P < 0.05$) DEGs from day 1 to day 4 were 891, 1573, 1638, and 1392 respectively, after comparing the RNA-seq data from two different strains. It corresponds to 10.11-18.59% of the total genes (Figure 4.1a). Later the DEGs were compared within the strain but on different days, *i.e.* day 1 vs day 2, day 2 vs day 3 and day 3 vs day 4 (Figure 4.1b). A list of 20 most up and down-regulated genes along with their functions is in appendix 1.

a)



b)

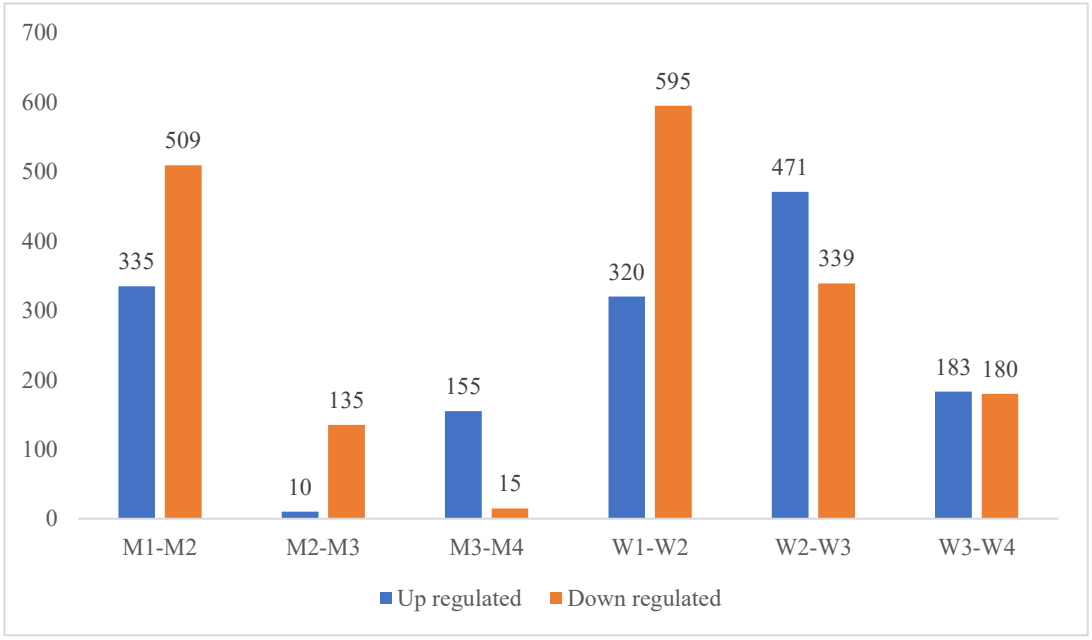
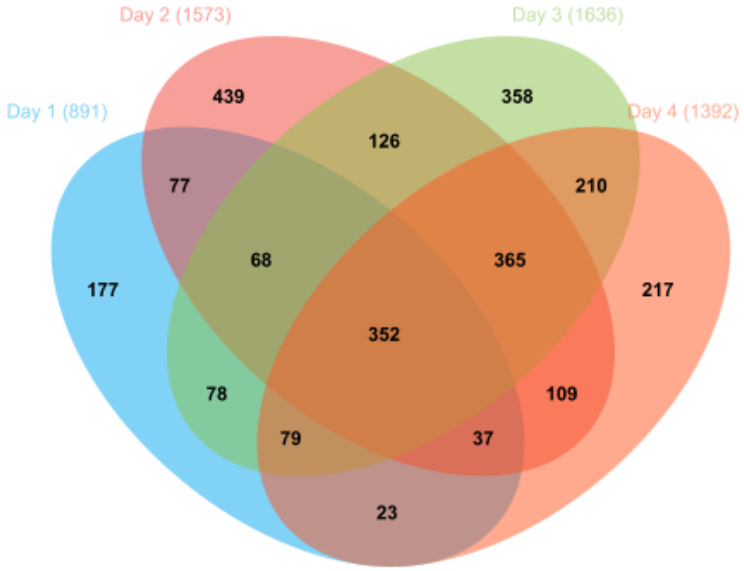
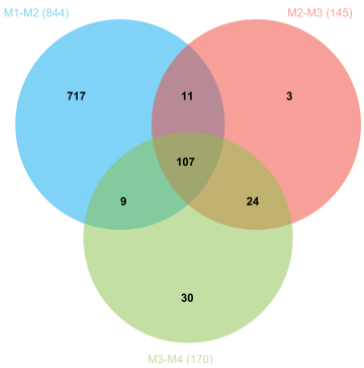


Figure 4. 1: Number of DEGs. a) compared between WT and MT and b) compared within the strain (in X-axis M/W denotes MT/WT and the number 1-4 denotes the days).

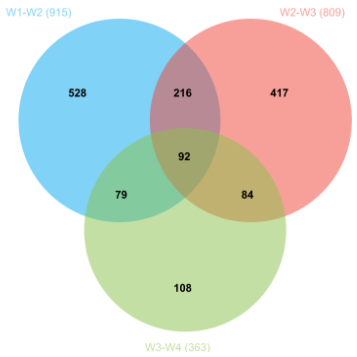
It is important to note that many genes overlapped over time among the DEGs mentioned above. The number of unique and overlapping genes from the different DEG pools is summarized below (Figure 4.2).


Day 2 (1573)
Day 3 (1636)
Day 1 (891)
Day 4 (1392)
439
358
126
77
210
68
365
177
217
352
78
109
79
37
23

(a)


M1-M2 (844)
M2-M3 (145)
717
11
3
107
9
24
30
M3-M4 (170)

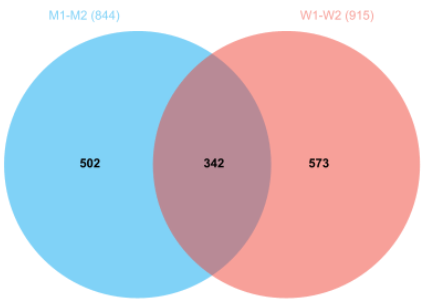
(b)


W1-W2 (915)
W2-W3 (809)
528
216
417
92
79
84
108
W3-W4 (363)

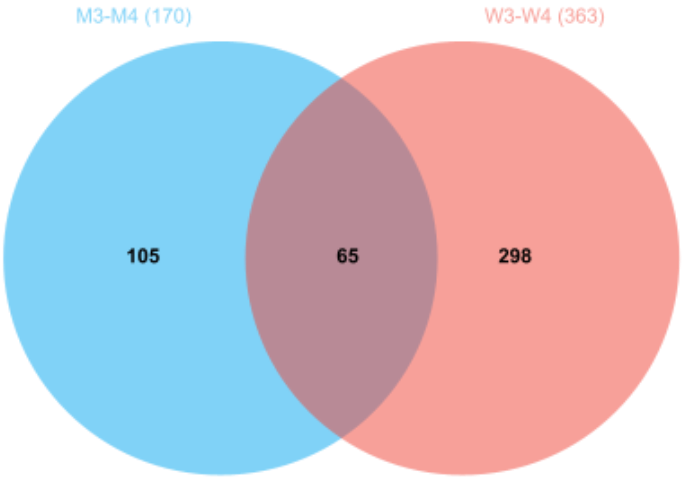
(c)


M2-M3 (145)
W2-W3 (809)
45
100
709

(d)


M1-M2 (844)
W1-W2 (915)
502
342
573

(e)



(f)

Figure 4. 2: Venn diagram of the number of unique and overlapping genes between/among different analyzed time points. a) represents DEGs between the strains, b) and c) represents the DEGs within the strain, and d), e) and f) represents the number of shared DEG at the same time points of MT and WT. The area of the diagrams does not represent the number of genes. W: Streptomyces galilaeus ATCC 31615 and M : Streptomyces galilaeus ATCC 31615 HO42; adjacent numerical value 1-4 represents the day of sampling.

Table 4. 1: Summary of the RNA-Seq reads for each sample. WT: Streptomyces galilaeus ATCC 31615, sample name refers to corresponding BAM file.

Strain	**Sample**	**Time point (day)**	**Number of times a read aligned with the WT genome**						**Total reads (millions)**	**Overall Alignment Rate**
			0		1		>1			
			No. Reads	%	No. Reads	%	No. Reads	%		%
MT	D1_42_1_S29_L003_R1_001.fastq	D1	429421	2.74	305736	1.95	14918507	95.3	15.6	97.26
	D2_42_2_S31_L003_R1_001.fastq	D2	366202	2.21	1211775	7.31	14989622	90.48	16.5	97.79
	D3_42_1_S33_L003_R1_001.fastq	D3	659621	3.61	1443887	7.91	16161084	88.48	18.3	96.39
	D4_42_3_S35_L003_R1_001.fastq	D4	398846	2.63	1217474	8.02	13573521	89.36	15.2	97.37
	Average		463522.5	2.79	1044718	6.29	14910683.5	90.9	16.4	97.25
WT	D1_WT_1_S28_L003_R1_001.fastq	D1	536348	3.25	837282	5.08	15118620	91.67	16.5	98.14
	D2_WT_3_S30_L003_R1_001.fastq	D2	595101	3.79	1695412	10.81	13396392	85.4	15.7	96.21
	D3_WT_2_S32_L003_R1_001.fastq	D3	400936	2.61	2738359	17.83	12215830	79.56	15.3	97.39
	D4_WT_3_S34_L003_R1_001.fastq	D4	350805	2.59	3200015	23.67	9970901	73.74	13.5	97.41
	Average		470797.5	3.06	2117767	14.3	12675435.8	82.5	15.3	97.28

4.2 Acl producing BGC prediction

The web application antiSMASH (version 5.1.2) had predicted 31 genomic regions (Table 4.2) involved in secondary metabolism across the WT genome, of which eight regions show 100% similarity with the known BGCs. This strain is best known for aclacinomycin production, which corresponds to region 13 (Table 4.2). This predicted BGC consists of 71 genes from 2133094 to 2204805 bp of the genome (Appendix 2).

Table 4. 2: Predicted genomic regions of WT for secondary metabolites production. LAP: Linear azol(in)e-containing peptides, hgIE-KS: heterocyst glycolipid synthase-like PKS, T: type.

Regions	Type	From	To	Most similar known cluster	Similarity
Region 1	LAP	44800	66996		
Region 2	Terpene	296516	316583		
Region 3	T1PKS, NRPS, Lassopeptide	389139	443213	Citrulassin D	100%
Region 4	T1PKS	491642	535706	Herboxidene	7%
Region 5	Terpene	603901	628502	Isorenieratene	100%
Region 6	Melanin	1143964	1152805	Melanin	57%
Region 7	Bactericin	1308299	1316450	Informatipeptin	42%
Region 8	hgIE-KS, T1PKS	1586530	1637245	Miharamycin A/B	11%
Region 9	Terpene	1669117	1695582	Hopene	92%
Region 10	Indole	1861238	1882386	5-isoprenylindole-3-carboxylate β-D-glycosyl ester	23%
Region 11	Oligosaccharide	2046082	2091954	Meiingmycin	5%
Region 12	Siderophore	2116328	2128615	Grincamycin	8%
Region 13	T2PKS	2133094	2204805	Cinerubin B	100%
Region 14	Terpene, butyrolactone	2308100	2329493	y-butyrolactone	100%

Region 15	Bactericin	2358421	2369752		
Region 16	Lassepeptide	2430086	2452607	SSV-2083	36%
Region 17	Siderophore	2613627	2525119		
Region 18	Terpene	3195093	3215132	Albaflavenone	100%
Region 19	NRPS	5131759	5185615	Ishigamide	66%
Region 20	Siderophore	5638454	5650259	Desferrioxamine	83%
Region 21	Melanine	5753528	5764088	Istamycin	4%
Region 22	NRPS	6073670	6137868	SCO-2138	71%
Region 23	transAT-PKS, NRPS	6313466	6395279	Leinamycin	12%
Region 24	Ectoine	6871301	6881705	Ecotine	100%
Region 25	T2PKS	7608932	7681447	Spore pigment	83%
Region 26	NRPS, T1PKS	7889667	7955546	Foxicins A-D	14%
Region 27	T3PKS	7980839	8022023	Germicidin	100%
Region 28	NRPS, T1PKS	8169813	8241883	Foxicins A-D	12%
Region 29	NRPS	8302091	8352417	Coelichelin	100%
Region 30	Terpene	8610846	8632924		
Region 31	NRPS	8712662	8824130	Enduracidin	14%

antiSMASH identified different regions of a BGC based on their corresponding function for BGC synthesis, *e.g.* core biosynthetic genes (CBG), additional biosynthetic genes (ABG), transport-related genes (TG), regulatory genes (RG) and other genes (OG) (Figure 4.3). The length of the genes and their functions are in appendix 2. Functional roles have been collected from the GenBank file of WT. Additionally, some functions had been deduced by blastP.

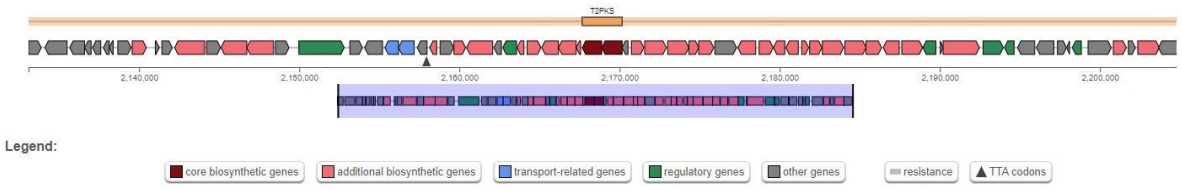


Figure 4. 3: Predicted genes in Acl producing BGC. The color of the legends shows the function of the predicted genes in the BGC and their direction.

4.2.1 DEGs from the predicted Acl producing cluster

The differential expression may vary since the RNA samples of three biological replicates were pooled together. Thus, the number of replicates for transcript analysis is one (N=1) and after that the significantly ($P < 0.05$) DEGs within and between the WT and MT were calculated. The comparative transcript count level within the strain revealed the significant ($P < 0.05$) DEGs between the transcriptome on different days. The level of differential expression was determined by the LogFC value ($P < 0.05$). A '+' value means up-regulation, whereas a '-' value denotes downregulation.

4.2.1.1 DEGs within the strains

The comparative DGE analysis (Table 4.3) demonstrated a pattern of expression, *i.e.,* most of the genes were downregulated at the first interval (D1-D2) and slightly upregulated during the third interval (D3-D4). In general, genes from the WT were more downregulated than their MT counterpart. However, WT showed slightly higher upregulation than MT strain.

4.2.1.1.1 DEGs within the WT strain

From the 72 genes that consist of the predicted Acl BGC, the number of significantly ($P < 0.05$) DEGs were 49 during the first interval. The number of the upregulated genes was 3 and the rest of the 46 genes were downregulated. There were only 4 DEGs ($P < 0.05$) during the second interval. There were only 4 significant ($P < 0.05$) DEGs during the second interval (D2-D3). Finally, there were 32 genes those are significantly ($P < 0.05$) differentially expressed during the third interval and all of them were upregulated (Table 4.3).

4.2.1.1.2 DEGs within MT strain

MT displayed a relatively similar pattern of differential gene expression. However, the degree of DE was different. There were 45 significantly ($P < 0.05$) DEGs during the first interval, of which only one was upregulated. The number of DEGs ($P < 0.05$) was lowered to 19 genes, and all of them were downregulated. However, the degree of downregulation was lower than the first interval and this direction of positive regulation was continued in the third interval. The number of DE genes that were significant is 16 ($P < 0.05$), and all of them were upregulated but their degree of upregulation was lower than WT at the same interval .

Table 4. 3: DEGs (P < 0.05) within the predicted Acl producing BGC of WT and MT strains on different days. DEs were determined within the strain. Genes without showing any LogFC value didn't differentially expressed (P < 0.05) and aren't shown. +/- denotes up/downregulation, respectively. D1, D2, D3, D4 represents the sampling day. DEGs with -1>LogFC>1 value was presented only.

Gene ID	LogFC					
	WT			MT		
	D1-D2	D2-D3	D3-D4	D1-D2	D2-D3	D3-D4
fig|33899.16.peg.2244	-2.039			-1.817		
fig|33899.16.peg.2245	-1.922			-2.178		
fig|33899.16.peg.2246	-1.934	-1.000		-3.432		
fig|33899.16.peg.2247	-2.345		1.165			
fig|33899.16.peg.2253	2.476	3.343		10.925	-1.592	1.160
fig|33899.16.peg.2256	-4.719			-5.935		
fig|33899.16.peg.2257	-5.527			-6.803		
fig|33899.16.peg.2260	2.027		1.546	-4.237		
fig|33899.16.peg.2261				-1.063		
fig|33899.16.peg.2264	-2.118			-3.113		
fig|33899.16.peg.2265	-3.947		1.696	-4.303		
fig|33899.16.peg.2266	-4.348	-1.015	2.437	-3.060	-1.126	
fig|33899.16.peg.2267	-4.132		1.888	-3.035	-1.091	1.040
fig|33899.16.peg.2268	-1.941			-1.803		
fig|33899.16.peg.2269	-2.859			-2.826		

fig|33899.16.peg.2270	-5.045		2.286	-3.241		
fig|33899.16.peg.2271	-5.388		2.402	-3.626		
fig|33899.16.peg.2273	-5.035		1.982	-2.800		
fig|33899.16.peg.2274	-3.700		1.955	-2.575	-1.027	
fig|33899.16.peg.2275	-4.527		1.764	-3.116		
fig|33899.16.peg.2276	-4.602		2.237	-3.193	-1.192	1.353
fig|33899.16.peg.2277	-4.637		2.248	-3.052	-1.227	1.203
fig|33899.16.peg.2278	-4.862		2.494	-3.618	-1.089	1.131
fig|33899.16.peg.2279	-4.214	-1.190	2.217	-3.292		
fig|33899.16.peg.2280	-4.654		2.060	-3.559	-1.202	1.136
fig|33899.16.peg.2281	-5.390		2.259	-3.705	-1.108	1.082
fig|33899.16.peg.2282	-5.506		2.573	-3.664	-1.113	1.243
fig|33899.16.peg.2283	-5.493		2.340	-3.181	-1.319	1.387
fig|33899.16.peg.2284	-5.058		2.504	-3.545	-1.260	1.081
fig|33899.16.peg.2285	-4.775		2.177	-3.243	-1.217	1.339
fig|33899.16.peg.2286	-4.364		2.238	-2.648	-1.233	1.078
fig|33899.16.peg.2287	-4.853		2.493	-3.252	-1.259	1.413
fig|33899.16.peg.2288	-4.272		2.021	-2.816		
fig|33899.16.peg.2289	-4.333		2.203	-2.873		1.084
fig|33899.16.peg.2290	-4.445		1.633	-4.092		

fig|33899.16.peg.2291	-5.096	2.363	-3.739		
fig|33899.16.peg.2292	-4.908	2.175	-3.603	-0.993	
fig|33899.16.peg.2293	-5.279	2.202	-3.406	-1.206	1.014
fig|33899.16.peg.2294	-4.394	1.619	-3.099	-1.040	
fig|33899.16.peg.2295	-3.779	2.181	-2.560		
fig|33899.16.peg.2296	-2.565				
fig|33899.16.peg.2297	-4.963	1.904	-3.835	-1.115	1.268
fig|33899.16.peg.2298	-3.351	1.151	-1.945		
fig|33899.16.peg.2299	-2.564		-1.547		
fig|33899.16.peg.2300	-2.734		-1.492		
fig|33899.16.peg.2301	-3.113		-2.977		
fig|33899.16.peg.2302	-1.125		-1.325		
fig|33899.16.peg.2307	1.227				
fig|33899.16.peg.2314	-1.999				

4.2.1.2 DEG of the predicted Acl producing BGC between WT and MT

The same transcriptomic data mentioned above was also analyzed to observe the differential gene expression between the strains. During the growth period, of the 72 genes that are involved in the predicted BGC, 40 (D1), 41 (D2), 41 (D3), and 37 (D4) genes were differentially expressed. The percentages of the DEGs against the number of predicted genes in the BGC for Acl synthesis were between 31.94 % to 52.77% . However most of the DEGs

were upregulated and the percentages of upregulated genes on different sampling days were 93.33% (D1), 91.89% (D2) 89.20% (D3), and 77.30% (D4) (Table 4.4).

Table 4. 4: DEGs (P<0.05) within the predicted Acl producing BGC of WT and MT strains on different days. Differential expression was determined between the strains from their transcriptomic data obtained on the same days. Genes without showing any LogFC value were not differentially expressed (P<0.05) and aren't mentioned. +/- denotes up/down-regulation, respectively. D1, D2, D3, D4 represents sampling day. DEGs with -1>LogFC>1 value was presented only.

Gene ID	LogFC			
	D1	D2	D3	D4
fig|33899.16.peg.2244	-1.005		-1.391	
fig|33899.16.peg.2245		-1.199		-1.020
fig|33899.16.peg.2246		-1.834		-1.169
fig|33899.16.peg.2247	-1.342			
fig|33899.16.peg.2253		2.708	-2.274	-1.293
fig|33899.16.peg.2254			1.853	
fig|33899.16.peg.2256	1.387			
fig|33899.16.peg.2257	1.512			
fig|33899.16.peg.2260	3.146	-3.211	-4.459	-4.921
fig|33899.16.peg.2261			-1.030	-1.366
fig|33899.16.peg.2264	1.777			
fig|33899.16.peg.2265	1.803	1.363	1.467	
fig|33899.16.peg.2266	1.219	2.423	2.264	0.807

fig|33899.16.peg.2267	1.334	2.348	1.919	1.084
fig|33899.16.peg.2270		2.297	1.958	
fig|33899.16.peg.2271		2.463	2.215	
fig|33899.16.peg.2273		2.957	2.145	1.072
fig|33899.16.peg.2274	1.093	2.135	1.772	
fig|33899.16.peg.2275	1.221	2.548	1.930	
fig|33899.16.peg.2276	1.252	2.577	2.312	1.442
fig|33899.16.peg.2277	1.125	2.626	1.939	
fig|33899.16.peg.2278	1.543	2.703	2.305	
fig|33899.16.peg.2279	1.427	2.266	2.621	1.337
fig|33899.16.peg.2280	1.524	2.535	2.062	1.151
fig|33899.16.peg.2281	1.511	3.112	2.299	1.135
fig|33899.16.peg.2282	1.191	2.949	2.410	1.093
fig|33899.16.peg.2283		3.207	1.888	
fig|33899.16.peg.2284	1.414	2.844	2.465	1.055
fig|33899.16.peg.2285	1.312	2.760	2.103	1.278
fig|33899.16.peg.2286		2.358	1.688	
fig|33899.16.peg.2287	1.633	3.150	2.235	1.168
fig|33899.16.peg.2288		2.365	2.346	1.326
fig|33899.16.peg.2289	1.074	2.451	2.178	1.073

fig|33899.16.peg.2290	1.850	2.118	1.873	1.018
fig|33899.16.peg.2291	1.182	2.455	2.054	
fig|33899.16.peg.2292	1.050	2.271	1.865	
fig|33899.16.peg.2293	1.065	2.855	2.016	
fig|33899.16.peg.2294	1.150	2.361	1.649	
fig|33899.16.peg.2295		1.717	1.747	
fig|33899.16.peg.2297	1.337	2.381	1.898	1.274
fig|33899.16.peg.2298		2.289	1.941	1.299
fig|33899.16.peg.2299	1.123	2.057	2.296	1.598
fig|33899.16.peg.2300		1.710	1.667	
fig|33899.16.peg.2303	1.746	2.306	1.911	1.896

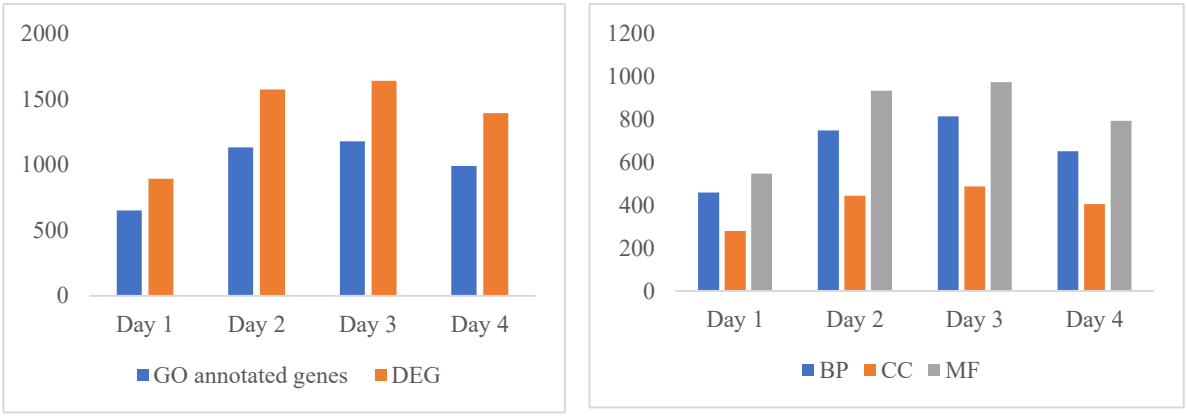
4.3 Functional annotation and GO enrichment analysis

4.3.1 Functional annotation based on GO

DEGs involved in MF, BP, and CC were identified by GO analysis. The DEGs from the between strain comparison was GO annotated and submitted to WEGO for classification. The number of GO annotated DEGs were 651 (73.06%) on day 1, 1132 (71.96%) on day 2, 1178 (71.92%) on day 3 and 989 (71.05%) on day 4 (Figure 4.4).

These GO annotated DEGs were classified into 36 GO categories at GO level 2. The distribution of the functional terms was as follows: 19 terms for BP, 9 terms for CC, and 8 terms for MF. In the CC category, cell, cell part, membrane, and membrane categories were the top four regarding the number of DEGs. In the MF category, the DEGs were primarily involved in catalytic activity and binding. Finally, metabolic process, cellular process, localization were the top two terms in the BP category, from day 1 to day 4 (Figure 4.5). The

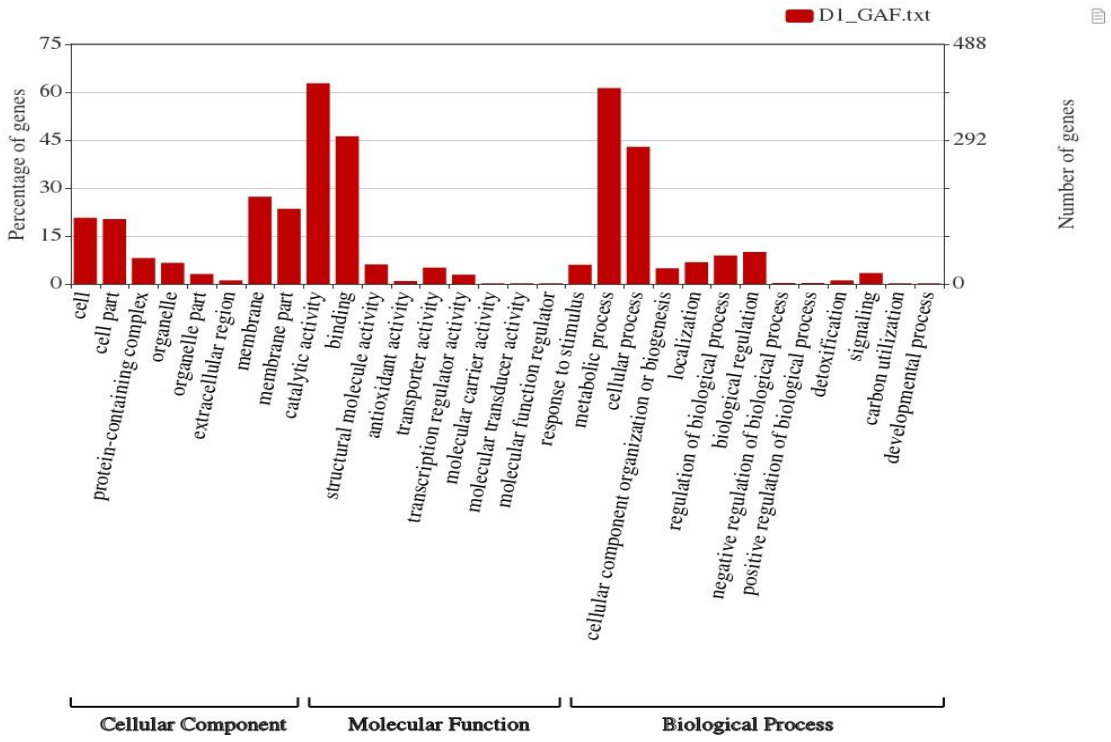
GO terms mentioned above were at level 2. According to the frequency of a GO term, the top 30 most specifically annotated GO terms and their corresponding DEG numbers were presented in appendix 3.



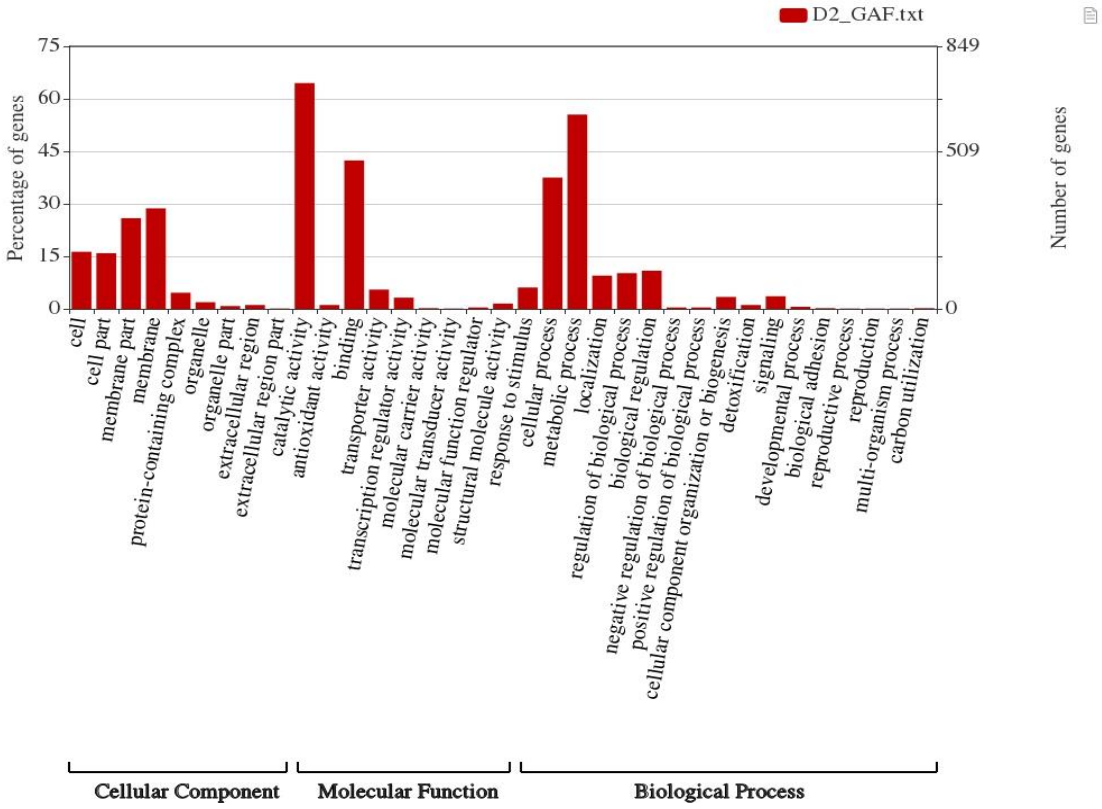
a) b)

Figure 4. 4: GO annotation of the DEG between the strains. a) the number of DEGs and the number of GO annotated DEGs, b) the categories of annotated GO to those DEGs. One single DEG can get multiple GO annotations.

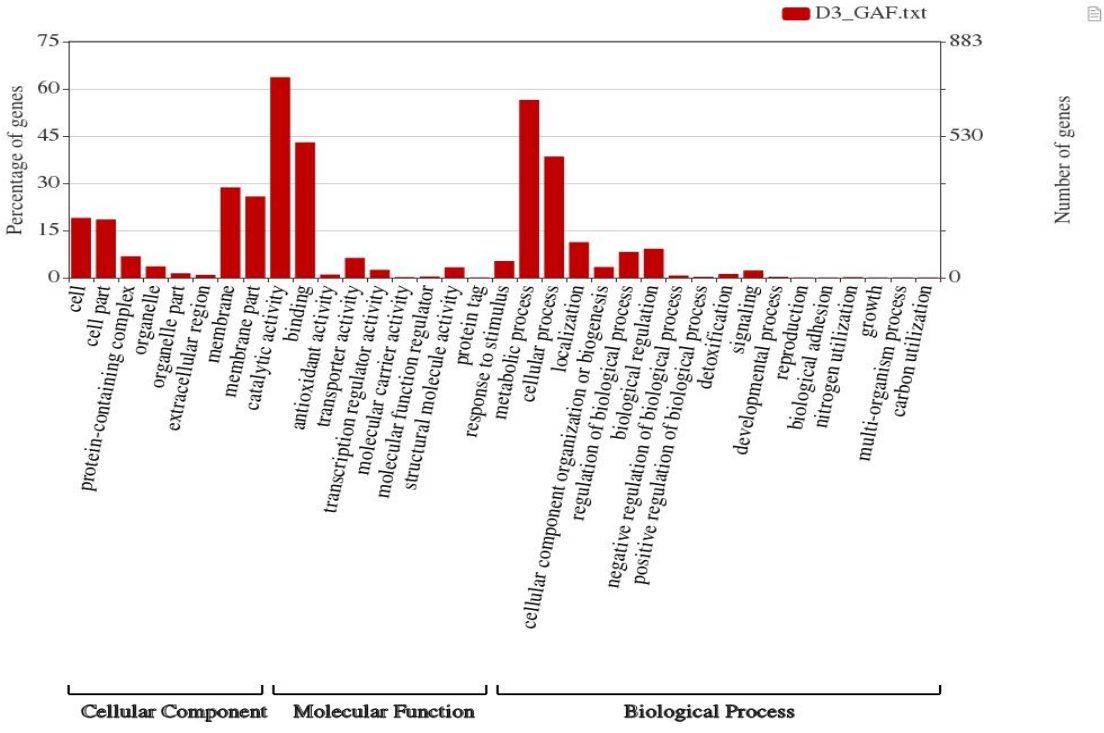
a)

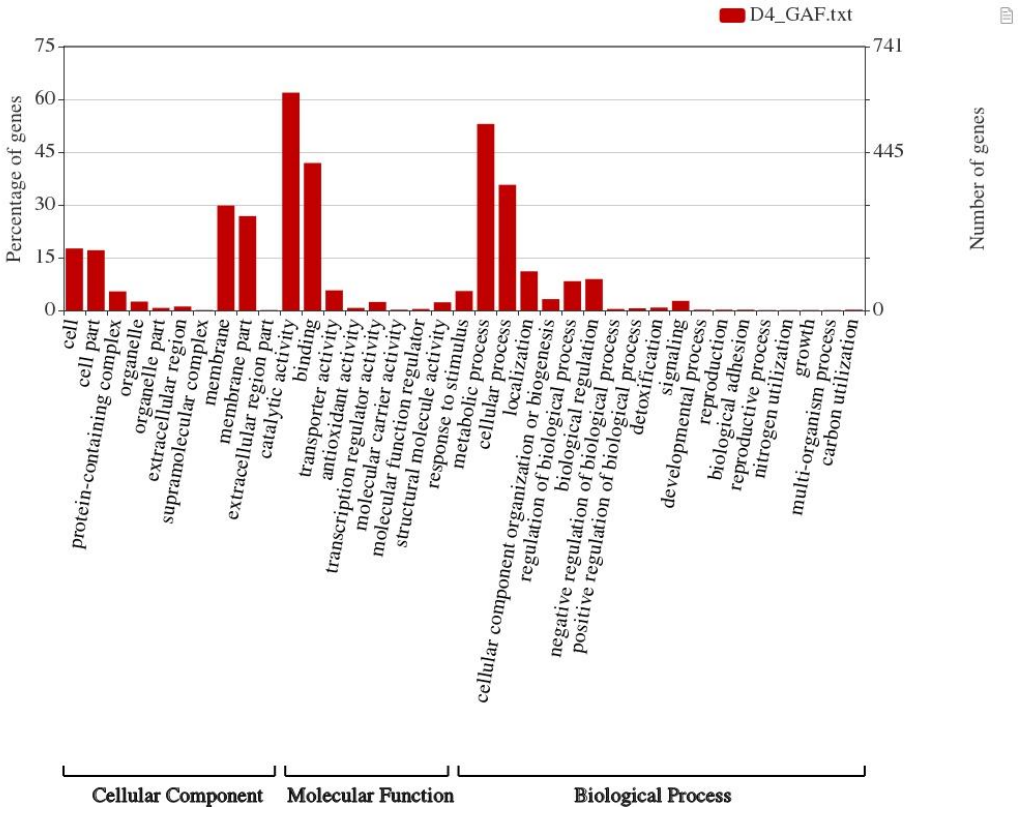


b)


D2_GAF.txt
Percentage of genes
Number of genes
75
60
45
30
15
0
849
509
0
cell
cell part
membrane part
membrane
protein-containing complex
organelle
organelle part
extracellular region
extracellular region part
catalytic activity
antioxidant activity
binding
transporter activity
transcription regulator activity
molecular carrier activity
molecular transducer activity
molecular function regulator
structural molecule activity
response to stimulus
cellular process
metabolic process
localization
regulation of biological process
biological regulation
negative regulation of biological process
positive regulation of biological process
cellular component organization or biogenesis
detoxification
signaling
developmental process
biological adhesion
reproductive process
reproduction
multi-organism process
carbon utilization
Cellular Component
Molecular Function
Biological Process

c)


D3_GAF.txt
Percentage of genes
Number of genes
75
60
45
30
15
0
883
530
0
cell
cell part
protein-containing complex
organelle
organelle part
extracellular region
membrane
membrane part
catalytic activity
binding
antioxidant activity
transporter activity
transcription regulator activity
molecular carrier activity
molecular function regulator
structural molecule activity
protein tag
response to stimulus
metabolic process
cellular process
cellular component organization or biogenesis
localization
regulation of biological process
biological regulation
negative regulation of biological process
positive regulation of biological process
detoxification
signaling
developmental process
reproduction
biological adhesion
nitrogen utilization
growth
multi-organism process
carbon utilization
Cellular Component
Molecular Function
Biological Process



d)

Figure 4. 5: GO categories of the DEGs between the strain comparison on different days. a) day 1, b)day 2, c) day 3 and d) day 4. The y-axis on the left represents the percent of the GO annotated DEGs that belong to a GO term, and the on the right side represents the number of DEGs.

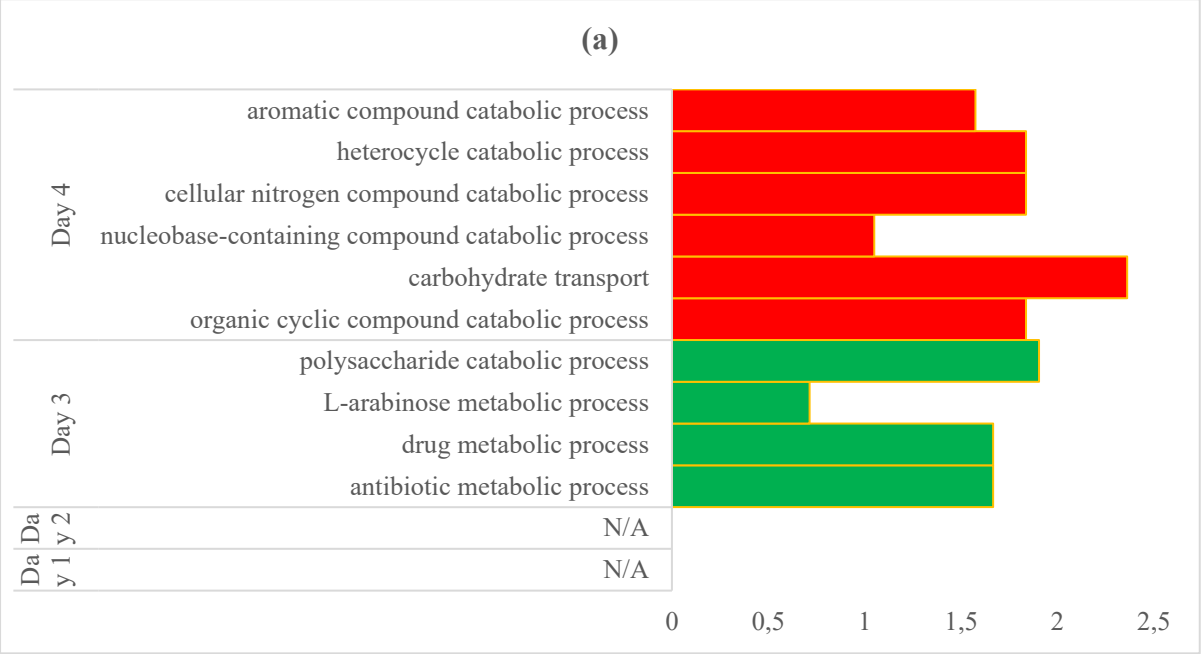
The DEGs that were successfully annotated with GO terms undergo GO enrichment analysis (0.05 > P) (Figure 4.6). The gene enrichment for the downregulated DEGs showed higher number of categories therefore categories enriched with more than 10% DEGs are mentioned for convenience.

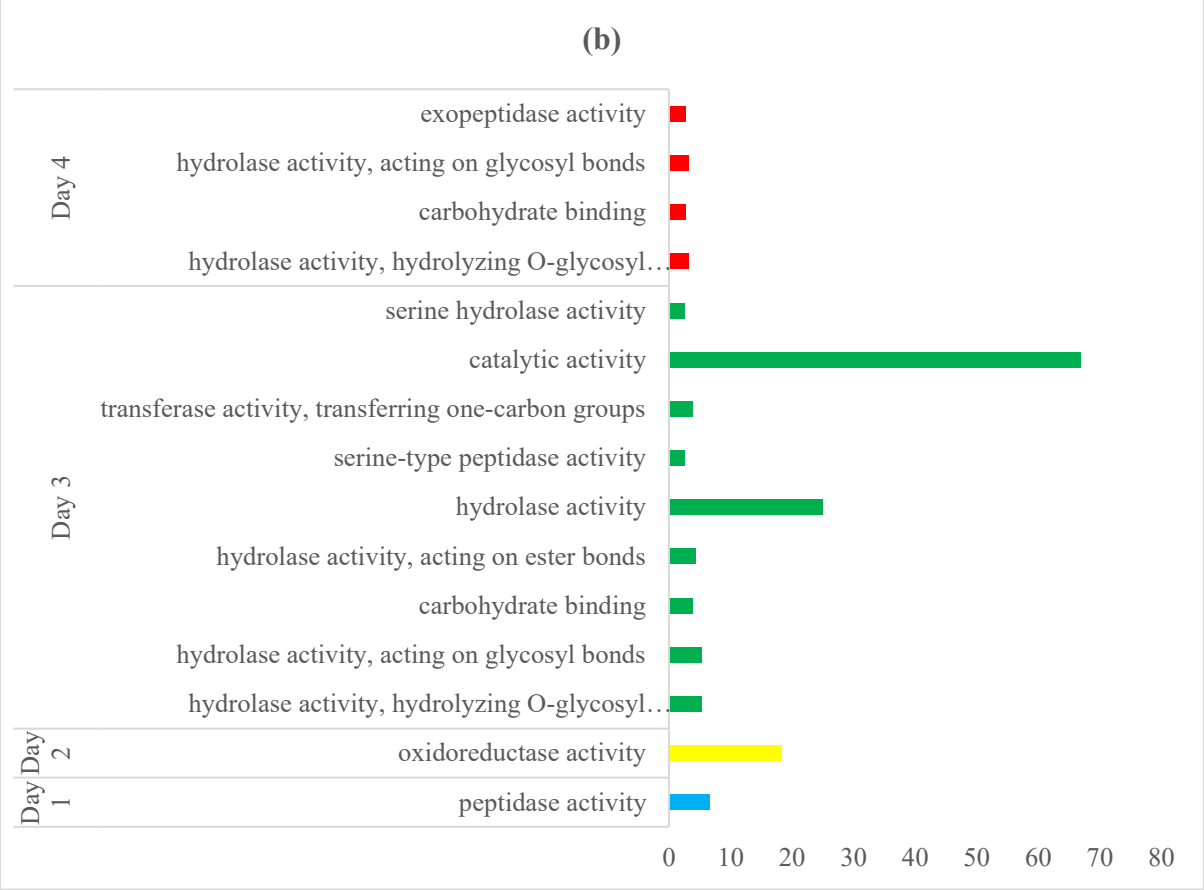
The upregulated DEGs were not enriched with any GO terms under the BP on the first two days. However, genes for the BP were enriched (among others) with antibiotic and drug metabolic process and polysaccharide catabolic process (day 3) and carbohydrate transport (day 4) (Figure 4.6 a). The downregulated genes were mainly enriched with the cellular process on days 1, 2, 3, and 4. Metabolic and biosynthesis process such as peptide, cellular protein, cellular macromolecule, cellular, organic substance, primary metabolic processes were enriched on day 1. Both localization, and establishment of localization were enriched on day 3 and 4. On day three it was observed that transport was enriched and on day 4 biosynthetic

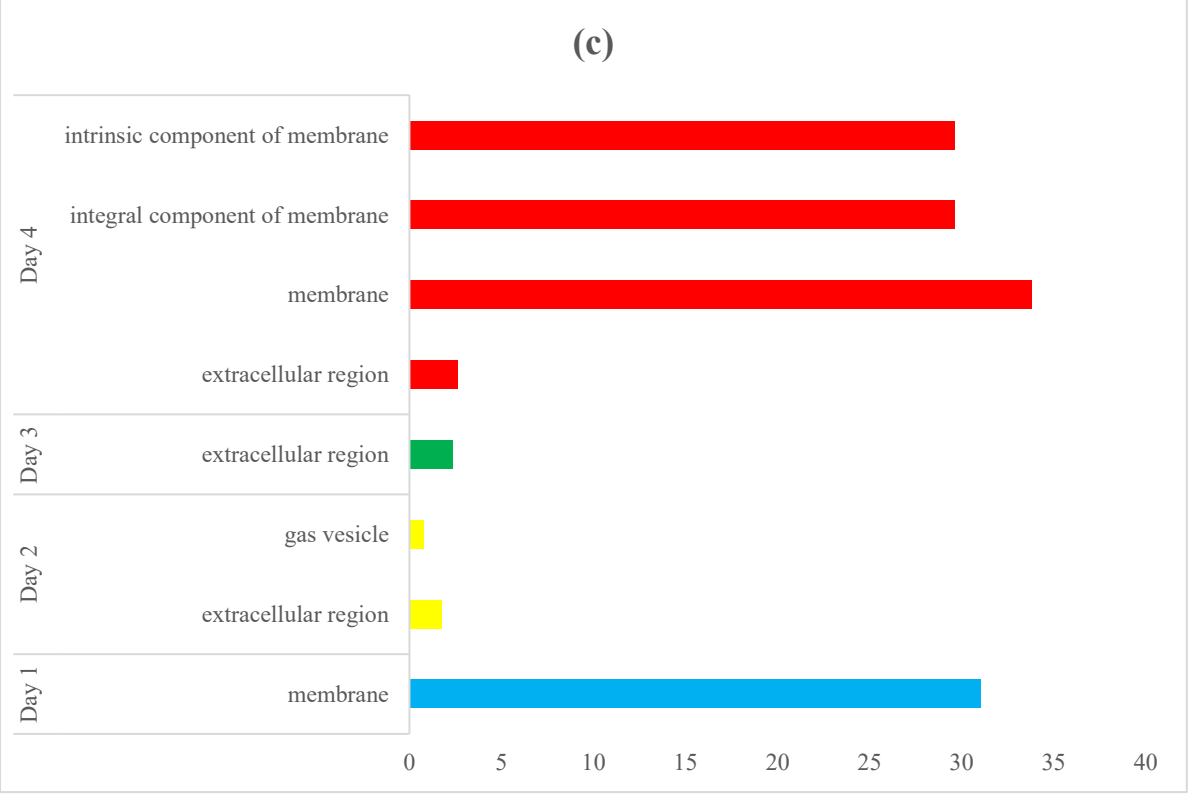
processes such as organic substances, cellular, biosynthetic process were enriched along with cellular metabolic process (Figure 4.6 d).

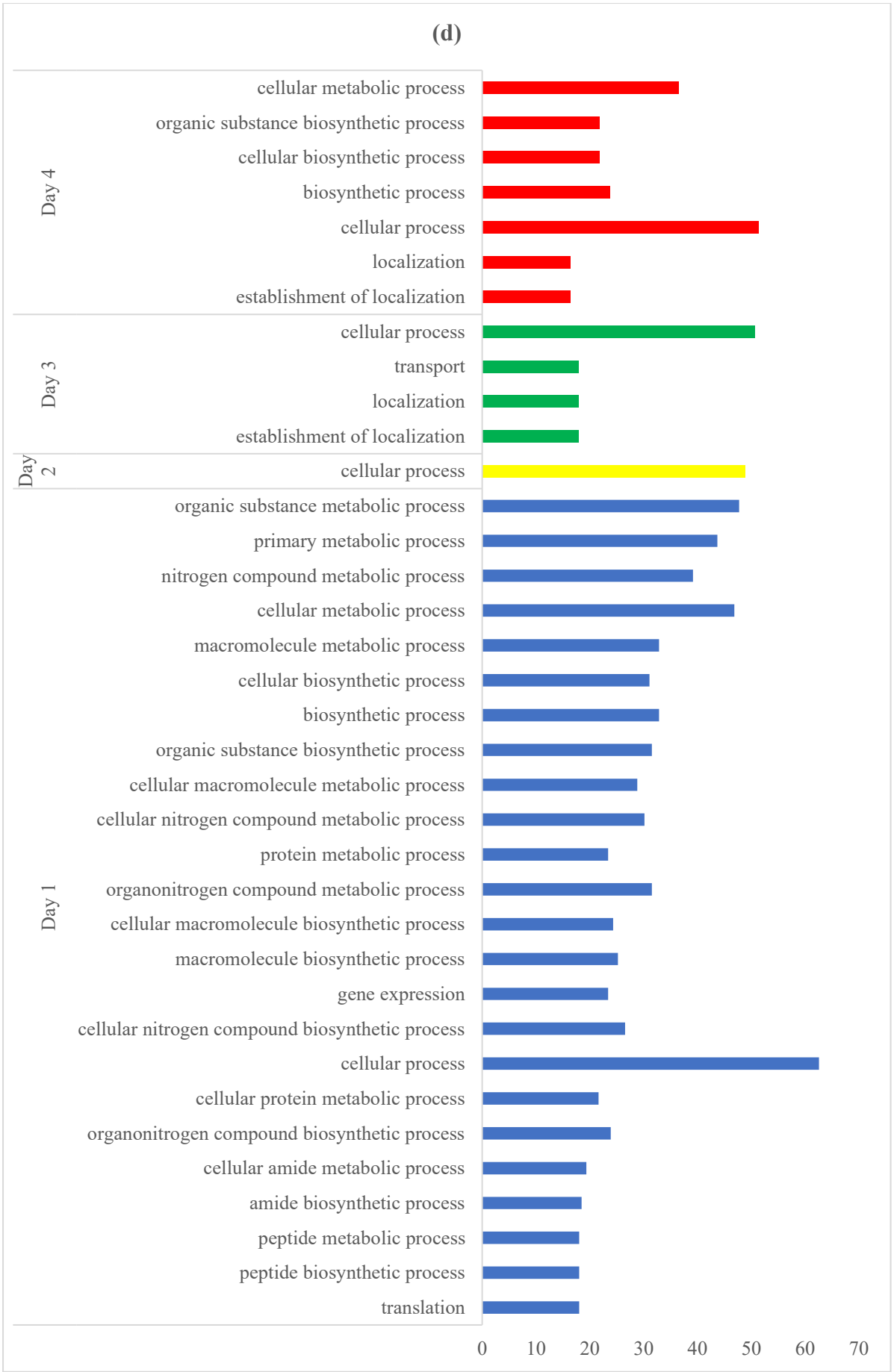
The GO enrichment for MF for the upregulated genes varied every day. The first day it was dominated by "peptidase activity," whereas "oxidoreductase activity" dominated on the second day. On the third day, the enrichment was broadened and significantly dominated by "catalytic activity" over other GO terms (Figure 4.6 b). However, the downregulated DEGs were enriched with broader GO terms for MF. MFs such as binding, heterocyclic, and organic cyclic compound binding were significantly enriched (among others) on day 1 (Figure 4.6 e). Day 3 was mainly enriched with different binding activities *e.g.* ion, nucleotide, anion, small molecule, and nucleoside phosphate binding. Binding activity was again enriched on day 4 along with heterocyclic compound, organic cyclic compound binding, and transferase activity (Figure 4.6 e).

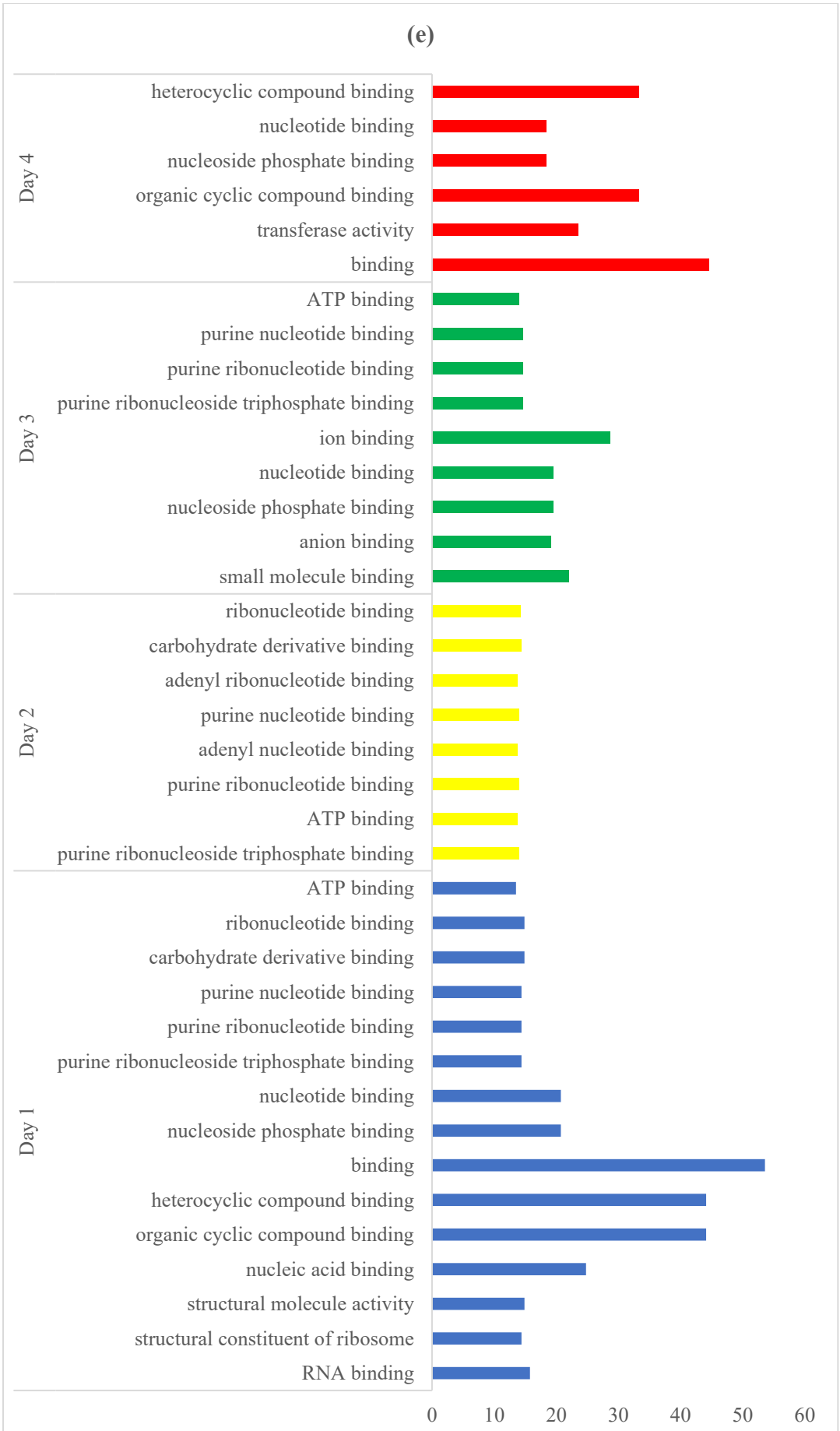
GO enrichment for the CC was much less diversified than the previously mentioned categories. The CC of the upregulated genes identified by GO enrichment analysis included (among others) membrane (day 1 and 4), an integral component of the membrane, and intrinsic component of membrane (day 4) (Figure 4.6 c), while the cellular component of the downregulated genes included (among others) cellular, anatomical entity (day 1), integral and intrinsic component of membrane (day 2), plasma membrane and cell periphery (day 3 and 4) (Figure 4.6 f).




(b)
Day 4
exopeptidase activity
hydrolase activity, acting on glycosyl bonds
carbohydrate binding
hydrolase activity, hydrolyzing O-glycosyl…
Day 3
serine hydrolase activity
catalytic activity
transferase activity, transferring one-carbon groups
serine-type peptidase activity
hydrolase activity
hydrolase activity, acting on ester bonds
carbohydrate binding
hydrolase activity, acting on glycosyl bonds
hydrolase activity, hydrolyzing O-glycosyl…
Day 2
oxidoreductase activity
Day 1
peptidase activity
0 10 20 30 40 50 60 70 80


(c)
Day 4
intrinsic component of membrane
integral component of membrane
membrane
extracellular region
Day 3
extracellular region
Day 2
gas vesicle
extracellular region
Day 1
membrane
0 5 10 15 20 25 30 35 40


(d)
Day 4
cellular metabolic process
organic substance biosynthetic process
cellular biosynthetic process
biosynthetic process
cellular process
localization
establishment of localization
Day 3
cellular process
transport
localization
establishment of localization
Day 2
cellular process
Day 1
organic substance metabolic process
primary metabolic process
nitrogen compound metabolic process
cellular metabolic process
macromolecule metabolic process
cellular biosynthetic process
biosynthetic process
organic substance biosynthetic process
cellular macromolecule metabolic process
cellular nitrogen compound metabolic process
protein metabolic process
organonitrogen compound metabolic process
cellular macromolecule biosynthetic process
macromolecule biosynthetic process
gene expression
cellular nitrogen compound biosynthetic process
cellular process
cellular protein metabolic process
organonitrogen compound biosynthetic process
cellular amide metabolic process
amide biosynthetic process
peptide metabolic process
peptide biosynthetic process
translation
0
10
20
30
40
50
60
70


(e)
Day 4
heterocyclic compound binding
nucleotide binding
nucleoside phosphate binding
organic cyclic compound binding
transferase activity
binding
Day 3
ATP binding
purine nucleotide binding
purine ribonucleotide binding
purine ribonucleoside triphosphate binding
ion binding
nucleotide binding
nucleoside phosphate binding
anion binding
small molecule binding
Day 2
ribonucleotide binding
carbohydrate derivative binding
adenyl ribonucleotide binding
purine nucleotide binding
adenyl nucleotide binding
purine ribonucleotide binding
ATP binding
purine ribonucleoside triphosphate binding
Day 1
ATP binding
ribonucleotide binding
carbohydrate derivative binding
purine nucleotide binding
purine ribonucleotide binding
purine ribonucleoside triphosphate binding
nucleotide binding
nucleoside phosphate binding
binding
heterocyclic compound binding
organic cyclic compound binding
nucleic acid binding
structural molecule activity
structural constituent of ribosome
RNA binding
0
10
20
30
40
50
60

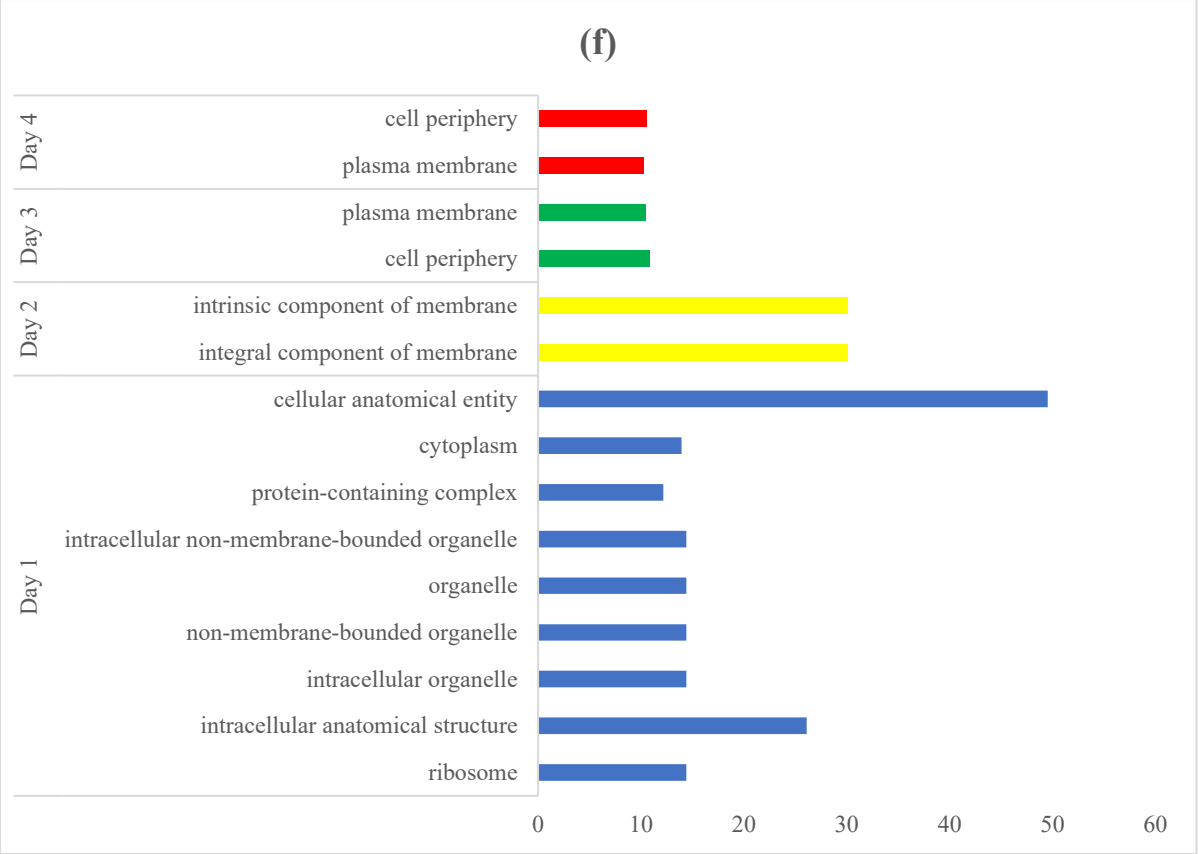


Figure 4. 6: GO enrichment (0.05>P-value) of the DEGs (between the strains) on different days. Gene enrichment from the upregulated genes involved in a)BP, b)MF and c) CC. Gene enrichment from the downregulated genes involved in d)BP, e)MF and f) CC. X-axis represents % of the DEG sequences that were enriched, and Y-axis represents GO.

4.4 Pathway analysis of DEGs from between and within strain comparisons

The functions of the DEGs were assigned to the KEGG pathways for a better understanding of their biological functions in a specific pathway. FASTA sequences contain both up and downregulated genes obtained from the DEG list after the comparisons between and within the strain were mapped to KEGG pathways together, and their role in metabolic pathways and biological behaviors were analyzed. The percentage of the mapped DEGs range between 52.01% to 37.93% (Figure 4.7). The mapped genes are mainly fit into four categories *i.e.* metabolism, genetic information processing, environmental information processing, and cellular process. Pathways having less than 1% of the mapped DEGs were excluded. They are further subdivided into more sub-classes according to the annotation of the DEGs of different days. However, most of the genes were annotated to different metabolic pathways. The global and overview metabolism of mapped genes is presented in Figure 4.8.

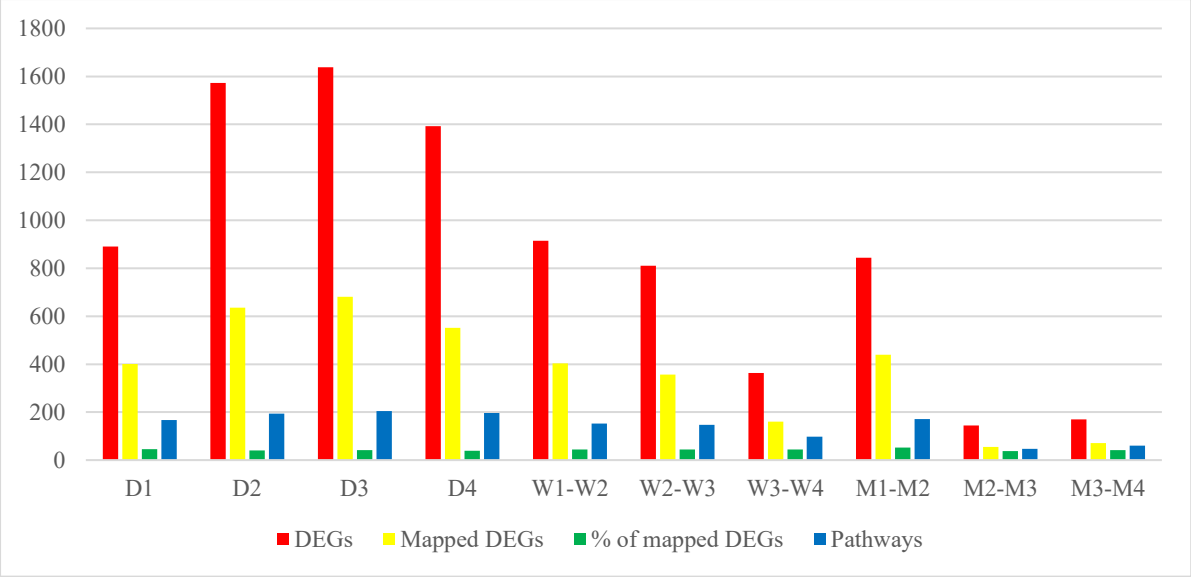
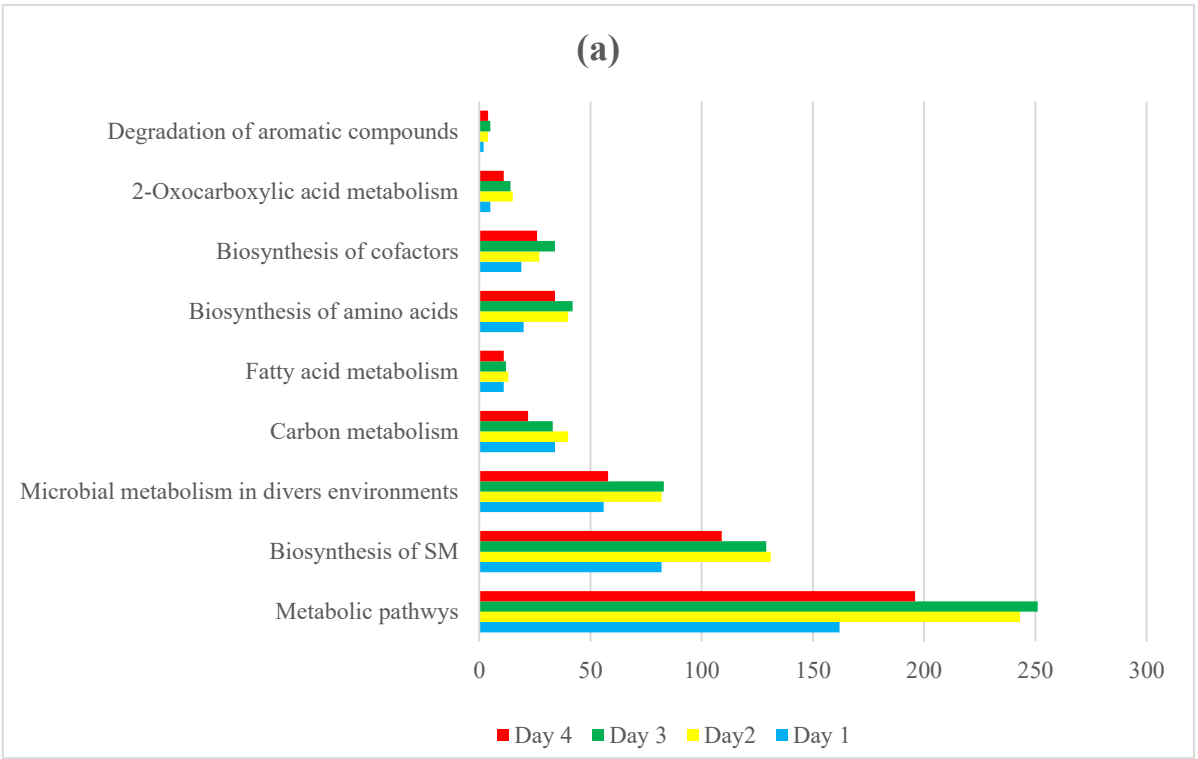


Figure 4. 7: Number and percent of the DEGs those are mapped by KEGG mapper KAAS and their corresponding number of mapped pathways. W1, W2 , W3 and W4 refers to the corresponding sampling day of the WT strain, and M1, M2,M3 and M4 refers to the sampling day from the mutant strain.



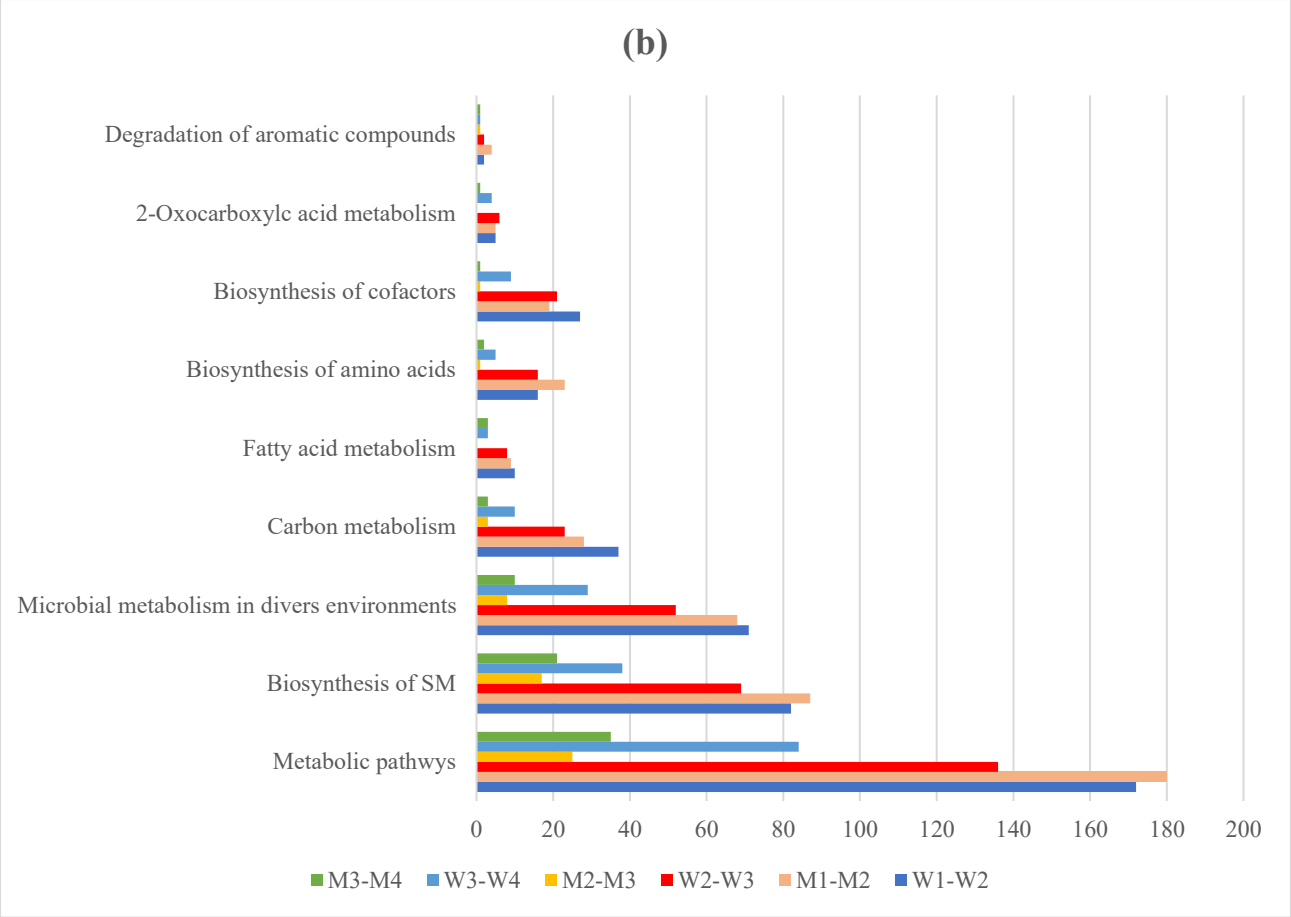
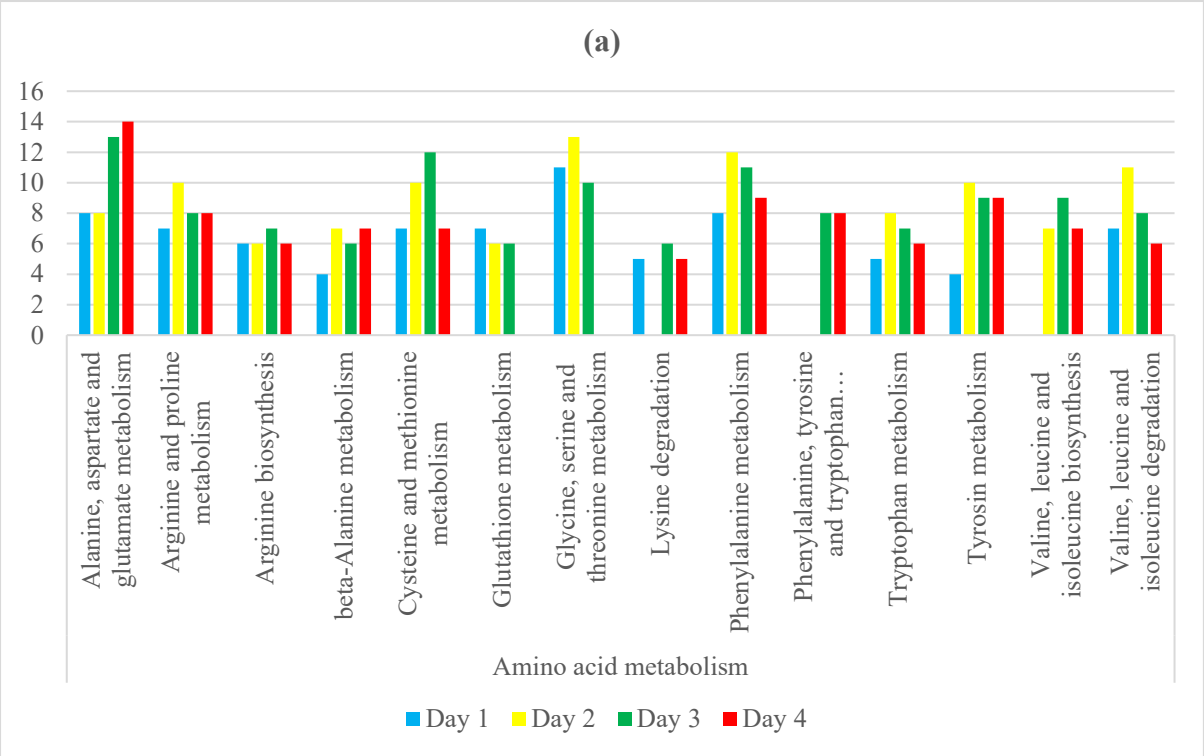


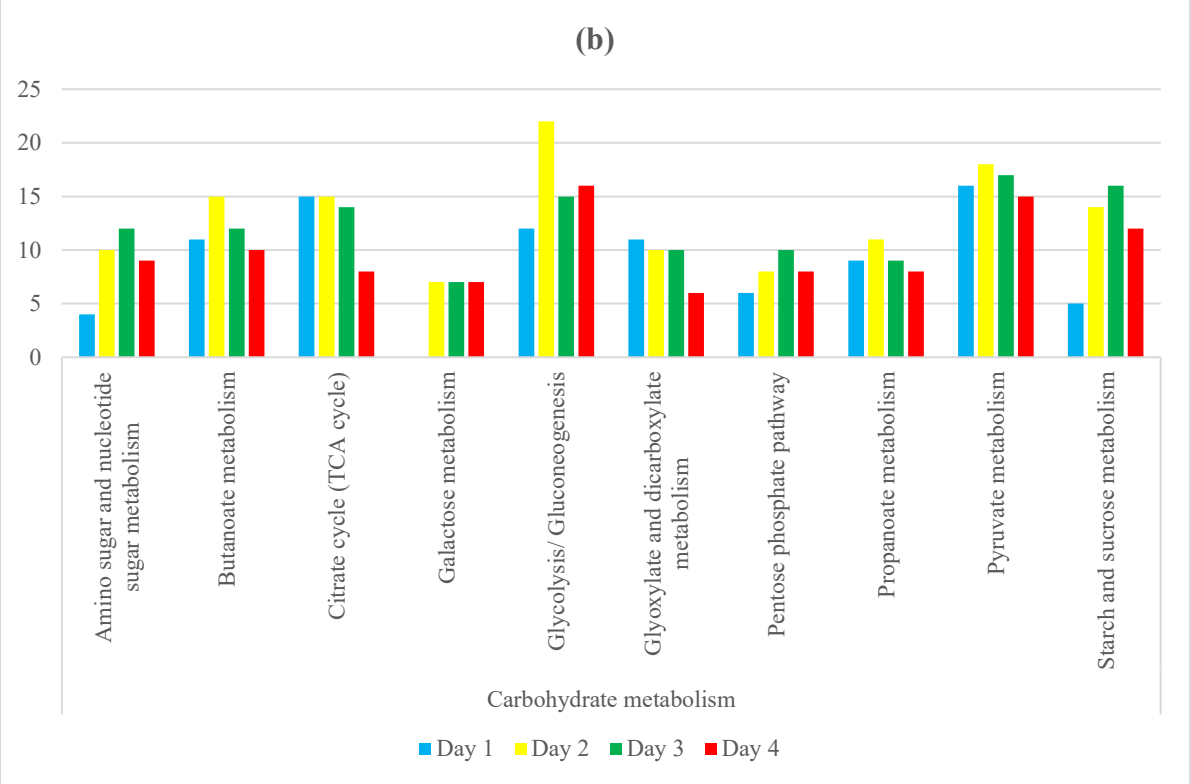
Figure 4. 8: The number of mapped DEGs to the global and overview map. (a) DEGs between the strains (b) DEGs within the strains. SM - secondary metabolites. W1, W2 , W3 and W4 refers to the corresponding sampling day of the WT strain, and M1, M2,M3 and M4 refers to the sampling day from the mutant strain.

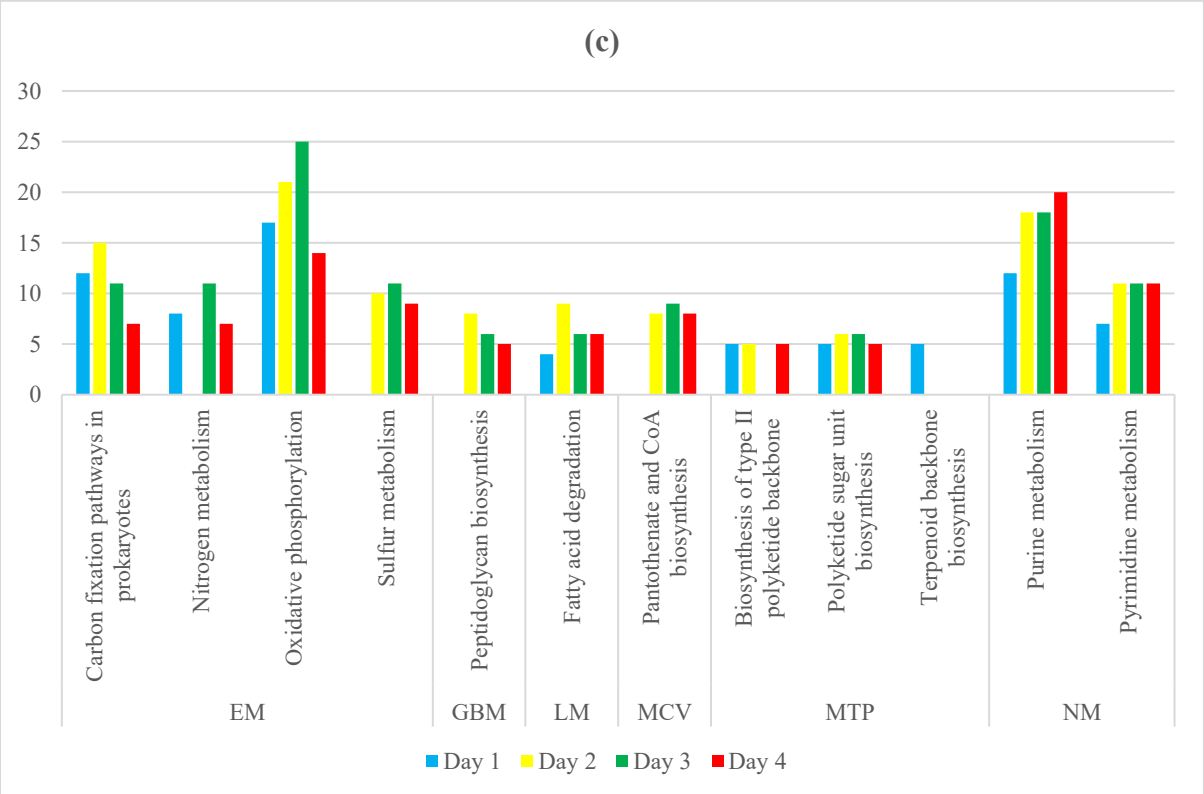
4.4.1 KEGG mapping of the DEGs between the strains

FASTA sequences of the DEGs between WT and MT obtained from four time points i.e. D1, D2, D3 and D4 were mapped to metabolism are categorized into 8 secondary classifications such as amino acid metabolism (14 pathways), carbohydrate metabolism (10 pathways), energy metabolism (4 pathways), metabolism of terpenoids and polyketides (3 pathways), nucleotide metabolism (2 pathways), glycan biosynthesis (1 pathway), metabolism of cofactors and vitamins (1 pathway) and lipid metabolism (1 pathway) (Figure 4.9 a, b, c). Amino acids are the building block of protein. The enriched pathways of the amino acid and carbohydrate metabolism reflected the growth cycle of the bacterial. Most of the mapped genes were either from day 2 or day 3 (Figure 4.9 a, b). The oxidative phosphorylation pathway under energy metabolism showed the highest number of mapped genes. There were 25 DEGs mapped to this

pathway on day 3. Nitrogen and sulfur metabolism pathways from this same category had the highest mapped DEGs on day 3 (Figure 4.9 c). Two pathways from genetic information processing category were enriched with the mapped DEGs. However, more than 30 genes were mapped to ribosomes under the translation category indicates a higher amount of activity from the translational machinery from day 1 and 3 (Figure 4.9 d). The intercellular antibiotic is transported by proton dependent transmembrane electrochemical system. ABC (ATP binding cassette) transporter protein is known to energizing this transport system. This transporter system had the highest number of mapped genes on day 2, 3, and 4 (highest on day 3) among the selected pathways (Figure 4.9 d).




(b)
25
20
15
10
5
0
Amino sugar and nucleotide sugar metabolism
Butanoate metabolism
Citrate cycle (TCA cycle)
Galactose metabolism
Glycolysis/ Gluconeogenesis
Glyoxylate and dicarboxylate metabolism
Pentose phosphate pathway
Propanoate metabolism
Pyruvate metabolism
Starch and sucrose metabolism
Carbohydrate metabolism
Day 1
Day 2
Day 3
Day 4


(c)
30
25
20
15
10
5
0
Carbon fixation pathways in prokaryotes
Nitrogen metabolism
Oxidative phosphorylation
Sulfur metabolism
Peptidoglycan biosynthesis
Fatty acid degradation
Pantothenate and CoA biosynthesis
Biosynthesis of type II polyketide backbone
Polyketide sugar unit biosynthesis
Terpenoid backbone biosynthesis
Purine metabolism
Pyrimidine metabolism
EM
GBM
LM
MCV
MTP
NM
Day 1
Day 2
Day 3
Day 4

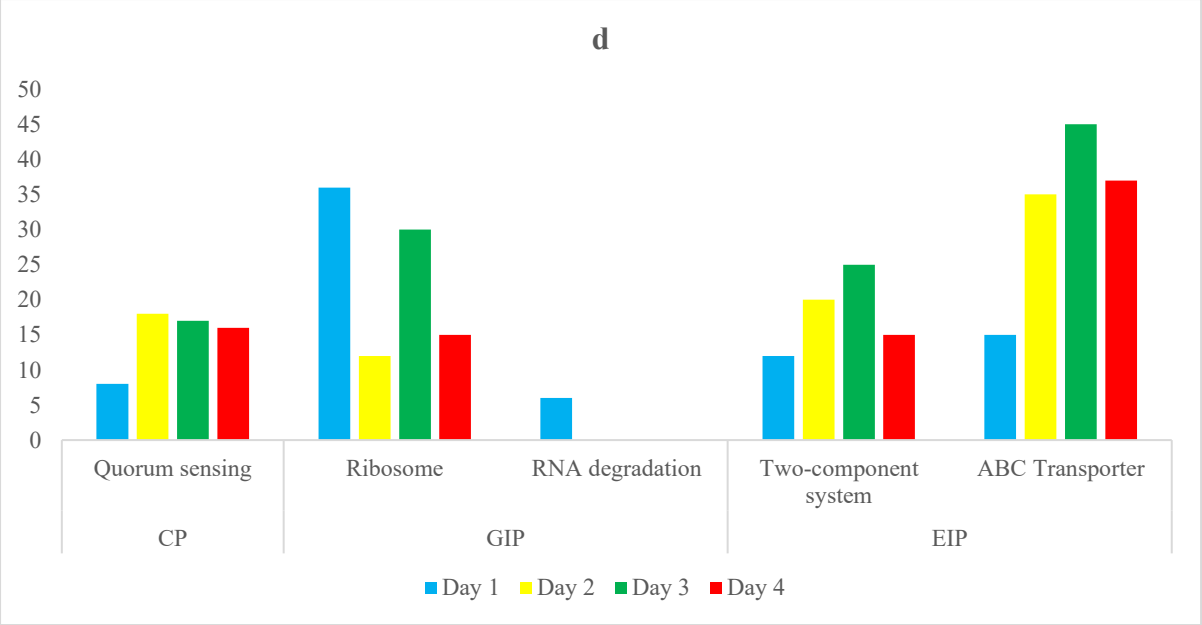


Figure 4. 9: Functional classifications of the mapped DEGs between the WT and MT strain on different days. DEGs mapped to different metabolic pathways of (a) amino acid, (b) carbohydrate, (c) energy, glycan biosynthesis, lipid, cofactors and vitamins, terpenoids and polyketides and nucleotides, (d) pathways involved in cellular processing (CP), genetic information processing (GIP) and environmental information processing (EIP). EM- energy metabolism, GBM-glycan biosynthesis and metabolism, LP-lipid metabolism, MCV-metabolism of cofactors and vitamins, MTP-metabolism ofterpenoids, and polyketides, and NM-nucleotide metabolism. Y-axis denotes the number of mapped DEGs into the pathway.

4.4.2 KEGG mapping of the DEGs within the strains

All the DEGs from the WT and MT strains from three-time intervals were mapped to different pathways. There were variations in their gene mapping. Physiologically critical pathways are in Figure 4.10, and the detailed pathway mapping result is presented in Appendix 4.

These DEGs from the WT were mapped to metabolism are categorized into eight secondary classifications such as amino acid metabolism (13 pathways), carbohydrate metabolism (10 pathways), energy metabolism (4 pathways), metabolism of terpenoids, and polyketides (4 pathways), nucleotide metabolism (2 pathways), lipid metabolism (2 pathways), other secondary metabolites (1 pathway), metabolism of cofactors and vitamins (1 pathway) (Appendix 4 a, b, c, d).

The number of mapped metabolic pathways related to amino acids was predominantly higher than other pathways. The pathways for carbohydrate metabolism were higher. Together, they showed that most of the DEGs were dedicated to cell growth (Appendix 4 a, b). However, 7

pathways belong to amino acid metabolism had shown a higher number of mapped DEGs on the first interval, *i.e.,* W1-W2 (Appendix 4 a). All the pathways for carbohydrate metabolism showed a higher number of mapped DEGs at the same period (Appendix 4 b).

The number of DEGs mapped in oxidative phosphorylation was highest at the second interval, *i.e.* W1-W2. This pathway was mapped by 27 DEGs during this time, and it was the highest number of DEGs among all the pathways and intervals (Figure 4.10 and Appendix 4) involved in WT. Pathways involved in NM showed that this strain had a higher number of mapped DEGs for purine and pyrimidine metabolism during the first interval, which decreased rapidly in the following intervals (Figure 4.10).

ABC transporter system from environmental information process categories was highest during the second interval. There were five metabolic pathways such as amino sugar and nucleotide sugar metabolism, arginine and proline metabolism, fructose and mannose metabolism, glutathione metabolism, terpenoid backbone biosynthesis, where DEGs only from WT were mapped to different intervals (Appendix 4). Additionally, WT had the DEGs mapped to the protein export pathway during the first interval in the genetic information processing category (Appendix 4 d), which was not present in the MT.

Comparison within the strain (MT) showed that the number of DEGs during the last two intervals was much lower than the first interval (Appendix 4 b) and comparatively lower than the WT. These DEGs were mapped to different metabolic pathways were categorized into 8 secondary classifications such as amino acid metabolism (10 pathways), carbohydrate metabolism (9 pathways), energy metabolism (4 pathways), metabolism of terpenoids, and polyketides (4 pathways), nucleotide metabolism (2 pathways), lipid metabolism (2 pathways), other secondary metabolism (2 pathways), metabolism of cofactors and vitamins (1 pathway) (Appendix 4 e, f, g, h).

All the pathways for carbohydrate metabolism and energy metabolism showed fewer mapped DEGs. The number of DEGs mapped to the pathways for glycolysis/gluconeogenesis, and pyruvate metabolism had sharply decreased after the first interval (Figure 4.10). The number of mapped DEGs belong to oxidative phosphorylation was 26 during the first interval, which was drastically lowered to 1 during the middle interval (Figure 4.10). However, the MT had shown some pathways that are not present in the WT strain, such as acarbose and validamycin biosynthesis, biosynthesis of vancomycin group antibiotics. Both cases mapped DEGs appear

only during the third interval, which roughly demonstrated that the mutant strain might have a different profile for antibiotic production and appeared lately (Appendix 4 g). Similarly, the number of mapped DEGs for nitrogen metabolism and ABC transporter system had decreased significantly during the second and third intervals (Figure 4.10).

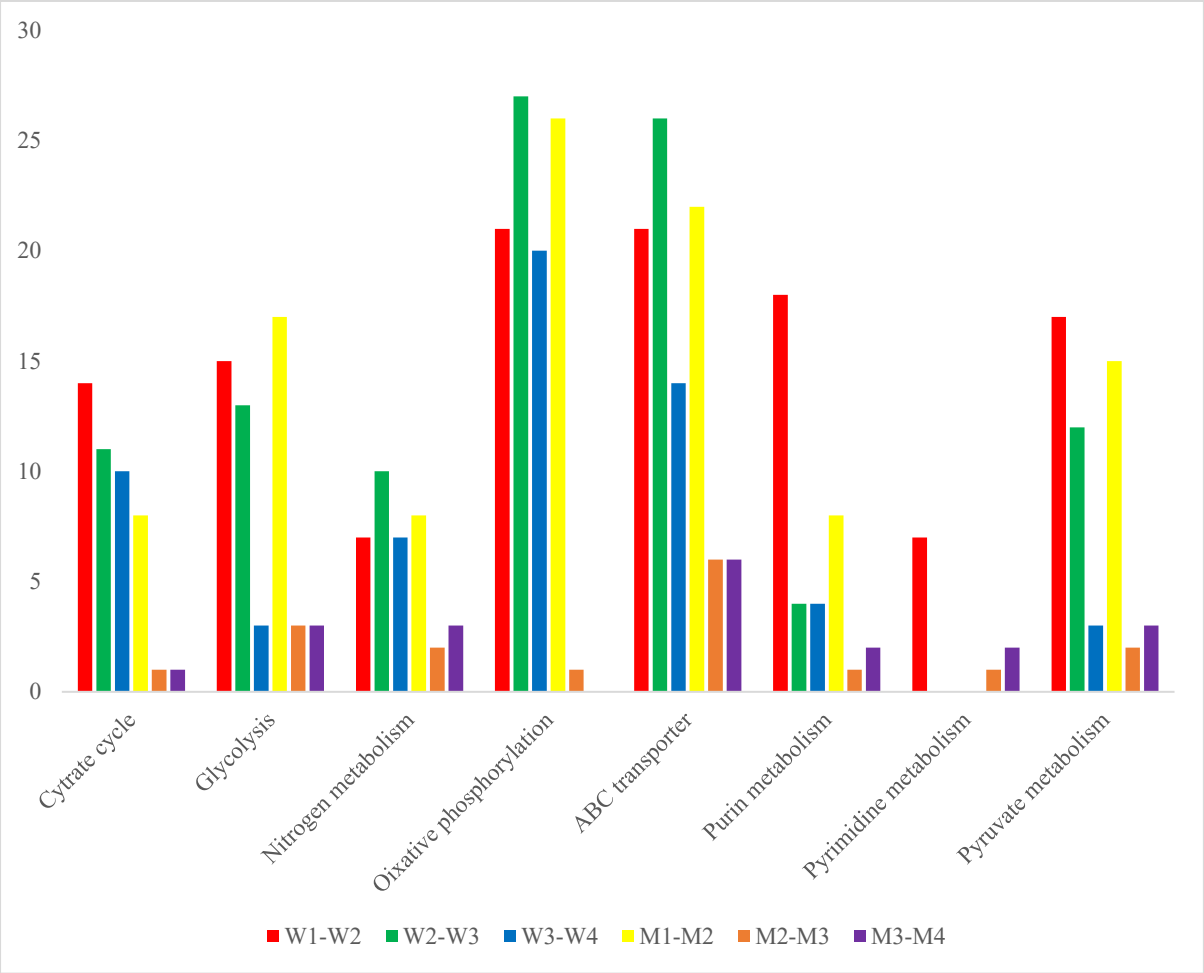


Figure 4. 10: KEGG Mapping of the DEGs from within strain analysis. Y-axis denotes the number of mapped DEGs into the corresponding pathways. W1, W2 , W3 and W4 refers to the corresponding sampling day of the WT strain, and M1, M2,M3 and M4 refers to the sampling day from the mutant strain.

5. Discussions

5.1 RNA-Seq for DEG analysis

RNA-Seq is now the most powerful and robust technology for genome-wide differential gene expression analysis. The analysis of the expression levels of all mRNAs in the transcriptome during cellular development can be a tool to elucidate regulatory pathways. Several transcriptomic studies used a mutant strain of *Streptomyces* to explain their gene expression pattern (Bignell *et al.,* 2005; Kato *et al.,* 2002 and Medema *et al.,* 2011). This thesis focuses on the gene expression pattern of the WT and MT strains. Analyzing the RNA-Seq data will make it easy to explain why the Acl producing WT and MT have different antibiotic synthesis profiles. In addition to this, GO enrichment analysis and KEGG pathway mapping of these DEGs may explain their functional categories and their expressed pathways to understand the recycling of Acl and why HO42 is an overproducer.

After the emergence of NGS techniques, comparative transcriptome analysis has widened the scope to study the biosynthetic gene clusters in biologically relevant organisms (Amos *et al.,* 2017). There are multiple tools available for a defined RNAseq pipeline, yet it is always challenging to choose the appropriate tools suitable for the objectives. Optimizing the sequencing depth (reads per sample) for these studies is crucial and a significant aspect to consider during an experimental design. Haas *et al.* (2012) suggested that the sequencing depth of 5-10 million rRNA depleted reads are enough to characterize the transcriptional activity from a wide variety of microbial species. In this study, average sequencing depths were 16.4 million (WT) and 15.3 million (MT). A much higher sequencing depth (> 30 million) can create noise from the highly expressed genes, although it can detect the transcripts with significantly lower abundance. In contrast, lower sequencing depth keeps these low abundance transcripts undetected (Tarazona *et al.,* 2011).

Estimation of within-group variability is necessary for making inferences about the conditions. The use of biological samples or replicates from different backgrounds is useful to draw such inferences (Auer & Doerge, 2010 and Fang & Cui, 2011). However, due to financial and technical constraints, researchers have to keep the number of biological replicates small (Auer & Doerge, 2010; McCarthy *et al.,* 2012 and Zhao *et al.,* 2016). Therefore, it is always critical to optimize the experimental design by balancing the sequencing depth and biological replicates (Auer & Doerge, 2010 and Fang & Cui, 2011). Pooling the RNA-Seq data from

different samples can mitigate the issue with an unbiased gene expression profile. Assefa *et al.* (2020) suggested that pooling of the samples can detect the biological effect of interest by retaining the variability of the sample. It is preferable to use three biological replicates for each experimental condition. Although three biological replicates due to economic limitations, they were pooled together. According to the edgeR (Robinson *et al.,* 2010) manual, dispersion value was set to 0.01 because the samples were pooled and variability was counted (Yunshun *et al.,* 2020).

In this experiment, time course DGE within the strain with 24hr intervals showed us the up and downregulation of primary and secondary metabolic pathways. Secondary metabolism is strongly connected to primary metabolism. Precursors and cofactors for secondary metabolites are derived from processes in the central carbon metabolism. However, precursor units might be synthesized through the degradation of stored macromolecules. Evidence suggested that the degradation of fatty acids and branched-chain amino acids (BCAAs) has been suggested to contribute to the acetyl-CoA supply for certain PKs in *Aspergillus* species (Richter *et al.,* 2014). Comparative analysis of regulatory genes (for example) may positively associate with higher production of Acl B in MT.

5.2 Predicted pathways in WT

The biosynthetic pathways for producing natural products *e.g.* antibiotics are often encoded by the genes living in close proximities and organized as BGCs. Their organization follows a highly conserved logic. With the help of this gene conservation logic, genome mining identifies the core bioactive molecules by homology searching (Blin *et al.,* 2019). Genes from these clusters are involved in precursor molecule biosynthesis, compound scaffold assembly, and modification. Besides these core functions, occasionally, genes for resistance, transport, and regulation are also involved (Nützmann *et al.,* 2016). A review study on BGC in *Streptomyces* had shown that the genomes could carry 8-83 BGCs (mean = 39.64 and SD = 11.40) (Belknap *et al.,* 2020). In this experiment, antiSMASH detected 33 BGCs in the WT genome. The computational identification of the BGCs starts from the core enzymes produced by core biosynthetic genes (for Acl producing BGC, gene ID: fig_33899.16.peg.2277 and fig_33899.16.peg.2278; mentioned Table 4.3), involved in the SM synthesis pathway. pHMM is a probabilistic model which was used by antiSMASH to detect this core catalytic molecule

producing genes. In the next step, by using manually curated BGC cluster rules, antiSMASH detected the other core genes (*e.g.* additional biosynthetic genes, genes related to transportation, regulatory genes, and other accessory genes) (Figure 4.3) (Blin *et al.,* 2019). However, many of those genes that reside in this predicted BGC were well characterized (Table 1.2).

Many *Streptomyces* contain particular BGCs responsible for synthesizing aromatic polyketide, which involves spore pigmentation (Iftime *et al.,* 2016) and induced during sporulation of aerial hyphae or as an indication of initiation of stationary phase (Kelemen *et al.,* 1998). Among the 31 predicted BGCs throughout the WT genome, region 1.25 (Table 4.2) showed similarity with multiple BGCs responsible for the synthesis of spore pigmenting polyketide *e.g.* BGC0000271 (Omura *et al.,* 2001) and BGC000272 (Martin *et al.,* 2001) with 83% and 85% similarity. Comparing the DEG data from within the strain analysis (appendix 8) likely demonstrate how these two strains *i.e.* WT and MT may differ from each other about the onset of sporulation and entering the stationary growth. The predicted region spans between the gene IDs fig:33899.16.peg.7322 to fig:33899.16.peg.7394. The DEG analysis clearly shows that a group of genes involved in transport related activities were upregulated during the last two intervals. However it must be noted that most of these DEGs had comparatively much higher upregulation. Since in-depth analysis of this BGC was not part of the thesis plan, it should be addressed in future experiments.

5.3 DE of the regulators

Although there are local and global regulators to influence the synthesis of an antibiotic, typically, genes inside the corresponding BGC have a superior effect on the production of the relevant antibiotic (Liu *et al.,* 2013). In addition to eliciting regulatory effects, some other genes within a BGC can also encode enzymes that catalyze the reaction for antibiotic formation (Wei *el al.,* 2018). These regulatory genes are often termed as cluster situated regulators (CSRs) (Liu *et al.,* 2013 and Wei *et al.,* 2018). However, the regulatory effects of those genes within the genome can exert upon an antibiotic producing BGC in a pathway-specific manner; they can also regulate the expression of the genes residing outside the BGC (Niu & Tan, 2013).

In contrast to this, pleiotropic regulators residing outside of the BGCs can also regulate the production of multiple antibiotics in addition to morphological development of the respective organism. Another type of regulator that can regulate the central metabolic system and those

regulatory genes mentioned above. They are termed as global regulators and situated throughout the genome (Liu *et al.,* 2013). Sigma factors and other related mechanisms, *e.g.* anti-sigma and anti-anti-sigma factors, contributed to an alternative regulation mode that commonly regulates antibiotic production in *Streptomyces* sp. (Zhou *et al.,* 2011). There are approximately 97 sigma factors, and such genes reside throughout the WT genome, of which few are just up and downstream of the predicted Acl producing BGC. This suggests a sigma factor-based regulation might exist in *Streptomyces* (Pinilla *et al.,* 2019). A regulator of sigma factor (gene ID=fig|33899.16.peg.2336) resides just downstream of the predicted Acl producing BGC was upregulated on day 2 (LogFC: 1.37) and day 3 (LogFC: 1.23). Since both WT and MT were grown in the same environment, this differentiation should not arise from environmental factors such as nutrition and temperature. However, sigma factors transcriptionally control the sporulation and other responses due to stress in bacteria *e.g. Bacillus* sp. (Zhou *et al.,* 2011).

5.4 Topmost DEGs

The 20 most up and downregulated genes are listed in appendix 1, and no DEG was reported from the Acl gene cluster when the comparison comes from between the strain. Additionally, many of the highly regulated genes were uncharacterized (hypothetical protein). Genes classified as hypothetical are predicted to code for proteins in an organism, but no experimental evidence for their function for the entire protein family (Ijaq *et al.,* 2019). *Streptomyces coelicolor* is considered a model organism for antibiotic-producing bacteria. but about 34% of its predicted protein-coding genes were hypothetical (Alam *et al.,* 2011).

One of the highly upregulated gene (fig|33899.16.peg.679) on D1 was a putative GT (appendix 1ai). This enzyme acts as a drug-tailoring tool and is involved in the glycosylation of natural products. Experimental evidence suggests that NDP-activated sugars were utilized by GT and stereo specifically transfer the sugar residue to the aglycone (Nguyen *et al.,* 2010 and Salem *et al.,* 2017). Long-chain-fatty acid-CoA ligase (fig|33899.16.peg.5946) activates the oxidation of complex fatty acids, and with the help of ATP and CoA, catalyzes the formation of fatty acyl-CoA (Watkins, 1997). It was also one of the most upregulated genes on D2 and D4, and other genes such as polyketide synthase modules and related proteins (fig|33899.16.peg.5943), ABC transporter permease protein 2 (fig|33899.16.peg.2475) (appendix 1aii and aiv).

5.5 DEGs of the Acl producing cluster

Genes residing inside the Acl producing BGC were differentially expressed and listed in Table 4.3 and 4.4. The TIGR04222 domain-containing membrane protein (gene ID: fig|33899.16.peg.2253) in the MT strain showed an intriguing expression pattern (Table 4.4). It had a predominantly higher expression during the first interval *i.e.* D1-D2, compared to the WT strain. This gene is may contribute to self-immunity by regulating the toxic activity (Baba & Schneewind, 1998; Flaherty *et al.,* 2014).

The gene for trypsin-like protease (TLP)/serin protease (gene ID: fig|33899.16.peg.2260) had shown an distinguished pattern of expression. In contrast to MT, it was upregulated at the first time point in WT. However, between the strain comparison shows that this gene was significantly downregulated after initial upregulation (Table 4.4). The secretion of proteolytic enzyme *e.g.* serin protease by *Streptomyces* is often shown at the onset of secondary metabolism. Thus, antibiotic synthesis and serine protease formation are coordinately regulated (Ginther, 1979).

Polyketide chain length factor (PCLF) (gene ID: fig|33899.16.peg.2277) forms a complex with β-keto acyl synthase (KS) (gene ID: fig|33899.16.peg.2278) and an acyl carrier protein (ACP) (gene ID: fig|33899.16.peg.2276) and makes up the minimal PKS. Within strain analysis (Table 4.3) demonstrated that genes involved in the process mentioned above were expressed differently between the strains (Table 4.3 and 4.4). This minimal PKS complex determines the PKS product's length (Bisang *et al.,* 1999; Watanabe & Ebizuka, 2004). Other accessory genes (in addition to minimal PKS) required for the synthesis of the intermediate aklavinone are polyketide ketoreductase (gene ID: fig|33899.16.peg.2280), monooxygenase (gene ID: fig|33899.16.peg.2279) and aromatase (gene ID: fig|33899.16.peg.2281) (Räty *et al.,* 2002b). These three genes had shown a similar pattern in both DEG analyses (Table 4.3 and 4.5), and except polyketide ketoreductase and monooxygenase, the other four genes did not show significant changes during intermediary duration *i.e.* D2-D3 (Table 4.3). Initially, nogalonic acid methyl ester cyclase (gene ID: fig|33899.16.peg.2271) was more downregulated in WT but upregulated later. Although MT had shown a similar trend in expression, the magnitude was lower than WT. This cyclase controls the stereochemistry of anthracyclines in *Streptomyces nogalater* (Torkkell *et al.,* 2000). The above-mentioned data apparently suggests that the WT strain has an efficient system for aklavinone production.

The tetracyclic aklavinone is glycosylated by a glycosyltransferase (encoded by *aknS*) (gene ID: fig|33899.16.peg.2293) and transfers the first deoxyhexose. An *in vitro* study suggests that its nearby the gene *aknT* (gene ID: fig|33899.16.peg.2292) strongly increases the expression of *aknS*, but it had expressed poorly in the absence of this activating gene (Lu *et al.,* 2005). Although the influencing role of *aknT* in this study was not detected but this gene-couple had showed observable differences, as mentioned earlier. However, DE of *aknT* in the MT strain was not significant at the later stage of growth.

5.6 GO enrichment analysis of DEGs

The characteristics of the biological attribute of the RNA-seq data can be identified by GO. I performed the GO classification and enrichment analysis based on data from the strain comparison, *i.e.* WT vs. MT. After that, GO enrichment analysis ($P < 0.05$) for the up and downregulated genes was completed and categorized into different functional classes. I have listed the top 30 GO classes (appendix 3). The more general GO annotations of the DEG between the strain from different days were in Figure 4.5. DEG on day 3 showed the highest number of annotated genes, *e.g.* ATP binding, DNA binding, metal ion binding, hydrolase activity, ATPase activity, etc., that belongs to MF. The annotated genes for BP showed that genes involved in the oxidation-reduction process were predominantly higher than other annotated genes throughout the growth cycle (appendix 3e-h), which demonstrate oxidoreductase activity and the control of oxidative stress (Pinilla *et al.,* 2019). The CC ontology terms were used to annotate the cellular location of the gene products. In this experiment, it has been observed that a significant portion of the DEGs was annotated to "integral component of membrane" (appendix 3: i-l) thus, It can be hypothesized that the product of most of the DEGs has at least some parts of their peptide embedded in the hydrophobic region of the membrane (Caspi *et al.,* 2018). Among other annotated GO terms in the CC category, the ABC transporter complex forms the central pore through the plasma membrane and transports the metabolites throughout the cell (Roncaglia *et al.,* 2013). This complex enables the transport of the antibiotic through the cell membrane to ATP hydrolysis and contributes to self-resistance against the produced antibiotics (Méndez & Salas, 2001).

The gene category overrepresentation analysis is a simple but widely used method to highlight any BP (Young *et al.,* 2010). WT is enriched with the genes involves in oxidoreductase activity.

In general, BGCs contain multiple oxidoreductase encoding genes, which play a critical role during the synthesis of its related natural product (Scott & Piel, 2019).

The upregulated genes involved in different hydrolase activity were enriched on day three (Figure 4.6b). Hydrolase enzymes catalyzed the bond cleavage reaction, *e.g.* aminolysis, esterification, polymerization, transesterification; by using water as a nucleophile (Paul & Fernández, 2016). It is suggested that the hydrolytic activity of different hydrolase enzymes has a correlation with the antibiotic synthesis. Hydrolytic activities such as protease, nuclease, amylase in streptomycin, novobiocin, levorin, and oleandomycin producing actinomycetes were significantly changed during the development of the cultures (Baskakova *et al.,* 1981). In gram-positive bacteria, a tightly organized interplay of hydrolytic and biosynthetic enzymatic activities actively remodeled the cell wall macromolecule peptidoglycan. This cell wall remodeling partially possess by hyphae and spore formation are significantly influenced by cell hydrolase enzymes, Out of 56 candidate cell wall hydrolase genes in *Streptomyces coelicolor* identified *in silico*, seven expressed during vegetive growth and sporulation (Haiser *et al.,* 2009). Gene enrichment analysis of my transcriptomic data showed that WT is enriched by genes with hydrolase activity on D3, which is possibly the outcome of its efficient cell differentiation. Additionally, WT also showed superior hydrolase activity on glycosyl bonds and catalytic activity (Figure 4.6b).

The cellular component of the enriched genes that were upregulated shows that on D1 and D4, the upregulated genes exclusively enriched the membrane part. In addition to this, some more sub-cellular regions, *e.g.* intrinsic and integral components of the membrane, were enriched in WT (Figure 4.6c). The hypha tip is considered an important area where membrane protein and lipids may be secreted (Flärdh & Buttner, 2009). Additionally, experiments suggested that secondary metabolism and cell differentiation are related in some *Streptomyces* (Li *et al.,* 2006 and Ou *et al.,* 2008) because membrane-bound enzymes catalyze the synthesis of cell-wall (Brötz-Oesterhelt & Brunner, 2008; Chopra *et al.,* 2002).

5.7 KEGG pathway mapping of the DEGs

Researchers get systematic insight into gene lists generated from omics-based experiments by pathway enrichment analysis which identifies biological pathways enriched with the listed genes significantly (Reimand *et al.,* 2019). Among other tools, scientists used KEGG pathway enrichment analysis tools in their experiments to get an overview of the metabolic network of different microorganisms (Jia *et al.,*, 2017; Malik *et al.,* 2020; Shen *et al.,* 2020).

The DEGs from within the strain and between the strain comparisons were annotated to KEGG pathways. Global and overview mapping of DEGs from the comparison between the strain showed that most of the annotated DEGs were from days 2 and 3. DEGs from the "within the strain" comparison demonstrated a similar pattern, that means the annotated genes come from mostly during the intermediate sampling time (Figure 4.7). The comparison between the strains demonstrated that different metabolic pathways related to primary metabolism, *e.g.* carbon, amino and fatty acids that could affect the secondary metabolism and biosynthesis of cofactors, were higher in WT than MT. It has known that the source of carbon influences secondary metabolite production and can act on different levels. Additionally, this influence directed the flow of carbon atoms to precursor amino acid, side-chain precursors (Rokem *et al.,* 2007).

Streptomyces utilize both glutamate and aspartate as sources of carbon and nitrogen (Corvini *et al.,* 2004). They also play a critical role in the central nitrogen metabolism (Hodgson, 2000). Aspartate transaminase transferred the amine group of aspartate to glutamate (Hodgson, 2000), and the carbon chain of glutamate and aspartate enters the TCA cycle after deamination (Borodina *et al.,* 2005). Alanine is considered to be a very important amino acid for *Streptomyces* due to its significant contribution to protein biosynthesis, peptidoglycan synthesis (Borodina *et al.,* 2005). DEGs (between the strain) mapped to alanine, aspartate and glutamate metabolism showed a clear indication that the number of DEGs involved into these pathways are increased (Figure 4.9a). However, it is also observed that the number of mapped DEGs in this metabolic class is dominated by the downregulated genes. Such as, aspartate aminotransferase (EC 2.6.1.1) catalyzes a reversible transfer of an amino group between aspartate and glutamate (Appendix 5). It also demonstrated that enzymes *e.g.* EC 6.3.1.2, EC 3.5.1.2 and EC1.4.1.13 multiple pathways directly connected with L-glutamate were mapped with downregulated DEGs. (Appendix 5).

Several carbohydrates-related metabolic pathways *e.g.* TCA cycle, glycolysis, pyruvate was also enriched by the DEGs (Figure 4.9b). Actinorhodin (ACT) is a type II polyketide produced by *Streptomyces coelicolor* A3(2) (Itoh *et al.,* 2007). Gene deletion experiment of glucose-6-phosphate dehydrogenase and phosphoglucomutase (PGM), and overexpression of acetyl coenzyme A carboxylase (ACCase) revealed that glucose-6-phosphate dehydrogenase plays an important role by regulating the carbon flux to ACT production. In contrast, PGM deletion resulted in glycogen overproduction coupled with lower ACT production. However, an efficient and increased supply of glucose to the ACT pathway can be achieved after overexpression of ACCase (Ryu *et al.,* 2006). Mapped data of the DEGs obtained from D2 showed that the downregulated gene PGM was mapped to the reversible conversion of glucose-1-phosphate and glucose-6-phosphate. It should be noted that most of the catalysts involved in the Embden-Meyerhof pathway (appendix 10) were downregulated; hence WT was reported to produce Acl, and upregulated gene ACCase was also mapped to the glycolytic pathways. Therefore there are certainly a diversion of carbon flow than expected. Phosphoglucose isomerase directs the carbohydrate catabolism to TCA cycle, subsequently create precursor molecules for secondary metabolite formation (Salas *et al.,* 1984). However, gene encode this enzyme didn't map to glycolysis metabolic pathway at any time point (data not shown) also suggests that this gene might have a stable expression.

A pan genomic study on *Streptacidiphilus* revealed that core genes (Callister *et al.,* 2008) mainly were mapped to the ribosome, purine metabolism, and oxidative phosphorylation pathways (Kim *et al.,* 2015; Malik *et al.,* 2020). KEGG mapping of the DEGs had shown a similar result as these pathways have one of few of the highest number of mapped genes over other metabolic pathways in my data (Figure 4.9).

It has also been observed that all the 17 strains of *Streptacidiphilus* showed the highest number of mapped gene in ABC transporter (Malik *et al.,* 2020). *Streptomyces* secreted the antibiotics through the cell membrane as part of its self-defense mechanism. Along with other molecules, ABC transporter system aids this secretory mechanism and also represents the largest protein family (Wilkens, 2015). Similarly, KEGG mapping has showed that highest number of the DEGs were mapped to "ABC transporter" (Figure 4.9d).

The two-component system (TCS) is a signal transduction system that helps *Streptomyces* to cope with the ever-changing environment. TCS acts as pleiotropic regulator for more than one

antibiotic production (Rodríguez *et al.,* 2013). Experiments suggested that TCS regulate the production of their respective antibiotics through some signal molecules, *e.g.* phosphokinase, or nutritional signal through carbon/nitrogen/phosphate or carbon/nitrogen ratio or pH value. The mapping of the DEGs for two-component shows that this system was enriched on day three. Carbon metabolism was slightly lower during this time (Figure 4.8a), but nitrogen metabolism was higher on day three (Figure 4.9c).

Comparative DEG within the strain of WT and MT (Figure 4.1b) shows that the number of DEGs in MT was much lower at the last two intervals. The number of upregulated genes was much lower than M2-M3, which reflects mapping of these DEGs into different KEGG pathways. However, the number of downregulated genes was higher in several cases. Surprisingly, the number of DEGs in the last interval, *i.e.,* M3-M4, was reversed. Additionally, many of those up/downregulated genes during M2-M3 were reversed at M3-M4. This depicts a scenario where a metabolic switching might take place in the mutant strain during this transition period, and therefore the number of significant DEGs was few.

Due to very a smaller number of DEGs during the last two intervals in MT strain *i.e.* M2-M3 and M3-M4 (appendix 4e), which demonstrated a fact that most part of the MT genome have a steady transcriptional activity. In house experiment on MT strain showed that it can produce anthracycline for a prolonged period of time (>10 days). The mutant strain also demonstrated a lack of oxidative phosphorylation (Figure 4.10), which means it was under continuous oxidative stress. Although it depends upon the species which type of metabolism will dominate throughout their growth (Millan-Oropeza *et al.,* 2017), seemingly, the mutant strain shows low glycolytic and oxidative metabolism (Figure 4.10) at the later stage of growth since they did not show significant differential expression.

Although the mutant strain shows a relatively low number of DEGs during M2-M3 and M3-M4 intervals, most of those DEGs were mapped into different pathways such as biosynthesis of acarbose and validamycin biosynthesis, streptomycin, vancomycin, type II polyketide backbone and polyketide sugar unit, and tetracycline (Appendix 4g). In its chemical structure, vancomycin is a glycopeptide antibiotic that is synthesized via non-ribosomal peptide assembly (Yim *et al.,* 2014). However, differences in the pathway annotations of DEGs to these pathways do not explicitly confirm their synthesis of those metabolites. Hence, it is required to test a hypothesis that why the DEG were mapped to some other metabolites only in MT

although it had remarkably low number of DEGs. The DEGs from M3-M4 were mapped to the pathways for type II polyketide backbone (appendix 6) and polyketide sugar unit (appendix 7) were from Acl synthesizing BGC.

6. Conclusions

Results of the present study would make a better understanding of the gene expression pattern between the WT and MT strains and provide several genes and pathways for further investigation. The differential expression of some genes, *e.g.* TIGR04222 domain-containing membrane protein, serine protease, hypothetical protein (gene ID: fig|33899.16.peg.2270)/transcriptional regulator, nogalonic acid methyl ester cyclase, *aknC, aknX, aknA, aknE1, aknT*, and *aknS* residing inside the Acl producing BGC from the experimental strains throughout their growth cycle provide a better insight on Acl biosynthesis. This study also demonstrates a possible role of a nearby global transcriptional regulator (sigma), which should be studied along with other global regulators. Although GO enrichment analysis did not make any conclusive remarks, it clearly shows the functional classifications of the DEG between the strains. The KEGG pathway mapping of the DEGs showed that most of the mapped pathways were enriched on D2 and D3. The lower number of mapped genes in pantothenate and CoA biosynthesis, glycolysis, oxidative phosphorylation, two-component system, and ABC transporter pathways in the mutant strain need further analysis.

There were a few shortcomings in this study, some of which are outside the scope of this thesis work, that make drawing concrete conclusions from the results challenging. There was a high percentage of transcripts that are aligned with the WT genome multiple times possibly due existence of rRNA transcripts even after rRNA depletion. So there is a possibility of that the sequencing depth was not optimized that ultimately reflected in the DGE analysis. Additionally, DEG analysis did not consider mutations in the recycling genes and that is certainly beyond the scope of this thesis work and should be addressed in a future study. Additionally, the effect of a gene finally reflects after its translation. So a study on their translatome might broaden the understanding of the recycling gene(s). The signal from rare transcripts which may be involved in recycling should be analyzed by increasing the sequencing depths. A large portion of the annotated genes throughout the WT genome, within and nearby the Acl BGC were uncharacterized (hypothetical genes) and demonstrates lack of annotation and makes this study a formidable task. It is also visible from DEG list that a large portion of list contains hypothetical protein (appendix 9). These hypothetical proteins need to be characterized for better GO analysis and KEGG pathway mapping. A WT genome with better annotations can help in the discovery of new structure and functions of a protein which ultimately allow us to discover additional protein pathways. A much better interpretation of

hypothetical proteins can lead us to an improved GO enrichment and KEGG pathway mapping analysis. In this current situation KEGG mapping apparently make more sense than GO enrichment because it clearly shows the significant pathways where the DEGs are mapped into. However, GO enrichment analysis sometimes beneficial if the objective is to determine any certain state of an organism *e.g.* when it utilizes more carbohydrates. Such assertion can make a clearer picture over an organism but not enough to make an appropriate conclusion. Finally, an improved genome annotation and further molecular experiments, computational and gene co-expression expression analysis will be required to clarify this computational study.

References

Al-Shahrour, F., Díaz-Uriarte, R., & Dopazo, J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics applications note*, *20*(4), 578–580.

Alam, M. T., Merlo, M. E., Hodgson, D. A., Wellington, E. M. H., Takano, E., & Breitling, R. (2010). Metabolic modeling and analysis of the metabolic switch in *Streptomyces coelicolor*. *BMC Genomics*, *11*(1), 202.

Alam, M., Takano, E., & Breitling, R. (2011). Prioritizing orphan proteins for further study using phylogenomics and gene expression profiles in *Streptomyces coelicolor*. *BMC Research Notes*, *4*(1), 1–9.

Kallio, M. A., Tuimala, J. T., Hupponen, T., Klemelä, P., Gentile, M., Scheinin, I., Koski, M., Käki, J., & Korpelainen, E. I. (2011). Chipster: user-friendly analysis software for microarray and other high-throughput data. *BMC Genomics*, *12*(507), 1–14.

Alexeev, I., Sultana, A., Mäntsälä, P., Niemi, J., & Schneider, G. (2007). Aclacinomycin oxidoreductase (AknOx) from the biosynthetic pathway of the antibiotic aclacinomycin is an unusual flavoenzyme with a dual active site. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(15), 6170–6175.

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

Amos, G. C. A., Awakawa, T., Tuttle, R. N., Letzel, A. C., Kim, M. C., Kudo, Y., Fenical, W., Moore, B. S., & Jensen, P. R. (2017). Comparative transcriptomics as a guide to natural product discovery and biosynthetic gene cluster functionality. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(52), E11121–E11130.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*, R106.

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169.

Andrews, S. (2010). FastQC a quality control tool for high throughput sequence data. Retrieved May 15, 2020, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Arnison, P. G., Bibb, M. J., Bierbaum, G., Bowers, A. A., Bugni, T. S., Bulaj, G., Camarero, J. A.,… Van Der Donk, W. A. (2013, December 10). Ribosomally synthesized and post-translationally modified peptide natural products: Overview and recommendations for a universal nomenclature. *Natural Product Reports, 30*(1), 108-160.

Assefa, A. T., Vandesompele, J., & Thas, O. (2020). On the utility of RNA sample pooling to optimize cost and statistical power in RNA sequencing experiments. *BMC Genomics*, *21*(1), 312.

Auer, P. L., & Doerge, R. W. (2010). Statistical design and analysis of RNA sequencing data. *Genetics, 185*(2), 405-416.

Baba, T., & Schneewind, O. (1998). Instruments of microbial warfare: Bacteriocin synthesis, toxicity and immunity. *Trends in Microbiology*, *6*(2), 66–71.

Banchio, C., & Gramajo, H. (2002). A stationary-phase acyl-coenzyme A synthetase of *Streptomyces coelicolor* A3(2) is necessary for the normal onset of antibiotic production.

Applied and Environmental Microbiology, *68*(9), 4240–4246.

Baral, B., Akhgari, A., & Metsä-Ketelä, M. (2018). Activation of microbial secondary metabolic pathways: Avenues and challenges. *Synthetic and Systems Biotechnology*, *3*, 163–178.

Barka, E. A., Vatsa, P., Sanchez, L., Gaveau-Vaillant, N., Jacquard, C., Klenk, H.-P., Clément, C., Ouhdouch, Y., & van Wezel, G. P. (2016). Taxonomy, physiology, and natural products of actinobacteria. *Microbiology and Molecular Biology Reviews*, *80*(1), 1–43.

Baskakova, A. A., Gurina, E. I., & Tsyganov, V. A. (1981). Hydrolase activity characteristics of actinomycetes during antibiotic biosynthesis. *Antibiotiki*, *26*(3), 88–92.

Belknap, K. C., Park, C. J., Barth, B. M., & Andam, C. P. (2020). Genome mining of biosynthetic and chemotherapeutic gene clusters in *Streptomyces* bacteria. *Scientific Reports*, *10*(2003), 1-9.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59.

Beretta, G. L., & Zunino, F. (2007). Molecular mechanisms of anthracycline activity. *Topics in Current Chemistry*, *283*, 1–19.

Bibb, M., & Hesketh, A. (2009). Chapter 4 Analyzing the regulation of antibiotic production in *Streptomycetes*. *Methods in Enzymology*, *458*, 93–116.

Bignell, D. R. D., Tahlan, K., Colvin, K. R., Jensen, S. E., & Leskiw, B. K. (2005). Expression of ccaR, encoding the positive activator of cephamycin C and clavulanic acid production in *Streptomyces clavuligerus*, is dependent on bldG. *Antimicrobial Agents and Chemotherapy*, *49*(4), 1529–1541.

Bisang, C., Long, P. F., Cortés, J., Westcott, J., Crosby, J., Matharu, A. L., … Leadlay, P. F. (1999). A chain initiation factor common to both modular and aromatic polyketide synthases. *Nature*, *401*(6752), 502–505.

Blin, K., Kim, H. U., Medema, M. H., & Weber, T. (2019). Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Briefings in Bioinformatics*, *20*(4), 1103–1113.

Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S. Y., Medema, M. H., & Weber, T. (2019). antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Research*, *47*(W1), 81–87.

Blin, K., Wolf, T., Chevrette, M. G., Lu, X., Schwalen, C. J., Kautsar, ….Medema, M. H. (2017). antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Research*, *45*, 36–41.

Borodina, I., Krabben, P., & Nielsen, J. (2005). Genome-scale analysis of Streptomyces coelicolor A3(2) metabolism. *Genome Research*, *15*(6), 820–829.

Bosso, J. A., Mauldin, P. D., & Salgado, C. D. (2010). The association between antibiotic use and resistance: The role of secondary antibiotics. *European Journal of Clinical Microbiology and Infectious Diseases*, *29*, 1125–1129.

Brötz-Oesterhelt, H., & Brunner, N. A. (2008). How many modes of action should an antibiotic have? *Current Opinion in Pharmacology*, *8*(5), 564–573.

Callister, S. J., McCue, L. A., Turse, J. E., Monroe, M. E., Auberry, K. J., Smith, R. D., Adkins, J. N., & Lipton, M. S. (2008). Comparative bacterial proteomics: analysis of the core genome concept. (J. Fraser, Ed.)*PLoS ONE*, *3*(2), e1542.

Carver, T., Harris, S. R., Berriman, M., Parkhill, J., & Mcquillan, J. A. (2012). Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*, *28*(4), 464–469.

Caspi, R., Billington, R., Fulcher, C. A., Keseler, I. M., Kothari, A., Krummenacker, M., … Karp, P. D. (2018). The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Research*, *46*(D1), D633–D639.

Chopra, I., Hesse, L., & O'Neill, A. J. (2002). Exploiting current understanding of antibiotic action for discovery of new drugs. *Journal of Applied Microbiology*, *92*(1), 4S-15S.

Christodoulou, D. C., Gorham, J. M., Herman, D. S., & Seidman, J. G. (2011). Construction of normalized rna-seq libraries for next-generation sequencing using the crab duplex-specific nuclease. In Frederick, M. A. (Ed.), *Current Protocols in Molecular Biology*,*94*, (4.12.1-4.12.11).

Chung, J. young, Fujii, I., Harada, S., Sankawa, U., & Ebizuka, Y. (2002). Expression, purification, and characterization of AknX anthrone oxygenase, which is involved in aklavinone biosynthesis in *Streptomyces galilaeus*. *Journal of Bacteriology*, *184*(22), 6115–6122.

Cloonan, N., Forrest, A. R. R., Kolle, G., Gardiner, B. B. A., Faulkner, G. J., Brown, M. K., T…Grimmond, S. M. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods*, *5*(7), 613–619.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatic Application Note*, *21*(18), 3674–3676. Retrieved May 12, 2020, from http://www.blast2go.de

Corvini, P. F. X., Delaunay, S., Maujean, F., Rondags, E., Vivier, H., Goergen, J. L., & Germain, P. (2004). Intracellular pH of *Streptomyces pristinaespiralis* is correlated to the sequential use of carbon sources during the pristinamycins-producing process. *Enzyme and Microbial Technology*, *34*(2), 101–107.

Costa-Silva, J., Domingues, D., & Lopes, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. (Z. Wei, Ed.)*PLOS ONE*, *12*(12), e0190152.

Culviner, P. H., Guegler, C. K., & Laub, M. T. (2020). A simple, cost-effective, and robust method for rrna depletion in rna-sequencing studies. *mBio*, *11*(2).

Cummings, M., Breitling, R., & Takano, E. (2014). Steps towards the synthetic biology of polyketide biosynthesis. *FEMS Microbiology Letters*, *351*(2), 116–125.

Dao, P., Numanagić, I., Lin, Y. Y., Hach, F., Karakoc, E., Donmez, N., Collins, C., Eichler, E. E., & Sahinalp, S. C. (2014). ORMAN: Optimal resolution of ambiguous RNA-Seq multimappings in the presence of novel isoforms. *Bioinformatics*, *30*(5), 644–651.

Dessimoz, C., & Škunca, N. (Eds.). (2017). *The Gene Ontology Handbook*. Methods in Molecular Biology, *1446*.

Dimmer, E. C., Huntley, R. P., Alam-Faruque, Y., Sawford, T., O'Donovan, C., Martin, M. J.,

…. Apweiler, R. (2012). The UniProt-GO Annotation database in 2011. *Nucleic Acids Research*, *40*(D1), D565–D570.

Dündar, F., Skrabanek, L., & Zumbo, P. (2015). *Introduction to differential gene expression analysis using RNA-seq*. Retrieved May 6, 2020, from https://chagall.med.cornell.edu/RNASEQcourse/Intro2RNAseq.pdf

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., ... Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, *323*(5910), 133–138.

Fang, Z., & Cui, X. (2011). Design and validation issues in RNA-seq experiments. *Briefings in Bioinformatics*, *12*(3), 280–287.

Flaherty, R. A., Freed, S. D., & Lee, S. W. (2014). The wide world of ribosomally encoded bacterial peptides. (Miller, V. Ed.)*PLoS Pathogens*, *10*(7), e1004221.

Flärdh, K., & Buttner, M. J. (2009). *Streptomyces* morphogenetics: Dissecting differentiation in a filamentous bacterium. *Nature Reviews Microbiology*, *7*(1), 36–49.

Fujii, I., & Ebizuka, Y. (1997). Anthracycline biosynthesis in *Streptomyces galilaeus. Chemical Reviews*, *97*(7), 2511–2523.

Gene Ontology Annotation. (n.d.). . Retrieved May 18, 2020, from http://docs.blast2go.com/user-manual/gene-ontology-annotation/

Ginther, C. L. (1979). Sporulation and the production of serine protease and cephamycin C by *Streptomyces lactamdurans*. *Antimicrobial Agents and Chemotherapy*, *15*(4), 522–526.

GöTz, S., García-GóMez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., … Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, *36*(10), 3420–3435.

Gräfe, U., Dornberger, K., Fleck, W. F., & Freysoldt, C. (1988). Compartmentation of enzymes interconverting aclacinomycins in *Streptomyces* species AM 33352. *Journal of Basic Microbiology*, *28*(1–2), 17–23.

Haas, B. J., Chin, M., Nusbaum, C., Birren, B. W., & Livny, J. (2012). How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? *BMC Genomics*, *13*(734).

Haiser, H. J., Yousef, M. R., & Elliot, M. A. (2009). Cell wall hydrolases affect germination, vegetative growth, and sporulation in *Streptomyces coelicolor. Journal of bacteriology*, *191*(21), 6501–6512.

Hertweck, C., Luzhetskyy, A., Rebets, Y., & Bechthold, A. (2007). Type II Polyketide Synthases: Gaining a Deeper Insight into Enzymatic Teamwork. *ChemInform*, *24*, 162–190.

Van Der Heul, H. U., Bilyk, B. L., McDowall, K. J., Seipke, R. F., & Van Wezel, G. P. (2018). Regulation of antibiotic production in *Actinobacteria*: New perspectives from the post-genomic era. *Natural Product Reports*, 1–30.

Hinderer, E. W., Flight, R. M., Dubey, R., MacLeod, J. N., & Moseley, H. N. B. (2019). Advances in gene ontology utilization improve statistical power of annotation enrichment. (K. Abe, Ed.)*PLOS ONE*, *14*(8), e0220728.

Hodgson, D. A. (2000). Primary metabolism and its control in *Streptomycetes*: A most unusual group of bacteria. *Advances in Microbial Physiology*.

Holmes, N. A., Innocent, T. M., Heine, D., Al Bassam, M., Worsley, S. F., Trottmann, F., …

Hutchings, M. I. (2016). Genome analysis of two *Pseudonocardia* phylotypes associated with acromyrmex leafcutter ants reveals their biosynthetic potential. *Frontiers in Microbiology*, *7*(2073).

Hölzer, M., & Marz, M. (2019). *De novo* transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. *GigaScience*, *8*(5), 1–16.

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, *37*(1), 1–13.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., …. Clamp, M. (2002). The Ensembl genome database project. *Nucleic Acids Research*, *30*(1), 38–41.

Huntley, R. P., Sawford, T., Martin, M. J., & O'Donovan, C. (2014). Understanding how and why the Gene Ontology and its annotations evolve: The GO within UniProt. *GigaScience*, *3*(4), 1–9.

Ijaq, J., Malik, G., Kumar, A., Das, P. S., Meena, N., Bethi, N., Sundararajan, V. S., & Suravajhala, P. (2019). A model to predict the function of hypothetical proteins through a nine-point classification scoring schema. *BMC Bioinformatics*, *20*(1), 14.

Iftime, D., Kulik, A., Härtner, T., Rohrer, S., Niedermeyer, T. H. J., Stegmann, E., Weber, T., & Wohlleben, W. (2016). Identification and activation of novel biosynthetic gene clusters by genome mining in the kirromycin producer *Streptomyces collinus* Tü 365. *Journal of Industrial Microbiology and Biotechnology*, *43*(2–3), 277–291.

Itoh, T., Taguchi, T., Kimberley, M. R., Booker-Milburn, K. I., Stephenson, G. R., Ebizuka, Y., & Ichinose, K. (2007). Actinorhodin biosynthesis: Structural requirements for post-PKS tailoring intermediates revealed by functional analysis of actVI-ORFl reductase. *Biochemistry*, *46*(27), 8181–8188.

Jia, N., Ding, M. Z., Luo, H., Gao, F., & Yuan, Y. J. (2017). Complete genome sequencing and antibiotics biosynthesis pathways analysis of *Streptomyces lydicus* 103. *Scientific Reports*, *7*(1), 1–8.

Kato, J. ya, Suzuki, A., Yamazaki, H., Ohnishi, Y., & Horinouchi, S. (2002). Control by A-factor of a metalloendopeptidase gene involved in aerial mycelium formation in *Streptomyces griseus*. *Journal of Bacteriology*, *184*(21), 6016–6025.

Kautsar, S. A., Blin, K., Shaw, S., Navarro-Muñoz, J. C., Muñoz, M., Terlouw, B. R., …Medema, M. H. (2020). MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Research*, *48*(D1), D454-D458.

Kelemen, G. H., Brian, P., Flärdh, K., Chamberlin, L., Chater, K. F., & Buttner, M. J. (1998). Developmental regulation of transcription of whiE, a locus specifying the polyketide spore pigment in *Streptomyces coelicolor* A3(2). *Journal of Bacteriology*, *180*(9), 2515–2521.

Khatri, P., & Dr, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatic review*, *21*(18), 3587–3595.

Kim, J. N., Kim, Y., Jeong, Y., Roe, J. H., Kim, B. G., & Cho, B. K. (2015). Comparative genomics reveals the core and accessory genomes of *Streptomyces* species. *Journal of Microbiology and Biotechnology*, *25*(10), 1599–1605.

Kjærbølling, I., Mortensen, U. H., Vesth, T., & Andersen, M. R. (2019). Strategies to establish

the link between biosynthetic gene clusters and secondary metabolites. *Fungal Genetics and Biology*, *130*, 107–121.

Kukurba, K. R., & Montgomery, S. B. (2015). RNA sequencing and analysis. *Cold Spring Harbor Protocols*, *2015*(11), 951–969.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359.

Li, W. V., & Li, J. J. (2018). Modeling and analysis of RNA-seq data: a review from a statistical perspective. *Quantitative Biology*, *6*(3), 195–209.

Li, W., Ying, X., Guo, Y., Yu, Z., Zhou, X., Deng, Z., … Tao, M. (2006). Identification of a gene negatively affecting antibiotic production and morphological differentiation in *Streptomyces coelicolor* A3(2). *Journal of Bacteriology*, *188*(24), 8368–8375.

de Lima Procópio, R. E., da Silva, I. R., Martins, M. K., de Azevedo, J. L., & de Araújo, J. M. (2012). Antibiotics produced by *Streptomyces*. *Brazilian Journal of Infectious Diseases*, *16*(6), 466–471.

Liu, G., Chater, K. F., Chandra, G., Niu, G., & Tan, H. (2013). Molecular regulation of antibiotic biosynthesis in *Streptomyces*. *Microbiology and Molecular Biology Reviews*, *77*(1), 112–143.

Loraine, A. E., Blakley, I. C., Jagadeesan, S., Harper, J., Miller, G., & Firon, N. (2015). Analysis and visualization of RNA-Seq expression data using rstudio, bioconductor, and integrated genome browser. *Plant Functional Genomics: Methods and Protocols: Second Edition, 1284*, (481–501).

Lu, W., Leimkuhler, C., Gatto, G. J., Kruger, R. G., Oberthür, M., Kahne, D., & Walsh, C. T. (2005). AknT is an activating protein for the glycosyltransferase AknS in L-aminodeoxysugar transfer to the aglycone of aclacinomycin A. *Chemistry and Biology*, *12*(5), 527–534.

Malik, A., Kim, Y. R., & Kim, S. B. (2020). Genome mining of the genus *Streptacidiphilus* for biosynthetic and biodegradation potential. *Genes*, *11*(10), 1–31.

Marguerat, S., & Bähler, J. (2010). RNA-seq: From technology to biology. *Cellular and Molecular Life Sciences*, *67*(4), 569–579.

Martin, R., Sterner, O., Alvarez, M. A., De Clercq, E., Bailey, J. E., & Minas, W. (2001). Collinone, a new recombinant angular polyketide antibiotic made by an engineered *Streptomyces* strain. *Journal of Antibiotics*, *54*(3), 239–249.

Matulova, M., Feckova, L., Novakova, R., Mingyar, E., Csolleiova, D., Zduriencikova, M., … Kormanec, J. (2019). A structural analysis of the angucycline-like antibiotic auricin from *Streptomyces lavendulae* subsp. Lavendulae CCM 3239 revealed its high similarity to griseusins. *Antibiotics*, *8*(102), 1–12.

McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, *40*(10), 4288–4297.

Medema, M. H., Alam, M. T., Heijne, W. H. M., van den Berg, M. A., Müller, U., Trefzer, A., Bovenberg, R. A. L., Breitling, R., & Takano, E. (2011). Genome-wide gene expression changes in an industrial clavulanic acid overproduction strain of *Streptomyces clavuligerus*. *Microbial Biotechnology*, *4*(2), 300–305.

Medema, M. H., Cimermancic, P., Sali, A., Takano, E., & Fischbach, M. A. (2014). A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. *PLoS Computational Biology*, *10*(12), 1–12.

Méndez, C., & Salas, J. A. (2001). The role of ABC transporters in antibiotic-producing organisms: Drug secretion and resistance mechanisms. *Research in Microbiology*, *152*(3–4), 341–350.

Metsä-Ketelä, M., Niemi, J., Mäntsälä, P., & Schneider, G. (2007). Anthracycline biosynthesis: genes, enzymes and mechanisms. *Anthracycline Chemistry and Biology I 282*, (101–140).

Millan-Oropeza, A., Henry, C., Blein-Nicolas, M., Aubert-Frambourg, A., Moussa, F., Bleton, J., & Virolle, M. J. (2017). quantitative proteomics analysis confirmed oxidative metabolism predominates in *Streptomyces coelicolor* versus glycolytic metabolism in *Streptomyces lividans*. *Journal of Proteome Research*, *16*(7), 2597–2613.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., & Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research*, *35*, 182–185.

Nelson, N. J. (2001). Microarrays Have Arrived: Gene Expression Tool Matures. *JNCI: Journal of the National Cancer Institute*, *93*(7), 492–494.

Netzker, T., Flak, M., Krespach, M. K., Stroe, M. C., Weber, J., Schroeckh, V., & Brakhage, A. A. (2018). Microbial interactions trigger the production of antibiotics. *Current Opinion in Microbiology, 45*, 117-123.

Nguyen, H. C., Karray, F., Lautru, S., Gagnat, J., Lebrihi, A., Huynh, T. D. H., & Pernodet, J. L. (2010). Glycosylation steps during spiramycin biosynthesis in *Streptomyces ambofaciens*: Involvement of three glycosyltransferases and their interplay with two auxiliary proteins. *Antimicrobial Agents and Chemotherapy*, *54*(7), 2830–2839.

Nieselt, K., Battke, F., Herbig, A., Bruheim, P., Wentzel, A., Jakobsen, Ø. M., … Wellington, E, M. H. (2010). The dynamic architecture of the metabolic switch in *Streptomyces coelicolor*. *BMC Genomics*, *11*, 1–9.

Niu, G. Q., & Tan, H. R. (2013). Biosynthesis and regulation of secondary metabolites in microorganisms. *Science China Life Sciences*, *56*(7), 581–583.

Nützmann, H. W., Huang, A., & Osbourn, A. (2016). Plant metabolic clusters – from genetics to genomics. *New Phytologist, 211*(3), 771-789.

Ogasawara, Y., Yackley, B. J., Greenberg, J. A., Rogelj, S., & Melançon, C. E. (2015). Expanding our understanding of sequence-function relationships of type ii polyketide biosynthetic gene clusters: bioinformatics-guided identification of frankiamicin a from *Frankia* sp. EAN1pec. *PLOS ONE*, *10*(4), e0121505.

Oki, T., Kitamura, I., Matsuzawa, Y., Shibamoto, N., Ogasawara, T., … Umezawa, H. (1979). Antitumor anthracycline antibiotics, aclacinomycin A and analogues. II. Structural determination. *The Journal of Antibiotics*, *32*(8), 801–819.

Oki, T., Matsuzawa, Y., Yoshimoto, A., Numata, K., Kitamura, I., Hori, S., … Takeuchi, T. (1975). New antitumor antibiotics, aclacinomycins A and B. *The Journal of Antibiotics*, *28*(10), 830–834.

Oliynyk, M., Samborskyy, M., Lester, J. B., Mironenko, T., Scott, N., Dickens, S., Haydock, S. F., & Leadlay, P. F. (2007). Complete genome sequence of the erythromycin-producing

bacterium *Saccharopolyspora erythraea* NRRL23338. *Nature Biotechnology*, *25*, 447–453.

OmicsBox | BioBam | Bioinformatics Made Easy. (2019, March 3). . Retrieved May 12, 2020, from https://www.biobam.com/omicsbox/

Omura, S., Ikeda, H., Ishikawa, J., Hanamoto, A., Takahashi, C., Shinose, M., … Hattori, M. (2001). Genome sequence of an industrial microorganism Streptomyces avermitilis: Deducing the ability of producing secondary metabolites. *Proceedings of the National Academy of Sciences of the United States of America*, *98*, 12215–12220.

Osbourn, A. (2010). Secondary metabolic gene clusters: Evolutionary toolkits for chemical innovation. *Trends in Genetics*, *26*, 449–457.

Oshlack, A., Robinson, M. D., & Young, M. D. (2010). From RNA-seq reads to differential expression results. *Genome Biology*, *11*(12), 1–10.

Ou, X., Zhang, B., Zhang, L., Dong, K., Liu, C., Zhao, G., & Ding, X. (2008). SarA influences the sporulation and secondary metabolism in *Streptomyces coelicolor* M145. *Acta Biochimica et Biophysica Sinica*, *40*(10), 877–882.

Paul, C. E., & Fernández, V. G. (2016). Biocatalysis and biotransformation in ionic liquids. *ionic liquids in lipid processing and analysis: opportunities and challenges,* 11–58.

Pinilla, L., Toro, L. F., Laing, E., Alzate, J. F., & Ríos-Estepa, R. (2019. Comparative transcriptome analysis of *Streptomyces clavuligerus* in response to favorable and restrictive nutritional conditions. *Antibiotics*, *8*(3), 96.

Raty, K., Hautala, A., Torkkell, S., Kantola, J., Ma, P., Hakala, J., & Ylihonko, K. (2002). Characterization of mutations in aclacinomycin A-non-producing *Streptomyces galilaeus* strains with altered glycosylation patterns. *Microbiology*, *148*, 3375–3384.

Räty, K., Kunnari, T., Hakala, J., Mäntsälä, P., & Ylihonko, K. (2000). A gene cluster from *Streptomyces galilaeus* involved in glycosylation of aclarubicin. *Molecular and General Genetics*, *264*(1–2), 164–172.

Räty, Kaj, Kantola, J., Hautala, A., Hakala, J., Ylihonko, K., & Mäntsälä, P. (2002). Cloning and characterization of *Streptomyces galilaeus* aclacinomycins polyketide synthase (PKS) cluster. *Gene*, *293*(1–2), 115–122.

Raveh, A., Delekta, P. C., Dobry, C. J., Peng, W., Schultz, P. J., Blakely, P. K., … Miller, D. J. (2013). Discovery of potent broad spectrum antivirals derived from marine actinobacteria. *PLoS ONE*, *8*(12), 82318.

Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., … Bader, G. D. (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and Enrichment Map. *Nature Protocols*, *14*(2), 482–517.

Relations in the Gene Ontology. (n.d.). . Retrieved May 8, 2020, from http://geneontology.org/docs/ontology-relations/

Richter, L., Wanka, F., Boecker, S., Storm, D., Kurt, T., Vural, Ö., Süßmuth, R., & Meyer, V. (2014). Engineering of *Aspergillus niger* for the production of secondary metabolites. *Fungal Biology and Biotechnology*, *1*(1), 4.

Ridley, C. P., Ho, Y. L., & Khosla, C. (2008). Evolution of polyketide synthases in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *105*,

4595–4600.

Risdian, C., Mozef, T., & Wink, J. (2019). Biosynthesis of polyketides in *Streptomyces*. *Microorganisms, 7*(5), 124.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47.

Robinson, M. D., Mccarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics Applications Note*, *26*(1), 139–140.

Rodríguez, H., Rico, S., Díaz, M., & Santamaría, R. I. (2013). Two-component systems in *Streptomyces*: key regulators of antibiotic complex pathways. *Microbial Cell Factories*, *12*(127).

Rokem, J. S., Lantz, A. E., & Nielsen, J. (2007a). Systems biology of antibiotic production by microorganisms. *Natural Product Reports*, *24*(6), 1262–1287.

Roncaglia, P., Martone, M. E., Hill, D. P., Berardini, T. Z., Foulger, R. E., Imam, F. T., … Lomax, J. (2013). The Gene Ontology (GO) Cellular Component Ontology: Integration with SAO (Subcellular Anatomy Ontology) and other recent developments. *Journal of Biomedical Semantics*, *4*(1), 20.

Ryu, Y. G., Butler, M. J., Chater, K. F., & Lee, K. J. (2006). Engineering of primary carbohydrate metabolism for increased production of actinorhodin in *Streptomyces coaicolor*. *Applied and Environmental Microbiology*, *72*(11), 7132–7139.

Salas, J. A., Quiros, L. M., & Hardisson, C. (1984). Pathways of glucose catabolism during germination of *Streptomyces* spores. *FEMS Microbiology Letters*, *22*(3), 229–233.

Salem, S. M., Weidenbach, S., & Rohr, J. (2017). Two cooperative glycosyltransferases are responsible for the sugar diversity of saquayamycins isolated from *Streptomyces* sp. KY 40-1. *ACS Chemical Biology*, *12*(10), 2529–2534.

Scherlach, K., Graupner, K., & Hertweck, C. (2013). Molecular bacteria-fungi interactions: effects on environment, food, and medicine. *Annual Review of Microbiology*, *67*(1), 375–397.

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., … Barton, G. J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, *22*(6), 839–851.

Schwecke, T., Aparicio, J. F., Molnár, I., König, A., Khaw, L. E., Haydock, S. F., … Leadlay, P. F. (1995). The biosynthetic gene cluster for the polyketide immunosuppressant rapamycin. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(17), 7839–7843.

Scott, T. A., & Piel, J. (2019). The hidden enzymology of bacterial natural product biosynthesis. *Nature Reviews Chemistry*, *3*(7), 404–425.

Seipke, R. F., Kaltenpoth, M., & Hutchings, M. I. (2012). *Streptomyces* as symbionts: An emerging and widespread theme? *FEMS Microbiology Reviews*, *36*(4), 862–876.

Shen, B. (2003). Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Current Opinion in Chemical Biology, 7*(2), 285-295.

Shen, W., Wang, D., Wei, L., & Zhang, Y. (2020). Fungal elicitor-induced transcriptional changes of genes related to branched-chain amino acid metabolism in *Streptomyces natalensis* HW-2. *Applied Microbiology and Biotechnology*, *104*(10), 4471–4482.

Škunca, N., Altenhoff, A., & Dessimoz, C. (2012). Quality of computationally inferred gene ontology annotations. *PLoS Computational Biology*, *8*(5).

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics*, *20*(11), 631–656.

Staunton, J., & Weissman, K. J. (2001). Polyketide biosynthesis: A millennium review. *Natural Product Reports*, *18*(4), 380–416.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., … Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550.

Tang, G. L., Zhang, Z., & Pan, H. X. (2017). New insights into bacterial type II polyketide biosynthesis. *F1000Research*, *6*, 172.

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., & Conesa, A. (2011). Differential expression in RNA-seq: A matter of depth. *Genome Research*, *21*(12), 2213–2223.

Torkkell, S., Kunnari, T., Palmu, K., Hakala, J., Mäntsälä, P., Mäntsä, M., … Ylihonko, K. (2000). Identification of a cyclase gene dictating the c-9 stereochemistry of anthracyclines from *Streptomyces nogalater*. *Antimicrobialagents and Chemotherapy*, *44*(2), 396–399.

Tran, P. N., Yen, M. R., Chiang, C. Y., Lin, H. C., & Chen, P. Y. (2019). Detecting and prioritizing biosynthetic gene clusters for bioactive compounds in bacteria and fungi. *Applied Microbiology and Biotechnology*, *103*, 3277–3287.

TUFTS. (n.d.). Genomics. Retrieved June 11, 2020, from http://tucf-genomics.tufts.edu/home/ordering

Velculescu, V. E., Zhang, L., Vogelstein, B., & Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, *270*(5235), 484–487.

Watanabe, A., & Ebizuka, Y. (2004). Unprecedented mechanism of chain length determination in fungal aromatic polyketide synthases. *Chemistry and Biology*, *11*(8), 1101–1106.

Watkins, P. A. (1997). Fatty acid activation. *Progress in Lipid Research*, *36*(1), 55–83.

Wei, J., He, L., & Niu, G. (2018). Regulation of antibiotic biosynthesis in actinomycetes: Perspectives and challenges. *Synthetic and Systems Biotechnology*, *3*(4), 229–235.

Westermann, A. J., Gorski, S. A., & Vogel, J. (2012). Dual RNA-seq of pathogen and host. *Nature Reviews Microbiology*, *10*(9), 618–630.

Van Wezel, G. P., & McDowall, K. J. (2011). The regulation of the secondary metabolism of *Streptomyces*: New links and experimental advances. *Natural Product Reports*, *28*(7), 1311–1333.

Wilkens, S. (2015). Structure and mechanism of ABC transporters. *F1000Prime Reports*, *7*.

Yim, G., Thaker, M. N., Koteva, K., & Wright, G. (2014). Glycopeptide antibiotic biosynthesis. *Journal of Antibiotics*, *67*(1), 31–41.

Ylihonko, K., Hakala, J., Niemi, J., & Lundel, J. (1994). Isolation and characterization of

aclacinomycin A-non-producing i (ATCC 31615) mutants. *Microbiology*, *140*, 1359–1365.

Young, M. D., Wakefield, M. J., Smyth, G. K., & Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, *11*(2), R14.

Yu, D., Xu, F., Zeng, J., & Zhan, J. (2012). Type III polyketide synthases in natural product biosynthesis. *IUBMB Life*, *64*(4), 285–295.

Yunshun, C., Davis, M., Matthew, R., Mark, R., Gordon. S. & Eliza, H. (2020). edgeR: differential analysis of sequence read count data User's Guide. Retrived from http://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf

Zhang, Z. H., Jhaveri, D. J., Marshall, V. M., Bauer, D. C., Edson, J., Narayanan, R. K., … Zhao, Q.-Y. (2014). A comparative study of techniques for differential expression analysis on rna-seq data. (Provero, P. Ed.)*PLoS ONE*, *9*(8), e103207.

Zhao, S., Zhang, B., Zhang, Y., Gordon, W., Du, S., Paradis, T., Vincent, M., & von Schack, D. (2016). Bioinformatics for rna-seq data analysis. *bioinformatics - updated features and applications*. *DOI: 10.5772/63267*

Zhou, T., Yao, J., & Liu, Z. (2017). Gene Ontology, Enrichment Analysis, and Pathway Analysis. *Bioinformatics in Aquaculture,* 150–168.

Zhou, Z., Gu, J., Du, Y.-L., Li, Y.-Q., & Wang, Y. (2011). The -omics Era- Toward a Systems-Level Understanding of *Streptomyces*. *Current Genomics*, *12*(6), 404–416.

Zyprych-Walczak, J., Szabelska, A., Handschuh, L., Górczak, K., Klamecka, K., Figlerowicz, M., & Siatkowski, I. (2015). The Impact of Normalization Methods on RNA-Seq Data Analysis. *BioMed Research International*, *2015*, 621–631.

Appendices

Appendix 1

list of top 20 genes those are up and downregulated most.

a) Between the strains (WT Vs MT)

i) Day 1

Gene ID	Length	LogFC	Gene product
fig|33899.16.peg.4096	456	8.540331	Hypothetical protein
fig|33899.16.peg.1277	825	8.006175	putative tyrosinase
fig|33899.16.peg.4311	153	7.714794	Hypothetical protein
fig|33899.16.peg.679	1581	7.714794	Putative glycosyltransferase
fig|33899.16.peg.508	1614	7.257936	alpha-L-arabinofuranosidase II
fig|33899.16.peg.546	450	7.128077	Hypothetical protein
fig|33899.16.peg.3941	618	6.846037	Hypothetical protein
fig|33899.16.peg.5923	1329	6.846037	Hypothetical protein
fig|33899.16.peg.5946	408	6.846037	Long-chain-fatty-acid--CoA ligase
fig|33899.16.peg.5948	1074	6.846037	CrtT-methyltransferase-like protein
fig|33899.16.peg.1637	1185	-9.2381	Hypothetical protein
fig|33899.16.peg.1646	807	-9.89699	Hypothetical protein
fig|33899.16.peg.1642	1725	-9.93482	C-5 cytosine-specific DNA methylase family protein
fig|33899.16.peg.1700	249	-9.98617	Hypothetical protein

fig|33899.16.peg.1690	318	-10.0496	Hypothetical protein
fig|33899.16.peg.1702	444	-10.2125	Phage protein
fig|33899.16.peg.1692	435	-10.9228	Hypothetical protein
fig|33899.16.peg.1681	1161	-11.0763	Hypothetical protein
fig|33899.16.peg.3899	696	-11.2149	Hypothetical protein
fig|33899.16.peg.1685	558	-12.1678	Hypothetical protein

ii) Day 2

Transcript Id	**Length**	**LogFC**	**Gene product**
fig|33899.16.peg.5924	1635	8.481446	Polyketide synthase modules and related proteins
fig|33899.16.peg.5946	408	8.068018	Long-chain-fatty-acid--CoA ligase
fig|33899.16.peg.1866	957	7.831675	Multicopper oxidase
fig|33899.16.peg.564	603	7.831675	Capsular polysaccharide biosynthesis protein
fig|33899.16.peg.5947	135	7.761602	Hypothetical protein
fig|33899.16.peg.8368	150	7.761602	Hypothetical protein
fig|33899.16.peg.5566	804	7.438068	Extracellular ribonuclease Bsn
fig|33899.16.peg.2475	831	7.348779	ABC transporter permease protein 2
fig|33899.16.peg.6551	1329	7.218092	lpha-glucosides-binding periplasmic protein AglE precursor

fig|33899.16.peg.2218	396	7.151316	Guanyl-specific ribonuclease
fig|33899.16.peg.1645	372	-7.32031	Hypothetical protein
fig|33899.16.peg.4803	495	-7.48284	Hypothetical protein
fig|33899.16.peg.1700	249	-7.58184	Hypothetical protein
fig|33899.16.peg.1702	444	-7.71865	Phage protein
fig|33899.16.peg.1697	435	-7.88297	Hypothetical protein
fig|33899.16.peg.4542	120	-7.88297	Hypothetical protein
fig|33899.16.peg.6153	1173	-8.09893	Hypothetical protein
fig|33899.16.peg.1628	243	-8.34427	Hypothetical protein
fig|33899.16.peg.4813	789	-9.31352	Transcriptional regulator
fig|33899.16.peg.4815	432	-10.3267	putative regulatory protein

iii) Day 3:

Transcript ID	**length**	**LogFC**	**Gene product**
fig|33899.16.peg.2475	831	7.828807	ABC transporter permease protein 2
fig|33899.16.peg.6703	699	7.279224	Putative oxidoreductase
fig|33899.16.peg.540	462	7.180351	Hypothetical protein
fig|33899.16.peg.1089	324	7.070788	Alpha-amylase inhibitor HAIM II precursor
fig|33899.16.peg.1156	1050	6.698898	Sorbitol dehydrogenase

fig|33899.16.peg.5002	981	6.548437	Hypothetical protein
fig|33899.16.peg.5937	1053	6.427157	Endo-1 4-beta-xylanase
fig|33899.16.peg.4373	888	6.380437	Hypothetical protein
fig|33899.16.peg.541	519	6.380437	Mobile element protein
fig|33899.16.peg.565	321	6.380437	Hypothetical protein
fig|33899.16.peg.1698	534	-7.17393	hypothetical protein
fig|33899.16.peg.5964	1065	-7.17963	transport regulator
fig|33899.16.peg.4814	432	-7.29101	Nudix hydrolase family protein
fig|33899.16.peg.4816	384	-7.30706	hypothetical protein
fig|33899.16.peg.1697	435	-7.31747	hypothetical protein
fig|33899.16.peg.8545	1047	-7.96393	ABC transporter substrate-binding protein
fig|33899.16.peg.4813	789	-9.98776	Transcriptional regulator KorSA GntR family
fig|33899.16.peg.1637	1185	-10.0425	hypothetical protein
fig|33899.16.peg.1629	192	-10.8658	hypothetical protein
fig|33899.16.peg.4815	432	-11.1866	Putative regulatory protein

iv) Day 4

Transcript ID	length	LogFC	Gene product
fig|33899.16.peg.5949	387	9.356733	hypothetical protein
fig|33899.16.peg.797	534	9.128841	hypothetical protein
fig|33899.16.peg.6552	1362	8.977472	ABC alpha-glucoside transporter inner membrane subunit AglF
fig|33899.16.peg.8368	150	8.524464	hypothetical protein
fig|33899.16.peg.2475	831	8.249654	ABC transporter permease protein 2
fig|33899.16.peg.5946	408	8.058871	Long-chain-fatty-acid--CoA ligase
fig|33899.16.peg.5943	3951	7.970473	Polyketide synthase modules and related proteins
fig|33899.16.peg.6551	1329	7.941416	Alpha-glucosides-binding periplasmic protein AglE precursor
fig|33899.16.peg.6553	840	7.777146	Alpha-glucoside transport system permease protein AglG
fig|33899.16.peg.5945	1053	7.700258	hypothetical protein
fig|33899.16.peg.1646	807	-8.82861	hypothetical protein
fig|33899.16.peg.1690	318	-9.32519	hypothetical protein
fig|33899.16.peg.1700	249	-9.38399	hypothetical protein
fig|33899.16.peg.1636	939	-9.39547	hypothetical protein
fig|33899.16.peg.1701	321	-9.9094	hypothetical protein
fig|33899.16.peg.4813	789	-10.0395	Transcriptional regulator

fig|33899.16.peg.1702	444	-10.152	phage protein
fig|33899.16.peg.727	171	-10.2501	hypothetical protein
fig|33899.16.peg.1629	192	-10.5045	hypothetical protein
fig|33899.16.peg.2966	1749	-10.6837	ABC transporter permease protein 1

b) Within the strain

i) Comparison between day 1 and 2 within MT.

Transcript ID	**length**	**LogFC**	**Gene product**
fig|33899.16.peg.2253	651	10.92496	Hypothetical Protein
fig|33899.16.peg.6551	1329	10.7782	Alpha-glucosides-binding periplasmic protein AglE precursor
fig|33899.16.peg.796	969	9.802582	ABC transporter ATP-binding protein
fig|33899.16.peg.6552	1362	9.714246	ABC alpha-glucoside transporter inner membrane subunit AglF
fig|33899.16.peg.555	369	9.519479	UPF0145 protein SCO3412
fig|33899.16.peg.8061	741	8.842586	Hypothetical Protein
fig|33899.16.peg.8078	897	8.521443	Universal stress protein
fig|33899.16.peg.6550	1674	8.403135	Alpha-glucosidase AglA
fig|33899.16.peg.6553	840	8.403135	Alpha-glucoside transport system permease protein AglG
fig|33899.16.peg.5558	1158	8.340132	Acyl-CoA dehydrogenase
fig|33899.16.peg.5163	879	-5.1501	Uncharacterized metal-dependent hydrolase

fig|33899.16.peg.5162	411	-5.50215	Hypothetical protein
fig|33899.16.peg.509	1647	-5.60125	alpha-galactosidase
fig|33899.16.peg.4714	741	-5.88076	Carbonic anhydrase
fig|33899.16.peg.2256	1608	-5.9352	Long-chain-fatty-acid--CoA ligase
fig|33899.16.peg.4608	1596	-6.39027	Two-component system sensor histidine kinase
fig|33899.16.peg.4748	1197	-6.4698	Protease
fig|33899.16.peg.2257	1671	-6.80255	Acetyl-CoA synthetase
fig|33899.16.peg.4746	732	-7.05743	GlnR-family transcriptional regulator
fig|33899.16.peg.4747	210	-8.72599	Hypothetical protein

ii) Comparison between day 2 and 3 in MT:

Transcript ID	**length**	**LogFC**	**Gene product**
fig|33899.16.peg.1833	942	2.894544	Cobalt-zinc-cadmium resistance protein CzcD
fig|33899.16.peg.5162	411	2.423708	Hypothetical Protein
fig|33899.16.peg.7186	1293	2.34684	Maltodextrin ABC transporter substrate-binding protein MdxE
fig|33899.16.peg.5163	879	1.727672	Uncharacterized metal-dependent hydrolase
fig|33899.16.peg.7417	1479	1.584328	Ricin-type carbohydrate-binding domain

fig|33899.16.peg.4615	2535	1.541028	Uncharacterized MFS-type transporter
fig|33899.16.peg.5788	162	1.254255	Hypothetical Protein
fig|33899.16.peg.6024	1017	1.219531	Dihydrofolate reductase
fig|33899.16.peg.7955	549	0.97958	Dihydrofolate reductase
fig|33899.16.peg.1331	2016	0.800832	putative secreted protein
fig|33899.16.peg.8073	747	-3.83974	Pyruvate formate-lyase activating enzyme
fig|33899.16.peg.8055	492	-3.86239	hypothetical protein
fig|33899.16.peg.1615	579	-3.87184	hypothetical protein
fig|33899.16.peg.1240	318	-3.93568	tRNA 2-thiouridine synthesis protein TusE
fig|33899.16.peg.8090	3699	-4.00232	Respiratory nitrate reductase alpha chain
fig|33899.16.peg.8076	1215	-4.14794	Flavohemoglobin / Nitric oxide dioxygenase
fig|33899.16.peg.1239	474	-4.1796	NADH dehydrogenase
fig|33899.16.peg.7669	2736	-4.41368	hypothetical protein
fig|33899.16.peg.8041	1023	-4.57976	Alcohol dehydrogenase
fig|33899.16.peg.8099	279	-4.69487	Respiratory nitrate reductase alpha chain

iii) Comparison between day 3 and 4 within MT

Transcript ID	length	LogFC	Gene product
fig|33899.16.peg.7669	2736	5.51832	Hypothetical protein
fig|33899.16.peg.8041	1023	5.514733	Alcohol dehydrogenase
fig|33899.16.peg.8055	492	5.349423	Hypothetical protein
fig|33899.16.peg.1239	474	5.246739	NADH dehydrogenase
fig|33899.16.peg.1240	318	5.162008	tRNA 2-thiouridine synthesis protein TusE
fig|33899.16.peg.8076	1215	5.070977	Flavohemoglobin / Nitric oxide dioxygenase
fig|33899.16.peg.8090	3699	4.99111	Respiratory nitrate reductase alpha chain
fig|33899.16.peg.8099	279	4.867178	Respiratory nitrate reductase alpha chain
fig|33899.16.peg.8073	747	4.748733	Pyruvate formate-lyase activating enzyme
fig|33899.16.peg.8048	2730	4.647366	Protein lysine acetyltransferase Pat
fig|33899.16.peg.5163	879	-1.46046	Uncharacterized metal-dependent hydrolase
fig|33899.16.peg.5283	630	-1.88018	Hypothetical Protein
fig|33899.16.peg.7417	1479	-2.0434	Ricin-type carbohydrate-binding domain
fig|33899.16.peg.5162	411	-2.21934	Hypothetical Protein
fig|33899.16.peg.4718	174	-2.39556	Hypothetical Protein

fig|33899.16.peg.1833	942	-2.575	Cobalt-zinc-cadmium resistance protein CzcD
fig|33899.16.peg.7186	1293	-2.60631	Maltodextrin ABC transporter substrate-binding protein MdxE
fig|33899.16.peg.3687	2310	-3.82072	beta-glucosidase
fig|33899.16.peg.2977	966	-6.4855	D-3-phosphoglycerate dehydrogenase
fig|33899.16.peg.6631	903	-6.86873	Ectoine hydroxylase

iv) Comparison between day 1 and 2 in WT

Transcript ID	length	LogFC	Gene product
fig|33899.16.peg.4280	2226	9.880863	Integral membrane protein
fig|33899.16.peg.797	534	9.422005	Tripartite tricarboxylate transporter TctB family;
fig|33899.16.peg.679	1581	8.350327	hypothetical protein
fig|33899.16.peg.4311	153	8.161912	hypothetical protein
fig|33899.16.peg.799	396	8.103228	UspA domain protein
fig|33899.16.peg.680	483	7.649496	hypothetical protein
fig|33899.16.peg.7098	456	7.277228	putative membrane protein
fig|33899.16.peg.3031	2235	6.983884	Limit dextrin alpha-1 C6-maltotetraose-hydrolase
fig|33899.16.peg.4315	1905	6.983884	Phage protein D

fig|33899.16.peg.8132	660	6.774179	Two-component transcriptional response regulator LuxR
fig|33899.16.peg.7305	648	-6.04858	Transcriptional regulator AcrR family
fig|33899.16.peg.7118	678	-6.28288	Integral membrane protein
fig|33899.16.peg.1656	123	-6.92249	hypothetical protein
fig|33899.16.peg.5554	1950	-7.0733	Methylcrotonyl-CoA carboxylase biotin-containing subunit
fig|33899.16.peg.3900	315	-7.20563	hypothetical protein
fig|33899.16.peg.1659	201	-7.33455	hypothetical protein
fig|33899.16.peg.3898	615	-7.59574	Transcriptional regulator Xre-family with cupin domain
fig|33899.16.peg.979	1020	-7.91653	Putative oxidoreductase
fig|33899.16.peg.1672	603	-8.26887	hypothetical protein
fig|33899.16.peg.7304	1014	-10.3318	putative membrane protein

v) Comparison between day 2 and 3 in WT

Transcript ID	**length**	**LogFC**	**Gene product**
fig|33899.16.peg.7376	1569	10.59694	ABC transporter substrate-binding protein
fig|33899.16.peg.4080	165	9.578356	ribosomal protein
fig|33899.16.peg.7370	1257	9.289268	hypothetical protein
fig|33899.16.peg.727	171	9.223787	ribosomal protein

fig|33899.16.peg.7372	1089	8.580242	Pyridoxal-5'-phosphate-dependent enzyme beta superfamily
fig|33899.16.peg.728	1173	8.51201	Metal chaperone involved in Zn homeostasis
fig|33899.16.peg.7371	831	8.414053	hypothetical protein
fig|33899.16.peg.5861	981	7.937384	Zinc ABC transporter substrate-binding protein ZnuA
fig|33899.16.peg.7379	1566	7.755993	ABC transporter ATP-binding protein
fig|33899.16.peg.7373	1032	7.609867	L-alanine-DL-glutamate epimerase
fig|33899.16.peg.721	882	-3.72311	Putative oxidoreductase
fig|33899.16.peg.710	999	-3.78439	Phytoene synthase
fig|33899.16.peg.5163	879	-4.04775	Uncharacterized metal-dependent hydrolase YcfH
fig|33899.16.peg.4714	741	-4.40582	Carbonic anhydrase beta class (
fig|33899.16.peg.5162	411	-4.52834	hypothetical protein
fig|33899.16.peg.1863	243	-4.64917	Transcriptional regulator WhiB family;
fig|33899.16.peg.799	396	-4.93157	UspA domain protein
fig|33899.16.peg.796	969	-5.21315	Tripartite tricarboxylate transporter TctB family
fig|33899.16.peg.797	534	-5.30089	Tripartite tricarboxylate transporter TctB family

fig|33899.16.peg.6931	1212	-6.25786	hypothetical protein

vi) Comparison between day 3 and 4 in WT

Transcript ID	**start**	**end**	**length**		**Gene product**
fig|33899.16.peg.1923	1791996	1793406	1410		putative secreted glucosidase
fig|33899.16.peg.205	177180	177399	219		hypothetical protein
fig|33899.16.peg.3573	3557437	3558379	942		hypothetical protein
fig|33899.16.peg.359	304144	305758	1614		hypothetical protein
fig|33899.16.peg.540	453379	453841	462		hypothetical protein
fig|33899.16.peg.6703	6948185	6948884	699		Putative oxidoreductase
fig|33899.16.peg.958	850978	851101	123		hypothetical protein
fig|33899.16.peg.7596	7880418	7880538	120		hypothetical protein
fig|33899.16.peg.2282	2173006	2174329	1323		hypothetical protein
fig|33899.16.peg.2284	2174985	2175882	897		UDP-glucose 4-epimerase
fig|33899.16.peg.1519	1409166	1409508	342		Transcriptional regulator ArsR family
fig|33899.16.peg.1518	1407439	1409170	1731		Sulfate permease
fig|33899.16.peg.798	689590	691090	1500		Universal stress protein family
fig|33899.16.peg.8049	8382171	8382678	507		Flavodoxin
fig|33899.16.peg.1481	1373445	1373646	201		hypothetical protein

fig|33899.16.peg.8089	8423869	8425426	1557		Respiratory nitrate reductase beta chain
fig|33899.16.peg.6550	6791217	6792891	1674		Alpha-glucosidase AglA
fig|33899.16.peg.6553	6795621	6796461	840		Alpha-glucoside transport system permease protein AglG
fig|33899.16.peg.6551	6792928	6794257	1329		Alpha-glucosides-binding periplasmic protein AglE precursor
fig|33899.16.peg.6552	6794263	6795625	1362		ABC alpha-glucoside transporter inner membrane subunit AglF

Appendix 2

Aclacinomycin producing BGC detected by antiSMASH and the corresponding gene ID.

Gene ID	Locus tag	Location				Function
		From	To	Length	strand	
fig|33899.16.peg.2243	ctg1_2065	2133093	2133885	792	+	Hypothetical protein/ DNA binding protein*
fig|33899.16.peg.2244	ctg1_2066	2134121	2135501	1380	-	Guanine deaminase
fig|33899.16.peg.2245	ctg1_2067	2135668	2136595	927	-	Urate oxidase
fig|33899.16.peg.2246	ctg1_2068	2136599	2137004	405	-	5-hydroxyisourate hydrolase
fig|33899.16.peg.2247	ctg1_2069	2137091	2137610	519	-	2-oxo-4-hydroxy-4-carboxy-5-ureidoimidazoline (OHCU) decarboxylase (Urate degradation)
fig|33899.16.peg.2248	ctg1_2070	2137766	2138135	369	-	DNA binding protein
fig|33899.16.peg.2249	ctg1_2071	2138131	2138380	249	-	Hypothetical protein
fig|33899.16.peg.2250	ctg1_2072	2138675	2139497	822	+	Hydroxypyruvate isomerase
fig|33899.16.peg.2251	ctg1_2073	2139560	2140451	891	+	2-hydroxy-3-oxopropionate reductase
fig|33899.16.peg.2252	ctg1_2074	2141016	2141262	246	+	Hypothetical protein
fig|33899.16.peg.2253	ctg1_2075	2141418	2142069	651	+	TIGR04222 domain-containing membrane protein*
fig|33899.16.peg.2254	ctg1_2076	2142218	2144099	1881	-	Glyoxylate carboligase
fig|33899.16.peg.2255	ctg1_2077	2144211	2145081	870	+	Hypothetical protein
fig|33899.16.peg.2256	ctg1_2078	2145117	2146725	1608	-	Long-chain-fatty-acid--CoA ligase
fig|33899.16.peg.2257	ctg1_2079	2146721	2148392	1671	-	Acetyle-CoA-synthetase
fig|33899.16.peg.2258	ctg1_2080	2148515	2149358	843	+	Transcriptional regulator
fig|33899.16.peg.2259	ctg1_2081	2149974	2152806	2832	+	Transcriptional regulator

fig|33899.16.peg.2260	ctg1_2082	2153155	2153947	792	+	Trypsin-like protease/serine protease*
fig|33899.16.peg.2261	ctg1_2083	2154084	2155203	1119	-	Hypothetical protein/ DNA binding domain*
fig|33899.16.peg.2262	ctg1_2084	2155344	2156199	855	-	ABC-2 type transporter
fig|33899.16.peg.2263	ctg1_2085	2156195	2157176	981	-	ABC transporter ATP-binding protein
fig|33899.16.peg.2264	ctg1_2086	2157366	2157975	609	-	Transcriptional regulator PadR family
fig|33899.16.peg.2265	ctg1_2087	2158140	2158578	438	-	AclR protein
fig|33899.16.peg.2266	ctg1_2088	2158795	2159641	846	+	Hypothetical protein/ NADPH-binding protein*
fig|33899.16.peg.2267	ctg1_2089	2159655	2160387	732	+	Hypothetical protein/ Methyltransferase*
fig|33899.16.peg.2268	ctg1_2090	2160448	2162080	1632	-	Putative oxidoreductase
fig|33899.16.peg.2269	ctg1_2091	2162153	2162624	471	-	Hypothetical protein/ Oxidoreductase*
fig|33899.16.peg.2270	ctg1_2092	2162708	2163563	855	-	Hypothetical protein/ Transcriptional regulator*
fig|33899.16.peg.2271	ctg1_2093	2163594	2164029	435	-	Nogalonic acid methyl ester cyclase/ Polyketide metabolic process*
fig|33899.16.peg.2272		2164022	2164147	125	-	Hypothetical protein*+
fig|33899.16.peg.2273	ctg1_2094	2164231	2165092	861	+	Methyltransferase*
fig|33899.16.peg.2274	ctg1_2095	2165145	2166189	1044	-	Modular polyketie synthase/ Acyltransferase*
fig|33899.16.peg.2275	ctg1_2096	2166185	2167292	1107	-	Acyl carrier protein synthase
fig|33899.16.peg.2276	ctg1_2097	2167288	2167564	276	-	Acyl carrier protein
fig|33899.16.peg.2277	ctg1_2098	2167667	2168891	1224	-	Polyketide chain length factor
fig|33899.16.peg.2278	ctg1_2099	2168887	2170159	1272	-	Beta-ketoacyl synthase

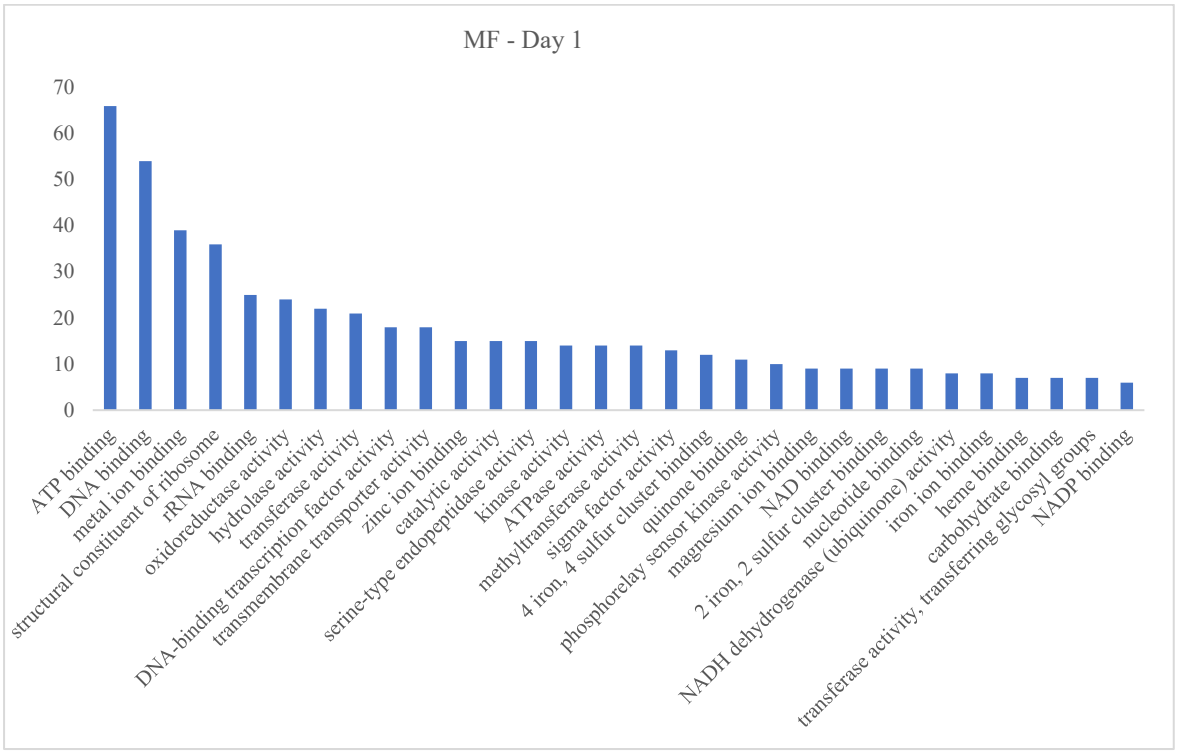
fig|33899.16.peg.2279	ctg1_2100	2170155	2170524	369	-	Monooxygenase
fig|33899.16.peg.2280	ctg1_2101	2170735	2171521	786	+	Acetoacyl-CoA reductase.
fig|33899.16.peg.2281	ctg1_2102	2171568	2172921	1353	+	Hypothetical protein/ Cyclase*
fig|33899.16.peg.2282	ctg1_2103	2173006	2174329	1323	+	Hypothetical protein/ Glycosyltransferase*
fig|33899.16.peg.2283	ctg1_2104	2174329	2174947	618	+	Epimerase
fig|33899.16.peg.2284	ctg1_2105	2174934	2175882	948	+	NAD-dependent epimerase
fig|33899.16.peg.2285	ctg1_2106	2175948	2177301	1353	+	Hypothetical protein/ Dehydratase (AclN)*
fig|33899.16.peg.2286	ctg1_2107	2177387	2178497	1110	-	Aminotransferase
fig|33899.16.peg.2287	ctg1_2108	2178719	2179595	876	+	Putative glucose synthase (AclY))
fig|33899.16.peg.2288	ctg1_2109	2179607	2180324	717	-	Methyletransferase
fig|33899.16.peg.2289	ctg1_2110	2180385	2181165	780	-	Putative cyclase
fig|33899.16.peg.2290	ctg1_2111	2181375	2181810	435	+	Cyclase/ Nuclear transport factor (AknV)*
fig|33899.16.peg.2291	ctg1_2112	2181868	2182639	771	+	Putative aklaviketone reductase
fig|33899.16.peg.2292	ctg1_2113	2182662	2183994	1332	+	Hypothetical protein/ P450-derived glycosyltransferase activator/AknT*
fig|33899.16.peg.2293	ctg1_2114	2184070	2185402	1332	+	Glycosyltransferase (AknS)
fig|33899.16.peg.2294	ctg1_2115	2185389	2186361	972	+	NAD-dependent epimerase/dehydratase
fig|33899.16.peg.2295	ctg1_2116	2186469	2187459	990	-	Oxidoreductase
fig|33899.16.peg.2296		2187468	2187587	119	-	Hypothetical protein
fig|33899.16.peg.2297	ctg1_2117	2187620	2188925	1305	+	Putative 3-dehydratase
fig|33899.16.peg.2298	ctg1_2118	2188938	2189757	819	-	Transcriptional regulator
fig|33899.16.peg.2299	ctg1_2119	2190024	2190204	180	+	Hypothetical protein

fig|33899.16.peg.2300	ctg1_2120	2190200	2192468	2268	+	AknN
fig|33899.16.peg.2301		2192465	2192641	176		Hypothetical Protein
fig|33899.16.peg.2302	ctg1_2121	2192703	2193963	1260	+	Hypothetical protein/ Histidine kinase*
fig|33899.16.peg.2303	ctg1_2122	2194035	2194641	606	-	Response regulator transcription factor
fig|33899.16.peg.2304	ctg1_2123	2194850	2195924	1074	-	NAD kinase
fig|33899.16.peg.2305	ctg1_2124	2196035	2197121	1086	-	Hypothetical protein/ N-acyltransferase*
fig|33899.16.peg.2306	ctg1_2125	2197341	2197908	567	+	Hypothetical protein/ ABATE domain-containing protein*
fig|33899.16.peg.2307	ctg1_2126	2197920	2198100	180	-	Hypothetical protein
fig|33899.16.peg.2308	ctg1_2127	2198267	2198837	570	-	Transcriptional regulator
fig|33899.16.peg.2309		2198997	2199203	206	+	Hypothetical Protein
fig|33899.16.peg.2310	ctg1_2128	2199252	2200704	1452	+	Glycosylhydrolase
fig|33899.16.peg.2311	ctg1_2129	2200836	2201637	801	+	Polysaccharide deacetylase
fig|33899.16.peg.2312	ctg1_2130	2201770	2202211	441	+	Hypothetical protein/ Ligand binding domain*
fig|33899.16.peg.2313	ctg1_2131	2202355	2203702	1347	+	Hypothetical protein/ Lanthionine synthetase C family protein*
fig|33899.16.peg.2314	ctg1_2132	2203653	2204805	1152	-	XdhC family protein

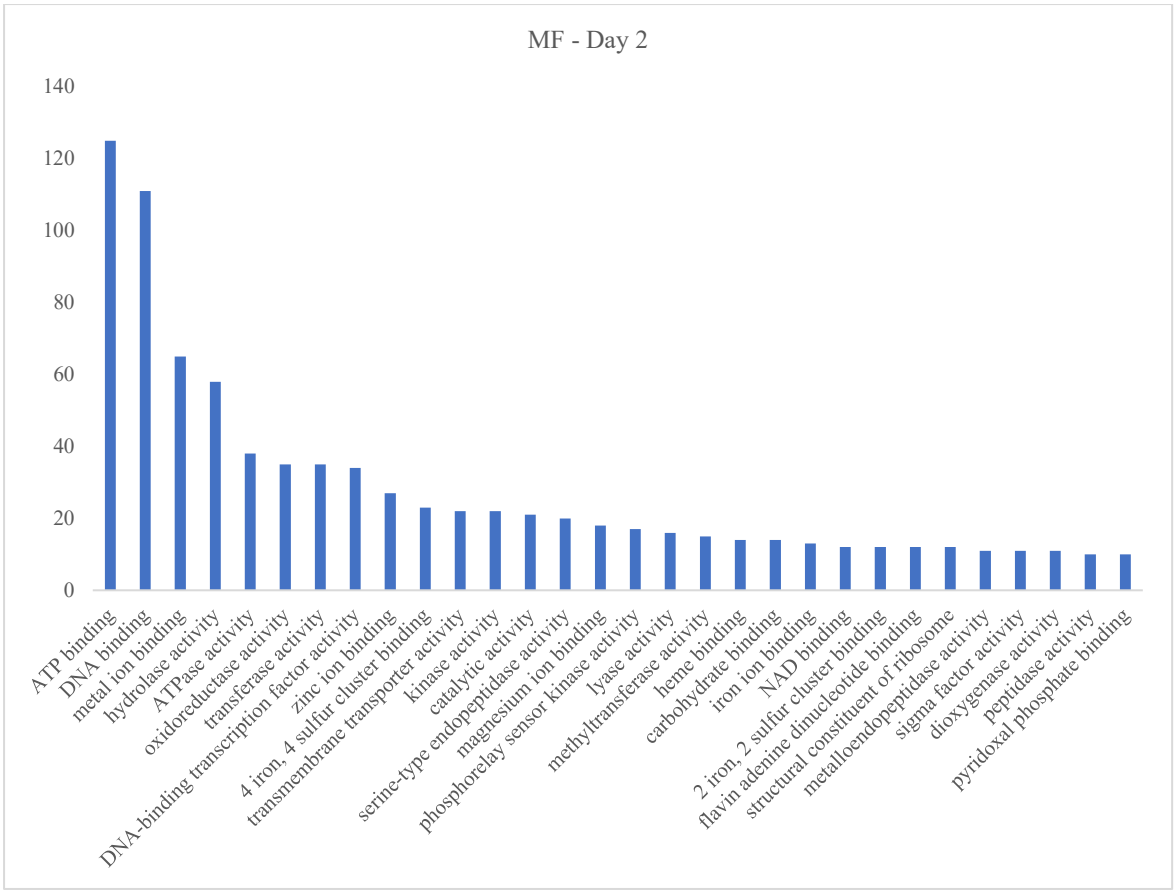
Appendix 3:

Top 30 annotated GO terms of the DEGs between the strains. Y-axis represents the number of annotated sequences.

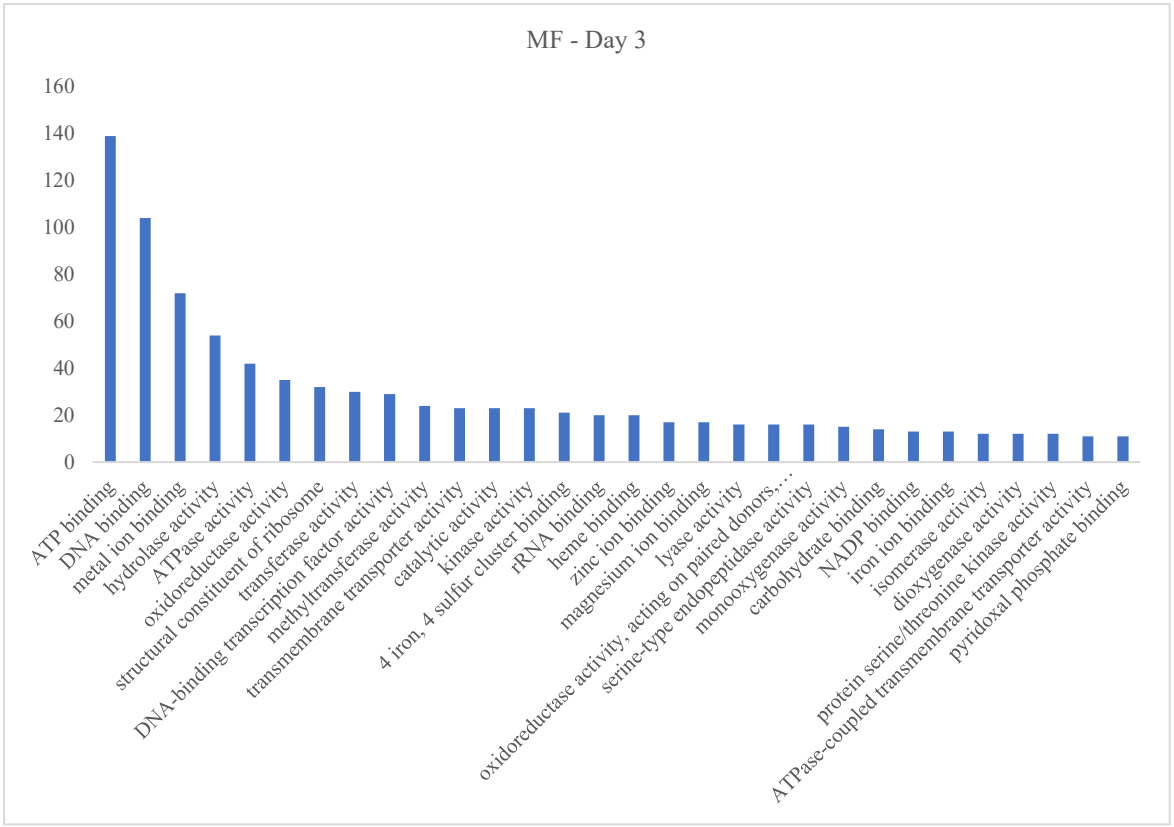
a)



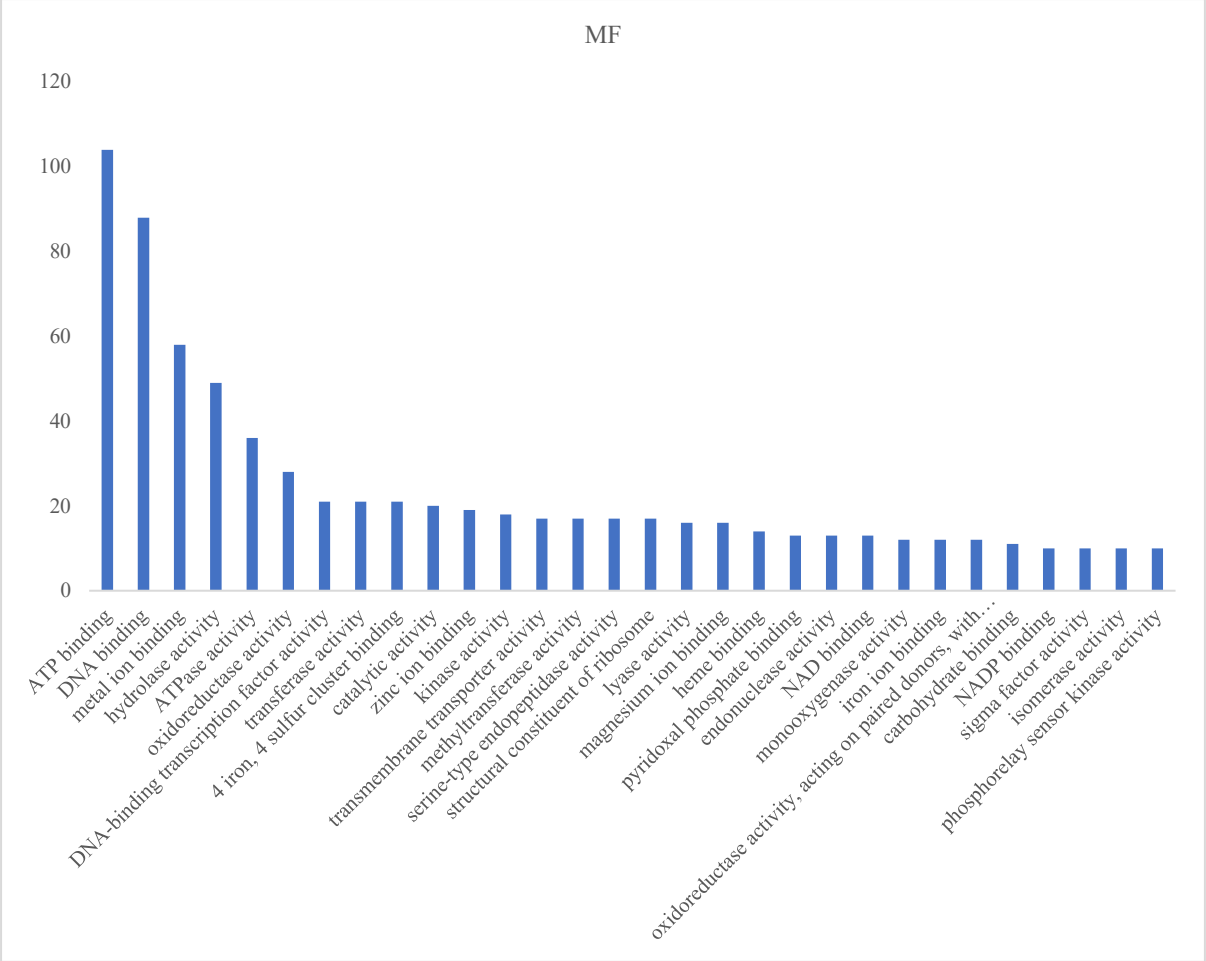
b)


MF - Day 2
140
120
100
80
60
40
20
0
ATP binding
DNA binding
metal ion binding
hydrolase activity
ATPase activity
oxidoreductase activity
transferase activity
DNA-binding transcription factor activity
zinc ion binding
4 iron, 4 sulfur cluster binding
transmembrane transporter activity
kinase activity
catalytic activity
serine-type endopeptidase activity
magnesium ion binding
phosphorelay sensor kinase activity
lyase activity
methyltransferase activity
heme binding
carbohydrate binding
iron ion binding
NAD binding
2 iron, 2 sulfur cluster binding
flavin adenine dinucleotide binding
structural constituent of ribosome
metalloendopeptidase activity
sigma factor activity
dioxygenase activity
peptidase activity
pyridoxal phosphate binding

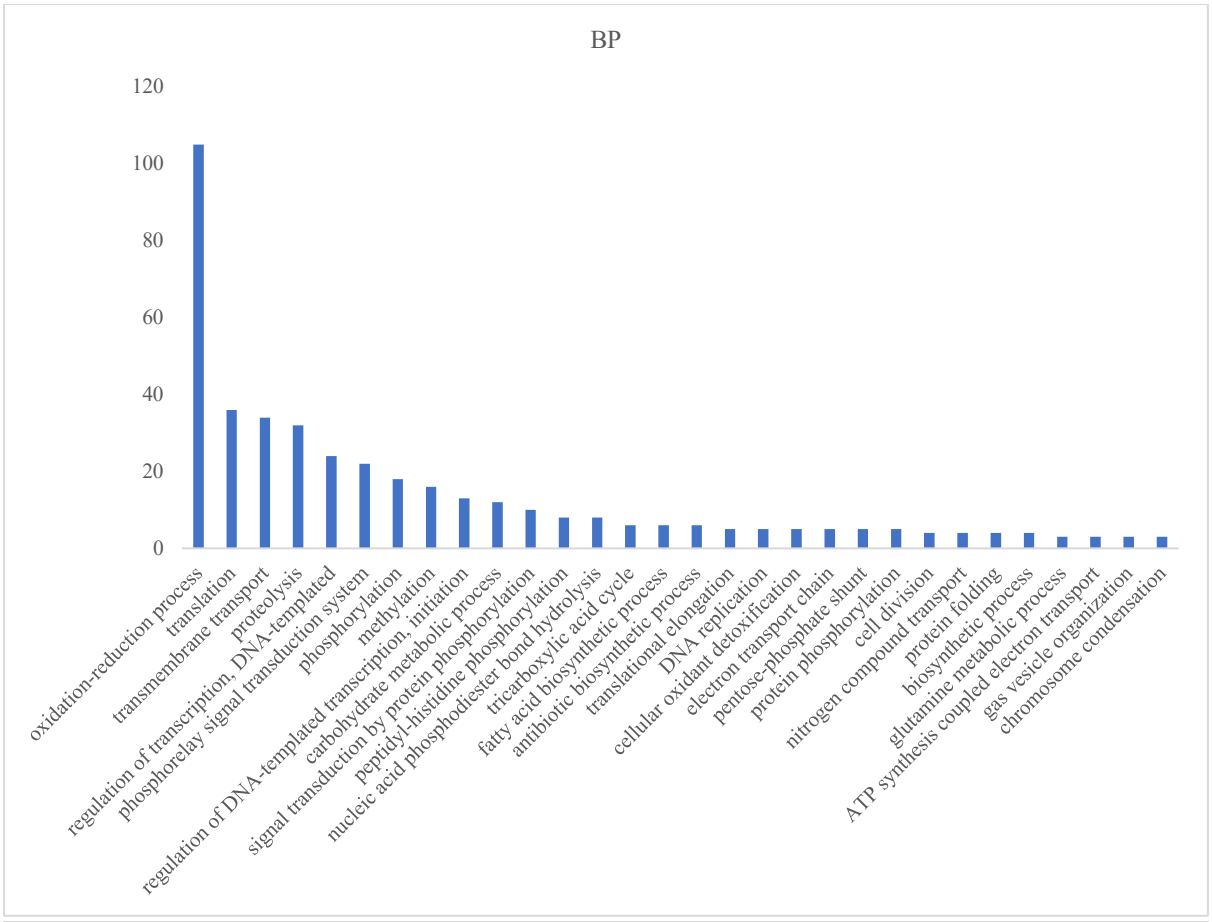
c)


MF - Day 3
160
140
120
100
80
60
40
20
0
ATP binding
DNA binding
metal ion binding
hydrolase activity
ATPase activity
oxidoreductase activity
structural constituent of ribosome
transferase activity
DNA-binding transcription factor activity
methyltransferase activity
transmembrane transporter activity
catalytic activity
kinase activity
4 iron, 4 sulfur cluster binding
rRNA binding
heme binding
zinc ion binding
magnesium ion binding
lyase activity
oxidoreductase activity, acting on paired donors,...
serine-type endopeptidase activity
monooxygenase activity
carbohydrate binding
NADP binding
iron ion binding
isomerase activity
dioxygenase activity
protein serine/threonine kinase activity
ATPase-coupled transmembrane transporter activity
pyridoxal phosphate binding

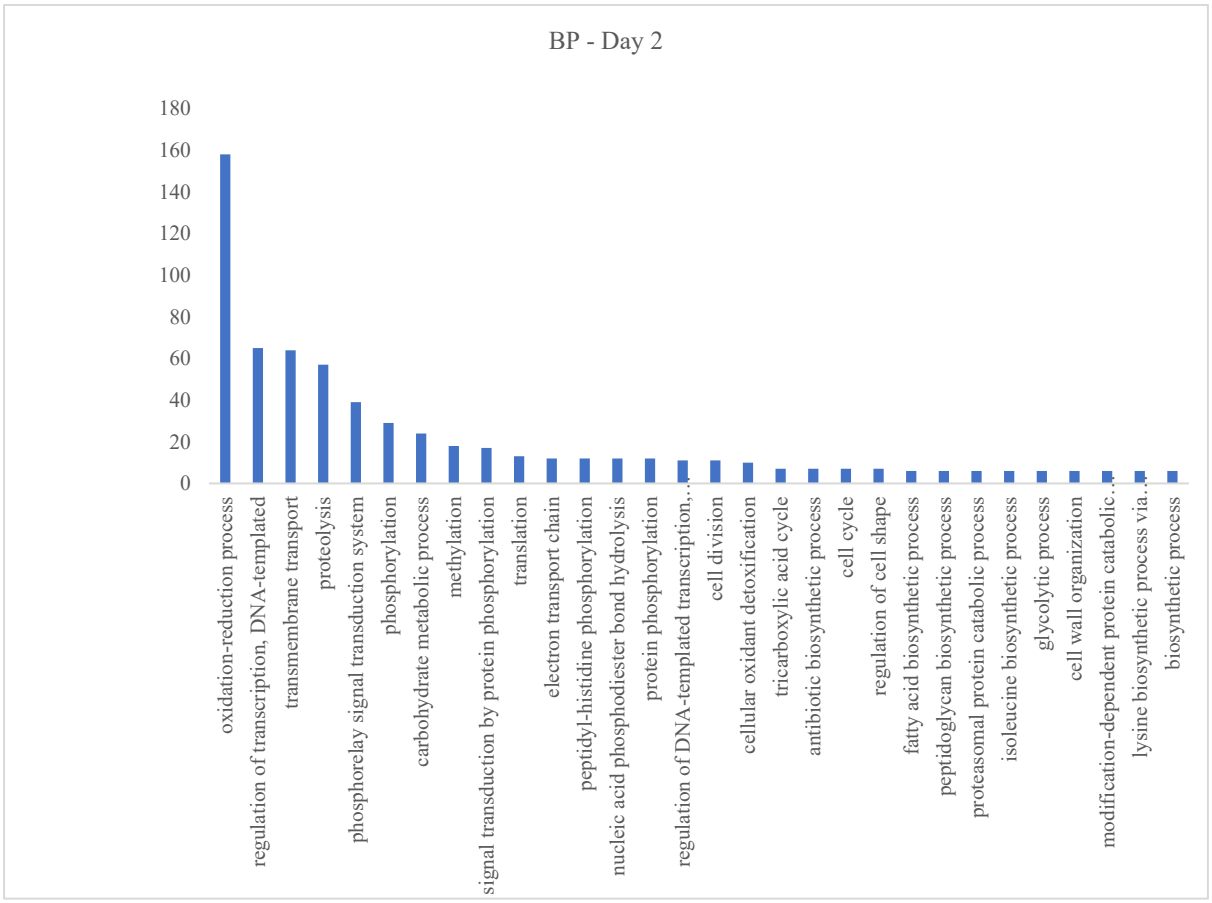
d)


MF
120
100
80
60
40
20
0
ATP binding
DNA binding
metal ion binding
hydrolase activity
ATPase activity
oxidoreductase activity
DNA-binding transcription factor activity
transferase activity
4 iron, 4 sulfur cluster binding
catalytic activity
zinc ion binding
kinase activity
transmembrane transporter activity
methyltransferase activity
serine-type endopeptidase activity
structural constituent of ribosome
lyase activity
magnesium ion binding
heme binding
pyridoxal phosphate binding
endonuclease activity
NAD binding
monooxygenase activity
iron ion binding
oxidoreductase activity, acting on paired donors, with…
carbohydrate binding
NADP binding
sigma factor activity
isomerase activity
phosphorelay sensor kinase activity

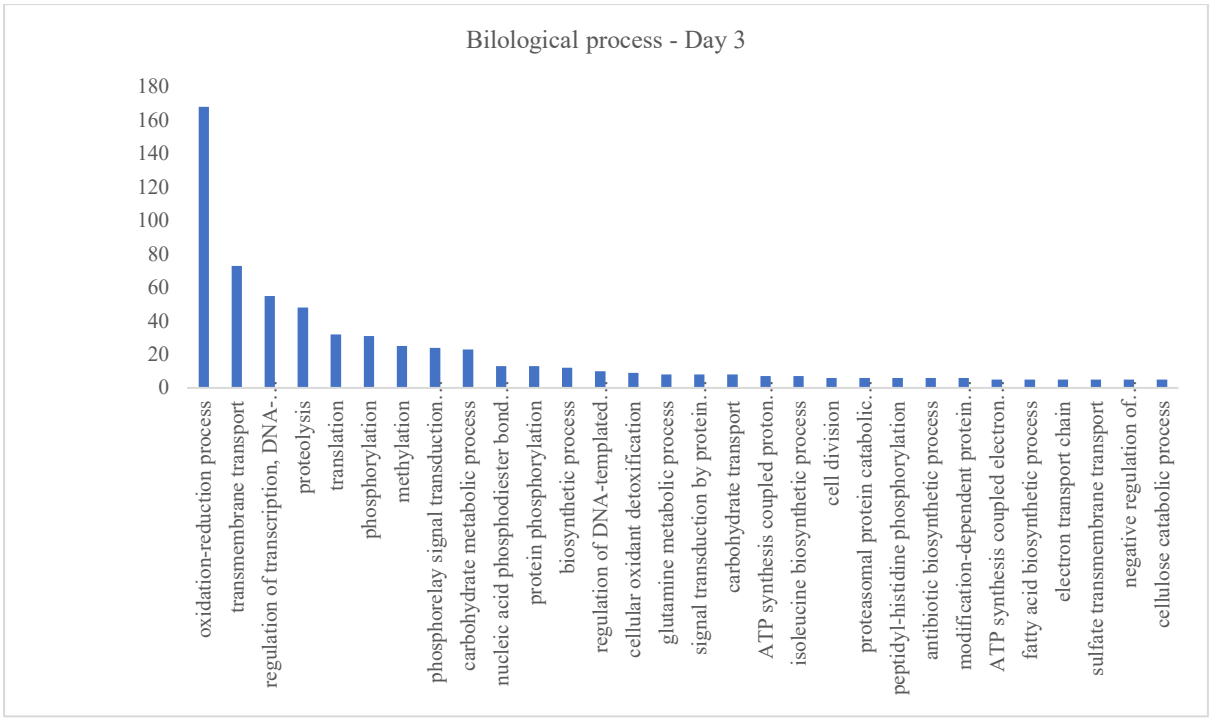
e)


BP
120
100
80
60
40
20
0
oxidation-reduction process
translation
transmembrane transport
proteolysis
regulation of transcription, DNA-templated
phosphorelay signal transduction system
phosphorylation
methylation
regulation of DNA-templated transcription, initiation
carbohydrate metabolic process
signal transduction by protein phosphorylation
peptidyl-histidine phosphorylation
nucleic acid phosphodiester bond hydrolysis
tricarboxylic acid cycle
fatty acid biosynthetic process
antibiotic biosynthetic process
translational elongation
DNA replication
cellular oxidant detoxification
electron transport chain
pentose-phosphate shunt
protein phosphorylation
cell division
nitrogen compound transport
protein folding
biosynthetic process
glutamine metabolic process
ATP synthesis coupled electron transport
gas vesicle organization
chromosome condensation

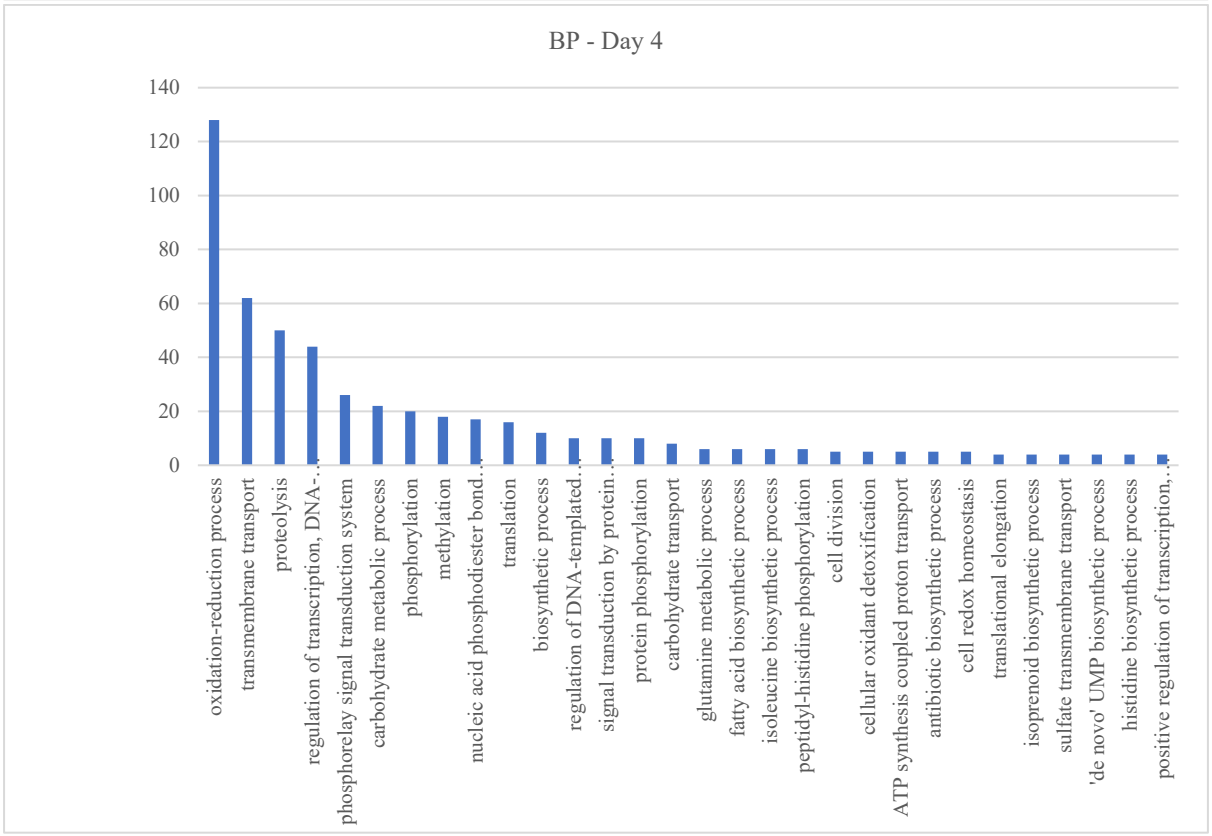
f)


BP - Day 2
180
160
140
120
100
80
60
40
20
0
oxidation-reduction process
regulation of transcription, DNA-templated
transmembrane transport
proteolysis
phosphorelay signal transduction system
phosphorylation
carbohydrate metabolic process
methylation
signal transduction by protein phosphorylation
translation
electron transport chain
peptidyl-histidine phosphorylation
nucleic acid phosphodiester bond hydrolysis
protein phosphorylation
regulation of DNA-templated transcription,...
cell division
cellular oxidant detoxification
tricarboxylic acid cycle
antibiotic biosynthetic process
cell cycle
regulation of cell shape
fatty acid biosynthetic process
peptidoglycan biosynthetic process
proteasomal protein catabolic process
isoleucine biosynthetic process
glycolytic process
cell wall organization
modification-dependent protein catabolic...
lysine biosynthetic process via...
biosynthetic process

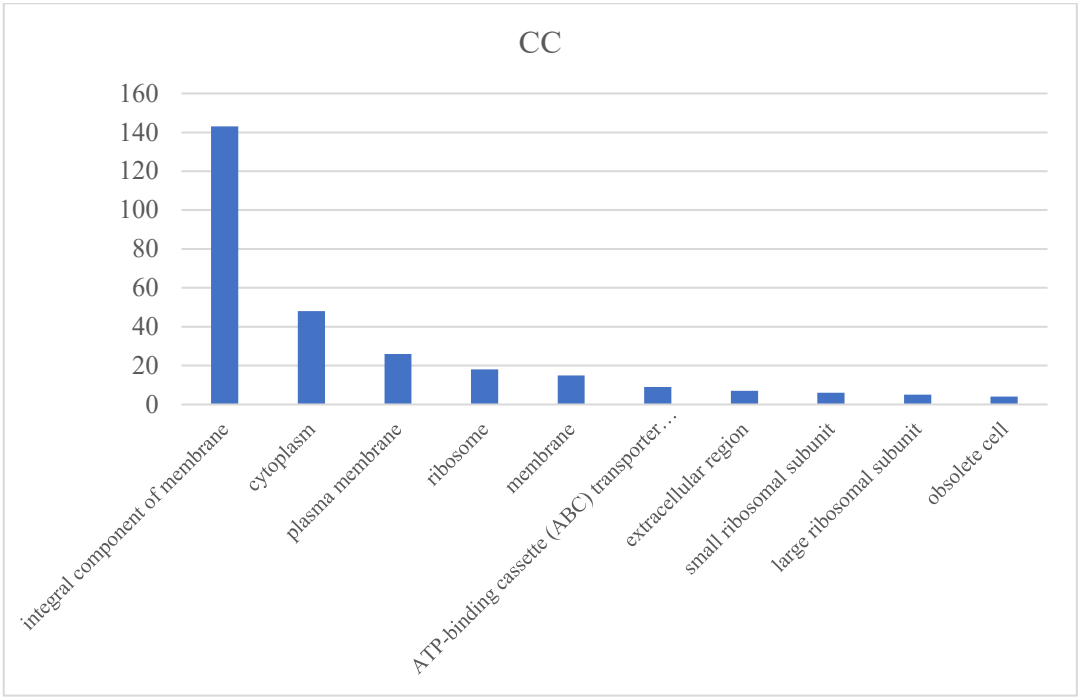
g)


Bilological process - Day 3
180
160
140
120
100
80
60
40
20
0
oxidation-reduction process
transmembrane transport
regulation of transcription, DNA-...
proteolysis
translation
phosphorylation
methylation
phosphorelay signal transduction...
carbohydrate metabolic process
nucleic acid phosphodiester bond...
protein phosphorylation
biosynthetic process
regulation of DNA-templated...
cellular oxidant detoxification
glutamine metabolic process
signal transduction by protein...
carbohydrate transport
ATP synthesis coupled proton...
isoleucine biosynthetic process
cell division
proteasomal protein catabolic...
peptidyl-histidine phosphorylation
antibiotic biosynthetic process
modification-dependent protein...
ATP synthesis coupled electron...
fatty acid biosynthetic process
electron transport chain
sulfate transmembrane transport
negative regulation of...
cellulose catabolic process

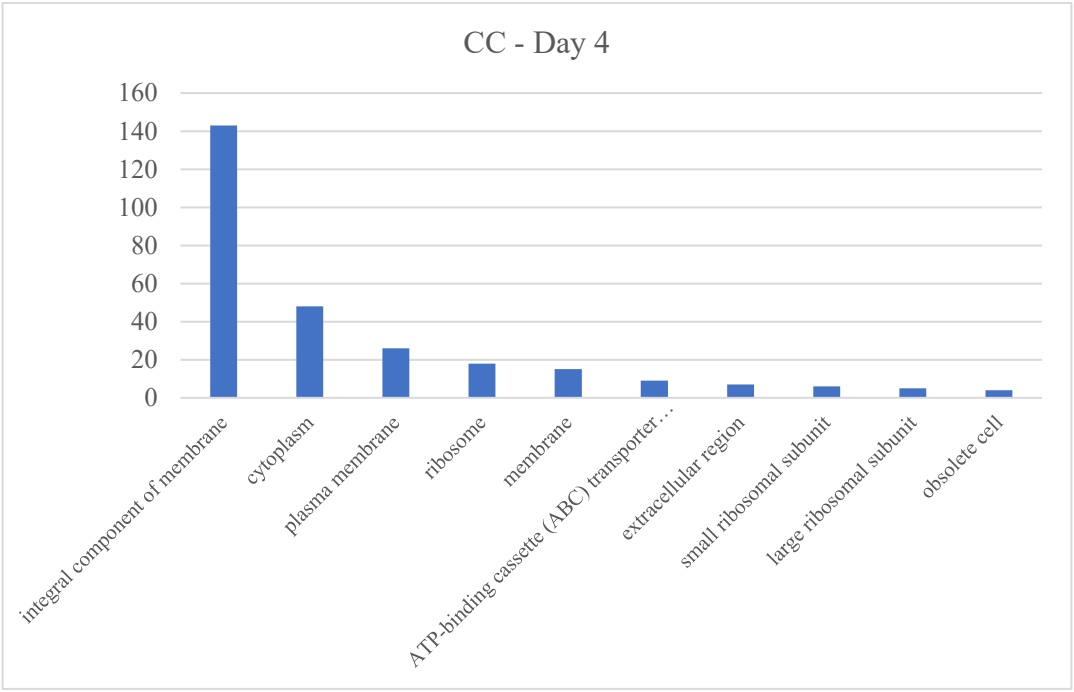
h)


BP - Day 4
140
120
100
80
60
40
20
0
oxidation-reduction process
transmembrane transport
proteolysis
regulation of transcription, DNA-...
phosphorelay signal transduction system
carbohydrate metabolic process
phosphorylation
methylation
nucleic acid phosphodiester bond...
translation
biosynthetic process
regulation of DNA-templated...
signal transduction by protein...
protein phosphorylation
carbohydrate transport
glutamine metabolic process
fatty acid biosynthetic process
isoleucine biosynthetic process
peptidyl-histidine phosphorylation
cell division
cellular oxidant detoxification
ATP synthesis coupled proton transport
antibiotic biosynthetic process
cell redox homeostasis
translational elongation
isoprenoid biosynthetic process
sulfate transmembrane transport
'de novo' UMP biosynthetic process
histidine biosynthetic process
positive regulation of transcription,...

k)


CC
160
140
120
100
80
60
40
20
0
integral component of membrane
cytoplasm
plasma membrane
ribosome
membrane
ATP-binding cassette (ABC) transporter…
extracellular region
small ribosomal subunit
large ribosomal subunit
obsolete cell

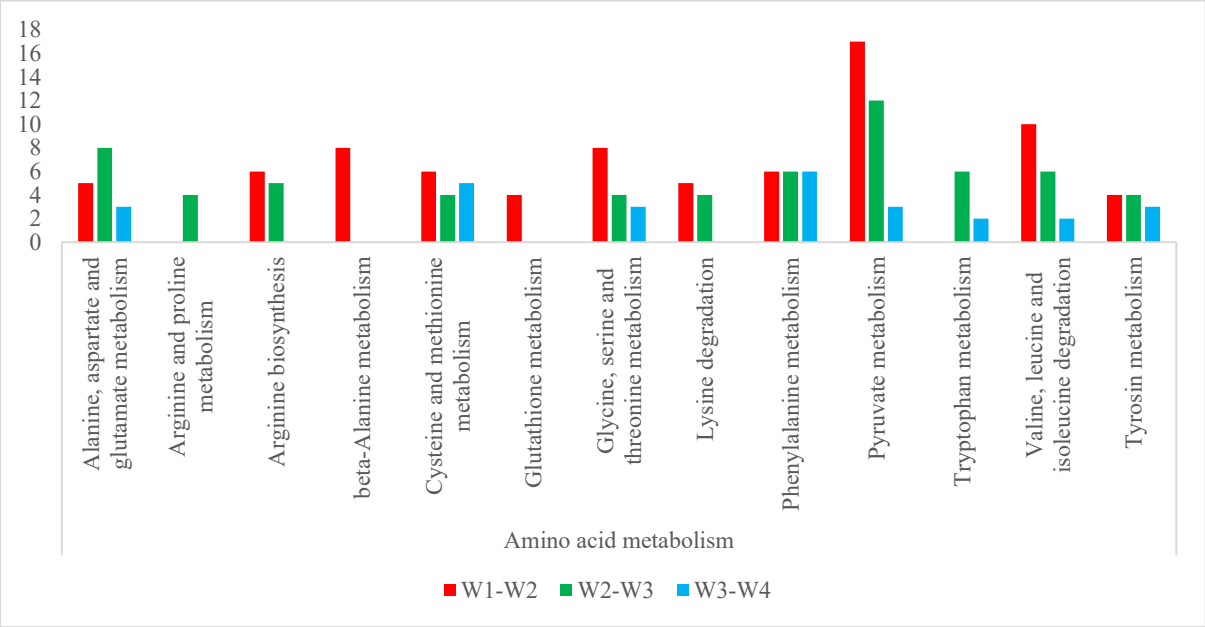
l)


CC - Day 4
160
140
120
100
80
60
40
20
0
integral component of membrane
cytoplasm
plasma membrane
ribosome
membrane
ATP-binding cassette (ABC) transporter…
extracellular region
small ribosomal subunit
large ribosomal subunit
obsolete cell

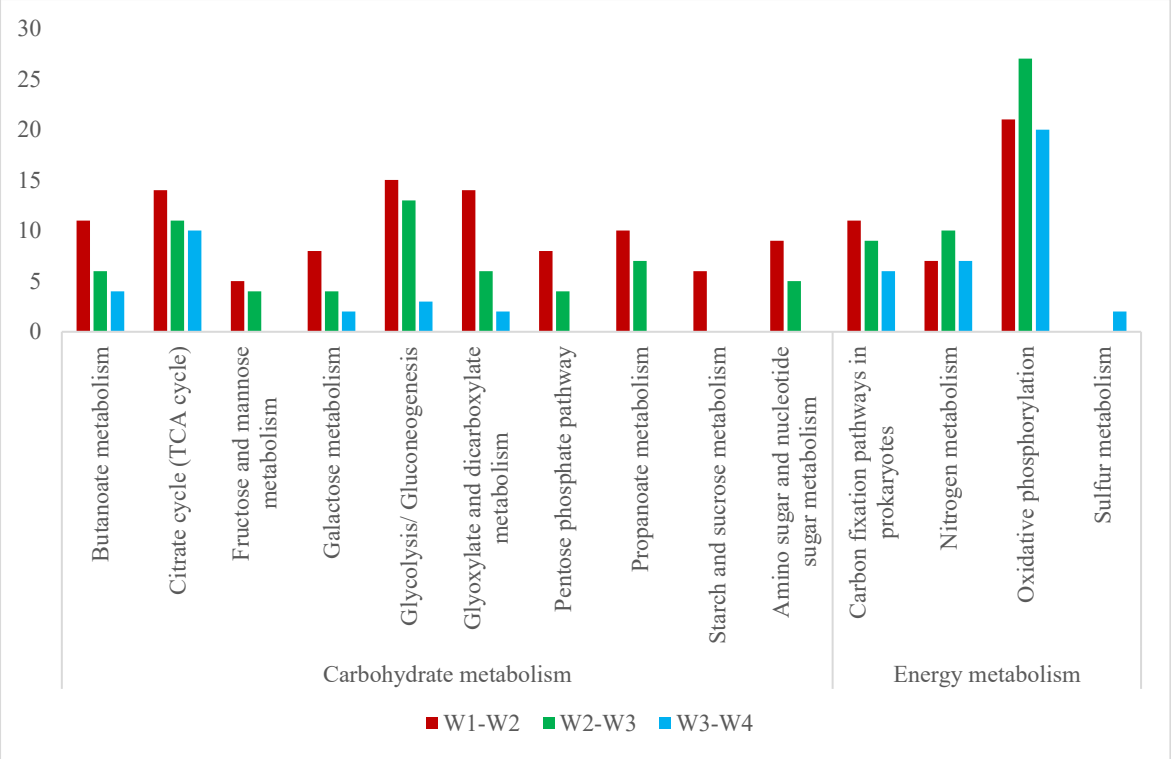
Appendix 4:

Functional classifications of the mapped DEGs within the wild (a-d) and mutant (e-h) strain on different day intervals. Y-axis denotes the number of mapped DEGs into the corresponding pathway. W1. W2 . W3 and W4 refers to the corresponding sampling day of the WT strain. and M1. M2.M3 and M4 refers to the sampling day from the mutant strain.

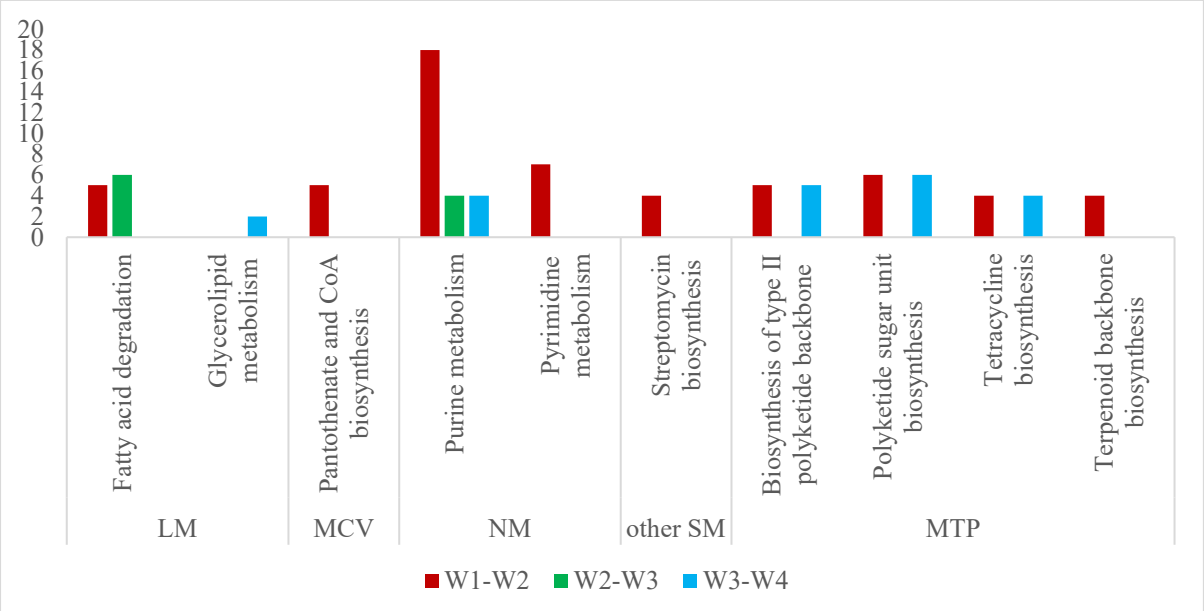
a) Amino acid metabolism:



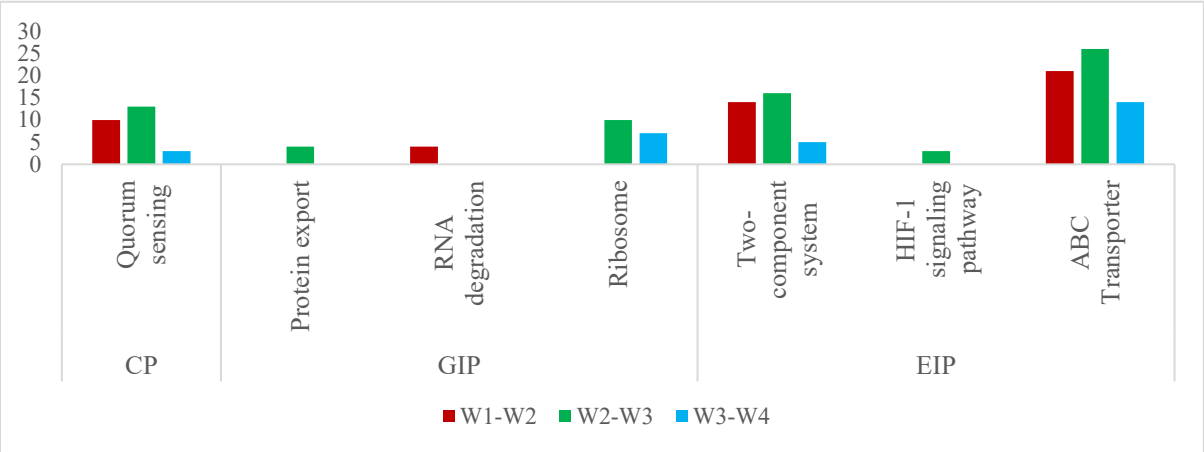
b) Carbohydrate and energy metabolism:



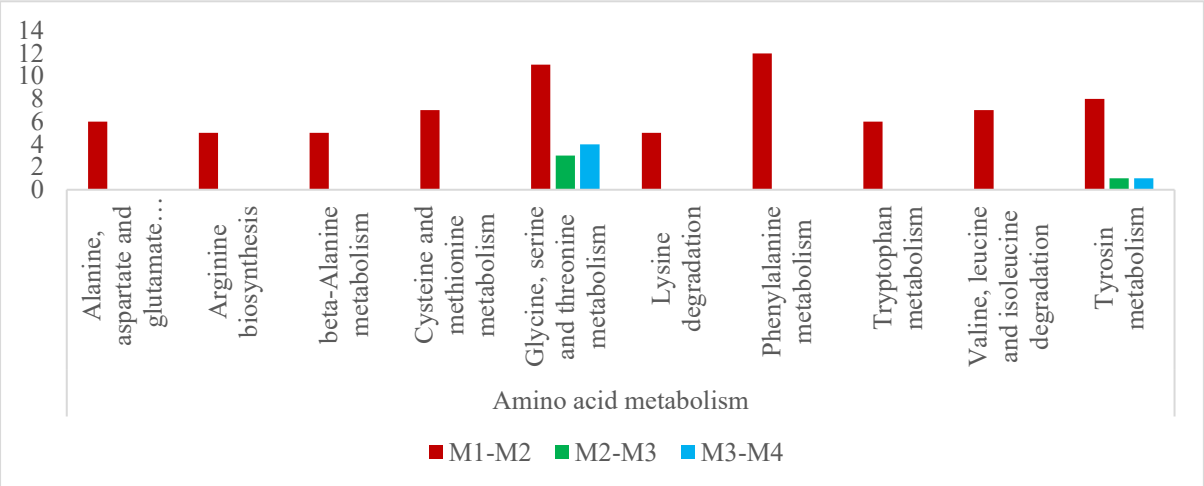
c) Lipid. cofactors. and vitamins. terpenoids and polyketides. nucleotides and other secondary metabolites:



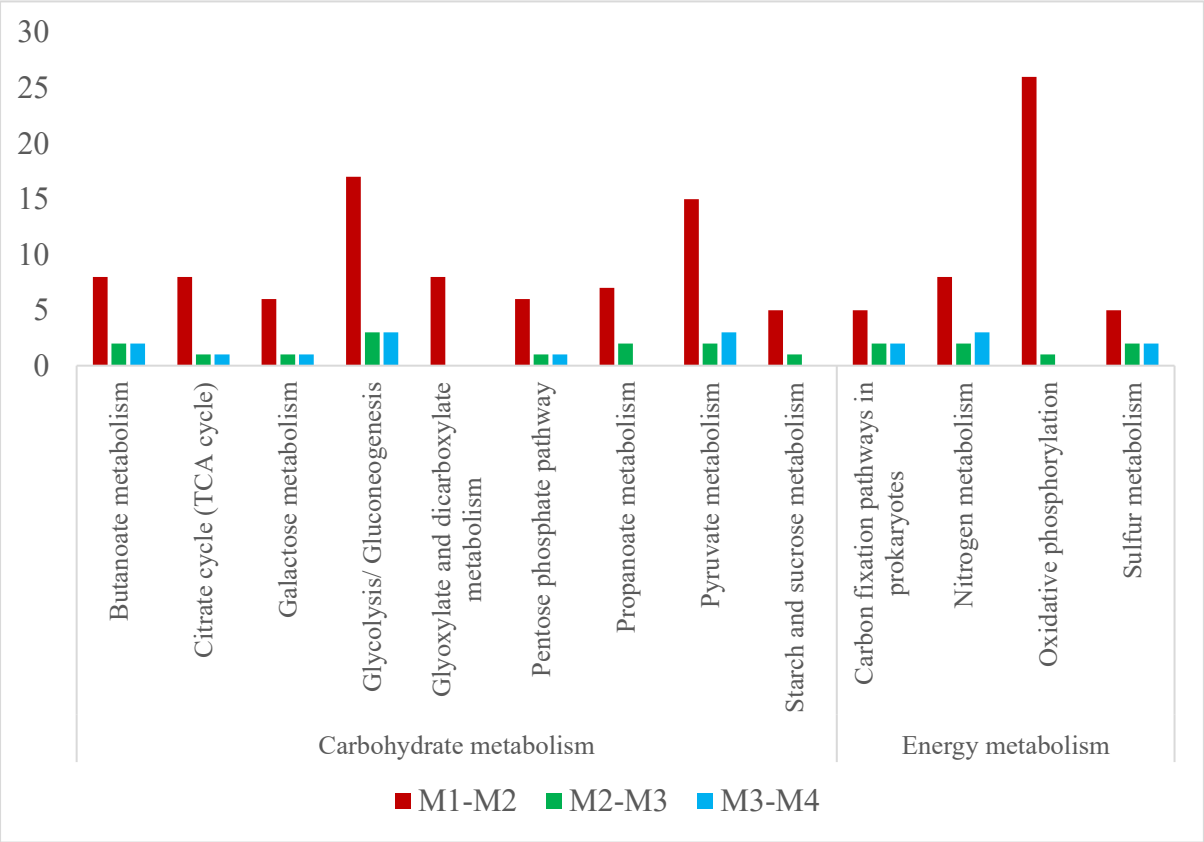
d) Pathways involved to cellular processing (CP). genetic information processing (GIP) and environmental information processing (EIP):



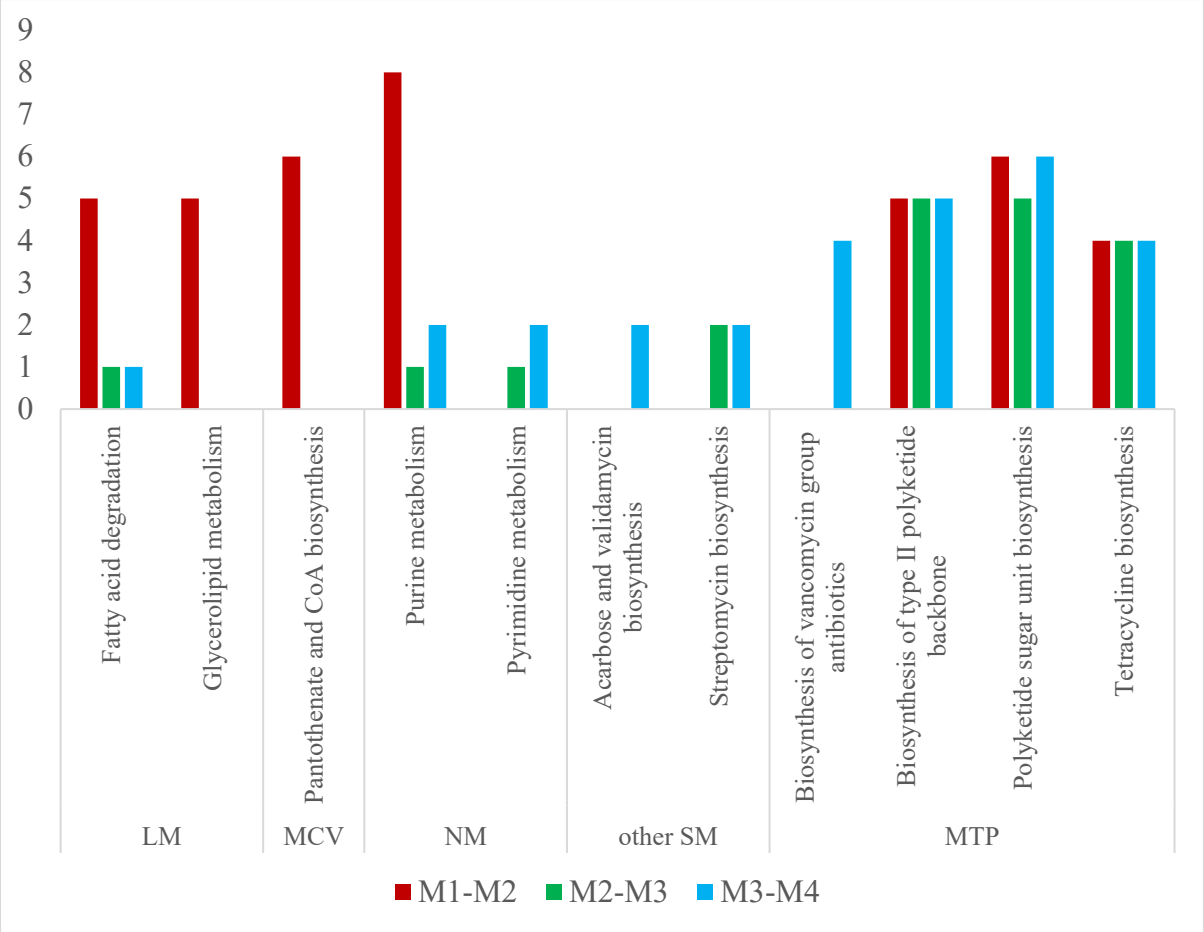
e) Amino acid metabolism:



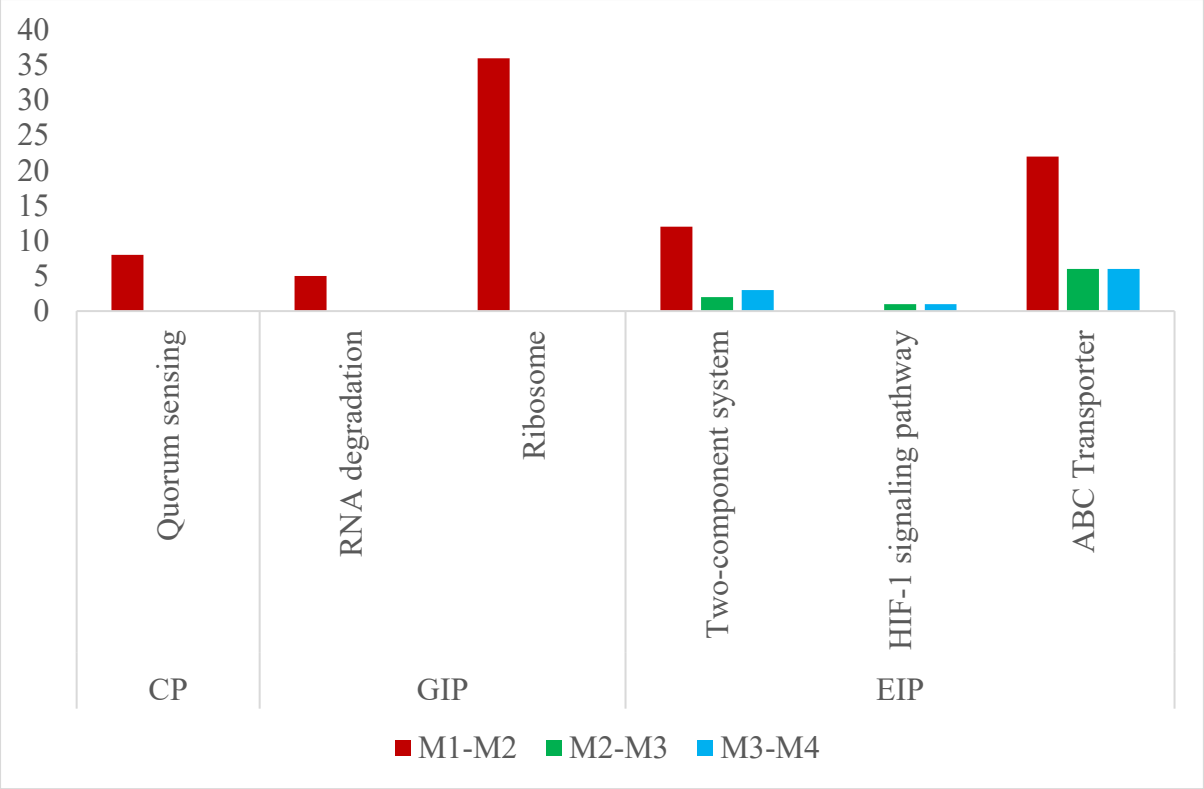
f) Carbohydrate and energy metabolism:



g) Lipid. cofactors and vitamins. terpenoids and polyketides. nucleotides and other secondary metabolites:

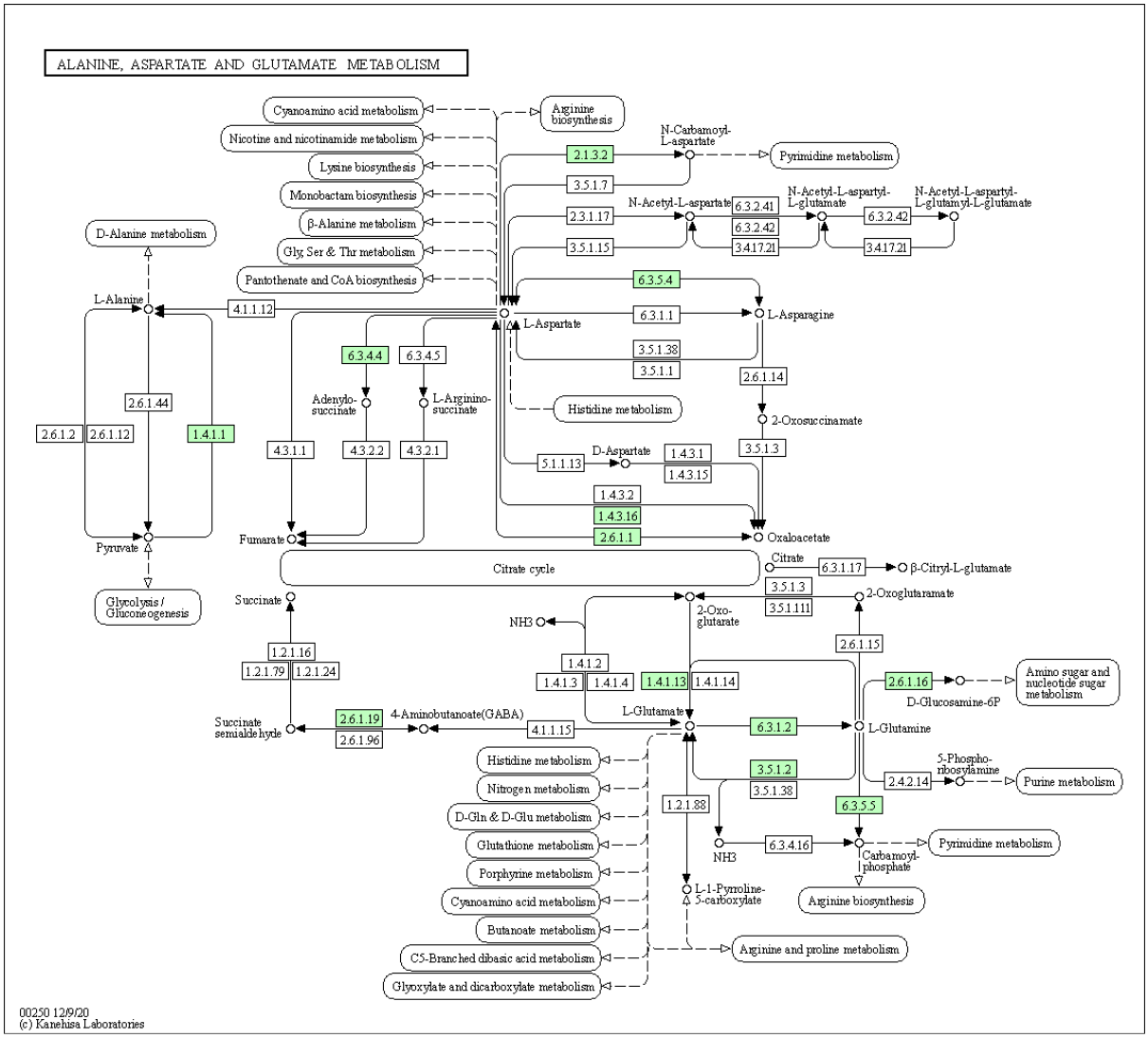


h) Pathways involved to cellular processing (CP). genetic information processing (GIP) and environmental information processing (EIP):


40
35
30
25
20
15
10
5
0
Quorum sensing
RNA degradation
Ribosome
Two-component system
HIF-1 signaling pathway
ABC Transporter
CP
GIP
EIP
M1-M2
M2-M3
M3-M4

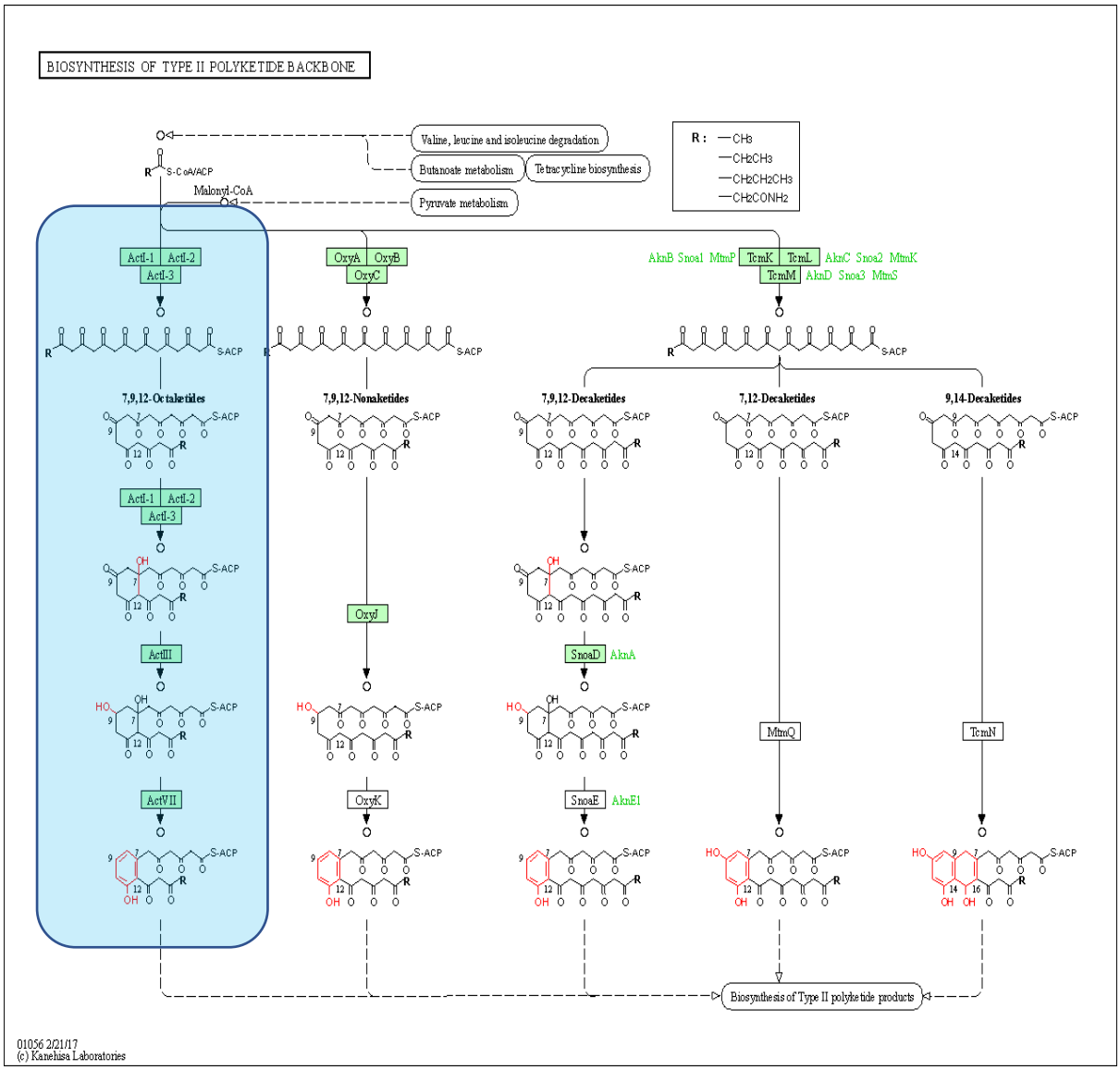
Appendix 5

DEGs mapped to aspartate, glutamate, and alanine metabolism (mapped DEGs are in yellow boxed with their corresponding EC (Enzyme code), EC: 6.3.4.4, EC: 6.3.5.4, EC: 2.6.1.16 and EC: 2.6.1.19 were upregulated:



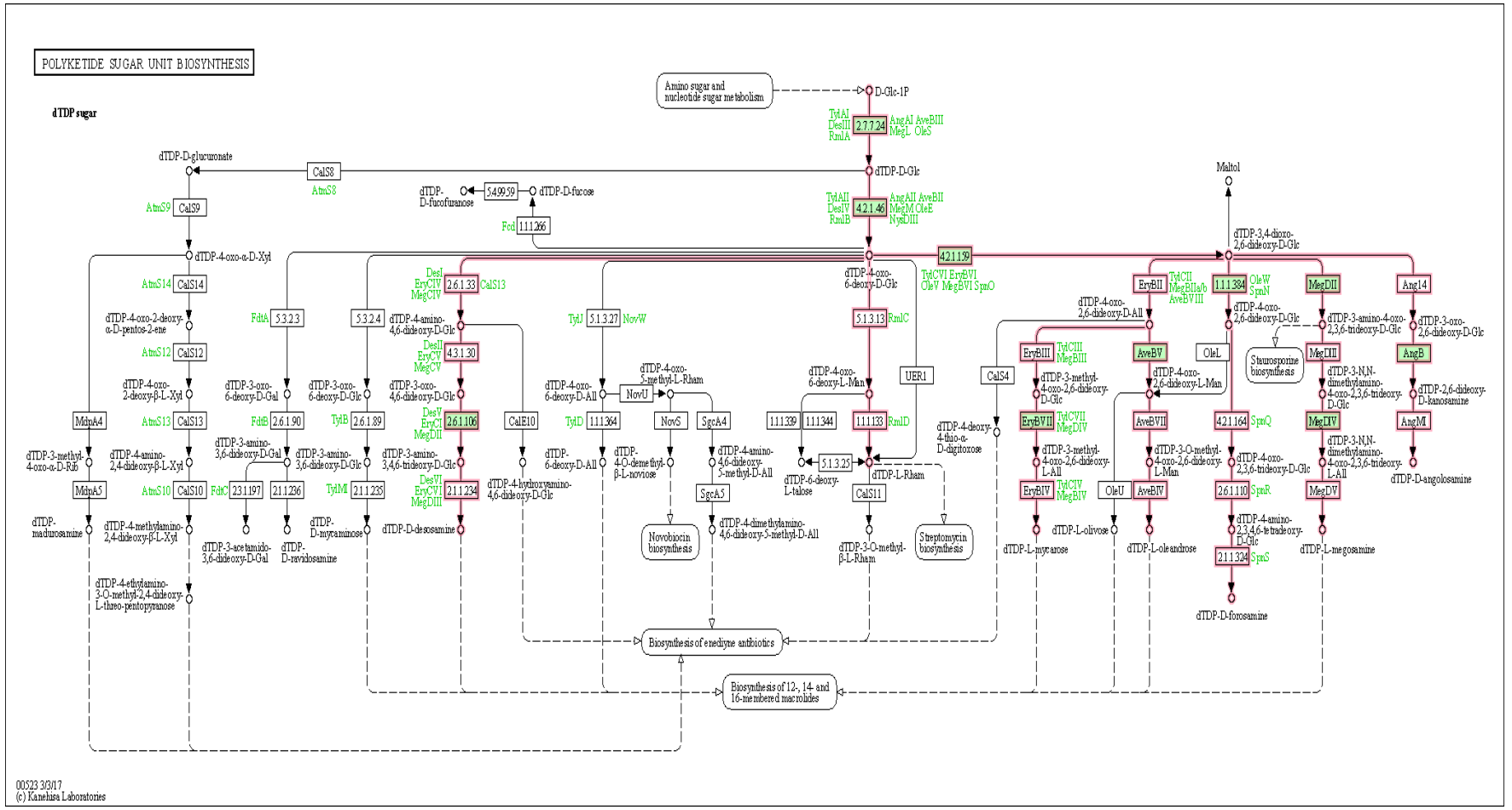
Appendix 6

DEGs mapped to type II polyketide backbone pathway (mapped DEGs are boxed in yellow, and the pathways involved are in blue background). actI1; minimal PKS ketosynthase (KS/KS alpha), actI2; minimal PKS chain-length factor (CLF/KS beta), actI3; minimal PKS acyl carrier protein, actVII; aromatase and actIII; ketoreductase.



Appendix 7

DEGs mapped to pathways for polyketide sugar unit biosynthesis (mapped DEGs are boxed in yellow within an arrow).



Appendix 8

List of DEGs from the BGC responsible for producing spore pigment.

	WT			MT		
Gene ID	W1-W2	W2-W3	W3-W4	M1-M2	M2-M3	M3-M4
fig|33899.16.peg.7325	0.05					
fig|33899.16.peg.7368		0.70				
fig|33899.16.peg.7370		9.29	1.43			
fig|33899.16.peg.7371		8.41	1.04			
fig|33899.16.peg.7372		8.58	1.28			
fig|33899.16.peg.7373		7.61				
fig|33899.16.peg.7374		4.17	1.64			
fig|33899.16.peg.7375		5.50	1.27			
fig|33899.16.peg.7376		10.60	1.10			
fig|33899.16.peg.7377		4.23				
fig|33899.16.peg.7378			2.32			
fig|33899.16.peg.7379		7.76	1.39			
fig|33899.16.peg.7385				1.69		

Appendix 9

Number of hypothetical proteins in the different DEG lists.

DEG list	Total	Total number of DEG	% of hypothetical protein
D1	234	891	26.26
D2	459	1573	29.17
D3	470	1638	28.69
D4	413	1392	29.66
W1-W2	251	915	27.43
W2-W3	201	810	24.81
W3-W4	103	353	29.17
M1-M2	187	844	22.10
M2-M3	42	145	28.96
M3-M4	46	170	27.05

Appendix 10

DEGs (from D2) mapped to glycolytic pathways are yellow box. The number represents their EC number. Embden-Meyerhof pathway showed in red color. DEGs within blue background were upregulated,