THESIS FOR THE DEGREE OF DOCTOR OF PHYLOSOPHY

An insight towards food-related microbial sets through metabolic modelling and functional analysis

SIMONAS MARCIŠAUSKAS

Department of Biology and Biological Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2020

An insight towards food-related microbial sets through metabolic modelling and functional analysis

SIMONAS MARCIŠAUSKAS

ISBN 978-91-7905-276-8
© Simonas Marcišauskas, 2020

Doktoravhandlingar vid Chalmers tekniska högskola
Ny serie nr 4743
ISSN 0346-718X

Division of Systems and Synthetic Biology
Department of Biology and Biological Engineering
Chalmers University of Technology
SE – 412 96 Gothenburg
Sweden
Telephone +46 (0) 31 772 1000

Printed by Chalmers Reproservice
Gothenburg, Sweden 2020

An insight towards food-related microbial sets through metabolic modelling and functional analysis

Simonas Marcišauskas
Department of Biology and Biological Engineering
Chalmers University of Technology

Abstract

The dietary food digestion depends on the human gastrointestinal tract, where host cells and gut microbes mutually interact. This interplay may also mediate host metabolism, as shown by microbial-derived secondary bile acids, needed for receptor signalling. Microbes are also crucial in the production of fermented foods, such as wine and dairy. Kefir is fermented milk processed by the symbiotic community of bacteria and yeasts. One such species is a yeast *Kluyveromyces marxianus*. Its thermotolerance is a desired trait in biotechnology since it may reduce the cooling demands during cultivation.

The systems biology tools allow analysing various size microbial communities under the different functional scope. For example, the homology prediction tools can give detailed functional insights when working with metagenomics data. The whole-cell metabolic processes can be summarised in genome-scale metabolic models (GEMs), which enable to predict the metabolic capabilities and allow for the integration of omics data.

The work shown in this thesis includes i) *in silico* analysis of food-related microbes; ii) the development of GEMs and RAVEN. With a focus on bile acid metabolism, hundreds of human gut microbes were annotated based on metagenomics data, thereby suggesting the differences in the potential for bile acid processing between healthy and diseased subjects. These findings may be exploitable once aiming to restore the bile acid metabolism for the patients having inflammatory bowel disease. Also, the metabolism of yeast *K. marxianus* was characterised in genome-scale. Two *K. marxianus* strains from kefir grains were isolated, sequenced, assembled, and functionally annotated. They were compared with the other ten strains, providing the core and dispensable physiological features for *K. marxianus*. Furthermore, the first GEM for *K. marxianus*, namely iSM996, was reconstructed. It was integrated with transcriptomics data to predict its metabolic capabilities in rich medium and high-temperature conditions. The results might be useful to optimise strain-specific medium for high-temperature applications. The final paper comprises the efforts to improve the usability for RAVEN, a toolbox for GEM reconstruction and analysis. Altogether the outcomes of this thesis suggest the potential applications for medicine and industrial biotechnology, which may be facilitated by the newly upgraded RAVEN toolbox.

Keywords: bile acids, comparative genomics, genome-scale metabolic model, gut, *Kluyveromyces marxianus*, next-generation sequencing, RAVEN, systems biology, thermotolerance, transcriptomics

List of publications

This thesis is based on the work contained in the following papers:

Paper I: *Metagenomic analysis of bile salt biotransformation in the human gut microbiome*. Promi Das, Simonas Marcišauskas, Boyang Ji, Jens Nielsen. BMC Genomics. 2019; 20:1–12.

Paper II: *The functional diversity between Kluyveromyces marxianus strains*. Simonas Marcišauskas, Yongkyu Kim, Sonja Blasche, Boyang Ji, Kiran Raosaheb Patil, Jens Nielsen [manuscript].

Paper III: *Reconstruction and analysis of a Kluyveromyces marxianus genome-scale metabolic model*. Simonas Marcišauskas, Boyang Ji, Jens Nielsen. BMC Bioinformatics. 2019; 20:551.

Paper IV: *RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor*. Hao Wang*, Simonas Marcišauskas*, Benjamín J. Sánchez, Iván Domenzain, Daniel Hermansson, Rasmus Agren, Jens Nielsen, Eduard J. Kerkhoven. PLoS Computational Biology. 2018;14.

Additional papers not included in the thesis:

Paper V: *A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism*. Hongzhong Lu, Feiran Li, Benjamín J. Sánchez, Zhengming Zhu, Gang Li, Iván Domenzain, Simonas Marcišauskas, Petre Mihail Anton, Dimitra Lappa, Christian Lieven, Moritz Emanuel Beber, Nikolaus Sonnenschein, Eduard J. Kerkhoven, Jens Nielsen. Nature Communications. 2019;10.

*Contributed equally

Contribution summary

Paper I: contributed to study design, contributed to reference dataset creation, contributed to metagenomics data analysis, contributed to writing and editing of the manuscript

Paper II: designed study, analysed data and wrote the manuscript

Paper III: designed study, analysed data and wrote the manuscript

Paper IV: updated RAVEN functions for the model import/export, KEGG reconstruction module, RAVEN-COBRA wrapper and other 20 functions; contributed to writing and editing of the manuscript

Preface

This dissertation serves as partial fulfilment of the requirements to obtain the degree of Doctor of Philosophy at the Department of Biology and Biological Engineering at Chalmers University of Technology. The PhD studies were carried out between February 2015 and February 2020 at the division of Systems and Synthetic Biology (SysBio) under the supervision of Jens Nielsen. The research was co-supervised by Boyang Ji and examined by Stefan Hohmann. It was funded by European Commission FP7-HEALTH project METACARDIS, ERASysAPP projects SysMilk and SYSTERACT (the latter provided by Västra Götalandsregionen), Novo Nordisk Foundation and the Knut and Alice Wallenberg Foundation.

Simonas Marcišauskas

September 2020

Table of contents

Introduction **1**

Background **3**

The human gut microbiome 3

Bile acid metabolism and its role in host physiology 5

Kefir and its microbial community 7

Kluyveromyces marxianus: a promising cell factory with controversial traits 10

Computational tools for species annotation and analysis 11

Part I: Comparative functional analysis of bile acid metabolism **17**

Paper I: Metagenomic study of bile acid biotransformation 17

Part II: *In silico* genomics analysis for yeast *Kluyveromyces marxianus* **23**

Paper II: Comparative genomics of 12 *K. marxianus* strains 23

Paper III: Reconstruction and analysis of *K. marxianus* GEM 28

Part III: RAVEN 2.0, a toolbox for GEM reconstruction and analysis **39**

Paper IV: Development of The RAVEN Toolbox 39

Conclusions and perspectives **43**

Acknowledgements **45**

References **47**

Abbreviations

ACP	Acyl carrier protein
ADP	Adenosine diphosphate
ATP	Adenosine triphosphate
BA	Bile acid
Bai	Bile acid-inducible
bp	Base pair
BS	Bile salt
BSBG	Bile acid biotransformation gene
BSBP	Bile acid biotransformation protein
CD	Crohn’s disease
CDS	A CoDing Sequence
CFU	Colony-forming unit
CoA	Coenzyme A
DNA	Deoxyribonucleic acid
EC	Enzyme Commission
FBA	Flux balance analysis
GAM	Growth-associated maintenance
gDW	Gram cell dry weight
GEM	Genome-scale metabolic model
GIT	Gastrointestinal tract
GPR	Gene-protein-reaction
GTP	Guanosine-5'-triphosphate
FAD	Flavin adenine dinucleotide
FMN	Flavin mononucleotide
IBD	Inflammatory bowel disease
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG Orthology
KOG	EuKaryotic Orthologous Group
NCBI	National Center for Biology Information
NGAM	Non-growth associated maintenance
NGS	Next-generation sequencing
P/O ratio	Phosphate/oxygen ratio
pH	Power of hydrogen
rRNA	Ribosomal ribonucleic acid
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
ORF	Open reading frame
SBML	Systems Biology Markup Language
UC	Ulcerative colitis
tRNA	Transfer ribonucleic acid
TSS-Seq	Transcription start site sequencing
YPD	Yeast extract peptone dextrose
YPX	Yeast extract peptone xylose

“It's not enough that you believe what you see. You must also understand what you see.”

- Leonardo da Vinci

Introduction

Nature presents numerous living forms that successfully adapt to their niche environments. Such cases are driven by specific capabilities of individual species or by the collective efforts of biological communities (Merino *et al.* 2019). Upon suitable environmental factors, the survival strategy requires an efficient assimilation system for essential nutrients. Such a system relies on the physiological capabilities for the species of interest and its co-inhabitants. In such an environment, a unicellular organism is an individual competitor which has a standalone ability to obtain the required metabolites, including their sensing, extracellular pre-processing, transporting into the cell and intracellular bioprocessing. Meanwhile, the higher (multicellular) eukaryotes have more complex environments for nutrient uptake. For instance, the human gut microbiome is a result of a symbiosis between the host cells and a high number of microbial species (Cani 2018). Although much effort has been made to annotate the adaptation patterns, there are still many microbial communities and individual species which are yet to be characterised. A better understanding of these novel organisms may reveal new candidates for the microbial cell factories or contribute to the higher quality of healthcare.

A variety of high-throughput experimental techniques, such as next-generation sequencing, metagenomic profiling, mass spectrometry and flow cytometry, can provide a detailed overview of individual microbes or microbial communities. One can utilise bioinformatics tools to pre-process and analyse experimental data, but such approaches are often limited to particular data types. This issue can be resolved by systems biology approaches, which enable the integration of several data types, thereby allowing to draw more solid insights or identify the complementary experiments required for further investigation. Genome-scale metabolic models (GEMs) are the systems biology platforms to predict the metabolic capabilities and allowing the integration of multi-omics data. However, many published GEMs have compatibility problems due to poor standardisation and lack of curation after release. While the functionality of these GEMs depends on their authors, the tools used for the GEM reconstruction and analysis may prevent the newly developed models from incompatibility problems.

This thesis comprises several functional level bioinformatic analyses for microbial species, which are in some way related to the human digestive system where the food breakdown takes place. The human gut is continuously modulated by ingested food through its microbial and nutritional content. Fermented foods are known with a positive impact on the human gut microbiome due to their probiotic and prebiotic properties, determined by their microbial and nutritional composition correspondingly. The species considered in this work inhabit the human gastrointestinal tract or contribute to the production of kefir, a fermented milk beverage acknowledged of its probiotic and prebiotic attributes. Also, the thesis aims to reduce GEM compatibility problems by an updated toolbox for GEM modelling.

Firstly, a metagenomic study was conducted for a cohort of gut microbes capable of modulating bile acid metabolism, which is known to facilitate the dietary fat uptake. A comparative approach was sought to estimate bile acid biotransformation potential between the groups of healthy and patients experiencing dysbiosis due to inflammatory bowel disease. This work aims

to comprehend the differences in bile acid metabolism between these groups and provide strategies to restore bile acid metabolism in the patients.

Secondly, an effort has been made to characterise a non-conventional yeast *Kluyveromyces marxianus*, found in kefir culture. Its thermotolerance, high growth rate and ability to metabolise a more extensive selection of substrates makes it a promising candidate of microbial cell factory. A functional comparison was performed for a dozen of *K. marxianus* strains followed by reconstruction of a consensus GEM for this species. Pangenome identification and investigation of the metabolic bottlenecks upon stress conditions serve as a basis for the further *K. marxianus* research and optimisation for industrial exploit.

Finally, the thesis presents RAVEN 2, an updated toolbox for metabolic modelling, which aims to facilitate GEM reconstruction, curation and analysis.

Background

The human gut microbiome

One can describe the human gastrointestinal tract (GIT) as a passageway where the food digestion, absorption, and excretion take place. Human GIT consists of the several specialised organs like mouth, stomach, small intestine and large intestine, which form the human digestive system together with other accessory organs, including the tongue, salivary glands, pancreas, liver and gallbladder (Cheng *et al.* 2010). The human GIT is densely colonised by microorganisms, including bacteria, archaea, eukaryotes and viruses. Known as the human gut microbiome (Thursby & Juge 2017), this microbial community is prevailed (97%) by bacterial phyla Firmicutes, Bacteroidetes, Proteobacteria and Actinobacteria (Rosenbaum *et al.* 2015). Although the human gut microbiota utilises food nutrients to ensure its viability, it significantly complements to the metabolic capabilities of the host. Such examples include energy harvest (den Besten *et al.* 2013), vitamin production (Conly & Stein 1992), gut homeostasis (Natividad & Verdu 2013), immune system maturation and modulation (Bäumler & Sperandio 2016; Gensollen *et al.* 2016). For instance, at the phylum level, Bacteroidetes degrade polysaccharides and produce acetate (Xu *et al.* 2007) while Firmicutes use the latter as a substrate to produce butyrate (Louis & Flint 2017), the primary energy source for colonocytes (Litvak *et al.* 2018).

Human GIT segments vary in respect of physiochemical factors like the power of hydrogen (pH) and the levels of digestive enzymes, bile salts and hydrochloric acid (Savage 1977). Such variations have a substantial impact on the local microbial density. The stomach and the part of small intestine comprise the upper GIT, which is responsible for the food digestion and absorption. For instance, the absorbed nutrients in the upper GIT are amino acids, lipids, carbohydrates and vitamins. While the stomach has a low pH due to the high level of hydrochloric acid, the small intestine has a neutral pH and is enriched with the digestive enzymes and bile acids. Such dynamic conditions affect the microbial colonisation, which is relatively low density in the upper GIT, having approximately 10^4-10^6 colony-forming units per millilitre (CFU/mL) (Bik *et al.* 2006; Zoetendal *et al.* 2012). Meanwhile, the lower GIT, comprising small and large intestine, has a higher food-derived nutrient availability. These conditions are much more favourable for the gut microbiome as shown by the significantly higher microbial density, being as high as 10^{11}-10^{12} CFU/mL (Claesson *et al.* 2011). Some microbial species from the lower GIT can produce vitamins, e.g. vitamin K2, which are subsequently absorbed by the host (Conly & Stein 1992).

Human gut colonisation begins during the infant age and continues for two years when the microbial composition reaches a similar state as in adults (Korpela & de Vos 2018). The gut microbiome is a complex, personal and highly dynamic ecosystem that is shaped by diet, probiotics, the host lifestyle and immune system, diseases and usage of antibiotics (Figure 1).

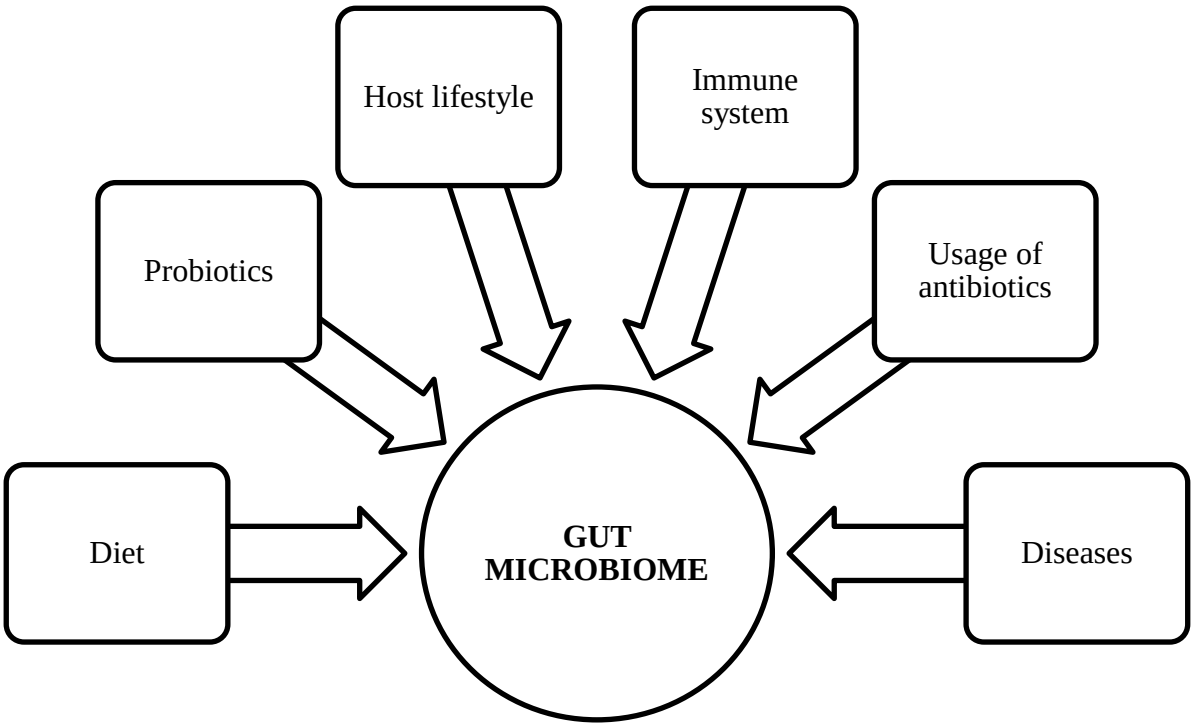


Figure 1. Schematic illustration of the critical factors which shape the human got microbiome. Diseases include neurological disorders, diabetes, inflammatory bowel disease, obesity and cystic fibrosis. Adapted from (Issa Isaac *et al.* 2019)

The primary factor which shapes the gut microbial composition is diet because the differences in nutrients may promote the growth for different microbial species. For example, higher fibre availability promotes gut microbial diversity and short-chain fatty acid production (Holscher 2017). Besides nutrients, the food may also contain probiotics, the bacterial species which have a positive impact on gut microbiome development. Lactobacilli and Bifidobacteria are probiotic species, which survive during the ingestion through the stomach and help to reduce the symptoms for diseases, like inflammatory bowel syndrome, and have the beneficial effects on host immune system (Khani *et al.* 2012). In addition to dietary factors, lifestyle impact is also significant. Previous studies showed that smoking might contribute to increased Bacteroides-Prevotella part in individuals (Lutgendorff *et al.* 2008). Moreover, smoking and low physical activity were associated with altered microbial composition in the colon (Huxley *et al.* 2009). And stress has an impact on colonic motor activity, leading to modified gut microbial composition and lower levels of Lactobacillus (Conlon & Bird 2014). There were also studies showing that antibiotics usage can result in microbial dysbiosis and thereby contribute to diseases like diabetes, obesity, asthma, rheumatoid arthritis, autism and inflammatory bowel disease (Zhang & Chen 2019). Naturally, diseases are also important factors affecting the microbial composition in the human gut in causing obesity (Ley *et al.* 2005) and cystic fibrosis (Burke *et al.* 2017).

Bile acid metabolism and its role in host physiology

Bile acids (BAs) are sterols which facilitate digestion and absorption of fats and fat-soluble vitamins in GIT. Primary BAs are synthesised in the liver and then conjugated with glycine or taurine. These conjugated are also known as bile salts (BSs), which are accumulated in the gallbladder and at the required time secreted into the duodenum, the first section of the small intestine (Figure 2). In humans, the most common BAs are cholate (CA) and chenodeoxycholate (CDCA).

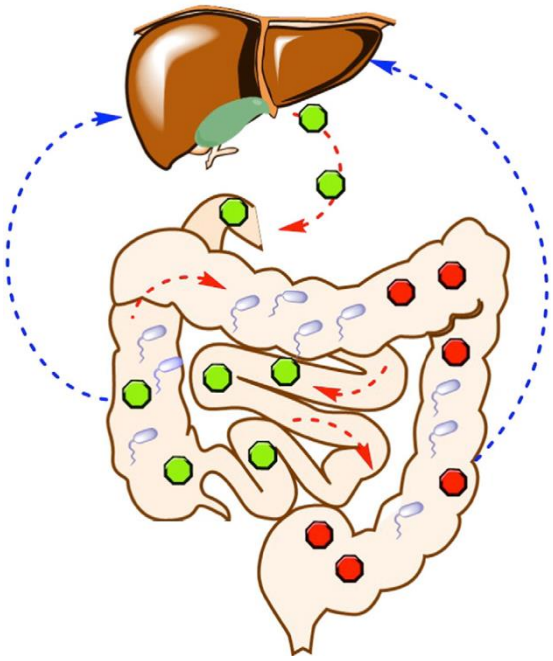


Figure 2. The relationship between gut microbiota and liver by enterohepatic circulation. Firstly, primary bile salts are produced in the liver and secreted into the gut, where they play an essential role in regulating the gut microbial composition. Secondly, some primary bile salts are functionally converted by gut bacteria into secondary bile acids, then reabsorbed into the circulation system and in the liver converted back to the primary bile salts. Circles filled with green colour show primary bile salts while secondary bile acids are shown as red-filled circles. Adapted from (Li *et al.* 2016)

When compared with BAs, BSs are less hydrophobic and more soluble in the small intestine. They, therefore, can act as emulsifiers by providing access for lipases to perform fat digestion. The vast majority (95%) of primary BAs are absorbed back via enterohepatic circulation pathway in the ileum, the third and the final section of the small intestine. Meanwhile, the remaining 5% are bio-transformed into secondary BAs or excreted to faeces (Mullish *et al.* 2018). Secondary BAs are BS derivatives having fewer functional groups and obtained through dehydroxylation, dehydrogenation or epimerisation (Dawson & Karpen 2015). An example of such a biotransformation process is shown in Figure 3.

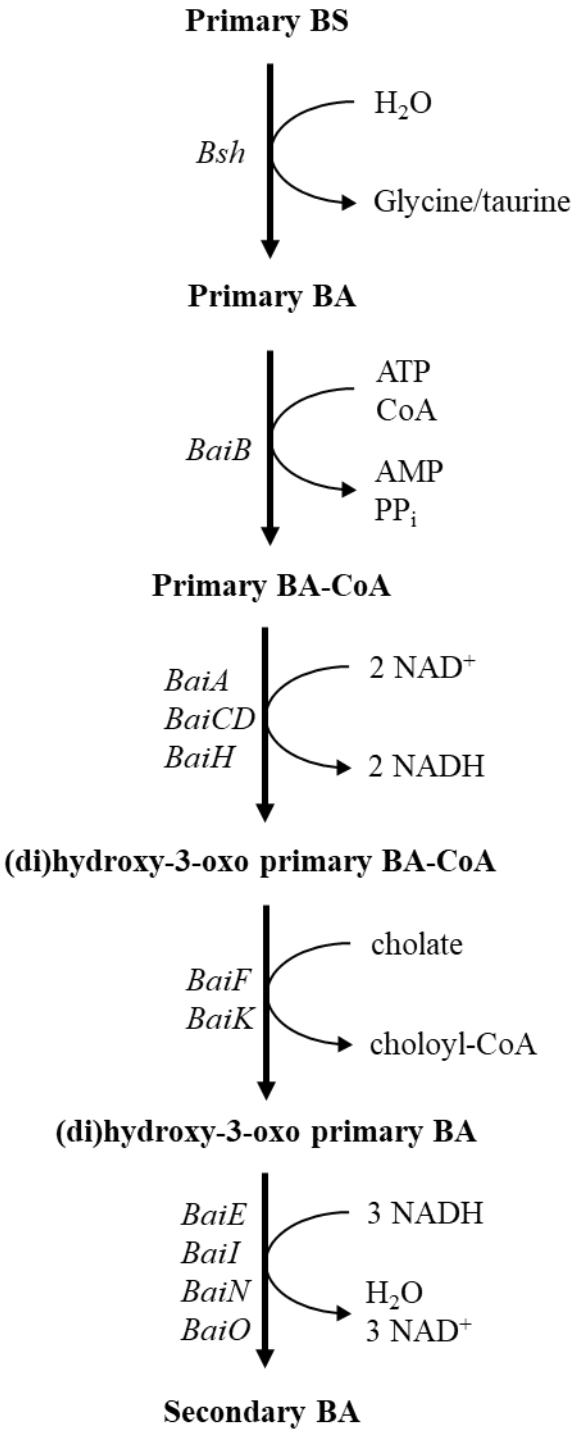


Figure 3. An overall scheme of primary bile salt (BS) biotransformation process in the large intestine. It starts with primary BS deconjugation, done by bacterial extracellular BS hydrolase (Bsh). Primary bile acids (BAs) are then transported into bacterial cells using BaiG membrane transporter, which is not shown in the scheme. The following intracellular modifications include BA conjugation with coenzyme A (CoA) followed by its oxidation. In the next step, CoA is transferred from BA to cholate. The last biotransformation steps are three reduction steps. Key intermediate metabolites are denoted in bold, the associated gene names are indicated in italics, and the other reactants/products are meant in regular font style.

The scheme suggests that BS biotransformation is beneficial to gut bacteria by providing the ability to ensure the redox balance. Although secondary BAs are more hydrophobic and therefore more toxic to gut microbiome and the host than primary ones, they mediate the host metabolic pathways using receptor signalling, including farnesoid X receptor (FXR), the liver X receptor (LXR), the G-protein coupled bile acid receptor TGR5 and vitamin D receptor (VDR) (de Aguiar Vallim *et al.* 2013). Most secondary BAs and BSs are reabsorbed by the host and transported to the liver, where they are converted back to the primary BSs.

Kefir and its microbial community

Human beings have utilised milk fermentation for thousands of years. This technology offers various beneficial properties to dairy products, such as extended shelf life, higher digestibility, enriched flavour and nutritional composition. Although fermented milk has been spontaneously made with empirical cultures for centuries, the research pioneered by Pasteur and Metchnikoff allowed to decipher the fermentation process and standardise the production for several dairy products (Kroger *et al.* 1992). Microbial fermentation starter cultures depend on the target dairy product, and lactic acid bacteria (LAB), including genera of *Lactobacillus*, *Lactococcus* and *Leuconostoc*, are always involved in the fermentation process of milk and other beverages (Oberman & Libudzisz 1998). However, LAB are seldom accompanied by acetic acid bacteria (AAB), yeasts or moulds.

Kefir is a fermented milk drink obtained using a mesophilic symbiotic culture of bacteria and yeasts (SCOBY) as inoculum. It is commonly produced from cow's, sheep's or goat's milk. A kefir SCOBY, also known as kefir grains, is a microbial cauliflower-like structure comprising a matrix of polysaccharides, proteins and lipids which embed microbes into a biofilm (Ninane *et al.* 2005). The most common microbial species appearing in kefir grains are LAB (e.g. genera *Lactobacillus*, *Lactococcus*, *Leuconostoc*, *Bifidobacterium*), AAB (e.g. genera *Acetobacter*, *Gluconobacter*) and yeasts (e.g. *Kluyveomyces marxianus*, *Saccharomyces cerevisiae*, *Kazachstania unispora*) (Prado *et al.* 2015). The size of kefir grains grow in each milk fermentation cycle (Leite *et al.* 2013; Lopitz-Otsoa *et al.* 2006). Kefir is generally produced overnight at room temperature. Its production process is illustrated in Figure 4.

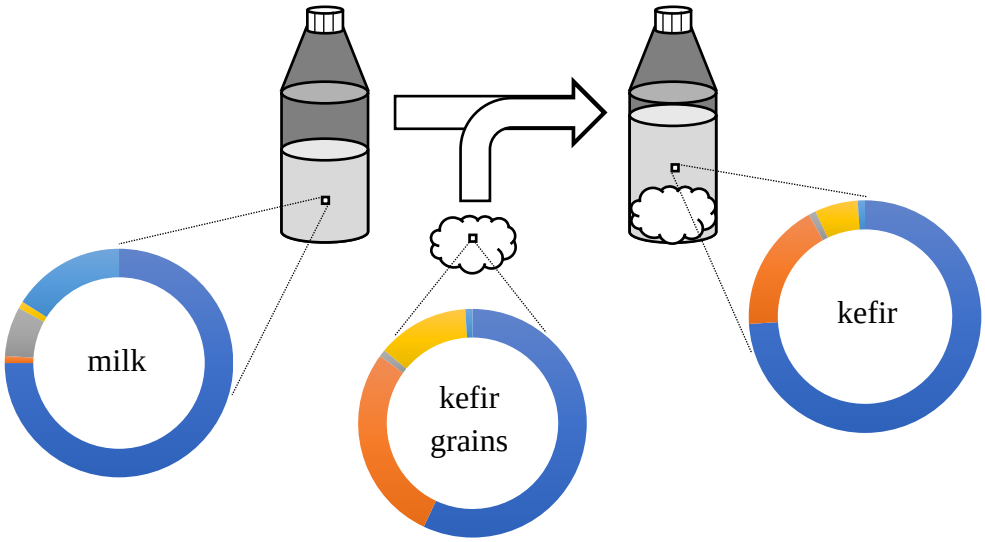


Figure 4. Kefir production scheme and the microbial composition of cow's milk, kefir grains and kefir obtained from cow's milk. The colours in the circle graphs denote species in the following categories: (i) lactic acid bacteria (LAB): *Streptococcus* spp. (blue), *Lactobacillus* spp. (orange), *Leuconostoc* spp. (grey); (ii) yeasts (yellow); (iii) other species (light blue). No composition data were available in (i) milk for *Lactococcus* spp. and yeasts; (ii) kefir grains and kefir for *Leuconostoc* spp. and other species. It was therefore assumed that in all such cases the abundance was equal to 1%. Figure generated using data from (Quigley *et al.* 2013; Simova *et al.* 2002).

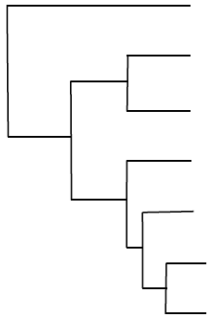
Like milk, kefir grains are mostly populated by LAB, but their genus level composition differs. Although *Streptococcus* spp. is the most dominant LAB in both environments, kefir grains contain more bacterial species from genus *Lactobacillus* than *Leuconostoc*. Kefir grains have also been reported to have a higher fraction of yeasts and a lower fraction of other species than milk. A kefir SCOBY is very stable and robust against contaminant species, mainly from *Pseudomonas* and *Escherichia* genera. Due to rapid acidification and kefir SCOBY species spread to milk, the kefir microbiological composition becomes more like kefir grains than milk. The nutritional profile of kefir depends on milk type and fermentation conditions, but when compared with unprocessed dairy, it is usually enriched with free amino acids, lipids (i.e. acylglycerols), carbohydrates (i.e. glucose, galactose), free fatty acids, vitamins (i.e. A, B, C, K), minerals (i.e. magnesium, calcium, phosphorus), ammonium, carbon dioxide, ethanol, acetate, biogenic amines (i.e. cadaverine, putrescine, spermidine) and flavour compounds (i.e. lactate, acetate, pyruvate, butyrate, diacetyl, acetaldehyde) (Rosa *et al.* 2017). Such a wide variety of nutrients is achieved by kefir microbial culture, whose major functional groups are summarised in Table 1.

Table 1. The major microbial functional groups of kefir culture.

Microbial Group	Function	Result	References
Homofermentative LAB	Ferment lactose into lactate	A lower, slightly acidic pH, preventing the growth of pathogens	(Issa & Tahergorabi 2019)
Lactobacillus kefiranofaciens	Produces kefiran	Facilitates the growth of kefir grains	(Zajšek *et al.* 2011)
Heterofermentative LAB	Ferment lactose into lactate, ethanol and CO_2	Antimicrobial activity for LAB due to CO_2	(Vardjan *et al.* 2013)
Leuconostoc mesenteroides, *Lactococcus lactis*	Ferment citrate into aroma compounds diacetyl and acetoin	Kefir with enriched flavour	(Leite *et al.* 2013)
AAB	Ferment lactose to acetate, produce vitamin B	Provide vitamins needed for other species to grow	(Rea *et al.* 1996)
Propionibacteria	Ferment lactose into propionate, ethanol and CO_2	Kefir enriched with propionate, which contributes to the lower cholesterol level once consumed	(Issa & Tahergorabi 2019)
Fungi	Hydrolyse milk proteins into amino acids, breakdown milk fats into fatty acids, synthesise complex B vitamins, and aroma compounds	Provide nutrients for other species to grow, kefir with enriched flavour	(Demir 2020; Lopitz-Otsoa *et al.* 2006)

Kluyveromyces marxianus: a promising cell factory with controversial traits

A non-conventional yeast *Kluyveromyces marxianus*, also recognised as *Candida kefyr*, is an aerobic homothallic organism known of its fast growth, thermotolerance and ability to utilise a wide range of sugars. Depending on the strain, *K. marxianus* has multiple chromosomes ranging from 6 to 12. Their genome sizes vary from 10.3 to 13.3 million base pairs (Mbp) but have a consistent guanine-cytosine (GC) content, ranging between 40.0-40.8%. Unlike its sister species *Kluyveromyces lactis*, *K. marxianus* is highly polymorphic and may have haploid, diploid or even triploid genome (Ortiz-Merino *et al.* 2018). A phylogenetic tree featuring *Saccharomyces cerevisiae* and other species from *Kluyveromyces* genus is shown in Figure 5.



Species	Lactose assimilation	Inulinase	μ_{max}, h^{-1}	Glucose fermentation	Maximum temperature, °C
Saccharomyces cerevisiae	-	-	0.37	+	35
Kluyveromyces aestaurii	-	-	n/a	+	35
Kluyveromyces nonfermentans	-	-	n/a	-	42
Kluyveromyces wikerhamii	-	-	0.43	+	37
Kluyveromyces lactis	+	-	0.50	+	37
Kluyveromyces marxianus	+	+	0.60	+	52
Kluyveromyces dobzhanskii	-	-	n/a	+	35

Figure 5. Evolutionary relationships between *K. marxianus* and other yeasts. The tree shows the phylogenetic relations between the genus *Kluyveromyces* and *Saccharomyces cerevisiae*. Some key phenotypic traits are also included in the table for comparison. Adapted from (Lane & Morrissey 2010)

While *K. marxianus* and *K. lactis* exhibit the fastest growth rate and distinctive lactose hydrolysation attribute, only *K. marxianus* is thermotolerant and able to assimilate inulin. However, *K. marxianus* does not metabolise cellulose and maltose due to the absence of α-galactosidase. This yeast is also known as Crabtree negative.

Since *K. marxianus* has been isolated mainly from dairy products, it contains GRAS (Generally Regarded as Safe) and QPS (Qualified Presumption of Safety) status and is therefore suitable for applications in the food and pharma industry. However, the past studies suggested a controversial image towards its impact on human health. Whereas B0399 strain was recognised as probiotic (Maccaferri *et al.* 2012), *K. marxianus* was reported to cause 0.2% of invasive candidiasis infection cases (Dufresne *et al.* 2014).

Thermotolerance, fast growth, ability to metabolise various sugars and high protein secretion capacity make *K. marxianus* a promising microbial cell factory proteins (Fonseca *et al.* 2008; Lane & Morrissey 2010). This species was therefore successfully applied in numerous studies for the production of the endogenous enzymes including inulinase (Bender *et al.* 2006), β-

galactosidase (Bansal *et al.* 2008), β-glucosidase (Barron *et al.* 1995) and β-xylosidase (Rajoka & Khan 2005).

Computational tools for species annotation and analysis

Next-generation sequencing. The rapid progress in high-throughput genome sequencing technologies significantly reduced sequencing costs, so that the ability to apply the sequencing is now in reach for individual research groups at an affordable price. The current sequencing technologies are known as next-generation sequencing (NGS), also referred to as the second generation of sequencing. One can use NGS methods to collect different omics data types, such as genomics, transcriptomics, epigenomics, for a given organism. NGS is also amenable in metagenomic analysis, allowing to identify and characterise microbial species from microbial communities of soil, seawater, food and the human gut. Figure 6 shows a typical workflow for NGS analysis.

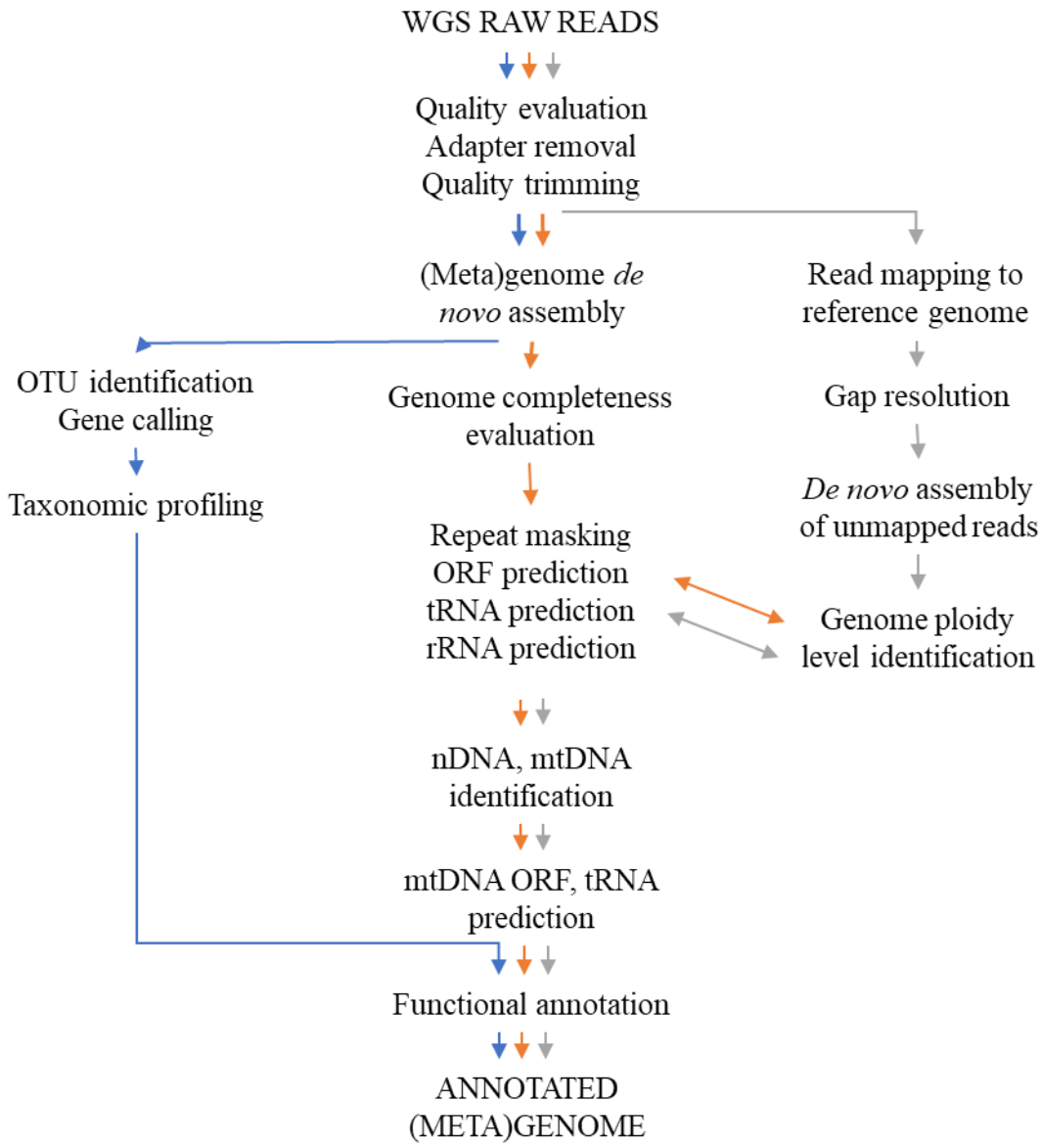


Figure 6. A generalised pipeline for (meta)genome annotation from whole-genome shotgun (WGS) sequencing data. Blue colour denotes metagenomic approach, green – approach based on *de novo* genome assembly, grey – a hybrid approach involving reference-based assembly and *de novo* assembly for unmapped reads. OTU, operational taxonomic unit; ORF, open reading frame; tRNA, transfer RNA; rRNA, ribosomal RNA; nDNA, nuclear DNA; mtDNA, mitochondrial DNA.

Similar to traditional Sanger sequencing, NGS methods rely on procuring polynucleotide chain fragments up to 500 base pairs long through the massive parallelisation. For Eukaryotic species, the total number of the sequenced nucleotides should be at least 150 times higher than the estimated target genome size. However, even though such data may provide a detailed insight into the physiology and metabolic capabilities for the target species, the additional experiments may still be needed to obtain the complete genome. This includes the identification of telomeric regions and the gaps indicating non-sequenced regions. Sanger sequencing is indispensable to close such gaps due to its higher accuracy than NGS. If gap closing is not considered, such sequencing approach is referred to as whole-genome shotgun (WGS).

Quality trimming. The generation of high-throughput NGS data requires the appropriate *in silico* tools for evaluation and processing. Computational tools like FastQC (Andrews & others 2010) and MultiQC (https://multiqc.info/) are indispensable for the initial assessment of the newly generated raw reads. In addition to reporting the overall data quality, it also checks for the adapter and contaminant sequences by checking their overall enrichment. Based on these results, one can estimate the genome size, thereby performing the adapter and quality trimming using trimmomatic (Bolger *et al.* 2014) and BBMap (Bushnell 2014) software. Regarding the paired-end data, both programs also identify the paired and unpaired reads (singletons) and then export them into separate files. Therefore, it is recommended to check the pre-processed data and make sure that adapter and other contaminant sequences are removed by using FastQC/MultiQC.

Genome assembly. In this step, the short reads are used as input to reconstruct the target genome. This can be achieved by using *de novo* or reference-based assembly tools, listed in Table 2.

Table 2. A list of the commonly used genome assembly software.

Name	Algorithm Type	Assembly Type	Input Reads	Reference
ABySS	De Bruijn Graph	*De novo*	Paired-end, single-end	(Jackman *et al.* 2017)
ALLPATHS-LG	Unipath Graph	*De novo*	Paired-end, single-end	(Gnerre *et al.* 2011)
MaSuRCA	Hybrid	*De novo*	Paired-end, single-end	(Zimin *et al.* 2013)
MEGAHIT	De Bruijn Graph	*De novo*	Paired-end, single-end	(Li *et al.* 2015)
MIRA	Hybrid	*De novo*, reference-based	Paired-end, single-end	(Mira *et al.* 2014)
PERGA	Greedy	*De novo*	Paired-end, single-end	(Zhu *et al.* 2014)
SGA	String Graph	*De novo*	Paired-end	(Simpson & Durbin 2012)
SOAPdenovo	De Bruijn Graph	*De novo*	Paired-end, single-end	(Luo *et al.* 2012)
SPAdes	De Bruijn Graph	*De novo*	Paired-end, single-end	(Bankevich *et al.* 2012)
SSAKE	Greedy	*De novo*	Paired-end, single-end	(Warren *et al.* 2007)
Velvet	De Bruijn Graph	*De novo*	Paired-end, single-end	(Zerbino 2010)

The resulting genome consists of a number of continuous DNA fragments, i.e. scaffolds, and its contiguity can be checked by using N50 and L50 metrics. Provided the list of the scaffolds sorted by length, the N50 statistic is defined as the length of the shortest scaffold which together with the longer scaffolds comprise 50% of the total scaffold length while L50 indicates the number of such scaffolds. In general, the assembled genome with the acceptable quality should have N50 higher than 5000 base pairs (bp) and L50 should not exceed 500. However, in addition

to contiguity, the assembled genome should also be assessed for completeness. Tools like QUAST (Gurevich *et al.* 2013) and BUSCO (Simao *et al.* 2015) calculate the number of universal single-copy genes, which are expected to be found in the assembly. A high-quality assembly should have as many such genes as possible. These results can be used as an indicator to tweak the parameter settings for genome assembly software.

Gene prediction and functional annotation. To facilitate gene prediction process, one should firstly annotate repetitive DNA sequences, such as low complexity regions and transposable elements, with programs RepeatMasker (Smit *et al.* 2015) or RepeatModeler (Smit & Hubley 2015). Open reading frames (ORFs) can be identified with prediction tools like AUGUSTUS (Stanke *et al.* 2008) and GeneMark (Ter-Hovhannisyan *et al.* 2008) while tRNAscan-SE (Lowe & Eddy 1997) and RNAmmer (Lagesen *et al.* 2007) are packages to locate tRNAs and rRNAs respectively. Once all predictions are complete and do not overlap in the genome, a ploidy check should be performed by ploidyNGS (https://github.com/diriano/ploidyNGS) or by clustering the predicted protein sequences with, e.g. CD-HIT (Fu *et al.* 2012). A significantly lower protein cluster number than the predicted proteins count indicates that the given species may be polyploid, therefore requiring manual curation for scaffolds before proceeding with gene prediction. It is also essential to identify the longest scaffolds, which do not have any ORFs predicted as they may be parts of mitochondrial DNA (mtDNA), which should be annotated separately with MITObim due to the different codon usage (Hahn *et al.* 2013).

Protein annotation can be performed through Gene Ontology (GO) terms, EuKaryotic Orthologous Groups (KOGs) or other database-specific terms, based on sequence similarity.

Metabolic modelling. Genome-scale metabolic models (GEMs) are valuable systems biology platforms as they allow to combine genome annotation, cultivation and other experimental data for a given species into a whole-cell metabolic network. This information is arranged in stoichiometric and gene-reaction matrixes and together with reaction constraints provide a comprehensive framework about gene-protein-reaction (GPR) associations and metabolic capabilities for a given species. Such a model is amenable for the further metabolic potential investigation through steady-state simulations, where only the biomass and several essential nutrients may be imbalanced. Flux balance analysis (FBA) is the most commonly used method in such simulations (Orth *et al.* 2010). Reaction constraints can be further modified according to simulated conditions, allowing to predict the metabolic outcomes upon these perturbations. GEMs have been successfully used to design strains and evaluate the cell capabilities under different conditions (Saha *et al.* 2014; Thiele & Palsson 2010).

Part I: Comparative functional analysis of bile acid metabolism

Paper I: Metagenomic study of bile acid biotransformation

OBJECTIVES

This study aimed:

- To identify bile acid biotransformation homologous proteins for gut microbial species and estimate their distribution at the phylum level
- To compare the bile acid biotransformation homologues between healthy and diseased subjects
- To compare primary and secondary bile acid abundances between healthy and diseased subjects

MOTIVATION

The previous studies hinted the relation between the human gut microbiome dysbiosis and altered bile acid levels (Marcobal *et al.* 2013; Wahlstrom *et al.* 2016, 2017). Although the efforts were made to analyse bile acid metabolism and its contribution to the human gut microbiome (Gothe *et al.* 2014; Jones *et al.* 2008), these analyses comprised only the partial sets of bile acid biotransformation gene types. The study shown in this section was therefore aimed to identify all the known bile acid biotransformation homologues in the human gut microbial species and provide the comprehensive insight into microbial-mediated bile acid role in the human gut microbiome. The work was based on metagenomic analysis for healthy and inflammatory bowel disease (IBD) patients, given that IBD patients had different gut microbiome composition when compared to the healthy ones (Jansson *et al.* 2009; Le Gall *et al.* 2011).

ANALYSIS, RESULTS AND DISCUSSION

I. Identification of bile salt biotransformation protein homologues

The starting point to identify bile salt bioprocessing protein (BSBP) homologues started with picking the list of experimentally verified BSBPs as a reference from *Clostridium scindens* and other bacterial species from the same genus. The identification of candidate BSBP homologues started with the BLASTP (Altschul *et al.* 1990) search by querying the reference BSBPs against UniProt ("UniProt: a worldwide hub of protein knowledge" 2019) database. Since this database already included proteins which were known as the ones having or not having the bile acid bioprocessing activity, they were used as the positive and negative controls when the threshold values for hit significance was optimised for each reference BSBP. For example, proteins from

Eggerthella lenta were considered as positive controls (Harris *et al.* 2018; Hirano & Masuda 1981), and proteins from *Helicobacter* and *Prevotella* genera were used as negative controls (Han *et al.* 1996; Itoh *et al.* 1999; Yokota *et al.* 2012). After that, the functional domains were identified for putative BSBP homologues during the HMM search (Eddy 2011) against Pfam v31.0 (Finn *et al.* 2014) database. The remaining set of putative BSBP homologues was then queried against eggNOG v4.5.1 database using *eggnog-mapper* local installation (Huerta-Cepas *et al.* 2017) which ran the homology search using DIAMOND blastp (Buchfink *et al.* 2015). The putative BSBP homologues that had the same protein domains and similar functional annotation data from eggNOG were extracted as the final reference database, comprising 10 613 BSBP homologues. These homologues had their taxonomic lineage annotated from UniProt and summarised into phylum-specific occurrence results (Table 3).

Table 3. Distribution of bile acid biotransformation gene types in several bacterial phyla. The values show the total number of strains in each category.

Phylum	**BaiA**	**BaiB**	**BaiCD**	**BaiE**	**BaiF**	**BaiG**	**BaiH**	**BaiI**	**BaiJ**	**BaiK**	**BaiL**	**Bsh**	**HSD**
Actinobacteria	86	317	25	538	146	475	67	0	15	68	44	36	182
Bacteroidetes	158	1	2	1	0	1	2	0	3	0	23	4	264
Firmicutes	385	13	435	20	96	244	424	2	130	92	576	889	409
Fusobacteria	5	0	2	0	0	0	2	0	44	0	0	0	11
Proteobacteria	114	152	92	104	460	73	85	0	31	534	84	2	598
Verrucomicrobia	1	0	0	1	0	0	0	0	0	0	2	0	2

The results suggested that many gut microbial species did not have all bile acid-inducible (Bai) and bile salt hydrolase (BSH) homologues when compared to *C. scindens*. On the other hand, one may hypothesise that the species which contained only a few BSBPs contained the proteins specific to substrates similar to bile acids. The list of putative BSBPs was also evaluated in the species level. Gut microbial strains having more than half bile BSBP categories were included in Figure 7.

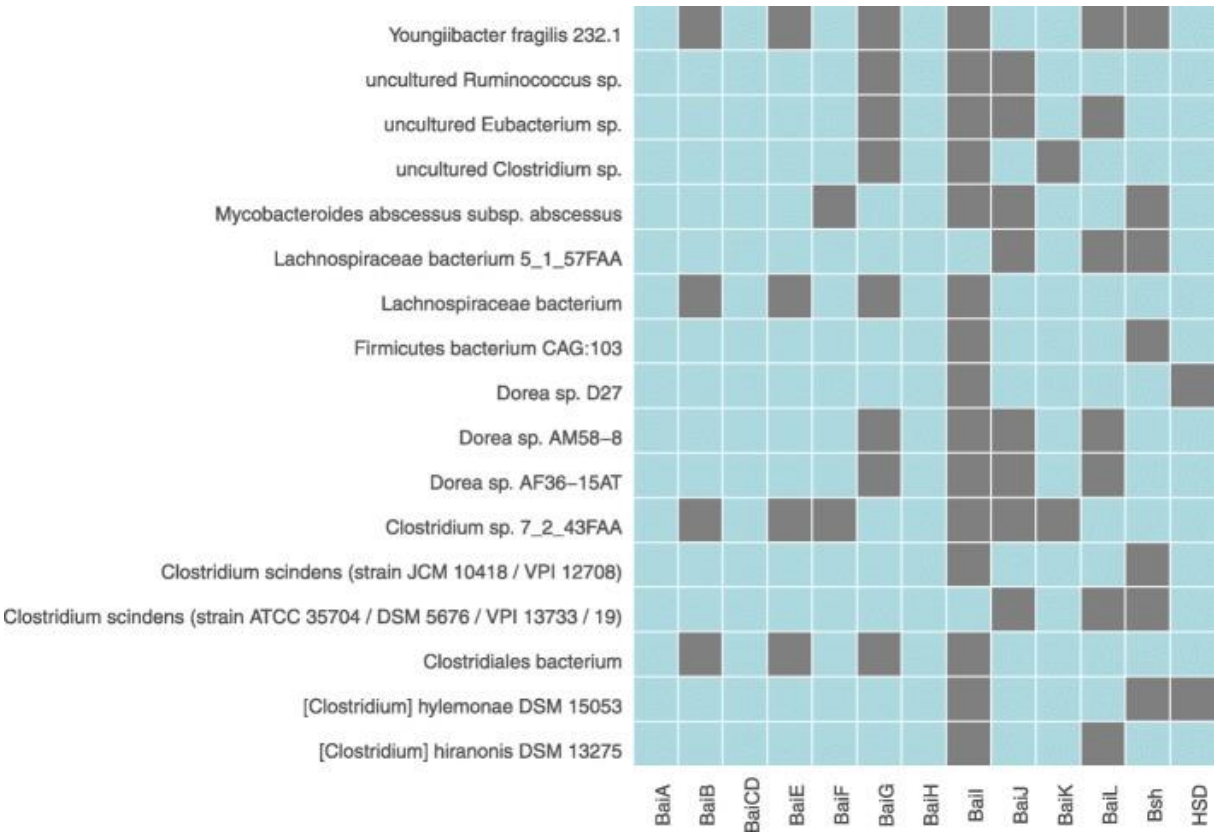


Figure 7. Prevalence of bile salt biotransformation protein homologues in bacterial strains, where more than half of the reference proteins were identified. The grey and blue colour indicate the absence and presence of the protein homologue, respectively.

The heatmap suggested that strains from the same genus had the similar pattern of BSBP homologues, but no species had all BSBP homologues appearing in *bai* operon from *C. scindens*.

II. Comparative analysis for bile acid biotransformation genes between healthy and diseased groups

This analysis involved the shotgun faecal metagenomic data mapping to corresponding genes for BSBPs identified in the earlier step. Such an approach allowed to obtain the differential abundance values for bile acid biotransformation genes (BSBGs). Both metagenomic datasets comprised healthy, IBD cohorts and were downloaded from European Nucleotide Archive at EMBL-EB under accession numbers PRJEB2054 and PRJNA389280. Firstly, the quality assessment was conducted for metagenomic reads with FastQC, and the singleton metagenomic reads were removed using BBMap. The abundance values for BSBGs were calculated with FMAP (Kim *et al.* 2016), which mapped genes to Kyoto Encyclopedia of Genes and Genomes (KEGG) (Ogata *et al.* 1999) pathway ko00121 by UniRef mapping to KEGG Orthologies (KOs). In the following step, the number of mapped reads was normalised in respect of the total number of paired reads in the corresponding metagenomic sample. This pipeline allowed to calculate BSBGs abundance in metagenome samples and compare bile acid biotransformation potential in gut microbiome between healthy and IBD subjects (Figure 8).

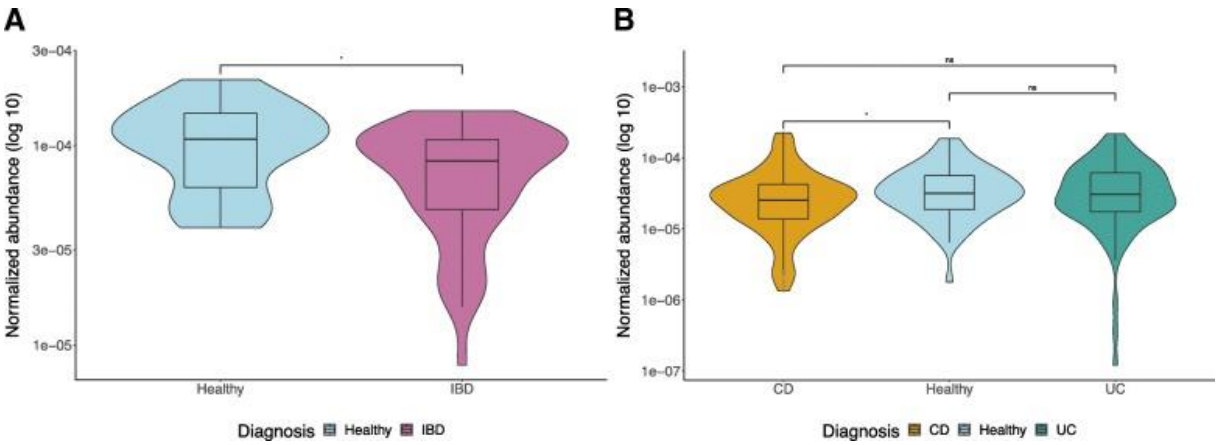


Figure 8. Quantitative comparison of normalised abundance of total BSBGs between healthy and IBD individuals in (a). Spanish cohort, and (b). American cohort. The shape refers to the kernel probability density of the data at different values. The boxplots inside the violin plot represent the interquartile range between the first and third quartiles with the median line inside the boxes, whereas the whiskers indicate the minimum and maximum values from the data distribution. The asterisks on the top indicate ns: p > 0.05, *: p < = 0.05, **: p < = 0.01, ***: p < = 0.001, ****: p < = 0.0001 (Mann-Whitney Wilcoxon test). IBD subjects diagnosed with subtype Crohn's disease and Ulcerative colitis is abbreviated as CD and UC respectively.

Spanish cohort (METAHIT project) comprised metagenomic data for 14 healthy and 25 IBD subjects, for which the abundance for BSBGs was calculated. The results suggested that the lower bile acid biotransformation potential in IBD patients than in a healthy group, since the mean for normalised BSBGs abundance values (7.8E-5) was lower than for healthy controls (1.1E-4). These findings coincided with the findings from another study which approached the same metagenomic dataset (Labbe *et al.* 2014). No comparisons between IBD subtypes (i.e., Ulcerative colitis (UC) and Crohn's disease (CD)) were considered due to having too few CD samples. Correspondingly, American cohort (iHMP project) comprised metagenomic data for 18 healthy and 65 IBD subjects, for which the abundance for BSBGs was calculated, but no significant differences between the healthy and IBD subject groups were found. The follow-up comparison of the healthy group against IBD subtypes showed the significantly lower BSBGs abundance in CD subjects (3.7E-5) than in healthy subjects (4.3E-5), suggesting the correspondingly decrease in bile acid bioprocessing potential.

The comparative analysis of BSBG abundance was also performed in several taxonomic levels. Results indicated that the most prevalent phylum was Firmicutes, coinciding with the literature data (Jones *et al.* 2008). The BSBG abundance values calculated in the phylum level are shown in Figure 9 shows the BSBG abundance values calculated in phylum level between healthy and diseased subject groups.

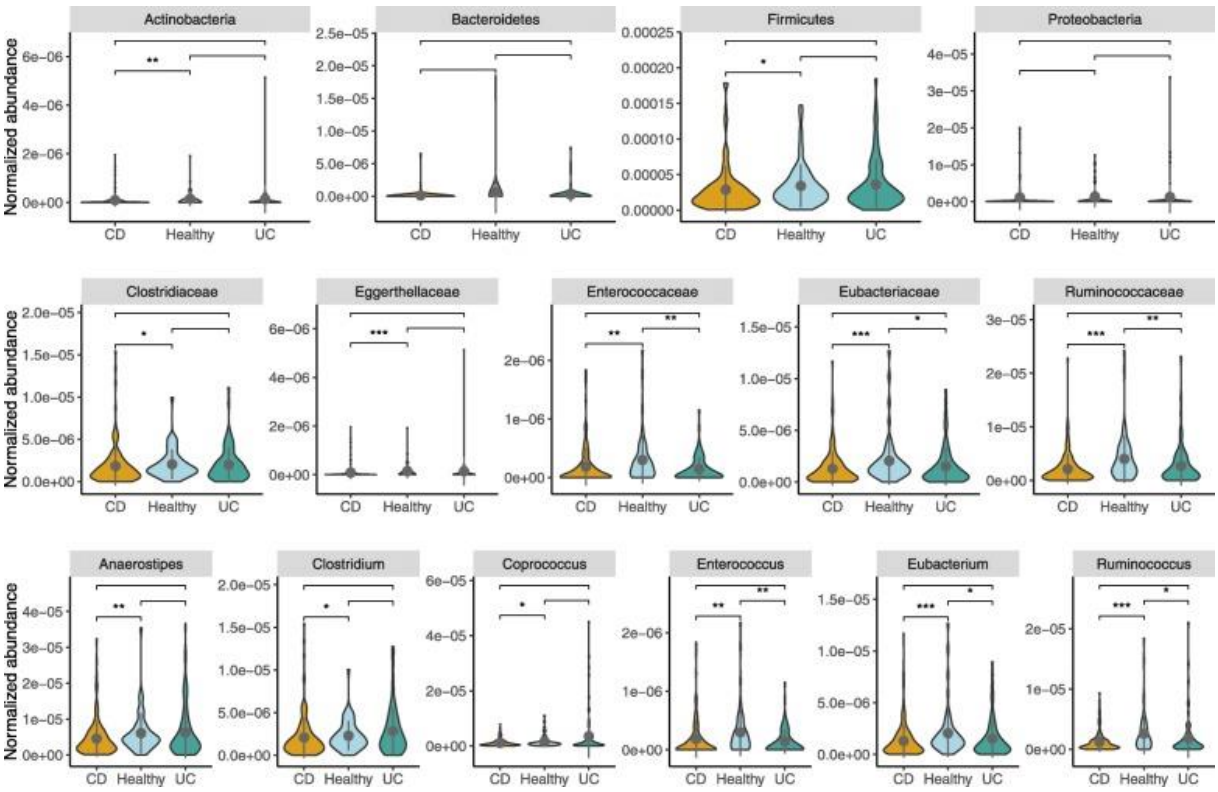


Figure 9. Quantitative comparison of normalised abundance of taxonomic lineage-specific BSBGs between healthy and IBD individuals in the American cohort. The shape refers to the kernel probability density of the data at different values. The point range refers to the mean and error range value of the data distribution. IBD subjects diagnosed with subtype Crohn's disease and Ulcerative colitis is abbreviated as CD and UC respectively.

The results showed that the CD patient group had fewer BSBGs in total from Firmicutes and Actinobacteria than the healthy group. It was, therefore, decided to check the differences of BSBG abundances in family and genus levels of Firmicutes, which is the phylum with the most abundant BSBGs. The analysis revealed that BSBGs originating from Enterococcaceae, Eubacteriaceae, and Ruminococcaceae had a lower prevalence in CD and UC subject groups compared to the healthy group. In contrast, Clostridiaceae and Eggerthellaceae BSBGs had the lower numbers only in CD subject group versus healthy group. Correspondingly, the genus level analysis showed that BSBGs abundances originating from genera *Enterococcus*, *Eubacterium*, and *Ruminococcus* were lower in CD and UC subject groups when compared to a healthy group. BSBGs from *Clostridium* and *Coprococcus* genera had lower abundance values in CD subjects in comparison to the healthy group. The findings in genus-level analysis suggested that genera mentioned above could be relevant descriptors of bile acid biotransformation potential in IBD subject group, as also suggested in the previous study (Martin *et al.* 2018).

III. Comparative bile acid metabolomics analysis between healthy and diseased groups

To check if IBD patients had significantly different levels of primary and secondary bile acids, metabolomics data from iHMP project was analysed. The levels for each bile acid were calculated as the proportion to the total bile acid level in the corresponding sample. The results showing the levels for various primary and secondary bile acids are included in Figure 10.

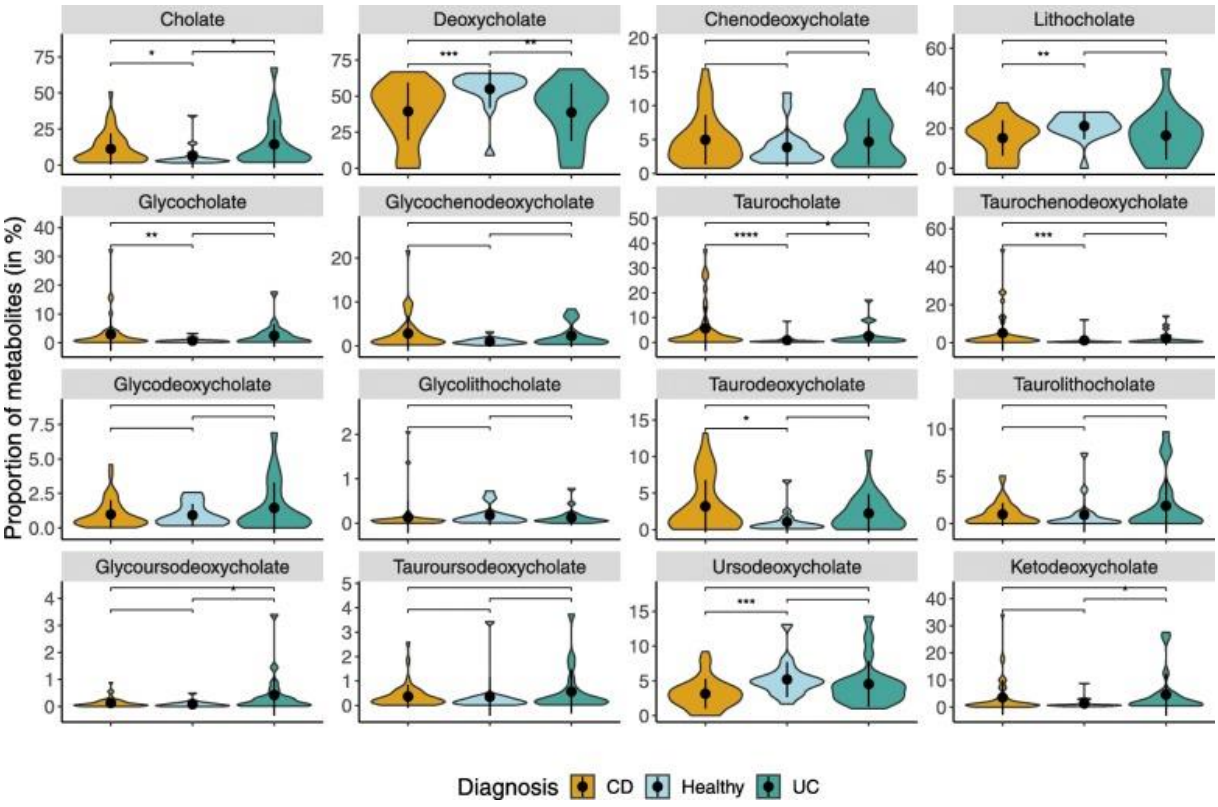


Figure 10. Quantitative comparison of bile acid metabolites between healthy and IBD subjects of American cohort. The shape refers to the kernel probability density of the data at different values. The point range refers to the mean and error range value of the data distribution. The asterisks on the top indicate ns: $p > 0.05$, *: $p <= 0.05$, **: $p <= 0.01$, ***: $p <= 0.001$, ****: $p <= 0.0001$ (Mann-Whitney Wilcoxon test). IBD subjects diagnosed with subtype Crohn's disease and Ulcerative colitis is abbreviated as CD and UC respectively.

Regarding the most abundant bile acids, primary bile acid (cholate) was found in higher levels in IBD patient group while secondary bile acids (deoxycholate and lithocholate) were found in lower levels when compared to the healthy group. Since such findings were expected for the cases in decreased bile acid biotransformation potential as previously shown in IBD patient group, it can be concluded that the current results are in line with the findings of metagenomic data analysis (Figure 8).

Part II: *In silico* genomics analysis for yeast *Kluyveromyces marxianus*

Paper II: Comparative genomics of 12 *K. marxianus* strains

OBJECTIVES

This study aimed:

- To assemble genomes for the two newly sequenced *K. marxianus* isolates
- To predict genes for the two assembled and the other eight published *K. marxianus* isolates
- To functionally annotate the newly annotated genes for ten *K. marxianus* strains and the genes for the other two published strains
- To perform the phylogenetic analysis for the 12 annotated strains
- To identify the *K. marxianus* pangenome and perform its functional analysis

MOTIVATION

Kluyveromyces marxianus is yeast with promising potential as a cell factory. Upon having two newly sequenced *K. marxianus* isolates from kefir grains, it seemed reasonable to analyse these isolates in the context of all known *K. marxianus* isolates. The previous study included the analysis for multiple *K. marxianus* strains in ploidy level (Ortiz-Merino *et al.* 2018). However, there were no computational approaches which were aimed to reconstruct the pangenome for *K. marxianus* and analyse it through the functional context. The reported results shown in this section can be utilised as a starting point towards more specific functional analysis.

ANALYSIS, RESULTS AND DISCUSSION

I. Genome assembly for two *K. marxianus* isolates

Two *K. marxianus* isolates, namely Olga-1 and Olga-2, were sequenced using Illumina HiSeq 2000 platform. The sequencing data comprised the 100 bp long pair-end reads having the insert size between 250 bp and 300 bp. FastQC was utilised to check the quality for the reads, the quality trimming was conducted using trimmomatic. Several *de novo* tools were used to assemble the reads, including ABySS, MIRA, SOAPdenovo and SPAdes. The testing for k-mer length value was applied for De Bruijn graph-based assembly tools. The tested values for k-mer length varied between 41 and 95. An evaluation for the resulting genome assemblies was made based on two criteria: genome contiguity and gene completeness. The genome contiguity was checked with QUAST while the gene completeness for single-copy orthologs was evaluated using BUSCO. The testing results showed that the most single-copy orthologs while

retaining the relatively high genome contiguity could be achieved when using the ABySS assembly tool and setting the value for k-mer length to 49. This setting was applied for both strains, and the corresponding assemblies were kept for the further analysis steps.

II. Gene prediction for ten *K. marxianus* strains

In addition to the two newly sequenced *K. marxianus* strains, gene prediction was sought for the other eight *K. marxianus* strains downloaded National Center for Biology Information (NCBI), including B0399, CBS4857, DMB1, IIPE453, KCTC17555, LHW-O, NRRLY-6860 and UFS-Y2791. RepeatMasker and RepeatModeler were used to identify repeat regions in the genomes. Protein coding sequences (CDSs) and transfer RNAs (tRNAs) were predicted with the *funannotate predict* function from funannotate (Palmer & Stajich 2019). The protein evidence needed for this process was obtained from Swiss-Prot. DIAMOND (Buchfink *et al.* 2015) was used to align the protein evidence to genomes and alignments were later refined by Exonerate (Slater & Birney 2005). After that, the alignment results were used as input for *ab initio* gene prediction tools GeneMark-ES and AUGUSTUS. EVidenceModeler (Haas *et al.* 2008) was utilised to combine the predicted gene models while bedtools (Quinlan & Hall 2010) allowed checking these models for the length, gaps and transposable elements. The prediction for tRNAs was achieved with tRNAscan-SE.

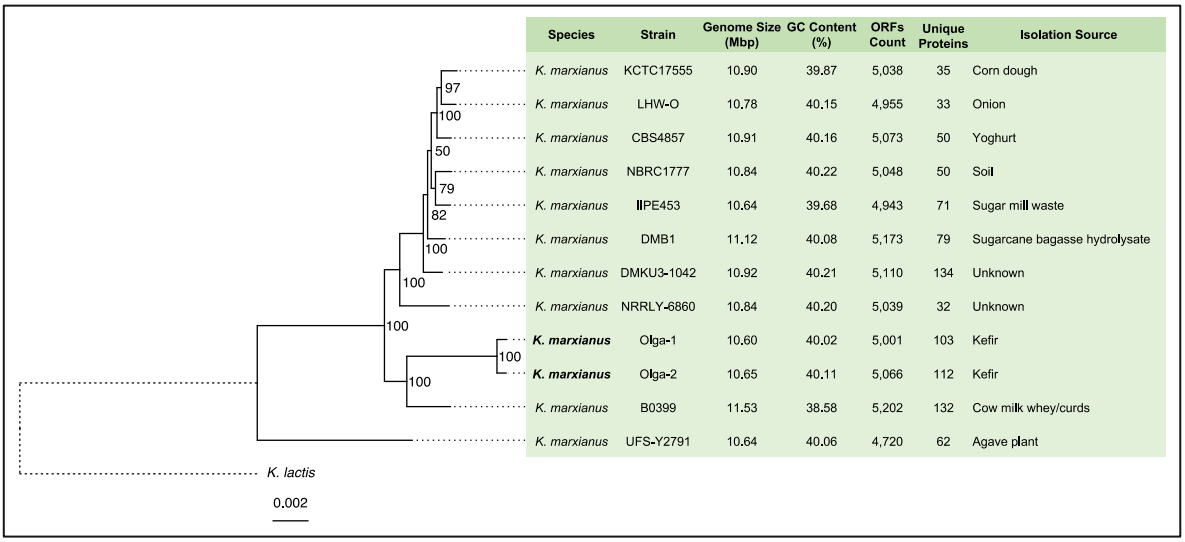
III. Functional annotation for 12 *K. marxianus* strains

Functional annotation was sought for ten *K. marxianus* strains having the newly predicted protein-coding sequences and for the two other strains downloaded from NCBI: DMKU3-1042 and NBRC1777. The annotation for the latter strains included only the identification for EuKaryotic Orthologous Groups (KOGs) using eggnog-mapper (Huerta-Cepas *et al.* 2017) tool.

The functional annotation was based on predicted protein sequences, which were used as input to the *funannotate annotate* function from funannotate. Gene names for proteins were fetched from the best hits during DIAMOND blastp search against Swiss-Prot proteins. The annotation for peptidases and biosynthetic gene clusters was obtained during DIAMOND blastp search against MEROPS v12.0 (Rawlings *et al.* 2018) and MIBiG v1.4 (Medema *et al.* 2015) databases respectively. The function *hmmsearch* from HMMER (Eddy 2011) allowed to identify and annotate protein families from Pfam v32.0 and carbohydrate-active enzymes (CAZymes) from dbCAN v7 (Yin *et al.* 2012). Also, GO, and InterPro protein families were annotated using InterProScan local installation (Jones *et al.* 2014). The annotation for KOGs was obtained from the eggNOG database using eggnog-mapper. Phobius (Kall *et al.* 2005) and SignalP were used to identify transmembrane and secreted proteins, respectively. Secondary metabolite biosynthetic gene clusters were located and annotated with antiSMASH local installation (Medema *et al.* 2011).

IV. Phylogenetic analysis for 12 *K. marxianus* strains

For *K. marxianus* phylogenetic analysis, the proteome for *Kluyveromyces lactis* NRRLY-1140 (downloaded from NCBI) was chosen as the outgroup species. The phylogenetic tree was built using the proteome from *K. lactis* and 12 *K. marxianus* strains as input in the *funannotate compare* function from funannotate. Proteinortho (Lechner *et al.* 2011) allowed to classify proteins into orthologous groups. The results showed that all genomes shared 1 068 single-copy genes. The corresponding protein sequences were used as input in multiple sequence alignment with MAFFT (Katoh & Standley 2013). Poorly aligned regions were trimmed with trimAl tool (Capella-Gutiérrez *et al.* 2009). The remaining sequences were concatenated into one sequence having 579 506 amino acids. RAxML (Stamatakis 2014) was used to generate 100 maximum likelihood phylogenetic trees. The PROTGAMMALG model was considered as the amino acid substitution model, as suggested by the maximum likelihood criterion. Bootstrap support values were calculated upon 100 iterations, including resampled data sets for each iteration. Figure 11 comprises the maximum likelihood tree with the highest likelihood value and some statistics for the genomes.



Species	Strain	Genome Size (Mbp)	GC Content (%)	ORFs Count	Unique Proteins	Isolation Source
K. marxianus	KCTC17555	10.90	39.87	5,038	35	Corn dough
K. marxianus	LHW-O	10.78	40.15	4,955	33	Onion
K. marxianus	CBS4857	10.91	40.16	5,073	50	Yoghurt
K. marxianus	NBRC1777	10.84	40.22	5,048	50	Soil
K. marxianus	IIPE453	10.64	39.68	4,943	71	Sugar mill waste
K. marxianus	DMB1	11.12	40.08	5,173	79	Sugarcane bagasse hydrolysate
K. marxianus	DMKU3-1042	10.92	40.21	5,110	134	Unknown
K. marxianus	NRRLY-6860	10.84	40.20	5,039	32	Unknown
K. marxianus	Olga-1	10.60	40.02	5,001	103	Kefir
K. marxianus	Olga-2	10.65	40.11	5,066	112	Kefir
K. marxianus	B0399	11.53	38.58	5,202	132	Cow milk whey/curds
K. marxianus	UFS-Y2791	10.64	40.06	4,720	62	Agave plant

Figure 11. Maximum likelihood phylogenetic tree and genome information for *Kluyveromyces marxianus* strains, for which the whole genome sequencing data was available. The tree was based on the concatenated protein sequence containing 1 068 genes present in all genomes. *Kluyveromyces lactis* was chosen as an outgroup. The bootstrap support is provided from 100 iterations while the scale bar indicated the expected substitutions per site. The species written in bold were sequenced in the current study. The branch length for the outgroup was scaled down 100-fold. ORFs, open reading frames.

The results showed that *K. marxianus* genomes were between 10.64 Mbp and 11.53 Mbp in size while two strains had the genomes longer than 11 Mbp. Regarding the GC content, *K. marxianus* B0399 had the lowest value (38.58%) while nine isolates had this ratio above 40%. Gene numbers of these strains differed between 4 720 and 5 202. The strain B0399 had the highest number of unique proteins (132) among all compared strains. Based on the phylogenetic tree, 11 of 12 *K. marxianus* strains could be classified into two clades. The first clade included eight strains appearing in the upper part of the figure, in which the members were isolated from

soil, crops, vegetables and yoghurt. Another clade comprised the strains isolated from dairy products: Olga-1, Olga-2 and B0399. Although one dairy isolate (CBS4857) appeared in the first clade, the suggested classification coincided with “A” and “B” haplotypes as described in *K. marxianus* ploidy variation study (Ortiz-Merino *et al.* 2018).

V. *K. marxianus* pangenome identification and its functional analysis

The orthologous groups (OGs) identified by Proteinortho in earlier step were used to identify *K. marxianus* pangenome and core genome (Figure 12). Among the 12 *K. marxianus* strains which comprised 60 368 protein sequences, the pangenome size was 5 804 OGs while the size for the core genome was 3 855 OGs. The results suggested that *K. marxianus* pangenome was of closed type. The ratio between the core genome and pangenome sizes (66%) was very similar to the same ratio for *Saccharomyces cerevisiae* (Li *et al.* 2019).

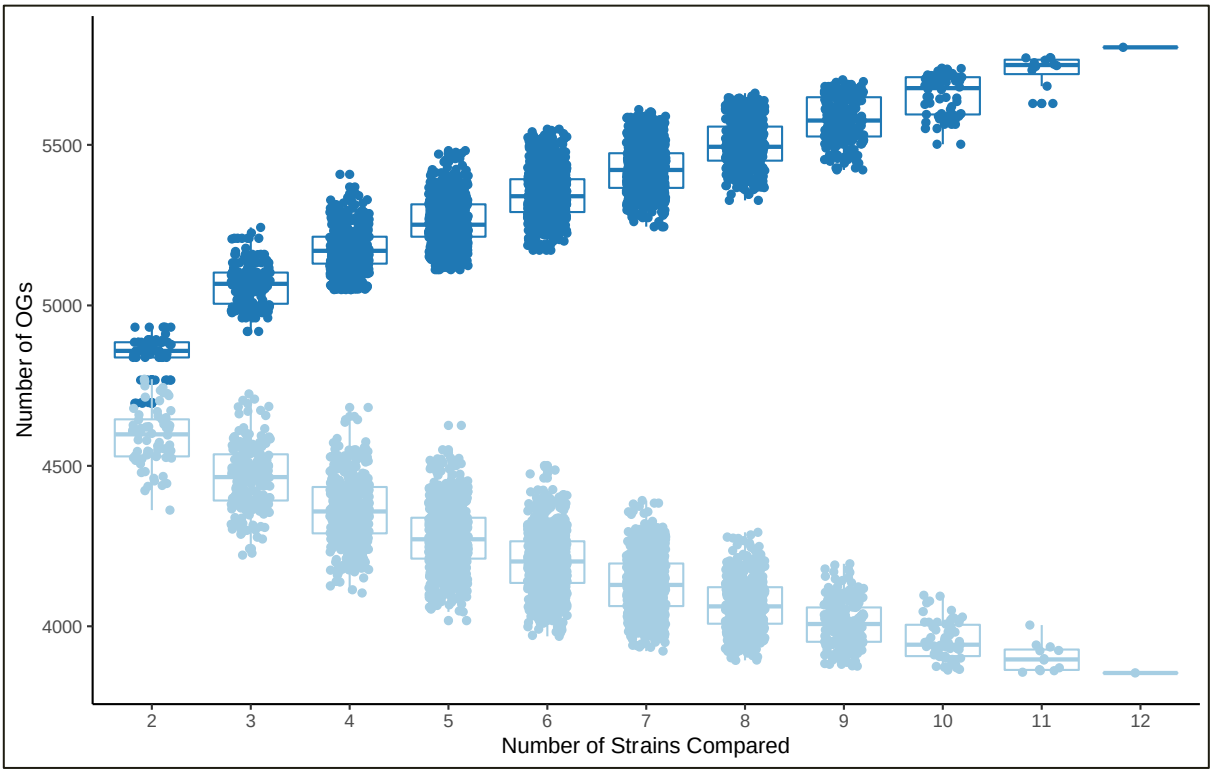


Figure 12. The size of *K. marxianus* pangenome and core genome depending on the number of the strains compared. For the strain combinations below 12, the calculations were based on subsampling all the possible strain combinations for particular strains number. Dark blue dots denote pangenome while light blue dots show core genome sizes. OGs, orthologous groups.

The pangenome, core and accessory genomes of *K. marxianus* were used for functional analysis involving KOGs (Figure 13). Regarding the distribution for the significant KOG categories, pangenome consisted of 26% metabolic, 29% information storage and processing, and 31% cellular processes and signalling KOGs. While the proportion for these same KOG categories was similar in core genome (27%, 28%, 32% correspondingly), the accessory genome had the higher part of KOGs related to information storage and processing (32%), having the smaller

part of metabolic (22%) and cellular processes and signalling KOGs (29%). Regarding the individual KOGs and their distribution between the core and accessory genomes, the majority of KOGs were included in the core genome, ranging between 72.96% and 91.68%. The most presented KOG categories in the core genome related to were lipid transport and metabolism (I) and nucleotide transport and metabolism (F), extracellular structures (W) while the most prevalent in the accessory genome were defence mechanisms (V) and translation, ribosomal structure and biogenesis (J) and inorganic ion transport and metabolism (P).

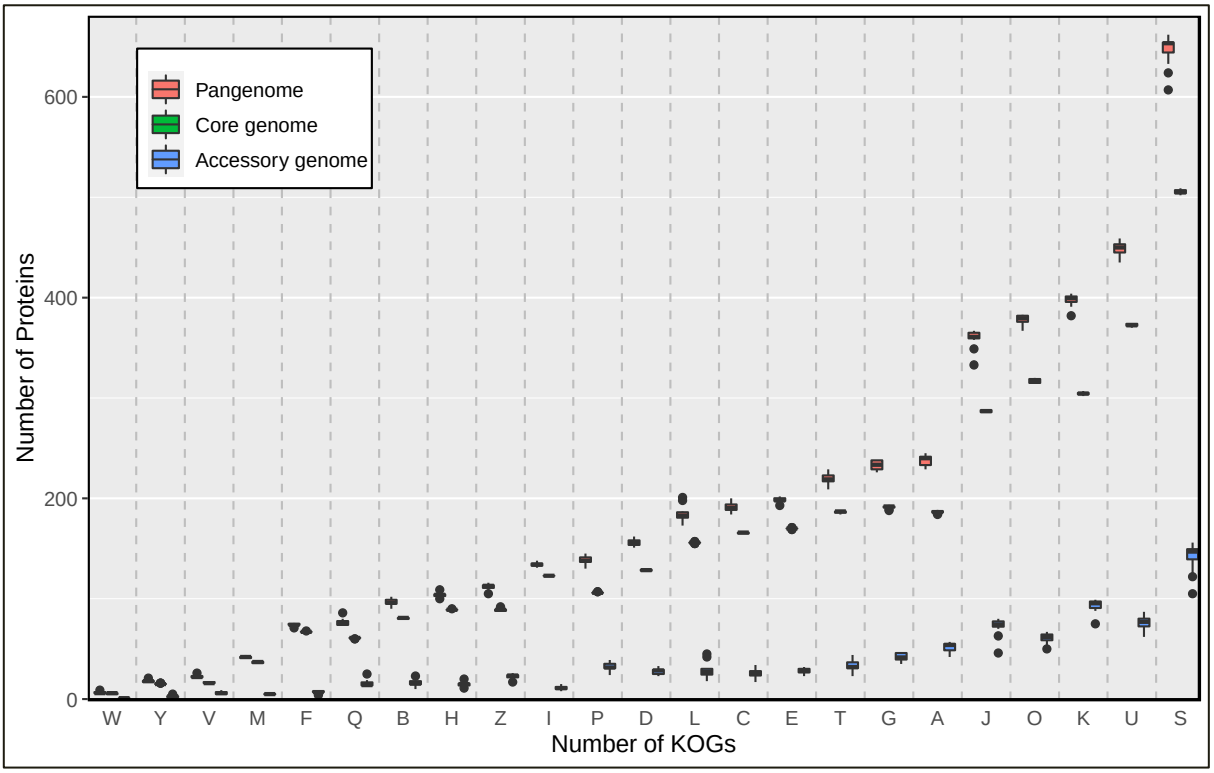


Figure 13. The distribution of *K. marxianus* proteins assignment to different EuKaryotic Orthologous Groups (KOGs) across its pangenome, core genome and accessory genome. KOGs were sorted by their average occurrence in pangenome. The letters in horizontal axis denote the following functional KOGs: (i) related to metabolism: C – energy production and conversion, E – amino acid transport and metabolism, F – nucleotide transport and metabolism, G – carbohydrate transport and metabolism, H – coenzyme transport and metabolism, I – lipid transport and metabolism, P – inorganic ion transport and metabolism, Q – secondary metabolites biosynthesis, transport and catabolism; (ii) related to cellular processes and signalling: D – cell cycle control, cell division, chromosome partitioning, M – cell wall/membrane/envelope biogenesis, O – posttranslational modification, protein turnover, chaperones, T – signal transduction mechanisms, U – intracellular trafficking, secretion and vesicular transport, V – defence mechanisms, W – extracellular structures, Y – nuclear structure, Z – cytoskeleton; (iii) related to information storage and processing: A – RNA processing and modification; B – chromatin structure and dynamics, J – translation, ribosomal structure and biogenesis, K – transcription, L – replication, recombination and repair; (iv) poorly characterised: unknown function (S).

Paper III: Reconstruction and analysis of *K. marxianus* GEM

OBJECTIVES

This study aimed:

- To reconstruct and validate the first GEM for *K. marxianus*
- To use the reconstructed GEM for metabolic capabilities evaluation in stress conditions
- To implement the continuous development for *K. marxianus* GEM

MOTIVATION

The decision to reconstruct the genome-scale model for *K. marxianus* was based on three reasons. Firstly, the genome features derived from Paper II showed the genetically determined physiological capabilities of the cell, but not their actual utilisation during homeostasis. Secondly, no published metabolic model in genome-scale was available for *K. marxianus*. Finally, only a few previous studies utilised GEMs to derive and compare temperature-specific models so that such an approach would provide the novel insight at systems level. The genome-scale set of metabolic reactions was therefore comprised into iSM996, the first publicly available GEM for *K. marxianus*. This model allowed to predict the *in silico* growth from various substrates and performed well when predicting various carbon source utilisation and the growth rates under different conditions. Condition-specific GEMs were further generated by integrating transcription start site sequencing (TSS-Seq) and *in silico* YPD data into the iSM996, which is publicly available as a GitHub repository and is updated under a regular basis to ensure its functionality and compatibility with the conventional constraint-based metabolic modelling tools.

ANALYSIS, RESULTS AND DISCUSSION

I. Reconstruction and validation of iSM996

a. Reconstruction of iSM996

The semi-automatic iSM996 model reconstruction was based on *K. marxianus* DMKU3-1042 proteome (NCBI accession PRJDA65233) and databases KEGG, MetaCyc (Caspi *et al.* 2018), TransportDB (Elbourne *et al.* 2017) and BRENDA (Schomburg *et al.* 2017). The template model used in this study was iOD907, the GEM for the species in the same genus *Kluyveromyces lactis* (Dias *et al.* 2014). The RAVEN Toolbox (Wang *et al.* 2018) was utilised to generate two draft models. The first draft model contained homologous reactions from iOD907 and was generated using the RAVEN function *getModelFromHomology*, which performed the bi-directional BLASTP search between *K. marxianus* and *K. lactis* proteomes

upon default threshold values (e-value 1E-30; identity 40%; alignment length 200). Spontaneous reactions and reactions associated with the short homologues (less than 250 amino acids) from the template model were also added to this draft model. The second draft model was generated using the RAVEN function *getKEGGModelForOrganism*, which applied the HMMER homology search while querying *K. marxianus* proteome against KO specific HMM sets upon default threshold value (e-value 1E-50). After that, this model was compartmentalised with the RAVEN function *predictLocalization*, which utilised WoLF PSORT (Horton *et al.* 2007) protein scores as input. Metabolite names and reaction reversibility information was imported to this draft model from iOD907, Yeast 7.6 (Aung *et al.* 2013) and HMR2 (Mardinoglu *et al.* 2014). The distinctive reactions from the second draft model were added to the first draft model while resolving all the metabolite compartmentalisation discrepancies in favour of the first draft model. The resulting model was considered for further reconstruction steps.

Since the full biomass composition data was not available for *K. marxianu*s, the corresponding information from iOD907 was fetched and modified with relevant bibliographic data where applicable. The mass composition for protein, carbohydrate, lipid, RNA and DNA content per gram cell dry weight (g/gDW) was integrated from the cultivation study (Fonseca *et al.* 2007). After that, stoichiometric coefficients for these biomass components were scaled-up to constitute exactly one gDW in biomass pseudo reaction. The cell wall carbohydrate composition for glucan, mannan and chitin was obtained from the cell wall study (Nguyen *et al.* 1998). Knowing the total carbohydrates mass part in one gDW, the mass for the remaining cell carbohydrates trehalose and amylose was calculated. The composition for nucleotides, deoxynucleotides and amino acids were calculated from the genome, transcriptome (including tRNA and rRNA) and proteome as described in the protocol (Thiele & Palsson 2010). No literature data was available for *K. marxianus* phosphate/oxygen ratio (P/O ratio), the growth-associated maintenance (GAM) and the non-growth associated maintenance (NGAM), so the corresponding data from the iOD907 was implemented to the model.

The semi-automatic gap filling was then performed to ensure that the model could produce all biomass components in Verduyn medium (Verduyn *et al.* 1992), which was considered as the minimal medium in the study. The primary gap-filling reaction sources were iOD907, Yeast 7.6 models and KEGG. The candidate reactions for the gap-filling were identified using the RAVEN function *fillGaps* and were kept in the model if no contradictions in literature were found.

The final iSM996 reconstruction steps involved the manual curation, where the efforts were made to implement *K. marxianus* literature data into the model. This included the addition of missing uptake pathways for several carbon (Lachance 2011) sources like inulin, L-arabinose and D-mannitol. Biosynthetic pathways were added for the known products, including 2-phenylethanol, phenethyl acetate and ethyl acetate. The iSM996 model was also curated for gene associations, EC numbers (Bairoch 2000), metabolite names, reaction elemental/charge balance data. The existing gene-protein-reaction (GPR) rules were checked for the substrate, cofactor usage and sub-cellular localisation relevance.

As a result, the final reconstructed metabolic network for *K. marxianus* contained the metabolic features imported from Yeast 7.6 (Aung *et al.* 2013), HMR2 (Mardinoglu *et al.* 2014),

BRENDA, TransportDB, KEGG and MetaCyc, however, the major part of the network was obtained from iOD907. In comparison with iOD907, iSM996 had higher genome coverage and more reactions occurring in the cytosol, albeit having fewer non-*S. cerevisiae* genes and reactions occurring outside the cell (Table 4). Both species shared 886 orthologous metabolic genes, thereby covering 97.7% iOD907, 88.9% iSM996 genes and resembling the pairwise similarity of their whole genomes (94%). The reconstructed model showed a noticeable increase in genome coverage and gene number since the decision was made to keep the genes in the model even from isolated subnetworks. Although it was not possible to connect such associated reactions to the main subnetwork during the reconstruction, these reactions still characterised the genetically determined metabolic features and can be re-wired to the main subnetwork upon the sufficient experimental evidence. The inclusion of such isolated subnetworks in iSM996 increased the blocked reactions percentage. It, therefore, seemed reasonable to compare iSM996 with the published model for phylogenetically close species with desirably more detailed literature knowledge available than *K. marxianus*. Yeast 8.3.4 (Sánchez *et al.* 2019), the yeast consensus model for the very well-studied budding yeast *Saccharomyces cerevisiae* was considered for the comparison using the GEM test suite *memote* (Lieven *et al.* 2018). As a result, there was a small difference in respect of the blocked reactions percentage, being 23.73% for iSM996 and 17.14% for Yeast 8.3.4. Moreover, iSM996 demonstrated higher overall consistency score than Yeast 8.3.4, being equal to 88 and 67, respectively. The higher score was due to the higher stoichiometric consistency and higher reactions percentage with mass balance. So, one could conclude that iSM996 is of similar quality like the yeast consensus model, which has been continuously refined since 2008.

Table 4. Comparison between *Kluyveromyces lactis* GEM iOD907 and *Kluyveromyces marxianus* GEM iSM996.

	iOD907	**iSM996**
Genes	907 (17.8%)	996 (20.1%)
S. cerevisiae homologues	691	916
Unique	216	80
Reactions	2 180	1 913
Extracellular	938	507
Cytosol	853	974
Mitochondria	359	390
Endoplasmic Reticulum	30	42
Metabolites	1 477	1 531
Extracellular	313	191
Cytosol	822	907
Mitochondria	296	359
Endoplasmic Reticulum	46	74

Albeit iSM996 had higher genes number, it featured fewer reactions and metabolites than iOD907 (Figure 14a). The main reason for this decrease was because all cytosolic metabolites in iOD907 had their corresponding counterparts in extracellular space while also having transport reactions for these metabolites between mentioned two compartments. The list of extracellular metabolites was revised according to the yeast consensus model, which allowed to decrease the number of transportable cytosolic metabolites from 313 to 182.

Regarding the featured metabolic pathways in iSM996, the most reactions linked to transport, exchange reactions and amino acid, lipid, carbohydrate metabolism (Figure 14b). The model contained 365 transport reactions between the cell and extracellular space and 140 reactions between intracellular compartments.

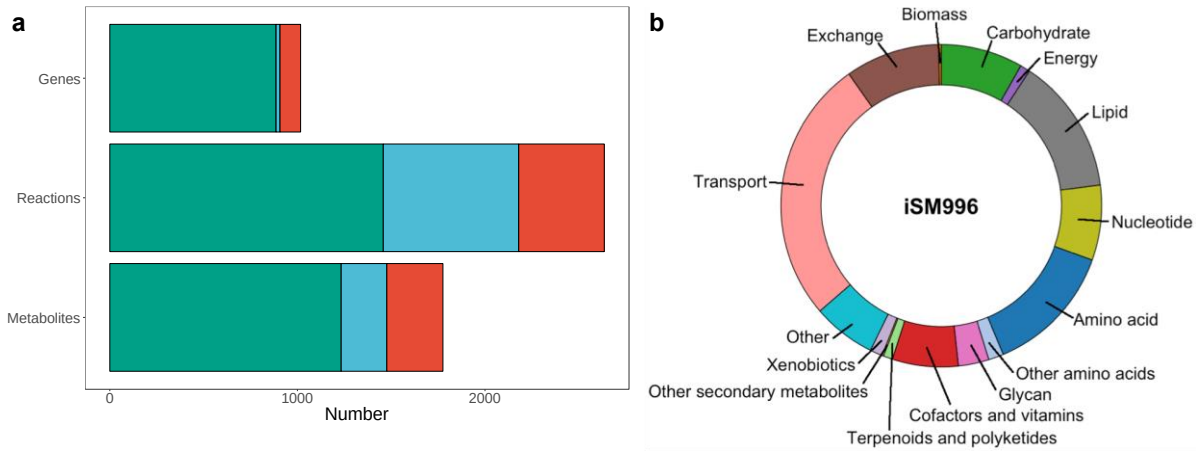


Figure 14. Overview of iSM996. (a) Comparison of genes, reactions and metabolites present in iSM996 and template model (iOD907). Green colour indicates overlapping entities, blue – specific to iOD907, red – specific to iSM996. (b) Distribution of reactions in each metabolic part.

b. Validation of iSM996

The validation for iSM996 was done with FBA upon minimal medium constraints. The qualitative checks for the *in silico* growth in various carbon and nitrogen sources were performed and compared with the literature knowledge (Figure 15a). During such testing, the composition for the minimal medium was modified in a way that tested substrates would be the sole sources for carbon or nitrogen. The model could predict the *in silic*o growth from various carbon sources, including glucose, galactose, D-xylose, sucrose, lactose, cellobiose and inulin. The growth could also be predicted in amino acid-free minimal medium, suggesting that the organism was capable of *de novo* synthesise all the necessary amino acids. However, no growth could be predicted when using L-lysine and cadaverine as the sole carbon sources. The LYSDEGII-PWY pathway for *Saccharomyces cerevisiae* in MetaCyc suggests the possible scenario for L-lysine degradation, showing the 6-step linear pathway from L-lysine to glutarate. Glutarate would then be converted to crotonoyl-CoA through two reactions, thereby reaching a fatty acid metabolic pathway. Regarding cadaverine, it is the product of L-lysine decarboxylation, but its further catabolism is unknown in fungi. However, one may suggest that cadaverine is converted to L-lysine through the carbon fixation and then processed through the LYSDEGII-PWY pathway like L-lysine. It was decided not to include these reactions due to the high uncertainty and since none of the candidate reactions was linked to any *S. cerevisiae* genes, making the homology search for *K. marxianus* impossible.

The iSM996 model was also used to find the computational explanation, why the target species cannot assimilate the non-growth carbon sources (Figure 15a). An investigation suggested that D-gluconate and N−acetyl−D−glucosamine cannot be assimilated, since the model was unable to transport them into the cell. The model could not predict the decomposition for melibiose, trehalose, starch, and myo-inositol, because the organism lacked glycoside hydrolases for these substrates. Although the growth could be predicted upon L-arabinose, D-arabinose cannot be assimilated, since *K. marxianus* lacks the gene linked to the redox-driven L-arabinose isomerisation to D-xylulose through L-arabinitol as intermediate. The growth upon D-

glucosamine as the only carbon source was not possible, because *K. marxianus* does not include glucosamine-6-phosphate deaminase like *S. cerevisiae* (Flores & Gancedo 2018).

The growth rate accuracy in minimal media was also evaluated for the iSM996 model. This was done by fixing the substrate uptake rates and comparing *in silico* growth rates with experimental values. The simulations (Figure 15b) suggested the strong correspondence between predicted and experimental growth rates, mainly when the experimental growth rates were below 0.3 h^{-1}. One may hypothesise that while the value for NGAM does not significantly impact the growth rates, the cell considers different mass proportion for biomass components and different GAM value. It was therefore not possible to re-use the same biomass composition in a wide range of the growth rates. The biomass reactions in GEMs are usually most relevant for the growth rates below 0.4 h^{-1} while to simulate the growth in higher growth rates, more specific experimental data for the biomass composition, GAM and NGAM values are needed.

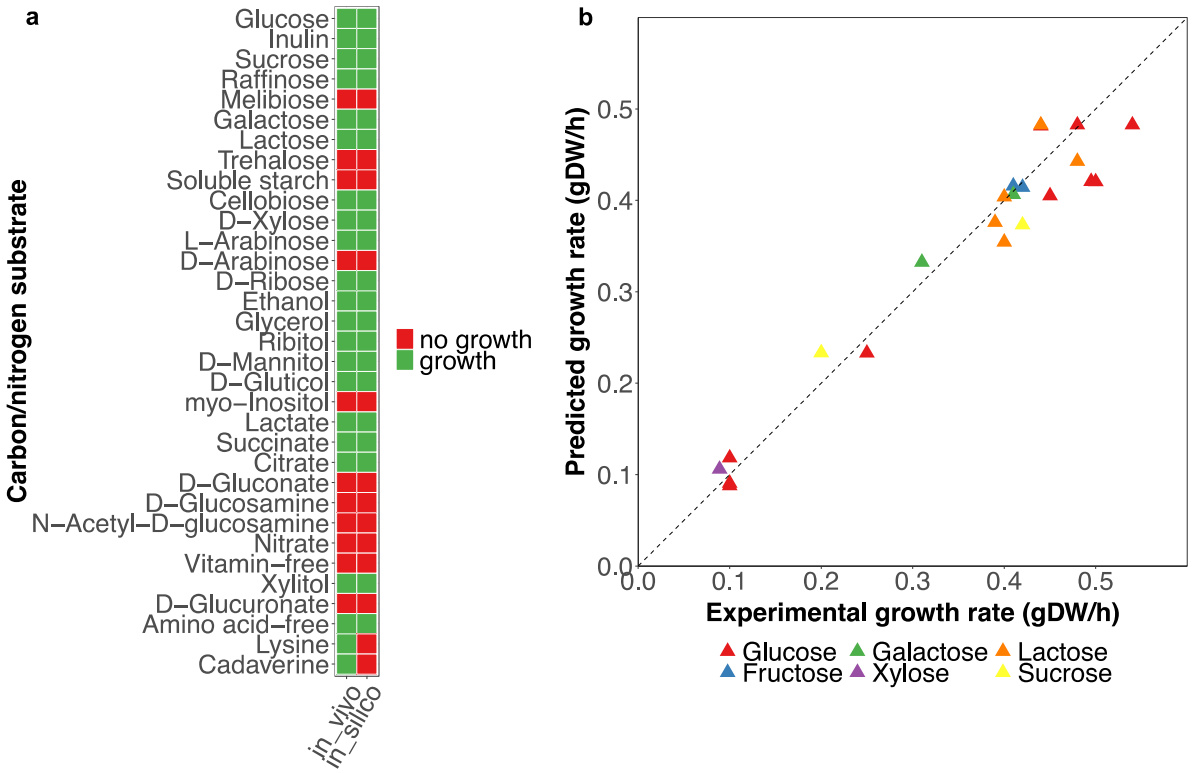


Figure 15. Validation results for iSM996 in minimal medium. (a) Comparison of the *in silico* growth for various carbon and nitrogen sources against literature data. Upon the simulations for nitrogen sources, glucose was considered as a carbon source. (b) Comparison of *in silico* and experimental growth rates for various carbon sources in minimal medium. The squared value of the Pearson correlation coefficient between experimental and predicted growth values was 0.9445.

II. Predicting *K. marxianus* metabolic capabilities in microaerobic and high-temperature conditions with iSM996

The iSM996 model was utilised to predict *K. marxianus* metabolic capabilities in microaerobic and high-temperature conditions. The model was therefore coupled with TSS-Seq data to deactivate reactions associated with inactive genes, thereby constructing the condition-specific models. Experimental data used in this analysis was obtained from Gene Expression Omnibus (GEO) (Edgar *et al.* 2002) under accession ID GSE66600. This dataset contained *K. marxianus* gene expression values from cultivation in rich medium: yeast extract peptone dextrose (YPD) and yeast extract peptone xylose (YPX). Four conditions were comprised in TSS-Seq dataset: 30°C YPD shaking (30D), 30°C YPX shaking (30X), 30°C YPD non-shaking (30DS) and 45°C YPD shaking (45D). The data for 30X condition was not included in the analysis, because the essential *KmXKS1* gene, responsible for D-xylose assimilation through D-xylulose phosphorylation was not active, thus making the model unable to predict growth. Nonetheless, all four conditions were used to identify the genes which were not active at least in one of four conditions. Consequently, three condition-specific models were obtained having the specific constraints for reactions linked with inactivated genes. For the more accurate and less speculative predictions, the genes which were inactive in all four conditions were not used to block the associated reactions. The gene was identified as inactive if it had zero expression values in all three replicates per condition. No threshold values for expression abundance values were implemented in this analysis. In total, TSS-Seq data was available for 988 from 996 genes in iSM996 while 115 genes were found inactive in at least one condition.

The condition-specific models were utilised to predict the maximal production capacity for the main biomass components (Figure 16). The objective function was fixed to 90% of the maximal growth rate and then maximised for production for biomass components. The integration results showed that the most zero-flux reactions (638) could be observed in the 45D condition, whereas 30D and 30DS correspondingly had 544 and 541 such reactions. The results also suggested the tighter regulation in 45D (80 genes turned off) than in 30D (24 genes turned off) and 30DS (15 genes turned off) conditions. FBA simulations showed the auxotrophy for riboflavin in 30DS and 45D. Besides, the ferroheme auxotrophy was observed in the 45D condition. To fix feasibility problems for these models and make then amenable for the comparison of metabolic profiles, the assumption was made that riboflavin and ferroheme were available in the medium and could be transported into the cell. These modifications allowed to make all three condition-specific models feasible and comparable to each other.

Reduced cost analysis suggested that L-cysteine in all three conditions was the primary growth-limiting substrate. The growth simulations also showed that *K. marxianus* could gather the amino acid pool for the same protein amount. The auxotrophies to L-arginine and L-histidine were found in 30D and 45D conditions. Meanwhile, the auxotrophies specific to the high temperature corresponded to the previously reported L-lysine and L-isoleucine auxotrophies (Yarimizu *et al.* 2013) and previously unobserved auxotrophies to L-alanine, L-phenylalanine and L-tyrosine. Given the inability to *de novo* synthesise seven amino acids at high temperature, the cell may conserve the precursor metabolites for these amino acids by allowing the higher flux through Embden–Meyerhof–Parnas (EMP) pathway and the citric acid (TCA) cycle. On the other hand, the lower predicted production pool at high temperature for L-alanine, L-

isoleucine, L-phenylalanine, L-tyrosine and L-arginine may reduce the availability for the biosynthesis of some proteins enriched by these amino acids.

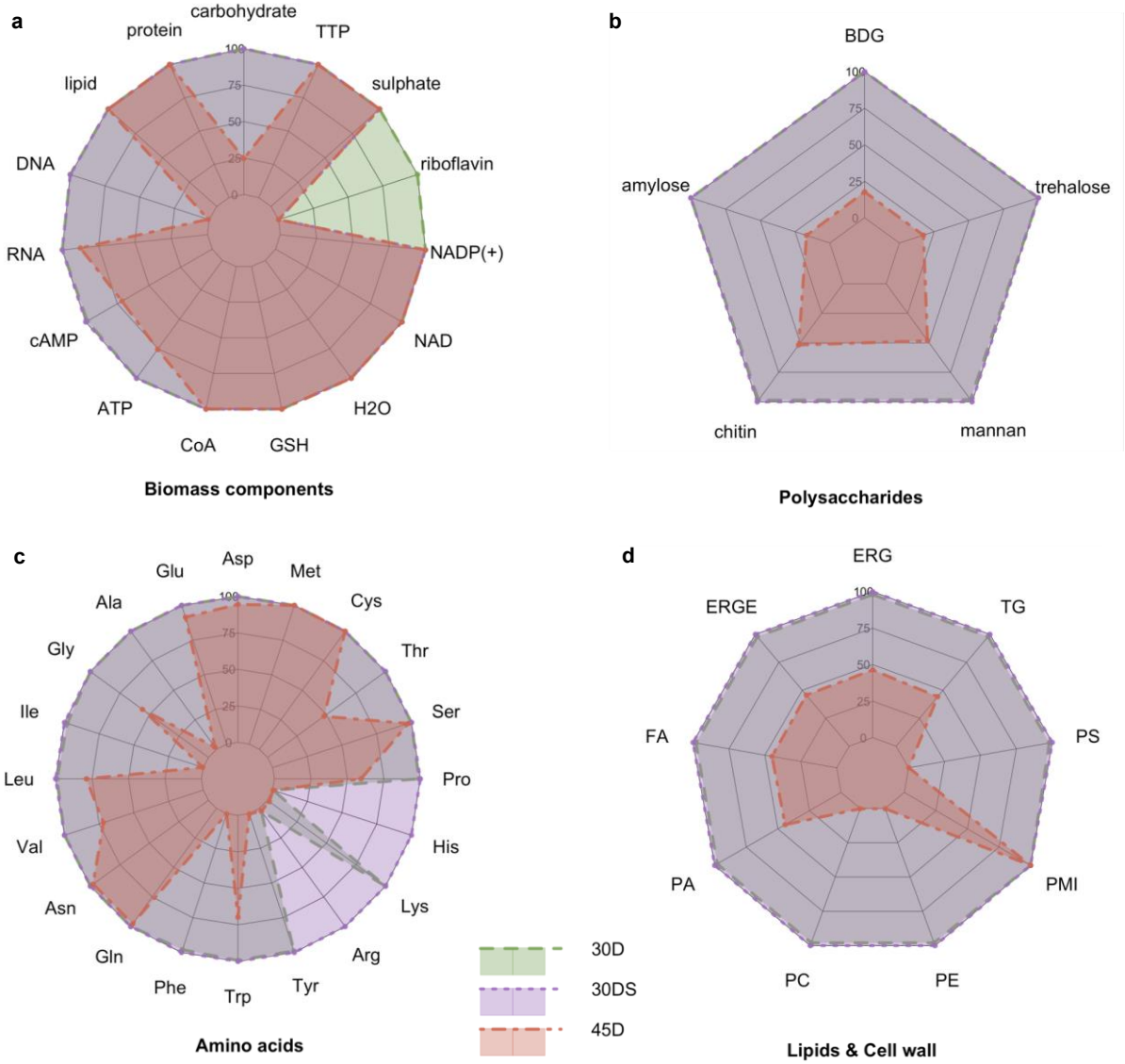


Figure 16. A radar chart showing the predicted potential for biomass precursors excessive production in 30D, 30DS and 45D conditions. As the magnitude is different for each metabolite, the relative production values are shown, where 100% indicates the largest production capacity between conditions. The data for the 30D condition is shown as the green polygon bordered with the dashed border while the corresponding data for the 30DS condition is in purple (dotted border) and the data for the 45D condition is in red (dot-dash border) colour. (a) Abbreviations: cAMP (3',5'-cyclic AMP), CoA (coenzyme A), GSH (reduced glutathione), TTP (deoxythymidine 5'-triphosphate). (b) Abbreviations: BDG ((1->3)-beta-D-glucan). (c) Abbreviations (by side-chain class): a) acid: Asp (L-aspartate), Glu (L-glutamate); b) aliphatic: Ala (L-alanine), Gly (glycine), Ile (L-isoleucine), Leu (L-leucine), Val (L-valine); c) amide: Asn (L-asparagine), Gln (L-glutamine); d) aromatic: Phe (L-phenylalanine), Trp (L-tryptophan), Tyr (L-tyrosine); e) basic: Arg (L-arginine), Lys (L-lysine); f) basic aromatic: His (L-histidine); g) Pro (L-proline); hydroxyl-containing: Ser (L-serine), Thr (L-threonine); h) sulphur containing: Cys (L-cysteine), Met (L-methionine). (d) Abbreviations: ergosterol (ERG), ergosterol ester (ERGE), FA (fatty acid), PA (phosphatidate), PC (phosphatidylcholine), PE (phosphatidylethanolamine), PMI (1-phosphatidyl-1D-myo-inositol), PS (phosphatidyl-L-serine), TG (triglyceride).

Regarding the riboflavin auxotrophy in 30DS and 45D, one may identify the other possible advantages in addition to just conserving the building blocks. Two condition-specific strategies for disabling *de novo* riboflavin synthesis were suggested. Figure 17 shows that the main riboflavin precursors GTP (guanosine-5'-triphosphate) and D-ribul 5-P (D-ribulose 5-phosphate) are metabolised in their linear pathways until their final products are used as precursors to produce 67dm81Drl (6,7-dimethyl-8-(1-D-ribityl)lumazine), the direct precursor for riboflavin. In 30DS, the expression of *KmRIB3* was inactivated. This gene catalyses the final reactions of GTP and D-ribul 5-P (D-ribulose 5-phosphate) linear pathways (r_0939 and r_0940 correspondingly). Meanwhile, the cell retained the interconversion ability between riboflavin and downstream metabolites FMN and FAD. Upon 45D, *KmRIB4* and *KmFMN1* genes were not expressed. *KmRIB4* is responsible for 67dm81Drl (6,7-dimethyl-8-(1-D-ribityl)lumazine) production (r_0938), while *KmFMN1* is involved in FMN conversion from riboflavin (r_0937). Upon high temperature, *K. marxianus* could therefore only convert between FMN and FAD or hydrolyse it back to riboflavin. These observations suggested the need for the organism to optimise the required amount for cofactors needed for ATP production (riboflavin) and other metabolic features associated with FMN and FAD. Upon microaerobic conditions, *K. marxianus* could freely produce all three cofactors. Meanwhile, at the high temperature due to the inability to convert riboflavin to FMN and FAD, one may hypothesise that the cell encounters ATP shortage and tries to maximise riboflavin amount. This trait may be further affected by limited ferroheme availability. Riboflavin is known to improve the high-temperature tolerance once added as a supplement upon the low ATP/ADP ratio (Chen *et al.* 2013). In comparison with 30D, the maximal ATP production in 45D was lower by 20%, so the increased ATP demand at high temperature was likely mainly related to NGAM processes if one assumed that ATP production was not limited by cofactors.

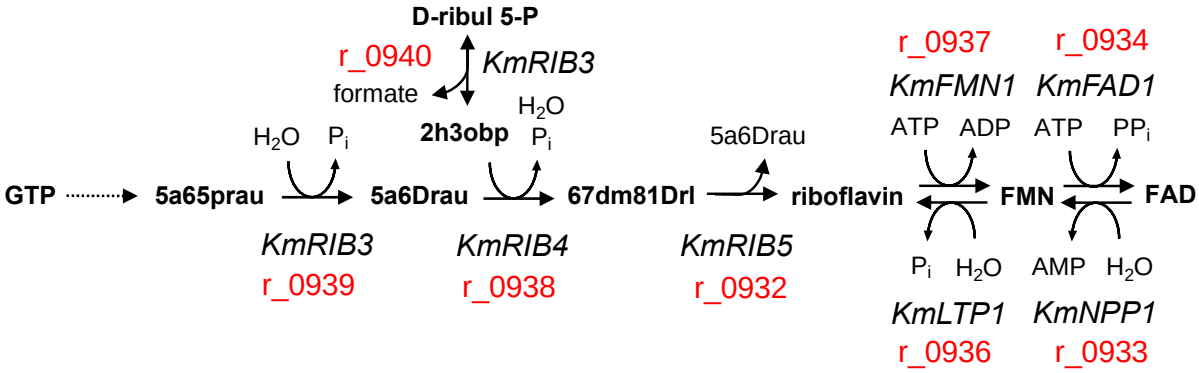


Figure 17. The riboflavin biosynthetic pathway. The gene names are written in italic, while the iSM996 reaction IDs are written in red colour. Abbreviations for metabolites: GTP (guanosine-5'-triphosphate), 5a65prau (5-amino-6-(5-phosphoribitylamino)uracil), 5a6Drau (5-amino-6-(D-ribitylamino)uracil, D-ribul 5-P (D-ribulose 5-phosphate), 2h3obp (2-hydroxy-3-oxobutyl phosphate), 67dm81Drl (6,7-dimethyl-8-(1-D-ribityl)lumazine production). Abbreviations for genes: *KmRIB3* (3,4-dihydroxy-2-butanone-4-phosphate synthase), *KmRIB4* (lumazine synthase), *KmRIB5* (riboflavin synthase), *KmFMN1* (riboflavin kinase), *KmFAD1* (FAD synthetase), *KmLTP1* (putative protein phosphotyrosine phosphatase), *KmNPP1* (nucleotide pyrophosphatase/phosphodiesterase).

Regarding the lipid pool, the same size could be predicted in all three conditions. No composition data were available for essential metabolite myo-inositol, so the arbitrary

abundance value (0.001 mmol/gDW/h) was used during the simulations, what could have the noticeable impact to the predicted lipid pool sizes. At microaerobic and high-temperature conditions the production for the remaining lipid components was restricted by turning off *KmDGK1* gene. This gene was reported to increase the sensitivity to heat when upregulated (Han *et al.* 2008).

The simulations also showed that upon 45D, the size for the deoxynucleotide pool was decreased more than 100-fold when compared with 30D. The reason for such a significant decrease was the lower production capability for dAMP. The suppression of genes *KmISN1* and *KmSDT1* prevented the cell to catabolise purine nucleotides for dAMP production, which played a significant role for 30D and 30DS conditions. This observation indicated that the cell tried to conserve nucleotides from degradation, coinciding with the study, which reported that genes linked to DNA repair were upregulated at high temperature (Lertwattanasakul *et al.* 2015).

Overall, the simulations using condition-specific models suggested that upon high-temperature *K. marxianus* inactivated some of its genes what allowed the cell to conserve the essential metabolites from degradation and to optimise nutrients procurement from medium through the introduced auxotrophies. These results may be used to the growth medium design upon the low oxygen availability and high temperature.

III. Establishing the further development pipeline for iSM996

Many published GEMs are available in standardised Systems Biology Markup Language (SBML) format (Olivier & Bergmann 2018), however, due to the differences in SBML levels, versions and libSBML (Bornstein *et al.* 2008) versions used to export GEMs, it is still a problematic issue to import these GEMs into metabolic modelling tools. To address this issue, the iSM996 model was placed in the GitHub repository (https://github.com/SysBioChalmers/Kluyveromyces_marxianus-GEM) and is further developed under the name of *Kluyveromyces_marxianus-GEM*. While the users can download the latest version of the model, they are also welcomed to report any compatibility or functionality issues, which would be responded promptly.

Part III: RAVEN 2.0, a toolbox for GEM reconstruction and analysis

Paper IV: Development of The RAVEN Toolbox

OBJECTIVES

This study aimed:

- To update RAVEN to version 2.0
- To establish the development policy for RAVEN

MOTIVATION

The RAVEN toolbox is a commonly used systems biology package for GEM reconstruction, analysis, visualisation and omics data integration (Agren *et al.* 2013). However, the absence of the new release since RAVEN v1.08 revealed an increasing number of compatibility issues and bugs. It is therefore reasonable to restore the development for this toolbox and establish a curation policy for fixing the existing bugs and compatibility issues while adding new functionalities.

ANALYSIS, RESULTS AND DISCUSSION

I. The RAVEN Toolbox 2.0

RAVEN 2.0 features the major revision of version 1.08 and is aimed to provide the most efficient *in silico* tools for genome-scale metabolic model reconstruction, curation, visualisation and analysis (Figure 18).

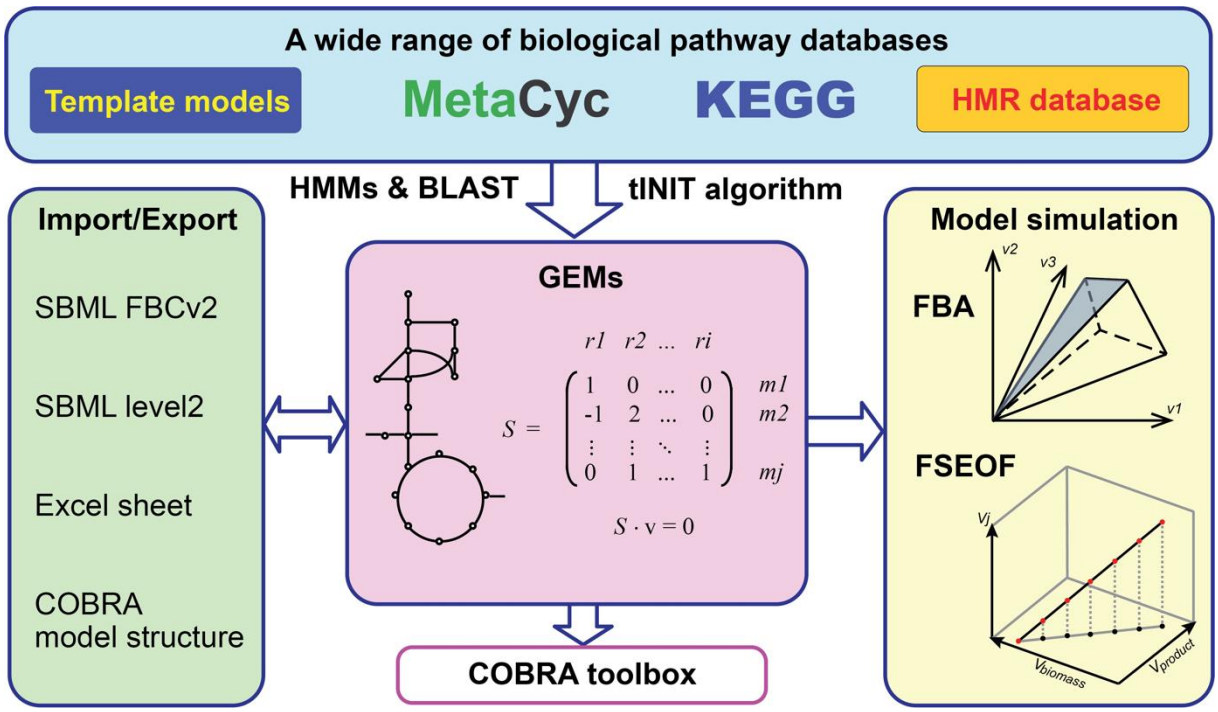


Figure 18. Schematic overview of the RAVEN toolbox version 2.0.

Unlike the previous RAVEN versions, RAVEN 2.0 is compatible with the other MATLAB ("MATLAB version 9.3.0.713579 (R2017b)" 2017) systems biology toolbox COBRA (Hannum *et al.* 2019). This was achieved by renaming the functions with overlapping names and introducing the new function *checkFunctionUniqueness*, which ensures that all the RAVEN function names are unique across all MATLAB paths.

The improvements have been made to the functions for GEM import and export. The function for the model import *importModel* is now compatible with SBML Level 3 and FBC package v2. In previous RAVEN versions, the model export function *exportModel* could only export models in SBML Level 2 Version 1 format but now is capable of exporting models in SBML Level 3 format as well. Since SBML format is not suitable for tracking changes between different model versions, the functions for the model export to plain-text and YAML formats were introduced. Also, the wrapper function *ravenCobraWrapper*, which allows to quickly convert between RAVEN and COBRA model structures in MATLAB environment, was provided in RAVEN.

RAVEN supports three strategies to reconstruct draft GEM: (i) based on tINIT algorithm (Agren *et al.* 2014), (ii) based on existing template model and (iii) based on public reaction databases. The first strategy utilises the *getINITModel* function, requires a GEM for the same species and metabolic tasks, as well as proteomics or transcriptomics data. The second approach requires a GEM for a phylogenetically close species. In such an approach, the functions *getBlast* and *getModelFromHomology* are utilised to run the bi-directional BLASTP search and then build the draft GEM. As an alternative, the users can now select the *getDiamond* function to run DIAMOND blastp search during homology search instead. The third approach involves the *de novo* reconstruction from KEGG or MetaCyc databases. Regarding the GEM reconstruction

from KEGG, users can run the *getKEGGModelForOrganism* function to build a GEM based on KEGG-supplied annotations (KEGG currently includes over 5000 species) or query its protein sequences for similarity to KO specific HMMs. An ability to reconstruct mode from MetaCyc is a new feature in RAVEN 2.0 and can be run with the *getMetaCycModelForOrganism* function that queries protein sequences with BLASTP/DIAMOND blastp for homology against curated enzymes by MetaCyc. The function *addSpontaneous* can be used to retrieve MetaCyc reactions which are not associated with any enzyme. The users who prefer to utilise both KEGG and MetaCyc for *de novo* GEM reconstruction can use the *combineMetaCycKEGGModels* function to merge both draft models.

Some users may also use RAVEN to update the existing GEMs. An example of such approach is shown in Figure 19, where KEGG-based and MetaCyc-based model reconstruction modules were able to suggest the noticeably high number of new candidate reactions to be incorporated into iMK1208 model for *Streptomyces coelicolor*.

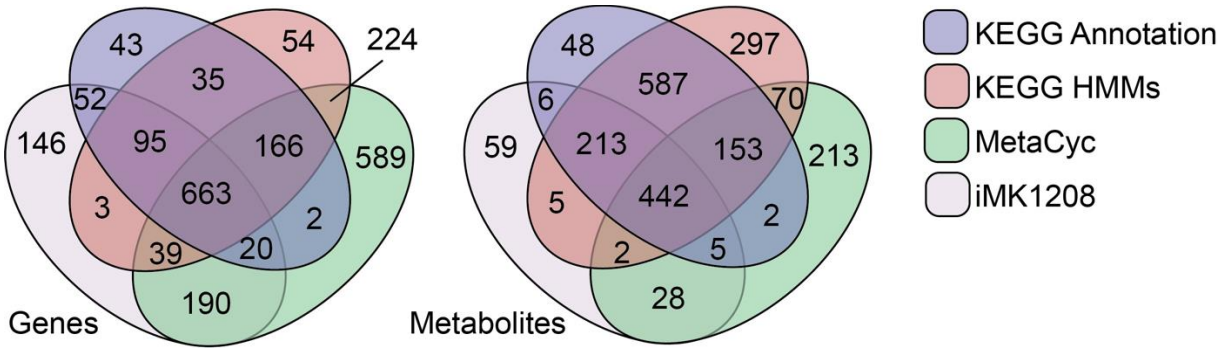


Figure 19. Venn diagrams comparing genes and metabolites between the three *de novo* draft reconstructions and the template GEM iMK1208 for *Streptomyces coelicolor*.

Regardless of the approach used to generate a draft GEM, the further semi-automatic curation is needed to ensure its functionality. It is recommended to start with the *gapReport* function, which performs the gap analysis and reports isolated subnetworks, dead-end reactions and the detailed report about metabolites which can be produced/consumed without any exchange reactions. Other mostly used functions for manual curation are related to the automatic gap filling (i.e. *gapFill*), reaction import from another GEM (i.e. *addRxnsGenesMets*) and GPR rules modification (i.e. *changeGeneAssoc*). Regarding *de novo* generated GEMs, as soon as the gap-filling step is complete, one can run the metabolite compartmentalisation with the *predictLocalization* function. In addition to WoLF PSORT, this function now also supports protein-specific scores from CELLO (Yu *et al.* 2006) and DeepLoc (Almagro Armenteros *et al.* 2017) subcellular localisation prediction tools, which use the protein sequences of target organism as input.

II. Development policy for RAVEN

RAVEN is available from the GitHub repository: https://github.com/SysBioChalmers/RAVEN, where the modelling community can download the latest versions, familiarise with the GEM reconstruction pipelines from the Wiki and report any functionality or compatibility through posting issues. In addition to the regular RAVEN updates once per several months, new released RAVEN version also includes the latest KEGG and MetaCyc versions.

Conclusions and perspectives

Part I showed an effort to estimate the bile acid biotransformation capacity in the human gut and compare it between healthy and IBD patient groups. Firstly, the curated list of candidate BSBP homologues was obtained from UniProt based on sequence homology and domain conservation. The abundance values for BSBGs were then estimated by mapping the faecal metagenomics data of healthy and IBD patient groups to putative BSBPs. The results suggested that the IBD patient group had a lower bile acid biotransformation potential than the healthy group. The follow-up faecal metabolomic analysis suggested the decreased levels of secondary bile acids in IBD patient group, therefore supporting the hypothesis that IBD patients had the lower bile acid bioprocessing potential than the healthy group. While the current study assumed that all BSBPs contributed equally to the bile acid biotransformation potential, it would be beneficial to investigate their kinetic parameters, substrate specificity and contribution in bile acid bioprocessing *in vivo* for further evaluation.

In Part II, thermotolerant yeast *K. marxianus* was annotated in genome-scale and comparatively analysed for the genomic insights of the 12 strains. A total of 5 804 and 3 855 OGs were identified for the pangenome and core genome, respectively. The functional core genome analysis suggested that the most conserved metabolic capabilities were associated to lipid and nucleotide metabolism. The future studies inspired by this approach may involve the more specific comparative genomics analysis, for instance, involving sugar transporters or heat-sensitive proteins.

Paper III comprised the reconstruction and analysis of the iSM996, the first publicly available GEM for *K. marxianus*. The model features 1913 reactions, 996 genes and 1531 metabolites. The iSM996 model could predict the carbon source utilisation and growth rates upon various media. This model is a reliable platform in computational studies requiring experimental data integration and strain design. To evaluate *K. marxianus* metabolic potential in microaerobic conditions and high temperature, the model was coupled with transcriptomics data upon YPD medium. In both microaerobic and high-temperature conditions, the auxotrophy to riboflavin was identified, suggesting that the cell had an increased demand for ATP and tried to ensure that respiration was not limited by cofactors availability. The results also suggested that in high-temperature *K. marxianus* turned off some genes to prevent essential metabolites degradation. Besides, several inactive genes also introduced auxotrophies to several amino acids, suggesting that these amino acids may still be available in the growth medium. Such findings may contribute to the growth medium design upon low oxygen availability and high temperature. Refined predictions may be achieved by integrating additional omics data into the model in the future.

In Part III the new version of RAVEN toolbox was presented. The updated toolbox includes the following key enhancements: (i) *de novo* reconstruction of GEMs based on the MetaCyc pathway database; (ii) a redesigned KEGG-based reconstruction pipeline; (iii) convergence of reconstructions from various sources; (iv) improved performance, usability, and compatibility with the COBRA Toolbox. The future updates of RAVEN will introduce the full compatibility with the COBRA model structure.

The results shown in this thesis did not cover all the approaches which were performed during the PhD project time, such as the identification and annotation of *K. marxianus* mitochondrial genomes. While it was possible to identify mitochondrial genomes, it seemed complicated to verify the predict mitochondrial gene structures, particularly the ones which comprised several subunits. Regarding metabolic modelling for *K. marxianus*, the efforts have been made to restrict the model solution space by adding the protein stability values calculated for the higher temperatures. However, such an approach was cancelled due to technical difficulties. Another cancelled approach involved the integration of the metabolomics data obtained during a kefir fermentation process. Such analysis would show the metabolic capabilities for *K. marxianus* during various time points of kefir fermentation, but this analysis was not continued due to the high uncertainty.

Overall, this thesis illustrated several *in silico* ways to approach food-related microbial species. In metagenomics study, the cohort of human gut microbial species was annotated in a context of bile acid metabolism, which enabled the evaluation of bile acid biotransformation potential and its effect to the human gut microbiome balance. The genome-scale functional analyses for the thermotolerant yeast *K. marxianus* included the comparative genomics for different strains, reconstruction of a genome-scale metabolic model and its utilisation to predict metabolic capabilities upon stress conditions. Also, the new version of RAVEN toolbox was introduced, allowing to perform genome-scale reconstruction and analysis for any newly sequenced species.

Acknowledgements

I hereby would like to thank my supervisor Jens sincerely for acceptance, inspiration, mentoring and continuous support during doctoral studies. Regardless of the difficulties, he always maintained a constructive mindset and highlighted the strengths of my work even when I thought the opposite. Special thanks to my co-supervisor Boyang, who was a good listener, always open for the discussions and gave me many great suggestions during the years.

It had been a pleasure to contribute to SysMilk project where we aimed to characterise the kefir microbial community. I wish to thank Yongkyu, Sanja, Kiran and other consortium members for experimental data and fruitful discussions during our meetings.

I want to express an appreciation to the great colleagues Hao and Eduard. I think we share the same long-term passion and commitment towards the development of RAVEN and hope our collaboration to continue. Hao, I would also like to thank you for taking the time to read the thesis, your comments allowed to improve the thesis quality significantly. A sincere thanks go to Promi for being an inspiring colleague whom I acknowledge for discussions, inspiration and collaboration in bile acid metabolism study – maybe it had a slow start but yielded nicely.

The division of Systems and Synthetic Biology is a multicultural community consisting of inspiring researchers and devoted administration, built to educate and science while respecting individualities. I cannot thank enough for being in such a brilliant group, administrative/dry lab support and all the feedback I received during the presentations, subgroup meetings and chatters.

I wish to thank with love to my wife Simona, parents, family members and friends for love, care, support, encouragement, understanding and acceptance during these years. I am blessed to have such support.

References

Agren, R., Liu, L., Shoaie, S., Vongsangnak, W., Nookaew, I., & Nielsen, J. (2013). The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for *Penicillium chrysogenum*. *PLoS Computational Biology*, **9**(3), e1002980.

Agren, R., Mardinoglu, A., Asplund, A., Kampf, C., Uhlen, M., & Nielsen, J. (2014). Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Molecular Systems Biology*, **10**(3), 721.

Almagro Armenteros, J. J., Sonderby, C. K., Sonderby, S. K., Nielsen, H., & Winther, O. (2017). DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics (Oxford, England)*, **33**(21), 3387–3395.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology*, **215**(3), 403–410.

Andrews, S., & others. (2010). FastQC: a quality control tool for high throughput sequence data, Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Aung, H. W., Henry, S. A., & Walker, L. P. (2013). Revising the Representation of Fatty Acid, Glycerolipid, and Glycerophospholipid Metabolism in the Consensus Model of Yeast Metabolism. *Industrial Biotechnology*, **9**(4), 215–228.

Bairoch, A. (2000). The ENZYME database in 2000. *Nucleic Acids Research*, **28**(1), 304–305.

Bankevich, A., Nurk, S., Antipov, D., … Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, **19**(5), 455–477.

Bansal, S., Oberoi, H. S., Dhillon, G. S., & Patil, R. T. (2008). Production of β-galactosidase by *Kluyveromyces marxianus* MTCC 1388 using whey and effect of four different methods of enzyme extraction on β-galactosidase activity. *Indian Journal of Microbiology*, **48**(3), 337–341.

Barron, N., Marchant, R., McHale, L., & McHale, A. P. (1995). Partial characterization of β-glucosidase activity produced by *Kluyveromyces marxianus* IMB3 during growth on cellobiose-containing media at 45°C. *Biotechnology Letters*, **17**(10), 1047–1050.

Bäumler, A. J., & Sperandio, V. (2016). Interactions between the microbiota and pathogenic bacteria in the gut. *Nature*, **535**(7610), 85–93.

Bender, J. P., Mazutti, M. A., De Oliveira, D., Di Luccio, M., & Treichel, H. (2006). Inulinase Production by *Kluyveromyces marxianus* NRRL Y-7571 Using Solid State Fermentation. *Applied Biochemistry and Biotechnology*, **132**(1–3), 951–958.

Bik, E. M., Eckburg, P. B., Gill, S. R., … Relman, D. A. (2006). Molecular analysis of the bacterial microbiota in the human stomach. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(3), 732–737.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**(15), 2114–2120.

Bornstein, B. J., Keating, S. M., Jouraku, A., & Hucka, M. (2008). LibSBML: an API library for SBML. *Bioinformatics*, **24**(6), 880–881.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, **12**(1), 59–60.

Burke, D. G., Fouhy, F., Harrison, M. J., … Ross, R. P. (2017). The altered gut microbiota in adults with cystic fibrosis. *BMC Microbiology*, **17**(1), 58.

Bushnell, B. (2014). *BBMap: a fast, accurate, splice-aware aligner*.

Cani, P. D. (2018). Human gut microbiome: hopes, threats and promises. *Gut*, **67**(9), 1716–1725.

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)*, **25**(15), 1972–1973.

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

Chen, J., Shen, J., Solem, C., & Jensen, P. R. (2013). Oxidative stress at high temperatures in *Lactococcus lactis* due to an insufficient supply of riboflavin. *Applied and Environmental Microbiology*, **79**(19), 6140–6147.

Cheng, L. K., O'Grady, G., Du, P., Egbuji, J. U., Windsor, J. A., & Pullan, A. J. (2010). Gastrointestinal system. *Wiley Interdisciplinary Reviews. Systems Biology and Medicine*, **2**(1), 65–79.

Claesson, M. J., Cusack, S., O'Sullivan, O., … O'Toole, P. W. (2011). Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proceedings of the National Academy of Sciences of the United States of America*, **108 Suppl**(Suppl 1), 4586–4591.

Conlon, M. A., & Bird, A. R. (2014). The impact of diet and lifestyle on gut microbiota and human health. *Nutrients*, **7**(1), 17–44.

Conly, J. M., & Stein, K. (1992). The production of menaquinones (vitamin K2) by intestinal bacteria and their role in maintaining coagulation homeostasis. *Progress in Food & Nutrition Science*, **16**(4), 307–343.

Dawson, P. A., & Karpen, S. J. (2015). Intestinal transport and metabolism of bile acids. *Journal of Lipid Research*, **56**(6), 1085–1099.

de Aguiar Vallim, T. Q., Tarling, E. J., & Edwards, P. A. (2013). Pleiotropic roles of bile acids in metabolism. *Cell Metabolism*, **17**(5), 657–669.

Demir, H. (2020). Comparison of traditional and commercial kefir microorganism

compositions and inhibitory effects on certain pathogens. *International Journal of Food Properties*, **23**(1), 375–386.

den Besten, G., van Eunen, K., Groen, A. K., Venema, K., Reijngoud, D.-J., & Bakker, B. M. (2013). The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. *Journal of Lipid Research*, **54**(9), 2325–2340.

Dias, O., Pereira, R., Gombert, A. K., Ferreira, E. C., & Rocha, I. (2014). iOD907, the first genome-scale metabolic model for the milk yeast *Kluyveromyces lactis*. *Biotechnology Journal*, **9**(6), 776–790.

Dufresne, S. F., Marr, K. A., Sydnor, E., … Neofytos, D. (2014). Epidemiology of *Candida kefyr* in patients with hematologic malignancies. *Journal of Clinical Microbiology*, **52**(6), 1830–1837.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, **7**(10), e1002195.

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, **30**(1), 207–210.

Elbourne, L. D. H., Tetu, S. G., Hassan, K. A., & Paulsen, I. T. (2017). TransportDB 2.0: A database for exploring membrane transporters in sequenced genomes from all domains of life. *Nucleic Acids Research*, **45**(D1), D320–D324.

Finn, R. D., Bateman, A., Clements, J., … Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Research*, **42**(Database issue), D222-30.

Flores, C. L., & Gancedo, C. (2018). Construction and characterization of a *Saccharomyces cerevisiae* strain able to grow on glucosamine as sole carbon and nitrogen source. *Scientific Reports*, **8**(1), 1–10.

Fonseca, G. G., Gombert, A. K., Heinzle, E., & Wittmann, C. (2007). Physiology of the yeast *Kluyveromyces marxianus* during batch and chemostat cultures with glucose as the sole carbon source. *FEMS Yeast Research*, **7**(3), 422–435.

Fonseca, G. G., Heinzle, E., Wittmann, C., & Gombert, A. K. (2008). The yeast *Kluyveromyces marxianus* and its biotechnological potential. *Applied Microbiology and Biotechnology*, **79**(3), 339–354.

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)*, **28**(23), 3150–3152.

Gensollen, T., Iyer, S. S., Kasper, D. L., & Blumberg, R. S. (2016). How colonization by microbiota in early life shapes the immune system. *Science (New York, N.Y.)*, **352**(6285), 539–544.

Gnerre, S., Maccallum, I., Przybylski, D., … Jaffe, D. B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(4), 1513–

1518.

Gothe, F., Beigel, F., Rust, C., Hajji, M., Koletzko, S., & Freudenberg, F. (2014). Bile acid malabsorption assessed by 7 alpha-hydroxy-4-cholesten-3-one in pediatric inflammatory bowel disease: correlation to clinical and laboratory findings. *Journal of Crohn's & Colitis*, **8**(9), 1072–1078.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics (Oxford, England)*, **29**(8), 1072–1075.

Haas, B. J., Salzberg, S. L., Zhu, W., … Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology*, **9**(1), R7.

Hahn, C., Bachmann, L., & Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads-a baiting and iterative mapping approach. *Nucleic Acids Research*, **41**(13), e129.

Han, G. S., O'Hara, L., Siniossoglou, S., & Carman, G. M. (2008). Characterization of the yeast DGK1-encoded CTP-dependent diacylglycerol kinase. *Journal of Biological Chemistry*, **283**(29), 20443–20453.

Han, S. W., Evans, D. G., El-Zaatari, F. A., Go, M. F., & Graham, D. Y. (1996). The interaction of pH, bile, and *Helicobacter pylori* may explain duodenal ulcer. *The American Journal of Gastroenterology*, **91**(6), 1135–1137.

Hannum, G., Mo, M. L., Palsson, B. Ø., Feist, A. M., Becker, S. A., & Herrgard, M. J. (2019). Creation and analysis of biochemical constraint- based models using the COBRA Toolbox v.3.0. *Nature Protocols*, **2**(3), 727–738.

Harris, S. C., Devendran, S., Mendez-Garcia, C., … Ridlon, J. M. (2018). Bile acid oxidation by *Eggerthella lenta* strains C592 and DSM 2243(T). *Gut Microbes*, **9**(6), 523–539.

Hirano, S., & Masuda, N. (1981). Transformation of bile acids by *Eubacterium lentum*. *Applied and Environmental Microbiology*, **42**(5), 912–915.

Holscher, H. D. (2017). Dietary fiber and prebiotics and the gastrointestinal microbiota. *Gut Microbes*, **8**(2), 172–184.

Horton, P., Park, K.-J., Obayashi, T., … Nakai, K. (2007). WoLF PSORT: protein localization predictor. *Nucleic Acids Research*, **35**, W585-7.

Huerta-Cepas, J., Forslund, K., Coelho, L. P., … Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular Biology and Evolution*, **34**(8), 2115–2122.

Huxley, R. R., Ansary-Moghaddam, A., Clifton, P., Czernichow, S., Parr, C. L., & Woodward, M. (2009). The impact of dietary and lifestyle risk factors on risk of colorectal cancer: a quantitative overview of the epidemiological evidence. *International Journal of Cancer*, **125**(1), 171–180.

Issa, A. T., & Tahergorabi, R. (2019). Milk Bacteria and Gastrointestinal Tract: Microbial Composition of Milk. Microbial Composition of Milk. *Dietary Interventions in Gastrointestinal Diseases: Foods, Nutrients, and Dietary Supplements*, 265–275.

Issa Isaac, N., Philippe, D., Nicholas, A., Raoult, D., & Eric, C. (2019). Metaproteomics of the human gut microbiota: Challenges and contributions to other OMICS. *Clinical Mass Spectrometry*, **14**, 18–30.

Itoh, M., Wada, K., Tan, S., Kitano, Y., Kai, J., & Makino, I. (1999). Antibacterial action of bile acids against *Helicobacter pylori* and changes in its ultrastructural morphology: effect of unconjugated dihydroxy bile acid. *Journal of Gastroenterology*, **34**(5), 571–576.

Jackman, S. D., Vandervalk, B. P., Mohamadi, H., … Birol, I. (2017). ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome Research*, **27**(15), 768–777.

Jansson, J., Willing, B., Lucio, M., … Schmitt-Kopplin, P. (2009). Metabolomics reveals metabolic biomarkers of Crohn's disease. *PloS One*, **4**(7), e6386.

Jones, P., Binns, D., Chang, H.-Y., … Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics (Oxford, England)*, **30**(9), 1236–1240.

Jones, B. V, Begley, M., Hill, C., Gahan, C. G. M., & Marchesi, J. R. (2008). Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. *Proceedings of the National Academy of Sciences of the United States of America*, **105**(36), 13580–13585.

Kall, L., Krogh, A., & Sonnhammer, E. L. L. (2005). An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics (Oxford, England)*, **21 Suppl 1**, i251-7.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**(4), 772–780.

Khani, S., Hosseini, H. M., Taheri, M., Nourani, M. R., & Imani Fooladi, A. A. (2012). Probiotics as an alternative strategy for prevention and treatment of human diseases: a review. *Inflammation & Allergy Drug Targets*, **11**(2), 79–89.

Kim, J., Kim, M. S., Koh, A. Y., Xie, Y., & Zhan, X. (2016). FMAP: Functional Mapping and Analysis Pipeline for metagenomics and metatranscriptomics studies. *BMC Bioinformatics*, **17**(1), 420.

Korpela, K., & de Vos, W. M. (2018). Early life colonization of the human gut: microbes matter everywhere. *Current Opinion in Microbiology*, **44**, 70–78.

Kroger, M., Kurmann, J. A., & Rasic, J. L. (1992). Fermented Milks—Past, Present, and Future. In *Applications of Biotechnology to Fermented Foods*, Washington, D.C.: National Academy Press, pp. 61–62.

Labbe, A., Ganopolsky, J. G., Martoni, C. J., Prakash, S., & Jones, M. L. (2014). Bacterial bile metabolising gene abundance in Crohn's, ulcerative colitis and type 2 diabetes metagenomes. *PloS One*, **9**(12), e115175.

Lachance, M. A. (2011). Kluyveromyces *van der Walt (1971)*. *The Yeasts*, Vol. 2, Elsevier B.V. doi:10.1016/B978-0-444-52149-1.00035-5

Lagesen, K., Hallin, P., Rødland, E. A., Staerfeldt, H.-H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, **35**(9), 3100–3108.

Lane, M. M., & Morrissey, J. P. (2010). *Kluyveromyces marxianus*: A yeast emerging from its sister's shadow. *Fungal Biology Reviews*, **24**(1–2), 17–26.

Le Gall, G., Noor, S. O., Ridgway, K., … Narbad, A. (2011). Metabolomics of fecal extracts detects altered metabolic activity of gut microbiota in ulcerative colitis and irritable bowel syndrome. *Journal of Proteome Research*, **10**(9), 4208–4218.

Lechner, M., Findeiss, S., Steiner, L., Marz, M., Stadler, P. F., & Prohaska, S. J. (2011). Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics*, **12**, 124.

Leite, A. M. de O., Miguel, M. A. L., Peixoto, R. S., Rosado, A. S., Silva, J. T., & Paschoalin, V. M. F. (2013). Microbiological, technological and therapeutic properties of kefir: A natural probiotic beverage. *Brazilian Journal of Microbiology*, **44**(2), 341–349.

Lertwattanasakul, N., Kosaka, T., Hosoyama, A., … Yamada, M. (2015). Genetic basis of the highly efficient yeast *Kluyveromyces marxianus*: complete genome sequence and transcriptome analyses. *Biotechnology for Biofuels*, **8**(1), 1–14.

Ley, R. E., Bäckhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D., & Gordon, J. I. (2005). Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(31), 11070–11075.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., & Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, **31**(10), 1674–1676.

Li, D., Wang, P., Wang, P., Hu, X., & Chen, F. (2016). The gut microbiota: A treasure for human health. *Biotechnology Advances*, **34**(7), 1210–1224.

Li, G., Ji, B., & Nielsen, J. (2019). The pan-genome of *Saccharomyces cerevisiae*. *FEMS Yeast Research*, **19**(7). doi:10.1093/femsyr/foz064

Lieven, C., Beber, M. E., Olivier, B. G., … Ma, H. (2018). Memote: A community driven effort towards a standardized genome-scale metabolic model test suite, 1–26.

Litvak, Y., Byndloss, M. X., & Bäumler, A. J. (2018). Colonocyte metabolism shapes the gut microbiota. *Science (New York, N.Y.)*, **362**(6418). doi:10.1126/science.aat9076

Lopitz-Otsoa, F., Rementeria, A., Elguezabal, N., & Garaizar, J. (2006). Kefir: a symbiotic

yeasts-bacteria community with alleged healthy capabilities. *Revista Iberoamericana de Micologia*, **23**(2), 67–74.

Louis, P., & Flint, H. J. (2017). Formation of propionate and butyrate by the human colonic microbiota. *Environmental Microbiology*, **19**(1), 29–41.

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, **25**(5), 955–964.

Luo, R., Liu, B., Xie, Y., … Wang, J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience*, **1**(1), 18.

Lutgendorff, F., Akkermans, L. M. A., & Söderholm, J. D. (2008). The role of microbiota and probiotics in stress-induced gastro-intestinal damage. *Current Molecular Medicine*, **8**(4), 282–298.

Maccaferri, S., Klinder, A., Brigidi, P., Cavina, P., & Costabile, A. (2012). Potential probiotic *Kluyveromyces marxianus* B0399 modulates the immune response in Caco-2 cells and peripheral blood mononuclear cells and impacts the human gut microbiota in an in vitro colonic model system. *Applied and Environmental Microbiology*, **78**(4), 956–964.

Marcobal, A., Kashyap, P. C., Nelson, T. A., … Sonnenburg, J. L. (2013). A metabolomic view of how the human gut microbiota impacts the host metabolome using humanized and gnotobiotic mice. *The ISME Journal*, **7**(10), 1933–1943.

Mardinoglu, A., Agren, R., Kampf, C., Asplund, A., Uhlen, M., & Nielsen, J. (2014). Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nature Communications*, **5**(May 2013), 1–11.

Martin, G., Kolida, S., Marchesi, J. R., Want, E., Sidaway, J. E., & Swann, J. R. (2018). <i>In Vitro<i/> Modeling of Bile Acid Processing by the Human Fecal Microbiota. *Frontiers in Microbiology*, **9**, 1153.

MATLAB version 9.3.0.713579 (R2017b). (2017), Natick, Massachusetts.

Medema, M. H., Blin, K., Cimermancic, P., … Breitling, R. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research*, **39**(Web Server issue), W339-46.

Medema, M. H., Kottmann, R., Yilmaz, P., … Glöckner, F. O. (2015). Minimum Information about a Biosynthetic Gene cluster. *Nature Chemical Biology*, **11**(9), 625–631.

Merino, N., Aronson, H. S., Bojanova, D. P., … Giovannelli, D. (2019). Living at the Extremes: Extremophiles and the Limits of Life in a Planetary Context. *Frontiers in Microbiology*, **10**, 780.

Mira, N. P., Münsterkötter, M., Dias-Valada, F., … Sá-Correia, I. (2014). The Genome Sequence of the Highly Acetic Acid-Tolerant *Zygosaccharomyces bailii*-Derived Interspecies Hybrid Strain ISA1307, Isolated from a Sparkling Wine Plant. *DNA Research*, **21**(3), 299–313.

Mullish, B. H., Pechlivanis, A., Barker, G. F., Thursz, M. R., Marchesi, J. R., & McDonald, J. A. K. (2018). Functional microbiomics: Evaluation of gut microbiota-bile acid metabolism interactions in health and disease. *Methods (San Diego, Calif.)*, **149**, 49–58.

Natividad, J. M. M., & Verdu, E. F. (2013). Modulation of intestinal barrier by intestinal microbiota: pathological and therapeutic implications. *Pharmacological Research*, **69**(1), 42–51.

Nguyen, T. H., Fleet, G. H., & Rogers, P. L. (1998). Composition of the cell walls of several yeast species. *Applied Microbiology and Biotechnology*, **50**(2), 206–212.

Ninane, V., Berben, G., Romnee, J. M., & Oger, R. (2005). Variability of the microbial abundance of a kefir grain starter cultivated in partially controlled conditions. *Biotechnology, Agronomy and Society and Environment*, **9**(3), 191–194.

Oberman, H., & Libudzisz, Z. (1998). Fermented milks. In B. J. B. Wood, ed., *Microbiology of Fermented Foods*, Boston, MA: Springer US, pp. 308–350.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, **27**(1), 29–34.

Olivier, B. G., & Bergmann, F. T. (2018). SBML Level 3 Package: Flux Balance Constraints version 2. *Journal of Integrative Bioinformatics*, **15**(1). doi:10.1515/jib-2017-0082

Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis? *Nature Biotechnology*, **28**(3), 245–8.

Ortiz-Merino, R. A., Varela, J. A., Coughlan, A. Y., … Morrissey, J. P. (2018). Ploidy Variation in *Kluyveromyces marxianus* Separates Dairy and Non-dairy Isolates. *Frontiers in Genetics*, **9**, 94.

Palmer, J., & Stajich, J. (2019). nextgenusfs/funannotate: funannotate v1.5.3. doi:10.5281/ZENODO.2604804

Prado, M. R., Blandón, L. M., Vandenberghe, L. P. S., … Soccol, C. R. (2015). Milk kefir: composition, microbial cultures, biological activities, and related products. *Frontiers in Microbiology*, **6**, 1177.

Quigley, L., McCarthy, R., O'Sullivan, O., … Cotter, P. D. (2013). The microbial content of raw and pasteurized cow milk as determined by molecular approaches. *Journal of Dairy Science*, **96**(8), 4928–4937.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, **26**(6), 841–842.

Rajoka, M. I., & Khan, S. (2005). Hyper-production of a thermotolerant β-xylosidase by a deoxy-D-glucose and cycloheximide resistant mutant derivative of *Kluyveromyces marxianus* PPY 125. *Electronic Journal of Biotechnology*, **8**(2), 177–184.

Rawlings, N. D., Barrett, A. J., Thomas, P. D., Huang, X., Bateman, A., & Finn, R. D. (2018). The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017

and a comparison with peptidases in the PANTHER database. *Nucleic Acids Research*, **46**(D1), D624–D632.

Rea, M. C., Lennartsson, T., Dillon, P., … Cogan, T. M. (1996). Irish kefir-like grains: Their structure, microbial composition and fermentation kinetics. *Journal of Applied Bacteriology*, **81**(1), 83–94.

Rosa, D. D., Dias, M. M. S., Grześkowiak, Ł. M., Reis, S. A., Conceição, L. L., & Peluzio, M. do C. G. (2017). Milk kefir: nutritional, microbiological and health benefits. *Nutrition Research Reviews*, **30**(1), 82–96.

Rosenbaum, M., Knight, R., & Leibel, R. L. (2015). The gut microbiota in human energy homeostasis and obesity. *Trends in Endocrinology and Metabolism: TEM*, **26**(9), 493–501.

Saha, R., Chowdhury, A., & Maranas, C. D. (2014). Recent advances in the reconstruction of metabolic models and integration of omics data. *Current Opinion in Biotechnology*, **29**(1), 39–45.

Sánchez, B., Li, F., Lu, H., Kerkhoven, E., & Nielsen, J. (2019). SysBioChalmers/yeast-GEM: yeast 8.3.4. doi:10.5281/ZENODO.3353593

Savage, D. C. (1977). Microbial ecology of the gastrointestinal tract. *Annual Review of Microbiology*, **31**, 107–133.

Schomburg, I., Jeske, L., Ulbrich, M., Placzek, S., Chang, A., & Schomburg, D. (2017). The BRENDA enzyme information system–From a database to an expert system. *Journal of Biotechnology*, **261**(April), 194–206.

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V, & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics (Oxford, England)*, **31**(19), 3210–3212.

Simova, E., Beshkova, D., Angelov, A., Hristozova, T., Frengova, G., & Spasov, Z. (2002). Lactic acid bacteria and yeasts in kefir grains and kefir made from them. *Journal of Industrial Microbiology & Biotechnology*, **28**(1), 1–6.

Simpson, J. T., & Durbin, R. (2012). Efficient *de novo* assembly of large genomes using compressed data structures. *Genome Research*, **22**(3), 549–556.

Slater, G., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.

Smit, A., & Hubley, R. (2015). RepeatModeler Open-1.0. Retrieved from http://www.repeatmasker.org

Smit, A., Hubley, R., & Green, P. (2015). RepeatMasker Open-4.0. Retrieved from http://www.repeatmasker.org

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)*, **30**(9), 1312–1313.

Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics (Oxford, England)*, **24**(5), 637–644.

Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O., & Borodovsky, M. (2008). Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. *Genome Research*, **18**(12), 1979–1990.

Thiele, I., & Palsson, B. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, **5**(1), 93–121.

Thursby, E., & Juge, N. (2017). Introduction to the human gut microbiota. *The Biochemical Journal*, **474**(11), 1823–1836.

UniProt: a worldwide hub of protein knowledge. (2019). *Nucleic Acids Research*, **47**(D1), D506–D515.

Vardjan, T., Mohar Lorbeg, P., Rogelj, I., & Čanžek Majhenič, A. (2013). Characterization and stability of lactobacilli and yeast microbiota in kefir grains. *Journal of Dairy Science*, **96**(5), 2729–2736.

Verduyn, C., Postma, E., Scheffers, W. A., & Van Dijken, J. P. (1992). Effect of Benzoic Acid on Metabolic Fluxes in Yeasts: A Continuous-Culture Study on the Regulation of Respiration and Alcoholic Fermentation. *Yeast*, **8**(7), 501–517.

Wahlstrom, A., Kovatcheva-Datchary, P., Stahlman, M., Backhed, F., & Marschall, H.-U. (2017). Crosstalk between Bile Acids and Gut Microbiota and Its Impact on Farnesoid X Receptor Signalling. *Digestive Diseases (Basel, Switzerland)*, **35**(3), 246–250.

Wahlstrom, A., Sayin, S. I., Marschall, H.-U., & Backhed, F. (2016). Intestinal Crosstalk between Bile Acids and Microbiota and Its Impact on Host Metabolism. *Cell Metabolism*, **24**(1), 41–50.

Wang, H., Marcišauskas, S., Sánchez, B. J., … Kerkhoven, E. J. (2018). RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on *Streptomyces coelicolor*. *PLoS Computational Biology*, **14**(10). doi:10.1371/journal.pcbi.1006541

Warren, R. L., Sutton, G. G., Jones, S. J. M., & Holt, R. A. (2007). Assembling millions of short DNA sequences using SSAKE. *Bioinformatics (Oxford, England)*, **23**(4), 500–501.

Xu, J., Mahowald, M. A., Ley, R. E., … Gordon, J. I. (2007). Evolution of symbiotic bacteria in the distal human intestine. *PLoS Biology*, **5**(7), e156.

Yarimizu, T., Nonklang, S., Nakamura, J., … Akada, R. (2013). Identification of auxotrophic mutants of the yeast *Kluyveromyces marxianus* by non-homologous end joining-mediated integrative transformation with genes from *Saccharomyces cerevisiae*. *Yeast*, **30**(12), 485–500.

Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., & Xu, Y. (2012). dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research*, **40**(Web Server issue), W445-51.

Yokota, A., Fukiya, S., Islam, K. B. M. S., … Ishizuka, S. (2012). Is bile acid a determinant of the gut microbiota on a high-fat diet? *Gut Microbes*, **3**(5), 455–459.

Yu, C.-S., Chen, Y.-C., Lu, C.-H., & Hwang, J.-K. (2006). Prediction of protein subcellular localization. *Proteins*, **64**(3), 643–651.

Zajšek, K., Kolar, M., & Goršek, A. (2011). Characterisation of the exopolysaccharide kefiran produced by lactic acid bacteria entrapped within natural kefir grains. *International Journal of Dairy Technology*, **64**(4), 544–548.

Zerbino, D. R. (2010). Using the Velvet *de novo* assembler for short-read sequencing technologies. *Current Protocols in Bioinformatics*, **Chapter 11**, Unit 11.5.

Zhang, S., & Chen, D.-C. (2019). Facing a new challenge: the adverse effects of antibiotics on gut microbiota and host immunity. *Chinese Medical Journal*, **132**(10), 1135–1138.

Zhu, X., Leung, H. C. M., Chin, F. Y. L., … Wang, Y. (2014). PERGA: a paired-end read guided de novo assembler for extending contigs using SVM and look ahead approach. *PloS One*, **9**(12), e114253.

Zimin, A. V, Marcais, G., Puiu, D., Roberts, M., Salzberg, S. L., & Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics (Oxford, England)*, **29**(21), 2669–2677.

Zoetendal, E. G., Raes, J., van den Bogert, B., … Kleerebezem, M. (2012). The human small intestinal microbiota is driven by rapid uptake and conversion of simple carbohydrates. *The ISME Journal*, **6**(7), 1415–1426.