BRNO UNIVERSITY OF TECHNOLOGY

Faculty of Electrical Engineering
and Communication

MASTER'S THESIS

Brno, 2020 Bc. Jana Schwarzerová

BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
FAKULTA ELEKTROTECHNIKY
A KOMUNIKAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF BIOMEDICAL ENGINEERING
ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

GENE REGULATION IN CLOSTRIDIUM BEIJERINCKII NRRL B-598
GENOVÁ REGULACE V CLOSTRIDIUM BEIJERINCKII NRRL B-598

MASTER'S THESIS
DIPLOMOVÁ PRÁCE

AUTHOR — **Bc. Jana Schwarzerová**
AUTOR PRÁCE

SUPERVISOR — **Mgr. Ing. Karel Sedlář, Ph.D.**
VEDOUCÍ PRÁCE

BRNO 2020

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

Master's Thesis

Master's study program **Biomedical Engineering and Bioinformatics**

Department of Biomedical Engineering

Student: Bc. Jana Schwarzerová ***ID:*** 186686

Year of study: 2 ***Academic year:*** 2019/20

TITLE OF THESIS:

Gene regulation in Clostridium beijerinckii NRRL B-598

INSTRUCTION:

1) Prepare a literature review of gene regulatory networks, network design methods, and possibilities of their static and dynamic analyses. 2) Study the possibility of using lab data for inference of new gene regulatory networks. Aim on lab techniques for measuring gene expression, primarily RNA-Seq. 3) Preprocess raw RNA-Seq data from lab experiments with the strain Clostridium beijerinckii NRRL B-598 and transform the data into a format suitable for gene regulatory network inference. 4) Construct a gene regulatory network for C. beijerinckii NRRL B-598 using various techniques of network inference. 5) Analyze the proposed network using static analyses of systems biology. 6) Discuss the results.

RECOMMENDED LITERATURE:

[1] MARBACH, Daniel, James C COSTELLO, Robert KÜFFNER, et al. Wisdom of crowds for robust gene network inference. Nature Methods. 2012, 9(8), 796-804.

[2] SEDLAR, Karel, Pavlina KOSCOVA, Maryna VASYLKIVSKA, Barbora BRANSKA, Jan KOLEK, Kristyna KUPKOVA, Petra PATAKOVA a Ivo PROVAZNIK. Transcription profiling of butanol producer Clostridium beijerinckii NRRL B-598 using RNA-Seq. BMC Genomics. 2018, 19(1), 415.

Date of project specification: 3.2.2020 ***Deadline for submission:*** 29.5.2020

Supervisor: Mgr. Ing. Karel Sedlář, Ph.D.

prof. Ing. Ivo Provazník, Ph.D.
Chair of study program board

Abstrakt

Diplomová práce se zabývá studiem genové regulace v *Clostridium beijerinckii* NRRL B-598, pro následné odvození genové regulační sítě bakterie *C. beijerinckii* NRRL B-598. V teoretické části této práce je uvedena obecná nomenklatura problematiky genové regulace se zaměřením na nomenklaturu genových regulačních sítí. Následně jsou zde popsané laboratorní metody, sloužící pro získání vhodných dat popisující expresi genů. Tato data jsou základem pro studium genové regulace a návrhy genových regulačních sítí. Práce se zaměřuje především na technologii RNA-Seq a stručný popis laboratorních dat získaných ze zmíněné bakterie *C. beijerinckii* NRRL B-598. V praktické části se práce zabývá předzpracováním těchto surových laboratorních dat a následným studiem genové regulace se zaměřením na odvození operonů a vytvoření prvních genových regulačních sítí pomocí různých přístupů pro *C. beijerinckii* NRRL B-598.

Klíčová slova

Genová regulace, Genová regulační síť, *Clostridium beijerinckii*, RNA-Seq, Transkriptom

Abstract

The master's thesis deals with the study of gene regulatory in *Clostridium beijerinckii* NRRL B-598 for inference gene regulatory network for *C. beijerinckii* NRRL B-598. The theoretic part describes basic nomenclature gene regulatory with the main focus on gene regulatory networks nomenclature. Laboratory methods which serve to obtain suitable gene describing express data are described there. These data are based on the study of gene regulatory and inference gene regulatory networks. The thesis is mainly focused on the RNA-Seq technology and brief description of laboratory data which were gathered using the strain *C. beijerinckii* NRRL B-598. In the practical part of the thesis pre-processing of these raw laboratory data and following gene regulatory research is performed which focuses on inference operons and creating first gene regulatory networks for *C. beijerinckii* NRRL B-598 using different approaches.

Keywords

Gene regulatory, Gene regulatory network, *Clostridium beijerinckii*, RNA-Seq, Transcriptome

Rozšířený abstrakt

V dnešní době je kladen velký důraz na výrobu biopaliv jako obnovitelných zdrojů kvůli stále větší snaze o ochranu životního prostředí. Využití bakterií produkujících butanol by mohlo být revolučním ekologickým řešením. Kmen *Clostridium beijerinckii* NRRL B-598 je známý jako producent butanolu a jeho genotyp i fenotyp je již za různých podmínek dobře popsaný. Nicméně spojení těchto dvou přístupů, které umožní popis mechanismu genové regulace chybí.

Diplomová práce, Genová regulace v *Clostridium beijerinckii* NRRL B-598, je zaměřena na vytvoření genové regulační sítě pro *C. beijerinckii* NRRL B-598. Genová regulační síť slouží právě pro popis a snadnější pochopení řízení genové exprese v jednotlivých organismech.

Mezi současné hlavní výzvy v oblasti systémové biologie patří studium genových regulací společně s popisem nemodelových organismů. Tyto hlavní cíle systémové biologie jsou novodobě řešeny *in silico* pomocí algoritmů založených na principech strojového učení. Jelikož se jedná o kombinaci bioinženýrství s umělou inteligencí, je práce pro snadnější pochopení tohoto širokého spektra rozložená do pěti hlavních kapitol, z nichž první tři lze považovat za souhrn potřebných teoretických znalostí a následné dvě popisují praktický výzkum analýzy genových expresních dat z kmene *C. beijerinckii* NRRL B-598.

První kapitola se zabývá obecnou nomenklaturou genové regulace se zaměřením na modelování genových regulačních sítí. Jedná se o stručný, avšak důležitý úvod do dané problematiky sloužící pro snadnější orientaci v práci. Dále zde nalezneme shrnutí potřebných teoretických znalostí s rozdělením do oblasti statického a dynamického modelování genových regulačních sítí.

Druhá kapitola se zaměřuje na popis získávání laboratorní dat, která slouží pro analýzu genové exprese. Jedná se především o popis sekvenačních metod využívající se v současné době. Hlavní důraz je kladen na objasnění sekvenační metody Illumina, jelikož právě touto metodou jsou získána data, která následně analyzujeme v této práci. Popis laboratorních metod se především zaměřuje na zisk RNA-Seq dat sloužící pro popis transkriptomu. V závěru druhé kapitoly je stručný avšak specifický úvod do mikrobiologie zaměřený na kmen *C. beijerinckii* NRRL B-598, z něhož jsou nasnímána konkrétní RNA-Seq data, která jsou zde zevrubně popsána.

Třetí část uvádí základní přehled algoritmů používaných v současné době pro odvození sítí se zaměřením na problematiku genové regulace. Algoritmy jsou založeny na principech strojového učení. Celá kapitola je výrazně inspirována metodami, které byly odvozeny v rámci projektu DREAM4, jehož hlavním cílem bylo odvození simulace genových regulačních sítí a predikace genové exprese. Jelikož se jedná o náročnou problematiku, která je v dnešní době ovšem základem pro studium genové regulace, je tato kapitola rozdělena do pěti části, v nichž každá část je vyhrazená pro metody založené na jednom principu strojového učení.

Následují kapitoly, ve kterých je popsán provedený výzkum genové exprese zaměřený na bakteriální kmen *C. beijerinckii* NRRL B-598. Praktická část je rozdělena na dvě hlavní kapitoly, konkrétně do čtvrté a páté kapitoly. Obě části jsou však pro snadnější pochopení systematicky rozdělené po jednotlivých krocích prováděné analýzy.

Čtvrtá kapitola pojednává o předzpracování surových laboratorních RNA-Seq dat. Surová laboratorní RNA-Seq data jsou velmi obsáhlá a jejich zpracování je výpočetně náročné. Celá tato část práce využívala výpočetní virtuální zdroje MetaCentrum. Podle postupu dodržující všeobecné podmínky pro zpracování RNA-Seq dat byly vytvořeny shell-skripty, které jsou reprodukovatelné i pro zpracování jiných surových laboratorních RNA-Seq dat pro různé bakteriální kmeny.

Postup předzpracování surových laboratorních RNA-Seq dat je rozdělen do šesti částí. První krok zahrnuje vygenerování reportu kvality nasnímaných dat, který je následně srovnáván s reporty vygenerovaných během a po předzpracování RNA-Seq dat. Pro vytváření reportu kvality jsme použili FastQC, aplikaci vytvořenou pro kontrolu kvality dat získaných z vysoce výkonného sekvenování. Výstup z této aplikace je report ve formátu HTML, kde je uživatelsky příznivě umožněn interaktivní pohled na kvalitu nasnímaných dat. Podrobnější přehled týkající se HTML reportů naleznete v kapitole 4.1, kde je odkazováno na přiloženou elektronickou přílohu HTML reportů z předzpracovaných dat.

Dalšími důležitými kroky v rámci předzpracování RNA-Seq dat jsou filtrace ribozomální RNA (rRNA), trimování dat a mapování čtení k referenčnímu genomu. Všechny tyto kroky jsou popsány v kapitolách 4.2 až 4.5. Po každém kroku byl vyhotoven HTML report, pro ověření správnosti provedeného kroku. Celý postup předzpracování byl implementován do shell skriptů. Výstupem z vytvořených shell skriptů jsou data ve formátu BAM, která jsou dále zpracována programovacím jazykem R do formátu *'count table'* pomocí dvou různých přístupů. Díky tomuto kroku si vytvoříme dva typy datasetů založených na odlišném přístupu, ale stejných datech, které dále porovnáváme a tím ověřujeme správnosti provedených analýz.

Posledním krokem při předzpracování bylo nutno provést normalizaci dat. V této práci jsme provedli tzv. RPKM normalizaci dat, která je v dnešní době velmi populární, nicméně není dokonalá, jelikož se příliš spoléhá na neomezený dynamický rozsah dat a proto byl následně implementován i jiný způsob normalizace dat založený na negativním bionickém rozdělení, tzv. DeSeq2. V dalších částech práce dále pracujeme s DeSeq2 normalizovanými daty. Obě tyto normalizace byly implementovány do programovacího jazyka R a byly vytvořeny R skripty s aplikací dostupných funkcí ze softwaru Bioconductor.

Vytvořený toolbox pro zpracování surových laboratorních RNA-Seq dat byl aplikován na RNA-Seq data získaná z kmene *C. beijerinckii* NRRL B-598. Díky získanému vícedimenzionálnímu datasetu jsme pro vizualizaci zvolili analýzu hlavních komponent (PCA). Výsledná zobrazení jsou popsána v kapitole 4.6 a uvedená v příloze Attachment B], Attachment C] a Attachment D]. Výsledná

zobrazení uvedená v příloze Attachment B], Attachment C] a Attachment D] jsou důkazem splnění teoretických předpokladů o správnosti a reprodukovatelnosti vytvořeného toolboxu, jelikož kontrolní vzorky, které byly snímané za stejných podmínek, avšak při jiných experimentech, vytvořili shluky. Po ověření reprodukovatelnosti vytvořeného toolboxu, složeného z shell skriptů a z R skriptů, byl toolbox nahrán pro veřejnou dostupnost na github pod /JanaSchwarzerova/Analytical-pipeline-rawRNA-Seq.

Poslední část diplomové práce popisuje výzkum genové regulace založený na expresních datech z *C. beijerinckii* NRRL B-598, jedná se o pátou kapitolu. Tato kapitola je rozdělena do tří hlavních částí. První část se zaměřuje na odvození operonů. Operony popisované v nemodelových organismech, jsou často odvozovány pouze pomocí programů založených na prohledávání databází či novodobějším přístupu programů založených na predikaci pomocí algoritmů strojového učení jako je online softwarový nástroj Operon-Mapper. V této práci byl vytvořen postup implementovaný v jazyce R, který kombinuje získanou informaci o predikaci operonů z již zmiňovaného dostupného online nástroje Operon-Mapperu a informaci získanou z nasnímaných a zpracovaných expresních RNA-Seq dat. Celkem jsme identifikovali 2 737 operonů.

Druhá část v páté kapitole se zabývá vytvořením co-expresních sítí pro *C. beijerinckii* NRRL B-598, které jsou založené na principu pomocí Pearsonova korelačního koeficientu a vzájemné informace. Oba tyto přístupy byly implementovány v programovacím jazyce R s využitím dostupných nástrojů z Bioconductoru. Vytvoření sítí z datasetu pro všech 5 276 získaných exprimujících genových informací není nic výpočetně jednoduchého, a tak i v této části byl využit virtuální výpočetní prostor, MetaCentrum. Co-expresní sítě byly vytvořené pro všechny datasety oběma přístupy. Tyto sítě byly dále zpracovány statickou analýzou a porovnány v diskuzích mezi sebou.

Následně byly vytvořeny první genové regulační sítě *C. beijerinckii* NRRL B-598, pomocí tří různých přístupu. První přístup byl založen na metodě bootstrapingu, druhý na stromové metodě uplatňujících se v algoritmů náhodných lesů a třetí pomocí metod založených na diferenciálních rovnicích. První dva přístupy byly implementovány v programovacím jazyce R a třetí přístup, který je založen na principu diferenciálních rovnic byl implementován v programovacím prostředí Matlab R2019b. Znovu byly vytvořeny sítě pro všechny datasety pomocí všech přístupů, které byly statisticky analyzovány a srovnány mezi sebou. Konečný výsledek vznikl sjednocení všech použitých přístupů k odvození genové regulační sítě a následného průniku sítí vytvořených pomocí odlišných datasetů. Jedná se o první genovou regulační síť pro *C. beijerinckii* NRRL B-598 tvořenou z 8 787 hran.

Bibliographic citation:

SCHWARZEROVÁ, Jana. *Gene regulation in Clostridium beijerinckii NRRL B-598*-[online]. Brno, 2020. Available from: https://www.vutbr.cz/studenti/zav-prace/detail/126847. Master's Thesis. Brno University of Technology, Faculty of Electrical Engineering and Communication, Department of Biomedical Engineering. Supervisor Karel Sedlář.

Declaration

I declare that I have elaborated my master's thesis on the theme of Gene regulation in *Clostidium beijerinckii* NRRL B-598 independently, under the master's thesis supervisor and with the use of technical literature and other sources of information which are all quoted in the thesis and detailed in the list of literature at the end of the thesis. As the author of the master's thesis I furthermore declare that, concerning the creation of this master's thesis, I have not infringed any copyright. In particular, I have not unlawfully encroached on anyone's personal copyright and I am fully aware of the consequences in the case of breaking Regulation § 11 and the following of the Copyright Act No 121/2000 Vol., including the possible consequences of criminal law from Regulation second part, head VI. part 4 of Criminal Act No 40/2009 Vol.

Brno 29th May 2020 ……

signature author

Acknowledgement

I would like to thank Bc. Martin Černý for helping with the completion of the knowledge from the microbiology field. I would like to thank Ing. Kateřina Jurečková for help to solve the final pre-processing step. I would like to thank Mr. Pavel Vítek for help with the correction of the thesis and for the progressive feedback. I would like to thank my brother Ing. Jan Schwarzer for six hours of his life which were devoted to the detailed correction in the thesis. The huge thank belongs to Mgr. Jana Jeřábková for the professional correction of English language. At the end I would like to thank Ing. Jana Musilová for the friendly discussion of interim procedure and results and my biggest thank belongs to my supervisor Mgr. Ing. Karel Sedlář Ph.D. for expert advice, help and patience during explaining necessary knowledge.

Computational resources were supplied by the Ministry of Education, Youth and Sports of the Czech Republic under the Projects CESNET (Project No. LM2015042) and CERIT-Scientific Cloud (Project No. LM2015085) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

Brno 29th May 2020 ……

signature author

List of figures

List of tables

List of equations

List of abbreviations

ANN	Artificial Neural Network
C1-FFL	Coherent Feedforward Loop
DAG	Directed Acyclic Graphical method
dCTP	deoxycytidine triphosphate
dUMP	deoxyuridine monophosphate
dUTP	deoxyuridine triphosphate
DREAM	Dialogue on Reverse Engineering Assessment and Methods project
FPKM	Fragments Per Kilobase of exon Model per million reads mapped
FTFF	Dataset with using parameters – False/True/False/False see Table 2
FTTT	Dataset with using parameters – False/True/True/True see Table 2
GRN	Gene Regulatory Network
IGV	Integrative Genomics Viewer
Lasso	Least Absolute Shrinkage and Selection Operator
MDS	Multidimensional scaling
mRNA	messenger RNA
NGS	Next Generation Sequencing
ODE	Ordinary Differential Equation
PCA	Principal Component Analysis
PYG	Peptone Yeast Extract Glucose
QA	Quality Assessment
QC	Quality Control
RNA-Seq	RNA – Sequencing
RPKM	Reads Per Kilobase of exon Model per million reads
rRNA	ribosomal RNA
SBS	Sequencing by S
SG	Systems Genetics
TPM	Transcripts Per kilobase Million
TSNI	Time Series Network Identification

Contents

Introduction

Clostridium beijerinckii NRRL B-598 is a bacterium, which belongs to the group of non-model organisms. System biology deals with description gene regulatory of organisms and the main challenge is description of non-model organisms. *C. beijerinckii* NRRL B-598 is a relatively well-described butanol producer, which is in demand, because nowadays there is a focus on sustainable microbial production of bio-based fuels.

The description of gene regulatory is performed by using gene regulatory networks. The gene regulatory networks are main challenge for explanation how exactly genomic sequence encodes the regulation of expression of the sets of genes. This thesis is focused on studying gene regulatory in *C. beijerinckii* NRRL B-598 which follows the creation of gene regulatory network for the strain *C. beijerinckii* NRRL B-598.

The thesis is divided into five parts for easy understanding. In the first part, there is description generally but important gene regulatory nomenclature. The second part mentions the laboratory method for gaining suitable data which are basis for proposal gene regulatory networks. There is the main focus on technologies of RNA-Seq. The description of laboratory data using the strain *C. beijerinckii* NRRL B-598 is also gathered in this part. The third part of the thesis describes the most used methods for inference gene regulatory networks currently which are usually based on machine learning principle.

The practical research is described in the fourth and fifth parts. In the fourth part the used pipeline for pre-processing raw RNA-Seq data and the following evaluation pre-processing analysis are written. This pipeline is allowed to obtain the same or even higher information value from RNA-Seq data than microarray data. The final fifth part is focused to research gene regulatory in *C. beijerinckii* NRRL B-598 from obtained express values of pre-processed RNA-Seq data. The first part of the research is focused on inference operons in *C. beijerinckii* NRRL B-598. Then co-expression networks were created by pre-processed RNA-Seq data and in the end of the research gene regulatory networks for *C. beijerinckii* NRRL B-598 was derived.

1 Gene regulatory networks

Gene regulation is a mechanism for controlling which gene gets expressed and at what level [1]. Gene regulation is a mechanism that operates to induce and repress the expression of a gene [2]. These mechanisms include structural and chemical changes to the genetic material, binding of proteins to specific DNA elements to regulate transcription, or mechanisms that modulate translation of messenger RNA (mRNA).

Each specific regulatory molecule controls a specific gene that is transcribed into mRNA [2]. These molecules help or block the transcription enzyme, RNA polymerase. There is a cluster of genes under control of a single promoter that is known as an operon. Operon [2] is a cluster on the chromosome, where related genes are located.

Gene regulatory networks (GRN) [3] are represented by the causality of developmental processes. Their main challenge is to explain exactly how genomic sequence encodes the regulation of expression of the sets of genes that increasingly generate developmental patterns and execute the construction of multiple states of differentiation.

1.1 Basic network terminology

Network represents complex systems which emerges from the orchestrated activity of many components that interact each other through pairwise interactions [4]. The components are reduced to the series of nodes that are connected to each other by links, with each link representing the interaction between two components.

Network consists of nodes and links. In formal mathematical language, it is referred as a graph. GRN has many ways for distribution e.g. we can divide GRN to bipartite or directional [5]. Bipartite GRN has two types of nodes. One type is a gene and the others are regulators. Although some genes are regulators of proteins or genes themselves. Directional means that regulators control genes and often not the other way around.

Establishing cellular networks is not simple. Physical interaction between molecules, such as protein-protein, protein-nucleic-acid and so on, can be described easily. The conceptualization of node-link nomenclature is used there. Nevertheless, more complex functional interactions can also be used within this representation. For example,

small-molecule substrates can be envisioned as the nodes of a metabolic network and the links as the enzyme-catalysing reactions that transform one metabolite into another [4].

Networks can be directed or undirected [4]. It depends on the interaction between two nodes, has a well-defined direction, which can be represented, for example the direction from substrate to a product in a metabolic reaction. These networks are called directed. In undirected networks, the links do not have an assigned direction.

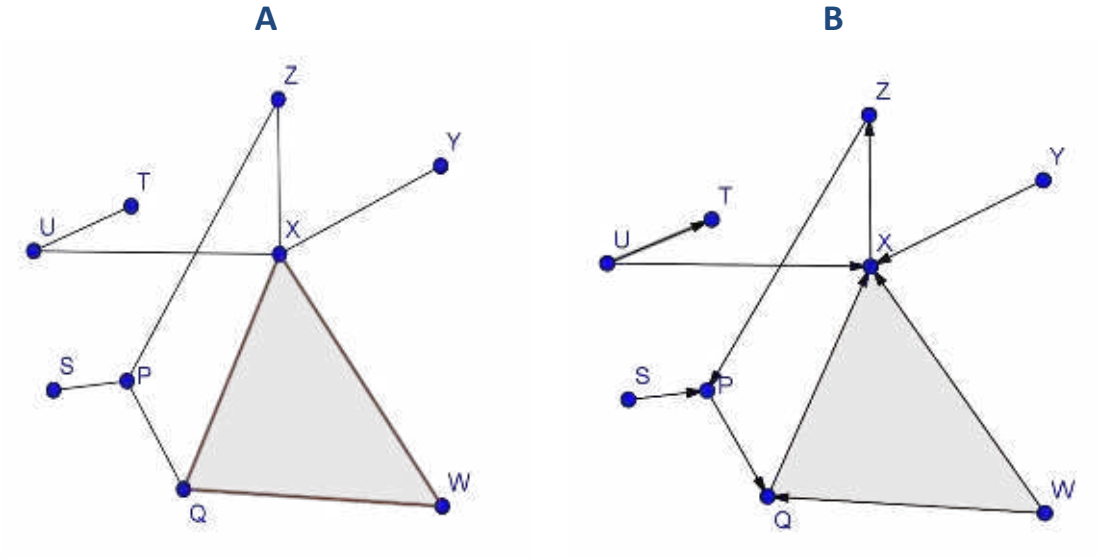


Figure 1: A is undirected network and B is directed network

The degree of node [4] is one of the most important elementary characteristics. The degree usually refers to k and it tells us how many links the node has to the other nodes. The degree distribution [4] refers to $P(k)$. It gives the probability that a selected node has accurately k links. There is also incoming and outgoing degree. Incoming degree [4] denotes the number of links which point, to a node. It refers to k_{in}. Outgoing degree [4] refers to k_{out} and denotes the number of links that start from it. In the Figure 1, we can see two nodes. On the right there is undirected network and on the left we can see directed network which has nine nodes. The undirected network has node X where degree is five. In directed network on the left the node X has four incoming degrees and one outcoming degree.

Biological networks are often scale-free [4]. It depends on their degree distribution which approximates a power law:

$$P(k) \sim k^{-\gamma}. \tag{1}$$

It indicates 'proportional to'. The value of γ refers to a lot of properties e.g. if $\gamma > 3$, the hubs are not relevant. If $2 > \gamma > 3$, there is a hierarchy of hubs and when $\gamma = 2$, a hub

and spoke network emerges. The usual properties of scale-free networks are not valid only for $\gamma < 3$. In this situation the dispersion of degree distribution is defined:

$$\sigma^2 = \langle k^2 \rangle - \langle k \rangle^2, \tag{2}$$

where σ is diverged and σ increases with the number of nodes meaning the series of unexpected features. It influences high degree of robustness against accidental node failures. $\langle k \rangle$ is the average degree.

Path length [4] tells us how many links we need to pass through to travel between two nodes. Many alternative pathways usually exist – paths between two nodes but there are one or more of the most important paths, being called the shortest path [4]. The shortest path is mostly the only one and means the path with the smallest number of links between the selected nodes. The path is connected with distance [4] which represents *l.* In directed networks, distance l_{XY} node *X* to node *Y* is usually different from *Y* to *X*, it is l_{YX}.

Clustering coefficient [4] is a phenomenon when node *X* is connected to node *Z* and node *Z* is connected to node *P*, see Figure 1. Clustering coefficient can be quantified as:

$$C_I = 2n_I/k(k-1), \tag{3}$$

where C_I is clustering, n_I is the number of links connecting k, k is a neighbour of node *I* to each other.

1.2 Network modelling

The network modelling is characterized by viewing cells in their underlying network structure at many different levels of detail, it is a cornerstone of systems biology [6]. Two emerging methodologies in network modelling provide invaluable insights into biological systems: static large-scale biological network modelling and dynamic quantitative modelling.

1.2.1 Static large-scale biological network modelling

Static large-scale biological network modelling [6] is focused on integrating, visualizing and topologically modelling of all kinds of omics data sets which are produced by

innovative high throughput screening biotechnologies. Static large-scale biological network modelling includes the following steps.

Firstly, the construction or inference functional biological network from omics and experimental data. These constructed networks give the whole view of biological systems. The next step is to integrate hetero-omics data across species and data type by network model. Researchers try to use constructed network models to integrate all kinds of type experimental data together.

The following step is to topologically analyse the biological network. Researchers try to connect the topological features of biological networks with biological function, design principles of regulation mechanism and evolution of the systems.

The next step is analysis of the biological network. The approach is fairly similar to BLAST [7] in comparative genomics. The strategy to compare the biological network across species or systems can offer a valuable framework for addressing many biological challenges such as deriving unknown biological function or elements by comparison a network model of well-researched systems with a network model of new research systems in systems biology.

At the end, visualization biological network and analytical results are done. Visualization is important because of helping to understand complex biological systems. However, the huge size of datasets with high heterogeneities is the reason why the visualization of large-scale biological networks is a current challenge. In this area, researchers have proposed a wide range of visualization methods, for example 2D, 2.5D and 3D and develop many software tools such as Cytoscape [8].

1.2.2 Dynamic quantitative modelling

Dynamic quantitative modelling [6] focuses on exploring the dynamics of biological systems by applying computational simulation and mathematical modelling. There are many options for developing quantitative model biological systems such as Boolean networks [9], Bayesian networks [10], Monte-Carlo simulation [11] etc.

In this methodology, a dynamic model is built according to the existing network structure, investigates system behaviour over time under various conditions and predicts complex behaviour in response to complex stimuli [6]. These rapid in silico experiments via dynamical modelling are used to gain first insights, form hypotheses and carry out meaningful tests. The dynamical modelling is used for understanding critical parameters,

biologists can technically and statistically design physical experiments for maximum efficacy. Resultant data from all experiments will be compared against simulations in various ways to test assumptions and hypotheses, identify new phenomena and spark new theories.

2 Laboratory data

Reconstructing GRN is a hard and long-standing challenge [12]. This challenge researches in the field of Systems Genetics (SG) [13]. SG solves complexity by integrating the questions and methods of system biology which is connected with the fundamental problem of interrelating genotype and phenotype in complex traits. SG data are genotyped data with other datasets that reflect the effect of a perturbation of the system caused by diverse genotypes.

Genetic studies are regular and consist of only genotype and phenotype data. It enables the identification of genetic loci which affects a given phenotype. Thus, measurements of thousands of molecular phenotypes enable algorithms to elucidate the regulatory networks. A lot of network inference methods have been proposed due to growing use of Next Generation Sequencing (NGS) [13].

NGS is a high-throughput technology that identifies the nucleic acid sequences and variants in a sample [14]. NGS is often subdivided into second-generation sequencing and third-generation sequencing. The high-throughput sequencing refers to the technologies without the physical separation of individual reactions into separate tubes, capillaries or lanes. Instead of it, the sequencing reactions occur parallelly on a solid surface, such as glass or beads, depending on the technology, and are only spatially separated [14]. These methods billions of sequencing reactions occur and are analysed at the same time. It improves the throughput and decreases the labour compared to the older methods as first-generation sequencing e.g. Sanger sequencing.

NGS is the time of commercial products, not famous scientific names. From a commercialization perspective the first NGS was introduced in 2004 by 454 Life Sciences [14]. Later it was purchased by Roche. Within 2 years, other platforms developed e.g. SOLiD [15], Illumina [16]. These platforms are going to be described in more details in the following chapters. In 2011 Iont Torrent [17] was introduced.

2.1 SOLiD

SOLiD is an enzymatic method of sequencing. The method uses DNA ligase, it is an enzyme with the ability to ligate double-stranded DNA strands [15]. Emulsion PCR is used to immobilise and amplify ssDNA primer-binding region. It is called an adapter.

The adapter has been conjugated to the target sequence on a bead. These beads are afterwards deposited onto a glass surface.

Once bead deposition has occurred, a primer of length N is hybridized to the adapter, then the beads are exposed to a library of 8-mer probes which have different fluorescent dye at the 5' end and a hydroxyl group at the 3' end [15]. A complementary probe will hybridize to the target sequence which is adjacent to the primer. DNA ligase is joined the 8-mer probe to the primer. A phosphorothioate linkage between bases 5 and 6 allows the fluorescent dye to be cleaved from the fragment using silver ions [15]. There are used four different fluorescent dyes which have different emission spectra. Thus, cleavage allows fluorescence to be measured and also generated a 5'-phosphate group which can undergo further ligation. The first round of sequencing is completed when the extension product is melted off. The second round of sequencing is performed with a primer of length N-1. The next round of sequencing is used by shorter primers, subsequently there are e.g. N-2, N-3 etc.

Measuring the fluorescence ensures that the target is sequenced. SOLiD is rarely used as a method of NGS, because it is slower than other NGS methods and has a problem with palindromic sequences. SOLiD is used for short read.

2.2 Illumina

Illumina sequencing is a method that generates millions of highly accurate reads making it much faster and cheaper than other sequencing methods [16]. Illumina sequencing instruments and reagents support massively parallel sequencing using a proprietary method that detects single bases as they are incorporated into growing DNA strands [18].

The procedure consists of a few steps. The first step is breaking up the DNA into more manageable fragments of around 200 to 600 base pairs [16]. Then the adaptors are attached to the DNA fragments and these fragments are made single-stranded. It is provided by incubating the fragments with sodium hydroxide. At the moment, when the fragments are prepared, the fragments are washed across the flowcell. On the surface of the flowcell there is a complementary DNA which binds to primers and therefore the DNA that is not attached is washed away. The DNA attached to the flowcell is then replicated to form small clusters of DNA with the same sequence [16]. During sequencing each cluster of DNA emits a signal that is detected by a camera. After unlabelled nucleotide bases, DNA polymerase are added to lengthen and join the strands of DNA

attached to the flowcell [16]. Thus, bridges of double create stranded DNA between the primers on the flowcell surface.

If we use heat, the double-stranded DNA is broken down into a single stranded DNA and it creates several million dense clusters of identical DNA sequences. In the next step primers and fluorescently-labelled terminators are added to the flowcell and the primer is attached to the DNA strand. Then the DNA polymerase binds to the primer and adds the fluorescently-labelled terminator to the DNA. If the base has been added, there are not more bases which can be added to the DNA strand until the terminator base is cut from the DNA. Lasers are used to activate the fluorescent label on the nucleotide base and this fluorescence is detected by a camera. Each of the terminator bases, it means A, C, G and T, gives off a different colour [16]. After the terminator is removed from the first base, the next fluorescently-labelled terminator base is added and the process continues.

The DNA sequence is analysed base-by-base during Illumina sequencing, making it a highly accurate method [16]. The Illumina sequencing is the cheapest sequencing technology in current [18]. On the other side this sequencing requires higher concentration of DNA and the placement of the clusters on the surface is random so clusters can be overlapped causing confusion nucleotides. The Illumina sequencing technology is divided into several system branches.

2.2.1 MiSeq

The MiSeq system is one of the Illumina sequencing systems which is used for small-genome sequencing. Using up-to-date reagents enables us to produce to 15 Gb of data output with 25 million sequencing reads and 2x300 bp read lengths.

The MiSeq System [19] leverages Illumina sequencing by synthesis (SBS) technology. This system is the first DNA-to-data sequencing platform, integrating cluster generation, amplification, sequencing and data analysis into a single instrument.

2.2.2 NextSeq

The NextSeq [20] is the second system of Illumina sequencing technology. This system is described in more detail because this thesis uses RNA-Seq data, which were sequenced in the NextSeq system. The NextSeq system gives the power of high-throughput

sequencing with the speed, affordability of a benchtop NGS. This system enables the exploration of the entire genome of any species for the deeper understanding of biology.

The NextSeq system allows sequencing a broad range of samples per run e.g. 1 to 12 exomes, 1 to 16 transcriptomes, 6 to 96 targeted panels and 12 to 14 gene profiling samples. This system provides support for paired-end sequencing. Read lengths 2 x 150 bp are defined.

The system is supported by the full suite of Illumina library preparation and target enrichment solutions, offering library compatibility across the Illumina sequencing portfolio [20]. SBS technology with the NextSeq gives exceptional accuracy. This proprietary, reversible, terminator-based method enables the parallel sequencing of millions of DNA fragments, detecting single bases as they are incorporated into growing DNA strands [20]. Thanks to it the method system eliminates errors associated with homopolymer.

2.2.3 HiSeq

The HiSeq system [21] is a high-throughput sequencing system which sequences high-quality data. This system uses Illumina SBS chemistry. HiSeq sequencing systems combine Illumina's proven and widely adopted, reversible terminator-based SBS chemistry with innovative engineering [21].

Nowadays HiSeq system is declared obsolete and it is written on the website by Illumina [18] that they will continue to provide full support of the instruments and supply the reagents through 2024. However, for the sake of the entirety of the thesis, it is necessary to mention this system which is used in the chapter 2.5.1.

2.2.4 NovaSeq

The NovaSeq [22] allows to get scalable throughput and flexibility for any sequencing method or genome. This system offers high-throughput sequencing across a wide range of applications [18]. The NovaSeq is about to leverage proven Illumina NGS technology, multiple flow cell types, two library loading workflows, and various read length combinations.

The NovaSeq allows more cost-effective manner for applications requiring large amounts of data, such as human whole-genome sequencing or ultradeep exome sequencing etc. The instrument can be combined with lower output of flow cells for using less intensive data methods. The NovaSeq system has replaced previously mentioned HiSeq system as shown in the thesis [23].

2.3 Ion Torrent

Ion Torrent technology [17] converts chemically encoded information to digital information on a semiconductor chip. This approach combinates simple chemistry with semiconductor technology.

The principle of the method incorporates nucleotides into a strand of DNA by a polymerase formation of a hydrogen ion which is released. In nature of this principle a formation of a hydrogen ion is a product. Thanks to this fact, there is a possibility to measure voltage which indicates chemically encoded information. If there are two identical bases on the DNA strand, the voltage will be doubled and the chip will record two identical bases [17]. Ion Torrent technology is a direct detection. It means no scanning, no camera, no light. Each nucleotide incorporation is recorded in a few seconds.

2.4 RNA-Seq methods

NGS platforms have a wide variety of methods for obtaining different outputs that is why sequencing methods are divided by differing inputs such as DNA or RNA samples [24]. Sequencing methods have many variants of libraries but the actual sequencing stage is the same, regardless of the method. The various preparation of libraries is destined for different types of sequencing e.g. whole-genome sequencing, RNA sequencing (RNA-Seq), targeted sequencing [24].

RNA-Seq [25] is a sequencing technique of NGS which analyses the transcriptome [26] of gene expression patterns encoded within our RNA. A transcriptome includes mRNA and the information about molecules expressed by an organism.

RNA-Seq is used for understanding the transcriptome which is basic for the exploration of the information hidden within genome with its functional protein

expression. Transcriptomic approaches are used to study transcriptomes whose main challenges of transcriptomics are to catalogue all species of transcript, including mRNAs, non-coding RNAs and small RNAs and also determine the transcriptional structure of gene, splicing patterns and other post-transcriptional modifications [27].

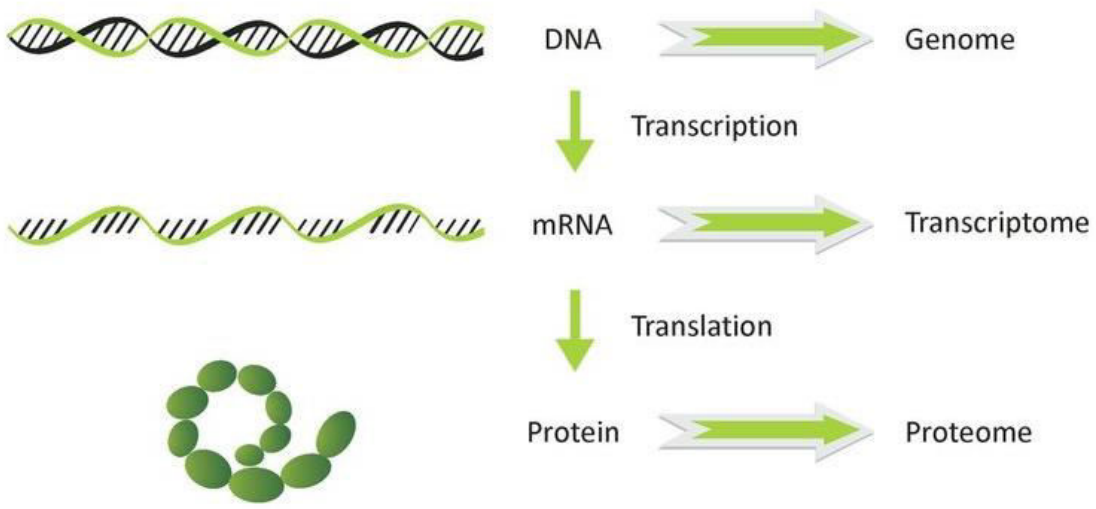


Figure 2: Central dogma of molecular biology where is visualized transcriptome [28]

The regulation of RNA transcription and processing directly affects protein synthesis [29]. This is the main reason why RNA-Seq is important for the description of gene regulation. RNA-Seq includes: post-translational modifications, RNA splicing, RNA bound to RNA-binding proteins, RNA expressed at various stages, unique RNA isoforms, RNA degradation and regulation of other RNA species [24].

While RNA-Seq is emerged as a powerful technology in transcriptome profiling, the main disadvantage of the standard RNA-Seq protocol is that it loses information about the strand of origin for each transcript [30]. If we lose strand information, it is impossible to quantify gene expression levels for gene with overlapping genomic loci accurately which are transcribed from opposite strands. Currently, there is possible to retain the strand information by modifying the RNA-Seq protocol known as strand-specific or stranded RNA-Seq. The comparison of stranded and non-stranded or unstranded RNA-Seq library methods and also their influence on the interpretation of an analysis is described in the study by Griffith et al. [31].

From the bioinformatics point of view RNA-Seq offers challenges which include the development of efficient methods to store retrieve and process large amounts of data [27]. These methods must reduce errors in analysis, base-calling and also eliminate

low-quality reads. Then high-quality reads have been obtained, the first step of data analysis maps the short reads from RNA-Seq to the reference genome or assembles them into contigs before aligning them to the genomic sequence to reveal transcription structure [27]. There are usually used programs for mapping reads to the genome but the short transcriptomic reads cannot be analysed in the same way. These short reads span exon junctions or contain poly(A) ends. Thus, some genomes have rare splicing. These genomes such as genome by *Saccharomces cervisiae* require special attention and only need to be given to poly(A) tails and to a small number of exon-exon junctions. Poly(A) tails can be identified simply by the presence of multiple and exon-exon junctions which can be identified by the presence of a specific sequence context and confirmed by the low expression of intronic sequences which are removed during splicing which is written in the study by Wang et al. [27].

2.5 *Clostridium beijerinckii* NRRL B-598

Clostridium [32] is one of the largest bacterial genera which includes several bacteria with enormous biotechnological potential and also a few well-known pathogens. Members of this genus are generally gram-positive and strictly anaerobic bacteria. Thus, clostridia are in avant-garde of industrially useful microbes.

Due to the required precautions for excluding oxygen during handling, clostridia were virtually inaccessible at the genetic level for a long time [32]. Fortunately, this situation has changed thanks to gene cloning, DNA transfer, gene expression modulation and gene knockout. However, this thesis studies *Clostridium beijerinckii* NRRL B-598 that is the reason why the attention is focussed on the *Clostridium beijerinckii*.

C. beijerinckii culture which is cultivated in Peptone Yeast Extract Glucose (PYG) broth is generally described as straight rods with rounded ends, being motile and peritrichous, measuring 0.5 – 1.7 μm × 1.7 – 0.8 μm [33]. These cells occur as single, in pairs or in short chains. This species is typically gram-stain-positive but become gram-stain-negative in older cultures. Their spores are oval, eccentric to subterminal and swell the cell with no exosporium or appendages [33].

Optimum temperature for the growth is 37 $°C$ and the growth is stimulated by a fermentable carbohydrate, inhibited by 6,5% NaCl or 20% bile acids. The strains of *C. beijerinckii* are nutritionally fastidious, requiring a complex mixture of growth factors. Abundant gas formation is detected in deep cultures in PYG agar [33]. These species are

used to produce industrial solvents. It has the ability to ferment saccharose and to utilize the alcohol sugars D- and L- arbitol, dulcitol and inositol but glycerol only weakly [34]. *C. beijerincekii* has all strains which are able to ferment methyl-glucopyranoside, turanose, dextrin and pectin.

In the past *C. beijerinckii* NRRL B-598 was wrongly classified as *Clostridium pasteurianum* NRRL B-598 as mentioned in the study by Sedlar et al. [35]. The strain *C. beijerinckii* NRRL B-598 is a relatively well-described butanol producer regarding its phenotype under various conditions [36].

In this thesis the laboratory data of *C. beijerinckii* NRRL B-598 which were sequenced using RNA-Seq are used. Following chapters describe conditions in which *C. beijerinckii* NRRL B-598 were sequenced and they also describe more information about laboratory data. There are used 7 replicates of RNA-Seq which are called A, B, C, D, E, F and G. All these replicates are sequenced in six time-points.

2.5.1 Standard cultivation transcriptome

The replicates A, B and C are described in the study by Sedlar et al. [36].The transcription profile of butanol producer *C. beijerinckii* NRRL B-598 is presented there. RNA-Seq dataset covering six time-points with the current highest dynamic range among solventogenic clostridia is used there.

Six time-points cover all metabolic stages within a period of 23 h. The last 24th hour was not analysed because there a was large number of dead and lysing cells. The result of it was the insufficient quality for RNA sequencing. Six time points are mentioned as $\{T_1, T_2, T_3, T_4, T_5, T_6\}$. Individual sampling points were selected based on the fermentation pattern which was monitored on-line as a change in a pH course [36].

Replicate A consists of reads that were 50 *bp* but series B and C consisted of reads that were 75 *bp* long. The final 2D representation in the study Sedlar et al. [36] shows that replicates are similar to each other at particular sampling time-points nevertheless replicates A were slightly more distant to replicates B and C. This is due to the type of sequencing because replicates A were sequenced using Illumina HiSeq and replicates B and C were sequenced using Illumina NextSeq whose principals HiSeq and NextSeq sequencing are contained in the chapter 2.2.

Replicates D and E are described in the study by Patakova et al. [37]. There are two biological replicates D and E. These technical replicates were analysed for changes

in the expression of individual gene and gene clusters. Biological meanings for these expression changes were sought in the study by Patakova et al. [37]. These replicates were selected in to six time-points, too.

The six time-points are selected as points of time in 3.5 *h*, 6 *h*, 8.5 *h*, 13 *h*, 18 *h* and 23 *h* on that account these time-points cell samples were taken from each bioreactor for RNA extraction. Thanks to this approach, the solventogenic phases of growth as well as the sporulation cycles were covered.

Library construction and sequencing of samples from technical replicates were performed by the CEITEC Genomics core facility (Brno, Czechia) on Illumina NextSeq500, single-end, 75 *bp* [37]. It is similar to replicates B and C. All these replicates B, C, D, E are obtained in the study by Vasylkivska et al. [38] as well. It is written that each broth samples were centrifuged, the cell pellet was washed with sterile distilled water and stored immediately at -70°*C*.

2.5.2 Butanol shock transcriptome

The replicates F and G are described in the study by Sedlar et al. [39]. Transcriptomic data of immediate and later responses towards a non-lethal butanol shock are described there. It was performed in the phase of transition between the late acidogenic phase and early start of the solventogenesis. Butanol was added directly after the sample collection at time 6 *h* [39].

RNA-Seq data set of *C. beijerinckii* NRRL B-598 is also selected to six time-points $\{T_{b1}, T_{b2}, T_{b3}, T_{b4}, T_{b5}, T_{b6}\}$. These time-points are {6 *h*, 6.5 *h*, 7 *h*, 8 *h*, 10 *h*, 12 *h*}. Library construction and sequencing of the sample was performed by CEITEC Genomics core facility (Brno, Czechia) on Illumina NextSeq, single-end, 75 *bp* [39].

3 Network inference methods

Systems Biology [13] has a target to decipher the complex behaviour of a living cell. The effective behaviour of the cell is probably defined through the multiple layers of interacting entities including DNA, mRNA, noncoding RNA, proteins, and metabolites [13]. This chapter is focused on the genetic genomics approach [13] that combines the power of genetics through the polymorphism and unscramble a GRN together with ability of gene expression. It means the representation of the gene-level interactions occurring under given conditions.

Genes are represented by vertices in GRN while directed edges represent the direct causal effect of genes to other genes through gene regulation [13]. Genes in gene regulation are usually called activators or repressors. We can use deciphering the data set of gene regulations and identify the most important and possible indirect players in GRN. These players influence a gene expression or phenotype and so link network structure to associated functional properties. Thus, more understanding of the way of the gene interactions appears that controls the overall cell behaviour. A variety of mathematical formalisms, continuous or discrete defined over time or in stationary states, have been proposed to represent the complex behaviour of known GRN [13].

GRN learning [13] has high-dimensionality where the number of genes in a typical genome is included. The number is larger than the number of samples that can be reasonably produced. Some algorithms decipher GRN structures based on genetic genomics data have used complex multivariate regression or Bayesian networks. The analysis of the output of different statistical methods is targeted at learning in a high dimension (based on the penalized linear regression or penalized Bayesian network structure learning) and it shows and defines the best performer on different datasets of simulated genetic genomics data, including up to 1,000 genes [13].

Different statistical models of gene regulation are described in this chapter. These models have been chosen for their ability to infer gene regulations from expression data automatically. Through the verification and fair assessment of algorithms, there is a high importance to learn which algorithms are the most useful for extracting biological insights from system genetic data [13]. This issue was solved by community effort which is mentioned in the DREAM project. The DREAM [12] is the Dialogue on Reverse

Engineering Assessment and Methods project, which is focused on the comprehensive blind assessment network inference methods.

3.1 Bootstrapping

Bootstrapping [40] is a method that is used to resample and assign measures of accuracy to sample estimates. Bootstrapping is usually used to estimate summary statistics such as the standard deviation or the mean. It is used in the applied machine learning to estimate the skill of machine learning models when making predictions on data not included in the training data [41].

Bootstrapping provides the estimation of the sampling distribution of any statistics and it uses only a simple resampling approach. Unfortunately, it means that repeated computations must be undergone [13]. Bootstrapping includes the approach of sampling with replacement [42] which selects a sample at random from the set which is returned to the set and after a second the other element is selected at random. Whenever a sample is selected, the set contains all data, so the sample can be selected more than once. There is no change in the size of dataset. We can assume that a sample of any size can be selected from the given population of any size [43].

In the GRN inference challenge there is each of replicate dataset that is obtained by random sampling with the replacement from the original sample. For each replicate dataset, the model is fitted and then it is possible to study the statistical properties of the distribution of the considered statistics on all resampled datasets [13].

The major use of bootstrapping is to contribute to the construction using the "confidence score" of edges in the predicted GRN. Bootstrapping is often used in the random forest algorithm where it allows us to avoid overfitting, thanks to the fact that bootstrapping has offered further opportunities. Since bootstrap datasets are obtained by sampling with replacement, each of them is deprived of around 1 – 0,632 = 36,8 % of the original samples [13]. The main disadvantages of bootstrapping are to multiply the computational burden and the loss of 36,8 % of the data, it may also affect the sharpness of estimates on every resampled dataset [13].

3.2 Regression

Regression analysis [41] is a process for an estimation of the relationships between outcome variables and predictors. It is divided into a linear and non-linear regression. In GRN inference it is often used to approach the linear regression. This approach to GRN inference is based on the assumption that the expression level of the transcription factors that directly regulates a target gene is the most informative, among all transcription factors, to predict the expression level of the target gene [12]. Regression analysis using inference challenge in the GRN is penalized by the linear regression.

A natural approach to solve the network inference problem considers each gene *g* individually from the others and also its expression value which can be represented as a linear function of all other gene expression levels and of all polymorphisms [13]. It represents:

$$E_g = \sum_{j=1}^{p} \alpha_{gj} M_j + \sum_{\substack{j=1 \\ j \neq g}}^{p} \beta_{gj} E_j + \varepsilon_g, \tag{4}$$

where M_j represents the allelic state of the polymorphism associated with gene *j,* α_g is the *p*-vector of linear effects of polymorphisms on E_g, E_j is the expression level of gene *j,* β_g is the *p*-vector of linear effects of other expression levels on E_g and ε_g is the Gaussian residual error term [13].

A dataset needs to be known for the explanation of the linear function which is defined for each GRN [13]. The dataset includes a sample of n recombinant inbred lines that are measured for p bi-allelic markers and p gene expression levels. Every polymorphism is associated with a single gene and may influence either its direct expression (*cis* polymorphism occurring in the regulatory region of the gene) or its ability to regulate other target genes (*trans* polymorphism in the transcribed gene region itself, influencing its affinity with other gene regulatory complexes) [13]. A dataset contains a $n \times p$ matrix e where e_{ij} is the steady-state expression level of gene j for the recombinant inbred lines individual *i* which is the real number and also a $n \times p$ matrix m where m_{ij} represents the allelic state of the polymorphism associated with gene *j* for recombinant inbred lines individual *i* which is a zero or one. Each e_{ij} is an observation of the random variable E_j and similarly each m_{ij} is an observation of the random variable M_j.

Parameters α_g and β_g are estimates for each gene g from matrix m and e using linear regression methods. The great advantage for this model is its simplicity which leads to $2p$ parameters (α_g and β_g) for E_g. It is a desirable property for the estimation in high-dimensionality settings. We can suppose that regulation networks are sparse so it is desirable to use some regularization [44] such as the Lasso regression. Penalized regression methods are used in GRN inference that lead a variable selection. These methods are described below.

3.2.1 Lasso Regression

Lasso (Least Absolute Shrinkage and Selection Operator) regression [44] is a linear regression. The linear regression can be explained in the case of a high number of input variables. It is typical for NGS inference. It is suitable to decrease the model complexity that is the input of variables or predictors. Removing predictors from the model can be seen as the setting of their coefficients to zero. Another way is penalizing them if they are far from zero. This attitude decreases model complexity while keeping all variables in the model.

The linear regression problem [13]:

$$Y = X\theta + \varepsilon, \tag{5}$$

where Y is the linear combination of r regressors $X = (X_1, \ldots, X_r)$ and ε is Gaussian noise. If we have a sample of size n and Gaussian distributions, the estimation of parameters is obtained by minimizing the residual sum of square but exclusively for the Lasso regression penalizes [13] is the residual sum of the square criteria by the sum of the absolute values:

$$\hat{\theta}^{lasso} = arg\min_{\theta} \|Y - X\theta\|_{l_2}^2 + \lambda\|\theta\|_{l_1}, \tag{6}$$

where $\hat{\theta}$ is the estimation of the parameters θ, l_1 represents norm using the absolute values of the parameters θ and l_2 represents minimizing the residual sum of squares.

The feature selection problem is solved in [12] with the Lasso procedure, too. The Lasso procedure can lead to obtaining a sparse linear model such as a model based only on a few transcription factors. The transcription factors selected by Lasso are therefore good candidates to regulate the target gene [12].

3.2.2 The Dantzig Selector

The Dantzig selector [45] is similar such as the penalized linear regression method. This method is based on l_1 norm penalization of the parameters subjected to the constraint bound on the maximum absolute correlation between the residuals and regressors [13]:

$$\hat{\theta}^{dantzig} = arg\ \underset{\theta}{min} \|\theta\|_{l_1}, \|X^T(Y - X\theta)\|_{l_\infty} \leq \delta, \quad (7)$$

where δ is the actual bound of the correlation among the residual and each regressor and X^T is the transpose of *X*. If the bound tends to zero, Dantzig selector imposes a null correlation among the residual and the regressors, else the other of these selectors set all coefficients to zero. This condition is satisfied by the RSS estimate, as it is equivalent to enforcing a null derivative of the RSS [13].

3.2.3 Confidence Scores

This chapter comments confidence scores and they are meant as confidence scores with penalized linear regressions and bootstrap [13]. These confidence scores are on the prediction of each oriented edge $j \longrightarrow g$ where gene *j* influences gene *g*. When α_{gj} is not zero, marker *j* has an impact on the expression of gene *g* hypothetically. The converse is impossible since expression levels cannot affect polymorphism [13]. If β_{gj} is not zero, a relationship exists between the expressions of genes *j* and *g*. However, the causal orientation is unknown. It means, we do not know, if *j* influences gene *g* or conversely.

Choosing the 'right' level of penalization in the Lasso regression or in the Dantzig selector is a difficult model selection problem [13]. Fluente [13] describes the choice of the penalty term λ such as non-fixed value. Nevertheless, all options for penalty values from zero value are explored, it is not penalization to a maximum value. The infimum[I] of the set of all λ precludes a single regressor to be included in any of the regressions. If the total of q use, there are different penalty values from the interval low penalty level $\frac{\lambda_{max}}{q}$ to the maximal penalty level λ_{max}.

[I] Infimum is the greatest lower bound of a set *S*, defined as a quantity *m* in a such way that no member of the set is less than *m,* but if ϵ is any positive quantity, however small, there is always one member that is less than $m + \epsilon$ [46].

A similar principle can be used for the Dantizig selector where the fraction of times was used as confidence score, see [13]. This fraction of times is over all penalizations and a regressor is presented with a nonzero parameter estimate.

3.3 Random Forests

GRN inference can be also used by nonlinear regression methods such as random forests [47]. However, nonlinear regression methods can be also considered, assuming that the expression level of E_g is a function of the remaining expression levels E^{-g} and of allelic states M [13]:

$$E_g = f_g(M, E^{-g}), \tag{8}$$

The use of the random forest for GRN reconstruction from expression data alone has been originally proposed in GENIE3 [13]. Random forests are the method where each node splitting considers only a random subset of features. Non-linear regression problem is between response Y and regressors X. This problem [13] is split recursively into the observed data with binary tests based on each single regressor variable where the variance of the response variable in the resulting subsets of samples should be as small as possible.

In each test there is a binary tree where a node which compares the input variable value with a threshold. This threshold is determined during the tree growing. The leaves of the tree represent the predicted value of the response variable. A random forest [13] includes trees which are grown thanks to two sources of randomness. Each tree is grown using a random bootstrapped sample of the data and the variable used at each split node which is selected only from a random subset of all variables [13].

The mean of all the regressions predicted by each tree is the random forest predicted value. The advantage of random forest is using the bootstrapping to estimate the importance of any or every regressor. After shuffling the values of the regressor considered in the samples that have not been used in each bootstrapped sub-sample, it is possible to compute the resulting increase in the variance of the regression error compared to non-permuted samples [13]. It gives an assessment of the regressor importance.

The confidence scores [13] by a random forest for oriented edge $k \longrightarrow l$ are [13]:

$$w_{kl}^m = 1 - \frac{r_{kl}^m - 1}{N}, \tag{9}$$

where r_{kl}^{m} is a global rank by the edge, N represents the largest overall rank. Ranks are produced by the importance of the factors such as f_g^m that is normalized by its standard deviation. For w_{kl}^{e} is an analogous definition.

3.4 Bayesian Networks

A Bayesian network is a directed acyclic graphical (DAG) model that captures the joint distribution probability over a set of variables by a factorization in local conditional probabilities linking one random variable with its 'parents' [13]. Bayesian network is defined such as a network which can derive directed graph in GRN inference challenge.

The DAG [13] implicitly captures a set of conditional independencies. These interdependencies are a joint probability distribution between variables and represents:

$$P(V) = \prod_{i=1}^{m} P(V_i | Pa(V_i)). \tag{10}$$

If B represents Bayesian network, it can be $B = (\vartheta, P_\vartheta)$ denoted where $\vartheta = (V, A)$ with vertices representing random discrete variables $V = \{V_1, \dots, V_m\}$ linked by a set of directed edges A and a set of conditional probability distributions $P_\vartheta = \{P_1, \dots, P_m\}$ and the variables are involved in each conditional probability table P_i. $Pa\,(V_i) = \{V_j \in V \mid (V_j, V_i) \in A\}$. Moreover, Pa is a set of parental nodes of V_i in ϑ [13].

Maximum likelihood estimates the parameters defining the conditional probability tables therefore it can be computed by simple counting. Then the GRN learning process is reduced to the problem of learning a DAG structure among these variables that maximizes $P(\vartheta|D) \propto P(D|\vartheta)P(\vartheta)$ where D represents the observed data [13]. Learning Bayesian networks is an NP problem and as a result GRN structure learning with bootstrapped greedy search is used in [13].

3.5 Other methods

Different statistical models of gene regulation have been already mentioned and have different ability to infer gene regulations from expression data automatically. Some methods which are chosen in [13] are described above. However, other methods for GNR

inference exist and are described in [12]. [12] mainly mentions methods such as Mutual information and Correlation.

Mutual information [48] defines how much one random variable tells us about another one. For two discrete variables X and Y whose joint probability distribution is $P_{XY}(x, y)$, the mutual information between them, denoted $I(X, Y)$, is given by [48]:

$$I(X,Y) = \sum_{x,y} P_{XY}(x,y) \log\left(\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}\right) = E_{P_{XY}} \log\left(\frac{P_{XY}}{P_X P_Y}\right), \tag{11}$$

where $E_{P_{XY}}$ is the expected value of the distribution P_{XY}. In the GRN implementation [12], X and Y represent a transcript factor and target gene.

Correlation in DREAM is used as Pearson's correlation and as Spearman's correlation. Pearson's correlation coefficient r was calculated among all transcription factors x and all target genes y in [12] as follows:

$$r_{xy} = \frac{n\sum x_i y_y - \sum x_i y_i}{\sqrt{n\ \sum x_i^2 - (\sum x_i)^2}\sqrt{n\ \sum y_i^2 - (\sum y_i)^2}}, \tag{12}$$

where n represents the number of measurements of x and y. Positively correlated gene pairs receive higher confidence.

Spearman's correlation was calculated among all transcription factors x and all target genes y in [12] as follow:

$$\rho_{xy} = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \tag{13}$$

where n represents the number of conditions that x and y have been sampled and d is the difference in the rank order between gene x and gene y over the n conditions [12]. The most correlated gene pairs were selected there.

4 Pre-processing RNA-Seq

RNA-Seq is a method for measuring gene expression, which gives us some information about mRNA as it was mentioned above, see chapter 2.4 for more information. RNA-Seq is more difficult for analysis than microarray data, however nowadays RNA-Seq has defined mature pipelines.

The pre-processing part deals with the pre-processing RNA-Seq process which is applied to pre-processing RNA-Seq data *C. beijerinckii* B-NRRL 598. These data were obtained under different conditions, more information is described in the chapter 2.5. In Attachment A] you can see a example of shell-script which is used for all samples. All created shell-scripts were computed using MetaCentrum Virtual Organization portal [49].

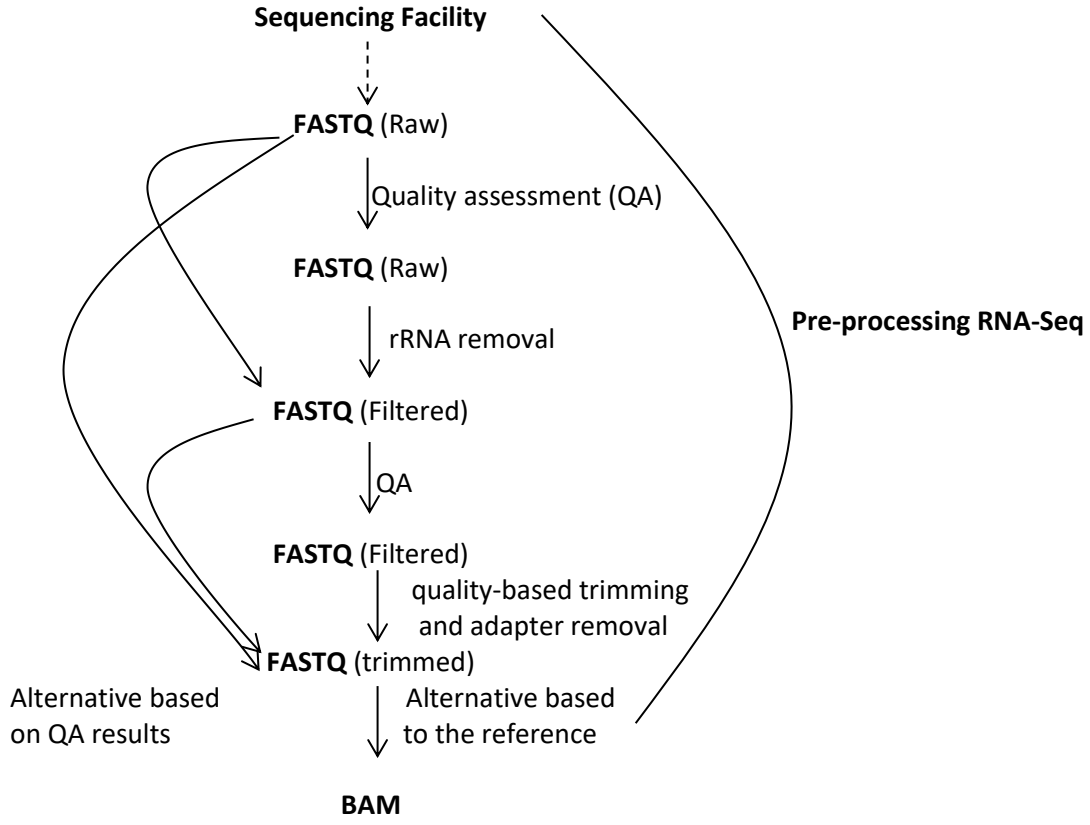


Figure 3: The first part of RNA-Seq pre-processing workflow [50]

This thesis is inspirited by pipelines from the study by Delhomme et al. [50]. These pipelines represent the standard workflows in pre-processing RNA-Seq. Figure 3 shows the graphic representation of the first part of the procedure. There are nodes that represent data format e.g. fastq, and edges represent the process treatment data.

The next part of RNA-Seq pre-processing workflow can be divided into QC, read mapping and alignment, quantification and differential gene expression analysis [51]. The quantification is an approach to quantifying gene expression by RNA-Seq to count the number of reads belonging to each gene [51]. The count table is created by R/Bioconductor featureCounts [52] in this thesis. However, read counts are influenced by factors such as transcript length or the total number of reads. Thus, it is necessary to normalise the read counts.

The normalisation is usually provided by the RPKM (Reads Per Kilobase of exon Model per million reads), FPKM (Fragments Per Kilobase of exon Model per million reads mapped) or TPM (Transcript Per kilobase Million) [52]. The correction for gene length is not necessary if we compare changes in the gene expression within the same samples but it is necessary for correct ranking gene expression levels within the sample to account for the correct long genes accumulate [51].

4.1 RNA Data QC Assessment

The first necessary step is the initial QC assessment [50]. The overall sequence quality, as GC percentage distribution and the presence or absence of overrepresented sequences are checked here. The output is an HTML document, where some sections show the specific metrics. HTML documents from each replicate are enclosed in the electronic attachment.

Sections in our HTML documents are divided to [50]:

a) **Basic Statistics:** There is self-explanatory metrics. The GC% should be the expected values for the sample species. Median GC% is 29,6 for *C. beijerinckii* NRRL B-598, it is nice agreement with *C. beijerinckii* [34] where GC% is 30.

b) **Per base sequence quality:** The first part is the Phred scale quality. It represents the probability that the base call is incorrect. The second part represents details of the Phred scaled quality as a function of the position in the read.

c) **Per sequence quality scores:** There is a quality distribution at the read level. The good quality data are represented as the histogram which is skewed to the right.

d) **Per base sequence content:** The average proportion of individual bases as a line across the length of the reads is visualized there.

e) **Per base GC content:** GC content is visualized as a function of the position in the read.

f) **Per sequence GC content:** The plot shows the distribution of GC content per read. The red curve represents data and the blue curve represents theoretical distribution. If the curve shows a region of high GC content, it means that the sample includes ribosomal RNA (rRNA). However, this effect can be represented by a contamination by an organism with higher GC content such as bacteria. A peak on the left side represents AT content per read.

g) **Per base N content:** There is a plot that shows the fraction of indistinguishable bases. It is represented as a function of the base position in the reads. This is expected to be close to zero if sequence data is high quality. Sequencing problems are represented as deviations from the expected values.

h) **Sequence length distribution:** It represents the distribution of read lengths. There should be only one peak located at the sequenced read length, prior to trimming.

i) **Sequence duplication level:** It represents the level of duplicate sequences in the library.

j) **Overrepresented sequences:** There is a table where are the sequences that are represented at the unusually large frequency in the reads.

k) **Kmer content:** This plot shows details about the occurrence of Kmers. These are nucleotide sequences of fixed *k* length. These nucleotide sequences are presented higher than the expected frequency as a function of their position within the read.

4.2 rRNA filtering

rRNA filtering is important to maximize the quality of the sequence data. It is necessary to remove as much rRNA as possible [54]. Wet-lab protocols usually include a rRNA removal step but it is recommended to do rRNA filtering there. SortMeRNA [55] was used for rRNA filtering. The algorithm is based on approximate seeds and allows fast and

sensitive analyses of NGS reads. It was originally developed to identify rRNA in metagenomics analyses.

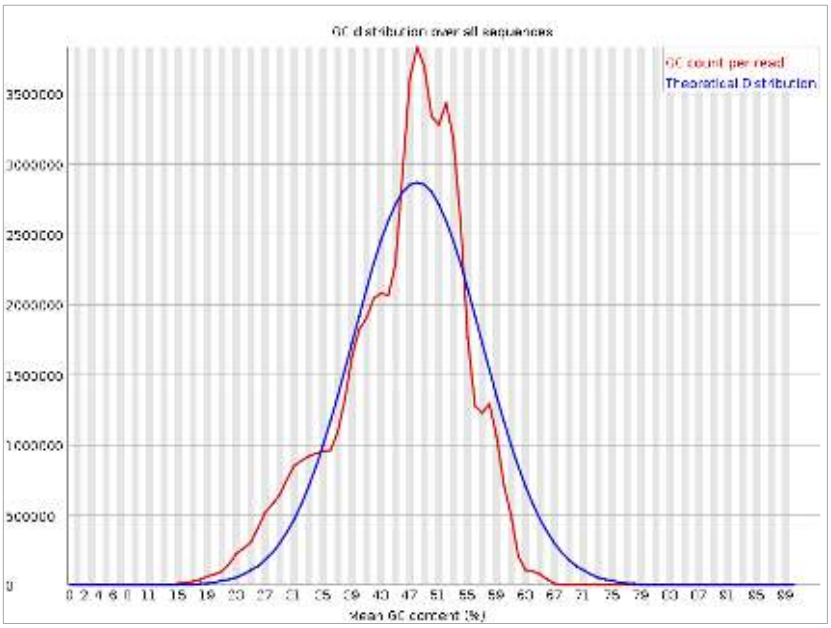


Figure 4: GC distribution data sequences in point-time 8,5 *h* from replicate C before filtering rRNA

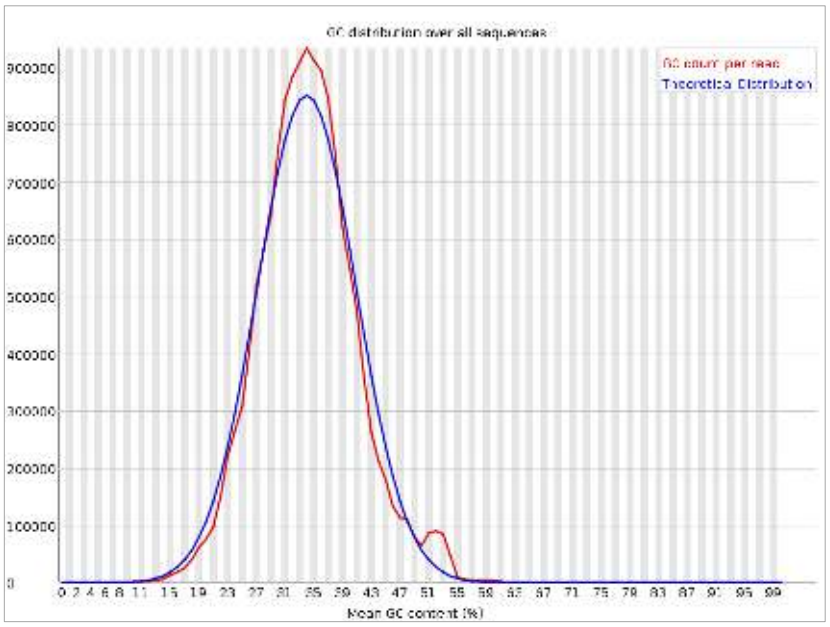


Figure 5: GC distribution data sequences in point-time 8,5 *h* from replicate C after filtering rRNA

The filtered data can be subjected to a QC assessment by FastQC again. The GC content plot should represent the biggest visual change because it is more closely to the theoretical normal distribution and GC curve should be closer to the expected GC value of the sample organism [50]. In Figure 4 and Figure 5 we can see graphs which represent GC distribution before and after rRNA filtering. The red curves are GC count per read and the blue curves expect theoretical normal distribution. After the application

filtered of rRNA step there is a visual change between these graphs where the plot represents our data and is more closely to the theoretical normal distribution as it is expected.

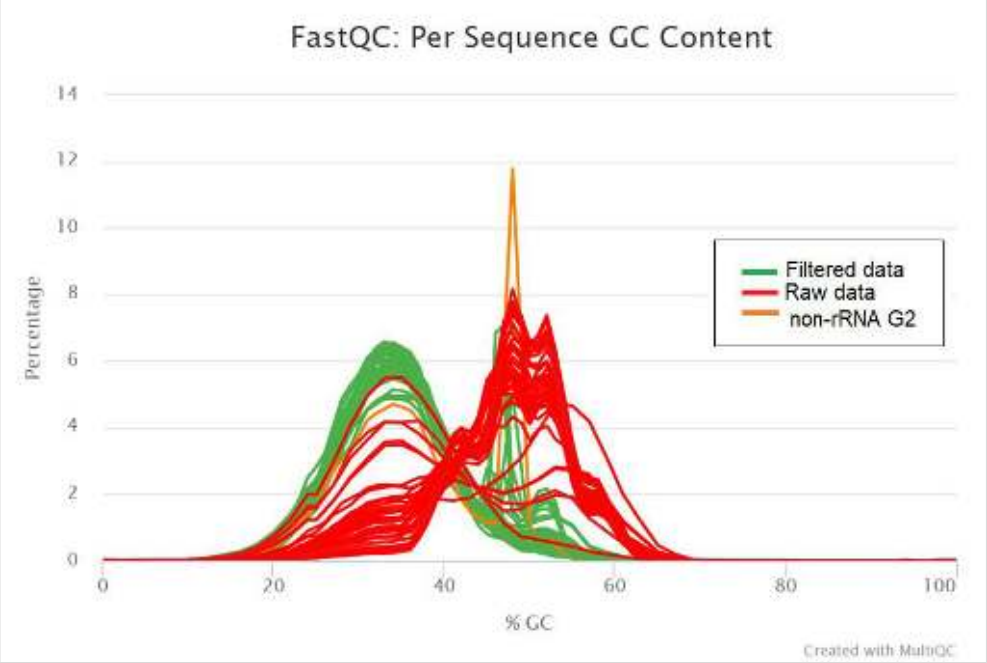


Figure 6: GC distribution data sequences before and after filtering rRNA

Figure 6 was created from all samples of all replicates. This plot is only the combination GC distribution data sequence before and after rRNA filtering. Per sequence GC content where green curve represents filtered data is more closely to the theoretical normal distribution of GC content, too. The modal GC content is calculated from the observed data and used to build a reference distribution. The green colour shows sufficient similarity with this theoretical distribution in contrast with the red colour which represents raw data.

Unfortunately, a specific sample from replicate G which was sequenced in 6.5 *h* time-point shows an unexpected result because the orange curve which represents this filtered samples is not closely to the theoretical normal distribution. This effect can be caused by contamination that can arise during wet-laboratory step. Filtering using bacterial rRNA database is provided and thus it can be assumed that with the help of using another rRNA database e.g. eucaryote rRNA database, it is possible to obtain a better result after filtering. However, this contamination is filtered during mapping on the genome in a following step.

4.3 Quality trimming and adapter removal

It is known on Illumina sequencers that the quality of a base pair is linked to its position in the read so bases in the last cycle of the sequencing process have a lower average quality than the earliest cycles [50]. There is a common approach to increase the mapping rate of reads by removing the low-quality bases, it is called quality trimming. These reads are trimmed from the 3' end until the quality which user selects as Phred-quality threshold is reached. A threshold of 20 is widely accepted [50].

Next issue connected to Illumina sequencing is the presence of partial adapter sequences within sequenced reads [50]. This effect occurs if the sample fragment size has a large variance and if fragments are shorter than the sequencer read-length. As the resulting reads contain a significant part of the adapter which may not be able to map such reads. Thus, the ability to identify adapters, follow clip, or trim them, may consequently significantly increase the aligned read proportion [50].

Trimmomatic tool [56] was used in the step and the data were subjected to QC, too. FastQC is performed to ensure the quality trimming and adapter removal steps. Several changes should become in comparison with the previous QC report such as per-base quality scores should be different, the per-sequence quality distribution should be shifted to higher scores and sequencing adapters are not identified as over-represented [50]. We can observe changes mainly in the sector Sequence length distribution which confirm correct trimming. All sequences from all samples satisfy theoretical prerequisites.

4.4 Read alignment to a reference

The final step is the reads alignment to a reference. This process is an active field of research and novel aligners are frequently published. Unfortunately, there is no 'silver bullet' so the choice of aligners will be dependent on the reference used in [50]. The aligner is usually chosen according to the type of available reference. The usage of STAR [57] is recommended for the genome based alignment of RNA-Seq data. Using e.g. BWT FM-index [58] is recommended for alignment of RNA-Seq data to a reference transcriptome.

Figure 7 shows the summary of the result alignment of all samples from all replicates. These results are written in Attachment B]. The mean of all samples from all

replicates is 8,3 million uniquely mapped reads. The maximum is 19,6 million uniquely mapped reads and the minimum is 1,3 uniquely mapped reads from all datasets. The standard deviation of uniquely mapped reads is 4,8 million.

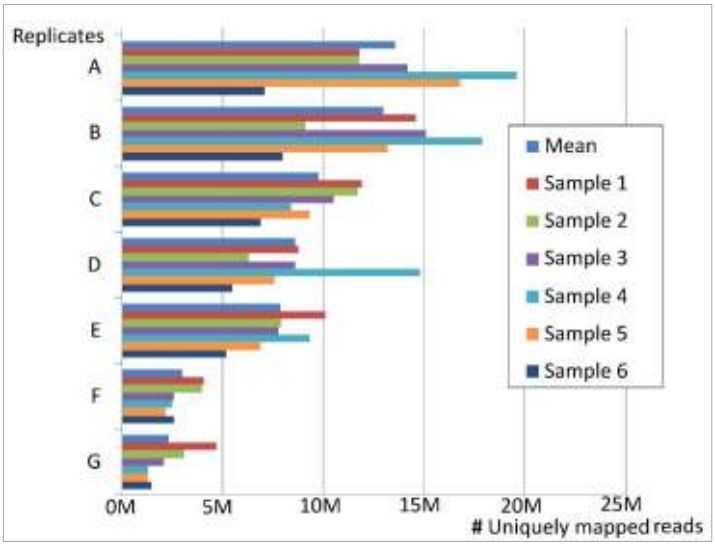


Figure 7: Summary alignment results of STAR tool

In this moment, the data using Integrative Genomics Viewer software (IGV) [59] is being checked and the replicates B, C, D, E, F and G are strand-specific RNA-Seq datasets are being revealed.

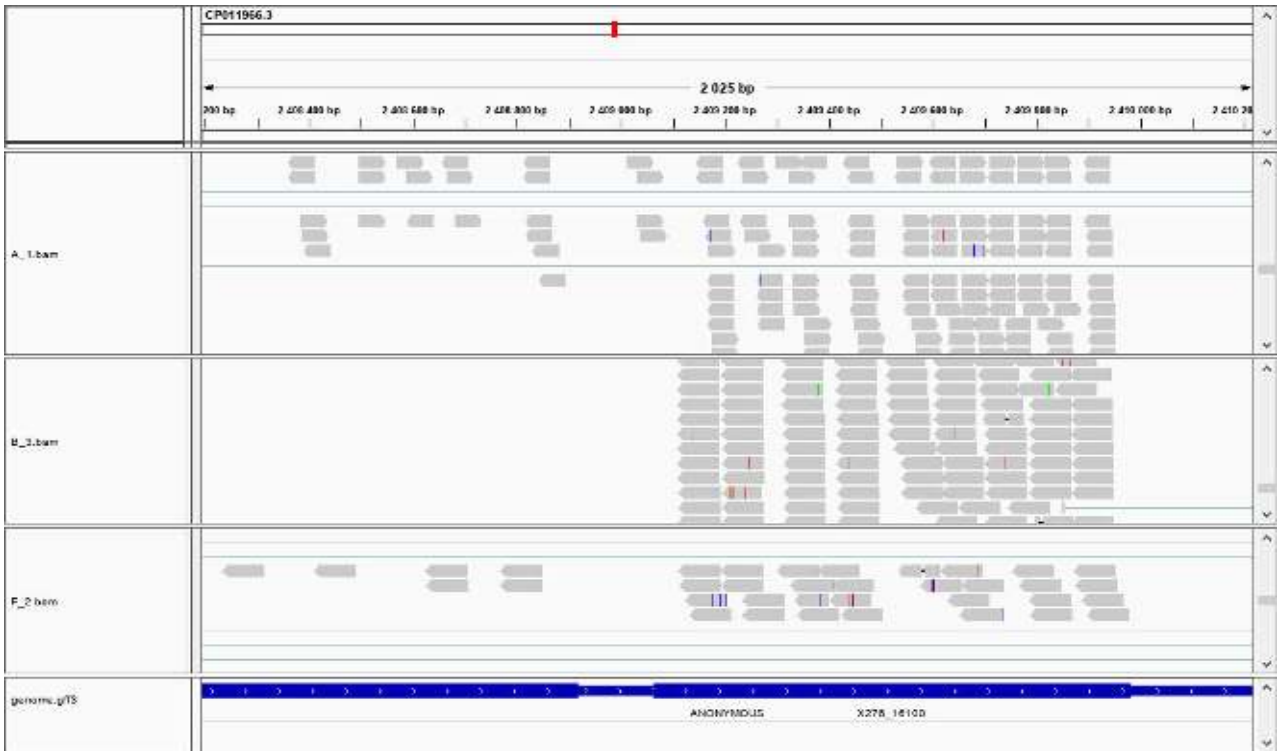


Figure 8: Example of samples A1, B3 and F2 visualization using IGV

In Figure 8 some specific samples – A1, B3 and F2 – and the visualization of these samples using IGV are chosen. We can see that in the section with replicates A there are reads which have random direction but in the section with replicates B or replicates F we can see that the direction of reads is exactly reverse as features in genome.gff3 – the last section in Figure 8. It means that during B, C, D, E, F and G sequencing the modifying of the RNA-Seq protocol known as stranded RNA-Seq was used how it is described in the chapter 2.4. In the end, it was decided to continue in the work without replicate A. Thus, the next reason for continuing the work without replicate A which is unstranded RNA-Seq dataset.

However, converting BAM files into count tables without stranded information cannot be considered as a false step. It is only under-utilization of the whole information so basic statistical analysis between non-stranded and reversely-strand obtained count tables is offered there. The evaluation differences between non-stranded and reversely stranded datasets is described using dimensionality reduction in the following chapter.

4.5 Normalization RNA-Seq

The step before normalization RNA-Seq data creates a count table from BAM files. The quantification approach is based on quantifying gene expression by the RNA-Seq count number of reads mapped to each gene. The approach to create a count table is based on R/Bioconductor featureCounts [52]. Two types of count tables are created there. The first type is based on the unique mapping reads which are counted only to one feature. The second approach is multimapping reads which is counted to standardized features.

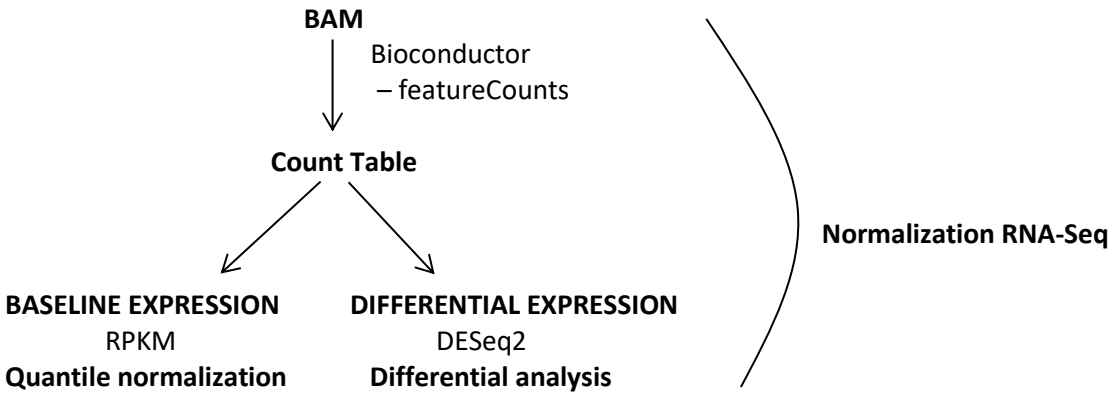


Figure 9: The second part of RNA-Seq pre-processing workflow [51]

Figure 9 shows the second part of RNA-Seq pre-processing workflow which is focused on normalization RNA-Seq data. The study by Schurch et. al. [60] describes the advantages and disadvantages among tools. This study recommends using edgeR or DESeq2 tool if you have fewer than 12 replicates.

RPKM normalization is typically used as a quantile normalization in the baseline expression. The tool R/edgeR rpkm [61] is used in this thesis. Currently, the RPKM is one of the most used methods for normalised data. This normalization is not perfect because it relies on the unlimited dynamic range of the RNA-Seq too much. It is not perfect for the comparison of the expression of one gene thought samples. However, RPKM is the most popular normalization and it is the reason why the RPKM is used. On the other side the decision of using this normalization in pre-processing data allows the possibility of future way for analysis RNA-Seq data from baseline expression view.

In the Table 1, there is an example of using normalization. We can see some samples of replicate A which were sequenced in time 8.5 *h*, 13 *h* and 18 *h*. On the left side of the Table 1 there are raw count data and on the right side there are RPKM normalised count data. Locus is the position of the described gene that is expressed. We used signification locus tag which is changeless and hence we can determine the properties of the expression of particular gene in the future correctly.

Table 1: The example using RPKM normalization

Locus tag	Replicates A					
	Raw count data			RPKM normalised data		
	T_3	T_4	T_5	T_3	T_4	T_5
X276_26735	54	103	52	40.60	57.59	35.03
X276_26730	58	117	76	43.61	65.42	51.19
X276_26725	49	98	47	36.84	54.80	31.66
X276_26720	83	88	44	11.01	8.68	5.23
X276_26715	1146	743	427	22.53	10.86	7.52
X276_26710	1322	1540	785	24.20	20.96	12.87

Another way is a differential expression analysis means taking the normalised count data and performing statistical analysis to discover quantitative changes in expression levels between experimental groups [51]. R/Bioconductor DESeq2 [62] can be called for differential expression analysis. This function prints a message for various steps, it performs such as the estimation of size factors, the estimation of dispersion values for each gene and fitting a generalized linear model [63].

DESeq2 function is based on the negative binomial as the reference distribution. The disadvantage of negative binomial distribution is the noise which is presented for

small-scale. In such cases, simpler methods are based on the Poisson distribution or empirical distributions although the absence of biological replication means no population inference and hence any *p* value calculation is invalid [52]. Fortunately, seven replicates $\{A, B, C, D, E, F, G\}$ are available for this research and other replicates are being prepared. Thus, DESeq2 function has been chosen for the analysis of our replicates.

4.6 Dimensionality reduction

In the fact our dataset includes two types of count tables of all replicates. One of them is based on the unique mapping reads that are counted only to one feature. These count tables are called FTFF. The second of them is multimapping reads which are counted to the standardized features. These types of count tables are called FTTT. In summary seven replicates exist which include six point-times samples and these seven replicates can be created by two approaches. Designation FTFF and FTTT was derived by the arguments which are different setting – True (T) or False (F) – between these two count tables, see in Table 2.

Table 2: Different setting arguments in R/Bioconductor featureCounts [52] between FTFF and FTTT.

Arguments	Description	Setting
useMetaFeatures	logical indicating whether the read summarization should be performed at the feature level or meta-feature level.	**F**alse
allowMultiOverlap	logical indicating if a read is allowed to be assigned to more than one feature (or meta-feature) if it is found to overlap with more than one feature.	**T**rue / **F**alse
countMultiMappingReads	logical indicating if multi-mapping reads should be counted.	**T**rue / **F**alse
Fraction	logical indicating if fractional counts will be produced for multi-mapping reads.	**T**rue / **F**alse

We have 5 276 protein coding genes that are signification as Locus Tag. Thus, there are 5 276 dimensionalities in our count tables. It is necessary to verify that pipeline which was used in the pre-processing RNA-Seq is correct. It is the reason for creating visualization which is based on the dimensionality reduction. Software packages that have functions to enable the visualization of the results such as DESeq2 in Bioconductor were

used. Considering the complexity of transcriptomes, the display of information is still a challenge. Thus, all of the tools are evolving rapidly and more comprehensive tools with desirable features can be expected in the future [52]. The global quality of the RNA-Seq dataset is essential to be assessed. It leads to checking the reproducibility among replicates. Reproducibility among technical replicates should be high but there are not any clear standards for biological replicates because they depend on the heterogeneity of the experimental system [52].

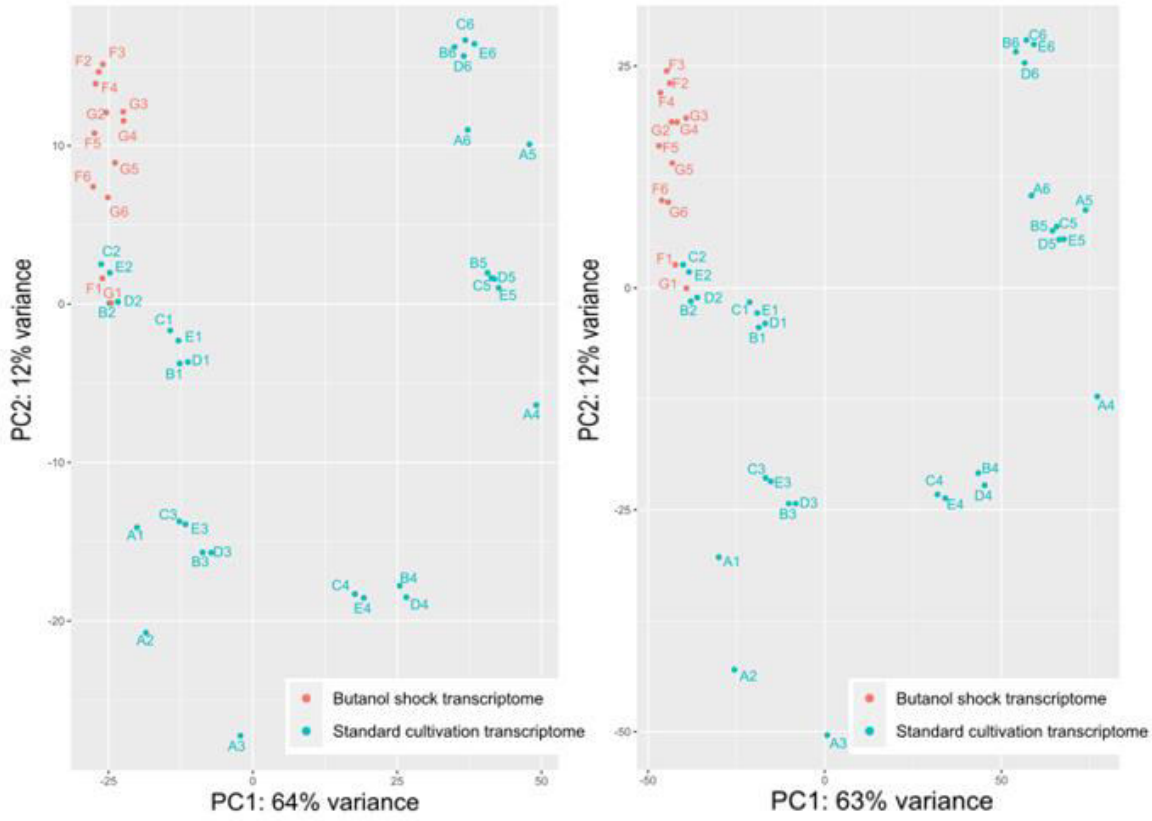


Figure 10: On the left there is visualization from FTFF non-stranded dataset using variance regularized transformation and on the right there is visualization from FTFF non-stranded dataset using variance stabilizing transformation

It is expected in the principal component analysis (PCA) that biological replicates of the same condition become clusters. It can be theoretical prerequisites for samples B2, C2, D2, E2, F1 and G1 because these samples are obtained from three different studies but same conditions, see in the chapters 2.5.1 and 2.5.2. PCA is used for the visualization of sample-to-sample distance [56]. The data points are projected onto the 2D plane. These two directions show most of the differences. The PC1 axis separates the data point the most and the PC2 axis represents the direction with the largest variance subjected to the constraint that must be orthogonal to the first direction [56].

In the Figure 10 we can see PCA plots which were created from FTFF dataset by all replicates. There is applied R/Bioconductor pcaMethods [64]. On the left there is a PCA plot in which variance regularized transformation was used, on the right we can see PCA plot where the approach with variance stabilizing transformation was used. We can see that using of regularized transformation has higher 1 % PC1 variance than using stabilizing transformation in this case. In the both of PCA plots we can see that B2, C2, D2, E2, F1 and G1 become a cluster. It can be considered a fulfilment theoretical prerequisite which was predicated. The samples from replicate A are shown on remote locations in the Figure 10. This fact is probably caused by different sequencing replicates. As it was mentioned above, the replicate A was sequenced using HiSeq method but other replicates were sequenced using NextSeq method. It is the reason why the work for analysis gene regulatory only with replicates B, C, D, E, F and G is going to be continued with.

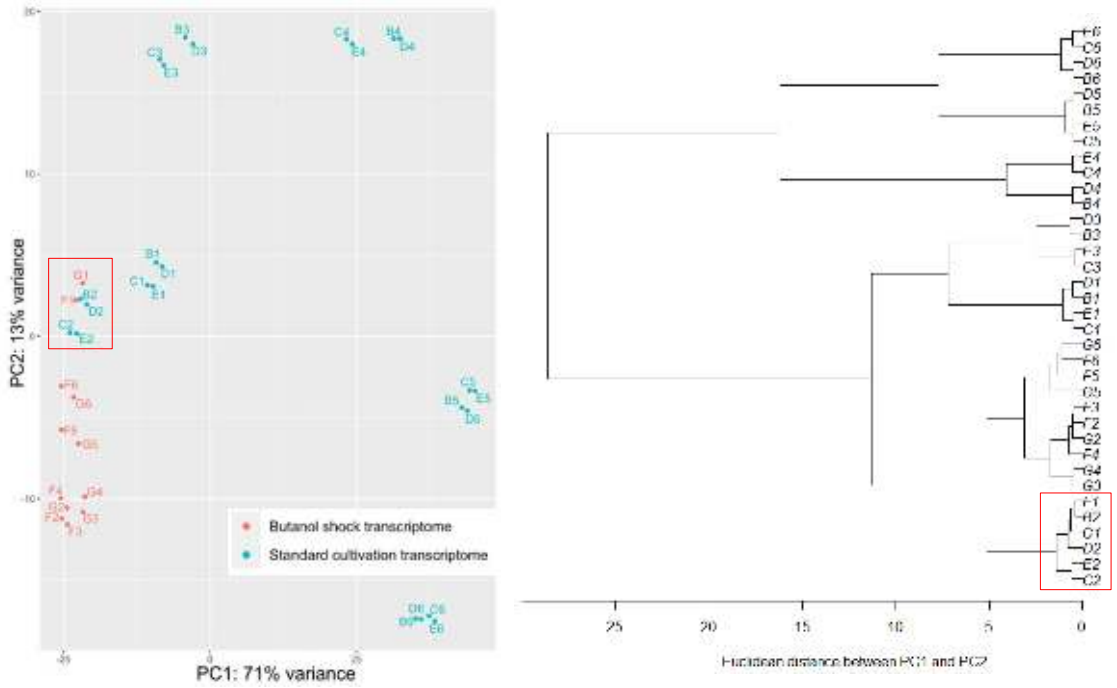


Figure 11: FTTT reversely stranded RNA-Seq dataset with used regularized transformation

Another utilization of PCA plots is shown differently between specific strand and non-strand dataset which is described by basic statistics using created PCA plots with their scree plots for all types of datasets which can be obtained. These plots are visualized in the Attachment C] and the Attachment D]. The scree plots confirm the correct using of PCA plots for our datasets because the first and the second principal components represent significantly larger variability than other principal components. Differences between unstranded and reversely stranded FTTT dataset are not huge. This fact shows

similarities in PCA plots in the Attachment C]. However, reversely stranded RNA-Seq count tables give more information than unstranded RNA-Seq count tables.

In the Attachment E] there are applications of UPGMA algorithm for evaluation PCA plots analysis. We used R/phangorn upgma [65] where the average from Euclidean distance between samples were applied. Our theoretical prerequisite that samples B2, C2, D2, E2, F1 and G1 become the cluster and it is confirmed in all obtained datasets. Figure 11 shows thanks to the red frame that our theoretical prerequisite has been confirmed. This prerequisite is confirmed by using the knowledge of Euclidean distance which in applied in the UPGMA method.

5 Gene regulation for *C. beijerinckii* NRRL B-598

Gene regulation for *C. beijerinckii* NRRL B-598 is derived only by reversely stranded FTFF and FTTT datasets. The omission of replicate A is justified in the chapter 4.6. It means that we have a great possibility to create two datasets which were sequenced from the same conditions and provide checking results.

5.1 Infer Operons dataset

Operon is the cluster of genes that have the same promoter and genes are transcribed and regulated as a single large mRNA including multiple structural genes as described in the chapter 1. We can see the operon in Figure 12. The operon structure [66] is one of features of prokaryotes.

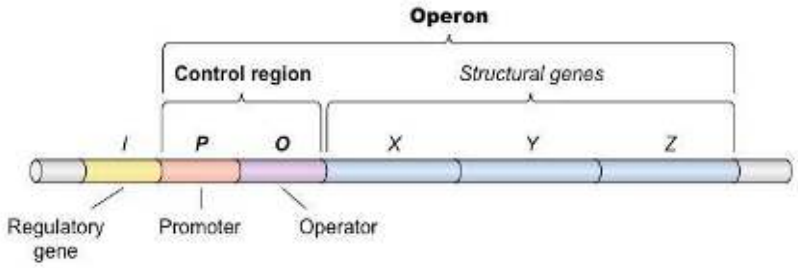


Figure 12: Operon structure [67]

Transcription unit (TU) [68] is a concept which was defined to make the understanding of operon function easier. TU is obtained from the genes that transcribe and regulate simultaneously. The identification of TUs is a challenge for resolving the understanding of the transcriptional regulation. TUs mapping can be used for the identification of new ruboswitches, non-coding RNA etc. The same TU can have multiples transcription start sites (TSSs) and transcription ends, alternative TSSs in bacteria are found for 15-60 % genes and operons [66]. We need to combine the information about transcription and translation with genomics data. There is a necessary complex where a new pipeline is created, see Figure 13, for prediction operons which is based on both pieces of information.

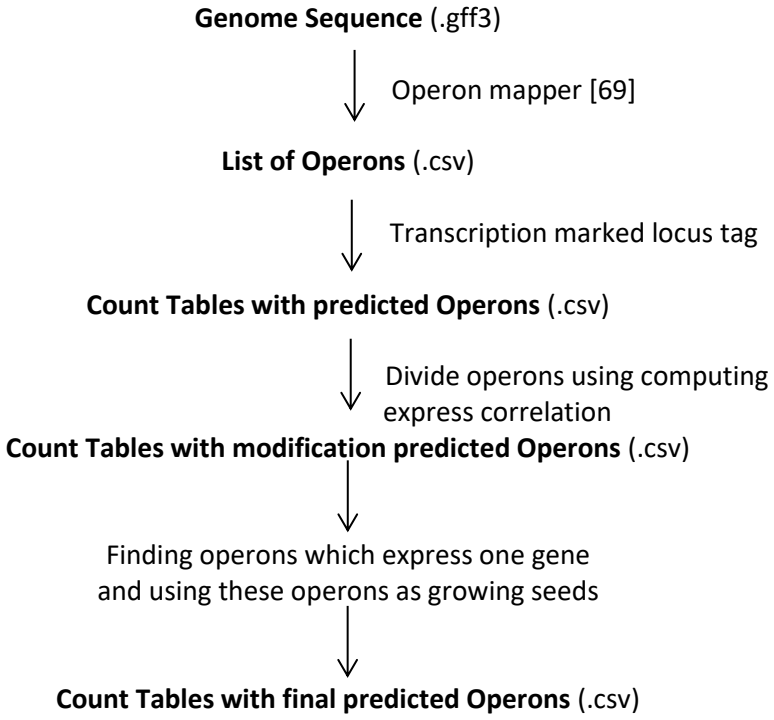


Figure 13: The processing workflow for obtaining operons

In this thesis Operon-mapper was used [69]. It is a web server that predicts the operons of any bacterial or archaeal genome sequence. This web server predicts operon using knowledges about intergenic distance of neighbouring genes as well as the functional relationships of their protein-coding products.

Operon-mapper is based on an artificial neural network (ANN). This algorithm was tested on a set of experimentally defined operons in *Escherichia coli* and *Bacillus subtilis* and reached accuracies of 94,6 % and 93,3 % [69]. ANN has inputs that are the intergenic distance of contiguous genes and score which thinks of the functional relationships between the protein products. Operon mapper predicts 3 357 operons in *C. beijerinckii* NRRL B-598. This prediction is based only on information that is obtained from genome format .gff3 number CP011966.3 [36]. After removing pseudogenes, tRNA and rRNA applied to the transcription mark using locus tag, 3 217 predicted operons, are obtained. However, we need a complex view where the transcription must be included which can be obtained from our created count tables. It was the reason for creating R scripts in the version 3.6.1.

This processing workflow for obtained operons was used to FTTT dataset which include B, C, D, E, F and G reversely stranded RNA-Seq replicates. This choice is based on prerequisite that using FTFF dataset can obtain more false positive results thanks to the loss of multimapping information. The first step is dividing predicted operons use to

computing correlation coefficient and dividing one predicted operon to two operons if the value of correlation coefficient is less than 30 %, we obtain 3249 operons.

Then we applied our created algorithm. It can be called growing operons seeds. The main idea is using precedent information predicted operons where operons are found that express only one gene and the express value from these operons are used to the next computing correlation between the express values from upstream predicted operons or downstream predicted operons in sequence. The choice of upstream predicted operons or downstream predicted operons depends on the higher value of correlation coefficient. If the correlation coefficient is more than 75 % and the distance between these operons is less than defined threshold in bp, we assume that it is one operon and these operons are concatenation. The threshold was defined as a mean distance between the predicted operons which is 374 *bp*. This process repeats until all correlation coefficients are more than 75 % or the distance between operons is higher than 5 000 *bp*. After the application of this algorithm, 2 737 operons are obtained. In Figure 14 there are the first five obtained operons that visualize using of the sunburst plot in MS Excel Office v. 16.

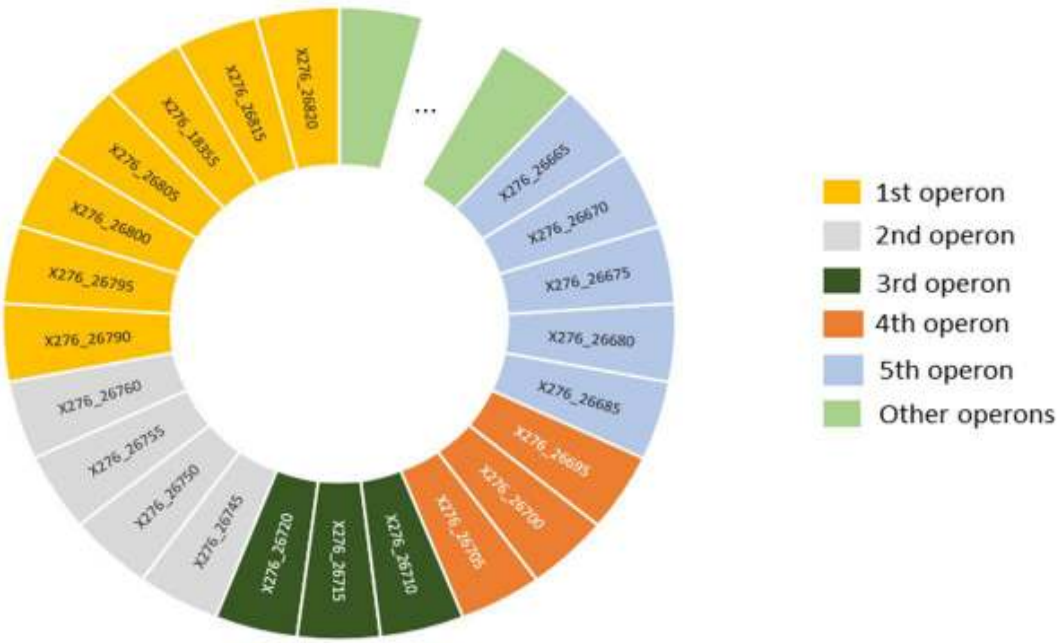


Figure 14: First five obtained operons, light green colour represents other operons

5.2 Gene co-expression network

Network inference methods have different approaches as it was mentioned in the chapter 3. This thesis is focus on the co-expression-based resources for infer GRN. Another type of network is shown importantly and is often used to describe gene regulatory, too. It is gene co-expression network (GCN) [70]. GCN are increasingly used to study the system level functionality of genes [70].

The concept of GCN construction is quite easy to understand. Nodes represent genes. Nodes are connected if the corresponding genes are significantly co-expressed across appropriately chosen tissues samples [70]. The main reason between GCN and GRN is the fact that GCN always obtains undirected edges. We choose several tools for the inference of GCNs and application to our datasets. Our data were divided to three datasets. The first dataset is BCDE dataset represented by standard cultivation transcriptome, see chapter 2.5.1. The second dataset is FG dataset represented by butanol shock transcriptome, see chapter 2.5.2 and final dataset is BCDEFG dataset which includes all these samples.

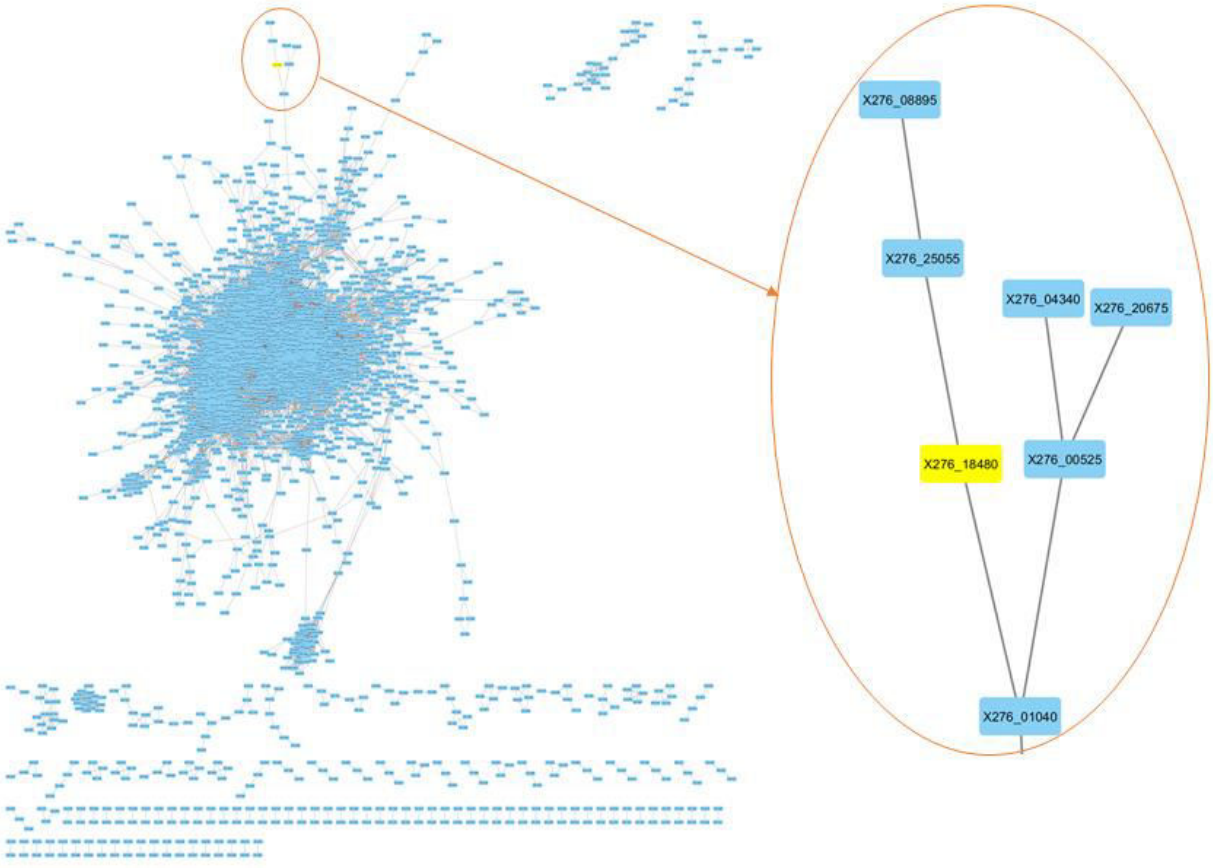


Figure 15: Example of gene co-expression network from RNA-Seq dataset containing gene expression profiles of 5276 genes from *C. beijerinckii* NRRL B-598

We have used R/Bioconductor CoexNet [71]. This package offers two different methods for reconstruction of co-expression networks. Pearson correlation coefficient and mutual information, these methods are described in the 3.5 chapter. The CoexNet package was applied to our datasets.

Figure 15 shows co-expression network which was created by R/Bioconductor CoexNet using MetaCentrum Virtual Organization portal and visualized by using Cytoscape. The computing was provided to input parameters which include a correlation method and threshold. We found the threshold using the function '*findThreshold*' which is included in the packages R/Bioconductor CoexNet. It finds the threshold value to establish the cut off to define the edges in the final co-expression network from two steps. Firstly, the threshold is obtained by the subtraction from clustering coefficient values of the real and random networks created from the possible threshold values in the correlation matrix. In the second step, a Kolmogorov-Smirnov test is used which has been made to evaluate the degree distribution respecting normality. In Figure 15 there is co-expression network which has threshold value 0.94. The network has 3 052 nodes and 82 634 edges. This network includes FTFF BCDE dataset, it is a standard cultivation transcriptome, see 2.5.1.

This locus tag X276_18480 indicates gene Spo0A which is known as a global regulator as mentioned in the study by Patakova [37]. Thus, the thesis is focused on this gene and this significant locus tag X276_18480 is visualized using the yellow colour. In the study by Sedlar [72] it is written that gene Spo0A is a sporulation initiator factor such as transcription factor for sol operon. However, the sporulation is not a necessary condition for solventogenesis and that sporulation can be achieved only under specific culture conditions [37]. Figure 15 shows that locus tag X276_18480 is adjacent to X276_25055 and X276_01040. Both of locus tags are protein coding. Specifically, X276_25055 codes A0A0K2MKB7 protein [73] whose length is 270 amino acids and X276_01040 is glgD gene which codes A0A0K2M7P6 protein [74].

Attachment F] shows examples of two co-express networks where locus tag X276_18480 is significant. Both of networks are created by the same parameters but the different dataset. One is from FTFF datasets and the other one is from FTTT datasets. This distribution of datasets was provided for checking and we can see that adjacent to X276_25055 and X276_01040 was correct because both of datasets have match.

Table 3: Basic static analysis co-expression networks based on correlation method

	BCDE		FG		BCDEFG	
	FTFF	FTTT	FTFF	FTTT	FTFF	FTTT
Number of nodes	3 052	3 143	5 249	2 746	5 013	2 233
Number of edges	82 634	85 603	478 451	20 109	1 637 789	35 873
Clustering coefficient	0.503	0.507	0.433	0.353	0.605	0.525
Connected components	114	134	2	131	9	168
Network density	0.018	0.017	0.035	0.005	0.130	0.014
Characteristic path length	4.592	4.639	3.156	6.418	2.370	4.853
Average no. of neighbours	54.151	54.472	181.302	14.646	653.417	32.130
Network heterogeneity	1.197	1.229	0.950	1.264	0.983	1.356

Table 4: Basic static analysis co-expression networks based on mutual information method

	BCDE		FG		BCDEFG	
	FTFF	FTTT	FTFF	FTTT	FTFF	FTTT
Number of nodes	2 543	2 622	1 141	1 171	711	766
Number of edges	713 312	731 518	137 993	148 509	12 019	12 261
Clustering coefficient	0.628	0.622	0.551	0.551	0.671	0.643
Connected components	4	8	4	8	49	53
Network density	0.220	0.213	0.212	0.217	0.048	0.042
Characteristic path length	1.875	1.885	1.797	1.792	3.540	3.350
Average no. of neighbours	560.339	557.985	241.881	253.654	33.809	32.013
Network heterogeneity	0.703	0.720	0.468	0.472	1.348	1.323

In sum we created 12 different co-expression networks. The co-expression networks were created by correlation and mutual information method, see in the chapter 3.5. Static network analysis was provided using NetworkAnalyzer [75]. The conclusion of results are shown in the Table 3 and Table 4. Table 3 shows basic static analysis from all co-expression networks which were created by correlation method and Table 4 shows the conclusion of results which has been created by mutual information method. Network parameters obtained by FTFF datasets and obtained by FTTT datasets are very similar. It points out to checking correction because we can assume similar results of these datasets created by same transcriptome data.

Number of nodes and edges are affected by threshold value which was obtained using '*findThreshold*'. The threshold values which are found by FTFF datasets are same as the threshold values which are found by FTTT. However calculated threshold values were different between correlation and mutual information approaches. It is reason why we can see huge difference between the number of nodes or edges between these different approaches. If we used threshold value such as 0.94, the final co-express networks have 3 052 nodes and 82 634 edges or 3 142 nodes and 85 603 edges, these networks were

created by BCDE datasets using correlation approach, see Table 3. On the other side, if we used threshold value such as 0.37, the final co-express networks have 1 141 nodes and 137 993 edges or 1 171 nodes and 148 509 edges, these networks are created by FG datasets using mutual information approach, see Table 4. Thus, when lower threshold value is used, the graph has more number of edges.

Clustering coefficient is defined in the chapter 1.1. Thus, the network clustering coefficient is the average of the clustering coefficients for all nodes in the network [76]. The premised is that nodes with less than two neighbours have a clustering coefficient of 0. The mean of clustering coefficient is higher in mutual information approach than correlation approach because the mean of clustering coefficients from correlation approach is 0.49 and the mean of clustering coefficients from mutual information approach is 0.61.

The most distinct difference between results of static analysis between correlation and mutual information methods is visible in the results of connected components. Number of connected components [76] indicates the connectivity of a network. It means that the lower number of connected components suggests a stronger connectivity. The mutual information approach has significant values of connected components then the correlation approach. Thus, networks which were created by mutual information approach has stronger connectivity.

The network density [76] is normalized version of the average number of neighbours. The average number of neighbours [76] indicates the average connectivity of a node in the network. The network density shows how densely the network is populated with edges. The mean of network density is higher using mutual information method. In opposite the mean of characteristic path length is higher using correlation approach. The characteristic path length [76] is also known as the average shortest path and gives the expected distance between two connected nodes.

The network heterogeneity [76] reflects the tendency of a network to contain hub nodes. The mean of network heterogeneity from the Table 3 is 1.16 and the mean of network heterogeneity from the Table 4 is 0.84. It means that co-expression networks using correlation approach have higher network heterogeneity than co-expression networks using mutual information method.

Final step in reconstruction co-express network for *Clostridium beijerinckii* NRRL B-598 is based on intersection above created co-express networks. The merging is visualization in Figure 16. This merging was made for all FTTT and FTFF co-

expression networks which was created by mutual information or correlation. The final results of co-express network are four co-expression networks.

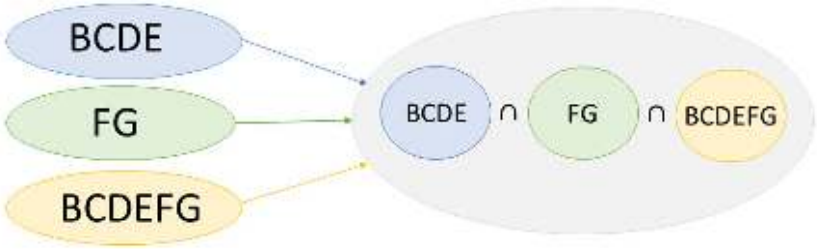


Figure 16: The merging co-expression networks

Table 5: Basic static analysis final co-expression networks

	Correlation		Mutual information	
	FTFF	FTTT	FTFF	FTTT
Number of nodes	3 039	1 430	139	144
Number of edges	10 543	1 516	102	98
Clustering coefficient	0.240	0.213	0.140	0.126
Connected components	1 316	743	63	70
Network density	0.002	0.001	0.011	0.010
Characteristic path length	6.648	5.667	3.892	4.481
Average no. of neighbours	6.938	2.120	1.468	1.361
Network heterogeneity	1.987	1.682	1.035	1.134

Table 5 shows final results of co-express networks which were created by RNA-Seq from *C. beijerinckii* NRRL B-598. We created four co-expression networks for *C. beijerinckii* NRRL B-598. Co-expression networks from FTTT and FTFF datasets were created for the verification of correct results because we assume that networks which have been created by the same method will be similar. The values of parameters network by FTFF and FTTT datasets in the Table 5 can be considered a fulfilment prerequisites.

The final comparison between correlation and mutual information shows that mutual information approach is stricter than correlation because co-expression networks which were created by mutual information method has significantly less number of nodes and edges than co-expression networks which were created by correlation. However, the co-expression networks based on mutual information has higher values of network density. Despite less number of nodes and edges the co-expression networks based on mutual information is higher populated with edges.

Unfortunately, co-expression networks based on mutual information lose information about genes which has not higher express value in samples such as Spo0A gene. Thus, after the final step merging information about this gene is lost in co-

expression networks based on mutual information. However, merging co-expression networks based on correlation preserved information about Spo0A gene so we can declare the dependence between Spo0A and glgD gene.

5.3 Gene regulatory network

The co-expression-based resources for infer GRN is used in this thesis. We showed GCN in the previous chapter where is mentioned that GCN is always an undirected graph. It is different between GCN and GRN. GRN is mentioned as a directed graph but co-expression based resources tools for infer GRN are insufficient for the infer directed graph.

Co-expression-based tools [77] have been widely adopted after the introduction of transcriptome scale quantification methods of transcript abundances. Two genes are deemed co-expressed if a significant dependency is determined between their transcript expression [77]. Currently, several available tools were created for the reverse engineering of GRNs based on this approach.

5.3.1 Bootstrap-based GRN

The bootstrapping approach is described in the chapter 3.1. There is used R/Bioconductor bc3net [78]. The main idea of bc3net is based on the bootstrap aggregation. Bc3net is an ensemble method that is based on bagging the C3NET algorithm, which means it corresponds to a Bayesian approach with non-informative priors [78].

Bc3net was applied to our datasets such as in the previous chapter 5.2. Unfortunately, bc3net is one of their tools which are not sufficient for the infer directed graph. Thus, all GRN which is created to using bc3net are undirected. We provided static analysis NetworkAnalyzer, see Table 6. Table 6 shows really similar network parameters. It means that all created GRNs based on the bootstrap report similar dependence between genes and this fact is not dependent to different obtaining datasets.

Table 6: Basic static analysis GRNs based on bootstrapping

	BCDE		FG		BCDEFG	
	FTFF	FTTT	FTFF	FTTT	FTFF	FTTT
Numbers of nodes	5276	5276	5 282	5 282	5 278	5 278
Numbers of edges	31 514	30 794	31 436	30 364	30 409	30 483
Clustering coefficient	0.165	0.170	0.082	0.083	0.161	0.168
Connected components	1	3	7	7	4	4
Network density	0.002	0.002	0.002	0.002	0.002	0.002
Characteristic path length	4.714	4.893	4.163	4.200	4.666	4.783
Average no. of neighbours	11.946	11.673	11.903	11.497	11.523	11.551
Network heterogeneity	0.343	0.361	0.325	0.310	0.393	0.403

GRNs were visualised using Cytoscape where we created example of the first neighbourhood of Spo0A, see Attachment H]. The Attachment H] shows that the connection between X276_01040 and X276_18480 is conformed in all first neighbourhood of Spo0A. Thanks to it fulfilment dependence between Spo0A and glgD gene, again.

The final GRNs which were created such as intersection above networks, the same principle is shown in Figure 16. It means we obtained two resulting networks. The first checking of results shows the similar networks parameters of FTFF datasets and FTTT datasets in the Table 7. The number of edges is 617 for FTFF datasets and 644 for FTTT datasets. If we compare edges in these datasets, we obtain 412 edges which are the same in the both final networks. It means we declare 412 edges which report dependence in GRNs based on the bootstrap approach.

Table 7: Basic static analysis for two final GRNs based on bootstrapping

	FTFF	FTTT
Numbers of nodes	5276	5276
Numbers of edges	617	644
Clustering coefficient	0.011	0.014
Connected components	4 698	4 680
Network density	0	0
Characteristic path length	2.058	1.643
Average no. of neighbours	0.234	0.244
Network heterogeneity	2.469	2.392

Table 7 shows that network parameters are different such as network parameter in the Table 6. The value of clustering coefficient is less in the Table 7 then in the Table 6 such as the value of average no. of neighbours, characteristic path length and network density. The value of network density is zero. It causes the high number of gene which is not connected with other genes after execution intersection. In the opposite value of

network heterogeneity is higher in the Table 7 then in the Table 6. It reports that the final networks which have been created by intersection have a higher tendency of the network to contain hub nodes.

We have focused on Spo0A in the other view of static analysis in GRNs which have been created by bootstrapped approach. In total is 412 edges which we have declared such as dependences among genes because these edges are the same in final FTFF and FTTT networks, see Table 7. Spo0A has been found and visualized by its neighbours. Unfortunately, the neighbourhoods have obtained only two genes but these genes are X276_01040 and X276_18480. The dependence between Spo0A and glgD gene are fulfilment, again.

5.3.2 Tree-based GRN

The decision tree-based method is described in the chapter 3.3. R/Bioconductor GENIE3 [79] has been applied to our datasets which was divided such as in the previous chapters. The GENIE3 is a decision tree-based method which has emerged as the best performer in the DREAM4 [12]. This inference method based on the variable selection of ensembled regression trees. It produces directed graphs of regulatory interactions allowing the presence of feedback loops in the network, it obtains realistic GRNs [77].

The GRNs which has been created by GENIE3 has had to compute using the MetaCentrum Virtual Organization portal. We have had to filter GRNs using empirically set threshold. This value has been set for a purpose obtained by GRNs which include about 5 000 edges because the GENIE3 output is adjacency list and this file format is computationally demanding for the visualization and following static analysis. The threshold value for samples from the standard cultivation transcriptome is 0.0095, the threshold value for samples from butanol shock transcriptome is 0.0074 and the threshold value for all samples is set as 0.0108. If there are more variable and larger dataset, we need to set higher threshold.

In sum 6 GRNs have been created whose approach has been based on the tree methods. The basic static analysis has been provided by NetworkAnalyzer and visualized in the Table 8. There is the same situation such as above where the first checking of results shows the similar networks parameters of FTFF datasets and FTTT datasets in the Table 8.

Networks parameters in the Table 8 is provided static analysis of GRN which is specific for directed graph. The difference between Table 6 and Table 8 is important. Number of nodes and edges in the Table 6 is much higher than in the Table 8. This fact is caused by a higher computational complexity during visualized directed networks than during the usage of undirected networks.

Table 8: Basic static analysis GRNs based on tree method

	BCDE		FG		BCDEFG	
	FTFF	FTTT	FTFF	FTTT	FTFF	FTTT
Numbers of nodes	2 404	2 490	2 854	2 758	2 693	2 756
Numbers of edges	5 063	4 879	5 070	3 709	5 025	4 814
Clustering coefficient	0.063	0.061	0.032	0.027	0.064	0.069
Connected components	147	139	209	166	251	291
Characteristic path length	11.593	13.421	12.904	7.616	8.301	7.738
Average no. of neighbours	3.831	3.583	3.180	2.613	3.286	3.091
Multi-edge node pairs	458	418	532	105	601	554

The values of clustering coefficient is less in the Table 8 than in the Table 6. However, it is important to mentioned that the clustering coefficient is defined in directed networks [76]:

$$C_n = \frac{e_n}{(k_n(k_n - 1))}, \tag{14}$$

where k_n is the number of neighbours of n and e_n is the number of connected pairs between all neighbours of n. The values of connected components are much higher in the directed networks, see Table 8 than in the undirected networks, see Table 6. Table 8 also shows network parameter which is typical for directed networks, it is multi-edge node pairs. This attribute indicates if n is a partner of node pairs with multiple edges [76]. The highest value of the multi-edge node pairs parameter belongs GRN which have been created by all samples from all replicates. It is caused to large datasets which creates GRN.

Figure 17 shows the example of GRN which has been created by the standard cultivation transcriptome obtained such as FTFF datasets. We visualized significant locus tag X276_18480 which represents Spo0A. GRN created by tree-based approach shows Spo0A with 7 indegree and 3 outdegree parameters. Unfortunately, the visualized neighbours do not show X276_01040 like in the undirected graphs. However, we can see others 8 possibilities of candidate for interaction with Spo0A gene. The description of these 8 candidates is in the Table 9.

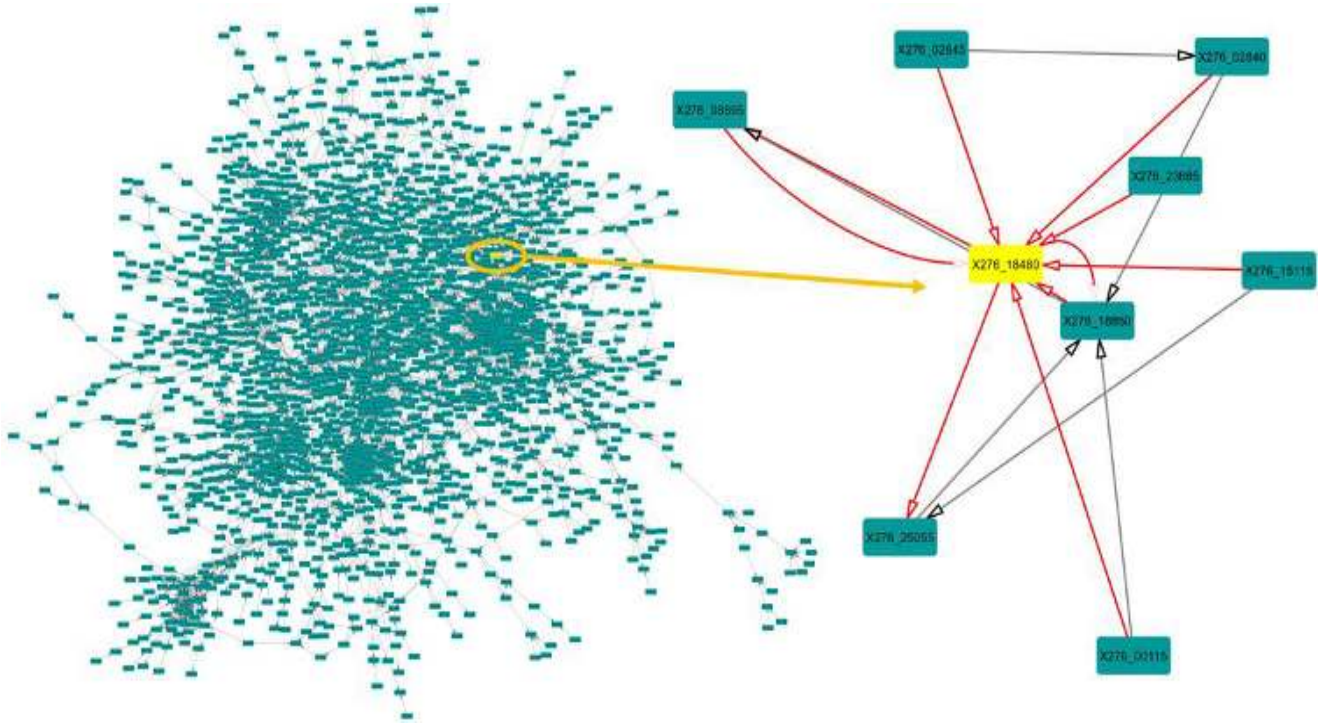


Figure 17: Example of GRN from RNA-Seq dataset containing gene expression profiles of 5276 genes from *C. beijerinckii* NRRL B-598

Table 9: Basic description of 8 candidates for interaction with Spo0A

Locus tag	Protein ID	Product
X276_02845	ALB44287.1	dCTP deaminase
X276_02840	ALB44286.1	hypothetical protein
X276_23685	ALB48059.1	Nitrogenase
X276_18850	ALB47156.1	alpha/beta hydrolase
X276_25055	ALB48296.1	lytic transglycosylase domain-containing protein
X276_00115	ALB43814.1	MazG-like family protein
X276_15115	ALB46480.1	cobyrinate a%2Cc-diamide synthase
X276_08895	ALB45387.2	class D beta-lactamase

Locus tags X276_02845 and X276_02840 are part of operon no. 1911 which is inference in chapter the 5.1. X276_02845 codes deoxycytidine triphosphate deaminase (dCTP deaminase) [80] product. It is an enzyme which is involved in the nucleotide metabolism. It catalyses the formation of deoxyuridine triphosphate (dUTP) which is turn in degraded by dUTPase to produce deoxyuridine monophosphate (dUMP). Dump is the immediate precursor of thymidine nucleotides [80]. In opposite the X276_02840 has the product which is described as a hypothetical protein. It means that the protein has been predicted nevertheless there is a lack of experimental evidence which is expressed in vivo. The products from other candidates for interaction with Spo0A are situated in the Table 9. Table 9 shows the description from CP011966.3 [36].

Attachment H] represents other examples of GRN which have been created by tree-based methods where visualized significant locus tag X276_18480 which represents Spo0A. In Attachment H] there is missed GRNs created by butanol shock transcriptome of FTFF datasets because in this GRN there is not visible locus tag X276_18480. It does not mean that these GRNs do not obtain the node which represents Spo0A but thanks to the usage of threshold we have persevered the nodes that have edges with higher weight.

However, in other GRNs which have been created by the tree-based approach X276_18480 is visible, see Attachment H]. In addition X276_18480 has common edge with X276_01040 in GRNs which have been created by all samples. This fact can show the correct prediction of dependence correlation between Spo0A and glgD genes.

Table 10: Basic static analysis two final GRNs based on random forest method

	FTFF	FTTT
Numbers of nodes	739	718
Numbers of edges	160	6
Clustering coefficient	0.011	0
Connected components	711	712
Characteristic path length	1.701	1.143
Average no. of neighbours	0.349	0.017
Multi-edge node pairs	31	0

The final GRNs which have been created such as the intersection above networks, again – see in the Figure 16. It means we have obtained two resulting networks. Table 10 shows static analysis for these GRNs. The first sight is obvious huge difference between the numbers of edges. GRN from FTTT datasets obtains only 6 edges and these edges are mismatched with edges in GRN from FTFF datasets. Thus, we can declare that we need more samples for inference tree-based GRN which can be causality.

Eventually, a selected part of GRN constructed as an intersection of FTFF datasets in shown the Figure 18. We highlighted locus X276_12610 which is a possible candidate for transcription factor coding protein ALB46023.1 annotated as transposase. Transposase [81] is the enzyme that cuts out the DNA and moves it to a different place. The predicted regulon of X276_12610 contains six genes: X276_04705, X276_19385, X276_00985, X276_10490, X276_10945, and X276_00235.

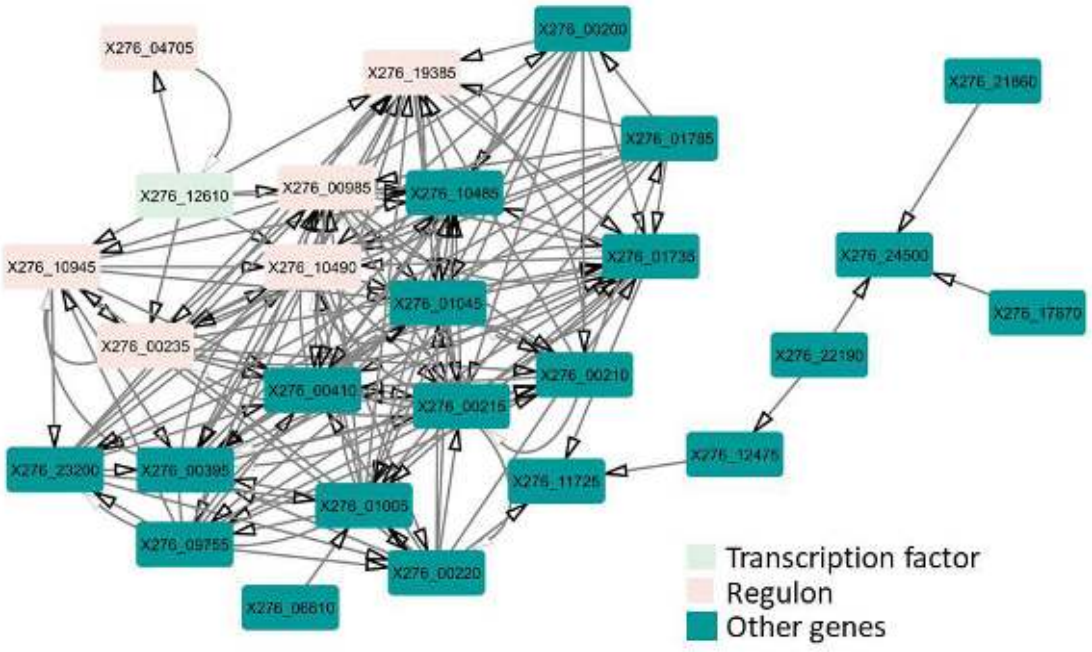


Figure 18: Sub-graph of intersection of tree-based GRN from FTFF datasets

Figure 18 shows example which have been analysed using CountTriplets [82]. It is the Cytoscape app that counts triplets motifs and computes z-scores against randomly generated networks. The CountTriplets performs abundance and significance analysis [82]. The triplets motifs are visualized in Attachment I]. The z-score [83] is defined as:

$$z(g_k) = \frac{f_{input} - \overline{f_{random}}}{\sqrt{\sigma^2_{random}}}, \tag{15}$$

where f_{input} is the sub-graph frequency in the input network $\overline{f_{random}}$ is the mean of frequencies of g_k in the random network. g_k represents sub-graph with a size k. Thus the z-score is the difference of f_{input} and $\overline{f_{random}}$ divided by the standard deviation σ^2_{random}. A motif is regarded as statistically significant if the associated z-score value is higher than 2.

The sub-graph in the Figure 18 has 27 nodes and 158 edges and the analysis of CountTriplets is visualized using triplets profile in Figure 19. It shows significance profile. The significance profile reports the profile of the network analysed as a line chart [82]. The significance profile has been computes against randomly generated sub-graph, see Figure 18. Figure 19 shows that the sub-graph in Figure 18 obtains 11 significant motifs because 11 motifs have higher z-score value than 2. FLLAAA was found among significant motifs, see in Attachment I]. The FFLAAA represents general Coherent Feedforward Loop (C1-FFL) motif which is found much

more frequently in the transcription network of *Escherichia coli* and yeast than the other types of motif [83].

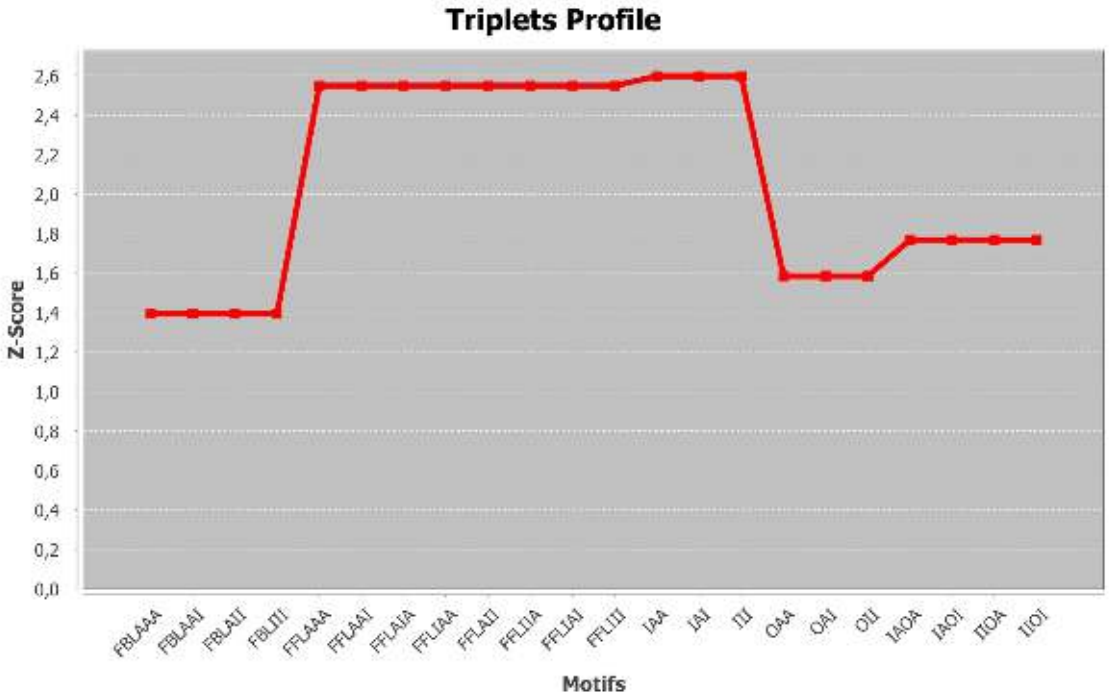


Figure 19: Triplets profile of example in Figure 18

5.3.3 Differential equation-based GRN

Differential equation-based GRN is the approach which is based on the ordinary differential equation (ODE) [77]. ODEs are learned from gene expression data, from multiple samples in GRN reconstruction. This approach is naturally suited to the model also non-linear relationships, because ODE methods are essentially RNA chemical reactions that can show a wide range of kinetic behaviours. Introducing the constraints of known kinetic parameters knowledge of GRN structure can be extremely beneficial to ODE-based methods [77].

In this thesis we used Time Series Network Identification (TSNI) algorithm which is a differential equations-based GRNs inference method and is available as MATLAB package [84]. The aim of this algorithm is to infer the local network of gene-gene interaction surrounding a gene of interest by measuring at multiple time points [77]. This package has been incorporated and used as a modification to our datasets. We have created *'GRN_main.m'* script which has included several steps.

The first step is the loading our datasets but in this infer GRN our dataset has been divided to six datasets which represents each of replicates. It was necessary because

inputs of *'tsni.m'* [84] requires chronologically arranged data. Thus, we have obtained six gene regulatory networks for each of replicates. In the following step we have created *'penetration.m'* function which has been applied to these six gene regulatory to obtain one final GRN in the main script. The procedure of penetration is visualization in the Figure 20. Figure 20 shows standard cultivation transcriptome as blue colour and butanol shock transcriptome as green colour.

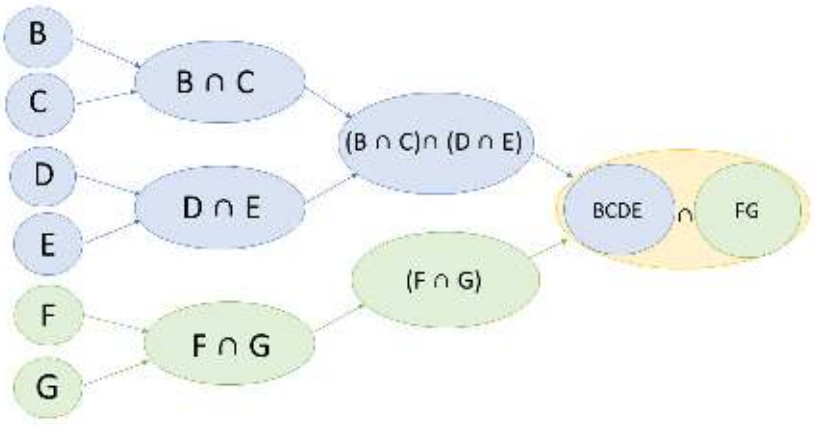


Figure 20: Diagram of penetration procedure for creation GRN

The evaluation of created GRNs has been provided to FTFF and FTTT datasets, see Table 11. The visualization GRNs has been used by Cytoscape and aMatReader [85]. The results have been obtained by using NetworkAnalyzer for basic static network analysis. We can see that results obtained by FTFF datasets that are similar to the FTTT datasets, again and so it is checking correction results.

Table 11: Basic static analysis GRNs based on differential equation

	FTFF	FTTT
Numbers of nodes	5 276	5 276
Numbers of edges	194 541	195 489
Clustering coefficient	0.602	0.566
Connected components	1 013	1 117
Characteristic path length	2.530	2.552
Average no. of neighbours	61.217	61.415
Multi-edge node pairs	27 774	28 200

We can see that we have the same numbers of nodes such as GRNs which is based on bootstrapping because the output of our *'tsni'* toolbox is weighted by edge matrix. Thus, we do not need high threshold like in the tree-based approach. There is a fact that we have 194 541 numbers of edges for FTFF datasets and 195 489 numbers of edges for FTTT datasets obtained by the highest values of numbers of edges from all used

approaches. However, these values have included edges which have reflected express influence themselves gene. It means that if we hide these themselves expressed edges we obtain about 190 000 edges.

Table 11 shows the parameters of directed GRNs. Thus the evaluation is provided between the Table 10 where parameters of tree based GRN are shown and the Table 11. The clustering of coefficient is significantly higher in the Table 11 than in the Table 10. The values of connected components and other parameters in the Table 11 are higher than parameters in the Table 10, too. This fact is affected by the numbers of nodes and edges. The ODE based networks are not computationally intensive and so the ODE based networks describe more gene regulatory information than tree based GRNs. The edges of ODE based networks are identical in 170 119 cases.

In this case it can be confused to locus tag X276_18480, again. In the Figure 21 there is X276_18480 such as a transcription factor with its first neighbourhood. We can predict that Spo0A is a transcription factor for these visible 27 genes including 5 operons which is obtained in the chapter 5.1.

Unfortunately, no genes which are connected above in others approaches are visible there. It is caused by a little number of time-points samples included only 6 time-points. Thus '*tsni*' method creates 6 initial gene regulatory networks which are followingly connected by a penetration, see Figure 20. If we have more samples sequenced in more time-points, the results of GRNs would be more causality. However, we can predict that Spo0A is a transcription factor which affects more operons and genes than only the sporulation sol operon such as written in the study by Sedlar [72].

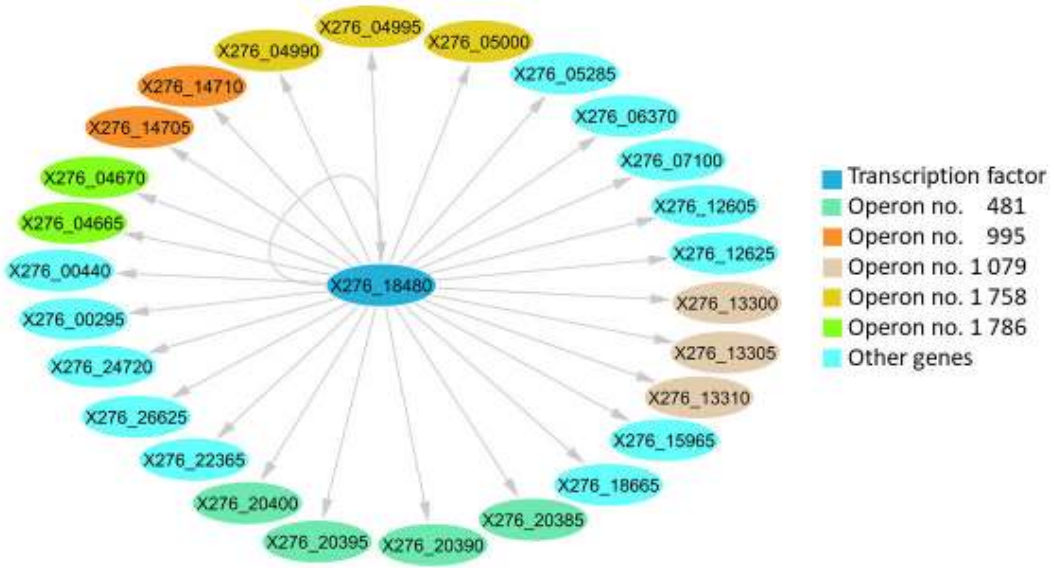


Figure 21: Part of GRN included locus tag X276_18480 with its 1st neighbourhood

Operon no. 1 786 includes X276_04665 and X276_04670 locus tags, the X276_04665 coding ALB44621.1 protein which produces glutamate synthase large subunit and X276_04670 coding ALB44622.1 which produces glutamate synthase subunit beta. As we see the similar products, we can adduce that the operon no. 1786 is correct inference. Figure 6 shows genes from operon no. 481 which is the only part of this operon because the predicted operon no. 481 includes X276_20380 locus tag, too. Its absence can cause small numbers of time-points sequencing, such as mentioned above.

In the next chapter we will predict the finial GRNs using the interconnection of all above mentioned approaches, we need to create ODE based networks which include about 10 000 edges. It means the 5 276 edges represent themselves express influence gene and other edges represent express influence among the different genes. We can see the basic static parameters of network in the Table 12. There are fewer parameters than in the Table 11. These networks are more strict than previous networks, see Table 11, because they are filtered by visible edges with higher weight value. The intersection of these networks obtains 8 785 edges which are identical.

Table 12: Basic static analysis filtered GRNs based on differential equation

	FTFF	FTTT
Numbers of nodes	5 276	5 276
Numbers of edges	9 639	9 332
Clustering coefficient	0.078	0.067
Connected components	4 566	4 645
Characteristic path length	3.291	3.199
Average no. of neighbours	1.460	1.326
Multi-edge node pairs	512	559

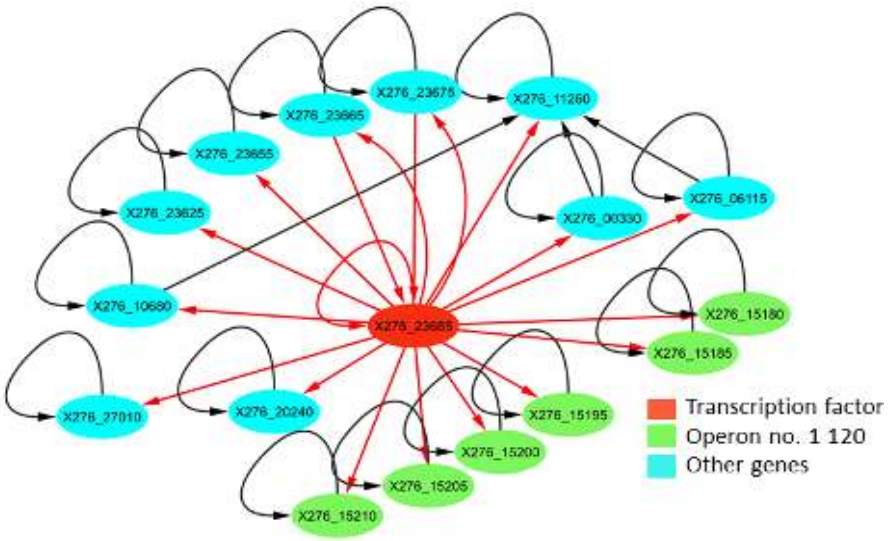


Figure 22: Sub-graph of intersection final ODE based networks

In Figure 22 there is a sub-graph of the intersection of final ODE based networks which shows the locus tag X276_23685 and its first neighbourhood. This locus tag codes protein ALB48059.1 and its product is nitrogenase. The locus tag X276_23685 is a possible candidate for transcription factor and its regulon is shown in Figure 22. The sub-graph in Figure 22 has 17 nodes and 21 edges. This network was analysed using CountTriplets similarly to the network in Figure 18. The significance profile has been computes against randomly generated sub-graph, see Figure 22. Figure 23 shows that the example of network in Figure 22 contains 10 significant motifs.

The most significant motifs are marked as IAA, IAI and III. In Attachment I] there are these motifs in part of linear triplets. The higher values of z-score for linear triplets motifs than closed triplets motifs have been caused by network architecture. It is caused by special example of sub-graph which has been created to visualize regulon of X276_23685. When we analysed the whole neighbourhood of X276_23685, we obtained 11 significant motifs, among them, also FFLAAA motif was detected, similarly to the sub-graph in Figure 18. It is caused by the size of network because the neighbourhood of X276_23685 contained 30 nodes and 143 edges. Here, triplets profile only for a selected sub-graph representing a possible regulon of X276_23685 is shown, see Figure 22. However, it is necessary to mention that the most significant motifs such as IAA, IAI and III have been found with the highest values of z-score in a sub-graph for neighbourhood of X276_23685, too.

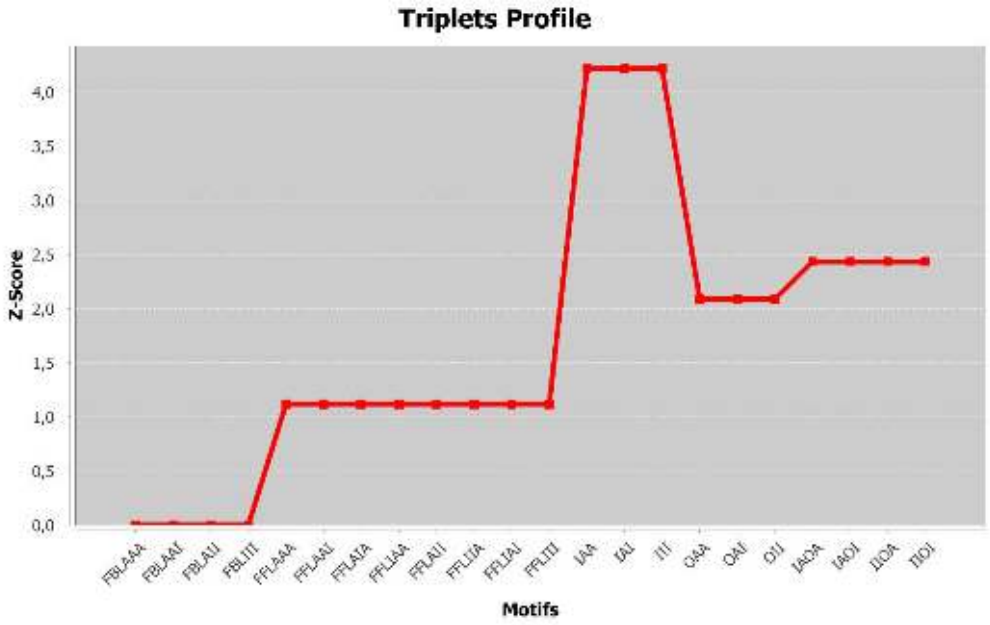


Figure 23: Triplets profile of example in Figure 22

5.3.4 Interconnection of approaches

The inference causality GRN for the non-model organism is important holistic approach and so we decided to use the interconnection of all used approaches in this thesis. Thus, the final results of GRNs are obtained by the interconnection of approaches whose description is described above. We have provided the union networks for all created networks from FTFF and FTTT datasets. This procedure is based on the merging GRNs based on bootstrapping, tree and ODE approaches.

The created final networks are undirected because the bootstrapping based networks are undirected and after applying merging to our GRNs in the Cytoscape the directed information is lost. Table 13 shows static parameters of networks from the final inference GRNs. The checking of FTFF and FTTT GRNs shows the correct approach because the parameters are similar. The checking is provided during the whole research for checking because we create GRN for the non-model organism and so we have any possibility for checking of our results.

GRNs include 10 416 edges for FTFF data and 9 982 edges for FTTT data. At the sight numbers of edges are huge but these numbers include edges which represent themselves express influence information. Thus, if we subtract these themselves edges, we obtain 5 140 edges among different genes for FTFF dataset and 4 706 edges among different genes for FTTT dataset. The part of example of final GRNs is shown in the Attachment J].

Figure 24 is the example of interconnected approach. We have taken the first neighbourhood of X276_12610 from FTTT dataset, see Attachment J]. The X276_12610 is locus tag which is visible and described in the chapter 5.3.2. We have also taken X276_18480 locus tag and created final sub-graph which visualises these specific genes such as the possible candidate of transcription factor. This sub-graph includes not only all approaches which are based on the different methods but also the information about predict operons. The information about the directed edges which come out of transcription factor are taken from R/bioconductor Genie3 method and *'tsni.m'* toolbox.

The X276_12610 is a possible candidate of the transcription factor for 6 operons and 17 other genes in the Figure 24. The X276_18480 which represents Spo0A is visible connection with X276_01040, again. Thus, we declare that Spo0A has significant dependency with glgD gene.

Table 13: Basic static analysis union GRNs

	FTFF	FTTT
Numbers of nodes	5 276	5 276
Numbers of edges	10 416	9 982
Clustering coefficient	0.101	0.092
Connected components	4 123	4190
Characteristic path length	3.044	2.914
Average no. of neighbours	1.733	1.559

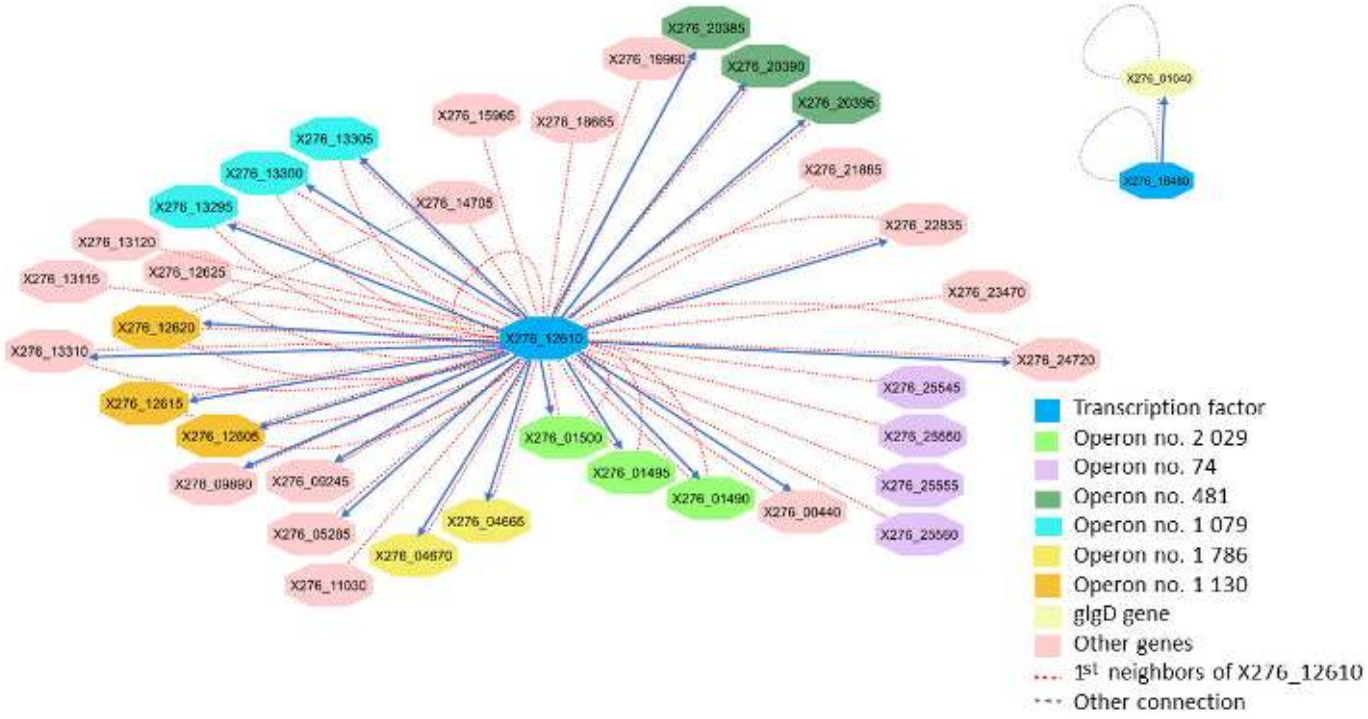


Figure 24: Example of interconnected approach

At the end we have created the adjacency list for the inference GRN which originated by the intersection FTTT and FTFF final networks. This adjacency list includes 8 787 edges which can be declared as significant edges in GRN for *C. beijerinckii* NRRL B-598. However, we need more replicates for obtaining more causality results which are sequenced in more time-points. Thus, in this thesis we have predicted the first GRN for *C. beijerinckii* NRRL B-598 including 8 787 possible candidates of edges which can be considered as the predecessor for more causality results in future.

Conclusion

The master's thesis deals with the study of gene regulation in *Clostridium beijerinckii* NRRL B-598. The thesis is focused on the description of gene regulatory nomenclature, inference gene regulatory networks, description of laboratory methods for obtained laboratory data which are usually used for studying gene regulatory. The described laboratory methods are mainly focused on the technologies of RNA-Seq and brief description laboratory data which have been got for examined bacterium *C. beijerinckii* NRRL B-598. The following part of the thesis is the theoretical description of the network inference methods which are the most used methods in current and also used in the chapters for inference gene regulatory networks for *C. beijerinckii* NRRL B-598.

The practical part of the thesis starts with pre-processing raw laboratory data obtained from *C. beijerinckii* NRRL B-598. This pipeline of pre-processing can be used for the pre-processing raw laboratory data of other bacterium therefore it is uploaded in the git hub /JanaSchwarzerova/Analytical-pipeline-rawRNA-Seq where is available. The results of pre-processing steps were evaluated using PCA plots where is shown that the samples sequenced in the same conditions become the clusters. Thus, it represents correctness of created analytical pipeline.

The outputs of pre-processing step are count tables which represent gene express values. We have created the count tables by two different approaches which we have called as FTFF and FTTT, more information in the chapter 4.6. The separation of data is done for following checking. We assumed the similar results of parameters gene regulatory networks. This checking is done during all procedure of inference the first gene regulatory networks for *C. beijerinckii* NRRL B-598.

The first step of the research of gene regulatory in *C. beijerinckii* NRRL B-598 is focused on the inference operons list. The inference operons list is obtained for the using of the combination machine learning approach which is included in the online tool Operon-mapper and expresses the gene value which has been obtained from *C. beijerinckii* NRRL B-598. We have predicted 2 737 operons. The final steps in this thesis have derived the first gene regulatory network for *C. beijerinckii* NRRL B-598 as an adjacency list which includes 8 787 edges. These edges are obtained by the interconnected different approaches and intersection of the two final gene regulatory networks from FTFF and FTTT datasets.

Literature

[1] *Nature research: Gene regulation*-[online Nature.com]. Springer Nature Publishing, 2019-[Accessed 2019-09-06]. Available from: https://www.nature.com/subjects/gene-regulation

[2] *Khan Academy: Gene regulation in bacteria*-[online www.khanacademy.org]. 2019-[Accessed 2019-09-06]. Available from: https://www.khanacademy.org/science/biology/gene-regulation/gene-regulation-in-bacteria/a/overview-gene-regulation-in-bacteria

[3] *NCBI: Gene regulatory networks*-[online https://www.ncbi.nlm.nih.gov]. USA, 2005-[Accessed 2019-09-06].Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC556010/

[4] BARABÁSI, Albert-László a Zoltán N. OLTVAI. NETWORK BIOLOGY: UNDERSTANDING THE CELL'S FUNCTIONAL ORGANIZATION. *Nature Reviews Genetics*-[online].2004, 101-113-[Accessed 2019-09-06]. DOI: https://doi.org/10.1038/nrg1272. ISSN 1471-0064. Available from: https://www.nature.com/articles/nrg1272

[5] WALHOUT, Albertha J.M. *Gene-Centered Regulatory Network Mapping*-[online]. 271-288-[Accessed 2019-10-06].DOI:10.1016/B978-0-12-544172-8.00010-4. Available from:https://www.sciencedirect.com/science/article/pii/B9780125441728000104

[6] XIA, Tian. *Network modeling in systems biology*. Ames, Iowa, United States, 2010. Graduate Theses and Dissertations. Iowa State University.

[7] *NCBI: Basic Local Alignment Search Tool*-[online]. USA: U.S. National Library of Medicine, 2019-[Accessed 2019-12-01]. Available from: https://blast.ncbi.nlm.nih.gov/Blast.cgi.

[8] SHANNON, Paul, Andrew MARKIEL, Owen OZIER, et al. *Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks*-[online]. 2003[Accessed 2020-04-05].DOI: 10.1101/gr.1239303. Available from: https://genome.cshlp.org/content/13/11/2498.short

[9] LÄHDESMÄKI, Harri, Ilya SHMULEVICH and Olli YLI-HARJA. On Learning Gene Regulatory Networks Under the Boolean Network Model. *Machine Learning*. Kluwer Academic Publishers. Manufactured in The Netherlands, 2003, s. 147–167. ISBN 1573-056

[10] XING, Linlin, Lei WANG, Chunyu WANG, Xiaoyan LIU, Maozu GUO a Yin ZHANG. An improved Bayesian network method for reconstructing gene regulatory network based on candidate auto selection. *BMC Genomics*-[online]. 17 November 2017-[Accessed 2019-12-02]. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-4228-y

[11] ANDRECUT, M., D. CLOUD a S.A. KAUFFMAN. Monte Carlo simulation of a simple gene network yields new evolutionary insights. *Journal of Theoretical Biology*-[online]. 7 February 2008, 468-474-[Accessed 2019-12-02]. DOI: https://doi.org/10.1016/j.jtbi.2007.10.035. Available from: https://www.sciencedirect.com/science/article/pii/S0022519307005425

[12] MARBACH, Daniel, James C COSTELLO, Robert KÜFFNER, et al. *Wisdom of crowds for robust gene network inference*-[online]. 15 July 2012, pages 796–804-[Accessed 2019-10-06]. DOI: 10.1038/nmeth.2016. Available from: https://www.nature.com/articles/nmeth.2016

[13] FUENTE, Alberto. Gene Network Inference. Germany: Springer-Verlag Berlin and Heidelberg GmbH & Co., 2014. ISBN 978-3-642-45160-7.

[14] ALEKSEYEV, Yuriy O., Roghayeh FAZELI, Shi YANG, Raveen BASRAN, Thomas MAHER, Nancy S. MILLER a Daniel REMICK. *A Next-Generation Sequencing Primer—How Does It Work and What Can It Do?*-[online]. 2018 May 6-[Accessed 2019-10-30]. DOI: 10.1177/2374289518766521.
Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5944141/

[15] *Atdbio: Next generation sequencing*-[online].-[Accessed 2019-10-30]. Available from: https://www.atdbio.com/content/58/Next-generation-sequencing#Sequencing-by-ligation-SOLiD

[16] *Your genome: What is the illumina method of dna sequencing*-[online]. [Accessed 2019-10-31]. Available from: https://www.yourgenome.org/facts/what-is-the-illumina-method-of-dna-sequencing

[17] *Thermofisher: Ion Torrent Next-Generation Sequencing Technology*-[online]. [Accessed 2019-11-03]. Available from: https://www.thermofisher.com/cz/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-technology.html

[18] *Illumina: Sequencing and array-based solutions for genetic research*-[online]. 2019-[Accessed 2019-12-03]. Available from: https://www.illumina.com/

[19] *MiSeq™ System; Illumina*-[online]. 2018-[Accessed 2019-11-01]. Available from: https://www.illumina.com/documents/products/datasheets/datasheet_miseq.pdf

[20] *NextSeq Sequencing System; Illumina*-[online].-[Accessed 2019-11-02]. Available from: https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/nextseq-550-system-spec-sheet-770-2013-053.pdf

[21] *HiSeq™ Sequencing Systems; Illumina*-[online]. 2014-[Accessed 2019-11-02]. Available from: https://www.illumina.com/documents/products/datasheets/datasheet_hiseq_systems.pdf

[22] *NovaSeq™ 6000 Sequencing System; Illumina*-[online]. 2019 [Accessed 2019-11-03]. Available from: https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/novaseq-6000-system-specification-sheet-770-2016-025.pdf

[23] *OPINIOMICS: HiSeq move over, here comes Nova! A first look at Illumina NovaSeq*-[online]. 2017-[Accessed 2019-11-03]. Available from: http://www.opiniomics.org/hiseq-move-over-here-comes-nova-a-first-look-at-illumina-novaseq/

[24] *An introduction to Next-Generation Sequencing Technology; Illumina*-[online]. 2017 [Accessed 2019-11-03]. Available from: https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

[25] J MACKENZIE, Ruairi. *RNA-seq: Basics, Applications and Protocol*-[online]. 2018-[Accessed 2019-12-04]. Available from: https://www.technologynetworks.com/genomics/articles/rna-seq-basics-applications-and-protocol-299461

[26] *Nature: transcriptome*-[online]. 2014-[Accessed 2019-12-04]. Available from: https://www.nature.com/scitable/definition/transcriptome-296/

[27] WANG, Zhong, Mark GERSTEIN a Michael SNYDER. *RNA-Seq: a revolutionary tool for transcriptomics*-[online]. 2009, 57–63-[Accessed 2019-12-03]. DOI: 10.1038/nrg2484. Available from: https://www.nature.com/articles/nrg2484

[28] *Transcriptomics*-[online].-[Accessed 2019-12-04]. Available from: http://alnelsongen564s17.weebly.com/transcriptome.html

[29] *RNA Sequencing methods collection; Illumina*-[online]. 2017-[Accessed 2019-11-03]. Available from: https://www.illumina.com/content/dam/illumina-marketing/documents/products/research_reviews/rna-sequencing-methods-review-web.pdf

[30] ZHAO, Shanrong, Ying ZHANG, William GORDON, Jie QUAN, Hualin XI, Sarah DU, David von SCHACK a Baohong ZHANG. *Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap*-[online]. 2015 [Accessed 2020-04-05]. DOI:10.1186/s12864-015-1876-7. ISSN 1471-2164. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-1876-7

[31] GRIFFITH, Malachi, Jason R. WALKER, Nicholas C. SPIES, Benjamin J. AINSCOUGH a Obi L. GRIFFITH. *Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud*-[online]. 2015-[Accessed 2020-04-05]. DOI: 10.1371/journal.pcbi.1004393.

[32] GOLDMAN, Emanuel a Lorrence H GREEN. *Practical Handbook of Microbiology*. 4 June 2015. Boca Raton: CRC Press, 2015. ISBN 9780429168932.

[33] BARNABY WHITMAN, William a Bergey's MANUAL TRUST. *Bergey's manual of systematics of archaea and bacteria*. USA:-[Hoboken, New Jersey]: Wiley,-[2015] ©2015, 2015. ISBN 9781118960608 1118960602.

[34] *NCBI: Clostridium beijerinckii*-[online].-[Accessed 2019-11-16]. Available from: https://www.ncbi.nlm.nih.gov/genome/?term=clostridium%20beijerinckii

[35] SEDLAR, K, J KOLEK, I PROVAZNIK a P PATAKOVA. *Reclassification of non-type strain Clostridium pasteurianum NRRL B-598 as Clostridium beijerinckii NRRL B-598*-[online]. 2017 Jan 19-[Accessed 2019-11-16]. DOI: 10.1016/j.jbiotec.2017.01.003. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28111164

[36] SEDLAR, Karel, Pavlina KOSCOVA, Maryna VAYLKIVSKA, Barbora BRANSKA, Jan KOLEK, Kristyna KUPKOVA, Petra PATAKOVA and Ivo PROVAZNIK. *Transcription profiling of butanol producer Clostridium beijerinckii NRRL B-598 using RNA-Seq*-[online]. 30 May 2018-[Accessed 2019-11-16]. DOI: 10.1186/s12864-018-4805-8. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-018-4805-8

[37] PATAKOVA, Petra, Barbora BRANSKA, Karel SEDLAR, Maryna VASYLKIVSKA, Katerina JURECKOVA, Jan KOLEK, Pavlina KOSCOVA and Ivo PROVAZNIK. *Acidogenesis, solventogenesis, metabolic stress response and life cycle changes in Clostridium beijerinckii NRRL B-598 at the transcriptomic level*-[online]. 04 February 2019-[Accessed 2019-11-16]. DOI: 10.1038/s41598-018-37679-0. Available from: https://www.nature.com/articles/s41598-018-37679-0

[38] VASYLKIVSKA, Maryna, Katerina JURECKOVA, Barbora BRANSKA, Karel SEDLAR, Jan KOLEK, Ivo PROVAZNIK a Petra PATAKOVA. *Transcriptional analysis of amino acid, metal ion, vitamin and carbohydrate uptake in butanol-producing Clostridium beijerinckii NRRL B-598*-[online]. November 7, 2019-[Accessed 2019-11-16]. DOI: 10.1371/journal.pone.0224560. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0224560

[39] SEDLAR, Karel, Jan KOLEK, Markus GRUBER, et al. *A transcriptional response of Clostridium beijerinckii NRRL B-598 to a butanol shock*-[online]. 13 October 2019 [Accessed 2019-11-16]. DOI: 10.1186/s13068-019-1584-7. Available from: https://biotechnologyforbiofuels.biomedcentral.com/articles/10.1186/s13068-019-1584-7

[40] EFRON, Bradley. *The jackknife, the bootstrap, and other resampling plans*-[online]. Stanford California: Stanford University, 1980-[Accessed 2019-12-07]. ISBN 978-0-89871-179-0. Available from: https://statistics.stanford.edu/sites/g/files/sbiybj6031/f/BIO%2063.pdf

[41] *Machine Learning Mastery: A Gentle Introduction to the Bootstrap Method*-[online]. August 8, 2019-[Accessed 2019-11-08]. Available from: https://machinelearningmastery.com/a-gentle-introduction-to-the-bootstrap-method/

[42] *E math zone: Sampling With Replacement*-[online].-[Accessed 2019-11-08]. Available from: https://www.emathzone.com/tutorials/basic-statistics/sampling-with-replacement.html

[43] CFI EDUCATION INC. *Regression Analysis*-[online]. 2015-[Accessed 2019-12-07]. Available from: https://corporatefinanceinstitute.com/resources/knowledge/finance/regression-analysis/

[44] Least Absolute Shrinkage and Selection Operator (LASSO). *Columbia University Mailman School of Public Health*-[online]. 722 West 168th St. NY, NY 10032-[Accessed 2019-12-07]. Available from: https://www.mailman.columbia.edu/research/population-health-methods/least-absolute-shrinkage-and-selection-operator-lasso

[45] CANDES, Emmanuel a Terence TAO. The Dantzig selector: Statistical esstimation when p is much larger than n1. *Institute of Mathematical Statistics*-[online]. 2007-[Accessed 2019-12-07].
DOI: 10.1214/009053606000001523. Available from: https://projecteuclid.org/download/pdfview_1/euclid.aos/1201012958

[46] *Wolfram Math World: Infimum*-[online mathworld.wolfram.com].-[Accessed 2019-11-09]. Available from: http://mathworld.wolfram.com/Infimum.html

[47] BREIMAN, Leo a Adele CUTLER. *Random Forests*-[online].-[Accessed 2019-12-07]. Available from: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#workings

[48] LATHAM, Peter E. a Yasser ROUDI. *Scholarpedia: Mutual information*-[online]. 2009-[Accessed 2019-11-11]. DOI: 10.4249/scholarpedia.1658. Available from: http://www.scholarpedia.org/article/Mutual_information

[49] *MetaCentrum-Virtual Organization*-[online].-[Accessed 2020-04-30]. Available from: https://metavo.metacentrum.cz/

[50] DELHOMME, Nicolas, Niklas MÄHLER, Bastian SCHIFFTHALER, David SUNDELL, Chanaka MANNAPPERUMA, Torgeir R. HVIDSTEN a Nathaniel R. STREET. Guidelines for RNA-Seq data analysis. *Epigenesys*-[online]. 17 November 2014-[Accessed 2019-12-24]. Available from:http://52.18.118.183:3000/materials/Lecture/Pre-course/20150303161357_p67.pdf

[51] Functional genomics (II): Common technologies and data analysis methods: RNA-Sequencing – Data analysis. *Train online*-[online].-[Accessed 2019-12-26]. Available from: https://www.ebi.ac.uk/training/online/course/functional-genomics-ii-common-technologies-and-data-analysis-methods/quantification

[52] Y1, Liao, Smyth GK a Shi W. *FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features.*-[online]. 2014-[Accessed 2020-04-06]. DOI: 10.1093/bioinformatics/btt656. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24227677

[53] CONESA, Ana, Pedro MADRIGAL, Sonia TARAZONA, et al. *A survey of best practices for RNA-seq data analysis*-[online]. In: 2016, 2016-[Accessed 2019-12-26]. DOI: 10.1186/s13059-016-0881-8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4728800/

[54] *GitHub: Ribosomal RNA removal using SORTMERNA*-[online]. 2017-[Accessed 2019-12-25]. Available from: https://github.com/biomendi/TRANSCRIPTOME-ASSEMBLY-PIPELINE/wiki/3.-Ribosomal-RNA-removal-using-SORTMERNA

[55] KOPYLOVA, Evguenia. *: SortMeRNA User Manual*-[online]. 2014-[Accessed 2019-12-25]. Available from: https://bioinfo.lifl.fr/RNA/sortmerna/code/SortMeRNA-user-manual-v2.0.pdf

[56] *THE USADEL LAB: Trimmomatic Manual: V0.32*-[online].-[Accessed 2019-12-25]. Available from: http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.32.pdf

[57] DOBIN, Alexander, Carrie A. DAVIS, Felix SCHLESINGER, et al. *STAR: ultrafast universal RNA-seq aligner*-[online]. 2013-[Accessed 2019-12-25]. DOI: 10.1093/bioinformatics/bts635. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3530905/

[58] LANGMEAD, Ben. Burrows-Wheeler Transform and FM Index. *Johns Hopkins University*-[online].-[Accessed 2019-12-25]. Available from: https://www.cs.jhu.edu/~langmea/resources/lecture_notes/bwt_and_fm_index.pdf

[59] ROBINSON, James T, Helga THORVALDSDÓTTIR, Wendy WINCKLER, Mitchell GUTTMAN, Eric S LANDER, Gad GETZ a Jill P MESIROV. *Integrative genomics viewer*-[online]. 10 January 2011-[Accessed 2020-04-08]. DOI: 10.1038/nbt.1754.

[60] SCHURCH, Nicholas J., Pietá SCHOFIELD, Marek GIERLIŃSKI, et al. *How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?*-[online]. 2016, 839–851-[Accessed 2020-04-08]. DOI: 10.1261/rna.053959.115.

[61] ROBINSON, Mark D., Davis J. MCCARTHY a Gordon K. SMYTH. *EdgeR: a Bioconductor package for differential expression analysis of digital gene expression data*-[online]. 1 January 2010, 139–140-[Accessed 2020-04-07]. DOI: doi.org/10.1093/bioinformatics/btp616.

[62] LOVE, Michael I., Simon ANDERS a Wolfgang HUBER. *Differential analysis of count data – the DESeq2 package*-[online]. May 25, 2016-[Accessed 2020-04-07]. DOI: doi.org/10.1186/s13059-014-0550-8.

[63] *RNA-seq workflow - differential expression*-[online]. April 9, 2019-[Accessed 2019-12-28]. Available from: http://52.18.118.183:3000/materials/Exercises/Day4/rnaseq_gene_EBI_advRNAseq_2019.html

[64] STACKLIES, Wolfram, Henning REDESTIG, Matthias SCHOLZ, Dirk WALTHER a Joachim SELBIG. *PcaMethods—a bioconductor package providing PCA methods for incomplete data*-[online]. 07 March 2007, 1164–1167-[Accessed 2020-04-08]. DOI: doi.org/10.1093/bioinformatics/btm069.

[65] SCHLIEP, Klaus Peter. *Phangorn: phylogenetic analysis in R*-[online]. 17 December 2010, 592–593-[Accessed 2020-04-09]. DOI: doi.org/10.1093/bioinformatics/btq706

[66] GARANINA, Irina A., Gleb Y. FISUNOV a Vadim M. GOVORUN. *BAC-BROWSER: The Tool for Visualization and Analysis of Prokaryotic Genomes*-[online]. 2018-[Accessed 2020-04-09]. DOI: 10.3389/fmicb.2018.02827. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6258810/

[67] *BioNinja: Operons*-[online].-[Accessed 2020-04-09]. Available from: https://ib.bioninja.com.au/higher-level/topic-7-nucleic-acids/72-transcription-and-gene/operons.html

[68] *Scitable by nature EDUCATION: transcription unit*-[online]. 2005-[Accessed 2020-04-09]. Available from: https://www.nature.com/scitable/definition/transcription-unit-260/

[69] B, Taboada, Estrada K, Ciria R a Merino E. *Operon-mapper: a web server for precise operon identification in bacterial and archaeal genomes.*-[online]. 2018-[Accessed 2020-04-09]. DOI: 10.1093/bioinformatics/bty496.

[70] ZHANG, Bin a Steve HORVATH. *A General Framework for Weighted Gene Co-Expression Network Analysis*-[online]. 2005-[Accessed 2020-04-19]. DOI: 10.2202/1544-6115.1128.

[71] HENAO, Juan David. *Coexnet: An R package to build CO-EXpression NETworks from Microarray Data*-[online]. 2017-[Accessed 2020-04-20]. Available from: https://bioc.ism.ac.jp/packages/3.6/bioc/vignettes/coexnet/inst/doc/coexnet.pdf

[72] SEDLAR, K, H SKUTKOVA, J KOLEK, P PATAKOVA a I PROVAZNIK. IDENTIFICATION AND CHARACTERIZATION OF SOL OPERON IN CLOSTRIDIUM PASTEURIANUM NRRL B-598 GENOME. *Conference: 2nd International Conference on Chemical Technology ICCT2014*-[online]. 2014-[Accessed 2020-04-23]. DOI: 10.13140/2.1.3688.6402.

[73] *InterPro Classification of protein families: A0A0K2MKB7*-[online].-[Accessed 2020-04-23]. Dostupné z: https://www.ebi.ac.uk/interpro/protein/UniProt/A0A0K2MKB7/

[74] *UniProt: UniProtKB - A0A0K2M7P6*-[online].-[Accessed 2020-04-23]. Available from: https://www.uniprot.org/uniprot/A0A0K2M7P6

[75] ASSENOV, Yassen, Fidel RAMÍREZ, Sven-Eric SCHELHORN, Thomas LENGAUER a Mario ALBRECHT. *Computing topological parameters of biological networks*-[online]. 2008-[Accessed 2020-04-24]. DOI: 10.1093/bioinformatics/btm554.

[76] *NetworkAnalyzer Online Help: NetworkAnalyzer Settings*-[online]. 2018-[cit. 2020-05-10]. Dostupné z: https://med.bioinf.mpi-inf.mpg.de/netanalyzer/help/2.7/

[77] MERCATELLI, F., L. SCALAMBRA, L. RAY, F. GIORGI a M. GIORGI. Gene regulatory network inference resources: A practical overview. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*-[online]. 2019-[Accessed 2020-04-18]. DOI: doi.org/10.1016/j.bbagrm.2019.194430.

[78] SIMOES, Ricardo de Matos a Frank EMMERT-STREIB. *Bagging Statistical Network Inference from Large-Scale Gene Expression Data*-[online]. 2012-[Accessed 2020-05-03]. DOI: doi.org/10.1371/journal.pone.0033624.

[79] HUYNH-THU, Vân Anh a Pierre GEURTS. *DynGENIE3: dynamical GENIE3 for the inference of gene networks from time series expression data*-[online]. 2018-[Accessed 2020-05-03]. DOI: doi.org/10.1038/s41598-018-21715-0.

[80] WANG, L a B WEISS. *Dcd (dCTP Deaminase) Gene of Escherichia Coli: Mapping, Cloning, Sequencing, and Identification as a Locus of Suppressors of Lethal Dut (dUTPase) Mutations*-[online]. 1992-[cit. 2020-05-11]. DOI: 10.1128/jb.174.17.5647-5653.1992.

[81] GOODSELL, David. *Transposase*-[online]. 2006-[Accessed 2020-05-12]. DOI: 10.2210/rcsb_pdb/mom_2006_12.

[82] CALDERONE, Alberto a Gianni CESARENI. *Analysis of Triplet Motifs in Biological Signed Oriented Graphs Suggests a Relationship Between Fine Topology and Function*-[online]. 2019-[Accessed 2020-05-26]. Available from: https://arxiv.org/pdf/1803.06520.pdf

[83] WONG, Elisabeth, Brittany BAUR, Saad QUADER a Chun-Hsi HUANG. *Biological network motif detection: principles and practice*-[online]. 2011-[cit. 2020-05-26]. DOI: doi.org/10.1093/bib/bbr033.

[84] *Di Bernardo Lab - Systems and Synthetic Biology Lab: Time Series Network Identification TSNI*-[online].-[Accessed 2020-05-03]. Available from: https://dibernardo.tigem.it/softwares/time-series-network-identification-tsni

[85] SETTLE, Brett, David OTASEK, John H MORRIS a Barry DEMCHAK. *AMatReader: Importing adjacency matrices via Cytoscape Automation*-[online]. 2018-[Accessed 2020-05-04]. DOI: 10.12688/f1000research.15146.2.

List of attachments

Attachments

Attachment A] Example of Pre-processing shell-script

```
## Pre-processing raw data RNA-Seq
##**************************************************************************
## X-replicate
##_________________________________________________________________________

## Quality assesment (QA) Raw data*****************************************
cd /auto/ … /RNASeq_repX
# Add module fastQC
module add fastQC-0.11.5
# Take all files "*.gz" and is done quality check;
# results will be write in to "raw_data_qa"
fastqc -o raw_data_qa *.gz
# Add multiQC
module add python36-modules-gcc
pip freeze | grep network
#networkx==2.0
# Go to raw_data_qa file
cd raw_data_qa
# run Multiqc
multiqc .
# add necessarilly modules
export LC_ALL=C.UTF-8
export LANG=C.UTF-8
# run Multiqc
multiqc .
gunzip *.gz

## Delete rRNA ************************************************************
cd /auto/…/RNASeq_C_beijerinckii_NRRL_598/sortmerna-2.1-linux-64

#samples: X01, X02, X03, X04, X05, X06
for i in 1 2 3 4 5 6
do
./sortmerna --ref ./rRNA_databases/silva-bac-16s-id90.fasta,./index/silva-bac-
16s-db:\
./rRNA_databases/silva-bac-23s-id98.fasta,./index/silva-bac-23s-db\
 --reads /auto/…/raw_X0${i}.fastq\
 --aligned /auto/…/X_${i} --fastx\
 --other /auto/…/X_${i}_non_RNA --log -v -a 10 -m 4096
done

## QA nonrRNA *************************************************************
# Similar procedure such as section QA raw data

## Trimming dat ***********************************************************
module add trimmomatic-0.36
for i in 1 2 3 4 5 6
```

```
do
java -jar /software/trimmomatic/0.36/dist/jar/trimmomatic-0.36.jar SE -threads
10 X_${i}_non_RNA.fastq X_${i}_non_RNA_trim.fq ILLUMINACLIP:TruSeq3-SE:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
done

## QA Non-rRNA-Trim ******************************************************
# Similar procedure such as section QA raw data

## Mapping sequences *****************************************************
# Add module STAR
module add star-2.5.2b
# Go to genome_annotation file
cd /auto/…/genome_annotation
module add cufflinks-2.2.1
gffread -E -O -T genome.gff3 -o genome.gtf

## Index genome ... ONLY ONCE!
STAR --runThreadN 10 --runMode genomeGenerate\
--genomeDir /auto/…/genome_annotation\
--genomeFastaFiles /auto/…/genome_annotation/genome.fasta\
--sjdbGTFfile /auto/…/genome_annotation/genome.gtf\
--sjdbOverhang 48 --sjdbGTFfeatureExon CDS

# Itself mapping to index genome
for i in 1 2 3 4 5 6
do
STAR --runThreadN 10 --genomeDir /auto/…/genome_annotation\
--readFilesIn /auto/…/X_${i}_non_RNA_trim.fq\
--outFileNamePrefix /auto/…/X${i}\
--outFilterMultimapNmax 5 --outReadsUnmapped Fastx
done

## Quality Assesment (QA) mapping ****************************************
# Go to Mapping_sequences file
cd /auto/…/index_genome

# run Multiqc
multiqc .

## Sort SAM to BAM  ******************************************************
## before building count table is necessary SAM sorts
module add samtools-1.4

for i in 1 2 3 4 5 6
do
samtools sort -l 9 -o X_${i}.bam X${i}Aligned.out.sam
done

## Creating count table  *************************************************
module add subread-1.5.2
subread-buildindex
subread-align --help

featureCounts -T 2 -a /auto/…/genome.gff3 -o /auto/…/Count_table_X.txt -t gene -
g locus_tag -O X_1.bam X_2.bam X_3.bam X_4.bam X_5.bam X_6.bam
```

Attachment B] Results of read alignment to a reference

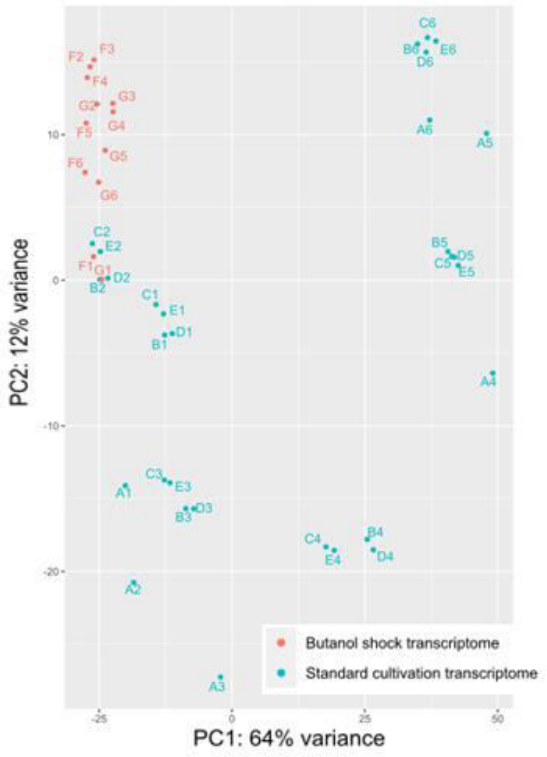
Table 14: Results of read alignment to reference from all samples

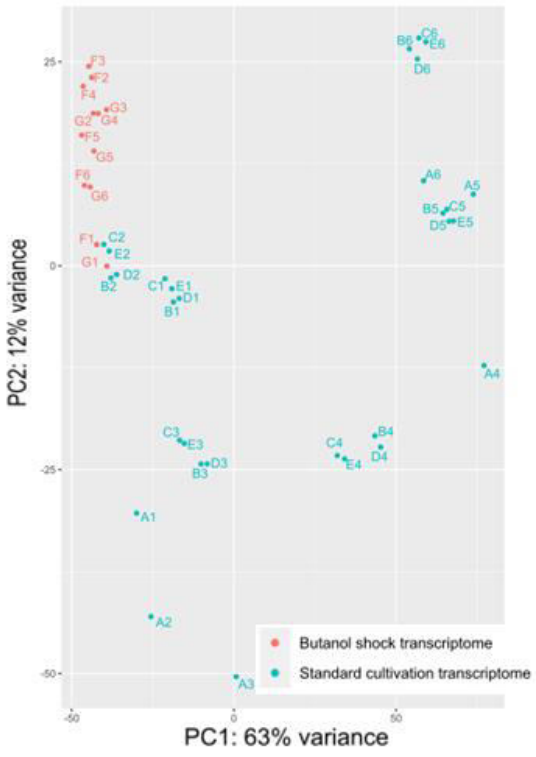
Sample Name	% Aligned	M Aligned	Sample Name	% Aligned	M Aligned
A1	95.9%	11,8	**E1**	96.3%	10,1
A2	96.3%	11,8	**E2**	96.4%	7,9
A3	95.7%	14,2	**E3**	96.2%	7,8
A4	96.2%	19,6	**E4**	95.7%	9,3
A5	94.4%	16,8	**E5**	95.9%	6,9
A6	92.8%	7,1	**E6**	94.4%	5,2
B1	95.9%	14,6	**F1**	84.4%	4,1
B2	96.2%	9,1	**F2**	80.5%	4,0
B3	96.6%	15,1	**F3**	89.6%	2,6
B4	96.0%	17,9	**F4**	89.2%	2,5
B5	96.4%	13,2	**F5**	77.7%	2,2
B6	95.7%	8,0	**F6**	91.9%	2,6
C1	95.3%	11,9	**G1**	89.5%	4,7
C2	96.0%	11,7	**G2**	78.5%	3,1
C3	96.1%	10,5	**G3**	93.3%	2,1
C4	95.8%	8,4	**G4**	94.6%	1,3
C5	96.0%	9,3	**G5**	94.8%	1,3
C6	95.6%	6,9	**G6**	90.8%	1,5
D1	95.5%	8,8			
D2	95.3%	6,3			**M Aligned**
D3	96.1%	8,6	**Mean**		8,3
D4	96.6%	14,8	**Maximum**		19,6
D5	95.8%	7,6	**Minimum**		1,3
D6	94.9%	5,5	**Standard deviation**		4,8

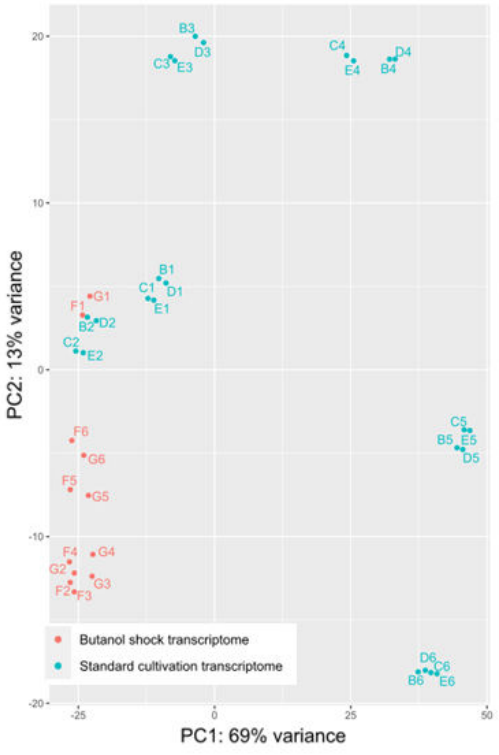
Table 15: Results of read alignment to reference from each replicates

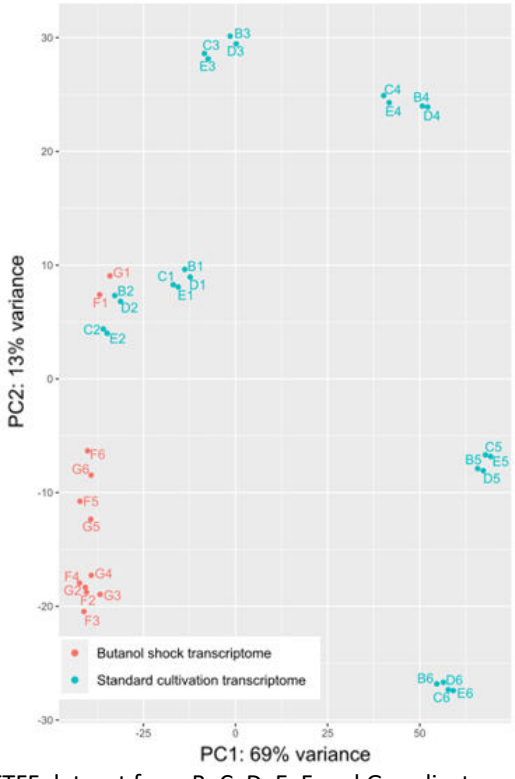
Replicate	Mean M Aligned	Maximum M Aligned	Minimum M Aligned	Standard deviation M Aligned
A	13,55	19,60	7,10	4,37
B	12,98	17,90	8,00	3,77
C	9,78	11,90	6,90	1,96
D	8,60	14,80	5,50	3,30
E	7,87	10,10	5,20	1,74
F	3,00	4,10	2,20	0,83
G	2,33	4,70	1,30	1,35

Attachment C] Figures PCA plots

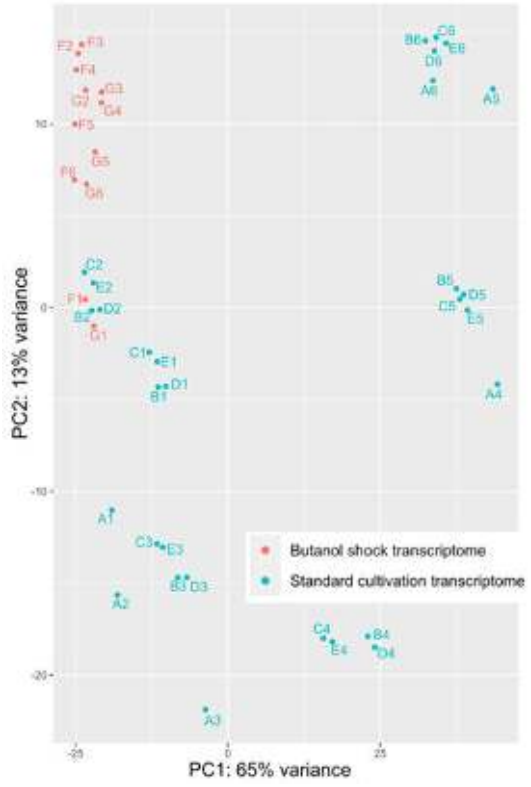
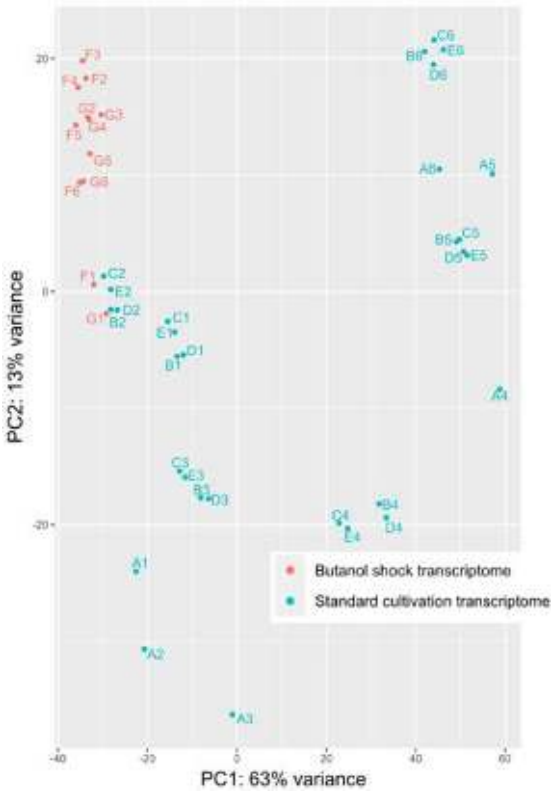
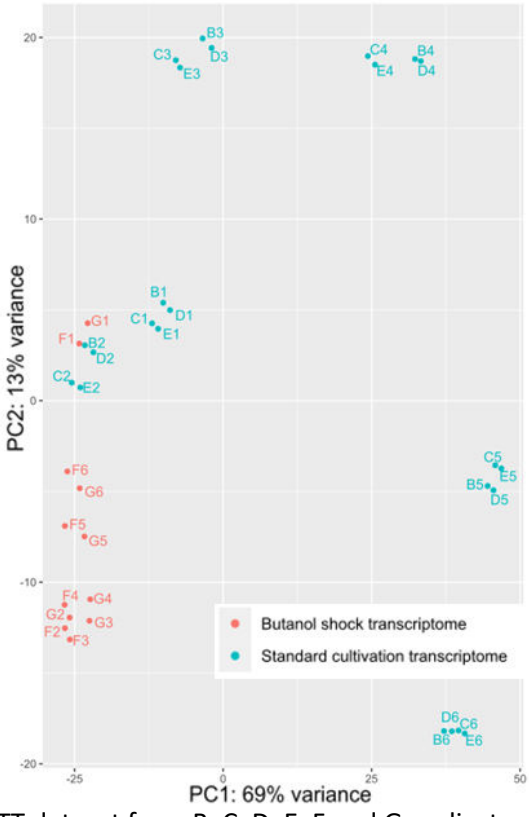
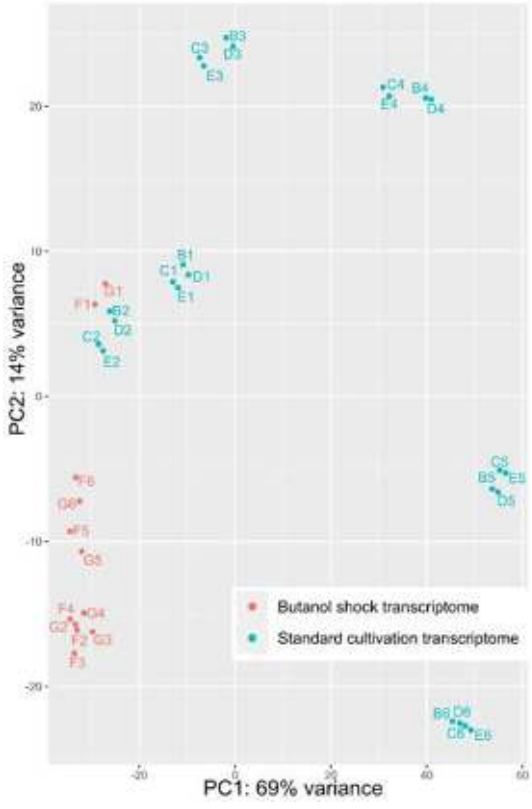
<table>
<tr><th colspan="2">Non-stranded dataset</th></tr>
<tr><th>Regularized transformation</th><th>Stabilized transformation</th></tr>
<tr><td> 

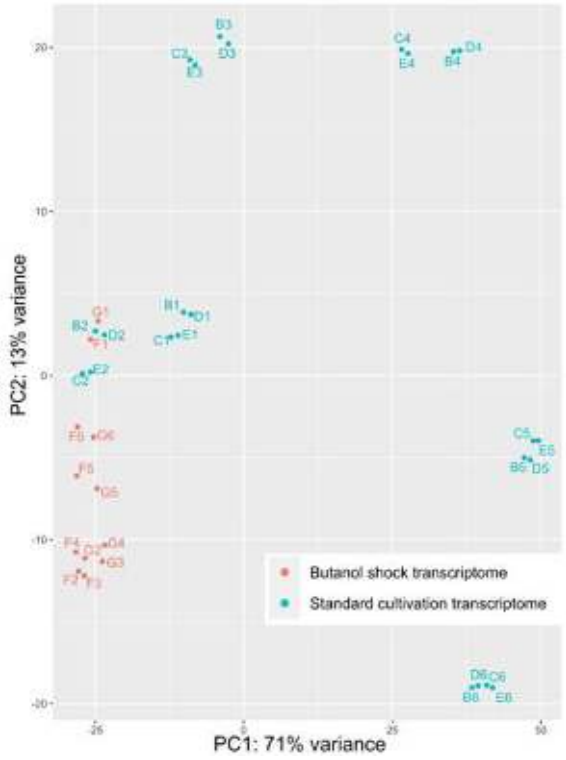
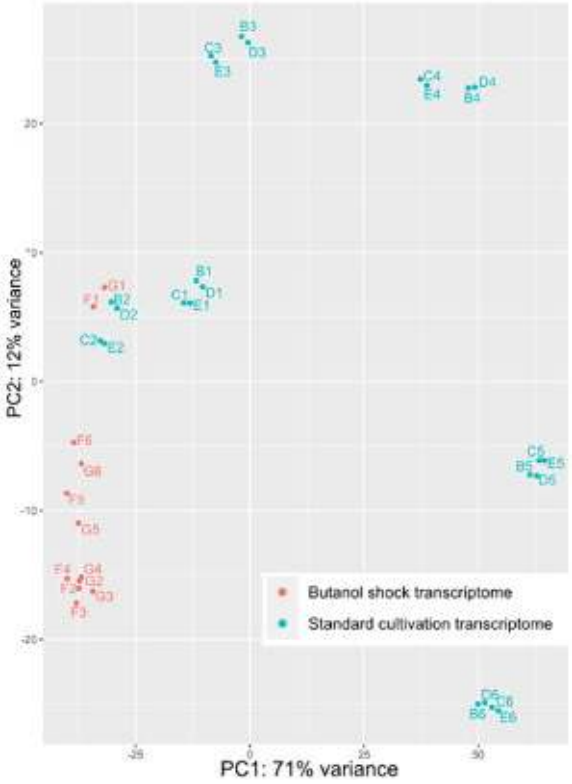
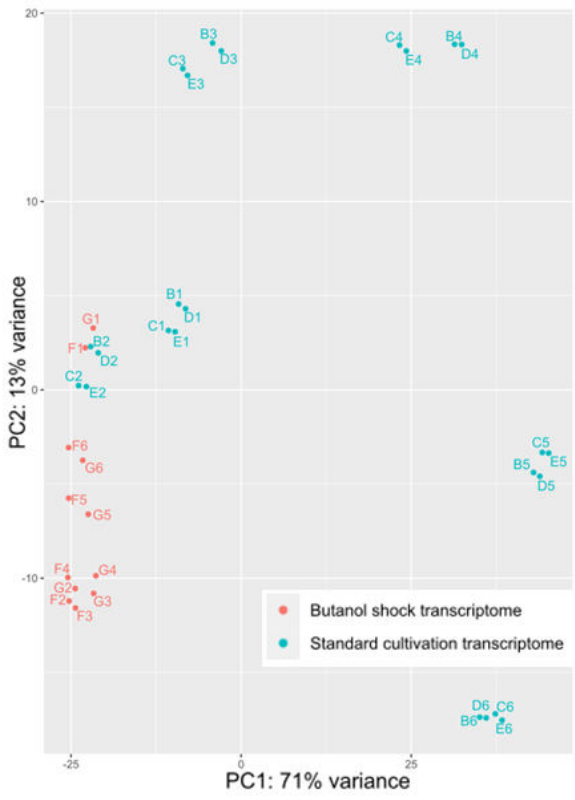
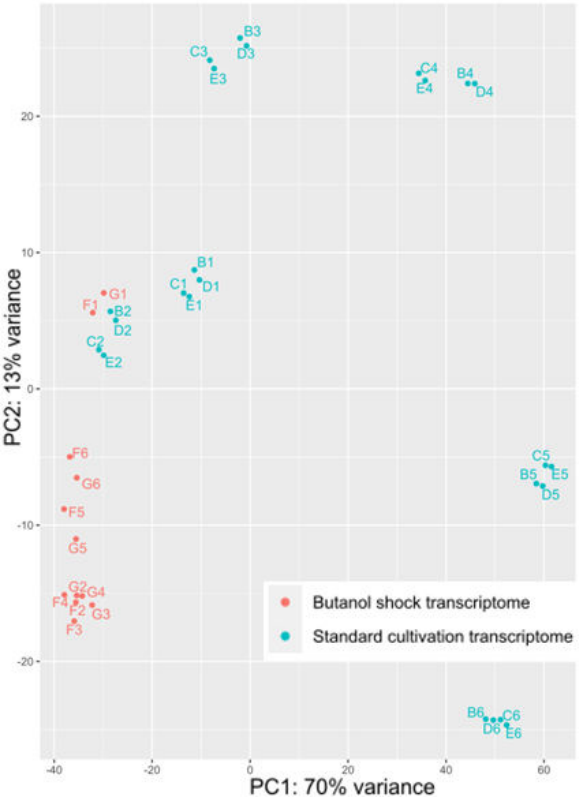
FTFF dataset from all replicates (A, B, C, D, E, F, G)</td><td> 

FTFF dataset from all replicates (A, B, C, D, E, F, G)</td></tr>
<tr><td> 

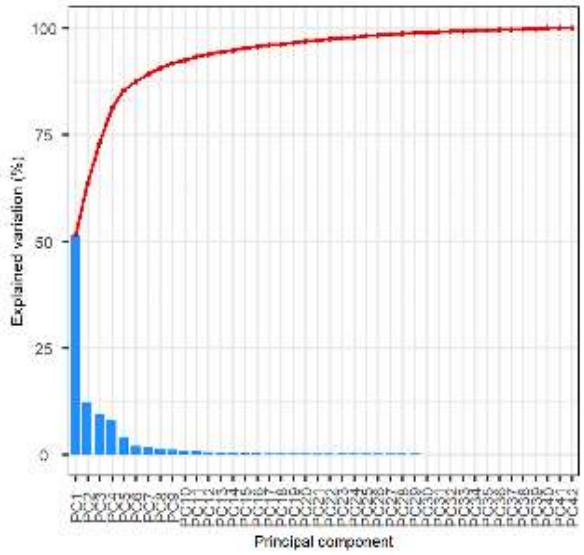
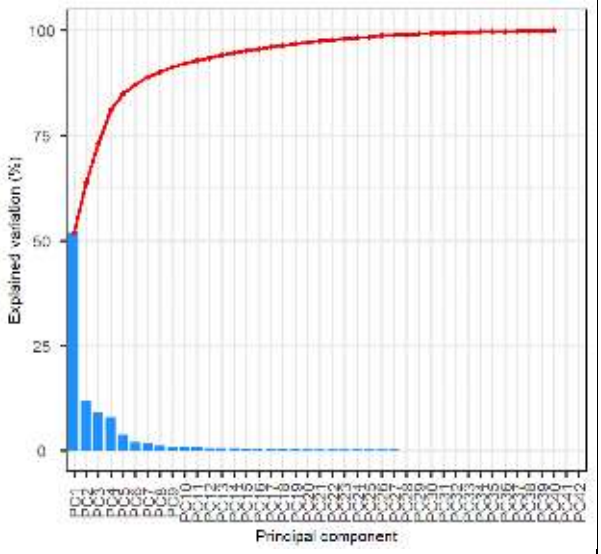
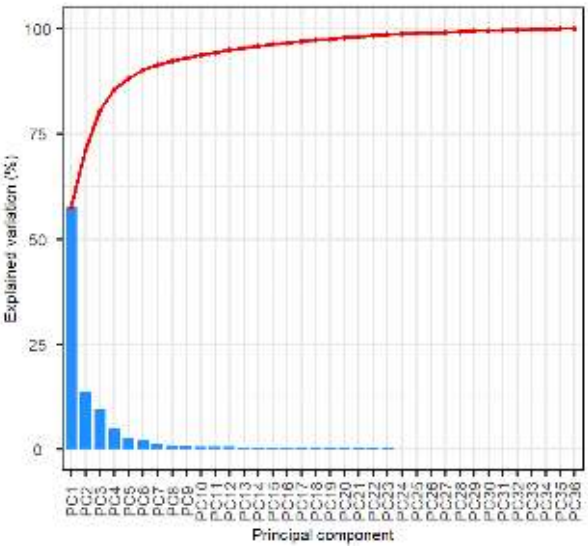
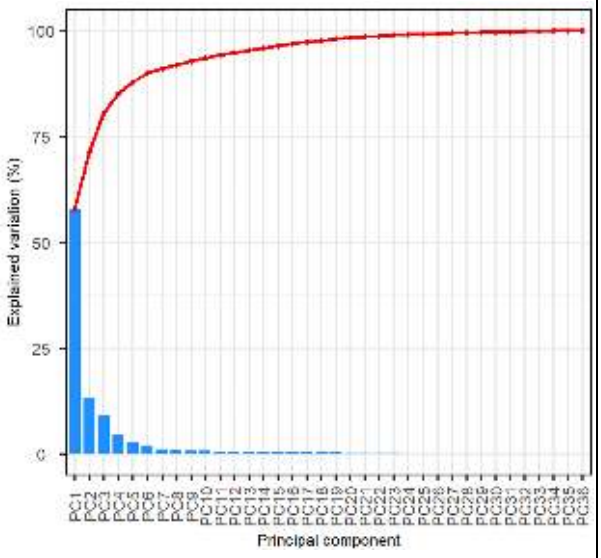
FTFF dataset from B, C, D, E, F and G replicates</td><td> 

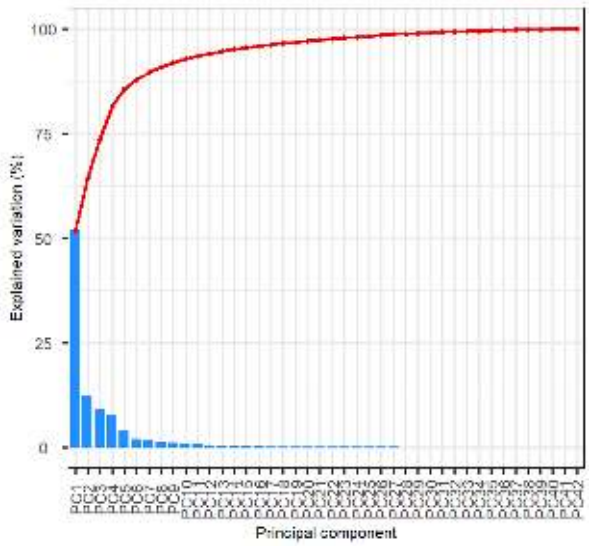
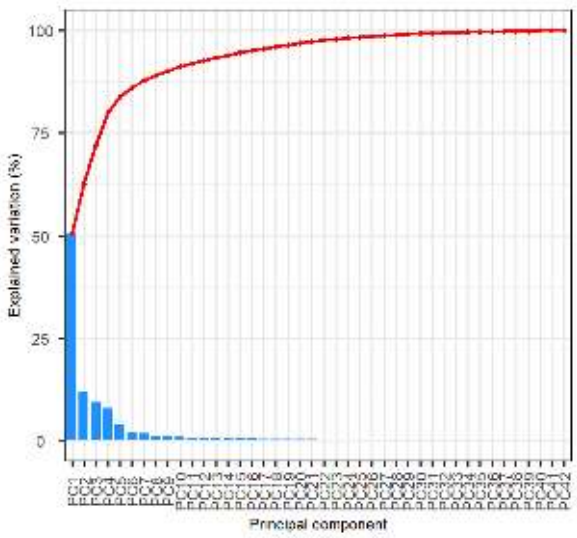
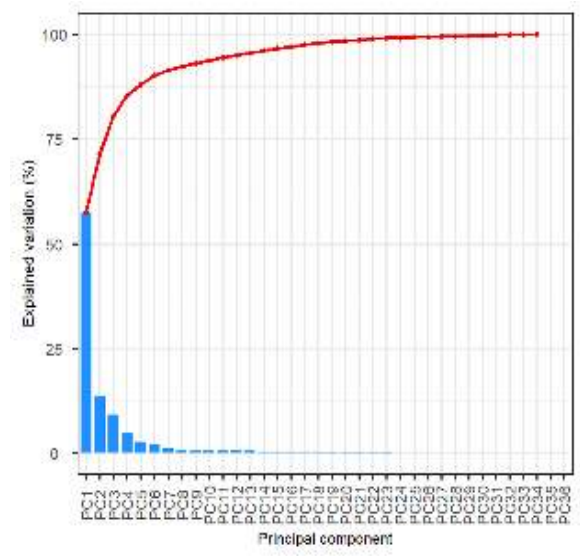
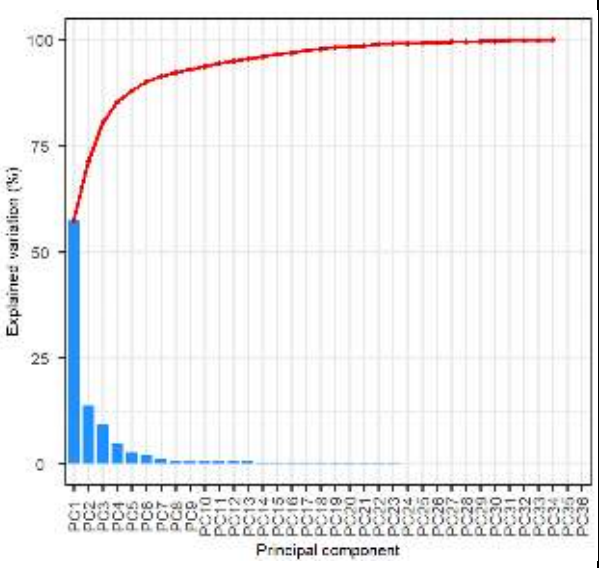
FTFF dataset from B, C, D, E, F and G replicates</td></tr>
</table>

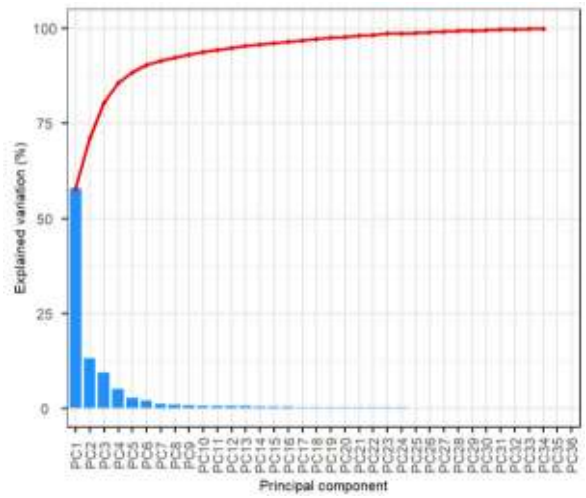
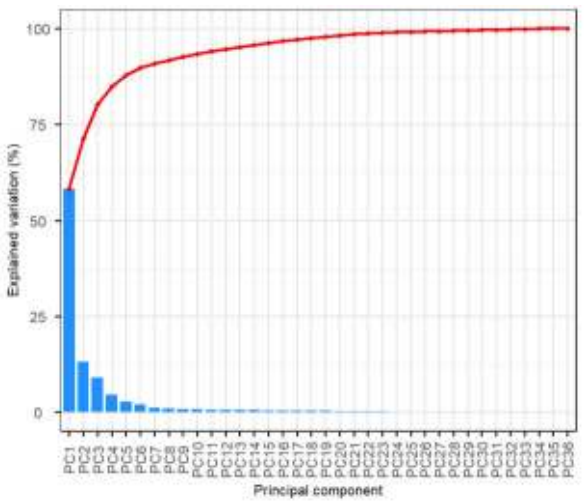
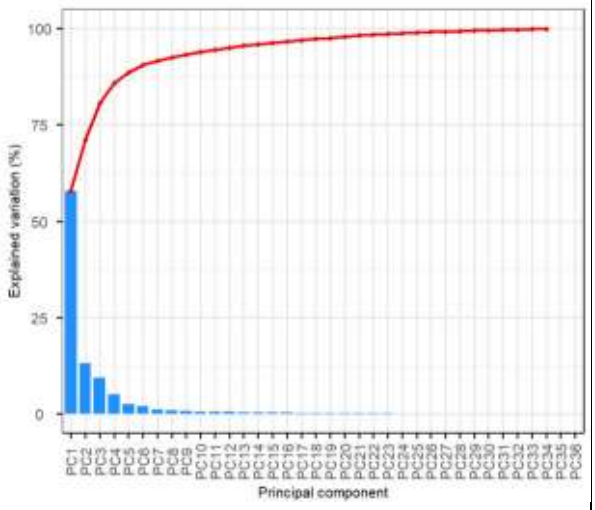
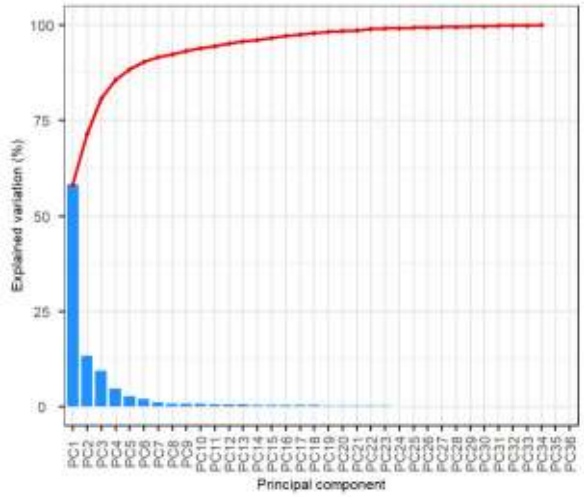
Non-stranded dataset	
Regularized transformation	**Stabilized transformation**
	
FTTT dataset from all replicates	FTTT dataset from all replicates
	
FTTT dataset from B, C, D, E, F and G replicates	FTTT dataset from B, C, D, E, F and G replicates

Reversely-strand dataset	
Regularized transformation	**Stabilized transformation**
 FTFF dataset from B, C, D, E, F and G replicates	 FTFF dataset from B, C, D, E, F and G replicates
 FTTT dataset from B, C, D, E, F and G replicates	 FTTT dataset from B, C, D, E, F and G replicates

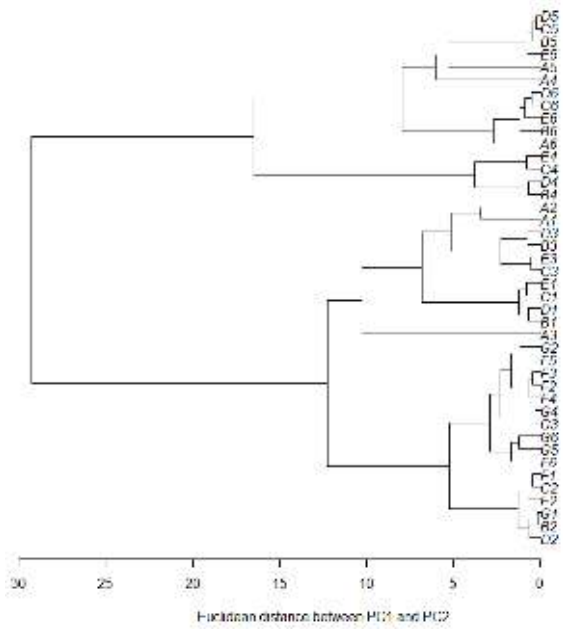
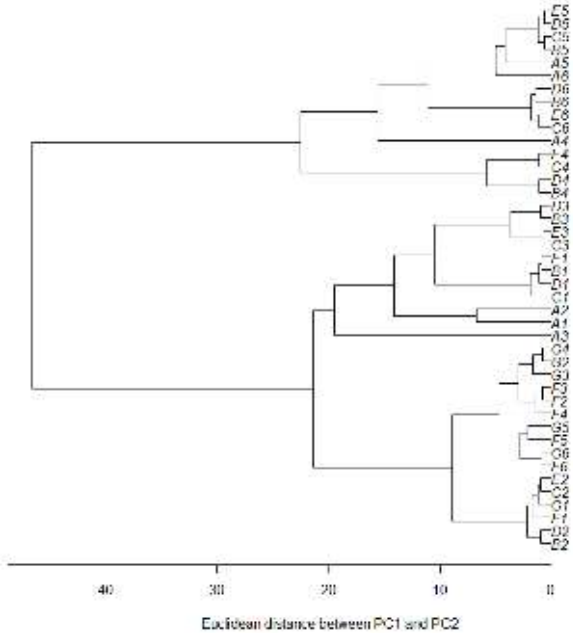
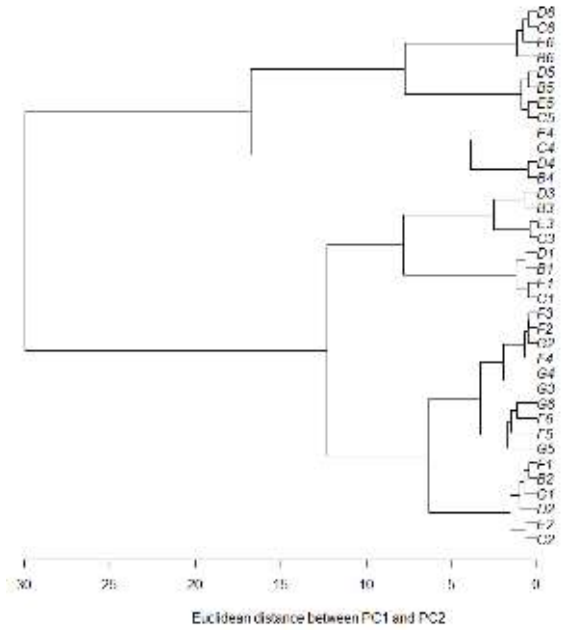
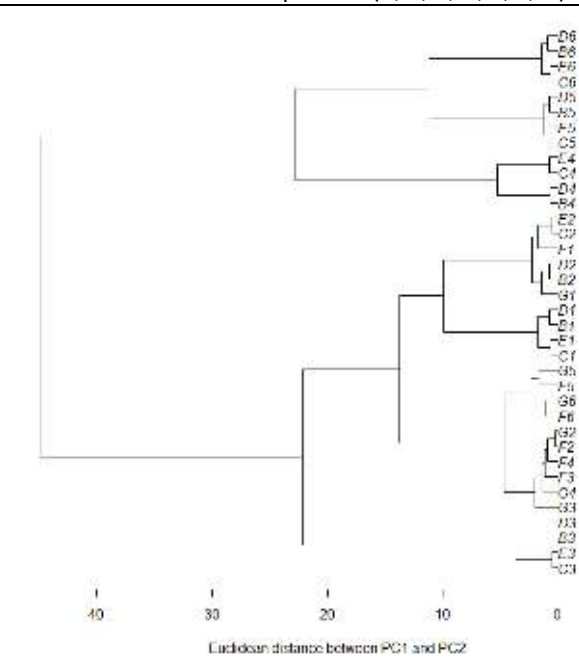
Attachment D] Figures Scree plots

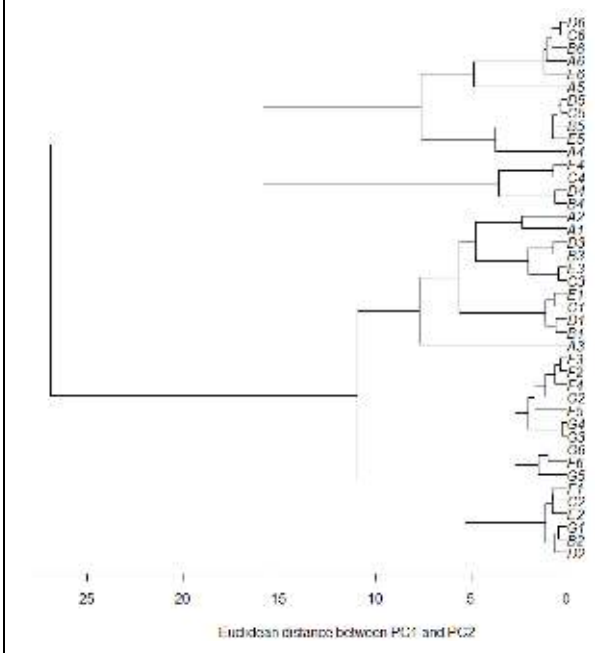
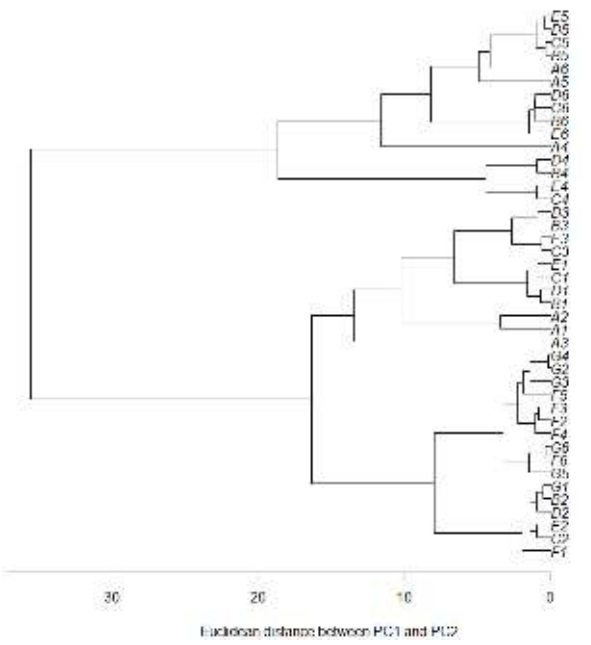
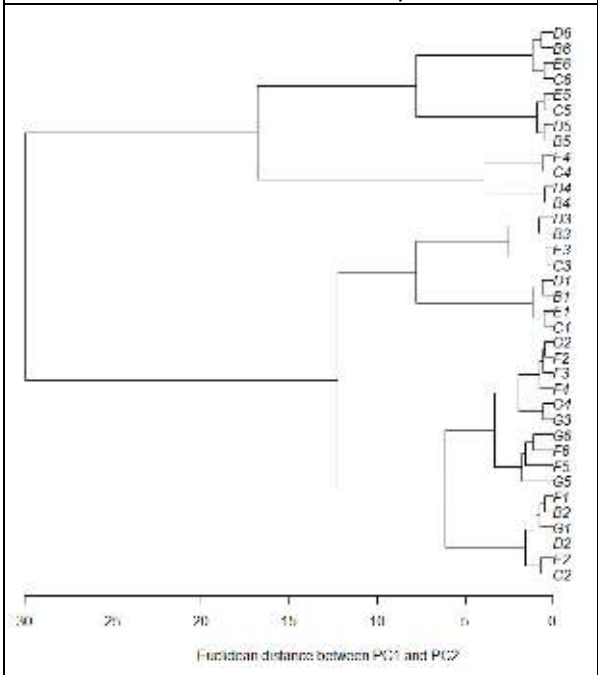
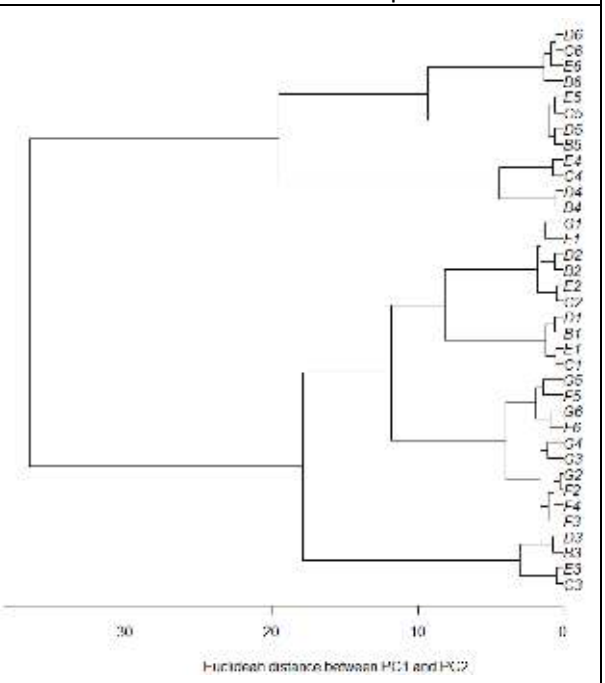
Non-stranded dataset	
Regularized transformation	**Stabilized transformation**
	
FTFF dataset from all replicates (A, B, C, D, E, F, G)	FTFF dataset from all replicates (A, B, C, D, E, F, G)
	
FTFF dataset from B, C, D, E, F and G replicates	FTFF dataset from B, C, D, E, F and G replicates

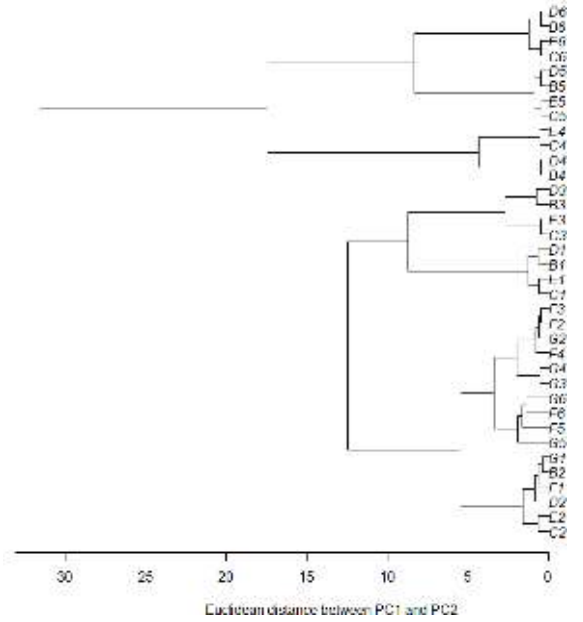
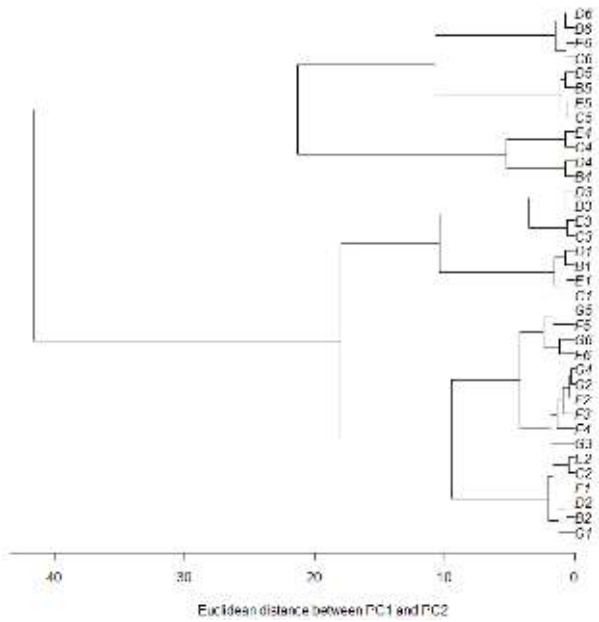
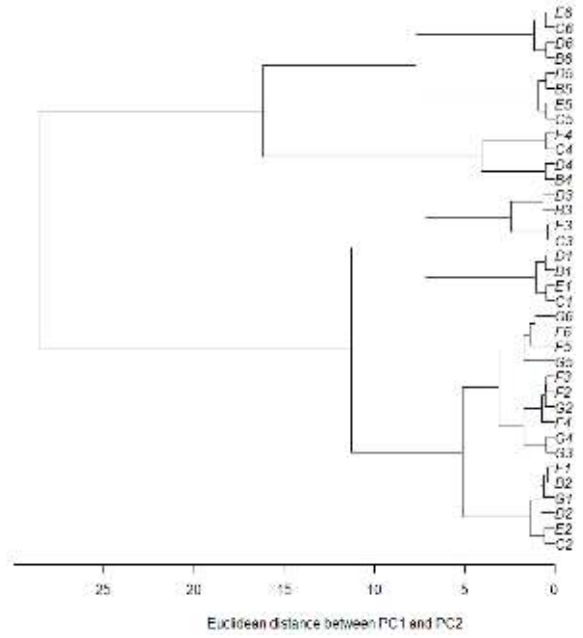
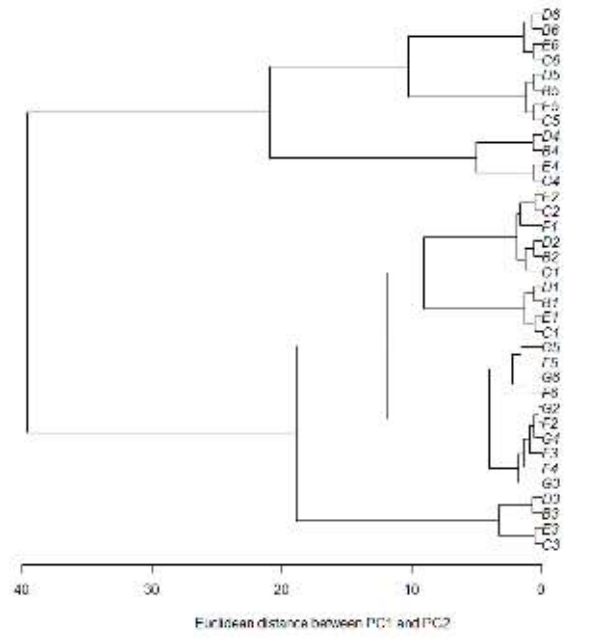
Non-stranded dataset	
Regularized transformation	**Stabilized transformation**
	
FTTT dataset from all replicates	FTTT dataset from all replicates
	
FTTT dataset from B, C, D, E, F and G replicates	FTTT dataset from B, C, D, E, F and G replicates

Reversely-strand dataset	
Regularized transformation	**Stabilized transformation**
	
FTFF dataset from B, C, D, E, F and G replicates	FTFF dataset from B, C, D, E, F and G replicates
	
FTTT dataset from B, C, D, E, F and G replicates	FTTT dataset from B, C, D, E, F and G replicates

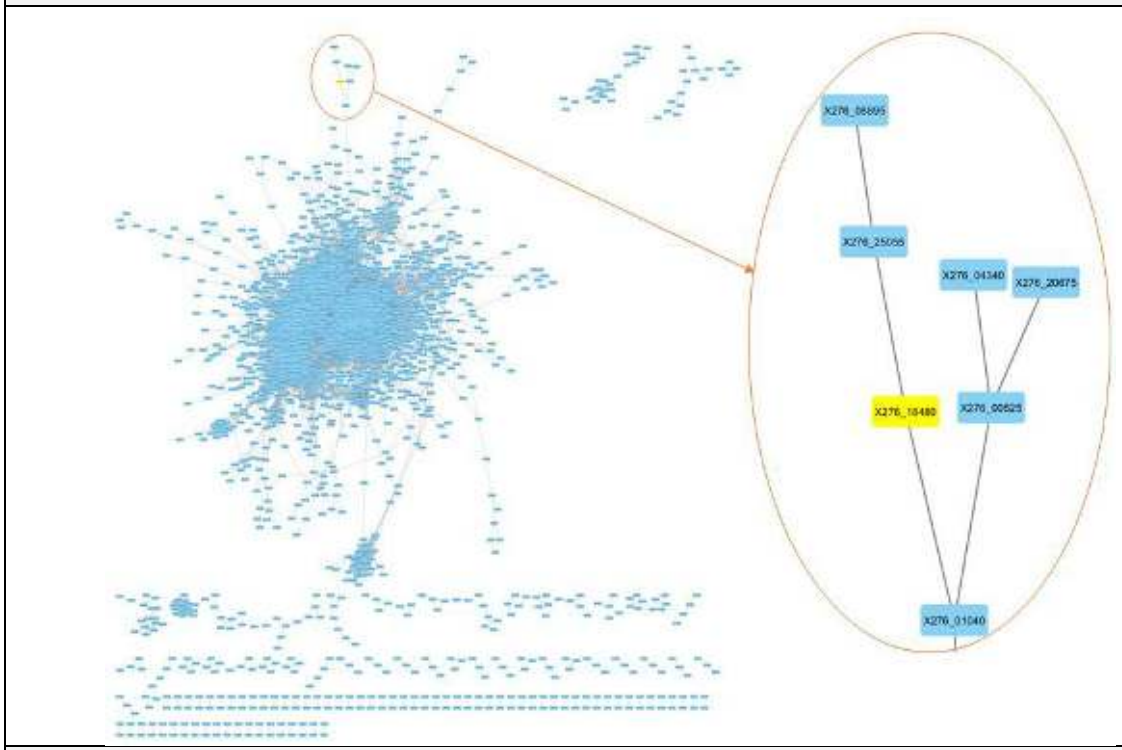
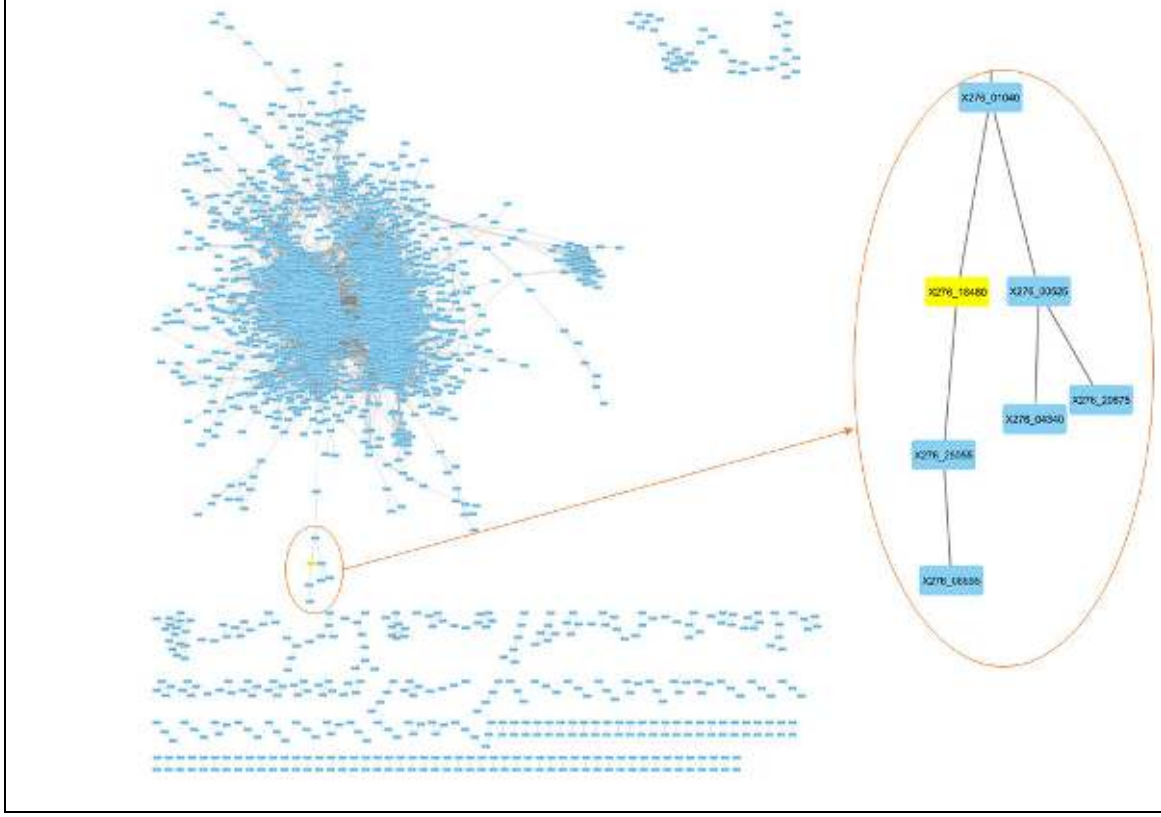
Attachment E] Figures UPGMA plots

<table>
<tr><th colspan="2">Non-stranded dataset</th></tr>
<tr><th>Regularized transformation</th><th>Stabilized transformation</th></tr>
<tr><td> 
FTFF dataset from all replicates (A, B, C, D, E, F, G)</td><td> 
FTFF dataset from all replicates (A, B, C, D, E, F, G)</td></tr>
<tr><td> 
FTFF dataset from B, C, D, E, F and G replicates</td><td> 
FTFF dataset from B, C, D, E, F and G replicates</td></tr>
</table>

Non-stranded dataset	
Regularized transformation	**Stabilized transformation**
	
FTTT dataset from all replicates	FTTT dataset from all replicates
	
FTTT dataset from B, C, D, E, F and G replicates	FTTT dataset from B, C, D, E, F and G replicates

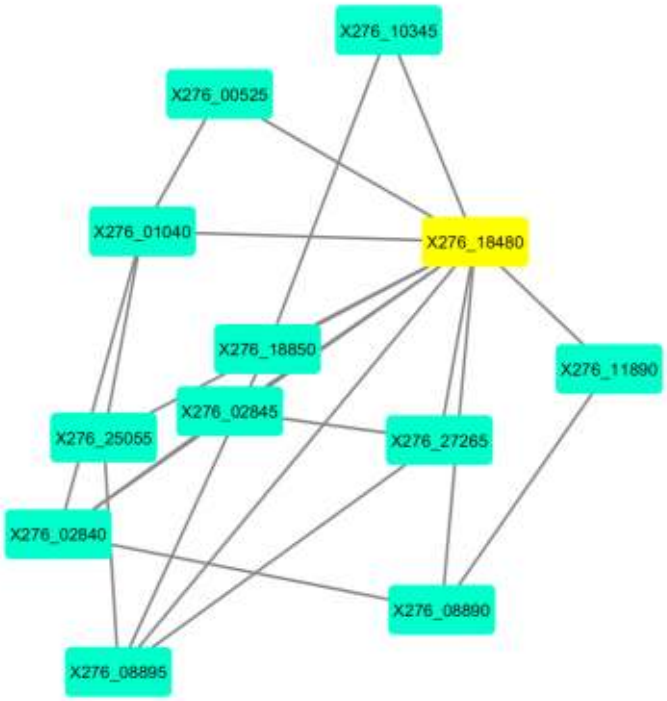
Reversely-strand dataset	
Regularized transformation	**Stabilized transformation**
	
FTFF dataset from B, C, D, E, F and G replicates	FTFF dataset from B, C, D, E, F and G replicates
	
FTTT dataset from B, C, D, E, F and G replicates	FTTT dataset from B, C, D, E, F and G replicates

Attachment F] Co-expression networks

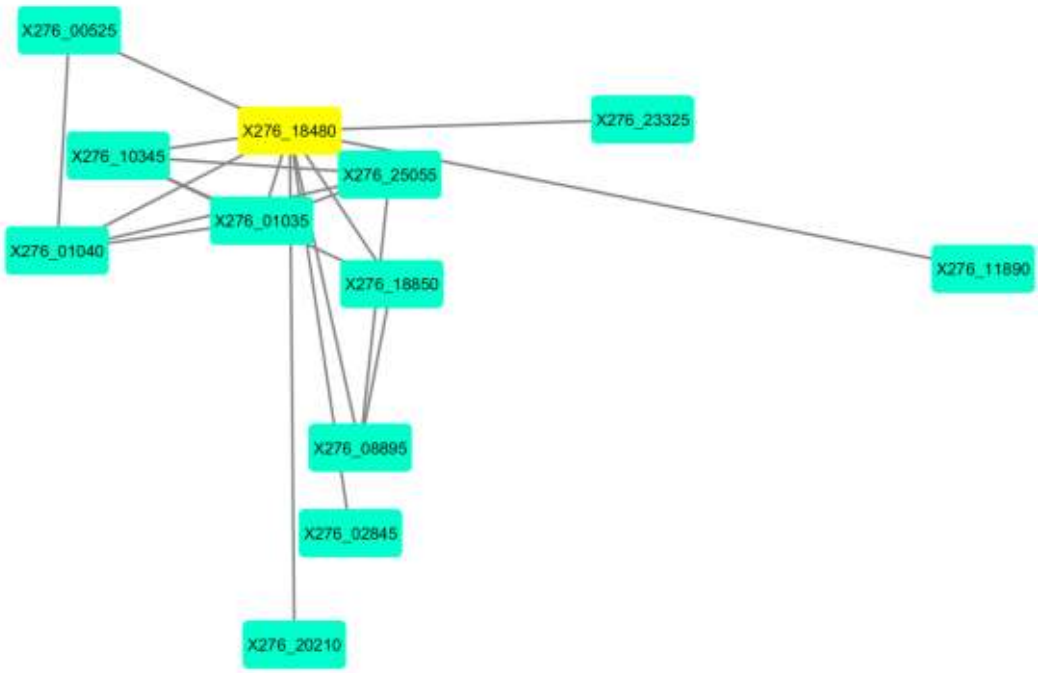
Standard cultivation transcriptome – BCDE dataset
FTFF

FTTT


Attachment G] Bootstrapped-based GRN

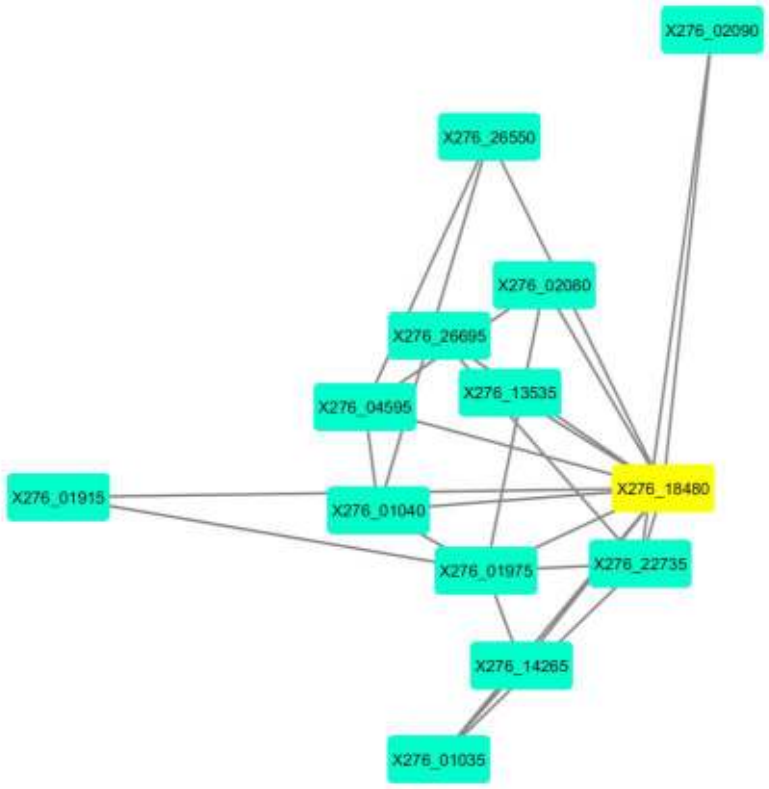
FTFF BCDE – The first neighbourhood of spo0A



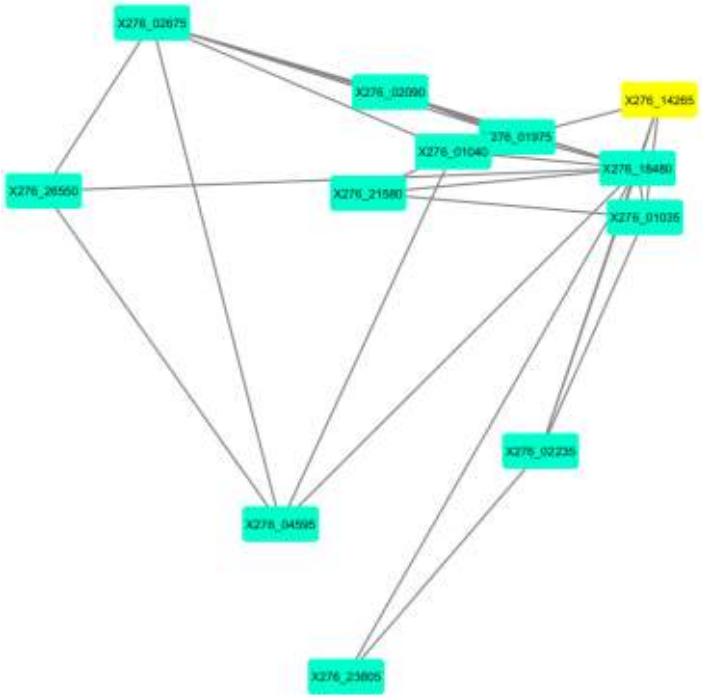
FTTT BCDE – The first neighbourhood of Spo0A



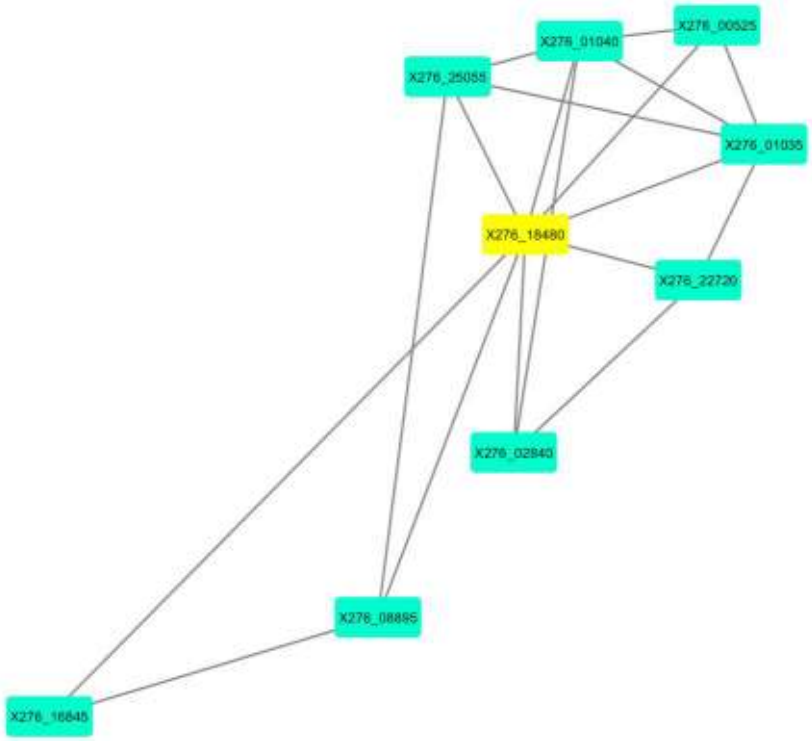
FTFF FG – The first neighbourhood of Spo0A



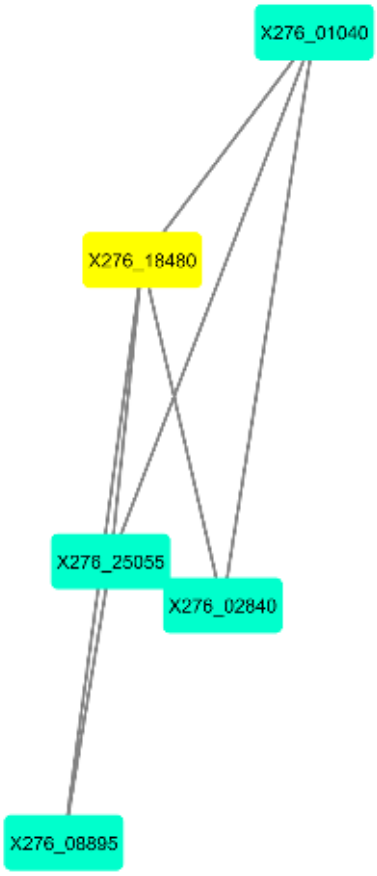
FTTT FG – The first neighbourhood of Spo0A



FTFF BCDEFG – The first neighbourhood of Spo0A

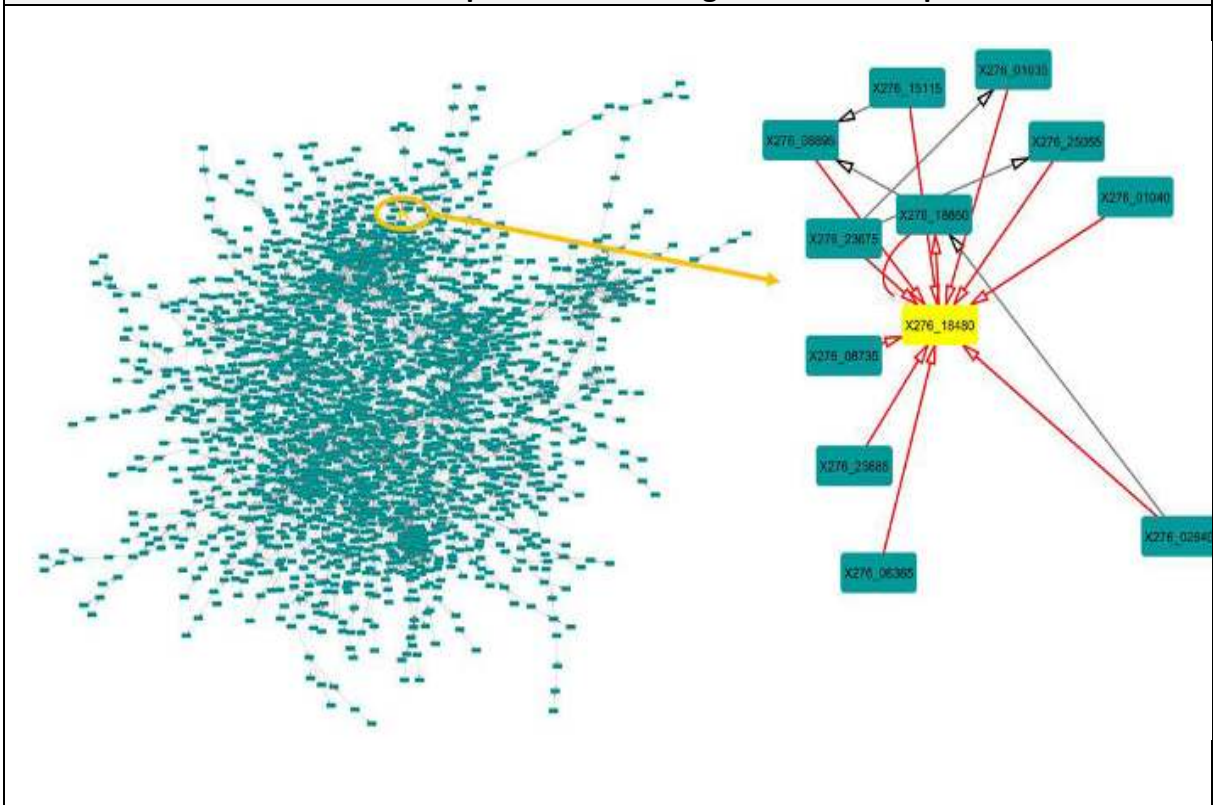


FTTT BCDEFG – The first neighbourhood of Spo0A

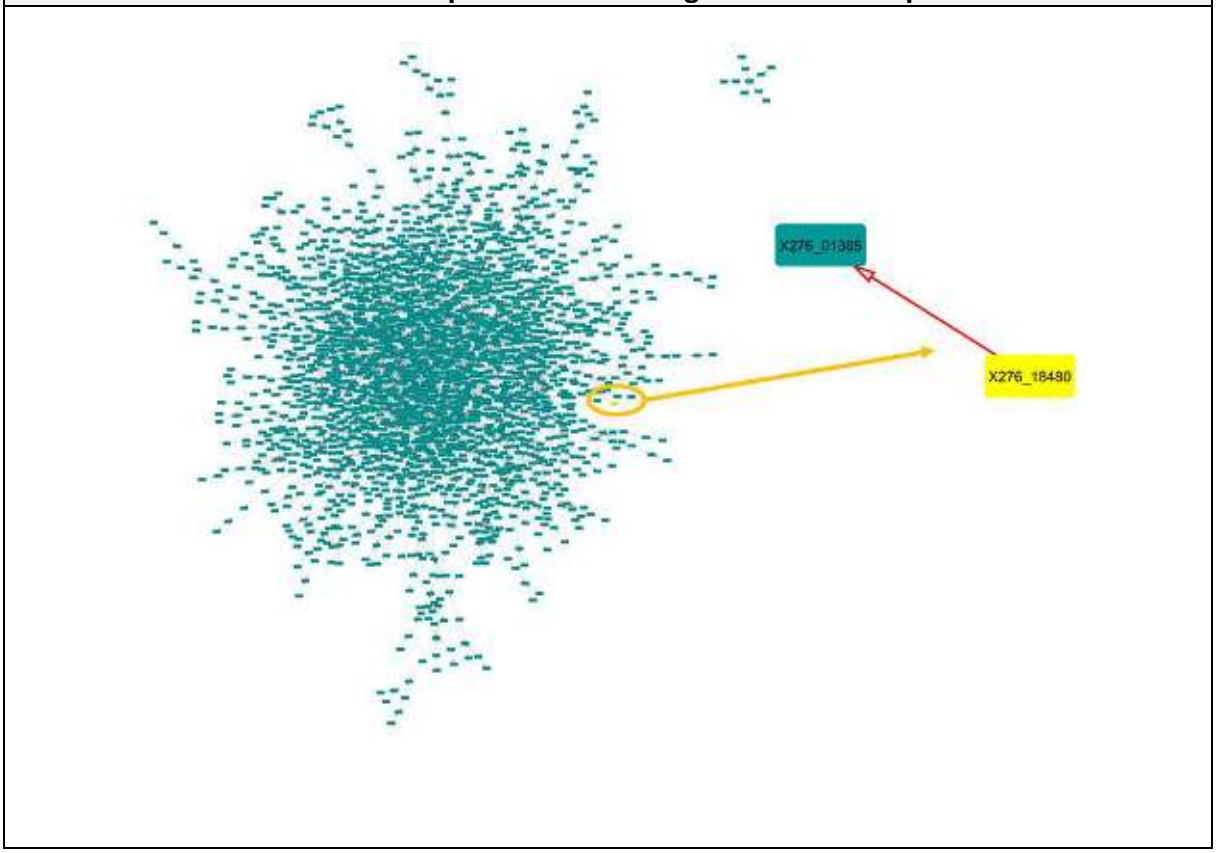


Attachment H] Tree-based GRN

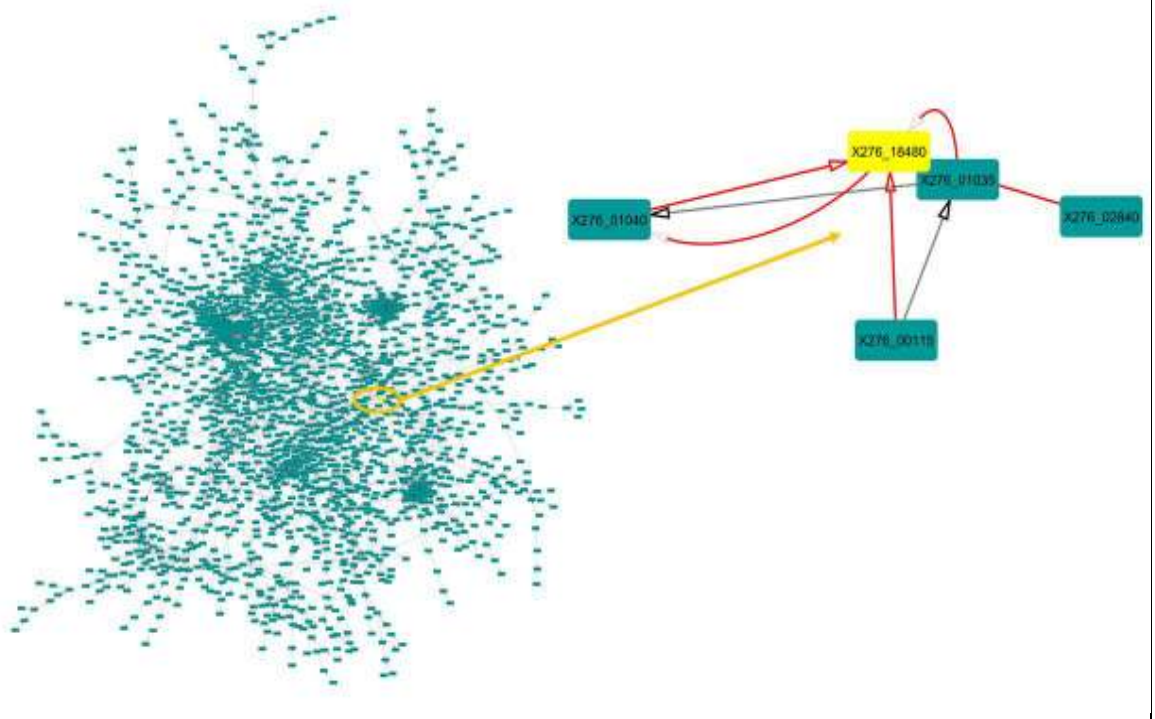
FTTT BCDE – Example of the first neighbourhood of Spo0A



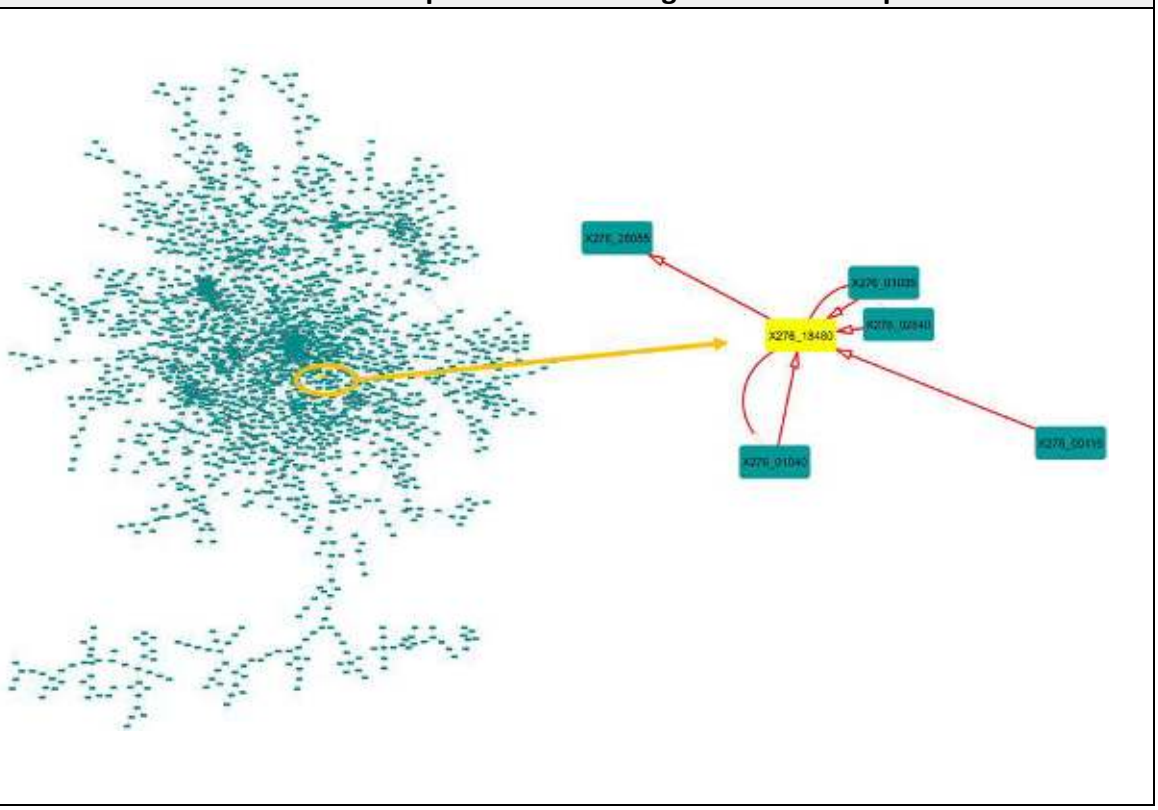
FTTT FG – Example of the first neighbourhood of Spo0A



FTFF BCDEFG – Example of the first neighbourhood of Spo0A



FTTT BCDEFG – Example of the first neighbourhood of Spo0A



Attachment I] Classification of triplet network motifs

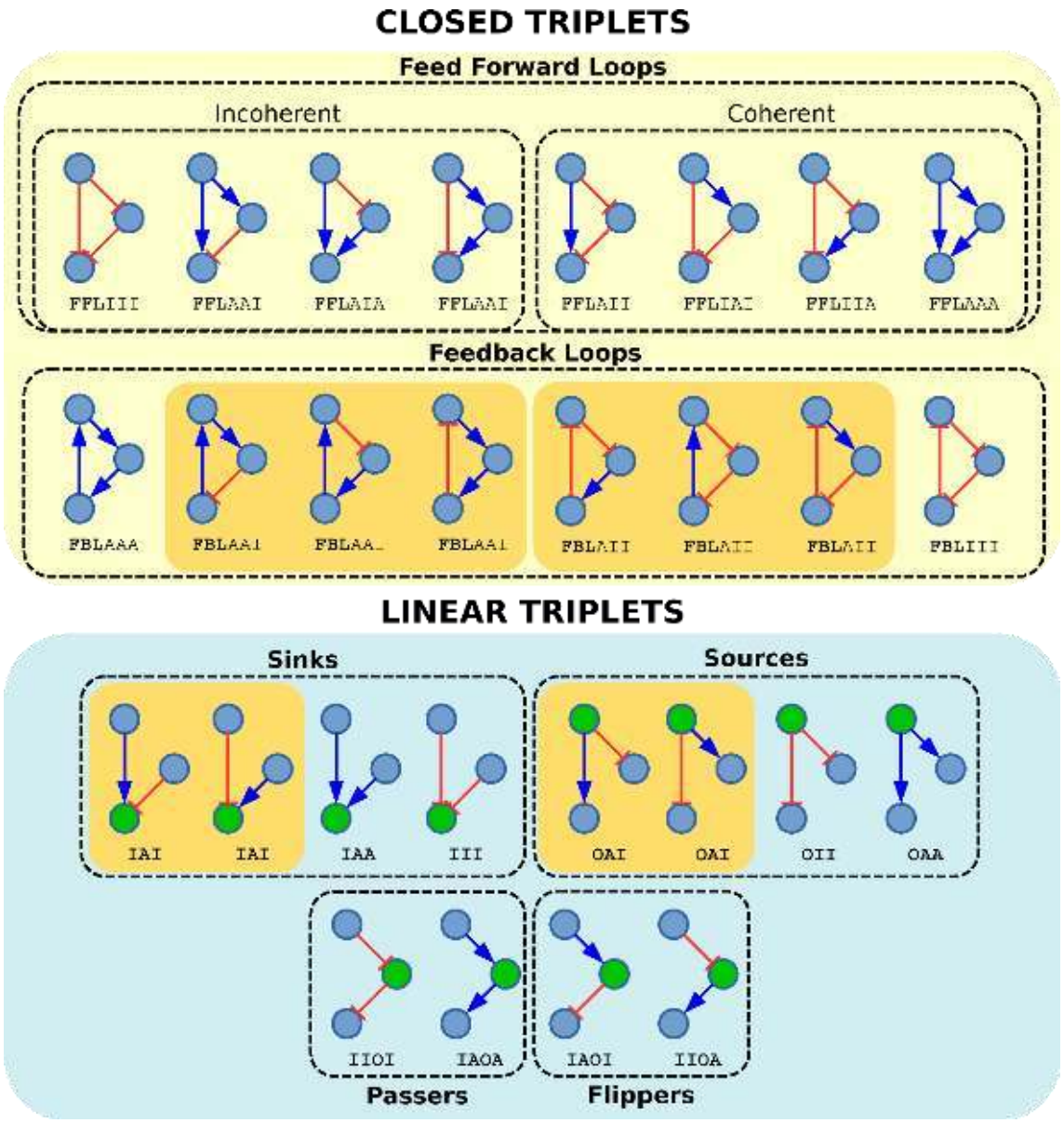
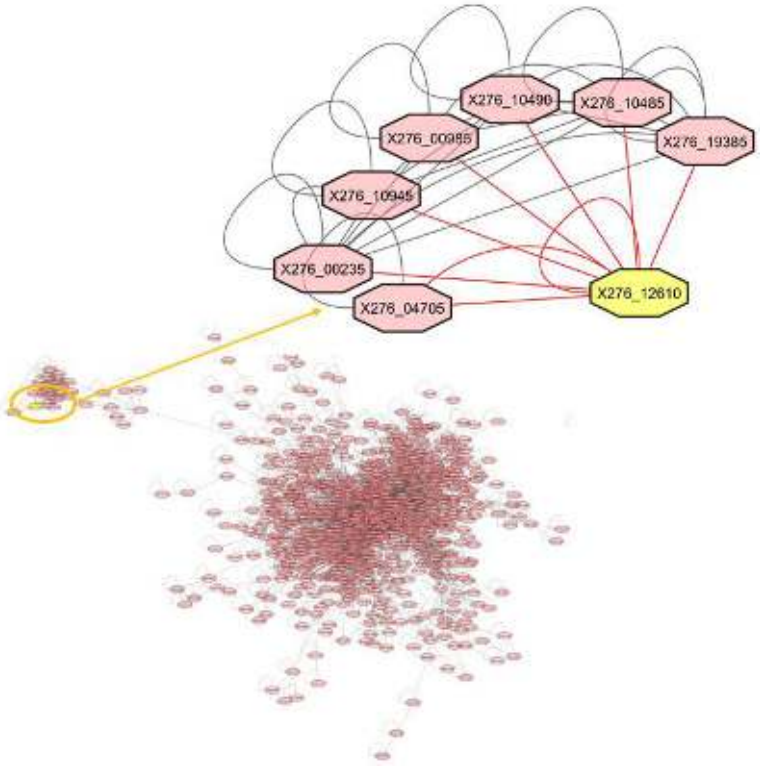


Figure 25: Classification of triplet network motifs [82]

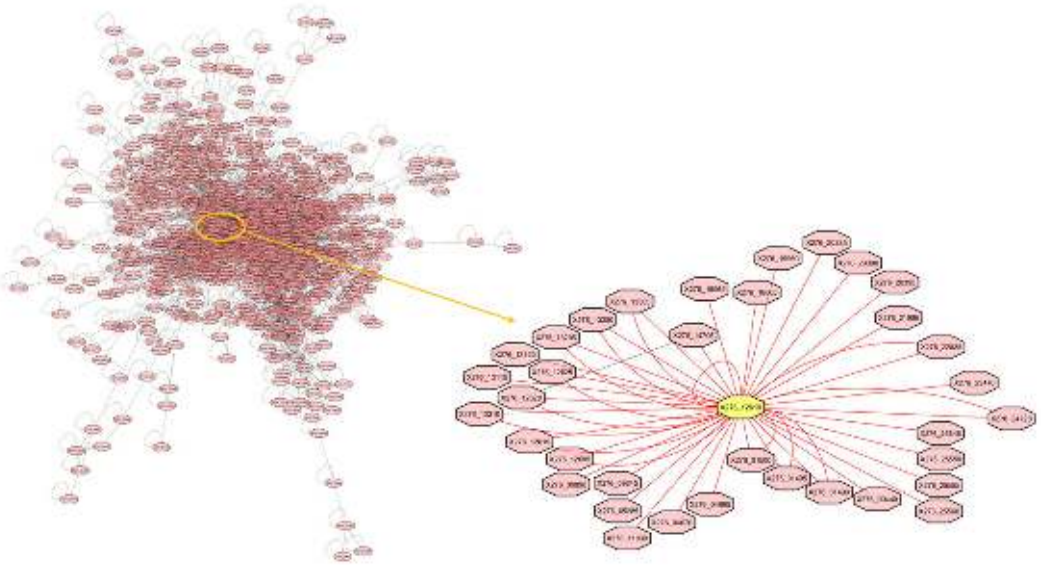
The two main classes are coloured in yellow (Closed Triplets) and blue (Linear Triplets). Motifs highlighted in orange are isomorphism and thus indistinguishable. Incoherent loops are loops where the target node [82]

Attachment J] Interconnected-based GRN

FTFF – Example of the first neighbourhood of X276_12610



FTTT – Example of the first neighbourhood of X276_12610



Attachment K] List of electronic attachments

- Brief description of electronic attachments:
  \Jana_Schwarzerova_MT_attachments\READ_ME.txt
- R.scripts which creates count table for different type of approach created datasets:
  \Jana_Schwarzerova_MT_attachments \ FTFF_&_FTTT
- All scripts which create gene regulatory network and final created network saves as adjacency list "Interconnected_GRN.csv":
  \Jana_Schwarzerova_MT_attachments \ GRN
- Shell scripts and R scripts with approach based on bootstrapping :
  \Jana_Schwarzerova_MT_attachments \ GRN \ Bc3net
- Shell scripts and R scripts with approach based on mutual information and correlation coefficient:
  \Jana_Schwarzerova_MT_attachments \ GRN \ CoexNet
- csv files which using such as input datasets (count tables):
  \Jana_Schwarzerova_MT_attachments \ GRN \ DataSet
- Shell scripts and R scripts with approach based on tree:
  \Jana_Schwarzerova_MT_attachments \ GRN \ GENIE3
- Shell scripts and R scripts with approach based on differential-equation:
  \Jana_Schwarzerova_MT_attachments \ GRN \ tsni
- The processing workflow for obtaining operons:
  \Jana_Schwarzerova_MT_attachments \ Operon
- R script and csv file in step where is add gene express information:
  \Jana_Schwarzerova_MT_attachments \ Operon \ Add_express_information
- Results obtained from online tool OperonMapper:
  \Jana_Schwarzerova_MT_attachments \ Operon \ Operon_mapper
- Transcription from obtained results of OperonMapper to useful format:
  \Jana_Schwarzerova_MT_attachments \ Operon \ R_transcription_to_LocusTag
- Results (such as HTML reports) and shell scripts for whole pre-processing part:
  \Jana_Schwarzerova_MT_attachments \ Pre-processing