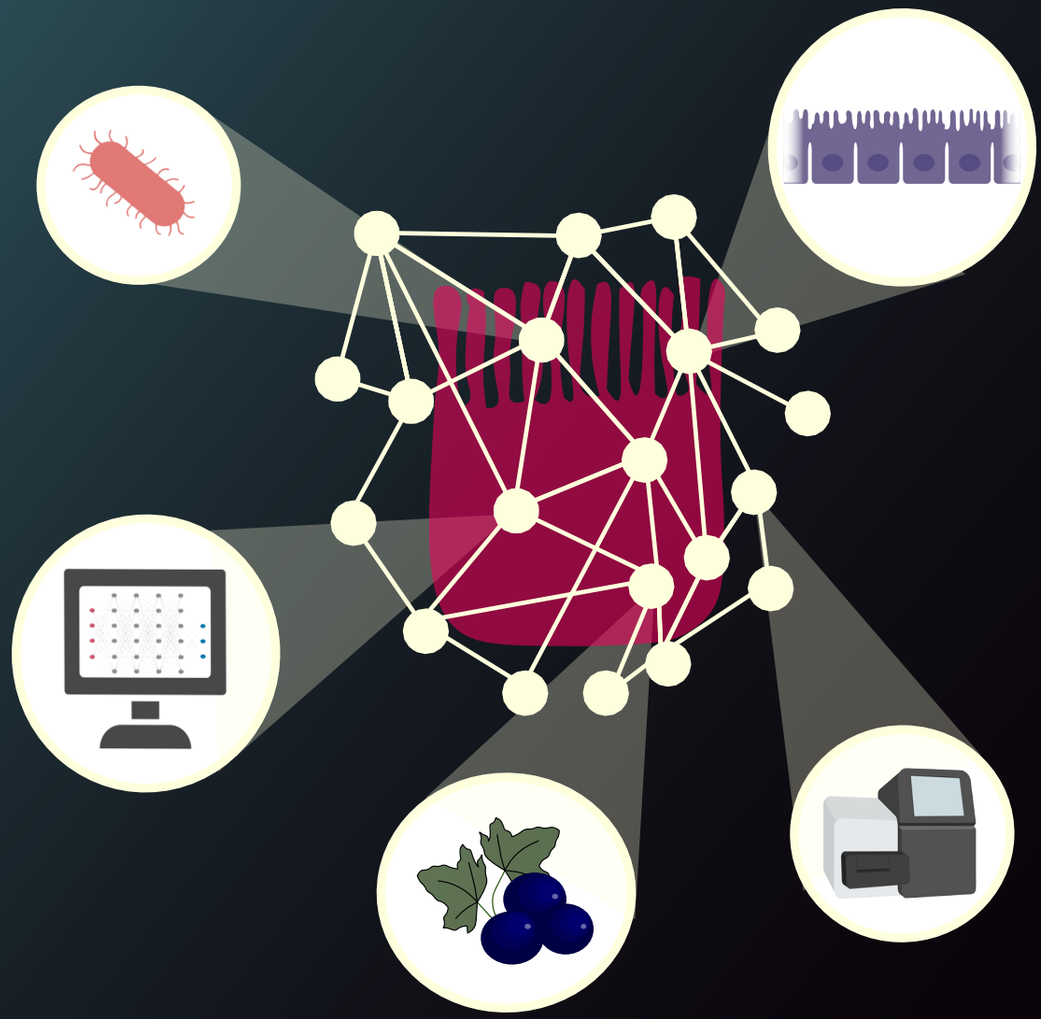
P.B. Venkatasubramanian

2018

A computational approach to study the transcriptional response of enterocytes exposed to luminal factors

A computational approach to study the transcriptional response of enterocytes exposed to luminal factors



Prashanna Balaji
Venkatasubramanian

Propositions:

1. Blackcurrant extracts protect against loss of epithelial integrity caused by *Clostridium difficile* toxins (this thesis).
2. Expression values expressed as fold-change ratios are better than raw values for data fusion and reanalysis of experiments with controls within a batch, particularly to build connectivity maps. (this thesis)
3. Future strategies to manipulate the microbiome can include vesicles containing host miRNAs (Liu *et al*., Cell Host Microbe. 2016 19(1): 32–43).
4. Computational Biology will gain ground over traditional biology in the near future (Nussinov *et al*., PLoS Comput Biol 2015 11(7): e1004318)
5. Biologists and computational biologists should pause and take time to standardise models and data before performing more experiments.
6. Similar to Newton's physical laws leading to the discovery of calculus in mathematics, the laws governing emergence and adaptability in biological systems will lead to a new mathematical technique.
7. Dutch can do without flavourful food but not without playing sports while Indians can do without playing sports but not without flavourful food.
8. Working as the sole computational biologist in a biology lab requires you to do odd jobs such as fixing computer hardware and software and installing computers linked to lab equipment.

Propositions belonging to the thesis, entitled:

A computational approach to study the transcriptional response of enterocytes exposed to luminal factors

Prashanna Balaji Venkatasubramanian

Wageningen, 11 June 2018

A computational approach to study the transcriptional response of enterocytes exposed to luminal factors

Prashanna Balaji Venkatasubramanian

Thesis committee

Promotor

Prof. Dr Jerry M. Wells

Professor of Host Microbe Interactomics

Wageningen University & Research

Co-promotors

Dr Jurriaan J. Mes

Senior Scientist, Fresh Food & Chains

Wageningen University & Research

Dr Nicole J.W. de Wit

Scientist, Fresh Food & Chains

Wageningen University & Research

Dr Edoardo Saccenti

Senior Postdoc, Systems and Synthetic Biology

Wageningen University & Research

Other members

Prof. Dr J. Molenaar, Wageningen University & Research

Dr M.M.W.B. Hendriks, Rosalind Franklin Biotechnology Centre, Delft

Dr A.A.C.M. Peijnenburg, RIKILT, Wageningen University & Research

Dr M.V. Boekschoten, Wageningen University & Research

This research was conducted under the auspices of the Graduate School Wageningen Institute of Animal Sciences

A computational approach to study the transcriptional response of enterocytes exposed to luminal factors

Prashanna Balaji Venkatasubramanian

Thesis

submitted in fulfilment of the requirements for the degree of doctor

at Wageningen University

by the authority of the Rector Magnificus,

Prof. Dr A.P.J Mol,

in the presence of the

Thesis Committee appointed by the Academic Board

to be defended in public

on Monday 11 June 2018

at 1:30 p.m. in the Aula.

Prashanna Balaji Venkatasubramanian

A computational approach to study the transcriptional response of enterocytes exposed to luminal factors,

214 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2018)

With references, summaries in English and Dutch

ISBN: 978-94-6343-874-2

DOI: 10.18174/448425

என் தாத்தா திரு. கணபதி ஐயர் (1928 – 2016) மற்றும்

என் அத்திம்பேர் திரு. கல்யாண்ராம் நாராயண் (1953-2015) அவர்களின் நினைவாக.

என்றென்றும் உங்கள் வாழ்த்துக்களுடன்

Index

General Introduction

The gastrointestinal tract

The gastrointestinal tract is a hollow organ running from the mouth to the anus. The organs that make up the GI tract are the mouth, oesophagus, stomach, small intestine, large intestine, and anus. In addition to the gastrointestinal tract, the liver, pancreas, and gallbladder play a role in the digestion of food and provision of nutrient to the body. The main function of the intestinal tract is the digestion of food and absorption of nutrients, which occurs primarily in the small intestine. Water which is secreted into the small intestine to aid digestion is reabsorbed in the colon to maintain water homeostasis. The large intestine is divided into cecum, colon and rectum which ends at the anus (Figure 1). The small intestine is roughly divided into three segments, the duodenum, jejunum, and ileum (Figure 1). The main function of the duodenum is the digestion and absorption of iron, calcium and water-soluble vitamins whereas the jejunum and ileum are important for absorption of other nutrients and the transport of bile acids and vitamin B12. This specialization in function is reflected in the expression of proteins with specialized functions at specific intestinal locations [1–3]. Additionally, there are location-specific differences in the cell type distribution resulting from stem cell differentiation along the crypt villus axis.

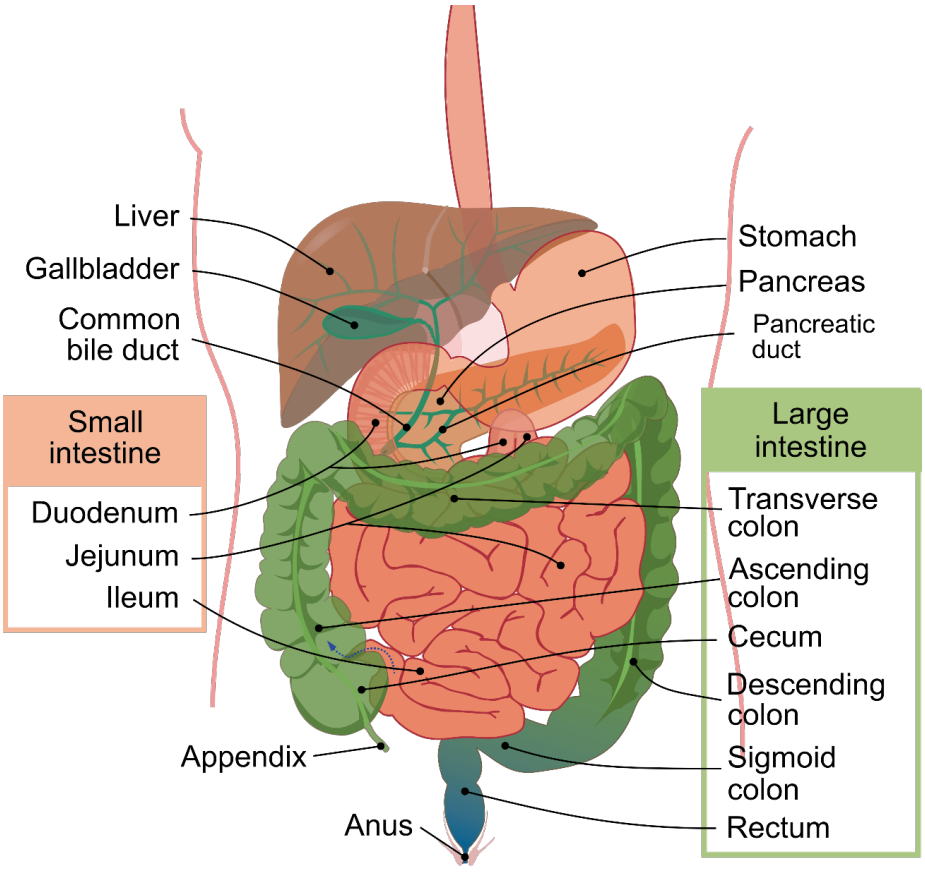


***Figure 1**: A schematic representation of human digestive system showing different segments of small and large intestine. Liver, Stomach, Gall bladder and Pancreas are also shown in figure. Figure adapted from Wikimedia produced by Mariana Ruiz Villarreal.*

Intestinal cell differentiation results from multi-lineage differentiation of adult stem cells residing at the bottom of the crypts. The intestinal epithelium is renewed every 3–5 days in humans, due to apoptosis and exfoliation of mature enterocytes and their replacement by proliferation from stem cells in the crypts. In the small intestine (SI), the epithelium contains goblet cells, enteroendocrine cells, Paneth cells, tuft cells, rapidly dividing enterocytes as well as mature enterocytes [4,5] (Figure 2). Mature enterocytes play a major role in absorption of nutrients [4,5] and consist of about 80 % of the small intestinal epithelia [6]. Goblet cells produce a secreted barrier

(mucous layer) to protect the epithelium from close contact with food and microbes [4]. Goblet cells comprise about 5% of the cells in the small intestinal epithelium and are distributed between the middle of the crypt to the end of the villus [6]. Enteroendocrine cells comprise about 1% of intestinal epithelial cells and produce hormones that are essential for regulating the functions of intestinal epithelium. Recently, enteroendocrine cells were shown to produce opioids in the gut, which might regulate exocrine and endocrine secretion to control digestive and metabolic processes [7,8].

Paneth cells are found in the bottom of epithelial crypts of the SI and possess cytoplasmic granules containing antimicrobial peptides and proteins. These cells play a key role in defence against pathogenic microorganisms, maintenance of stem cells and homeostasis [4]. Tuft cells normally make up about 0.4% of the small intestinal epithelium and possess distinctive 'tufts' of microvilli facing into the intestinal lumen. Two different subtypes of tuft cells and diversity in endocrine cells has been reported in a study using single cell RNAsequencing technique of epithelial cells obtained from mice and mice organoids [9].

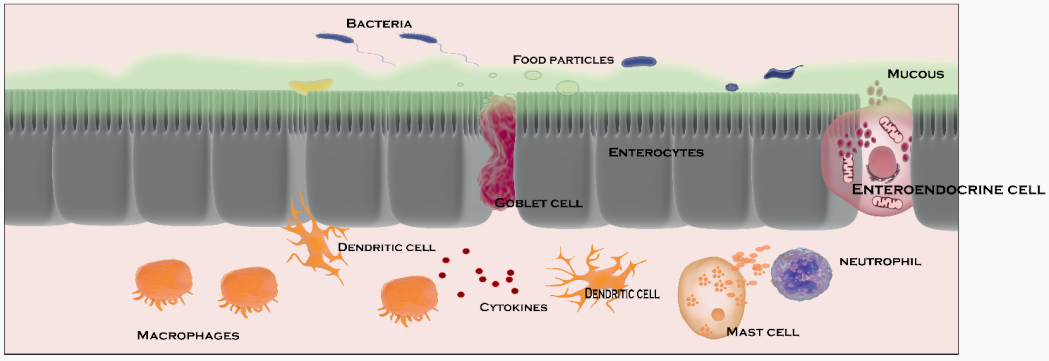


***Figure 2**: A diagram of the intestinal epithelia containing different epithelial cell lineages, the mucous layer and underlying immune cells in the lamina propria.*

Gatekeeper and communication roles of the intestinal epithelium

Apart from its key role in the absorption of nutrients, the intestinal epithelium prevents loss of water, electrolytes and forms a protective barrier against the entry of microorganisms. The co-evolution of mammals with symbiotic intestinal microbiota has resulted in specialised adaptations that avoid chronic inflammatory responses to commensal and symbiotic microbes while maintaining the capacity to fight off pathogens and promote adaptive immune responses. Cells of the epithelial lineage form an innate physical and chemical barrier against microbes, a key element of which is the secreted mucus produced by goblet cells [10] (Figure 2). In the small and large intestine of humans the secreted mucus consists predominantly of MUC2 which is highly glycosylated and negatively charged. Mucus is overall negatively charged and assembled into a network through disulphide bonds linking cysteine rich terminal domains with other monomers to form a hydrogel which limits permeation of microbes, particles and molecules. Constant production and removal of mucus by peristalsis also contributes to its barrier function.

Apart from secreted mucin, intestinal cells produce transmembrane mucins, which are crucial components of the glycocalyx on the apical surface of mucosal epithelium. Expression of the different members of this family of mucins varies along the intestinal tract. Like secreted mucins, transmembrane mucins are extensively O-glycosylated on the extracellular domains to sterically hinder bacterial binding to the cell–surface. Although certain oligosaccharides found on cell-surface mucin may mimic ligands for microbial adhesins they are shed upon binding and thus act as a mechanism to release pathogens from the surface [10].

Intestinal permeability is defined as the functional capacity to regulate passage of molecules across the intestinal wall through the paracellular space between adjacent cells [11]. The paracellular permeability of the epithelium is controlled by protein complexes known as tight junctions (TJs), which reside near the apical surface of adjacent epithelial cells (Figure 3).

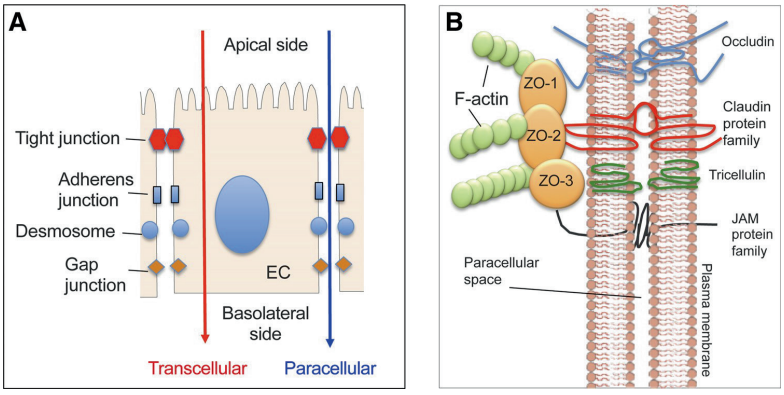


Figure 3: A) *Schematic diagram displaying the cellular junctions between enterocytes. Tight junctions (TJs) allow selective translocation of substances across the epithelial barrier through the paracellular space and regulate permeation of various molecules. Adherens junctions are involved in cell-cell adhesion and intracellular signalling. Other basolateral epithelial junctions include desmosomes and gap junctions, which are involved in cell-cell adhesion and intracellular communication, respectively.* ***B)*** *shows the composition of the TJs, namely: occludin, junctional adhesion molecule (JAM) proteins, and members of the claudin protein family. (Reproduced from Wells et al, 2016* [12]*, Licensed under Creative Commons Attribution CC-BY 3.0:* © *the American Physiological Society)*

TJs prevent the paracellular passage of large molecules through the epithelium while allowing diffusion of ions, water, and small compounds [12]. Other cell junctions are the adherens junctions, desmosomes, and gap junctions, which are involved in cell-cell adhesion and intra-cellular signalling [12]. Epithelial cells can transport receptors to apical or basolateral membranes and TJs prevent their lateral diffusion allowing specific transport or signalling functions to be polarised in the epithelium. Expression of the protein components of TJs varies along the intestinal tract, resulting in location-specific differences in ion transport and water absorption. The occludin and claudins that make up tight junctions, interact with the zonula occludens (ZO) proteins located on the intracellular side of the plasma membrane thereby anchoring them to the actin cytoskeleton.

The breakdown of the barrier function via tight-junction regulation is common to many diseases and infections, particularly in case of bacterial infections caused *E. coli, H. pylori, V. cholera* and *C. difficile* among others [13–15]. Several mice model studies to investigate the relationship between gut barrier impairment and intestinal inflammation has confirmed that leaky gut leads to inflammatory signals using genes responsible for uncontrolled cell death signalling. Abnormality in tight junction structure and increased intestinal permeability are also found to play a role in celiac disease, intestinal bowel disease (IBD), inflammatory bowel syndrome (IBS) [13].

The epithelium plays an important role in transporting secretory immunoglobulin A (sIgA) into the intestinal lumen. Every day an estimated 3 gram of sIgA is secreted into the intestinal lumen highlighting its important role in protecting the mucosal surface. Secretory IgA contributes to the barrier function of the epithelial primarily via agglutination and exclusion of bacteria from the epithelial surface [11].

Another key function of the epithelium which contributes to its gatekeeper function is secretion of several antimicrobial peptides and proteins (AMPs) into the lumen [12,16]. The largest producers AMPs are the enterocytes and Paneth cells found in the small intestine [16]. Expression can be constitutive or induced by innate receptor recognition of microbe-associated molecular patterns. AMPs appear to be retained by the mucus layer covering the epithelium, and only minor levels of these molecules are found in the lumen [16]. Evidence from knock-out and transgenic mice has revealed an important role of AMPs in limiting penetrability of the mucus by pathogenic and commensal bacteria and susceptibility to infection with enteric pathogens [17–19].

Epithelial cells are crucial regulators of immune homeostasis in the mucosa. For example, interactions between microbes, components released by microbes including their metabolites and innate or other receptors expressed by epithelial cells can trigger expression of a surprising diversity of chemokines, cytokines, and peptide hormones by specific cell types in the epithelium [20]. Several studies in knockout mice that demonstrate TLR signalling in the epithelium has a profoundly beneficial role in maintaining homeostasis. This cross-talk regulates several aspects

of immunity including, antigen sampling, B cell function, sIgA production, DC function, and innate immunity.

Caco-2 as an intestinal enterocyte model

Knowledge of the structure, function and role of gut enterocytes is important for gaining a better understanding of the role of the intestinal epithelium in nutrient acquisition, energy metabolism and innate immunity. In a living organism, studying the specific roles of epithelial enterocytes in biological processes is difficult due to the complexity of the intestinal system and therefore model systems are frequently utilized. One such model cell system is the Caco-2 cell line, which has been used in numerous studies to investigate enterocyte functions [21,22]. Caco-2 cells were originally derived from a colon tumour but when they are maintained as confluent monolayer for more than 16 days they resemble the morphology and function of mature absorptive enterocytes as found in the small intestine [23]. Furthermore, monolayers of Caco-2 cells form tight junctions between adjacent cells, allowing the special separation of apical and basolateral receptors, as observed *in vivo*.

In the field of drug discovery, Caco-2 cell line has been useful for studying drug absorption into host cells and in understanding epithelial barrier function [21,22,24–28]. Caco-2 cells have also been used to study the effect of nutritional factors on human intestinal epithelial proliferation, brush border enzyme activity, and motility and other important functions of the intestine [29]. Tallkvist and others investigated the role of specific trans-epithelial membrane proteins responsible for iron transport using Caco-2 model cell system [30]. Another example is the study by Ling and colleagues, who investigated the effect of *Bifidobacterium* on an LPS-stimulated Caco-2 monolayer and repeated the experiment in mice model. The outcome of the Caco-2 experiment indicated that the *Bifidobacterium* may protect against intestinal barrier dysfunction and the same conclusions were drawn from the *in vivo* experiments [31].

In addition to the above mentioned studies, Caco-2 cells have been exposed to various luminal factors and vast amount of microarray based experimental data is available [32]. Some of these experiments include exposure of Caco-2 cells to: vegetables including paprika, broccoli, onion; food compounds sulforaphane and quercetin; pathogens such as *L. monocytogenes* and *C. difficile*; prebiotics such as GOS and fibres; probiotics including *L. casei* and *B. subtilis.* Such a varied dataset could be exploited using systems biology techniques to mine for knowledge and generate novel hypotheses. This is explained in detail in later sections of this chapter.

Clostridium difficile and CDI

The community of microbial organisms that reside within and on multicellular organisms including plants and animals are called microbiota. The microbial colonies present in the microbiota could be commensal, symbiotic or pathogenic in relationship to the host organism. In humans, microbiota communities are found in skin, in the respiratory tracts, mammary glands, uterus, ovarian follicles, placenta, bodily fluids, conjunctiva and the gastrointestinal tract. The intestinal microbiota have been shown to profoundly influence the physiology and metabolism of their hosts [33]. Gut microbial population and their diversity have been studied in detail owing to their contribution to health and disease pathogenicity [34]. The widely accepted dogma related to human microbiome is that the foetus is sterile and the first contact with the microbiota is during birth [34,35]. However, recent studies indicate the presence of microbes in foetus even before birth, in the form of foetal microbiota [35]. Dietary content of fibre, fat and protein are major factors driving the composition and metabolism of the intestinal microbiota [36]. Harmful pathogens can also be ingested via contaminated food or liquids [5,37–39] leading to intestinal disorder and sometimes systemic disease for example, *Campylobacter jejuni, Salmonella typhimurium, Listeria monocytogenes*, pathogenic *Escherichia coli* and *Clostridium difficile*.

Clostridium difficile is a spore-forming gram positive bacteria that is present in air, soil and water and some strains are toxigenic and harmful to humans [40–42]. *C. difficile* infection (CDI) can cause

diseases ranging from diarrhoea to life-threatening inflammation of the colon. Illness from *C. difficile* most commonly affects patients in hospitals or in long-term care facilities and typically occurs after use of antibiotic medications. The use of antibiotics disrupts the ecology of the human microbiome and the recurrence of *C. difficile* infection is associated with the low microbiome diversity [43]. Studies indicate that antibiotic treatment diminishes the products of bacterial metabolism (*e.g.* short-chain fatty acids and secondary bile acids) and there is an increase in the precursors of these by-products such as succinate. *C. difficile* exploits several metabolic pathways to metabolize succinate into butyrate [44]. Investigations are in progress to understand alternative metabolic routes that enable *C. difficile* population expansion after antibiotic treatment [43]. This leaves a narrow set of antibiotics that can be used to treat CDI [45]. Recent studies also indicate emergence of drug resistance in *C. difficile* strains and multi-drug resistance strains have been reported [46] which further emphasizes the need for alternative treatments.

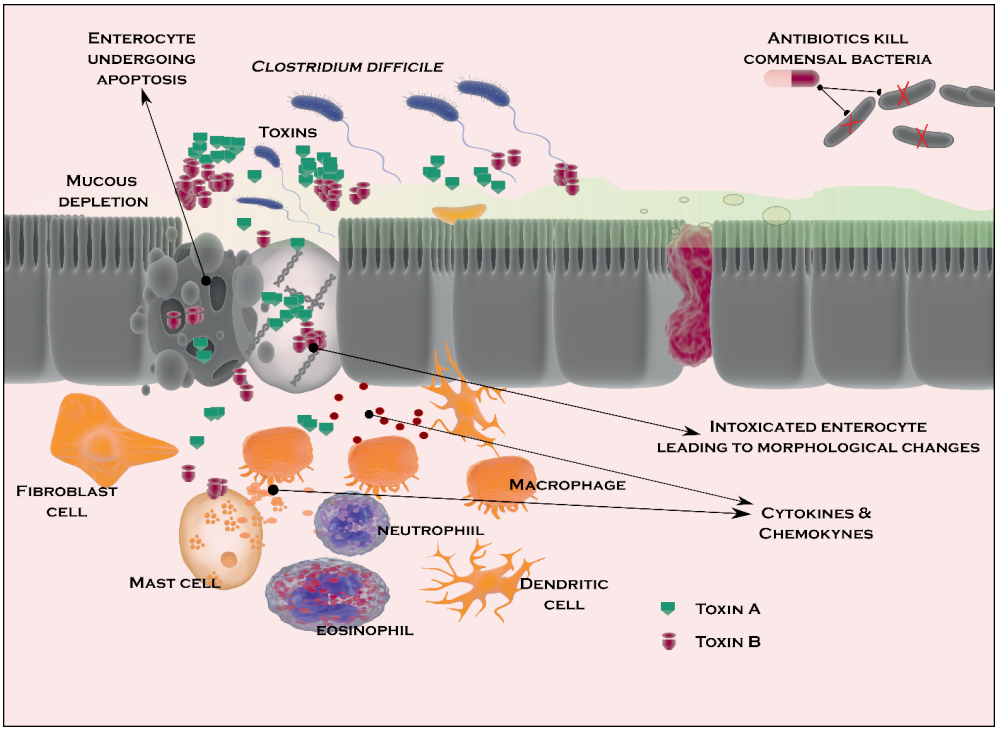


***Figure 4**: A simplified scheme showing the effects of antibiotics killing commensal bacteria followed by C. difficile invasion of intestinal epithelia. The diagram shows a depletion in the mucus layer, rounding up of an enterocyte and another undergoing apoptosis. The release of toxins by C. difficile, intoxication of toxins, disruption of cell structure and release of cytokines and chemokines are also shown along with the accumulation of immune cells.*

Once *C. difficile* populates the gut in sufficient numbers, the activities of enterotoxin - Toxin A (ToxA or TcdA) and cytotoxin - Toxin B (ToxB or TcdB) [47,48] cause diarrhoea and gastroenteritis. ToxA (308 kDa) and ToxB (270 kDa) share high similarity in sequence and functional homology [47] and are both glucosyltransferases that inactivate RHO, RAC and CDC42 proteins within target cells. The toxins show similar substrate specificity owing to the 74% homology in their N-terminal domains. The genes responsible for these two toxins are located on the same pathogenicity locus and are situated close to each other [47]. The general mode of action for both ToxA and ToxB is via receptor mediated endocytosis after which they go on to deactivate RHO, RAC and CDC42 and thus disrupt key cell signalling systems [47]. This leads to break down of the actin dependent cytoskeleton and also dissolution of the tight junctions. The cells start to lose their polarity, followed by rounding up leading to a failure of intestinal barrier. Simultaneously, signals that lead to inflammation and programmed cell death are triggered.

The effects of *C. difficile* and the two toxins have been of increased interests to scientist in recent times, due to the emergence of antibiotic resistant strains. The effects of *C. difficile* toxins on gene expression in the enterocytes has been studied using multiple organisms including human cell model [49–51,15,52,53]. In addition to these studies, *C. difficile* has also been co-cultured with Caco-2 cells (colon like) to study effects on gene expression for up to two hours [15]. Although, the above-mentioned studies have focused on the expression of mRNA in C. difficile induced enterocytes, the role of miRNA and their regulation of mRNA is not known. Moreover, the emergence of drug resistant *C. difficile* strains and the recent discovery of CDI incidence in small intestine warrants an urgent effort to find new effective treatments.

The role of miRNAs in cellular functions and in disease

The miRNAs are 20-25 base pairs in length and are non-coding RNAs [54–56]. They are considered to be mainly involved in mRNA post-transcriptional regulation by targeting 6–8 nucleotides (called as the seed region) [57,58]. miRNAs are initially transcribed as pri-miRNA transcripts, about 60-70 nucleotides (nt) long, forming short hairpin loops by folding back on themselves [57] which are further processed by a micro-processor complex that cleaves the pri-miRNA to form pre-miRNA. Pre-miRNA, which retains the hairpin loop from pri-mRNA, is translocated into the cytosol from the nucleus and the hair pin loop is cleaved by dicer to form a double stranded miRNA. One of the strands of the double stranded miRNA forms a RISC-complex with other proteins which binds to mRNA based on nucleotide complementarity. The transcription is then either mitigated or the mRNA is cleaved and discarded [59,55,57,60].

About 1000 miRNAs have been found in humans. MiRNAs are shown to play a major role in biological processes including metabolism, apoptosis and cell proliferation [61]. Additionally, there is ample evidence of malfunctioning of miRNA in regulation of processes in cancer cells. The miRNAs related to cancer are dysregulated by several mechanisms, including deletion/amplification, epigenetic modifications and impairment of miRNA biogenesis. MiRNAs like miR-15a and miR-16-1 are known to be tumour suppressors that induce apoptosis. Studies revealed them to be deleted in case of B-cell chronic lymphocytic leukaemia [62]. In the case of acute myeloid leukaemia, the expression of miR-223 was silenced by AML1/ETO via epigenetic modification (CpG methylation). Similarly, in case of cardiac hypertrophy, miR-23a, miR-23b, miR-24, miR-195, miR-199a, and miR-214 were reported to be regulated [63]. Studies also report the role of miRNA in immune and autoimmune disorders, liver diseases and neurodevelopmental disorders [61]. In infectious diseases, the role of miRNAs have been elicited by viral infections with Hepatitis C virus (HCV), Human Immunodeficiency Virus-1 (HIV-1), influenza virus [64] among others and in bacterial infections caused by several species of *Mycobacterium* genus, pathogenic *Salmonella*, *L. monocytogenes* and some species of the

genus *Francisella* [65]. Moreover, a recent study indicates that the host secretes miRNA in order to regulate the microbiome population in the gut [66].

miRNAs are being explored as therapeutic targets in treatment against cancer and hepatitis. The expression of miRNA are either suppressed or are targeted using drug molecules that interfere with their normal functioning or miRNAs are used as tumour suppressors [67]. Some therapeutics have reached clinical trials highlighting the potential to treat other diseases through targeting miRNA regulation [67].

Physiological effects of bio-actives present in food

The interaction of food with the digestive system has been studied for a long time because healthy eating habits have a huge potential to mitigate common diseases. This concept was proclaimed in the time of Hippocrates (4th century B.C.E.) and is evident in the historical records of ancient civilisations including Indians (Ayurveda and Siddha systems of medicine), Chinese and Greeks. Today the World Health Organisation (WHO) recommends healthy eating habits as an effective approach to reduce chronic diseases [68]. In addition to food consumed on daily basis, there are functional foods which provide beneficial effects beyond basic nutrition [36,69]. These are usually fortified foods that offer supplementary nutrients such as vitamins or minerals [69].

Bioactive compounds of regular plant-based food and their beneficial effects has been investigated over the years**.** The benefits exerted by several key components like polyphenols, cell wall (fibres), carotenoids, glucosinolates among others, on human health have been studied in detail. One of the hypothesis is that polyphenols exert their health benefits by eliminating free radicals, by protecting and regenerating dietary antioxidants (e.g. vitamin E) and by chelating pro-oxidant metals. Studies indicate that the polyphenols have anti-cancer, anti-microbial properties and also help in wound healing [70]. Consumption of food containing high levels of polyphenols, such as wine, has been linked to low incidence of coronary heart diseases. The

polyphenols in wine prevent platelet aggregation and protect the LDLs from free radicals [70]. Similarly, reactive oxidative species (ROS) and reactive nitrogen species (RNS) could oxidise glucose into carbonyl toxic product. This could further attack amino acids forming glycation end products (AGEs) which may trigger inflammation by release of cytokines. The polyphenols act as antioxidants, keeping ROS/RNS in control [70].

Another major beneficial compound obtained from vegetables and fruits are the dietary fibres obtained from the plant cell walls. The plant cell wall is composed of cellulose, hemicellulose, lignin and pectin. Soluble dietary fibres have a high water holding capacity and can form gel in the stomach and thereby increase the viscosity of food digesta. This causes a delay in gastric emptying and slows down the movement of food from stomach to duodenum and transport rates of nutrients to the small intestine [71]. Due to the laxative effects of insoluble dietary fibres, they are recommended as the first line of treatment in case of constipation. Moreover, the incidence of certain cancers have been observed to be lesser in countries with larger consumption of plant based diet in comparison against countries with more meat based diet [71]. It has also been found that the absorption and bioavailability of food substances varies from individual to individual. This may be due to several factors such as epi-genetics, gut microbiota, sex, age and lifestyle among others [72]. These variations may, in turn, reflect in studies involving food related exposure experiments but do not become vividly distinguishable.

Thus, the data derived from exposure experiments requires computational analysis that considers the challenges from various parameters involved in study design. Such studies could be performed at a holistic level with a systems approach. Computational systems biology techniques could prove to be a promising route to unravel the mechanisms behind functional food components on enterocytes.

Systems Biology as an emerging facet of biological sciences

Systems Biology is the study of biological systems as a whole rather than studying the individual parts. Although systems biology and the techniques used in systems biology are not clearly and precisely defined, the above definition forms the basis of systems biology. Systems biology is an amalgamation of several different techniques ranging from biological high-throughput techniques such as RNAseq, microarray, mass spectrometry, metabolomics, proteomics to statistics, mathematics and computational model development and simulation studies.

The Institute of Systems Biology in Seattle states that "Systems biology is based on the understanding that the whole is greater than the sum of the parts. It is a holistic approach to deciphering the complexity of biological systems that starts from the understanding that the networks that form the whole of living organisms are more than the sum of their parts. It is collaborative, integrating many scientific disciplines – biology, computer science, engineering, bioinformatics, physics and others – to predict how these systems change over time and under varying conditions, and to develop solutions to the world's most pressing health and environmental issues."

The origin of systems biology as an idea could be attributed to the French physiologist Claude Bernard in mid nineteenth century [73]. Later, Austrian biologist Ludwig von Bertalanffy put forth his General Systems Theory. Simultaneously, study of enzyme kinetics as a separate branch was already in vogue, in the early 20th century. The work of Hodgkin and Huxley and that of Denis Noble further established the idea of using mathematical models for systemic study [74,75]. Systems theorist Mihajlo Mesarovic in 1966 helped establish systems biology as a distinct discipline of study in the conference titled "*Systems Theory and Biology*" [76].

Systems biology has clearly evolved over the years with the advent of new technologies and high throughput techniques in experimental biology assisted by powerful computers and advanced statistical and mathematical methods. In recent times, high-throughput techniques

like whole genome/transcriptome sequencing and microarray technologies have been coupled with computational procedures for large scale data analysis and together have yielded deeper insights into understanding biological systems [77,78].

Systems biology has been applied in studying diseases like cancer [79], tuberculosis [80,81], arthritis [82] among others; in studying cell systems of both prokaryotes [83–85] and eukaryotes; in understanding organ systems like liver [86], heart [87], gut [88]; and in understanding single cells as systems [89]. Applications of systems biology and bioinformatics in food related studies has been recently termed as 'foodomics' and it's applications have been limited [90]. Additionally, different scales of biological interactions are studied in systems biology using high throughput data.

High-throughput technologies and analysis techniques

Systems biology works on the data – model – hypothesis – data cycle (Figure 5). The new data is used to update the model and generate novel hypothesis. Several high-throughput techniques are being used by systems biologist to gain new insights into biological systems and for generating novel hypotheses that can further be tested by biological experiments. Different scales of analysis are performed in systems biology, namely: genomic, transcriptomic, proteomic, epigenomic, phenomic, metabolomic, cellular, tissue level, organismal level, spatial, temporal, spatiotemporal and others. Also, multiple scales of data and analysis may be combined together in an integromics study [91].

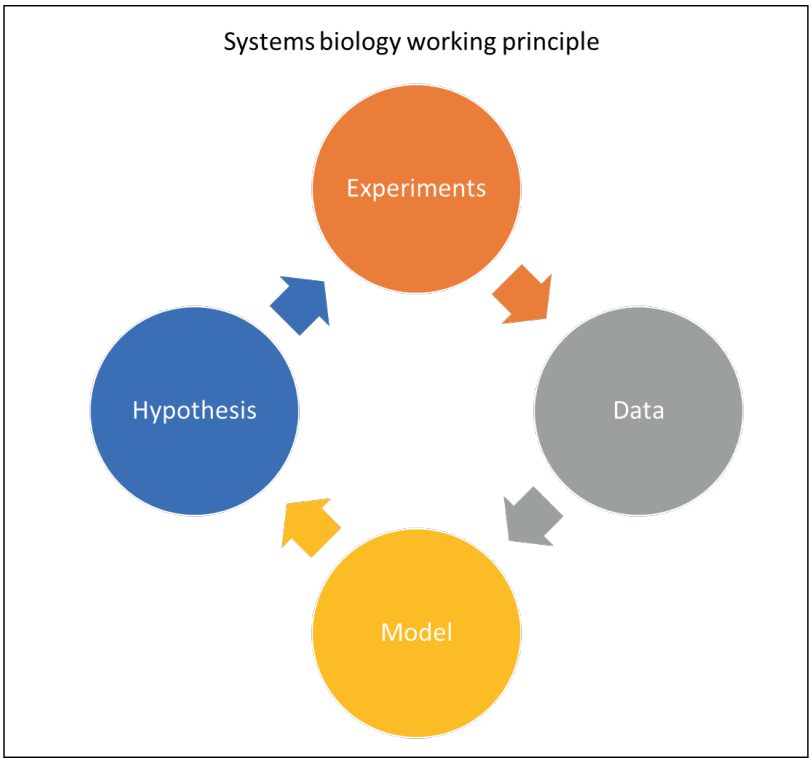


Figure 5*: Representative figure showing the principle of systems biology. System biology works on the principle of an iterative experiment cycle involving experiments – data – model – and hypothesis*

In the past biologists used whole genome sequencing techniques based on capillary sequencing for genome-scale model development and analysis and these have been replaced by high-throughput sequencing technologies such as Illumina dye sequencing and single molecule real time (SMRT) sequencing. Such recent technologies generate vast quantities of data and can be used to discover and study the deoxyribose nucleic acid (DNA) of previously unknown organisms *via* de novo assembly of sequences [92]. Additionally, chromatin immuno-precipitation (ChiP) on DNA microarray chip (ChIP on chip) and ChIP-seq (which is a next generation sequencing technology) allows investigation of protein-DNA binding. Such data can be

combined with transcriptome data to build transcription factor networks and other multiscale networks [93].

The transcriptome is defined as the complete set of Ribose Nucleic Acid (RNA) material present in an organism, cell or tissue. There are many types of RNA and multiple methods are employed to gather the transcriptome information. The most commonly studied RNAs are the coding RNAs that are potentially transcribed into proteins. Polymerase Chain Reaction (PCR) is commonly used for assessing the presence of a piece DNA in a cell [94]. A more robust method for quantifying RNA expression study is quantitative real-time PCR (qRT-PCR) [94]. PCR is a laborious technique to apply at a genome scale *i.e.* to quantify the amount of all genes in a biological sample. Overtime, DNA microarrays revolutionized the field of transcriptome identification and quantification analysis. They could be used to study the content of the entire transcriptome of a biological sample, provided the sequence of the genes were already known [95]. They were used extensively in the past but have recently been surpassed by RNAseq, a next generation sequencing (NGS) technique [96]. However, microarray technology is still commonly used in food or infection related Caco-2 exposure experiments and has been the focus of this thesis.

Methods of data analysis in systems biology

Systems biology involves a wide range of techniques for data analysis and knowledge retrieval depending on the question and data at hand [97–100]. Systems biology approaches are broadly divided into bottom-up and top-down approaches (Figure 6). Bottom-up approaches usually encompass studying and developing models that define lower level abstractions of biological systems from molecular level all the way to organism [101,102]. Typically, such model development approaches begin with generating biological networks for example, gene regulatory networks or signalling networks using molecular data collected from databases and scientific publications. Generally, networks are generated to study a small segment of a cellular process, e.g. a cytokine inducing mechanism. The networks are then used to build computational and

mathematical models that can be utilized to generate an experimentally verifiable hypothesis [102]. The bottom-up approach models typically begin at studying an organelle (*e.g.* mitochondria, nucleus, *etc.)* or a cell type (*e.g.* enterocytes, neurons, *etc.*) and is further expanded into studying tissues and organs [101].

The other common systems biology approach is the top-down modelling approach, which comprises of reverse engineering and developing statistical models of a system using high-throughput data generated from targeted experiments [102]. These experiments are generally performed at whole genome scale and the networks generated from the data address the biological processes within the whole system (genome of a cell/organism).

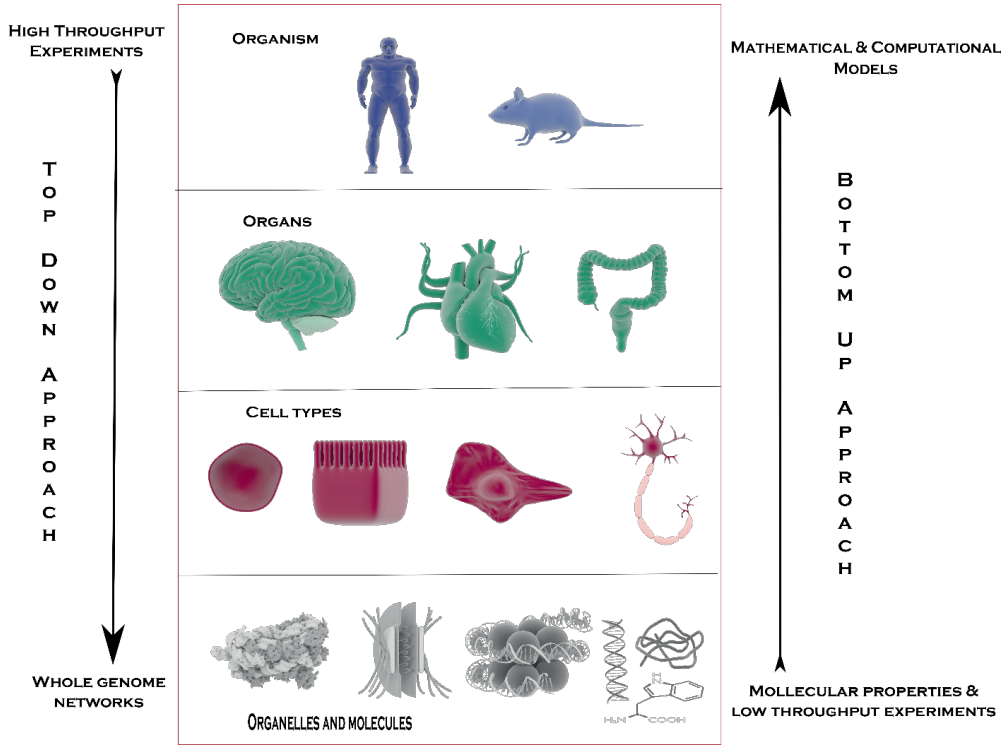


Figure 6: *Illustration of top-down and bottom-up approaches. The top-down approaches usually start with high-throughput experimental data to infer genome-scale network models. Bottom-up approaches start with pathways (or small networks) constructed using properties of different*

molecules in a biological sample and models are developed to generate hypothesis that are used to fill in the knowledge gaps (pathways).

In order to analyse enormous amounts of data generated using high-throughput techniques and large scale low-throughput experiments, statistical procedures are widely used [101]. Statistical techniques such as regression analysis, hypothesis testing, clustering, biclustering and principal component analysis (PCA) are commonly used [103]. Biclustering technique is a data mining technique that enables simultaneous clustering of the samples (columns) and genes (rows) of a biological data matrix [104]. Principal component analysis is a statistical data transformation protocol that converts a set of observations of correlated variables into a set of values of new variables that are linearly uncorrelated. The new variables are called principle components and are obtained using orthogonal transformation of original observations [105].

In addition to these, targeted statistical network generation algorithms are used, for example, ARACNE [106], WGCNA [107] and CLR [108]. ARACNE and CLR use mutual information between genes while WGCNA uses weighted correlation as the measure to determine interaction. While these methods are efficient in single cell organisms, they are not efficient in eukaryotes [109]. Systems biology also involves use of machine learning techniques based on statistics in analysing high-throughput data [110]. Classification models based on Support Vector Machines (SVM) or random forests are known to be useful in systemic analysis [110,111].

Similarly, kinetic models using ordinary differential equations (ODEs) and partial differential equations (PDEs) are extensively used in analysing systems that have rate reactions, usually in drug discovery [112]. Multiple studies have applied control systems theory in reverse engineering biological systems [113,114]. This application also leads to modification of molecular circuits in biological systems and in synthesis of artificial organisms by generating new synthetic genomes [115,116]. In addition to using dynamic models, static models are also used to analyse metabolic networks. Methods like Flux balance analysis (FBA), Flux variability analysis (FVA, an extension

of FBA) and elementary mode analysis use constraint based modelling to generate fluxes at steady states [117].

Batch effects during analysis of large scale data

Living organisms are a complex system and they exhibit properties such as emergence and adaptability. Understanding such systems within the current framework of technologies requires integration of data from different sources. Data collected from diverse sources, from different individuals of same species, inhabiting varied environments, sampled by different people could lead to something known as “batch effects”. The batch effects pose a strong statistical challenge that need to be reconciled before analysis of the data [118–120]. Batch effects are more prominently found in transcriptomics data, particularly from microarray data. Microarray data also have platform-based batch effects and cross-platform analysis of data prove to be difficult. Several methods have been developed to overcome batch effects, although each have their own limitations [121,122]. Most of these techniques focus on data generated from disease samples and are therefore often best suited to mitigate batch effects in disease samples.

Thesis Overview

The aim of this thesis was to develop additional tools and methods to exploit existing transcriptomic data derived from exposure experiments conducted using the intestinal epithelial cell line Caco-2 and utilize the data to find novel applications of food. We used transcriptomic data from exposure of Caco-2 cells to food, food components or microorganisms primarily derived from Affymetrix© based microarrays. To analyse these transcriptomics networks, a protocol was developed to remove batch effects when metadata about an experimental protocol was unknown.

In **Chapter 2**, microarray data collected from transcriptomics studies on Caco-2 cells were used in a correlation-based method to find genes of interest in a pathway for qPCR studies. We

identified genes that are expressed in oxidative stress response pathways (AhR and Nrf2 pathways) in Caco-2 cells on exposure to different varieties of coffee. **Chapter 3** deals with studying the impact of *C. difficile* toxins on colon-like Caco-2 cells, cultured for 7 days, at the molecular level and integrated data from RNAseq experiments and miRNA arrays. Additionally, the impact of *Clostridium* toxins on small intestine-like Caco-2 cells, cultured for 21 days, were explored in **Chapter 4**. Further, a model was developed using microarray data related to food exposure experiments on Caco-2, largely derived from **Chapter 2**, and PCA. This model was applied to find beneficial food substances that may mitigate the effects of *C. difficile* toxins. The Caco-2 microarray data used in this thesis comprised of arrays in which each batch had multiple exposure experiments performed within them with a common control and the metadata information related to the experiments were scarce or unreliable. Combining data obtained from different sources could lead to batch effects. In **Chapter 5**, a solution was developed to address the problems related to batch effects when combining large sets of microarray data. The method was further tested on the Caco-2 microarray data used in earlier chapters (chapter 1 and 3) and on sample data generated in a study on arthritis patients.

References

1. Bates, M. D. *et al.* Novel genes and functional relationships in the adult mouse gastrointestinal tract identified by microarray analysis. *Gastroenterology* **122,** 1467–1482 (2002).

2. Anderle, P. *et al.* Changes in the transcriptional profile of transporters in the intestine along the anterior-posterior and crypt-villus axes. *BMC Genomics* **6,** 69 (2005).

3. Comelli, E. M. *et al.* Biomarkers of human gastrointestinal tract regions. *Mamm. Genome Off. J. Int. Mamm. Genome Soc.* **20,** 516–527 (2009).

4. Gerbe, F., Legraverend, C. & Jay, P. The intestinal epithelium tuft cells: specification and function. *Cell. Mol. Life Sci.* **69,** 2907–2917 (2012).

5. Santaolalla, R. & Abreu, M. T. Innate immunity in the small intestine. *Curr. Opin. Gastroenterol.* **28,** 124–129 (2012).

6. De Santa Barbara, P., Van Den Brink, G. R. & Roberts, D. J. Development and differentiation of the intestinal epithelium. *Cell. Mol. Life Sci.* **60,** 1322–1332 (2003).

7. Bezençon, C. *et al.* Murine intestinal cells expressing Trpm5 are mostly brush cells and express markers of neuronal and inflammatory cells. *J. Comp. Neurol.* **509,** 514–525 (2008).

8. Kokrashvili, Z. *et al.* Release of endogenous opioids from duodenal enteroendocrine cells requires Trpm5. *Gastroenterology* **137,** 598–606.e2 (2009).

9. Haber, A. L. *et al.* A single-cell survey of the small intestinal epithelium. *Nature* **551,** 333–339 (2017).

10. Johansson, M. E. V. & Hansson, G. C. Immunological aspects of intestinal mucus and mucins. *Nat. Rev. Immunol.* **16,** 639–649 (2016).

11. Bischoff, S. C. *et al.* Intestinal permeability – a new target for disease prevention and therapy. *BMC Gastroenterol.* **14,** (2014).

12. Wells, J. M. *et al.* Homeostasis of the gut barrier and potential biomarkers. *Am. J. Physiol.-Gastrointest. Liver Physiol.* **312,** G171–G193 (2016).

13. König, J. *et al.* Human Intestinal Barrier Function in Health and Disease. *Clin. Transl. Gastroenterol.* **7,** e196 (2016).

14. Nusrat, A. *et al.* Clostridium difficile Toxins Disrupt Epithelial Barrier Function by Altering Membrane Microdomain Localization of Tight Junction Proteins. *Infect. Immun.* **69,** 1329–1336 (2001).

15. Janvilisri, T., Scaria, J. & Chang, Y.-F. Transcriptional profiling of Clostridium difficile and Caco-2 cells during infection. *J. Infect. Dis.* **202,** 282–290 (2010).

16. Meyer-Hoffert, U. *et al.* Secreted enteric antimicrobial activity localises to the mucus surface layer. *Gut* **57,** 764–771 (2008).

17. van Ampting, M. T. J. *et al.* Intestinally Secreted C-Type Lectin Reg3b Attenuates Salmonellosis but Not Listeriosis in Mice. *Infect. Immun.* **80,** 1115–1120 (2012).

18. Salzman, N. H., Ghosh, D., Huttner, K. M., Paterson, Y. & Bevins, C. L. Protection against enteric salmonellosis in transgenic mice expressing a human intestinal defensin. *Nature* **422,** 522–526 (2003).

19. Vaishnava, S. *et al.* The antibacterial lectin RegIIIγ promotes the spatial segregation of microbiota and host in the intestine. *Science* **334,** 255–258 (2011).

20. Wells, J. M., Rossi, O., Meijerink, M. & Baarlen, P. van. Epithelial crosstalk at the microbiota–mucosal interface. *Proc. Natl. Acad. Sci.* **108,** 4607–4614 (2011).

21. Artursson, P., Palm, K. & Luthman, K. Caco-2 monolayers in experimental and theoretical predictions of drug transport. *Adv. Drug Deliv. Rev.* **64, Supplement,** 280–289 (2012).

22. Sambuy, Y. *et al.* The Caco-2 cell line as a model of the intestinal barrier: influence of cell and culture-related factors on Caco-2 cell functional characteristics. *Cell Biol. Toxicol.* **21,** 1–26 (2005).

23. Engle, M. J., Goetz, G. S. & Alpers, D. H. Caco-2 cells express a combination of colonocyte and enterocyte phenotypes. *J. Cell. Physiol.* **174,** 362–369 (1998).

24. Artursson, P. Epithelial transport of drugs in cell culture. I: A model for studying the passive diffusion of drugs over intestinal absorbtive (Caco-2) cells. *J. Pharm. Sci.* **79,** 476–482 (1990).

25. Wilson, G. *et al.* Transport and permeability properties of human Caco-2 cells: An in vitro model of the intestinal epithelial cell barrier. *J. Controlled Release* **11,** 25–40 (1990).

26. Meunier, V., Bourrié, M., Berger, Y. & Fabre, G. The human intestinal epithelial cell line Caco-2; pharmacological and pharmacokinetic applications. *Cell Biol. Toxicol.* **11,** 187–194 (1995).

27. Delie, F. & Rubas, W. A Human Colonic Cell Line Sharing Similarities With Enterocytes as a Model to Examine Oral Absorption: Advantages and Limitations of the Caco-2 Model. *Crit. Rev. Ther. Drug Carr. Syst.* **14,** (1997).

28. Gan, L.-S. L. & Thakker, D. R. Applications of the Caco-2 model in the design and development of orally active drugs: elucidation of biochemical and physical barriers posed by the intestinal epithelium. *Adv. Drug Deliv. Rev.* **23,** 77–98 (1997).

29. Perdikis, D. A. & Basson, M. D. Basal nutrition promotes human intestinal epithelial (Caco-2) proliferation, brush border enzyme activity, and motility. *Crit. Care Med.* **25,** 159–165 (1997).

30. Tallkvist, J., Bowlus, C. L. & Lönnerdal, B. Functional and molecular responses of human intestinal Caco-2 cells to iron treatment. *Am. J. Clin. Nutr.* **72,** 770–775 (2000).

31. Ling, X., Linglong, P., Weixia, D. & Hong, W. Protective Effects of Bifidobacterium on Intestinal Barrier Function in LPS-Induced Enterocyte Barrier Injury of Caco-2 Monolayers and in a Rat NEC Model. *PLOS ONE* **11,** e0161635 (2016).

32. Venkatasubramanian, P. B. *et al.* Use of Microarray Datasets to generate Caco-2-dedicated Networks and to identify Reporter Genes of Specific Pathway Activity. *Sci. Rep.* **7,** 6778 (2017).

33. Sender, R., Fuchs, S. & Milo, R. Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. *Cell* **164,** 337–340 (2016).

34. Quigley, E. M. M. Gut Bacteria in Health and Disease. *Gastroenterol. Hepatol.* **9,** 560–569 (2013).

35. Rodríguez, J. M. *et al.* The composition of the gut microbiota throughout life, with an emphasis on early life. *Microb. Ecol. Health Dis.* **26,** (2015).

36. Laparra, J. M. & Sanz, Y. Interactions of gut microbiota with functional food components and nutraceuticals. *Pharmacol. Res.* **61,** 219–225 (2010).

37. Groschwitz, K. R. & Hogan, S. P. Intestinal Barrier Function: Molecular Regulation and Disease Pathogenesis. *J. Allergy Clin. Immunol.* **124,** 3–22 (2009).

38. Snoeck, V., Goddeeris, B. & Cox, E. The role of enterocytes in the intestinal barrier function and antigen uptake. *Microbes Infect.* **7,** 997–1004 (2005).

39. Tomasello, E. & Bedoui, S. Intestinal innate immune cells in gut homeostasis and immunosurveillance. *Immunol. Cell Biol.* **91,** 201–203 (2013).

40. Hensgens, M. P. M. *et al.* Clostridium difficile infection in the community: a zoonotic disease? *Clin. Microbiol. Infect.* **18,** 635–645 (2012).

41. Sammons, J. S., Toltzis, P. & Zaoutis, T. E. Clostridium difficile Infection in Children. *JAMA Pediatr.* **167,** 567–573 (2013).

42. Natarajan, M., Walk, S. T., Young, V. B. & Aronoff, D. M. A Clinical and Epidemiological Review of Non-toxigenic Clostridium difficile. *Anaerobe* **22,** 1–5 (2013).

43. Langdon, A., Crook, N. & Dantas, G. The effects of antibiotics on the microbiome throughout development and alternative approaches for therapeutic modulation. *Genome Med.* **8,** (2016).

44. Ferreyra, J. A. *et al.* Gut microbiota-produced succinate promotes C. difficile infection after antibiotic treatment or motility disturbance. *Cell Host Microbe* **16,** 770–777 (2014).

45. Nelson, R. L., Suda, K. J. & Evans, C. T. Antibiotic treatment for Clostridium difficile-associated diarrhoea in adults. *Cochrane Database Syst. Rev.* **3,** CD004610 (2017).

46. Spigaglia, P. Recent advances in the understanding of antibiotic resistance in Clostridium difficile infection. *Ther. Adv. Infect. Dis.* **3,** 23–42 (2016).

47. Voth, D. E. & Ballard, J. D. Clostridium difficile Toxins: Mechanism of Action and Role in Disease. *Clin. Microbiol. Rev.* **18,** 247–263 (2005).

48. Ryan, K. J., Ray, C. G., Sherris, J. C., Systems (Firm), T. D. & service), S. *Sherris medical microbiology : an introduction to infectious diseases*. (New York : McGraw-Hill, 2004).

49. Mitchell, T. J. *et al.* Effect of toxin A and B of Clostridium difficile on rabbit ileum and colon. *Gut* **27,** 78–85 (1986).

50. Goulding, D. *et al.* Distinctive Profiles of Infection and Pathology in Hamsters Infected with Clostridium difficile Strains 630 and B1. *Infect. Immun.* **77,** 5478–5485 (2009).

51. Buckley, A. M. *et al.* Susceptibility of Hamsters to Clostridium difficile Isolates of Differing Toxinotype. *PLOS ONE* **8,** e64121 (2013).

52. D'Auria, K. M. *et al.* Systems analysis of the transcriptional response of human ileocecal epithelial cells to Clostridium difficile toxins and effects on cell cycle control. *BMC Syst. Biol.* **6,** 2 (2012).

53. D'Auria, K. M. *et al.* In Vivo Physiological and Transcriptional Profiling Reveals Host Responses to Clostridium difficile Toxin A and Toxin B. *Infect. Immun.* **81,** 3814–3824 (2013).

54. Fabian, M. R., Sonenberg, N. & Filipowicz, W. Regulation of mRNA Translation and Stability by microRNAs. *Annu. Rev. Biochem.* **79,** 351–379 (2010).

55. Filipowicz, W., Bhattacharyya, S. N. & Sonenberg, N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat. Rev. Genet.* **9,** 102–114 (2008).

56. Kloosterman, W. P. & Plasterk, R. H. A. The Diverse Functions of MicroRNAs in Animal Development and Disease. *Dev. Cell* **11,** 441–450 (2006).

57. Bartel, D. P. MicroRNA Target Recognition and Regulatory Functions. *Cell* **136,** 215–233 (2009).

58. Ellwanger, D. C., Büttner, F. A., Mewes, H.-W. & Stümpflen, V. The sufficient minimal set of miRNA seed types. *Bioinformatics* **27,** 1346–1350 (2011).

59. Rana, T. M. Illuminating the silence: understanding the structure and function of small RNAs. *Nat. Rev. Mol. Cell Biol.* **8,** 23 (2007).

60. Bartel, D. P. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell* **116,** 281–297 (2004).

61. Ardekani, A. M. & Naeini, M. M. The Role of MicroRNAs in Human Diseases. *Avicenna J. Med. Biotechnol.* **2,** 161–179 (2010).

62. Peng, Y. & Croce, C. M. The role of MicroRNAs in human cancer. *Signal Transduct. Target. Ther.* **1,** 15004 (2016).

63. Wang, J., Liew, O. W., Richards, A. M. & Chen, Y.-T. Overview of MicroRNAs in Cardiac Hypertrophy, Fibrosis, and Apoptosis. *Int. J. Mol. Sci.* **17,** (2016).

64. Li, Y. & Kowdley, K. V. MicroRNAs in Common Human Diseases. *Genomics Proteomics Bioinformatics* **10,** 246–253 (2012).

65. Das, K., Garnica, O. & Dhandayuthapani, S. Modulation of Host miRNAs by Intracellular Bacterial Pathogens. *Front. Cell. Infect. Microbiol.* **6,** (2016).

66. Liu, S. *et al.* The Host Shapes the Gut Microbiota via Fecal microRNA. *Cell Host Microbe* **19,** 32–43 (2016).

67. Rupaimoole, R. & Slack, F. J. MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nat. Rev. Drug Discov.* **16,** 203 (2017).

68. Willett, W. C. *et al. Prevention of Chronic Disease by Means of Diet and Lifestyle Changes*. (The International Bank for Reconstruction and Development / The World Bank, 2006).

69. Hasler, C. M. Functional Foods: Benefits, Concerns and Challenges—A Position Paper from the American Council on Science and Health. *J. Nutr.* **132,** 3772–3781 (2002).

70. Lima, G. P. P., Vianello, F., Corrêa, C. R., Campos, R. A. da S. & Borguini, M. G. Polyphenols in Fruits and Vegetables and Its Effect on Human Health. *Food Nutr. Sci.* **5,** 1065 (2014).

71. Padayachee, A., Day, L., Howell, K. & Gidley, M. J. Complexity and health functionality of plant cell wall fibers from fruits and vegetables. *Crit. Rev. Food Sci. Nutr.* **57,** 59–81 (2017).

72. Manach, C. *et al.* Addressing the inter-individual variation in response to consumption of plant food bioactives: Towards a better understanding of their role in healthy aging and cardiometabolic risk reduction. *Mol. Nutr. Food Res.* **61,** n/a-n/a (2017).

73. Noble, D. Claude Bernard, the first systems biologist, and the future of physiology. *Exp. Physiol.* **93,** 16–26 (2008).

74. Hodgkin, A. L. & Huxley, A. F. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* **117,** 500–544 (1952).

75. Noble, D. Cardiac action and pacemaker potentials based on the Hodgkin-Huxley equations. *Nature* **188,** 495–497 (1960).

76. Wolkenhauer, O. Systems biology: The reincarnation of systems theory applied in biology? *Brief. Bioinform.* **2,** 258–270 (2001).

77. Frese, K. S., Katus, H. A. & Meder, B. Next-Generation Sequencing: From Understanding Biology to Personalized Medicine. *Biology* **2,** 378–398 (2013).

78. Salvioli, A. & Bonfante, P. Systems biology and 'omics' tools: A cooperation for next-generation mycorrhizal studies. *Plant Sci.* **203–204,** 107–114 (2013).

79. Werner, H. M. J., Mills, G. B. & Ram, P. T. Cancer Systems Biology: a peek into the future of patient care? *Nat. Rev. Clin. Oncol.* **11,** 167–176 (2014).

80. Beste, D. J. V. & McFadden, J. Metabolism of Mycobacterium tuberculosis. in *Systems Biology of Tuberculosis* 55–78 (Springer, New York, NY, 2013).

81. Rocco, A., Kierzek, A. & McFadden, J. Stochastic Gene Expression in Bacterial Pathogens: A Mechanism for Persistence? in *Systems Biology of Tuberculosis* 157–177 (Springer, New York, NY, 2013).

82. You, S. *et al.* A Systems Approach to Rheumatoid Arthritis. *PLOS ONE* **7,** e51508 (2012).

83. Bundalovic-Torma, C. & Parkinson, J. Comparative Genomics and Evolutionary Modularity of Prokaryotes. in *Prokaryotic Systems Biology* 77–96 (Springer, Cham, 2015).

84. Chan, C. S. & Turner, R. J. Biogenesis of Escherichia coli DMSO Reductase: A Network of Participants for Protein Folding and Complex Enzyme Maturation. in *Prokaryotic Systems Biology* 215–234 (Springer, Cham, 2015).

85. Deineko, V., Kumar, A., Vlasblom, J. & Babu, M. Quantitative and Systems-Based Approaches for Deciphering Bacterial Membrane Interactome and Gene Function. in *Prokaryotic Systems Biology* 135–154 (Springer, Cham, 2015).

86. Bogle, I. D. L. *et al.* Systems Biology of the Liver. in *Reviews in Cell Biology and Molecular Medicine* (Wiley-VCH Verlag GmbH & Co. KGaA, 2006).

87. Mcculloch, A. D. & Paternostro, G. Cardiac Systems Biology. *Ann. N. Y. Acad. Sci.* **1047,** 283–295 (2005).

88. Martins dos Santos, V., Müller, M. & de Vos, W. M. Systems biology of the gut: the interplay of food, microbiota and host at the mucosal interface. *Curr. Opin. Biotechnol.* **21,** 539–550 (2010).

89. Lubeck, E. & Cai, L. Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nat. Methods* **9,** 743–748 (2012).

90. Capozzi, F. & Bordoni, A. Foodomics: a new comprehensive approach to food and nutrition. *Genes Nutr.* **8,** 1–4 (2013).

91. Venkatesh, T. & Harlow, H. B. Integromics: challenges in data integration. *Genome Biol.* **3,** reports4027.1-reports4027.3 (2002).

92. Paszkiewicz, K. & Studholme, D. J. De novo assembly of short sequence reads. *Brief. Bioinform.* **11,** 457–472 (2010).

93. Wang, S. *et al.* Target analysis by integration of transcriptome and ChIP-seq data with BETA. *Nat. Protoc.* **8,** 2502–2515 (2013).

94. Garibyan, L. & Avashia, N. Research Techniques Made Simple: Polymerase Chain Reaction (PCR). *J. Invest. Dermatol.* **133,** e6 (2013).

95. Lodish, H. *et al.* DNA Microarrays: Analyzing Genome-Wide Expression. (2000).

96. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).

97. Ali, W., Deane, C. & Reinert, G. Protein Interaction Networks and Their Statistical Analysis. in *Handbook of Statistical Systems Biology* (eds. Stumpf, M. P. H., Balding, D. J. & Girolami, rk) 200–234 (John Wiley & Sons, Ltd, 2011).

98. Barenco, M., Brewer, D., Callard, R. & Hubank, M. Modelling Transcription Factor Activity. in *Handbook of Statistical Systems Biology* (eds. Stumpf, M. P. H., Balding, D. J. & Girolami, rk) 440–450 (John Wiley & Sons, Ltd, 2011).

99. Coolen, A. C. C., Fraternali, F., Annibale, A., Fernandes, L. & Kleinjung, J. Modelling Biological Networks via Tailored Random Graphs. in *Handbook of Statistical Systems Biology* (eds. Stumpf, M. P. H., Balding, D. J. & Girolami, rk) 309–329 (John Wiley & Sons, Ltd, 2011).

100. Ebbels, T. M. D. & De Iorio, M. Statistical Data Analysis in Metabolomics. in *Handbook of Statistical Systems Biology* (eds. Stumpf, M. P. H., Balding, D. J. & Girolami, rk) 163–180 (John Wiley & Sons, Ltd, 2011).

101. Edwards, L. M. & Thiele, I. Applying systems biology methods to the study of human physiology in extreme environments. *Extreme Physiol. Med.* **2,** 8 (2013).

102. Shahzad, K. & Loor, J. J. Application of Top-Down and Bottom-up Systems Approaches in Ruminant Physiology and Metabolism. *Curr. Genomics* **13,** 379–394 (2012).

103. Linderman, G. C., Chance, M. R. & Bebek, G. MAGNET: MicroArray Gene expression and Network Evaluation Toolkit. *Nucleic Acids Res.* gks526 (2012). doi:10.1093/nar/gks526

104. Reiss, D. J., Baliga, N. S. & Bonneau, R. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics* **7,** 280 (2006).

105. Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Philos. Transact. A Math. Phys. Eng. Sci.* **374,** (2016).

106. Margolin, A. A. *et al.* ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics* **7,** S7 (2006).

107. Zhang, B. & Horvath, S. A General Framework for Weighted Gene Co-Expression Network Analysis. *Stat. Appl. Genet. Mol. Biol.* **4,** (2005).

108. Faith, J. J. *et al.* Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* **5,** e8 (2007).

109. Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nat. Methods* **9,** 796–804 (2012).

110. Touw, W. G. *et al.* Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Brief. Bioinform.* **14,** 315–326 (2013).

111. Tarca, A. L., Carey, V. J., Chen, X., Romero, R. & Drăghici, S. Machine Learning and Its Applications to Biology. *PLOS Comput. Biol.* **3,** e116 (2007).

112. Resat, H., Petzold, L. & Pettigrew, M. F. Kinetic Modeling of Biological Systems. *Methods Mol. Biol. Clifton NJ* **541,** 311–335 (2009).

113. Montefusco, F., Cosentino, C. & Bates, D. G. Control Engineering Approaches to Reverse Engineering Biomolecular Networks. in *Handbook of Statistical Systems Biology* (eds. Stumpf, M. P. H., Balding, D. J. & Girolami, rk) 83–113 (John Wiley & Sons, Ltd, 2011).

114. Yıldırım, M. A. & Vidal, M. Systems engineering to systems biology. *Mol. Syst. Biol.* **4,** 185 (2008).

115. Gibson, D. G. *et al.* Complete Chemical Synthesis, Assembly, and Cloning of a Mycoplasma genitalium Genome. *Science* **319,** 1215–1220 (2008).

116. Khalil, A. S. *et al.* A Synthetic Biology Framework for Programming Eukaryotic Transcription Functions. *Cell* **150,** 647–658 (2012).

117. Haggart, C. R., Bartell, J. A., Saucerman, J. J. & Papin, J. A. Whole-genome metabolic network reconstruction and constraint-based modeling. *Methods Enzymol.* **500,** 411–433 (2011).

118. Chen, C. *et al.* Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. *PLOS ONE* **6,** e17238 (2011).

119. Giordan, M. A Two-Stage Procedure for the Removal of Batch Effects in Microarray Studies. *Stat. Biosci.* **6,** 73–84 (2014).

120. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8,** 118–127 (2007).

121. Luo, J. *et al.* A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *Pharmacogenomics J.* **10,** 278–291 (2010).

122. Xia, X.-Q. *et al.* WebArrayDB: cross-platform microarray data analysis and public data repository. *Bioinformatics* **25,** 2425–2429 (2009).

Use of Microarray Datasets to generate Caco-2-dedicated Networks and to identify Reporter Genes of Specific Pathway Activity

Prashanna Balaji Venkatasubramanian[1], Gamze Toydemir[1,2], Nicole de Wit[1], Edoardo Saccenti[3], Vitor A. P. Martins Dos Santos[3,4], Peter van Baarlen[5], Jerry M. Wells[5], Maria Suarez-Diez[3] and *Jurriaan J. Mes[1].

Published in Scientific Reports – 2017.

Abstract

Intestinal epithelial cells, like Caco-2, are commonly used to study the interaction between food, other luminal factors and the host, often supported by microarray analysis to study the changes in gene expression as a result of the exposure. However, no compiled dataset for Caco-2 have ever been initiated and Caco-2-dedicated gene expression networks are barely available. Here, 341 Caco-2-specific microarray samples were collected from public databases and from in-house experiments pertaining to Caco-2 cells exposed to pathogens, probiotics and several food compounds. Using these datasets, a gene functional association network specific for Caco-2 was generated containing 8937 nodes 129711 edges. Two *in silico* methods, a modified version of bi-clustering and the new Differential Expression Correlation Analysis, were developed to identify Caco-2-specific gene targets within a pathway of interest. These methods were subsequently applied to the AhR and Nrf2 signalling pathways and altered expression of the predicted target genes was validated by qPCR in Caco-2 cells exposed to coffee extracts, known to activate both AhR and Nrf2 pathways. The datasets and *in silico* method(s) to identify and predict responsive target genes can be used to more efficiently design experiments to study Caco-2/intestinal epithelial-relevant biological processes.

Introduction

Biological networks are representational interactions between genes, proteins, and other biomolecules. Different kinds of biological networks (*e.g.* protein-protein interaction or signalling networks) represent different features of a cell [1]. Such networks can be usefully exploited to gain key insights into biological systems [2,3]. Exploration of tissue and cell type specific networks has demonstrated the effects of tissue specific regulation on the remodelling of biological networks [4]. Differential network analysis has also been used to compare topological characteristics of networks corresponding to normal or tumorous cells and to isolate characteristics of distinct cancer subtypes, which in turn has led to the prediction of cancer subtype-specific drug targets [5]. One important biological system is the epithelial cells lining the small and large intestine. The role of diet and the response of host towards diet and its compounds is challenging to be studied *in vivo* due to the complexity of biological systems and inter-individual variability. Thus, a reductionist approach using the human Caco-2 intestinal epithelial cell line is a widely accepted laboratory model to understand the response of enterocytes exposed to nutrition and microbes [6–8]. Although Caco-2 cells were derived from a colon carcinoma, when cultured as confluent monolayers for 2-3 weeks, they functionally resemble the enterocytes lining the small intestine [9]. Caco-2 cells have been used in numerous experiments to study effects of food products and compounds [6,7,10–13], probiotics [8,14], pathogens [15–17] and other studies [18–20], using microarrays. Comparative proteomic analysis of Caco-2 cells and scrapings of the human intestinal epithelium support the usability of this *in* vitro model [21], although Caco-2 cells appear to over-express as well as under-express certain proteins which needs to be considered in the interpretation of *in vitro* data and translation of results to the *in vivo* situation [21].

A compendium of Caco-2 gene expression profiles under a broad number of conditions can be instrumental in building dedicated network models describing gene interactions in human enterocytes and in providing new insights on their functioning. Although, gene profiles tuned for

selected tissues [22–24] are present, to the best of our knowledge, no broad compendium of Caco-2 microarray experiments have been initiated, limited data on metabolic networks is available [25,26] and no gene/protein association networks are available for Caco-2/ enterocytes. Another commonly faced problem is the identification of Genes Of Interest (GOI) in the pathways investigated for a specific cell type. Thus identification of candidate sets of GOI could help study the impact of treatments on specific pathways of interest in a given cell type.

Intestinal epithelial cells, apart from major functions like digestion and absorption of nutrients, minerals and water [27,28], play an important role in the exclusion or detoxification of xenobiotics and regulating oxidative stresses. The AhR and Nrf2 pathways are involved in the metabolism of xenobiotics and protection against oxidative stress [29,30]. AhR is an important regulator of Phase I and Phase II enzymes and other enzymes which metabolize compounds such as dioxins, polycyclic aromatic hydrocarbons, plant polyphenols and tryptophan photoproducts [31]. Nrf2 has been designated the “master regulator” of the adaptive response to oxidative stress [29] and regulates the expression of antioxidant proteins that protect against oxidative damage triggered by injury and inflammation.

In this study, we aim to i) exploit the knowledge accumulated in the publicly available datasets on Caco-2 cells exposed to different treatments in order to generate a dedicated network model accounting for gene associations specific to enterocytes and ii) to develop workflows to reliably select genes for studying intestinal enterocyte-specific pathways. The proposed strategies were experimentally validated by focussing on GOI in the Nrf2 and AhR pathways using Caco-2 cells exposed to coffee to induce the gene responses within these pathways. The obtained networks are provided as supplementary files and R scripts for the identification of GOI are also made available as supplementary files with a working example.

Materials and Methods

Data Processing

Caco-2 microarray gene expression data were obtained from public repository, Array Express (www.ebi.ac.uk/arrayexpress) and from in-house experiments performed using Affymetrix© 1.1 ST array platform. In-house data was obtained by exposure of Caco-2 cells grown on transwells with different preparations of food-related compounds in experiments conducted over several years. Publicly available data was restricted to experiments on Affymetrix platform. Data and associated metadata were manually curated using the following inclusion criteria: i) experiments that did not induce genetic mutations, ii) experiments performed on Caco-2 cell monolayers that were grown for at least seven days and iii) arrays probing for at least 17000 genes (annotated in Chip Definition Files), thereby leaving out old arrays. Based on these criteria 341 arrays were selected corresponding to 22 experimental batches encompassing 85 different treatments (Table 1). GSE accession numbers of publicly available datasets and other relevant descriptions are given in Supplementary file (Supplementary Table S1).

Table 1. *Summary of collected dataset*

Total Arrays	341
Total Experiments	88
From the lab of Jurriaan Mes	173
From Array Express	168
Type of Exposure	
Vegetables	9
Fruits	20
Fibres	22
Probiotics	7
Pathogens	11
Others	6
Food compounds	10

The consolidated data of 341 arrays were normalized using the SCAN algorithm before network construction and biclustering analysis, as this is a method that performs well for cross comparison [52]. RMA normalization was used for differential expression (DE) analysis, as is considered as standard for this calculation [53]. All the normalization procedures were performed using R Bioconductor packages *SCAN.UPC* [54] and *affy* [55]. Microarray probes were matched to gene identifiers using the CDF array annotation (version 18) provided by the University of Michigan microarray© lab [56]. After both normalization procedures, a combined set of 21996 genes was obtained. All statistical programming were performed using statistical language R (version 3.2.3).

Identification of genes expressed in Caco-2 cells

Universal exPression Code was used to obtain a standardized score describing the active/inactive state of each gene in each array of our data compendium [54]. Genes with a UPC value greater than 0.5 in at least one array were considered to be expressed in Caco-2 cells and therefore used in the analysis. This step was applied to the matrix of 21996 genes and 341 arrays reducing it to a matrix of 12849 genes and 341 arrays. In this matrix there were some genes with some values missing, likely due to platform differences. Therefore, genes with missing values in more than half the total number of arrays (*ie.* 170 arrays) were discarded. Remaining missing values were imputed using KNN algorithm from the 'impute' R package in [57,58] with default parameters. The final data matrix contained values for 10831 genes over 341 arrays.

Caco-2 cell specific network generation

The database STRING (version 10) [59] was used for the retrieval of high confidence human specific protein association and a combined score cut-off value of 700 was used as recommended by STRING. Nodes representing genes identified as not being expressed by

Caco-2 cells were removed from the network. The network (in edgelist format) is available as supplementary file (Caco2_Network) and at http://semantics.systemsbiology.nl/. Edgelist contains pairs of interacting genes (first two columns) and in this file genes are denoted by their Entrez Ids. The third column refers to the weight of each edge, which is however empty in the given file, as the edges have no weights. The networkx (python package) was used for network topological analysis [60].

Biclustering Algorithm

The Biclustering algorithm of cMonkey [45] adapted by van Dam et al. [47] was used to find biclusters (*i.e.* groups of co-expressed genes in a subset of conditions [61,62]). In our implementation a pre-defined set of genes, called seed genes, together with additional genes from a second list called gene pool were used to find biclusters. Seed genes were selected using the following two approaches: i) from literature on Caco-2 expression in response to different types of coffee (SQSTM1, HMOX1, NRF2 and ABCC1 for the Nrf2 pathway and CYP1A1, TIPARP and AHR for AhR pathway). ii) from Weighted Gene Correlation Network Analysis [63] (WGCNA). The WGCNA method partitions genes expressed in Caco-2 cell lines into groups enriched for topological overlap based on their expression profiles. These groups are then assessed for enrichment in genes belonging to the selected pathways using Ingenuity Pathways Analysis (IPA) (http://www.ingenuity.com, release March 2014). Genes assigned to the selected pathways in the enriched modules (FDR < 0.05) were further included in the seed gene list (DNAJB1 and ENC1 for Nrf2 pathway and ARNT and PRKCA for AhR pathway). To build the gene pool, genes expected to be in the pathway of interest were retrieved from pathway database IPA (Ahr and Nrf2 consensus pathway).The gene pool list contained 87 genes for Nrf2 pathway and 48 genes for AhR pathway.

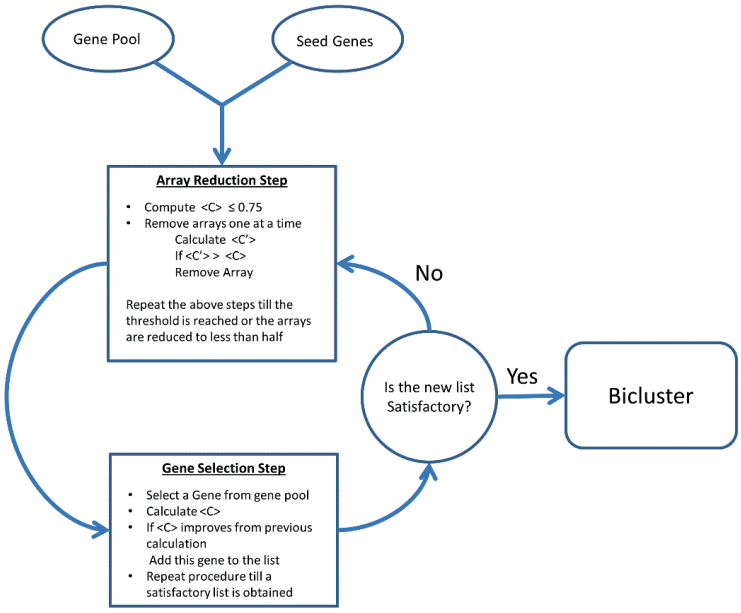


Figure 1. *Flow diagram describing Biclustering algorithm. Seed genes are a predefined group of genes. The gene pool is the set of genes to be tested for inclusion in the bicluster. <C> indicates mean pairwise correlation, <C'> indicates new mean pairwise calculation*

Biclustering was performed using R implementing the iterative procedure depicted in Figure 1. In the first step, the data compendium is explored to select arrays for which the seed genes show a high degree of mean pairwise correlation between each other. This selection is performed by iteratively removing one array from the list and comparing the average pairwise correlation between seed genes computed considering the full array list and the array list without the selected one. If removal of the considered array leads to an increase of this correlation, the array is permanently removed from the array list. This process is iterated until either the average correlation between seed genes is greater than or equal to a threshold value, $C_T = 0.75$ or half of the initial arrays have been removed.

Once the reduced array set has been established, an additional iterative procedure to search for candidate genes is performed. In the initialisation step, a new list of genes is built containing the seed genes. Then a new gene is selected from the gene pool and the mean correlation between this new gene along with the genes in the current list is calculated. If such correlation value is greater than previous correlation value, the new gene is added. This procedure is iterated till no new genes remain. The full procedure of array reduction and gene addition is continued until a bicluster with the desired properties is obtained.

Differential Expression Correlation Analysis (DECA)

We implemented a new algorithm, Differential Expression Correlation Analysis (DECA) to find GOI using differential expression (DE) values from microarray datasets. The DECA algorithm works by calculating correlation values between seed genes and other DE genes identified using the UPC algorithm. DE values were calculated for 85 experimental setups (3 of which could not be used as they lacked sufficient replicates or controls) giving a total of 21996 genes. For each of these genes the treatments were compared to their respective controls using Bioconductor package *limma* [64]. Following this, UPC filtering was applied and the DE matrix (a matrix containing the DE values with genes along the rows and experimental comparisons along the columns) was reduced to 12849 genes. Genes that were missing expression values for more than 56 conditions (roughly two third conditions) were excluded and then remaining missing data were imputed using KNN impute as mentioned above. This resulted in a matrix of DE values for 12462 genes and 85 conditions. All corresponding missing P-values were substituted with 1.

The next step in DECA is the selection of seed genes from literature. Seed genes were chosen in such a way that they showed strong and significant (absolute fold change ≥ 2 and p-value < 0.01) DE in stimulations associated to the chosen pathway (SQSTM1, NQO1 and HMOX1 for the Nrf2 pathway and CYP1A1 and TIPARP for AhR pathway).

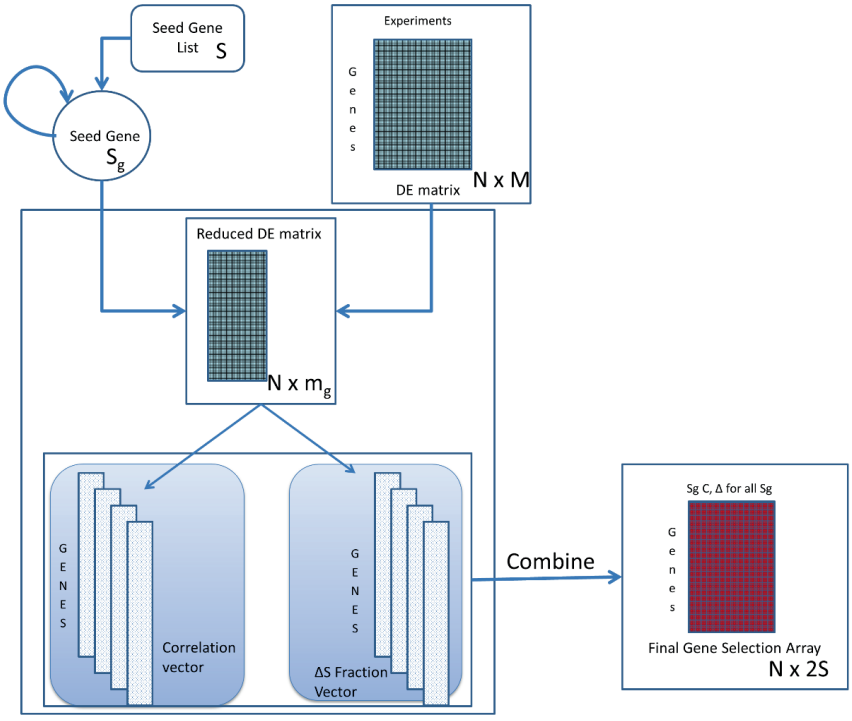


Figure 2. *Flow diagram describing DECA (Differential Expression Correlation Analysis). Seed gene list refers to the starting gene selection. DE matrix is the input data matrix. The algorithm outputs a ranked list of genes which are highly correlated with the input genes.*

The workflow of the procedure is described in Figure 2 and implemented in R. Seed genes were then randomly considered one at a time. The DE matrix is reduced by the algorithm to contain only the comparisons in which the seed gene under consideration is found to have significant DE. Correlation values are calculated between the seed gene and each gene in the gene pool using the reduced DE matrix. The fraction of reduced comparisons in which each gene has significant DE (p-value < 0.01) is recorded and is termed significance fraction. Finally, correlations and fractions for each seed gene, are combined in a matrix format and a selection criterion for absolute correlation values and significance fraction was set at 0.6. A list of genes that have either absolute correlation value or significance fraction above the threshold for any of the seed gene were selected. Subsequently, this new list of genes were ranked depending

on their individual absolute correlation values and significance fraction for each seed gene, thereby providing *2n* ranks (where *n* is the number of seed genes). A final rank was calculated by estimating the geometric mean of the 2n ranks for each gene.

All R scripts used in this paper are also available at http://semantics.systemsbiology.nl/index.php/download-page/.

DECA comprehensive *in silico* assessment

10 pathways were chosen at random for assessment of DECA algorithm. These pathways are ABC transporters pathway, Adherens junction pathway, Fat Absorption pathway, Gap junction pathway, Glycerolipid metabolism pathway Glycerophospholipid metabolism pathway, Nfk-β signalling pathway, p53 signalling pathway, PPAR signalling pathway and TLR signalling pathway. Some of these pathways are known to be associated with intestinal epithelia [65–67]. The genes associated to each of the 10 pathways were selected form KEGG pathway database [38]. For each of these pathways, 3 seed genes were chosen at random. The chosen seed genes were ensured for significant differential expression in at least 15 experiments. The seed genes were then used in DECA and the resulting gene list was ranked as mentioned above. The number of genes present in the top 10% of the ranked list belonging to the pathway were calculated. In addition to this, a Welch two sample t-test was performed to assess if the average ranks of the pathway related genes had a better rank compared against the average ranks of the rest of the genes in the ranked list. The protocol was iterated 10 times for each pathway. The results are provided as supplementary file (Supplementary Table S2).

Culturing & experimental exposure of Caco-2 cells

The Caco-2 cells were cultured for 7 days until they reach confluence in DMEM (Dulbecco's Modified Eagle Medium) (Control media) prior to exposure to coffee extracts (Turkish coffee [TC], Brasil Espirito [BE], Java Preanger [JP], Nescafe© [NC]) or TCDD. The RNA was

harvested and primers were developed for qPCR. The detailed description of the protocol is provided in supplementary file (Supplementary Text F1).

Results

Cell/Tissue-specific Gene Expression Profiles aid the identification of reporter genes for specific pathway activity

In this study, we develop strategies to generate dedicated gene network models for Caco-2 and identify specific gene responses to nutrition related exposures. This was illustrated using Ahr and Nrf2 pathways. We have independently validated our results through a new experimental setup on which Caco-2 cells were exposed to coffee extracts, which have previously been shown to induce the Ahr and Nrf2 pathways [32]. Coffee extracts have a great chemical diversity and the components vary according to the cultivar, treatment, processing, storage, *etc.* [33–36]. We have tested induction of these pathways using four coffee types.

To identify reporter genes for the AhR and Nrf2 pathways, scientific literature was searched and we investigated whether these genes were also responsive to oxidative stress in our Caco-2 model after exposure to TCDD (2,3,7,8-Tetrachlorodibenzo-*p*-dioxin) or coffee. 16 genes that are frequently used as indicators for AhR and Nrf2 signalling, were selected from the literature (Table 2) for validation. Caco-2 cells were exposed to coffee extracts (Turkish coffee, Brasil Espirito, Java Preanger, Nescafe©) and TCDD and relative expression of the selected genes was measured by qPCR. Out of the 16 genes tested, 3 genes were not detectable (CT values ≥ 35) and 5 genes showed no DE (a fold change threshold of 1.5 folds up or down in at least two of the coffee samples), indicating that 50% of the genes selected from literature are not useful for studying the activities of the AhR and Nrf2 pathways in enterocytes.

Table 2. *Expression changes upon coffee/xenobiotics exposure of initial set of genes selected based on existing literature. '*' indicates genes found to be significantly differentially expressed (Fold change > ±1.5) in Turkish Coffee only. Genes were considered to be responsive, if they were expressed in at least two coffee samples.*

Gene Name	**Pathway**	**Reference**	**Significant change in expression (Fold Change larger/smaller than ± 1.5)**
SQSTM1	Nrf2	[Jain et al, 2010] [68]	Yes
HMOX1	Nrf2	[Bøhn et al, 2014] [33]	Yes
Nrf2	Nrf2	[Bøhn et al, 2014] [33]	No
ABCC1	Nrf2	[Adachi et al, 2007] [69]	No
ABCC2	Nrf2	[Adachi et al, 2007] [69]	No
NQO1	Nrf2	[Bøhn et al, 2014] [33]	Yes
ABCG2	Nrf2	[Isshiki et al, 2011] [70]	No *
GSTP1	Nrf2	[Steinkellner et al, 2005] [71]	Yes
ARNT	AhR	[Ishikawa et al, 2014; Yeager et al, 2009] [32,72]	No
AhR	AhR	[Kalthoff et al, 2010] [73]	Yes
CYP1A1	AhR	[Ishikawa et al, 2014] [32]	Yes
TiPARP	AhR	[Diani-Moore et al, 2010] [74]	Yes
UGT1A6	AhR	[Yeager et al, 2009] [72]	Yes
CYP1A2	AhR	[Ishikawa et al, 2014] [32]	Not detected
CYP1B1	AhR	[Ishikawa et al, 2014] [32]	Not detected
AHRR	AhR	[Mimura et al, 2003; Abel et al, 2010] [30,31]	Not detected

Compendium of Caco-2 experimental data supports cell-specific genes selection

A data compendium was generated using Affymetrix expression profiles of 341 arrays from 85 Caco-2 exposure experiments (Table 1). UPC filtering procedure was used to identify genes that are actively expressed in Caco-2 and 12849 genes were identified to be expressed. These genes were then used to generate a cell-specific network dedicated to Caco-2 intestinal epithelial cells. Supplementary file (Supplementary Table S3) presents the comparison between network topological properties of the full interaction network retrieved from STRING (converted to Entrez Ids) and the Caco-2 specific network. The same cut-off (≥ 700) related to the reliability of the interactions (STRING combined score) was selected for both networks. The Caco-2 network is composed of 8937 nodes and 129711 edges and can be explored using common network visualization tools such as Cytoscape [37]. Notice the differences in the number of nodes and edges between the two networks. Out of the 16 genes that we previously selected based on literature, ABCC1, ABCG2 and TIPARP are removed from the network of functional associations. This indicates that in the overall network they are connected only to nodes that show no (active) expression in our compendium. However, even after this reduction, still large number of genes remain (77 nodes for Nrf2 pathway and 42 nodes for AhR pathway) to probe for each pathway and therefore we wanted to optimize our approach to identify GOI.

Biclustering analysis improves gene selection

The biclustering method works based on identification of genes that are co-expressed with seed genes (*i.e.* genes well known to be responsive in Caco-2 cells to a specific perturbation). In order to identify Caco-2-responsive genes within the Nrf2 pathway, we used a full list of genes that are involved in this pathway (derived from generic IPA consensus pathway). SQSTM1, HMOX1, NRF2, ABCC1, DNAJB1 and ENC1 were selected as seed genes. The seed genes were used to identify co-expressed genes within the compendium of microarrays. The initial average correlation threshold for array selection was set at 0.75 (default value). In this way, only

arrays that showed a high degree of correlation with the seed genes were included for GOI identification.

Table 3. *Expression changes upon coffee exposure of genes selected using the biclustering algorithm. '-' indicates genes that were not the target of experimental validation. Genes were considered to be responsive, if they were differentially expressed in at least two coffee samples.*

Gene Name	Pathway	Seed Genes	Found from	Significant change in expression (Fold Change more than ± 1.5)
DNAJB1	Nrf2	Yes	WGCNA	-
SQSTM1	Nrf2	Yes	Literature	Yes
HMOX1	Nrf2	Yes	Literature	Yes
ENC1	Nrf2	Yes	WGCNA	No
Nrf2	Nrf2	Yes	Literature	No
ABCC1	Nrf2	Yes	Literature	No
CDC34	Nrf2	No	Biclustering	-
DNAJC4	Nrf2	No	Biclustering	-
GTR	Nrf2	No	Biclustering	-
ATF4	Nrf2	No	Biclustering	Yes
GSTA2	Both	No	Biclustering	Yes
GSTM4	Both	No	Biclustering	Yes
MAPK8	AhR	No	Biclustering	-
MED1	AhR	No	Biclustering	-
NCOR2	AhR	No	Biclustering	-
NFIA	AhR	No	Biclustering	No
ARNT	AhR	Yes	WGCNA	No
AhR	AhR	Yes	Literature	Yes
CYP1A1	AhR	Yes	Literature	Yes
PRKCA	AhR	Yes	WGCNA	No
TiPARP	AhR	Yes	Literature	Yes

The biclustering analysis reduced the 341 arrays (the initial number of arrays) to 229 arrays and the following genes were obtained as GOI: CDC34, DNAJC4, GTR, ATF4, GSTA2 and GSTM4. Together with the seed genes this resulted in a total of 12 potential responsive genes for the Nrf2 pathway (Table 3). These genes had an average correlation of 0.79 in the arrays included in this analysis.

Similarly, CYP1A1, TIPARP, AHR, ARNT and PRKCA were chosen as seed genes for AhR pathway. Owing to the small number of seed genes, mean correlation threshold for array selection was set at a more stringent value of 0.8. The biclustering analysis reduced the initial 341 arrays to 274 arrays and predicted GSTA2, GSTM4, MAPK8, MED1, NCOR2 and NFIA as GOI for the AhR pathway. This procedure reduced the number of potential responsive genes to 11 for AhR pathway (Table 3), including seed genes.

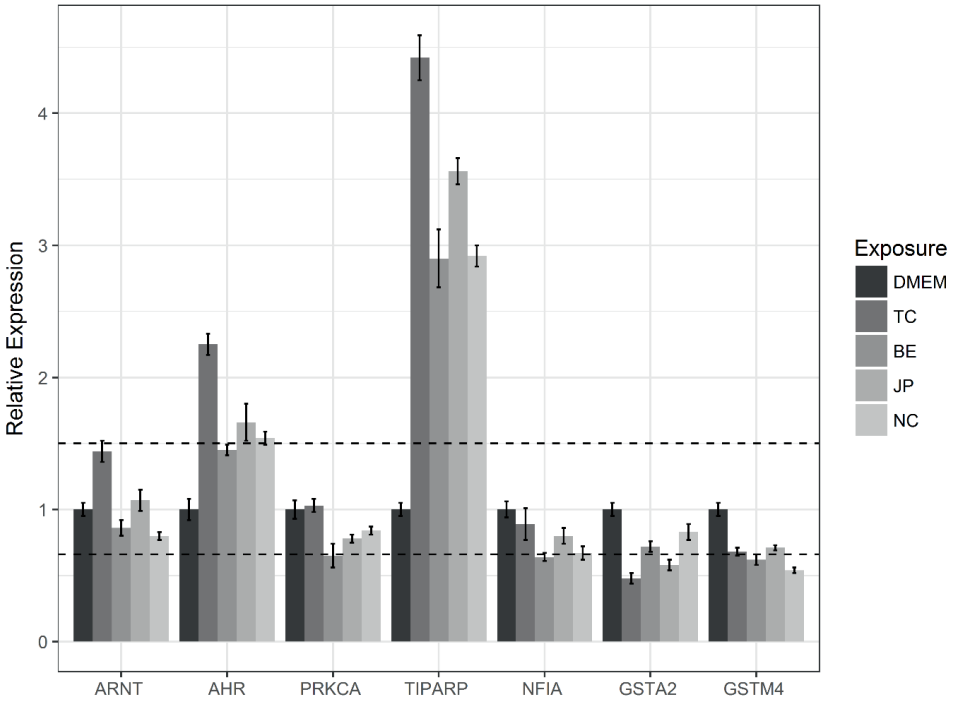


Figure 3. *qPCR results for AhR Pathway genes predicted using biclustering algorithm. The plot shows the relative gene expression level (control vs treatment) of several genes associated with AhR pathway. Results have been normalized to control (DMEM) values. Values and error bars represent average and standard deviation of three replicates. Dashed lines represent the fold change cut-off limits (1.5 for up regulation and 0.6 for down regulation). CYP1A1 is not shown here as it exceeds the plot limits. TC indicates Turkish coffee, BE indicates Brasil Espirito, JP indicates Java Preanger and NC indicates Nescafe©*

We selected 14 genes for experimental verification using Caco-2 cells exposed to coffee extracts (Figures 3 and 4). Of these, 6 genes were specific to AhR pathway, 6 specific to Nrf2 pathway and 2 common to both pathways. Four of these genes have been predicted by the algorithm ("Biclustering" see Table 3). All 4 genes were found to be expressed in Caco-2 cells of which 3 showed substantial changes in expression (Fold Change > 1.5) between control and treatment (Figures 3 and 4).

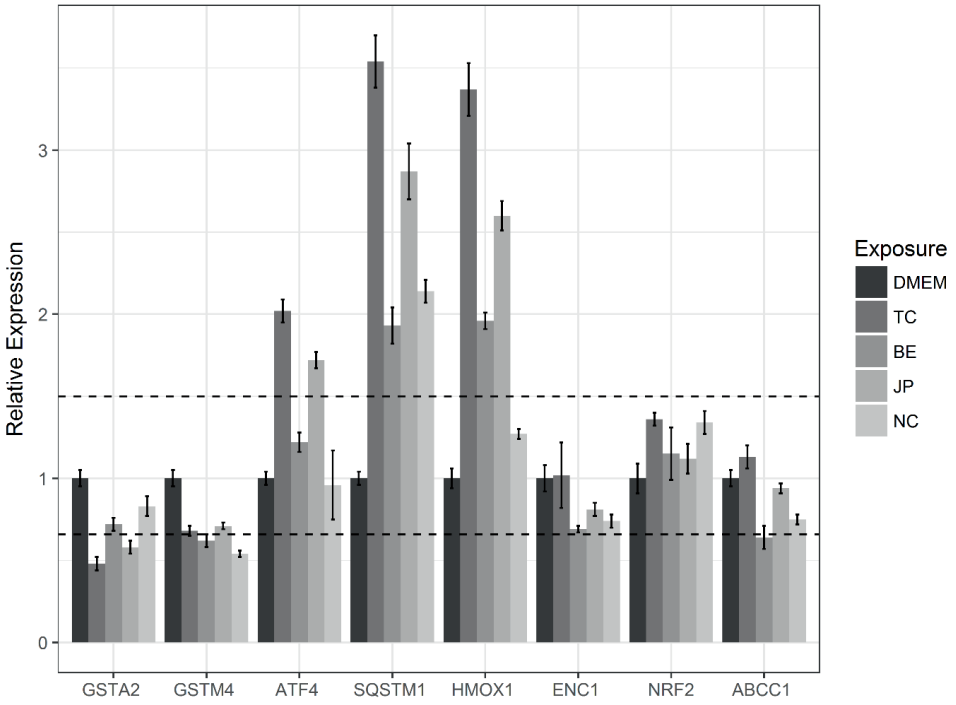


Figure 4. *qPCR results for Nrf2 Pathway genes predicted using biclustering algorithm. The plot shows the relative gene expression level (control vs treatment) of several genes associated with Nrf2 pathway. The line represents the fold change cut-off limits (1.5 for up regulation and 0.6 for down regulation). TC indicates Turkish coffee, BE indicates Brasil Espirito, JP indicates Java Preanger and NC indicates Nescafe©*

Based on these results, we concluded that this strategy constitutes a useful addition to the literature data for gene selection. Selected genes extracted from the literature can be combined with the ones selected using the proposed approach. In those cases where literature provides an ample list of genes for experimental validation, our approach serves to further refine the selection of genes which are differentially expressed by Caco-2 cells in a specific pathway.

Differential Expression Correlation Analysis further enhances gene selection

An assessment of DECA algorithm was performed using 10 pathways from the KEGG database [38] that are of interest to intestinal epithelia. For each pathway 10 runs were performed using three randomly selected genes from the pathways as seed genes. Genes known to be in the target pathways were found to be significantly better ranked than genes not in the pathway, as indicated by the enrichment p-values. On average ~9 % of genes related to each pathway could be predicted as target genes on analysing the top 10% ranked genes using DECA algorithm. The performance of the algorithm varied according to the pathway from 6% to 15%. This result indicates that without any further literature considerations DECA is able to retrieve genes associated to the pathway. In this assessment seed genes were chosen at random, however careful selection of seed genes is required to obtain more reliable prediction of target genes. As in the previous case, this approach would work best when combined with pre-existing knowledge. The results of the *in silico* assessment are provided in supplementary file (Supplementary Table S2). The DECA method was applied to find a global set of genes (amongst all genes expressed in Caco-2) associated with Nrf2 and AhR pathways which are

responsive to altered pathway activity. SQSTM1, NQO1 and HMOX1, involved in the Nrf2 pathway were used as seed genes for the DECA algorithm. 2834 genes were found to have correlation values or significance fractions above the 0.6 threshold against each seed gene. The genes were ranked as mentioned in Materials and Methods section and top ranked genes were considered for further analysis. From this list, GCLM [39], TXNRD1 [40], SOX9 and KCTD5 [41] were selected for further experimental validation via qPCR as there is some evidence of involvement in this pathway. In addition, BAG3 [42] gene which did not belong to the top ranking genes was randomly chosen as a negative control (Table 4).

Table 4. *Expression changes upon coffee exposure of genes identified using the DECA algorithm in AhR and Nrf2 pathways. '*' indicates genes found to be significantly differentially expressed (Fold change > ±1.5) in Turkish Coffee only. '^' indicates genes found to be significantly differentially expressed (Fold change > ±1.5) in Nescafe only. N/A indicates genes that were not the target of experimental validation. Genes were considered to be responsive, if they were expressed in at least two coffee samples.*

Gene Name	Pathway	Type	Significant change in expression (Fold Change more than ± 1.5)
CYP1A1	AhR	Seed Genes	Yes
TIPARP	AhR	Seed Genes	N/A
ATP9A	AhR	Predicted	No
UGCG	AhR	Predicted	Yes
CHMP1B	AhR	Predicted	Yes
EREG	AhR	Predicted	Yes
RND3	AhR	Predicted	Yes
SQSTM1	Nrf2	Seed Genes	Yes
HMOX1	Nrf2	Seed Genes	N/A
NQO1	Nrf2	Seed Genes	N/A
BAG3	Nrf2	Predicted	No *
SOX9	Nrf2	Predicted	Yes *^
TXNRD	Nrf2	Predicted	Yes
GCLM	Nrf2	Predicted	Yes
KCTD5	Nrf2	Predicted	Yes *^

A similar approach was used to predict the GOI in the AhR pathway. Only two genes, CYP1A1 and TIPARP were chosen as the seed genes for the DECA algorithm which resulted in a list of 398 ranked genes. From this list, UGCG [43], EREG [44], RND3, CHMP1B were chosen for experimental verification as evidence from scientific literature associated few of them with the AhR pathway. ATP9A was randomly selected as a negative control (Table 4).

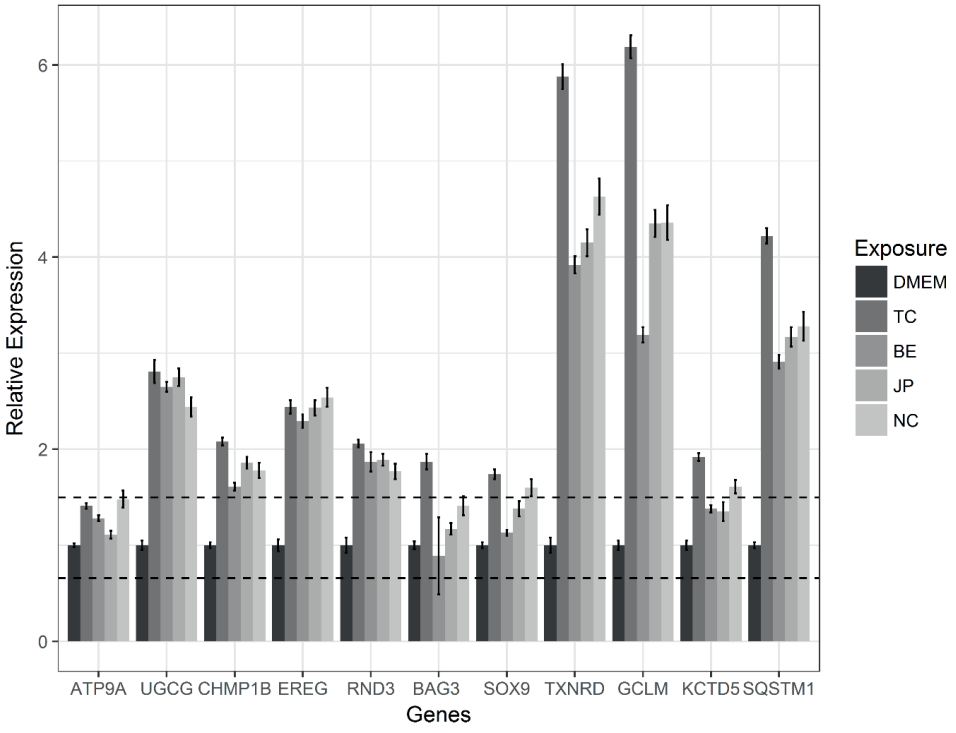


Figure 5. *qPCR results of both AhR and Nrf2 pathways provided together for genes predicted using DECA algorithm. The line represents the fold change cut-off limits (1.5 for up regulation and 0.6 for down regulation). CYP1A1 is not shown here as it exceeds the plot limits. TC indicates Turkish coffee, BE indicates Brasil Espirito, JP indicates Java Preanger and NC indicates Nescafe©*

The above mentioned 10 genes along with a seed gene for each pathway were experimentally verified using qPCR analysis in Caco-2 cells exposed to coffee samples (Figure 5). The results

indicate that 75% of the selected GOI showed a substantial relative difference in expression (absolute fold change > 1.5) in all tested samples, 2 genes (SOX9 and KCTD5) were differentially expressed upon exposure to two of the coffee extracts (Turkish and Nescafe, absolute fold change > 1.5) while the control genes showed no significant change in expression in most coffee extracts, as expected.

These results indicate that the DECA is a substantially improved strategy to identify GOI compared to other methods discussed in this paper and moreover do not require prior knowledge of the genes within the pathway except for the seed genes.

Discussion

Initially we focussed on developing an intestinal enterocyte-specific association network using expression data from Caco-2 cells exposed to different nutrients and stimuli. The network was constructed by selecting 12849 genes (actively) expressed in Caco-2 based on UPC filtering. This is consistent with previous observations of 11559 [26] and 14113 genes [24] based on RNAseq data (Caco-2 cells grown under controls). Differences could be attributed to different selection procedures or experimental approaches. Additionally, the gene list and network provided in this paper are based on a compendium of transcriptomics data from exposure of Caco-2 cells to different nutrients and stimuli.

When applying our Caco-2-specific selection to STRING network the number of edges and nodes was reduced considerably (~50%). The number of connected components is reduced by over 60% and the local network structure is preserved with similar values of clustering coefficient, which suggests a more compact network, as expected for gene that are functionally closely related. The degree assortativity decreases indicating less redundancy on gene associations when the network is restricted to Caco-2. Incidentally STRING could support dedicated data analysis by enabling seamless tissue specific gene selection.

Biclustering simultaneously clusters both genes and samples to arrive at the identification of genes with similar expression profiles in a subset of the samples. Existing biclustering algorithms do not allow targeting a particular pathway [45,46], instead they generally try to find biclusters that cover either a broad range of genes or conditions. Similarly WGCNA based clustering does not focus on a particular pathway but looks for modules of co-expressed genes that may belong to more than one pathway. Here we present a bi-clustering approach, that represents a modification of that in van dam et al, that allows the user to select or pre-select the seed genes and thus a pathway [47]. Nevertheless, biclustering performed poorly as the identified GOI did not show significant DE, indicating little responsiveness of Caco-2 cells to coffee exposures.

Therefore, DECA algorithm was used, resulting in a list of responsive gene candidates and a set of criteria to further rank them. From the ranked list, genes were selected for experimental verification in Caco-2 cells exposed to coffee and we found association with AhR and Nrf2 pathways. The verified genes were not in these pathways as defined in IPA. It might be that some of these genes have an indirect association to these pathways. The DECA ranking can be combined with existing knowledge, for instance, adding weight to genes on the basis of literature evidence. Of the 5 genes predicted for Nrf2 pathway, GCLM and TXNRD1 are previously known downstream gene targets of NRF2 [39,40]. KCTD5 is likely to have an indirect interaction mediated by CUL3 [41] and BAG3 (negative control gene) has been associated with Nrf2 pathway [42] while we find that only Turkish coffee induces this gene. Similarly for the genes predicted for AhR pathway, UGCG is indirectly linked to AhR pathway via ARNT [43] and EREG is reported as a target gene for AHR [44].

Seed genes play a critical role in predicting responsive genes in a certain pathway and should be carefully considered and accurately selected. As an example, Nrf2 gene was initially included among the seed genes for the biclustering algorithm. However, experimental verification showed transcript levels of this gene not to be responsive to coffee exposure. It was later not used as seed gene for DECA algorithm and was replaced with NQO1. One optimal way to select

seed genes is to select two or three highly differentially expressed genes (Fold Change > 3) associated to the pathway of interest from literature (eg. CYP1A1 and TIPARP for AhR pathway), verify their altered expression in response to activation or repression of the pathway and use these as seed genes.

The biclustering algorithm requires a further selection of genes to be considered, the gene pool set. This selection was performed by aggregating non cell type specific pathway level information. On the other hand, DECA has no such constraint and the whole set of expressed genes are considered. Therefore DECA is our method of choice to identify GOI in pathways for which little information is available. One could also argue that, when combining such a large set of array data collected over different batches, batch correction techniques should be applied. However, here each experiment has its own control in the same batch. As a result batch effects and experimental effects might be confounded and usually applied correction methods such as ComBat and SVA are not effective [48,49]. Instead, we have used a higher level integration approach, in which data from each study is compared with the corresponding control. This way we bypass the need for additional batch corrections as we study only correlations between changes in gene expression.

In addition to predicting GOI, the compendium presented in this paper can be used for other purposes. For instance, a systematic categorization of the treatments based on expression profile, similar to the approach taken in Connectivity map [50] and thus could select food components that have effects on certain genes and pathways. Such datasets can also be used to predict key regulators and/or gene hubs [2]. Additionally, the database can be expanded further by adding data from future experiments, even from technologies like RNAseq. The provided Caco-2 specific network also serves as a platform to understand future experiments. Gene expression data from a new experiment could be integrated with this network by using algorithms for network mining and active module identification [3]. The Caco-2 cell type specific network can also be used to develop networks associated to different conditions such as Caco-2 exposure to pathogens or pathogenic toxins, then these networks can be used to identify

potential drug targets by applying statistical methods and identifying hub genes using similar strategies as the one successfully used in cancer research [51]. This paper can therefore be seen as a first important step to improve current analysis tools for Caco-2 and thereby elicit a better understanding of the interaction between our intestinal epithelium and luminal (nutritional) compounds.

Conclusion

Caco-2 cell lines are increasingly used as model systems to study the interaction of food and other luminal factors with the intestinal system of the host, which is difficult to study *in vivo*. As the availability of experimental datasets will grow further we believe that this work is the first step in generation of a Caco-2 specific database and tissue specific research tools and strategies to extract more knowledge from these data. One of the research tools for which we make an important step is the dedicated protein-protein association network using gene expression data for Caco-2. The network provided in this paper could be the basis to be implemented in other software tools like IPA and STRING and can be further updated when more data become available in the future. The modified biclustering and DECA methods should additionally provide the necessary tools to extract genes of a desired pathways and can be applied, by the codes provided, to a similar dataset of any cell type of interest.

In the future, a comprehensive Caco-2 transcriptome database should include microarray data from other platforms such as agilent, illumina, etc but more importantly should include RNAseq data which will provide additional information on splice isoforms. We believe that such a cohesive database would provide finer results regarding the genes of interest in Caco-2 and can support the analysis and understanding of future Caco-2 cell based analysis. The dataset can additionally be used for building classifiers using genetic profiling and in finding therapeutic food solutions.

Acknowledgement

We would like to thank Renata Ariens for helping with qPCR experiments. The project was financial supported by the Dutch Ministry of Economic Affairs within the Systems Biology programme 'Virtual Gut', KB-17-003.02-021 and by The Scientific and Technological Research Council of Turkey (TUBITAK)

References

1. Saccenti, E., Suarez-Diez, M., Luchinat, C., Santucci, C. & Tenori, L. Probabilistic Networks of Blood Metabolites in Healthy Subjects As Indicators of Latent Cardiovascular Risk. *J. Proteome Res.* **14,** 1101–1111 (2015).

2. Li, R.-H. *et al.* Multiple differential expression networks identify key genes in rectal cancer. *Cancer Biomark.* **16,** 435–444 (2016).

3. Mitra, K., Carvunis, A.-R., Ramesh, S. K. & Ideker, T. Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet.* **14,** 719–732 (2013).

4. Ellis, J. D. *et al.* Tissue-Specific Alternative Splicing Remodels Protein-Protein Interaction Networks. *Mol. Cell* **46,** 884–892 (2012).

5. Altay, G., Asim, M., Markowetz, F. & Neal, D. E. Differential C3NET reveals disease networks of direct physical interactions. *BMC Bioinformatics* **12,** 296 (2011).

6. Cs, F., Rn, B., Jr, B., G, L. & Rf, M. Polyamine metabolism and transforming growth factor-beta signaling are affected in Caco-2 cells by differentially cooked broccoli extracts. *J. Nutr.* **138,** 1840–1845 (2008).

7. Murphy, E. F., Hooiveld, G. J., Muller, M., Calogero, R. A. & Cashman, K. D. Conjugated linoleic acid alters global gene expression in human intestinal-like Caco-2 cells in an isomer-specific manner. *J. Nutr.* **137,** 2359–2365 (2007).

8. Matsuki, T. *et al.* Epithelial Cell Proliferation Arrest Induced by Lactate and Acetate from Lactobacillus casei and Bifidobacterium breve. *PLOS ONE* **8,** e63053 (2013).

9. Engle, M. J., Goetz, G. S. & Alpers, D. H. Caco-2 cells express a combination of colonocyte and enterocyte phenotypes. *J. Cell. Physiol.* **174,** 362–369 (1998).

10. Nakano, E. *et al.* Riboflavin Depletion Impairs Cell Proliferation in Adult Human Duodenum: Identification of Potential Effectors. *Dig. Dis. Sci.* **56,** 1007–1019 (2010).

11. Dihal, A. A. *et al.* Pathway and single gene analyses of inhibited Caco-2 differentiation by ascorbate-stabilized quercetin suggest enhancement of cellular processes associated with development of colon cancer. *Mol. Nutr. Food Res.* **51,** 1031–1045 (2007).

12. Traka, M. *et al.* Transcriptome analysis of human colon Caco-2 cells exposed to sulforaphane. *J. Nutr.* **135,** 1865–1872 (2005).

13. Pereira-Caro, G. *et al.* Hydroxytyrosyl ethyl ether exhibits stronger intestinal anticarcinogenic potency and effects on transcript profiles compared to hydroxytyrosol. *Food Chem.* **138,** 1172–1182 (2013).

14. Turroni, F. *et al.* Genome analysis of Bifidobacterium bifidum PRL2010 reveals metabolic pathways for host-derived glycan foraging. *Proc. Natl. Acad. Sci.* **107,** 19514–19519 (2010).

15. He, X., Mishchuk, D. O., Shah, J., Weimer, B. C. & Slupsky, C. M. Cross-talk between E. coli strains and a human colorectal adenocarcinoma-derived cell line. *Sci. Rep.* **3,** (2013).

16. Arbibe, L. *et al.* An injected bacterial effector targets chromatin access for transcription factor NF-κB to alter transcription of host genes involved in immune responses. *Nat. Immunol.* **8,** 47–56 (2007).

17. Eskandarian, H. A. *et al.* A Role for SIRT2-Dependent Histone H3K18 Deacetylation in Bacterial Infection. *Science* **341,** 1238858 (2013).

18. Ishimoto, Y., Nakai, Y., Satsu, H., Totsuka, M. & Shimizu, M. Transient up-regulation of immunity- and apoptosis-related genes in Caco-2 cells cocultured with THP-1 cells evaluated by DNA microarray analysis. *Biosci. Biotechnol. Biochem.* **74,** 437–439 (2010).

19. Christensen, J. *et al.* Defining new criteria for selection of cell-based intestinal models using publicly available databases. *BMC Genomics* **13,** 274 (2012).

20. Eyking, A. *et al.* Toll-like Receptor 4 Variant D299G Induces Features of Neoplastic Progression in Caco-2 Intestinal Cells and Is Associated With Advanced Human Colon Cancer. *Gastroenterology* **141,** 2154–2165 (2011).

21. Lenaerts, K., Bouwman, F. G., Lamers, W. H., Renes, J. & Mariman, E. C. Comparative proteomic analysis of cell lines and scrapings of the human intestinal epithelium. *BMC Genomics* **8,** 91 (2007).

22. Liu, X., Yu, X., Zack, D. J., Zhu, H. & Qian, J. TiGER: A database for tissue-specific gene expression and regulation. *BMC Bioinformatics* **9,** 271 (2008).

23. Petryszak, R. *et al.* Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res.* **44,** D746–D752 (2016).

24. Zhong, J. *et al.* Resolving Chromosome-Centric Human Proteome with Translating mRNA Analysis: A Strategic Demonstration. *J. Proteome Res.* **13,** 50–59 (2014).

25. Sahoo, S. & Thiele, I. Predicting the impact of diet and enzymopathies on human small intestinal epithelial cells. *Hum. Mol. Genet.* **22,** 2705–2722 (2013).

26. Ghaffari, P. *et al.* Identifying anti-growth factors for human cancer cell lines through genome-scale metabolic modeling. *Sci. Rep.* **5,** 8183 (2015).

27. Wells, J. M., Rossi, O., Meijerink, M. & Baarlen, P. van. Epithelial crosstalk at the microbiota–mucosal interface. *Proc. Natl. Acad. Sci.* **108,** 4607–4614 (2011).

28. Sokolis, D. P. & Sassani, S. G. Microstructure-based constitutive modeling for the large intestine validated by histological observations. *J. Mech. Behav. Biomed. Mater.* **21,** 149–166 (2013).

29. Hybertson, B. M., Gao, B., Bose, S. K. & McCord, J. M. Oxidative stress in health and disease: The therapeutic potential of Nrf2 activation. *Mol. Aspects Med.* **32,** 234–246 (2011).

30. Mimura, J. & Fujii-Kuriyama, Y. Functional role of AhR in the expression of toxic effects by TCDD. *Biochim. Biophys. Acta* **1619,** 263–268 (2003).

31. Abel, J. & Haarmann-Stemmann, T. An introduction to the molecular basics of aryl hydrocarbon receptor biology. *Biol. Chem.* **391,** 1235–1248 (2010).

32. Ishikawa, T., Takahashi, S., Morita, K., Okinaga, H. & Teramoto, T. Induction of AhR-Mediated Gene Transcription by Coffee. *PLOS ONE* **9,** e102152 (2014).

33. Bøhn, S. K., Blomhoff, R. & Paur, I. Coffee and cancer risk, epidemiological evidence, and molecular mechanisms. *Mol. Nutr. Food Res.* **58,** 915–930 (2014).

34. Boettler, U. *et al.* Coffees rich in chlorogenic acid or N-methylpyridinium induce chemopreventive phase II-enzymes via the Nrf2/ARE pathway in vitro and in vivo. *Mol. Nutr. Food Res.* **55,** 798–802 (2011).

35. Paur, I., Balstad, T. R. & Blomhoff, R. Degree of roasting is the main determinant of the effects of coffee on NF-κB and EpRE. *Free Radic. Biol. Med.* **48,** 1218–1227 (2010).

36. Somoza, V. Five years of research on health risks and benefits of Maillard reaction products: An update. *Mol. Nutr. Food Res.* **49,** 663–672 (2005).

37. Lopes, C. T. *et al.* Cytoscape Web: an interactive web-based network browser. *Bioinformatics* **26,** 2347–2348 (2010).

38. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45,** D353–D361 (2017).

39. Loboda, A., Damulewicz, M., Pyza, E., Jozkowicz, A. & Dulak, J. Role of Nrf2/HO-1 system in development, oxidative stress response and diseases: an evolutionarily conserved mechanism. *Cell. Mol. Life Sci.* **73,** 3221–3247 (2016).

40. Solis, W. A. *et al.* Glutamate–cysteine ligase modifier subunit: mouse Gclm gene structure and regulation by agents that cause oxidative stress. *Biochem. Pharmacol.* **63,** 1739–1754 (2002).

41. Liu, Z., Xiang, Y. & Sun, G. The KCTD family of proteins: structure, function, disease relevance. *Cell Biosci.* **3,** 45 (2013).

42. Kwak, M.-K. *et al.* Modulation of gene expression by cancer chemopreventive dithiolethiones through the Keap1-Nrf2 pathway. Identification of novel gene clusters for cell survival. *J. Biol. Chem.* **278,** 8135–8145 (2003).

43. Sutter, C. H. *et al.* EGF receptor signaling blocks aryl hydrocarbon receptor-mediated transcription and cell differentiation in human epidermal keratinocytes. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 4266–4271 (2009).

44. Haarmann-Stemmann, T., Bothe, H. & Abel, J. Growth factors, cytokines and their receptors as downstream targets of arylhydrocarbon receptor (AhR) signaling pathways. *Biochem. Pharmacol.* **77,** 508–520 (2009).

45. Reiss, D. J., Baliga, N. S. & Bonneau, R. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics* **7,** 280 (2006).

46. Huttenhower, C. *et al.* Detailing regulatory networks through large scale data integration. *Bioinformatics* **25,** 3267–3274 (2009).

47. Dam, J. C. van, Schaap, P. J., Santos, V. A. M. dos & Suárez-Diez, M. Integration of heterogeneous molecular networks to unravel gene-regulation in Mycobacterium tuberculosis. *BMC Syst. Biol.* **8,** 111 (2014).

48. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28,** 882–883 (2012).

49. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8,** 118–127 (2007).

50. Lamb, J. *et al.* The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science* **313,** 1929–1935 (2006).

51. Zaman, N. *et al.* Signaling Network Assessment of Mutations and Copy Number Variations Predict Breast Cancer Subtype-Specific Drug Targets. *Cell Rep.* **5,** 216–223 (2013).

52. Piccolo, S. R. *et al.* A single-sample microarray normalization method to facilitate personalized-medicine workflows. *Genomics* **100,** 337–344 (2012).

53. Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4,** 249–264 (2003).

54. Piccolo, S. R., Withers, M. R., Francis, O. E., Bild, A. H. & Johnson, W. E. Multiplatform single-sample estimates of transcriptional activation. *Proc. Natl. Acad. Sci.* **110,** 17778–17783 (2013).

55. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20,** 307–315 (2004).

56. Dai, M. *et al.* Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **33,** e175–e175 (2005).

57. Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17,** 520–525 (2001).

58. Trevor Hastie, Robert Tibshirani, Balasubramanian Narasimhan & Gilbert Chu. impute: impute: Imputation for microarray data. R.

59. Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43,** D447-452 (2015).

60. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring Network Structure, Dynamics, and Function using NetworkX. in *Proceedings of the 7th Python in Science Conference* (eds. Varoquaux, G., Vaught, T. & Millman, J.) 11–15 (2008).

61. Remondini, D. *et al.* Complex patterns of gene expression in human T cells during in vivo aging. *Mol. BioSyst.* **6,** 1983–1992 (2010).

62. Novokmet, M. *et al.* Changes in IgG and total plasma protein glycomes in acute systemic inflammation. *Sci. Rep.* **4,** (2014).

63. Zhang, B. & Horvath, S. A General Framework for Weighted Gene Co-Expression Network Analysis. *Stat. Appl. Genet. Mol. Biol.* **4,** (2005).

64. Smyth, G. K. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Stat. Appl. Genet. Mol. Biol.* **3,** 1–25 (2004).

65. Eun, C. S. *et al.* Attenuation of colonic inflammation by PPARgamma in intestinal epithelial cells: effect on Toll-like receptor pathway. *Dig. Dis. Sci.* **51,** 693–697 (2006).

66. Jeon, M. K., Klaus, C., Kaemmerer, E. & Gassler, N. Intestinal barrier: Molecular pathways and modifiers. *World J. Gastrointest. Pathophysiol.* **4,** 94–99 (2013).

67. Wullaert, A., Bonnet, M. C. & Pasparakis, M. NF-κB in the regulation of epithelial homeostasis and inflammation. *Cell Res.* **21,** 146–158 (2011).

68. Jain, A. *et al.* p62/SQSTM1 Is a Target Gene for Transcription Factor NRF2 and Creates a Positive Feedback Loop by Inducing Antioxidant Response Element-driven Gene Transcription. *J. Biol. Chem.* **285,** 22576–22591 (2010).

69. Adachi, T. *et al.* Nrf2-dependent and -independent induction of ABC transporters ABCC1, ABCC2, and ABCG2 in HepG2 cells under oxidative stress. *J. Exp. Ther. Oncol.* **6,** 335–348 (2007).

70. Isshiki, M., Umezawa, K. & Tamura, H. Coffee Induces Breast Cancer Resistance Protein Expression in Caco-2 Cells. *Biol. Pharm. Bull.* **34,** 1624–1627 (2011).

71. Steinkellner, H. *et al.* Coffee consumption induces GSTP in plasma and protects lymphocytes against (±)-anti-benzo[a]pyrene-7,8-dihydrodiol-9,10-epoxide induced DNA-damage: Results of controlled human intervention trials. *Mutat. Res. Mol. Mech. Mutagen.* **591,** 264–275 (2005).

72. Yeager, R. L., Reisman, S. A., Aleksunes, L. M. & Klaassen, C. D. Introducing the 'TCDD-Inducible AhR-Nrf2 Gene Battery'. *Toxicol. Sci.* **111,** 238–246 (2009).

73. Kalthoff, S., Ehmer, U., Freiberg, N., Manns, M. P. & Strassburg, C. P. Coffee induces expression of glucuronosyltransferases by the aryl hydrocarbon receptor and Nrf2 in liver and stomach. *Gastroenterology* **139,** 1699–1710, 1710–2 (2010).

74. Diani-Moore, S. *et al.* Identification of the Aryl Hydrocarbon Receptor Target Gene TiPARP as a Mediator of Suppression of Hepatic Gluconeogenesis by 2,3,7,8-Tetrachlorodibenzo-p-dioxin and of Nicotinamide as a Corrective Agent for This Effect. *J. Biol. Chem.* **285,** 38801–38810 (2010).

Supplementary tables and figures

Supplementary Table S1*: Table contains the details of the experimental data included in the Publication along with their Accession number and ArrayExpress URL where available. Assays used implies the number of samples from each accession that are used in making the publication.*

Experiment name	Title	Platform	Accession	Assays used	Total Number of Assays	Release Date
broccoli extracts that had been cooked for different lengths of time	Transcription profiling of human colon Caco2 cells treated with extracts from broccoli that had been cooked for different lengths of time	hg 133 plus 2.0	E-MEXP-1372	12	12	30-09-2008
Caco-2 cells co-cultivated with B. animalis subsp. lactis BB-12	Expression data from Caco-2 cells co-cultivated with B. animalis subsp. lactis BB-12	hg 133 plus 2.0	E-GEOD-21930	3	3	30-07-2010
Caco-2 with THP1 coculture	Caco-2 cocultured with THP-1, time course	hg 133 plus 2.0	E-GEOD-17625	6	6	06-04-2010
conjugated linoleic acid (CLA)	Transcription profiling of human Caco-2 cells treated with conjugated linoleic acid (CLA)	hg 133 plus 2.0	E-GEOD-6518	9	9	14-06-2008
E. coli strains	Cross-talk between E. coli strains and a human colorectal adenocarcinoma-derived cell line	hg 133 plus 2.0	E-GEOD-50040	18	18	23-08-2013
Establishment of objective criteria for selecting relevant intestinal cell-based models	Establishment of objective criteria for selecting relevant intestinal cell-based models	hg 133 plus 2.0	E-GEOD-30292	15	43	23-06-2012
hydroxytyrosol (HTy) and hydroxytyrosyl ethyl ether (HTy-Et)	Transcription profiling of human colon Caco-2 cells treated with hydroxytyrosol (HTy) and hydroxytyrosyl ethyl ether (HTy-Et)	hg 133 plus 2.0	E-GEOD-38833	9	9	19-06-2012
Polydextrose fermentation metabolite effect	Polydextrose fermentation metabolite effect on Caco-2 colon cancer cells	hg 133 plus 2.0	E-GEOD-28792	15	15	22-04-2011

Riboflavin depletion	Riboflavin depletion impairs cell proliferation in intestinal cells: Identification of mechanisms and consequences	hg 133 plus 2.0	E-GEOD-15132	18	18	06-03-2009
Bifidobacterium bifidum PRL2010 on gene expression in intestinal epithelial cells	Effects of Bifidobacterium bifidum PRL2010 on gene expression in intestinal epithelial cells	Nugo	E-GEOD-21976	8	8	01-10-2010
Ascorbate-stabilized quercetin	Transcription profiling of human Caco-2 cell differentiation by ascorbate-stabilized quercetin	Hg-133a-2.0	E-GEOD-7259	8	8	15-06-2008
Caco-2 co-culture with Lactobacillus casei and Bifidobacterium breve	Caco-2 cell gene expression following co-culture with Lactobacillus casei and Bifidobacterium breve	Hg-133a-2.0	E-GEOD-37369	9	9	09-04-2013
sulforaphane (SF)	Transcription profiling of human colon Caco-2 cells treated with sulforaphane (SF)	hg-u133a	E-MEXP-170	8	8	01-07-2005
wild type Shigella flexneri and a OspF mutant	Transcription profiling of human Caco-2 intestinal epithelial cell line infected with wild type Shigella flexneri and a OspF mutantt reveals an injected bacterial effector targets chromatin access for NF-kB as a strategy to shape transcription of immune genes	hg-u133a	E-GEOD-6082	8	8	14-06-2008
Expression data from Caco-2 cells expressing TLR4 and associated mutants	Expression data from Caco-2 cells expressing TLR4 and associated mutants	hugene-1.0ST	E-GEOD-26226	12	12	10-10-2011
Listeria monocytogenes	Transcription profiling by array of Caco-2 cells infected with Listeria monocytogenes, treated and untreated with AGK2.	hugene-1.0ST	E-MEXP-3912	10	10	01-06-2013
Berry experiment	In preparation For Submission	Hugene-1.1 ST	From the lab of J. Mes	66	In preparation	In preparation
Dietary fibres	In preparation For Submission	Hugene-1.1 ST	From the lab of J. Mes	72	In preparation	In preparation
Probiotics	In preparation For Submission	Hugene-1.1 ST	From the lab of J. Mes	23	In preparation	In preparation
onion experiment	In preparation For Submission	Hugene-1.1 ST	From the lab of J. Mes	24	In preparation	In preparation

***Supplementary Table S2**: Table below shows the values computed for comprehensive assessment of DECA algorithm using 10 pathways mentioned in the materials and methods*

Pathway	Seed Gene 1	Seed Gene 2	Seed Gene 3	Total genes	Pwy Genes top 10 percent	P.val unadj	Ratio	Average ratio per pathway
TLR_signalling	5608_at	3665_at	148022_at	106	8	0.03873	0.08	0.07
TLR_signalling	2353_at	3665_at	3654_at	106	6	0.03703	0.06	
TLR_signalling	5970_at	3725_at	2353_at	106	6	0.036	0.06	
TLR_signalling	3576_at	5970_at	3665_at	106	10	0.01676	0.09	
TLR_signalling	3725_at	4792_at	3576_at	106	4	0.02181	0.04	
TLR_signalling	2353_at	3665_at	5970_at	106	8	0.04128	0.08	
TLR_signalling	148022_at	4792_at	3654_at	106	6	0.12846	0.06	
TLR_signalling	3576_at	3725_at	148022_at	106	8	0.04208	0.08	
TLR_signalling	3576_at	5608_at	3665_at	106	7	0.00961	0.07	
TLR_signalling	5970_at	3576_at	5608_at	106	11	0.0138	0.1	
ABC_transporters	23461_at	368_at	4363_at	45	5	0.0292	0.11	0.09
ABC_transporters	1080_at	9429_at	19_at	45	4	0.02106	0.09	
ABC_transporters	4363_at	9429_at	10057_at	45	4	0.00584	0.09	
ABC_transporters	4363_at	10057_at	23461_at	45	4	0.12114	0.09	
ABC_transporters	23461_at	1080_at	4363_at	45	4	0.35114	0.09	
ABC_transporters	1080_at	19_at	4363_at	45	3	0.28679	0.07	
ABC_transporters	19_at	10057_at	1080_at	45	3	0.24994	0.07	
ABC_transporters	19_at	4363_at	368_at	45	3	0.02165	0.07	
ABC_transporters	1080_at	23461_at	9429_at	45	5	0.03535	0.11	
ABC_transporters	23461_at	368_at	19_at	45	4	0.0947	0.09	
Adherens_juction	8936_at	2260_at	5819_at	74	3	0.40711	0.04	0.08
Adherens_juction	5797_at	6934_at	71_at	74	11	0.00085	0.15	
Adherens_juction	8936_at	5797_at	4088_at	74	6	0.05678	0.08	
Adherens_juction	5819_at	5797_at	6934_at	74	5	0.03866	0.07	
Adherens_juction	71_at	8936_at	6934_at	74	9	0.00021	0.12	
Adherens_juction	4088_at	5819_at	2260_at	74	5	0.20803	0.07	
Adherens_juction	8936_at	5797_at	2260_at	74	4	0.30792	0.05	
Adherens_juction	71_at	8936_at	4088_at	74	7	0.07895	0.09	
Adherens_juction	5819_at	5797_at	2260_at	74	3	0.40129	0.04	
Adherens_juction	71_at	8936_at	2260_at	74	5	0.16672	0.07	
Fat_absorption	80168_at	10555_at	19_at	41	4	0.0109	0.1	0.1
Fat_absorption	10555_at	8611_at	80168_at	41	4	0.04471	0.1	
Fat_absorption	8611_at	10555_at	19_at	41	5	0.01328	0.12	
Fat_absorption	80168_at	19_at	10555_at	41	4	0.0109	0.1	
Fat_absorption	80168_at	19_at	10555_at	41	4	0.0109	0.1	
Fat_absorption	80168_at	19_at	10555_at	41	4	0.0109	0.1	
Fat_absorption	19_at	8611_at	80168_at	41	3	0.33412	0.07	
Fat_absorption	80168_at	8611_at	10555_at	41	4	0.04471	0.1	

Fat_absorption	10555_at	80168_at	19_at	41	4	0.0109	0.1	
Fat_absorption	80168_at	8611_at	10555_at	41	4	0.04471	0.1	
Gap_junction	7277_at	10381_at	347733_at	88	7	0.00259	0.08	0.1
Gap_junction	84617_at	7280_at	6654_at	88	10	0.29133	0.11	
Gap_junction	10381_at	6654_at	2697_at	88	8	0.11547	0.09	
Gap_junction	347733_at	7280_at	84790_at	88	11	0.01228	0.13	
Gap_junction	7280_at	10381_at	84617_at	88	7	0.0115	0.08	
Gap_junction	6654_at	84790_at	56034_at	88	8	0.1911	0.09	
Gap_junction	10381_at	84790_at	56034_at	88	10	0.01484	0.11	
Gap_junction	7277_at	7846_at	2697_at	88	6	0.01016	0.07	
Gap_junction	84617_at	7846_at	1813_at	88	9	0.03125	0.1	
Gap_junction	84790_at	347733_at	7846_at	88	8	0.01634	0.09	
Glycerolipid metabolism	9388_at	84803_at	80339_at	59	5	0.00376	0.08	0.1
Glycerolipid metabolism	8611_at	10555_at	9388_at	59	8	0.09312	0.14	
Glycerolipid metabolism	80339_at	132158_at	8611_at	59	4	0.0104	0.07	
Glycerolipid metabolism	84803_at	80339_at	9388_at	59	5	0.00376	0.08	
Glycerolipid metabolism	80339_at	10555_at	9388_at	59	8	0.02462	0.14	
Glycerolipid metabolism	80339_at	11343_at	84803_at	59	6	0.00979	0.1	
Glycerolipid metabolism	8527_at	132158_at	57678_at	59	6	0.18471	0.1	
Glycerolipid metabolism	8611_at	10555_at	84803_at	59	6	0.03608	0.1	
Glycerolipid metabolism	9388_at	84803_at	132158_at	59	4	0.02052	0.07	
Glycerolipid metabolism	9388_at	8611_at	57678_at	59	8	0.11261	0.14	
Glycerophospholipid metabolism	10555_at	8611_at	79143_at	95	6	0.16953	0.06	0.07
Glycerophospholipid metabolism	57678_at	56261_at	23761_at	95	9	0.86616	0.09	
Glycerophospholipid metabolism	57678_at	2819_at	5337_at	95	9	0.27249	0.09	
Glycerophospholipid metabolism	10555_at	2819_at	8611_at	95	4	0.12988	0.04	
Glycerophospholipid metabolism	79888_at	10555_at	8611_at	95	7	0.0221	0.07	
Glycerophospholipid metabolism	5337_at	23761_at	2819_at	95	8	0.23537	0.08	
Glycerophospholipid metabolism	57678_at	79143_at	84803_at	95	7	0.5983	0.07	
Glycerophospholipid metabolism	8611_at	8527_at	2819_at	95	3	0.96371	0.03	
Glycerophospholipid metabolism	2819_at	79143_at	23761_at	95	6	0.242	0.06	
Glycerophospholipid metabolism	2819_at	79143_at	79888_at	95	7	0.04837	0.07	
NFKB-signalling	5971_at	148022_at	330_at	93	16	0.00029	0.17	0.15
NFKB-signalling	51588_at	5328_at	9020_at	93	12	0.00198	0.13	

NFKB-signalling	598_at	9020_at	5743_at	93	13	0.00028	0.14	
NFKB-signalling	598_at	3383_at	5328_at	93	17	0.00211	0.18	
NFKB-signalling	5328_at	7132_at	5743_at	93	14	0.00189	0.15	
NFKB-signalling	3383_at	9020_at	598_at	93	16	0.0022	0.17	
NFKB-signalling	7128_at	51588_at	598_at	93	15	0.00376	0.16	
NFKB-signalling	148022_at	5743_at	4616_at	93	12	0.01239	0.13	
NFKB-signalling	5970_at	598_at	4616_at	93	12	0.00728	0.13	
NFKB-signalling	4616_at	5971_at	51588_at	93	11	0.00081	0.12	
p53_signalling	595_at	1026_at	5054_at	69	9	0.01606	0.13	0.11
p53_signalling	595_at	7057_at	545_at	69	7	0.65044	0.1	
p53_signalling	1026_at	595_at	1647_at	69	7	0.01273	0.1	
p53_signalling	27244_at	545_at	5054_at	69	10	0.02123	0.14	
p53_signalling	4616_at	8795_at	901_at	69	4	0.0006	0.06	
p53_signalling	901_at	545_at	5366_at	69	9	0.06978	0.13	
p53_signalling	2810_at	5366_at	27244_at	69	8	0.0008	0.12	
p53_signalling	8795_at	7057_at	1647_at	69	4	0.00828	0.06	
p53_signalling	5054_at	5366_at	9133_at	69	11	0.00032	0.16	
p53_signalling	7057_at	9133_at	901_at	69	10	0.00639	0.14	
PPAR-signalling	123_at	1376_at	1374_at	72	4	0.34538	0.06	0.06
PPAR-signalling	8309_at	1376_at	1622_at	72	4	0.03086	0.06	
PPAR-signalling	123_at	4312_at	1376_at	72	4	0.09617	0.06	
PPAR-signalling	4312_at	1962_at	1622_at	72	5	0.16514	0.07	
PPAR-signalling	1962_at	1376_at	4312_at	72	5	0.02223	0.07	
PPAR-signalling	123_at	4312_at	1962_at	72	5	0.03471	0.07	
PPAR-signalling	1622_at	8309_at	1374_at	72	4	0.09354	0.06	
PPAR-signalling	4312_at	51703_at	123_at	72	4	0.22211	0.06	
PPAR-signalling	51703_at	4312_at	1374_at	72	5	0.21034	0.07	
PPAR-signalling	51703_at	1622_at	1962_at	72	5	0.02004	0.07	

Supplementary Table S3*: Network statistics of the Caco-2 specific protein-protein interaction network.*

Metrics	STRING network (Entrez Id)	Caco-2 Network
# of nodes	13762	8937
# of edges	272903	129711
# of connected components	131	54
Density of Graph	0.0029	0.0032
Average degree	39.66	29.03
Giant component nodes	13462	8824
Giant component edges	272698	129648
Degree Assortativity	0.57	0.13
Average Clustering Coefficient	0.34	0.34

Supplementary text F1

Culturing & experimental exposure of Caco-2 cells

The Caco-2 cell line was obtained from the American-Type Culture Collection (ATCC HTB-37TM; USA). The cells were routinely grown in 75 cm^2 tissue culture flasks (with canted neck and 0.2 µm vented cap, Corning, 430641) using Dulbecco's Modified Eagle's Medium (DMEM) (Invitrogen, 42430-082; with 4.5 g/L glucose, no pyruvate, 4 mM L-glutamine, and 25 mM HEPES) supplemented with 9.1% Fetal Bovine Serum (FBS, Hyclone Perbio) (Fischer Scientific CH 30160.03; heat inactivated at 56ºC for 45 min). Cells were maintained in a humidified atmosphere of 5% CO_2 in air at 37ºC and sub-cultured at 80-90% confluence. For exposure experiments, Caco-2 cells were grown on transwells (Greiner bio-one, 662640, translucent, 0.4 µm pores, $1x10^8$ pores/cm, 0.312 cm^2 surface area for cell growth) in 24-well plates (Greiner bio-one Cellstar plates, 662102, Alphen a/d Rijn, The Netherlands). Cells, having a passage number between 30 and 45, were seeded at a concentration of $0.225x10^6$ cells/mL and grown in DMEM supplemented with 10% FBS, at 37ºC and 5% CO_2 in air. Cells were allowed to grow for 7 days and the culture medium was replaced every two days. To ensure that the monolayers exhibit the properties of a tight biological barrier, transepithelial electrical resistance (TEER) was monitored using a MilliCell-ERS voltohmmeter (Millipore Co., United States). Monolayers with TEER values exceeding 300 $\Omega.cm^2$ were used exclusively for the experiments.

Turkish coffee (obtained from a local market in Turkey), 2 types of filtered coffees (Java Preanger and Brasil Espirito) (obtained from a local market in The Netherlands), and instant coffee (Nescafe Gold Blend obtained from a local market in The Netherlands) samples were brewed without any sugar and/or milk addition. For Turkish coffee brew, 10 g of ground coffee sample was cooked with 130 mL of MQ water until the boiling point (53). For the filtered coffee brews, 8 g of powder was extracted with 140 mL of boiled MQ water (54) and for the instant coffee brew, 2 g of instant coffee was solubilized in 150 mL of boiled MQ water (55). All coffee

brews were first filtered (Whatman Filter Paper, 589/1, ashless, Whatman, U.K.), 2 times, and then freeze-dried, and stored at -80ºC until analysis.

For the Caco-2 cell exposure experiments, the freeze-dried coffee extracts were re-dissolved in cell culture medium (DMEM with 9.1% FBS) to give the same dry-weight concentrations as in the original coffee brews. For the sample treatments, the culture medium was first removed from the well, and then coffee samples were added, in duplicates, to the apical side of the cells, with a volume of 150 µL; while the basolateral side was refreshed with culture medium only, with a volume of 700 µL. Cells were incubated with samples for 24 h at culture conditions. During sample incubations, the cell monolayer integrity was checked with TEER measurements. After completion of the exposure experiments, the cells were harvested for RNA extraction.

After exposure, total RNA was isolated from the Caco-2 cells by using 200 µL of TriZol (Invitrogen, 15596-026, Paisley, UK). The TriZol extracts were subsequently treated with DNase (Qiagen, RNase free DNase set, #79254, Hilden, Germany), and purified with RNeasy mini columns (Qiagen, Hilden, Germany), using the protocol supplied by the manufacturer. The concentration and purity of the RNA samples were determined spectrophotometrically using a NanoDrop (ND-1000 Spectrophotometer, Thermo Fisher Scientific, Wilmington, USA).

One microgram of total RNA was reverse transcribed into cDNA using iScript™ cDNA Synthesis Kit (Bio-Rad, Hercules, CA, USA) in a final volume of 20 µL. Primers used for the amplification of reference genes (β-actin, GAPDH, RPLP0) and the target genes (AhR, ARNT, CYP1A1, TiPARP, ABCC1, ABCC2, ABCG2, Nrf2, NQO1, GSTP1, GSTM4, GSTA2, UGT1A6, HMOX-1, SQSTM-1, ATF4, NFIA, PRKCA, ENC1 (NRPB), UGCG, EREG, RND3, CHMP1B, ATP9A, GCLM, TXNRD1, SOX9 KCTD5, BAG3) are given in Table T1 provided below.

qPCR amplification was performed with 5 µL diluted (40 times diluted) cDNA sample, 2.5 µL of each primer (3.2 µM for CYP1A1; 0.8 µM for the other primers) and 10 µL of SYBR Green Supermix (Bio-Rad, Cat# 172-5006CUST, Hercules, USA) in a final volume of 20 µL. Every sample was run in technical duplicates. Gene expression analysis was conducted on a BioRad

CFX96 Real-Time System with C1000 Thermal Cycler. Gene expression levels were calculated using Biogazelle qbase^{plus} (Zwijnaarde, Belgium) program and the expression values of the selected target genes were normalized using the reference genes β-actin, RPLP0, and GAPDH (Table T1, provided below)

Table T1: *Details of the gene names that were used in qPCR experiments and their corresponding Primers*

Gene Name	Gene Symbol	Forward Primer (5¢-3¢)	Reverse Primer (5¢-3¢)
Beta Actin	ACTB	CTGGAACGGTGAAGGTGACA	AAGGGACTTCCTGTAACAATGCA
Glyceraldehyde-3-phosphate dehydrogenase	GAPDH	TGCACCACCAACTGCTTAGC	GGCATGGACTGTGGTCATGAG
Ribosomal protein, large, P0	RPLP0	GCAATGTTGCCAGTGTCTG	GCCTTGACCTTTTCAGCAA
Aryl hydrocarbon receptor	AhR	ACATCACCTACGCCAGTCG	CGCTTGGAAGGATTTGACTTGA
Aryl hydrocarbon receptor nuclear translocator	ARNT	GGAACAAGATGACAGCCTAC	CAGAAAGCCATCTGCTGCC
Cytchrome P450, family 1, subfamily A, polypeptide 1	CYP1A1	TCTTTGGAGCTGGGTTTG	ACTGTGTCTAGCTCCTCTTG
TCDD-inducible poly (ADP-ribose) polymerase	TiPARP	AGAACGAGTGGTTCCAATCCA	TGGGTGCAAAAGATCAGTCTG
ATP-binding casette, sub-family C, member1	ABCC1	CTCTATCTCTCCCGACATGACC	AGCAGACGATCCACAGCAAAA
ATP-binding casette, sub-family C, member2	ABCC2	TCTCTCGATACTCTGTGGCAC	CTGGAATCCGTAGGAGATGAAGA
ATP-binding cassette, sub-family G, member 2	ABCG2	ACGAACGGATTAACAGGGTCA	CTCCAGACACACCACGGAT
Nuclear factor (erythroid-derived 2)-like 2	Nrf2	TCCAGTCAGAAACCAGTGGAT	GAATGTCTGCGCCAAAAGCTG
NAD(P)H dehydrogenase, quinone 1	NQO1	GGGATCCACGGGGACATGAATG	ATTTGAATTCGGGCGTCTGCTG
Glutathione S-transferase pi	GSTP1	TGCAAATACATCTCCCTCATCTACA	CGGGCAGTGCCTTCACAT
Glutathione S-transferase mu 4	GSTM4	AGAGGAGAAGATTCGTGTGGA	TGCTGCATCATTGTAGGAAGTT
Glutathione S-transferase alpha 2	GSTA2	TACTCCAATATACGGGGCAGAA	TCCTCAGGTTGACTAAAGGGC
UDP-glucuronosyltransferase 1-6	UGT1A6	TGATCCTGGCTGAGTATTTGGG	TGGGAATGTAGGACACAGGGT
Heme oxygenase (decycling) 1	HMOX-1	TCTCTTGGCTGGCTTCCTTA	ATTGCCTGGATGTGCTTTTC
Sequestosome 1	SQSTM1	GCACCCCAATGTGATCTGC	CGCTACACAAGTCGTAGTCTGG
Activating transcription factor 4	ATF4	ATGACCGAAATGAGCTTCCTG	GCTGGAGAACCCATGAGGT
Nuclear factor I/A	NFIA	GCAGGCCCGAAAACGAAAATA	TTTGCCAGAAGTCGAGATGCC
Protein kinase C, alpha	PRKCA	GTCCACAAGAGGTGCCATGAA	AAGGTGGGGCTTCCGTAAGT
Ectodermal neural cortex 1	ENC1 (NRPB)	GCTGCTGTCTGATGCACAC	AGAGTTGCACTACCATGTCCT
Rho family GTPase 3	RND3	GCTCCATGTCTTCGCCAAG	AAAACTGGCCGTGTAATTCTCA
UDP-glucose ceramide glucosyltransferase	UGCG	GAATGGCCGTCTTCGGGTT	AGGTGTAATCGGGTGTAGATGAT
ATPase, class II, type 9A	ATP9A	AAGTCAACTCCCAGGTCTACAG	CGCTGGTTCTTTTCAACGATGA
charged multivesicular body protein 1B	CHMP1B	GAATGAGTGCGCGAGTCGAT	GGTCTTCAATGTCGCATCCAT
epiregulin	EREG	GGACAGTGCATCTATCTGGTGG	TTGGTGGACGGTTAAAAAGAAGT
glutamate-cysteine ligase, modifier subunit	GCLM	CATTTACAGCCTTACTGGGAGG	ATGCAGTCAAATCTGGTGGCA
glutamate-cysteine ligase, modifier subunit	TXNRD	CATTTACAGCCTTACTGGGAGG	ATGCAGTCAAATCTGGTGGCA
SRY (sex determining region Y)-box 9	SOX9	AGCGAACGCACATCAAGAC	CTGTAGGCGATCTGTTGGGG
potassium channel tetramerization domain containing 5	KCTD5	AACGAGACAGCAAAACATCGC	TGACCAACTGCTCGAACTTCC
BCL2-associated athanogene 3	BAG3	TGGGAGATCAAGATCGACCC	GGGCCATTGGCAGAGGATG

R Codes and networks

The R codes along with examples for biclustering and DECA and the Caco-2 specific PPIN (as edgelist) are available at http://semantics.systemsbiology.nl/

Exploring the role of miRNAs in regulation of the Caco-2 cell transcriptional response to *Clostridium difficile* toxins

Prashanna Balaji Venkatasubramanian[1], Renata Ariens[1], Els Oosterink[1], Edoardo Saccent[2], Jurriaan J. Mes[1] and *Nicole de Wit[1].

Manuscript in final stages of preparation

Abstract

Clostridium difficile produces two key toxins called toxin A and toxin B which are known to be important for cytopathic and cytotoxic effects of *Clostridium difficile* infection (CDI). The molecular effects of these toxins on enterocytes have been studied with primary focus on the changes induced in mRNA (gene) expression of the host cells. The impact of the toxins on miRNA expression of the host cells have largely been unexplored.

We have investigated the impact of the *C. difficile* toxins on the expression of mRNAs and miRNAs in enterocytes and particularly focused on the potential role of miRNAs in gene expression regulation for which we used Caco-2 cells as the enterocyte model system. We further mapped the interactions between miRNAs and mRNAs that were inversely regulated by the toxins using the miRNA-mRNA interaction database, miRWalk, to identify potential miRNA target genes. Subsequently, we performed network analysis to identify hub miRNAs and mRNAs. Moreover, our pathway enrichment analysis using IPA® showed that miRNAs might have a role in *C. difficile*-induced changes in cell proliferation, intestinal barrier function and immune responses.

Introduction

Clostridium difficile is an anaerobic, motile, gram positive bacteria present in the intestine of many mammals. Some strains of *C. difficile* cause a symptomatic infection in humans called *Clostridium difficile* Infection (CDI). Incidence of CDI and subsequent mortality has increased in recent years in USA, Canada and Europe [1]. The pathogenesis due to CDI , and the mechanism of action of toxin A (toxA) and toxin B (toxB) has been an area of intense investigation [2–8]. To gain more insight into the cellular effects and mechanisms of action of the toxins, gene expression studies have been conducted in *in vitro* using human cell lines HCT-8 [5] and Caco-2 [9], as well as *in vivo* in the mouse cecum [7].

The *C. difficile* toxins A and B bind to surface receptors on host cells, resulting in receptor-mediated endocytosis and further intoxicate the cells by disrupting Rho-GTPase dependent signalling [10] leading to altered cell morphology and apoptosis. Studies indicate that cellular pathways related to cell cycle/proliferation [5], cholesterol metabolism, fatty acid biosynthesis [7], interleukin signalling related to innate immunity are enriched based on transcriptomics studies after exposure to toxins.

Transcription factors are generally known to play a key role in regulation of gene expression, but in recent years it has become clear that non-coding RNA (ncRNAs) especially microRNAs (miRNA), also play an important regulatory role. The miRNA are non-coding RNAs of 20-25 base pairs in length [11–13] and are mainly involved in mRNA post-transcriptional regulation likely by targeting 6–8 nucleotides (called the seed region) in the 5' end of the miRNA [14,15]. The regulation of mRNA by miRNA is carried out by a multitude of steps that finally ends in the formation of RNA-induced silencing complex (RISC) complex [14]. The miRNA along with other proteins forms the RISC complex and binds to mRNA based on complementarity of the nucleotides. The mRNA bound to RISC complex is destabilized via the degradation of poly A-tail and thus the transcription is reduced [16]. To date it is clear that miRNAs contribute to the

regulation of protein synthesis via mRNA destabilization or via translational repression. There are multiple, contradictory mechanisms reported in different tissues and conditions [11] and several models are still hypothesized for the translational inhibition by miRNAs which need further in depth research [17].

About one thousand miRNAs are predicted in humans[12]. Investigations in the past have been reported on the role of miRNA in bacterial infection in intestine and in influencing commensal bacterial population [18–20]. Host microRNA expression changes have been observed following infection with bacteria like *Salmonella enterica*, *Listeria monocytogenes* and also pathogens from *Mycobacterium* and *Francisella* species. These changes play a role in triggering immune responses against the bacterial infection [18] and mitigation of excessive inflammation and have shown to involve miRNAs such as miR-146, miR-155, miR-125, let-7 and miR-21 [18,19]. Liu and others have investigated the role of intestinal epithelial miRNA in shaping the host microbiota [20]. To the best of our knowledge the role of miRNA in regulation of transcription response to *C. difficile* toxins has not been investigated in detail.

In this study, we investigated the role of miRNA in *C. difficile* toxin-induced changes in gene expression and the linked biological processes. For this, the Caco-2 cell line model was used and the effect of simultaneous exposure to both toxA and toxB. The miRNA expression was investigated using miRNA microarrays and mRNA expression was investigated using RNAseq. The results from these two experiments were used along with existing knowledge from public databases to probe miRNA-mRNA interaction. We demonstrated that a substantial number of biological processes related to Caco-2 cells in response to *C. difficile* toxins were likely controlled by miRNAs.

Methods and Materials

The overall experimental and computational analysis workflow used in this study is illustrated in Figure 1. Caco-2 cells exposed to *C. difficile* toxins were harvested for isolation of mRNA and

small RNAs. The expression of mRNA was measured using RNAseq and miRNA expression was assessed using miRNA microarrays. The resulting data was processed and differentially expressed (DE) miRNAs and mRNAs were identified. The interactions between DE mRNAs and miRNAs were obtained from miRWalk 2.0 database and network analysis performed using Cytoscape. The DE mRNA was used for pathway enrichment analysis using IPA©. Similarly, pathway analysis was performed using the target genes of the miRNAs. An overlap between the two enriched lists of pathways were compared for a final analysis.

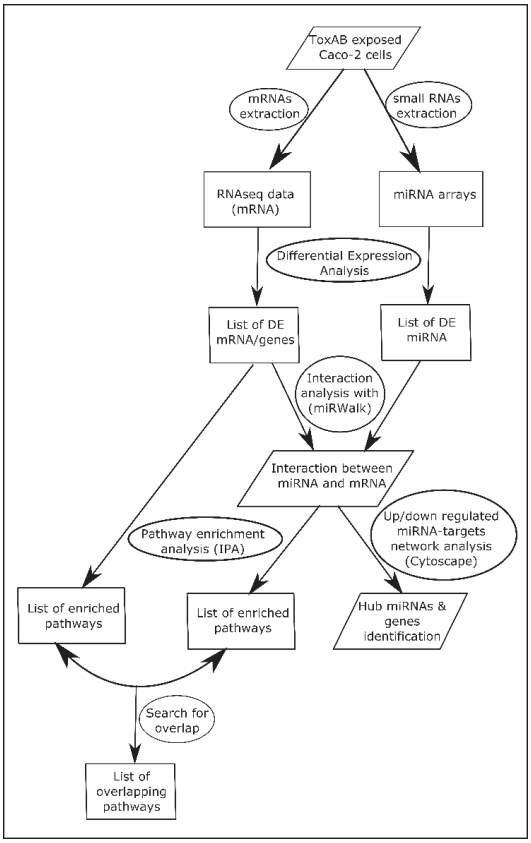


***Figure 1**: Flow diagram showing the methods used in this study.*

Cell culture, toxin exposure and TEER estimation

ATCC derived Caco-2 cells were cultured in Dulbecco's modified Eagle's medium (DMEM; Gibco-Invitrogen, Bleiswijk, The Netherlands) with 4.5 g/L glucose, 0.58 g/L glutamine, no pyruvate, supplemented with 10% heat inactivated FBS (Hyclone Perbio, Etten-Leur, The Netherlands) and used with passage numbers between 30 and 40. For transwell assays, 330,000 cells were seeded on ThinCert transwells (Greiner Bio-one) with 33.6 mm^2 membranes and 0.4 µm pores in 24-well suspension culture plates. Cells were grown for 7 days at 5% CO_2 and 37°C and apical (150 µL) and basolateral (700 µL) medium was replaced every other day.

All *C. difficile* toxins were derived from List Biological Laboratories, Inc. (Campbell, California, USA; Toxin A (#152), Toxin B (#155) and dilutions were made in DMEM/FBS. For toxin incubation, 7 days differentiated Caco-2 cells were exposed to 0.25µg/ml toxin A + 0.25µg/ml toxin B (ToxAB) in triplicates. These conditions were chosen based on a dose-response pilot study. One day before the exposure experiments, medium was refreshed and at the day of exposure, medium was removed from the apical and basal compartments and toxin samples were added to the apical compartment while fresh DMEM/FBS medium was added to the basal compartment. To monitor the integrity of the Caco-2 monolayer, transepithelial electrical resistance (TEER) was measured at 37ºC using a MilliCell-ERS Ω meter (Millipore, Molsheim, France). The exposure experiments were performed three times to obtain three independent biological replicates.

RNA extraction and RNAseq

TEER was measured every hour and cells were harvested at a ~35% drop in TEER after exposure to the toxins which was approximately 4.5 hours after initial exposure. This threshold was used since a drop in TEER value below 35% resulted in irreversible damage to the cells and activation of cell-death related processes. Caco-2 cells were lysed with 300 µL TRIzol

(Invitrogen, Life Technologies, Bleiswijk, Netherlands) and the triplicates in each experiment were pooled for RNA isolation.

RNA-isolation

Caco-2 cells were lysed with 300 µl Trizol (Invitrogen, Life Technologies, Bleiswijk, the Netherlands) and the triplicates in each experiment were pooled for RNA isolation. Total RNA was isolated using TRIzol Reagent according to the manufacturer's instructions until the RNA was in the aqueous phase. Subsequently the RNA was further isolated and purified with the RNeasy Mini kits (Qiagen, Venlo, the Netherlands) following the manufacturers 'protocol with the addition of a DNase treatment (Qiagen). RNA concentration and purity were measured using the Nanodrop ND-1000 spectrophotometer (Nanodrop Technologies, Wilmington, USA) [24].

For the isolation of miRNA the same steps were followed with slight adaptations to the RNeasy kit protocol to ensure that the miRNAs were present in the column. 70% Ethanol was changed to 100% Ethanol and buffer RW1 was replaced by buffer RWT (Qiagen, Venlo, the Netherlands).

RNAseq data analysis

The raw mRNA data obtained after Illumina sequencing were processed using CLC Genomics® Workbench (www.qiagenbioinformatics.com/) software version 8.5.1 for alignment to transcriptome, RPKM calculation and differential expression analysis. The RNAseq reads were aligned to hg38 whole human transcriptome dataset (gene and mRNA tracks) obtained from Ensemble database [21]. The transcriptome dataset contained 60,448 transcripts. The reads were aligned only to gene regions with a mismatch cost set at 2. The RPKM values were obtained for 23000 genes (rest of the 60448 transcript annotations were ncRNAs and small RNAs). The transcripts were further filtered using Universal expression protocol (UPC) algorithm to assess their presence/ absence. Transcripts were assumed to be present if they crossed a threshold cut-off of UPC value > 0.5. This was followed by differential expression analysis to detect the gene expression changes between control and toxA + toxB (toxAB) exposed Caco-2 cells. The

statistical significance was assessed using a two-tailed paired t-test (provided as Gaussian t-test in CLC Genomics®). Differential expression significance was set at FDR corrected p-value < 0.01.

miRNA analysis

Small RNAs were hybridised using miRNA arrays (miRNA 4.1) from Affymetrix® as mentioned above and the resulting CEL files were processed using transcriptome analysis console software. The data was normalized using expression console (now part of transcriptome analysis console) and the normalized probeset data was converted to miRNA level data using annotation file obtained from Affymetrix website (miRNA-4_1-st-v1). This resulted in 6631 miRNAs and differential expression analysis between the control and toxAB induced Caco-2 cells was performed using transcriptome analysis console. The data was tested for significance using one-way paired ANOVA and miRNAs with p-value < 0.01 were considered significant.

miRNA- mRNA interaction analysis

To identify target genes of differentially expressed miRNAs, miRWalk 2.0 [22,23] was used. miRWalk is a database and tool containing information on both validated and predicted miRNA - target interactions. The miRWalk validated targets resulted from text mining search obtained from 4 databases [23]. miRWalk predicted interactions are from sequence homology searches. In addition to miRWalk predicted list of genes data was obtained from 3 other prediction databases namely, miRanda, RNA22 and Targetscan. This list was combined with predicted targets from miRWalk and genes were chosen for further analysis if they were predicted as a target by at least 3 of the above-mentioned databases. In order to identify the target genes that are responsible for direct (primary) miRNA-mRNA interactions, target genes that were significantly regulated in the opposite direction to the miRNA regulation were chosen for further analysis.

The miRNA-mRNA interactions were subjected to network analysis using Cytoscape [24,25]. The interaction network was loaded on Cytoscape and all self-loops were removed from the network.

The network analyser plugin was used to calculate network metrics like degree of nodes and other centrality measures.

Pathway enrichment analysis

Pathway enrichment analysis was performed using Ingenuity® Pathway Analysis [26] (IPA) tool (QIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis). The pathway analysis was performed with default setting. Significance values for enrichment were calculated using Fisher's exact test right-tailed and significant pathways were estimated based on FDR corrected p-value < 0.05.

Results

In our study, Caco-2 cells were exposed to both *C. difficile* toxins toxA and toxB (toxAB) and after a 35% drop in TEER harvested to analyse ToxAB induced changes in gene expression. The harvested mRNAs were subject to RNAseq and small RNAs were subjected to miRNA arrays. Differential expression analysis and pathway enrichment analysis were performed on the resulting data. The miRNA-mRNA interaction was further studied as well.

RNAseq – UPC

Using the data from RNAseq technology and Universal exPression Code (UPC) protocol (UPC value > 0.5), 11890 genes were found to be transcriptionally active in Caco-2 cells. After subjecting the data to differential expression analysis, 826 transcripts were found to be differentially expressed between DMEM control and ToxAB (p-value < 0.01). Further analysis identified 492 transcripts which were down-regulated and 334 that were up-regulated. The top 15 up and top 15 down-regulated genes are presented in table 1.

Table 1: *Top 15 up regulated and top 15 down regulated mRNAs along with their Fold Change. '*' indicates mRNAs regulated by one or more miRNAs.*

Gene Symbol	Fold Change
*PTGER4	12.71
RHOB	12.58
*KLF6	10.32
JUN	8.61
*ATF3	6.04
*MAFF	5.99
GEM	5.17
*HAS3	4.83
BTG2	4.60
PLK3	4.22
ARL14	4.21
*NOS2	4.16
*WEE1	4.06
*PLAUR	3.93
*IER3	3.92
AC091729.9	-2.45
DCP1B	-2.46
RPUSD2	-2.47
JADE1	-2.61
*MEF2C	-2.64
CCDC51	-2.66
*FFAR4	-2.67
PRKCQ-AS1	-2.67
*TIGD2	-2.67
*AMOTL2	-2.90
CSRP2BP	-2.93
*SP8	-3.39
*DFFB	-3.62
*TTC30B	-4.56
CASS4	-6.67

Among the highly downregulated genes, most are involved in cell cycle/apoptosis (e.g. AMOTL2, DFFB, JADE1, MEF2C, CSRP2BP) and influencing cell adhesion/migration (e.g Cass4). Among the highly upregulated genes are RHOB (involved in actin and cytoskeletal dynamics, cell movement), JUN and ATF3, all previously found to be upregulated on exposure to *C. difficile* toxins in HCT-8 and mouse cecum. These genes are known to be involved in cell morphology and stress-induced responses and confirming that Caco-2 combined with RNAseq have been performed as could be expected based on literature.

The differential expression gene list was then analysed for pathway enrichment using IPA® to gain better insight into the pathways affected by *C. difficile* infection in the intestine. 71 pathways were found to be significantly (p-value < 0.05) enriched in total, among which are pathways related cell cycle/proliferation, innate-immunity, and protein biosynthesis. The complete list of enriched pathways are provided in the supplementary information table 1. Figure 2 shows the top 20 enriched canonical pathways based on their measure of significance (p-value). It seemed that pathways related to cell cycle/proliferation are most inhibited, whereas pathways related to immune function are stimulated. Among protein biosynthesis pathways, we found that genes involved in mitochondrial translation were particularly downregulated.

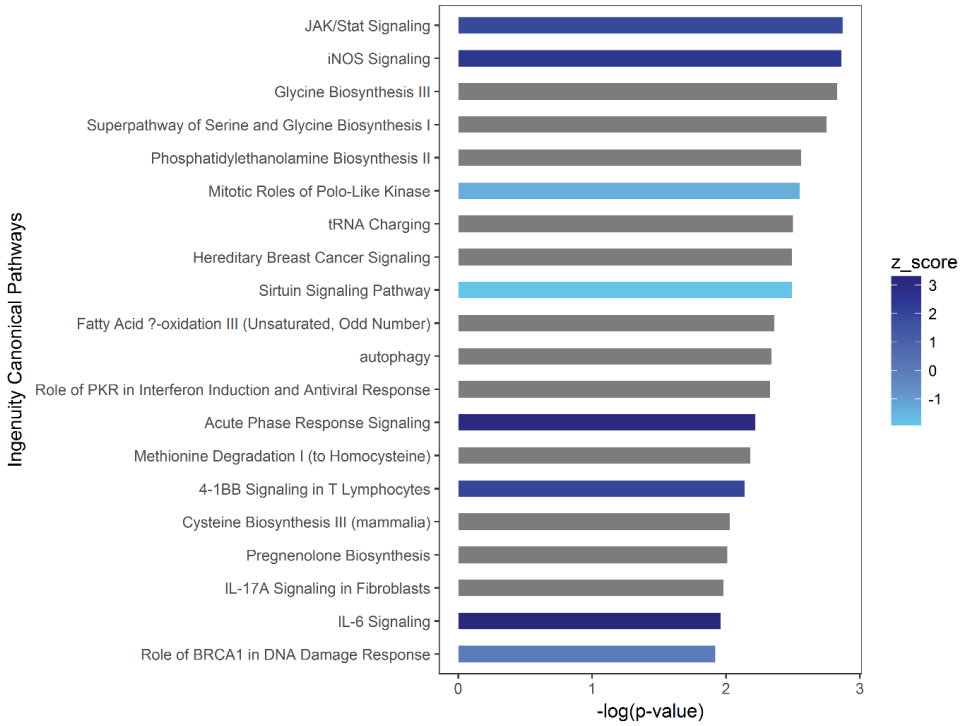


***Figure 2**: Shows the top 20 Ingenuity canonical pathways that were enriched based on the differentially expressed mRNA (gene) list. The z score indicates the inhibition (negative score) or stimulation (positive score) of a pathway. The grey bars indicate missing z-scores.*

***Table 2:** Significantly regulated miRNAs with Fold Change and validated target genes. Validated targets are based on an intersection between data obtained from miRWalk 2.0 and significantly differentially expressed in opposite direction based on RNAseq data.*

miRNA (miR id)	Fold Change	Validated targets - Entrez ID (based on miRWalk 2.0 and RNAseq)
miR-1343-3p	1.99	51421, 1678, 140461, 64149, 51808
mir-6804	1.36	
miR-6730-3p	1.27	23649
miR-4482-5p	1.25	10196
mir-4315-1	1.24	
mir-4315-2	1.24	
miR-200b-3p	1.21	54453
mir-548ag-1	1.2	5828, 158471, 83876

miR-4783-5p	1.19	
mir-6499	1.19	57677, 1678, 199857, 5828, 29099, 3659, 2868, 23404, 63931, 83876, 10196, 647087, 26227, 26146, 88745, 25896, 23593, 55696, 23649
miR-361-5p	1.14	51645, 4839, 51528, 25980, 84881
miR-128-3p	1.12	166815, 257407, 199857, 253635, 669, 65080, 64149, 55204, 63931, 11200, 51163, 58488, 4839, 22944, 29916
miR-4529-5p	1.11	
mir-375	1.1	1678, 56912, 79717, 1326, 22796, 9166, 339175, 55907, 53371, 84881
mir-4481	1.1	23404, 10714, 11331, 79621
miR-4781-5p	1.1	
miR-95-3p	1.09	
mir-16-2	1.07	5723, 51524, 8604, 54554, 4913, 55347, 51058, 84864, 55028, 8835, 64149, 81602, 55204, 7265, 60625, 51645, 80742, 65983, 1284, 25926, 57088, 63931, 51204, 51249, 10732, 57180, 79568, 11212, 90522, 55003, 57621, 10240, 80254, 10946, 51067, 57418, 84259, 440, 22944, 5565, 88745, 26273, 217, 85365, 23438, 29916
miR-218-1-3p	1.07	57677
mir-4435-1	-1.06	7832, 81631, 51621, 25987, 1969, 1454
mir-4435-2	-1.06	7832, 81631, 51621, 25987, 1969, 1454
miR-518e-3p	-1.08	
mir-194-2	-1.09	
mir-6785	-1.09	
mir-2052	-1.11	
miR-1297	-1.12	133, 1604, 387893, 1025, 79647
mir-548ba	-1.14	79647, 1027
miR-301a-5p	-1.16	4084, 11176, 80271, 11098, 9776
miR-6880-3p	-1.57	
miR-4767	-1.89	
miR-3200-5p	-1.93	11343
miR-3065-5p	-2.06	27242, 79791, 8440
miR-23a-5p	-2.62	257629, 1454
miR-27a-5p	-2.95	207

miRNA Differential Expression Analysis

miRNA expression was assessed and the miRNAs that showed differential expression between the control and treatment (toxin AB) were identified. After significance testing, 35 differentially

expressed miRNAs were identified. 16 miRNAs were found to be downregulated and 19 miRNAs were upregulated. Table 2 shows all the up and down regulated miRNAs.

miRNA – mRNA targets interaction reveals key miRNA regulation

To study the role of miRNAs in *C. difficile* toxin-induced cytotoxicity in enterocytes, target genes of the differentially expressed miRNAs were first identified using the miRWalk 2.0 tool as mentioned in the methods section. This filtering resulted in 5267 validated gene targets and 67264 predicted target genes for the 19 upregulated miRNAs and 1448 validated targets and 55202 predicted targets for the 16 downregulated miRNAs. The validated interactions in this list contained redundant interactions owing to multiple experimental evidences of interactions and all redundancies were removed in subsequent analysis. Further, target genes (both predicted and validated) that were not significantly regulated based on the RNAseq data, were removed. This resulted in 1435 predicted and 54 validated up-regulated targets genes for the downregulated miRNAs (opposite direction of gene expression) and 1648 predicted and 175 validated down-regulated target genes for the upregulated miRNAs.

The miRNA and their interacting target genes were subjected to network analysis using Cytoscape and the visualised networks are shown in Figure 3**.** The key results of the network analysis and their node degrees are provided in Table 3. Figure 3a shows upregulated miRNAs along with their target genes and Figure 3b shows the downregulated miRNAs and their target genes. Among the upregulated miRNAs, hsa-miR-16-5p, hsa-miR-128-3p and hsa-miR-1343-3p had the largest number of target genes (91, 77, 56 respectively) while among downregulated miRNAs, hsa-miR-194-3p, hsa-miR-3065-5p and hsa-miR-4435 had the largest number of target genes (65, 56 and 41 respectively). Among the target genes, RAB3B (upregulated expression) and UMPS (downregulated expression) were the genes that were most extensively regulated by multiple miRNAs. Both genes are connected to 10 miRNAs.

***Table 3**: Table containing key regulated miRNAs and target mRNAs. miRNAs with degree greater than 50 and mRNA with degree greater than 5 are shown. Table 3a contains downregulated miRNAs and upregulated target mRNAs. Table 3b contains upregulated miRNAs and downregulated target mRNAs.*

Down Regulated miRNA	Degree
miR-194-3p	65
miR-3065-5p	56
miR-4435	56

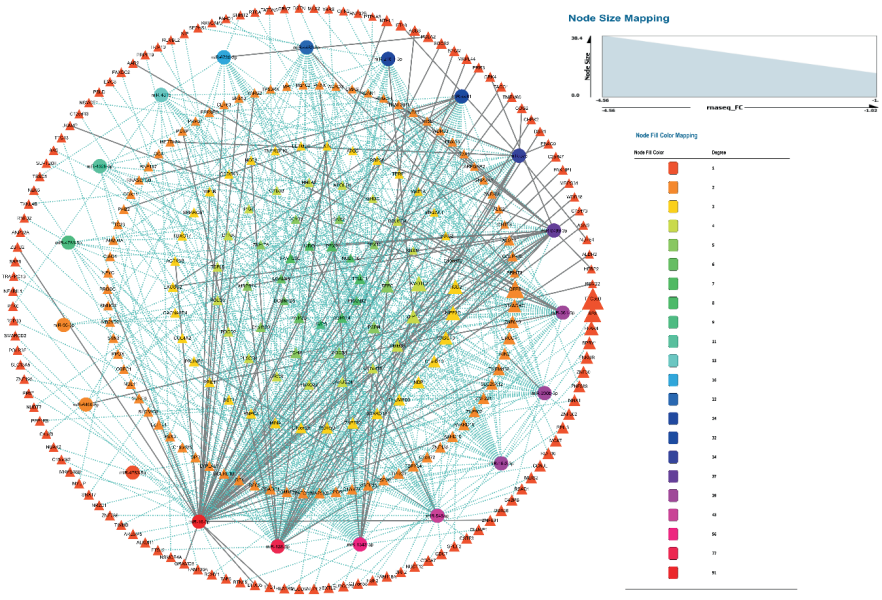
Up Regulated mRNA	Degree
RAB3B	10
THSD4	6
ATRN	6
MTF1	6
SMURF1	6
FAM64	6

a.

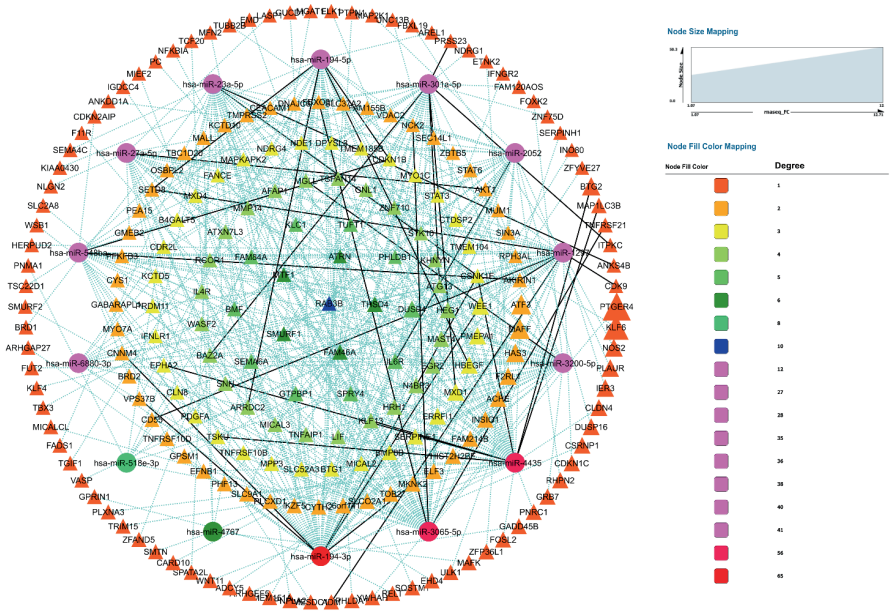
Up Regulated miRNA	Degree
miR-16-5p	91
miR-128-3p	77
miR-1343-3p	56

Down Regulated mRNA	Degree
UMPS	9
ÀGPAT4	8
PRKAB2	8
TTLL11	8
NUDT16	8
PEX2	7
MRO	7
FAM120C	7
LGALS8	6
DCUN1D5	6
PAIP2B	6

b.


Node Size Mapping
Node Size
maseq_FC
Node Fill Color Mapping
Node Fill Color
Degree

a.


Node Size Mapping
Node Size
maseq_FC
Node Fill Color Mapping
Node Fill Color
Degree
hsa-miR-194-5p
hsa-miR-23a-5p
hsa-miR-301a-5p
hsa-miR-27a-5p
hsa-miR-2052
hsa-miR-548k
hsa-miR-3200-5p
hsa-miR-5680-3p
hsa-miR-518e-3p
hsa-miR-4435
hsa-miR-4767
hsa-miR-3065-5p
hsa-miR-194-3p

b.

***Figure 3**: Figure represents the interaction between miRNA – mRNA. The miRNAs are represented by circles and mRNAs are represented by triangles. The solid black lines indicate validated interactions while light blue lines indicate predicted interactions. The size of the triangles indicates the fold change of the mRNA, i.e. a larger size implies a higher fold change. The nodes are arranged in concentric circles based on their degree. The outer most ring of triangles indicate mRNAs that have an interaction with only one miRNA (degree one). The next ring of triangles (within the ring of circles) are regulated by two miRNAs (degree two). The following concentric rings of triangles have higher degrees and the colours indicate the degree of each node. The miRNAs (represented by circles in the second ring) are arranged in an increasing order of degree in clockwise direction such that the miRNAs with maximum degree are located at the bottom of the ring. (a) Indicates the interaction between up-regulated miRNAs and their target genes (down regulated mRNAs). The largest triangle in the outer ring is that of PTGER4 gene. The gene at the centre is RAB3B. Additional key insights regarding the genes and miRNAs are given in table 3. (b) The interaction between down-regulated miRNAs and their target genes (up regulated mRNAs) are shown. The largest triangle in the outer ring is that of PTGER4 gene. Additional key insights regarding the genes and miRNAs are given in table 3.*

Among the down regulated mRNAs that are targets of upregulated miRNAs, UMPS was observed to be regulated by 9 miRNAs. Similarly, IL6R was found to be targeted by 5 miRNAs. Among the top upregulated genes, PTGER4 and KLF6 was regulated by mir-3065-5p and mir-23a-5p respectively, while ATF3 was regulated by 2 miRNAs. Interestingly, we did not find any miRNA that regulated RHOB and JUN. AGPAT4 (involved in biosynthetic process based on gene ontology) and NUDT16 (associated with positive regulation of cell proliferation) were found to be regulated by 8 miRNAs that were upregulated. Among the top down regulated genes (mRNAs), AMOTL2 was regulated by 4 miRNAs, MEF2C was regulated by 3 miRNAs, DFFB is regulated by 2 miRNAs and FFAR4 was regulated by miR-128-3p.

The target genes (both validated and predicted) obtained from miRWalk were further probed for pathway enrichment using IPA®. This resulted in 32 pathways significantly enriched based on

the target genes. Among these pathways, biosynthesis related pathways were found to be predominantly regulated by miRNAs and those related to mitochondrial functioning like phosphatidylethanolamine biosynthesis II and protein biosynthesis related pathways of serine and glycine biosynthesis I were also among the top regulated pathways.

Comparison of this result with the pathway enrichment list (71 pathways) obtained using RNAseq data revealed that 25 of the 71 pathways overlapped with miRNA target genes enriched pathways (Table 4). Of the top 20 pathways enriched using RNAseq data about 65% was found to be overlapping with miRNA target genes pathway enrichment results, among which are many immune response pathways, proliferation/cell cycle pathways.

Among the top pathways (Table 5) regulated by the miRNAs (based on p-value) are JAK/STAT signalling and Acute Phase Response Signalling which are involved in cell cycle/proliferation; Protein biosynthesis pathways like Phosphatidylethanolamine Biosynthesis II pathway and Superpathway of Serine and Glycine Biosynthesis I; immune related pathways like IL6 signalling and iNOS signalling. These pathways are found to play a role in *C. difficile* toxins induced enterocytes from mRNA based pathway analysis.

***Table 4**: The RNAseq data enriched pathways overlapping with miRNA target genes are provided along with p-values of enrichment*

RNAseq pathways	P-values
JAK/Stat Signalling	2.87
iNOS Signalling	2.86
Superpathway of Serine and Glycine Biosynthesis I	2.75
Phosphatidylethanolamine Biosynthesis II	2.56
Hereditary Breast Cancer Signalling	2.49
Sirtuin Signalling Pathway	2.49
autophagy	2.34
Acute Phase Response Signalling	2.22
Methionine Degradation I (to Homocysteine)	2.18
Cysteine Biosynthesis III (mammalia)	2.03
Pregnenolone Biosynthesis	2.01

IL-6 Signalling	1.96
Role of BRCA1 in DNA Damage Response	1.92
Serine Biosynthesis	1.86
Pancreatic Adenocarcinoma Signalling	1.75
Neuregulin Signalling	1.74
TGF-β Signalling	1.72
Cell Cycle Control of Chromosomal Replication	1.69
PI3K Signalling in B Lymphocytes	1.61
Asparagine Biosynthesis I	1.41
D-mannose Degradation	1.41
Oncostatin M Signalling	1.39
ERK5 Signalling	1.38
Heme Biosynthesis II	1.35
ErbB2-ErbB3 Signalling	1.33

Table 5: *Top 20 enriched pathways based on miRNA target genes*

miRNA target genes enriched pathways	**-log(p-value)**
Superpathway of Serine and Glycine Biosynthesis I	3.33
Phosphatidylethanolamine Biosynthesis II	3.14
Neuregulin Signalling	2.31
Serine Biosynthesis	2.25
Sirtuin Signalling Pathway	2.01
autophagy	2
Methionine Degradation I (to Homocysteine)	1.92
Cysteine Biosynthesis III (mammalia)	1.81
JAK/Stat Signalling	1.81
Heme Biosynthesis II	1.72
TGF-β Signalling	1.71
iNOS Signalling	1.66
Hereditary Breast Cancer Signalling	1.66
ERK5 Signalling	1.66
Acute Phase Response Signalling	1.64
Asparagine Biosynthesis I	1.61
D-mannose Degradation	1.61
ErbB2-ErbB3 Signalling	1.61
Sertoli Cell-Sertoli Cell Junction Signalling	1.59

Discussion

Several studies in the past have focused on the effects of *C. difficile* and its toxins on large intestine enterocytes and cecum gene expression (mRNA expression) [5,7,9]. These studies have highlighted the genes that showed significant expression changes and the concordant pathways that were enriched.

PTGER4 (EP4), a GPCR family gene, was found to be the most upregulated gene (Fold Change > 12). This gene has been found to be important for maintaining intestinal barrier, in suppressing mucosal damage [27] and in promoting intestinal repair through adaptive immune response [28]. Additionally, genes like RHOB, KLF6, ATF3 and JUN which are also reported in earlier studies using *C. difficile* toxin exposure to HCT-8 [5] cell line and mice cecum [7] have been observed in the top 10 upregulated genes (based on fold change) in our study. The RHOB is an important component of the RHO-GTPase signalling. In addition to it, other GTPases like ARL14, GEM (both among top 10 (Table 1)) and CDC42 were also observed to be upregulated. *C. difficile* toxins are known to transfer glucose molecule to the RHO-GTPases and thus affect the actin dependent processes [29]. The RHO-GTPases act as molecular switches regulating their downstream processes. *C. difficile* toxins glucosylate the RHO and CDC42 and thus make these proteins inactive and their functions are affected. KLF6, ATF3 and JUN are transcription factors and are found to be involved in cellular stress response and apoptosis [30–32]. Interestingly, CASS4 (HEPL) gene is the most downregulated gene (Fold change < -6) and is involved in cell adhesion [33]. CASS4 has not been previously associated with *C. difficile* toxins induced enterocytes and this may be due to the lesser sensitivity of arrays compared against RNAseq. Additionally, AMOTL2 which is indirectly involved in cellular adhesion by controlling transport of actin filaments is also downregulated [34] (among top 10, shown in Table 1). CSRP2BP (KAT14) is down regulated and is part of the ATAC complex playing a role in cell cycle [35,36]. All these results together indicate that the toxins induce a total loss of cell integrity by targeting the cell morphology and cell-cell adhesion.

Pathway analysis of RNAseq data shows inhibition of cell cycle/proliferation, which is likely in line with the drop in TEER that we found after exposure to the *C. difficile* toxins. This disruption of the intestinal integrity is indicative for the previously reported *C. difficile* toxin-induced cell death [10]. Stimulation of (innate) immune function through release of pro-inflammatory IL-6 signalling and anti-inflammatory IL-10 signalling were both found to be enriched pathways and have been previously observed [37,38]. *C. difficile* cytotoxin (toxB) was previously found to affect protein biosynthesis in intestinal epithelium [39] and we found that mitochondrial protein biosynthesis in particular, was affected. This was also previously reported for other bacterial toxins [40].

There are very few studies describing in detail the impact of the toxins on miRNA expression and the miRNA – mRNA interaction in the host. Li and colleagues have utilized whole genome microarray data from an earlier study involving Caco-2 co-culture with *C. difficile* bacteria [9]. They applied big data mining techniques to investigate the cross-talk between the host (Caco-2) and microbes (*C. difficile*) [38] and have predicted the potential impact of the miRNAs on the pathogens, expressed and secreted by the host, on the luminal pathogens. The study by Li et al, was focused on early to mid-stage infection (30 min to 120 min) while we focussed on later time point (4.5 hours). We found similar pathways to their late stage infection study, like inflammation related pathways and cellular biosynthesis. While earlier studies have focused on the effect of transcriptome during *C. difficile* infection and have incorporated elements on miRNA regulation, none of them studied in detail the impact of the toxins on miRNA expression of enterocytes and have not studied the miRNA – mRNA interaction in the host in detail. We explored the role of miRNA in host responses to *C. difficile* toxins in detail.

Previous studies have also reported involvement of miRNAs in bacterial infections. Among the miRNAs we have found to change upon exposure to the toxins are miRNA that were earlier reported to be regulated in host pathogen interaction. miR-128 has been reported to be upregulated in *Salmonella typhimurium* infection [41,42] and was found to be upregulated in our analysis. Studies indicate that this miRNA affects the cytoskeleton indirectly by altering CDC42,

a RHO-GTPase [42]. miR-16 was upregulated in *L. monocytogenes* infected epithelial cells [43] and we observed miR-16-2, a part of miR-16 microRNA precursor family, to be upregulated. The miR-16 is reported to target AU-rich sequences of TNF-A, IL6 and IL8 and degrade them rapidly [42]. While miR-23a was identified in relation to a *M. bovis* infection of *Bos taurus* [44], we observed miR-23a-5p to be downregulated. Interestingly, mir-3065 and mir-361 were found down regulated and up regulated in many infections in macrophages, respectively [45] but their potential role in enterocytes and infection is not studied yet. We also found miR-3065-5p to be downregulated and mir-361-5p to be upregulated.

In our study, the miRNA-mRNA interaction was explored using the miRWalk 2.0 database and the resulting interactions were screened for opposite regulation (up regulated miRNA with down regulated mRNA and *vice versa*). miRNA regulation of mRNA could either be primary or secondary. Primary regulation is direct interaction between an miRNA and mRNA while secondary regulation is when an miRNA regulates the upstream regulators of an mRNA. In order to ensure that only primary regulation of mRNA by miRNA were captured, we considered only those interactions where the differential expression of miRNA was in opposite direction to the mRNA.

The results of miRNA-mRNA interaction network analysis using Cytoscape reveals useful insight. Interestingly, RAB3B (which was upregulated) was found to be targeted by 10 miRNAs. RAB3B is a GTPase protein involved in epithelial polarization and tight junction regulation [46], which indicates a role in intestinal barrier function. Furthermore, Rab GTPases are reported to play a role in bacterial infections, however the exact role for RAB3B in this is not yet known [47]. RAB3B was previously reported to be regulated by several miRNAs [48–50]. Among the down regulated mRNAs that are targets of upregulated miRNAs, UMPS was observed to be regulated by 9 miRNAs. UMPS is reported to play a role in pyrimidine biosynthesis [51]. Previous studies have shown that pathogens require pyrimidine for colonisation in the intestine. Reduction of pyrimidine biosynthesis might be a defence response by the host cell [52,53].

We have identified mRNAs potentially regulated by multiple miRNAs and identified miRNAs with a large number of potential target genes. Such investigations shed light on miRNAs and encourage research towards miRNAs that can be used as potential therapeutic targets or agents in which miRNAs are either suppressed or mimicked. Several miRNA based therapeutics for cancer and hepatitis have already reached clinical trial phases [54]. However, this should be done with great caution as the miRNAs often regulate multiple mRNAs simultaneously and the interactions between specific miRNAs and mRNA have not been completely unravelled.

Pathway enrichment analysis of miRNA target genes resulted in several key pathways that were also regulated by mRNAs, *e.g.* JAK/STAT signalling, sirtuin signalling, *etc.* These results indicate that miRNA regulate significant number of toxin induced biological functions and thus play a substantial role in *C. difficile* toxin-induced cytopathologic effects in enterocytes.

Acknowledgement

The project was financial supported by the Dutch Ministry of Economic Affairs within the Systems Biology programme 'Virtual Gut', KB-17-003.02-021 and the TO2Flex programme.

References

1. Lessa, F. C., Gould, C. V. & McDonald, L. C. Current Status of Clostridium difficile Infection Epidemiology. *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.* **55,** S65–S70 (2012).

2. Triadafilopoulos, G., Pothoulakis, C., O'Brien, M. J. & LaMont, J. T. Differential effects of Clostridium difficile toxins A and B on rabbit ileum. *Gastroenterology* **93,** 273–279 (1987).

3. Riegler, M. *et al.* Clostridium difficile toxin B is more potent than toxin A in damaging human colonic epithelium in vitro. *J. Clin. Invest.* **95,** 2004–2011 (1995).

4. Hecht, G., Pothoulakis, C., LaMont, J. T. & Madara, J. L. Clostridium difficile toxin A perturbs cytoskeletal structure and tight junction permeability of cultured human intestinal epithelial monolayers. *J. Clin. Invest.* **82,** 1516–1524 (1988).

5. D'Auria, K. M. *et al.* Systems analysis of the transcriptional response of human ileocecal epithelial cells to Clostridium difficile toxins and effects on cell cycle control. *BMC Syst. Biol.* **6,** 2 (2012).

6. Du, T. & Alfa, M. J. Translocation of Clostridium difficile toxin B across polarized Caco-2 cell monolayers is enhanced by toxin A. *Can. J. Infect. Dis.* **15,** 83–88 (2004).

7. D'Auria, K. M. *et al.* In Vivo Physiological and Transcriptional Profiling Reveals Host Responses to Clostridium difficile Toxin A and Toxin B. *Infect. Immun.* **81,** 3814–3824 (2013).

8. Chumbler, N. M., Farrow, M. A., Lapierre, L. A., Franklin, J. L. & Lacy, D. B. Clostridium difficile Toxins TcdA and TcdB Cause Colonic Tissue Damage by Distinct Mechanisms. *Infect. Immun.* **84,** 2871–2877 (2016).

9. Janvilisri, T., Scaria, J. & Chang, Y.-F. Transcriptional profiling of Clostridium difficile and Caco-2 cells during infection. *J. Infect. Dis.* **202,** 282–290 (2010).

10. Voth, D. E. & Ballard, J. D. Clostridium difficile Toxins: Mechanism of Action and Role in Disease. *Clin. Microbiol. Rev.* **18,** 247–263 (2005).

11. Fabian, M. R., Sonenberg, N. & Filipowicz, W. Regulation of mRNA Translation and Stability by microRNAs. *Annu. Rev. Biochem.* **79,** 351–379 (2010).

12. Filipowicz, W., Bhattacharyya, S. N. & Sonenberg, N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat. Rev. Genet.* **9,** 102–114 (2008).

13. Kloosterman, W. P. & Plasterk, R. H. A. The Diverse Functions of MicroRNAs in Animal Development and Disease. *Dev. Cell* **11,** 441–450 (2006).

14. Bartel, D. P. MicroRNA Target Recognition and Regulatory Functions. *Cell* **136,** 215–233 (2009).

15. Ellwanger, D. C., Büttner, F. A., Mewes, H.-W. & Stümpflen, V. The sufficient minimal set of miRNA seed types. *Bioinformatics* **27,** 1346–1350 (2011).

16. Bartel, D. P. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell* **116,** 281–297 (2004).

17. Oliveto, S., Mancino, M., Manfrini, N. & Biffo, S. Role of microRNAs in translation regulation and cancer. *World J. Biol. Chem.* **8,** 45–56 (2017).

18. Eulalio, A., Schulte, L. & Vogel, J. The mammalian microRNA response to bacterial infections. *RNA Biol.* **9,** 742–750 (2012).

19. Staedel, C. & Darfeuille, F. MicroRNAs and bacterial infection. *Cell. Microbiol.* **15,** 1496–1507 (2013).

20. Liu, S. *et al.* The Host Shapes the Gut Microbiota via Fecal microRNA. *Cell Host Microbe* **19,** 32–43 (2016).

21. Aken, B. L. *et al.* The Ensembl gene annotation system. *Database* **2016,** (2016).

22. Dweep, H., Sticht, C., Pandey, P. & Gretz, N. miRWalk--database: prediction of possible miRNA binding sites by 'walking' the genes of three genomes. *J. Biomed. Inform.* **44,** 839–847 (2011).

23. Dweep, H. & Gretz, N. miRWalk2.0: a comprehensive atlas of microRNA-target interactions. *Nat. Methods* **12,** 697 (2015).

24. Lopes, C. T. *et al.* Cytoscape Web: an interactive web-based network browser. *Bioinformatics* **26,** 2347–2348 (2010).

25. Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinforma. Oxf. Engl.* **27,** 431–432 (2011).

26. Krämer, A., Green, J., Pollard, J. & Tugendreich, S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **30,** 523–530 (2014).

27. Kabashima, K. *et al.* The prostaglandin receptor EP4 suppresses colitis, mucosal damage and CD4 cell activation in the gut. *J. Clin. Invest.* **109,** 883–893 (2002).

28. Miyoshi, H. *et al.* Prostaglandin E2 promotes intestinal repair through an adaptive cellular response of the epithelium. *EMBO J.* **36,** 5–24 (2017).

29. Schoentaube, J., Olling, A., Tatge, H., Just, I. & Gerhard, R. Serine-71 phosphorylation of Rac1/Cdc42 diminishes the pathogenic effect of Clostridium difficile toxin A. *Cell. Microbiol.* **11,** 1816–1826 (2009).

30. Leppä, S. & Bohmann, D. Diverse functions of JNK signalling and c-Jun in stress response and apoptosis. *Oncogene* **18,** 6158 (1999).

31. Hai, T., Wolfgang, C. D., Marsee, D. K., Allen, A. E. & Sivaprasad, U. ATF3 and stress responses. *Gene Expr.* **7,** 321–335 (1999).

32. Andreoli, V., Gehrau, R. C. & Bocco, J. L. Biology of Krüppel-like factor 6 transcriptional regulator in cell life and death. *IUBMB Life* **62,** 896–905 (2010).

33. Nikonova, A. S., Gaponova, A. V., Kudinov, A. E. & Golemis, E. A. CAS proteins in health and disease: an update. *IUBMB Life* **66,** 387–395 (2014).

34. Hultin, S. *et al.* AmotL2 links VE-cadherin to contractile actin fibres necessary for aortic lumen expansion. *Nat. Commun.* **5,** 3743 (2014).

35. Guelman, S. *et al.* The double-histone-acetyltransferase complex ATAC is essential for mammalian development., The Double-Histone-Acetyltransferase Complex ATAC Is Essential for Mammalian Development. *Mol. Cell. Biol. Mol. Cell. Biol.* **29, 29,** 1176, 1176–1188 (2009).

36. Suganuma, T. *et al.* The ATAC Acetyltransferase Complex Coordinates MAP Kinases to Regulate JNK Target Genes. *Cell* **142,** 726–736 (2010).

37. Solomon, K. The host immune response to Clostridium difficile infection. *Ther. Adv. Infect. Dis.* **1,** 19–35 (2013).

38. Li, C.-W., Su, M.-H. & Chen, B.-S. Investigation of the Cross-talk Mechanism in Caco-2 Cells during Clostridium difficile Infection through Genetic-and-Epigenetic Interspecies Networks: Big Data Mining and Genome-Wide Identification. *Front. Immunol.* **8,** (2017).

39. Pothoulakis, C., Triadafilopoulos, G., Clark, M., Franzblau, C. & LaMont, J. T. Clostridium difficile cytotoxin inhibits protein synthesis in fibroblasts and intestinal mucosa. *Gastroenterology* **91,** 1147–1153 (1986).

40. Jiang, J.-H., Tong, J. & Gabriel, K. Hijacking Mitochondria: Bacterial Toxins that Modulate Mitochondrial Function. *IUBMB Life* **64,** 397–401 (2012).

41. Zhang, T. *et al.* Salmonella enterica Serovar Enteritidis Modulates Intestinal Epithelial miR-128 Levels to Decrease Macrophage Recruitment via Macrophage Colony-Stimulating Factor. *J. Infect. Dis.* **209,** 2000–2011 (2014).

42. Das, K., Garnica, O. & Dhandayuthapani, S. Modulation of Host miRNAs by Intracellular Bacterial Pathogens. *Front. Cell. Infect. Microbiol.* **6,** (2016).

43. Izar, B., Mannala, G. K., Mraheil, M. A., Chakraborty, T. & Hain, T. microRNA Response to Listeria monocytogenes Infection in Epithelial Cells. *Int. J. Mol. Sci.* **13,** 1173 (2012).

44. Vegh, P. *et al.* MicroRNA profiling of the bovine alveolar macrophage response to Mycobacterium bovis infection suggests pathogen survival is enhanced by microRNA regulation of endocytosis and lysosome trafficking. *Tuberculosis* **95,** 60–67

45. Siddle, K. J. *et al.* Bacterial Infection Drives the Expression Dynamics of microRNAs and Their isomiRs. *PLoS Genet.* **11,** (2015).

46. Yamamoto, Y. *et al.* Distinct roles of Rab3B and Rab13 in the polarized transport of apical, basolateral, and tight junctional membrane proteins to the plasma membrane. *Biochem. Biophys. Res. Commun.* **308,** 270–275 (2003).

47. Stein, M.-P., Müller, M. P. & Wandinger-Ness, A. Bacterial Pathogens Commandeer Rab GTPases to Establish Intracellular Niches. *Traffic Cph. Den.* **13,** 1565–1588 (2012).

48. Ye, F. *et al.* miR-200b as a prognostic factor in breast cancer targets multiple members of RAB family. *J. Transl. Med.* **12,** 17 (2014).

49. Wakabayashi, K. *et al.* Analysis of microRNA from archived formalin-fixed paraffin-embedded specimens of amyotrophic lateral sclerosis. *Acta Neuropathol. Commun.* **2,** (2014).

50. Obayashi, M. *et al.* microRNA-203 suppresses invasion and epithelial-mesenchymal transition induction via targeting NUAK1 in head and neck cancer. *Oncotarget* **7,** 8223–8239 (2016).

51. Evans, D. R. & Guy, H. I. Mammalian Pyrimidine Biosynthesis: Fresh Insights into an Ancient Pathway. *J. Biol. Chem.* **279,** 33035–33038 (2004).

52. Vogel-Scheel, J., Alpert, C., Engst, W., Loh, G. & Blaut, M. Requirement of Purine and Pyrimidine Synthesis for Colonization of the Mouse Intestine by Escherichia coli. *Appl. Environ. Microbiol.* **76,** 5181–5187 (2010).

53. Yang, H.-J., Bogomolnaya, L., McClelland, M. & Andrews-Polymenis, H. De novo pyrimidine synthesis is necessary for intestinal colonization of Salmonella Typhimurium in chicks. *PLOS ONE* **12,** e0183751 (2017).

54. Rupaimoole, R. & Slack, F. J. MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nat. Rev. Drug Discov.* **16,** 203 (2017).

Supplementary data

***Supplementary table 1**: All significantly enriched pathways obtained from mRNA and miRNA target genes enrichment studies are shown in this table. The 'NA' indicates that z-scores are not available for those pathways. The molecules column indicates genes that were found enriched in each pathway.*

RNAseq data enriched pathways				
Ingenuity Canonical Pathways	**-log(p-value)**	**Ratio**	**z-score**	**Molecules**
JAK/Stat Signalling	2.87	0.12	1.897	RELA, STAT6, AKT1, JUN, PTPN1, CDKN1A, SOCS2, STAT3, NFKB2, MAP2K1
iNOS Signalling	2.86	0.159	2.449	RELA, NFKBIA, JUN, IFNGR2, NFKB2, NOS2, IRF1
Glycine Biosynthesis III	2.83	1	NA	AGXT, AGXT2
Superpathway of Serine and Glycine Biosynthesis I	2.75	0.429	NA	PSPH, PHGDH, SHMT2
Phosphatidylethanolamine Biosynthesis II	2.56	0.375	NA	ETNK2, PIGF, PCYT2
Mitotic Roles of Polo-Like Kinase	2.55	0.127	-1.342	ANAPC4, PLK3, WEE1, PPP2R5B, ANAPC5, CDC7, CDC16, CHEK2
tRNA Charging	2.5	0.158	NA	TARS2, YARS2, YARS, CARS2, HARS2, QARS
Hereditary Breast Cancer Signalling	2.49	0.0922	NA	GADD45B, WEE1, SMARCE1, SMARCD2, RFC5, FANCL, HDAC5, FANCE, AKT1, RFC4, CDKN1A, UBC, CHEK2
Sirtuin Signalling Pathway	2.49	0.0745	-1.886	TIMM8A, ATG5, PFKFB3, RELA, GADD45B, TIMM9, NDRG1, NR1H3, STAT3, NFKB2, BPGM, XPA, VDAC2, ATG13, JUN, AKT1, GABARAPL1, IDH2, MAP1LC3B, MAPK7, NOS2
Fatty Acid β-oxidation III (Unsaturated, Odd Number)	2.36	0.667	NA	ECI2, ECI1
autophagy	2.34	0.13	NA	ATG13, ATG5, ULK1, MAP1LC3B, SQSTM1, ACE, VPS11
Role of PKR in Interferon Induction and Antiviral Response	2.33	0.146	NA	RELA, NFKBIA, AKT1, TNFRSF1A, NFKB2, IRF1
Acute Phase Response Signalling	2.22	0.0828	3.051	RELA, HPX, C4BPB, TNFRSF1A, IL6R, NFKB2, STAT3, JUN, AKT1, NFKBIA, SOCS2, SERPINE1, ELK1, MAP2K1
Methionine Degradation I (to Homocysteine)	2.18	0.2	NA	MAT1A, PRMT3, FTSJ1, MRM2

4-1BB Signalling in T Lymphocytes	2.14	0.156	2	RELA, NFKBIA, JUN, NFKB2, MAP2K1
Cysteine Biosynthesis III (mammalia)	2.03	0.182	NA	MAT1A, PRMT3, FTSJ1, MRM2
Pregnenolone Biosynthesis	2.01	0.25	NA	MICAL1, MICAL2, MICAL3
IL-17A Signalling in Fibroblasts	1.98	0.143	NA	RELA, NFKBIA, JUN, NFKB2, MMP1
IL-6 Signalling	1.96	0.0859	3.317	RELA, NFKBIA, AKT1, JUN, TNFRSF1A, IL6R, STAT3, NFKB2, ELK1, MAPKAPK2, MAP2K1
Role of BRCA1 in DNA Damage Response	1.92	0.1	0	FANCE, RFC4, CDKN1A, SMARCE1, SMARCD2, RFC5, CHEK2, FANCL
Role of JAK1, JAK2 and TYK2 in Interferon Signalling	1.9	0.167	NA	RELA, IFNGR2, STAT3, NFKB2
Superpathway of D-myo-inositol (1,4,5)-trisphosphate Metabolism	1.9	0.167	NA	ITPKC, IMPA2, ITPKA, PMPCA
HMGB1 Signalling	1.89	0.084	3.162	RELA, LIF, AKT1, JUN, RHOB, TNFRSF1A, IFNGR2, NFKB2, ELK1, SERPINE1, MAP2K1
Serine Biosynthesis	1.86	0.4	NA	PSPH, PHGDH
CD27 Signalling in Lymphocytes	1.83	0.115	1	RELA, NFKBIA, JUN, MAP3K8, NFKB2, MAP2K1
Phenylalanine Degradation IV (Mammalian, via Side Chain)	1.82	0.214	NA	AASDH, ALDH2, MAOB
April Mediated Signalling	1.78	0.128	1.342	RELA, NFKBIA, JUN, NFKB2, ELK1
IL-10 Signalling	1.77	0.101	NA	RELA, IL4R, NFKBIA, JUN, STAT3, NFKB2, ELK1
Role of IL-17A in Arthritis	1.77	0.101	NA	RELA, NFKBIA, NFKB2, MAPKAPK2, NOS2, MAP2K1, MMP1
Pancreatic Adenocarcinoma Signalling	1.75	0.0833	2.828	RELA, AKT1, CDKN1A, HBEGF, CDKN1B, STAT3, NFKB2, ELK1, MAP2K1, SIN3A
Neuregulin Signalling	1.74	0.093	1.414	AKT1, GRB7, HBEGF, ERRFI1, CDKN1B, ELK1, MAP2K1, AREG
Histidine Degradation VI	1.73	0.2	NA	MICAL1, MICAL2, MICAL3
TGF-β Signalling	1.72	0.092	0.378	JUN, SMURF2, TFE3, SERPINE1, MAP2K1, TGIF1, SMURF1, PMEPA1
B Cell Activating Factor Signalling	1.69	0.122	2	RELA, NFKBIA, JUN, NFKB2, ELK1
Chronic Myeloid Leukemia Signalling	1.69	0.0857	NA	RELA, AKT1, MECOM, CDKN1A, CDKN1B, NFKB2, MAP2K1, SIN3A, HDAC5
Cell Cycle Control of Chromosomal Replication	1.69	0.107	NA	CDC45, CDC7, POLA2, CDK9, CHEK2, ORC1
PI3K/AKT Signalling	1.68	0.0813	1.265	RELA, NFKBIA, AKT1, YWHAH, CDKN1A, PPP2R5B, MAP3K8, CDKN1B, NFKB2, MAP2K1
IL-15 Production	1.66	0.143	NA	RELA, PTK6, NFKB2, IRF1
Telomerase Signalling	1.62	0.0833	1.342	ELF3, AKT1, ETS2, CDKN1A, PPP2R5B, TPP1, MAP2K1, TERF1, HDAC5

Type II Diabetes Mellitus Signalling	1.61	0.0794	1.89	RELA, NFKBIA, AKT1, PRKAB2, TNFRSF1A, PKM, SOCS2, NFKB2, ADIPOR2, SMPD2
PI3K Signalling in B Lymphocytes	1.61	0.0794	3.162	RELA, IL4R, ATF3, NFKBIA, AKT1, JUN, CARD10, NFKB2, ELK1, MAP2K1
MIF Regulation of Innate Immunity	1.61	0.116	2.236	RELA, NFKBIA, JUN, NFKB2, NOS2
TNFR2 Signalling	1.61	0.138	2	RELA, NFKBIA, JUN, NFKB2
Death Receptor Signalling	1.59	0.087	1.414	TNFRSF21, RELA, NFKBIA, TNFRSF1A, TNFRSF10B, DFFB, NFKB2, TNFRSF10A
PPAR Signalling	1.59	0.087	-2.121	RELA, NFKBIA, JUN, PDGFA, TNFRSF1A, NR1H3, NFKB2, MAP2K1
Ceramide Signalling	1.56	0.086	1.89	RELA, AKT1, JUN, TNFRSF1A, PPP2R5B, NFKB2, MAP2K1, SMPD2
RAR Activation	1.52	0.0695	NA	RELA, AKT1, JUN, ADCY5, RDH14, PNRC1, RDH16, SMARCE1, SMARCD2, NFKB2, MAPKAPK2, MAP2K1, MMP1
Prostate Cancer Signalling	1.51	0.0842	NA	RELA, NFKBIA, AKT1, CDKN1A, CDKN1B, NFKB2, MAP2K1, SIN3A
CD40 Signalling	1.51	0.0897	1.89	RELA, NFKBIA, JUN, STAT3, NFKB2, MAPKAPK2, MAP2K1
Sumoylation Pathway	1.49	0.0833	0.816	NFKBIA, JUN, RFC4, RHOB, RCOR1, NFKB2, RFC5, ISG20
HGF Signalling	1.49	0.0789	1.134	ELF3, AKT1, JUN, ETS2, CDKN1A, MAP3K8, STAT3, ELK1, MAP2K1
Ubiquinol-10 Biosynthesis (Eukaryotic)	1.45	0.158	NA	MICAL1, MICAL2, MICAL3
1D-myo-inositol Hexakisphosphate Biosynthesis II (Mammalian)	1.45	0.158	NA	ITPKC, ITPKA, PMPCA
D-myo-inositol (1,3,4)-trisphosphate Biosynthesis	1.45	0.158	NA	ITPKC, ITPKA, PMPCA
Gα12/13 Signalling	1.44	0.0741	2.333	RELA, NFKBIA, AKT1, JUN, F2RL1, MEF2C, MAPK7, NFKB2, ELK1, MAP2K1
Erythropoietin Signalling	1.43	0.0864	NA	RELA, NFKBIA, AKT1, JUN, NFKB2, ELK1, MAP2K1
TNFR1 Signalling	1.43	0.104	2.236	RELA, NFKBIA, JUN, TNFRSF1A, NFKB2
Production of Nitric Oxide and Reactive Oxygen Species in Macrophages	1.43	0.0674	1.941	RELA, NFKBIA, AKT1, JUN, RHOB, TNFRSF1A, PPP2R5B, IFNGR2, MAP3K8, NFKB2, NOS2, MAP2K1, IRF1
Thio-molybdenum Cofactor Biosynthesis	1.41	1	NA	MOCOS
Asparagine Biosynthesis I	1.41	1	NA	ASNS
D-mannose Degradation	1.41	1	NA	MPI

RANK Signalling in Osteoclasts	1.4	0.08	1.89	RELA, NFKBIA, AKT1, JUN, MAP3K8, NFKB2, ELK1, MAP2K1
Tryptophan Degradation X (Mammalian, via Tryptamine)	1.4	0.15	NA	ALDH2, MAOB, RDH14
Oncostatin M Signalling	1.39	0.118	2	STAT3, ELK1, MAP2K1, MMP1
ERK5 Signalling	1.38	0.0909	0.816	LIF, AKT1, YWHAH, MEF2C, MAP3K8, MAPK7
Heme Biosynthesis II	1.35	0.222	NA	PPOX, UROS
Noradrenaline and Adrenaline Degradation	1.35	0.114	NA	HSD17B10, ALDH2, MAOB, RDH14
Superpathway of Methionine Degradation	1.35	0.114	NA	MAT1A, PRMT3, FTSJ1, MRM2
IL-17 Signalling	1.34	0.0824	NA	RELA, AKT1, JUN, ELK1, MAPKAPK2, NOS2, MAP2K1

miRNA targeted mRNAs enriched pathways

Ingenuity Canonical Pathways	-log(p-value)	Ratio	z-score	Molecules
Superpathway of Serine and Glycine Biosynthesis I	3.33	0.429	NA	PSPH, PHGDH, SHMT2
Phosphatidylethanolamine Biosynthesis II	3.14	0.375	NA	ETNK2, PIGF, PCYT2
Neuregulin Signalling	2.31	0.0814	1.134	AKT1, GRB7, HBEGF, ERRFI1, CDKN1B, ELK1, MAP2K1
Serine Biosynthesis	2.25	0.4	NA	PSPH, PHGDH
Sirtuin Signalling Pathway	2.01	0.0495	-0.302	TIMM8A, ATG5, PFKFB3, GADD45B, TIMM9, NDRG1, STAT3, BPGM, VDAC2, ATG13, AKT1, GABARAPL1, MAP1LC3B, NOS2
autophagy	2	0.0926	NA	ATG13, ATG5, ULK1, MAP1LC3B, SQSTM1
Methionine Degradation I (to Homocysteine)	1.92	0.15	NA	MAT1A, PRMT3, MRM2
Cysteine Biosynthesis III (mammalia)	1.81	0.136	NA	MAT1A, PRMT3, MRM2
JAK/Stat Signalling	1.81	0.0723	1.633	STAT6, AKT1, PTPN1, SOCS2, STAT3, MAP2K1
Heme Biosynthesis II	1.72	0.222	NA	PPOX, UROS
TGF-β Signalling	1.71	0.069	0	SMURF2, SERPINE1, MAP2K1, TGIF1, SMURF1, PMEPA1
iNOS Signalling	1.66	0.0909	NA	NFKBIA, IFNGR2, NOS2, IRF1
Hereditary Breast Cancer Signalling	1.66	0.0567	NA	FANCE, GADD45B, AKT1, WEE1, SMARCE1, SMARCD2, RFC5, CHEK2
ERK5 Signalling	1.66	0.0758	0.447	LIF, AKT1, YWHAH, MEF2C, MAP3K8
Acute Phase Response Signalling	1.64	0.0533	2.828	NFKBIA, AKT1, C4BPB, IL6R, SOCS2, STAT3, ELK1, SERPINE1, MAP2K1
Asparagine Biosynthesis I	1.61	1	NA	ASNS

D-mannose Degradation	1.61	1	NA	MPI
ErbB2-ErbB3 Signalling	1.61	0.0735	2	AKT1, CDKN1B, STAT3, ELK1, MAP2K1
Sertoli Cell-Sertoli Cell Junction Signalling	1.59	0.052	NA	F11R, AKT1, CLDN4, MYO7A, MAP3K8, ELK1, NOS2, MAP2K1, TUBB2B
T Helper Cell Differentiation	1.58	0.0725	NA	STAT6, IL4R, IL6R, IFNGR2, STAT3
Pancreatic Adenocarcinoma Signalling	1.56	0.0583	2.236	AKT1, HBEGF, CDKN1B, STAT3, ELK1, MAP2K1, SIN3A
Cleavage and Polyadenylation of Pre-mRNA	1.48	0.167	NA	CSTF3, WDR33
Pregnenolone Biosynthesis	1.48	0.167	NA	MICAL2, MICAL3
PI3K Signalling in B Lymphocytes	1.46	0.0556	2.646	IL4R, ATF3, NFKBIA, AKT1, CARD10, ELK1, MAP2K1
IL-6 Signalling	1.43	0.0547	2.646	NFKBIA, AKT1, IL6R, STAT3, ELK1, MAPKAPK2, MAP2K1
Role of JAK2 in Hormone-like Cytokine Signalling	1.38	0.0938	NA	PTPN1, SOCS2, STAT3
IGF-1 Signalling	1.35	0.0566	2.236	AKT1, YWHAH, SOCS2, STAT3, ELK1, MAP2K1
Role of BRCA1 in DNA Damage Response	1.34	0.0625	NA	FANCE, SMARCE1, SMARCD2, RFC5, CHEK2

Identification of food compounds attenuating the cytopathic effects of *Clostridium difficile* toxins using transcriptomics datasets

Prashanna Balaji Venkatasubramanian[1], Els Oosterink[1], Monic Tomassen[1], Maria Suarez-Diez[2], Edoardo Saccenti[2], Jurriaan Mes[1], Nicole de Wit[1]*

This work has been submitted to BMC Genomics

Abstract

Clostridium difficile is an anaerobic, spore-forming bacterium that can cause diarrhoea and fulminant colitis. *C. difficile* toxins A and B play a key role in the pathogenesis. Besides effects in the colon, recent studies indicate prevalence of *C. difficile* infection (CDI) in the small intestine. In this paper, we explored the impact of *C. difficile* toxins on the small intestine using an *in vitro* approach and used systems biology techniques along with large data integration to identify food compounds that can reduce their cytopathic impact. Differentiated Caco-2 cells were exposed to *C. difficile* toxins and the transcriptomic changes were studied. For the identification of foods with potential counteracting effects, these transcriptomic data were combined with a data compendium containing microarrays from Caco-2 cells exposed to various food compounds to conduct principle component analysis (PCA). Food candidates were selected and further examined for their counteracting effect on toxin-induced disruption of cell integrity and translocation of toxins. Based on PCA we hypothesized that blackcurrant, strawberry and yellow onions would attenuate the cytotoxic effects of *C. difficile* toxins and verified these predictions *in vitro*. The verification was done using trans-epithelial electrical resistance (TEER) measurements and translocation of the toxins. These results might lead to novel strategies for treating *C. difficile* infection in patients receiving antibiotic therapy.

Introduction

Clostridium difficile infection (CDI) affects hundreds of thousands of people each year and causes a broad spectrum of symptoms that range from watery diarrhoea to fulminant colitis [1–3]. *C. difficile* spores are able to endure extreme conditions and can to colonize the intestine, especially when normal microbiota has been disturbed by the antibiotics [1,2,4]. Most *C. difficile* strains produce two major exotoxins, toxin A and toxin B (ToxA and ToxB, also called TcdA and TcdB, respectively) which play a major role in pathogenesis of CDI [4]. ToxA and ToxB share 44% sequence identity between each other and are found to have overlapping enzymatic activities [4]. Studies indicate that these exotoxins initially bind to the cell surface receptors and this is followed by the toxins being internalized into the host cells. On internalization, the toxins target the Ras superfamily of small GTPases for modification via glycosylation, leading to irreversible inactivation of vital signalling pathways in the cell [4]. Low doses of ToxA have been found to alter cell polarity by inducing plasma membrane components redistribution in 3-D and 2-D intestinal epithelial cell model systems [5]. Earlier study on the effects of the toxins on Caco-2 monolayer indicates that ToxA assist translocation of the toxins more than ToxB alone [6]. On the other hand, the cytotoxicity of ToxB has been found to be 100 - 10,000 times more potent than ToxA for several cell types [4]. Small differences found for pathway mechanism of the toxins might be explained by the use of different animal models, different toxin dosages or unidentified genetic variations in *C. difficile* strains [4].

For many years, CDI studies have focused on CDI in large intestine as *C. difficile* was initially thought to solely colonize the large intestine [2]. However, recent studies indicate that *C. difficile* infections also affect small intestine and are on a rise [7]. There is also increasing incidence of CDI in patients suffering from Intestinal Bowel Disease (IBD) who have undergone surgical procedures [8]. IBD is an umbrella term for intestinal diseases that consists of two primary types: Ulcerative colitis and Crohn's disease. Ulcerative colitis is known to affect only the colon while Crohn's disease affects both colon as well as the terminal ends of the small intestine [9]. In the

light of these evidences, understanding the impact of *C. difficile* on small intestinal epithelia is of need and importance. In addition, new treatments and preventive strategies for CDI are needed [10] [11] as there has been a rise in antibiotic resistant strains [12]. A food based strategy which can easily be applied on a daily basis could be a way forward and sometimes support medicine based therapeutic treatment, similar to the method applied for *Helicobacter pylori* [13].

The aim of this research was to identify food compounds that might prevent/reduce the toxic effects of the *C. difficile* toxins on intestinal epithelial cells using Caco-2 cell monolayers cultured for 21 days as model system. An integrative biology approach was used to analyse the transcriptional response of Caco-2 cell monolayers to toxin exposure and existing transcriptomics data for Caco-2 cells exposed to food in order to select candidate food substances to attenuate cytopathic effect of the *C. difficile* toxins. Finally, the predicted foods were tested for their capacity to attenuate effect of *C. difficile* toxins on integrity of Caco-2 intestinal cell monolayers *in vitro*.

Methods

Caco-2 cell culture

ATCC derived Caco-2 cells were cultured in Dulbecco's modified Eagle's medium (DMEM; Gibco-Invitrogen, Bleiswijk, The Netherlands) with 4.5 g/L glucose, 0.58 g/L glutamine, no pyruvate, supplemented with 10% heat inactivated FBS (Hyclone Perbio, Etten-Leur, The Netherlands) and used with passage numbers between 30 and 40. For transwell assays, 330,000 cells were seeded on ThinCert transwells with 33.6 mm^2 membranes and 0.4 µm pores in 24-well suspension culture plates. Cells were grown for 21 days at 5% CO_2 and 37°C and apical (150 µL) and basolateral (700 µL) medium were replaced three times per week.

In vitro digestion

A total of 15 grams of white onion, yellow onion, blackcurrants (Ben Finlay cultivar) and strawberries (Sabrina cultivar) were mixed with an equal amount of 140 mM NaCl / 5 mM KCl and mashed with an ultra torax. The *in vitro* digestion protocol was mainly based on the paper of Vreeburg *et al*. with some slight modifications and in line with the standardized protocol as proposed by Minekus *et al*. [14,15]. In more detail, 20 g of sample were transferred into a 50 mL tube, the pH was adjusted to 2 with HCl and 0.667 mL of 40 g/L porcine pepsin in 0.1 M HCl was added. The samples were then incubated for 30 min at 37°C. Subsequently, 1 M $NaHCO_3$ was added to raise the pH to at least 5.8, followed by 0.95 mL of 4 g/L porcine pancreatin in 0.1 M $NaHCO_3$ and 0.5 mL of a mixture of sodium taurocholate and sodium glycodeoxycholate (176 mM of each) in 0.1 M $NaHCO_3$. The pH of the sample was adjusted to 6.5 with 1 M $NaHCO_3$, flushed with nitrogen and the sample was subsequently incubated for 60 min at 37°C. After this incubation, the pH of the sample was adjusted to 7.5 with 1 M $NaHCO_3$ and the weight of the sample was adjusted to 30 g with 140 mM NaCl / 5 mM KCl. Samples were centrifuged for 45 min at 3023 × g at 4°C. The supernatant was taken, flushed with nitrogen and stored at −80°C until further use. For preparing the *in vitro* digestion control, 140 mM NaCl / 5 mM KCl was used to replace food compounds.

Exposure of differentiated Caco-2 cells to toxins and various food compounds

All *C. difficile* toxins were derived from List Biological Laboratories, Inc. (Campbell, California, USA; Toxin A (#152), Toxin B (#155), Toxin A Toxoid (#153), Toxin B Toxoid (#154)) and dilutions were made in DMEM/FBS. For toxin incubation, 21 day differentiated Caco-2 cells were exposed to 0.25µg/ml toxin A (ToxA) or 0.25µg/ml toxin B (ToxB) or 0.25µg/ml toxin A + 0.25µg/ml toxin B (ToxAB). Toxin A and B toxoids (0.25µg/ml), meaning formaldehyde inactivated toxins, were also included in the exposure experiments as a control for the cytotoxic

effects of the *C. difficile* toxins. One day before the exposure experiments, medium was refreshed and at the day of exposure, medium was removed from the apical and basal compartments and toxin samples were added to the apical compartment while fresh DMEM/FBS medium was added to the basal compartment. To monitor the integrity of the Caco-2 monolayer, transepithelial electrical resistance (TEER) was measured at 37ºC using a MilliCell-ERS Ω meter (Millipore, Molsheim, France). In following experiments, Caco-2 cells were co-incubated with *C. difficile* toxins (0.25µg/ml toxin A or 0.25µg/ml toxin B or 0.25µg/ml toxin A + 0.25µg/ml toxin B) and various (digested) food compounds. Therefore, digested white onion (WOd), yellow onion (YOd), blackcurrant (Ben Finlay), Strawberry (Sabrina) and digestion control samples (control digest) were diluted 1:4 in DMEM/FBS medium. Undigested galacto-oligosaccharides (GOS; Friesland Campina, Amersfoort, the Netherlands) were added to the Caco-2 cells in a concentration of 0.5mg/ml and DMEM/FBS medium was used as control for this exposure. One day before the exposure experiments, medium was refreshed and at the day of exposure, medium was removed from the apical and basal compartments and diluted samples (toxins + food compounds) were added to the apical compartment while fresh DMEM/FBS medium was added to the basal compartment. TEER was determined before and at 0, 1, 3, 4, 5, 6 and 24 h after addition of samples to check integrity of the intestinal monolayer. In order to be able to reverse the toxic effect of *C. difficile* toxins by food compounds, we focused on a ~35% drop in TEER after exposure to toxins, since a TEER drop below 35% seems to damage the cells irreversibly. At that moment (~35% drop in TEER by one of the toxin treatments), the Caco-2 cells were lysed with 300 µL TRIzol (Invitrogen, Life Technologies, Bleiswijk, Netherlands) and the triplicates in each experiment were pooled for RNA isolation. These RNA samples were used for microarray analysis. Toxin translocation was measured using the *Clostridium difficile* toxin A or B ELISA kit (Catalog No. ABIN1098189, Antibodies-online GmbH, Aachen, Germany). In each experiment, the digested food and control samples were exposed to Caco-2 cells in triplicate and three independent exposure experiments were performed on different days.

RNA isolation

For Caco-2 cells, cell lysates of triplicates per experiment were pooled and total RNA was extracted using the QIAshredder and RNeasy Mini kits (Qiagen, Venlo, The Netherlands) following the manufacturers' protocols [14]. Briefly, TriZol (Invitrogen) extraction from ThinCerts transwells was performed with 300 µL TriZol and followed by DNase-I treatment (Sigma-Aldrich, Zwijndrecht, the Netherlands) and RNeasy clean-up (Qiagen, Venlo, The Netherlands). Quality and amount of RNA was evaluated by UV spectrometry (260 and 280 nm wavelength) on the Nanodrop spectrophotometer (Thermo Scientific, Wilmington, DE, USA).

Microarray Experiments

RNAs of each independent Caco-2 experiment with toxin exposures (n=3 per treatment) were hybridized to Affymetrix© Human Gene 1.1 ST according to standard Affymetrix© protocols. Quality control of the datasets was performed using Bioconductor packages [16,17] integrated in an on-line pipeline [17]. Microarray data have been submitted to GEO and can be found under accession code GSE100541.

Data preprocessing, Differential Expression and Pathway Analysis

Array data were normalized using the Robust Multiarray Average (RMA) M-estimator method [18,19], probe sets were defined according to Dai *et al*. [20]. To exclude interference of non-expressed or very lowly expressed genes, Universal exPression Code (UPC) [21] was computed. Genes with UPC value lower than 0.5 in none of the replicates were considered to be non-expressed and removed from further analysis. To identify differential gene expression induced by the toxins, pair-wise comparison analyses were performed (toxins versus DMEM control) and genes with p-value <0.01 were considered to be differentially expressed. Pathway enrichment analysis was performed using Ingenuity Pathway Analysis (IPA) (Ingenuity© Systems, www.ingenuity.com). Significance values for enrichment were calculated using Fisher's exact test right-tailed.

Data Compendium

Data compendium comprised of microarray data collected from previous food exposure studies reported in Venkatasubramanian *et al* [22]. The dataset was restricted to experiments where the exposure was food and food related substances like white and yellow onion, sulforaphane (present in broccoli) and probiotics like *L. casei* and *B. breve* [22]. In total, 73 experiments conducted over 15 batches were included in the data compendium and most experiments had been performed in triplicates. Supplementary Table T1 describes the dataset used in the data compendium and their GEO accession number. The data was RMA normalized in batches of experiments using *affy* and *Oligo* packages [23,24] of Bioconductor [16] in R (version 2.3.2) and the differential expression values were calculated using *Limma* package [25,26].

Principal Component Analysis

Principal Component Analysis (PCA) is a statistical technique which is used to reduce dimensionality while conserving significantly large variability in the data [27,28]. This is achieved by identifying new variables (principal components) that are linear combinations of the original variables. Principal components are orthogonal and each component capture decreasing amounts of information. Basically, PCA results in the projection of the original data on a lower dimensional space (PCA space): observation that are closer to each one in the PCA space are expected to show similar characteristics, in this case gene expression levels. Based on this, PCA was used to select beneficial compounds, under the rationale that foods most distant from toxins and toxoids would be the most efficient in countering the effects of toxins and thus prove to be beneficial against *C. difficile* toxin-induced disruption of intestinal integrity. Meanwhile, it was expected that food compounds closer to toxoids in the PCA space have least beneficial effects.

PCA was performed on the differential expression values of selected genes (*e.g.* genes from Sumoylation pathway, as derived from IPA or top n-number of regulated genes) from the

compendium along with the differential expression values for toxins and toxoids and plots were generated (Figure 4). PCA was performed in R using '*prcomp*' (R version 3.2.3). Data dimensionality was assessed using a Tracy-Widom test at a 0.01 confidence level [29,30]. The optimal number of components was 11 for PCA with sumoylation genes and 17 for PCA with top 50 genes. In all cases the selected number of components accounted for >99% of variance explained. For visualization purposes only the first two components are shown.

Other statistical Analysis

TEER measurements and toxin translocation measurements were conducted in biological triplicates. The data was visualized as mean +/- standard deviation and was statistically tested for significant differences between control and treatment using a paired student's t-test. Results were considered statistically significant if their p-values were below 0.05.

Results

Gene expression profiling of Caco-2 cells exposed to *C. difficile* toxins.

Caco-2 cells, differentiated for 21 days into a small intestinal phenotype, were exposed to *C. difficile* ToxA, ToxB, ToxA and ToxB together (ToxAB) or the formaldehyde inactivated toxoids A and B and controls. Transepithelial electrical resistance (TEER) was used to measure integrity of the Caco-2 monolayers. On average in the triplicate experiments, after 4.1h (± 0.6h) incubation, a ~35% drop in TEER was reached on exposure to ToxA and/or ToxAB and RNA of Caco-2 cells was harvested for microarray analyses. Within this time frame, neither ToxB nor the toxoids induced a drop in TEER and so did not affect the monolayer integrity (Figure 1). Differential gene expression was assessed by comparing the *C. difficile* toxin treatments versus the DMEM control. To further select the genes that are most likely involved in the (cyto)toxic effects of the toxins, we excluded genes that were also differentially expressed by the toxoids compared to DMEM control (n=562 genes). This resulted in the differential expression of 1407

genes for ToxA, 215 genes for ToxB and 1013 genes for ToxA and B. Figure 2a shows the top 10 up and down regulated genes for each toxin treatment, based on fold change (FC) (p-value < 0.01), and combined in a heatmap. For the upregulated genes there is large overlap between the ToxA and ToxAB, but ToxB commonly induced lower FC than ToxA and ToxAB (Figure 2b and Figure 2c). The most upregulated gene for all toxins was RHOB (Ras homolog gene family, member B), which is a member of the Rho GTP-binding protein family. Among the downregulated genes, larger variation of genes was observed between the toxins. For ToxA and ToxB individually, ANKRD1 (Ankyrin Repeat Domain 1) was the most downregulated gene, whereas for the combination of toxin A and B (ToxAB), CCL15 (Chemokine (C-C motif) ligand 15) showed the strongest downregulation (Figure 2a).

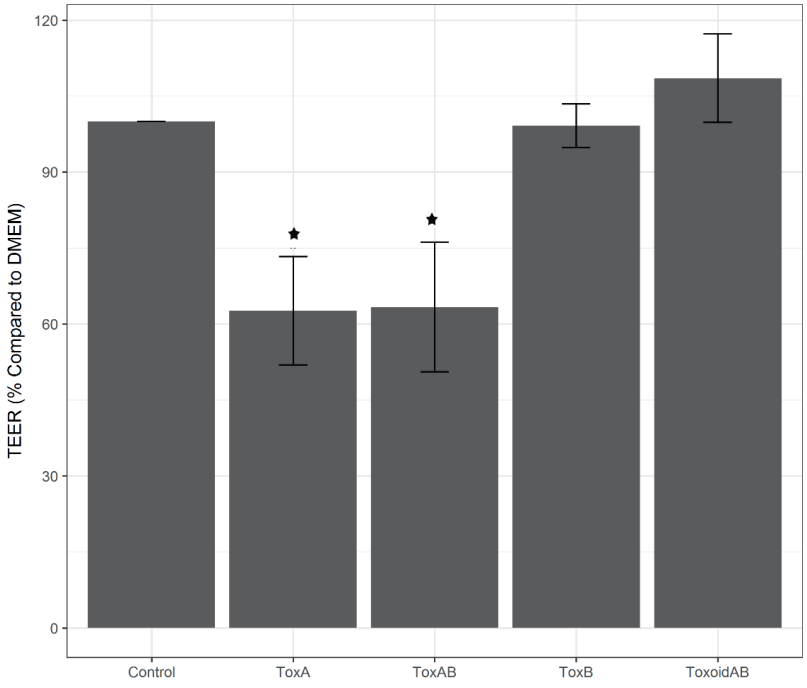


***Figure 1**: TEER drop measured after exposure of Caco-2 cells to ToxA, ToxA and ToxB (ToxAB), ToxB and ToxoidAB compared to DMEM control which was set to 100%. '*' indicates* significant results (p-value < 0.05).

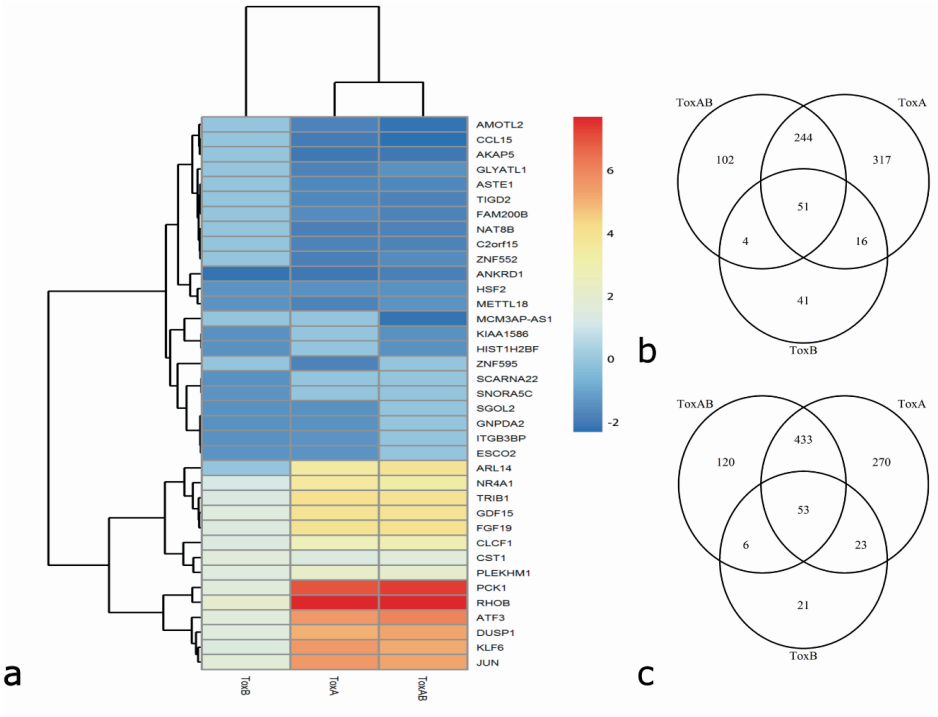


***Figure 2**: a) Heatmap of top 10 up and 10 down regulated genes from experiments where Caco-2 cells were exposed to ToxA, ToxB and ToxAB (ToxA + ToxB) (p-value < 0.01), combined for all three toxin exposures. b & c) Differential expressed genes induced by ToxA, ToxB and ToxAB* b) down regulated, c) up regulated.

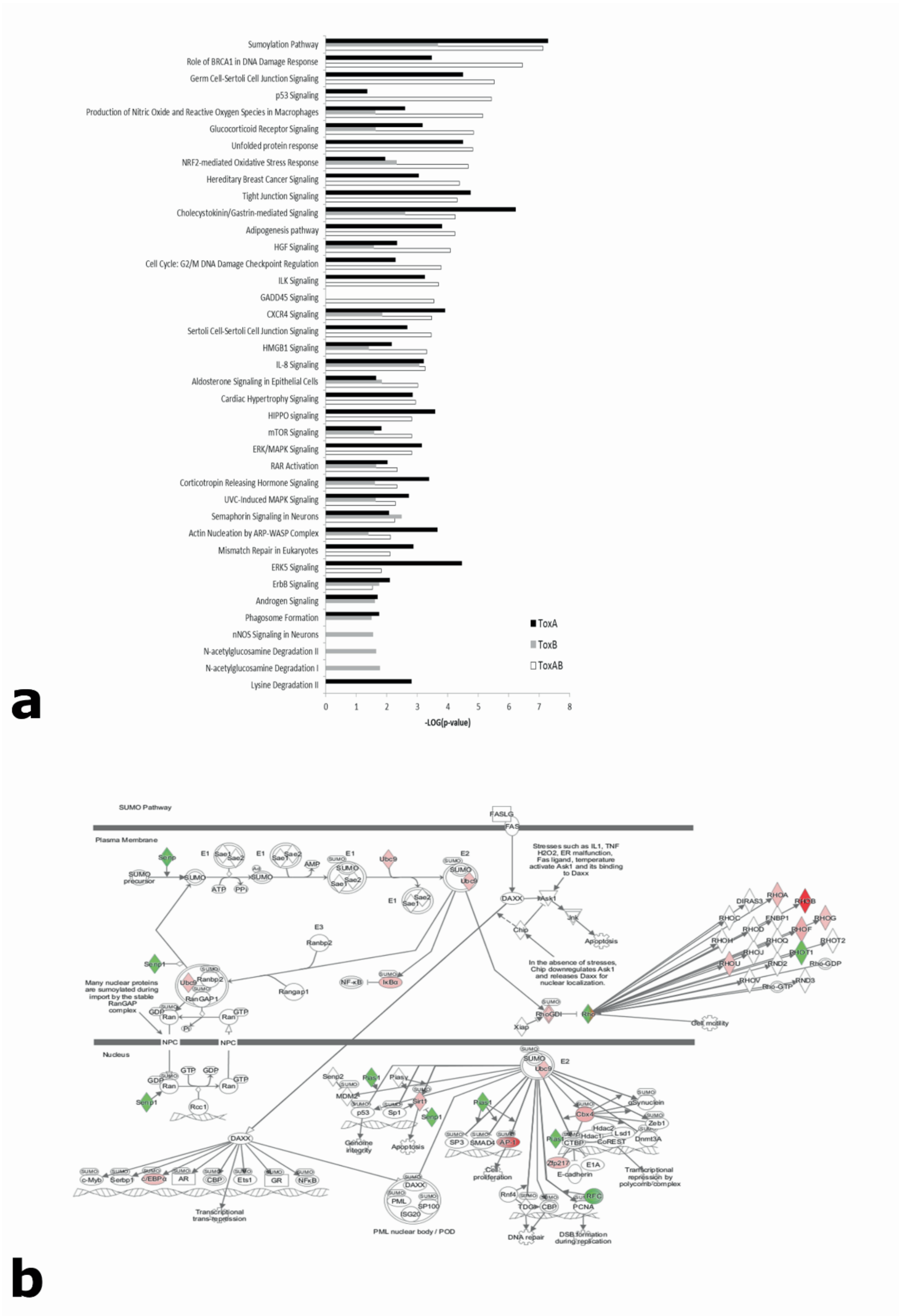

Sumoylation Pathway
Role of BRCA1 in DNA Damage Response
Germ Cell-Sertoli Cell Junction Signaling
p53 Signaling
Production of Nitric Oxide and Reactive Oxygen Species in Macrophages
Glucocorticoid Receptor Signaling
Unfolded protein response
NRF2-mediated Oxidative Stress Response
Hereditary Breast Cancer Signaling
Tight Junction Signaling
Cholecystokinin/Gastrin-mediated Signaling
Adipogenesis pathway
HGF Signaling
Cell Cycle: G2/M DNA Damage Checkpoint Regulation
ILK Signaling
GADD45 Signaling
CXCR4 Signaling
Sertoli Cell-Sertoli Cell Junction Signaling
HMGB1 Signaling
IL-8 Signaling
Aldosterone Signaling in Epithelial Cells
Cardiac Hypertrophy Signaling
HIPPO signaling
mTOR Signaling
ERK/MAPK Signaling
RAR Activation
Corticotropin Releasing Hormone Signaling
UVC-Induced MAPK Signaling
Semaphorin Signaling in Neurons
Actin Nucleation by ARP-WASP Complex
Mismatch Repair in Eukaryotes
ERK5 Signaling
ErbB Signaling
Androgen Signaling
Phagosome Formation
nNOS Signaling in Neurons
N-acetylglucosamine Degradation II
N-acetylglucosamine Degradation I
Lysine Degradation II
ToxA
ToxB
ToxAB
0 1 2 3 4 5 6 7 8
-LOG(p-value)
a
SUMO Pathway
Plasma Membrane
Nucleus
Stresses such as IL1, TNF H2O2, ER malfunction, Fas ligand, temperature activate Ask1 and its binding to Daxx
In the absence of stresses, Chip downregulates Ask1 and releases Daxx for nuclear localization.
Many nuclear proteins are sumoylated during import by the stable RanGAP complex
Transcriptional trans-repression
PML nuclear body / POD
Transcriptional repression by polycomb complex
DSB formation during replication
DNA repair
Cell proliferation
Genome integrity
Apoptosis
Cell motility
b

***Figure 3**: a) Pathways enriched in Caco-2 based on differentially expressed genes induced by ToxA, ToxB and ToxAB. Pathways have been ordered according to p-values of enrichment analysis. Only top 20 pathways are shown (combined for all three toxin exposures). b) Differentially expressed genes by ToxAB in Sumoylation pathway as represented by IPA are shown. Red indicates up regulated genes and green indicate down regulated genes. Darker shades indicate larger fold changes.*

Next, differentially expressed genes (p-value < 0.01) were selected for pathway enrichment analysis. The (canonical) pathways were ranked based on significance of the enrichment score and the top 20 pathways for all toxins are presented in Figure 3a. The sumoylation pathway showed highest significance in all three perturbations (ToxA, ToxB, ToxAB) and therefore we considered it as an important pathway in Caco-2 cells in response to *C. difficile* toxin exposure. Sumoylation is a posttranslational modification and has a role in various cellular processes, such as stress and injury. Figure 3b shows the gene expression changes in the sumoylation pathway induced by ToxAB in more detail. Among the genes from this pathway that were differentially expressed, it was found that the Rho GTP-binding protein family genes were substantially affected by the toxins. As *C. difficile* toxins are known to affect Rho GTPases this is not surprising [31,32]. Additionally, p53 signalling (apoptosis-related), cell junction and tight junction signalling pathways are observed to be significantly affected. These pathways might be linked to the drop in TEER and thus the monolayer integrity that we found in our exposure experiments. Pathways related to cytokine signalling like, IL-8 (interleukin-8) signalling and ILK (integrin linked kinase) signalling are other interesting signalling pathways that are activated, indicative for a potential effect of toxins on immune-related responses.

Identification of food compounds which might attenuate *C. difficile* toxin-induced disruption of intestinal integrity.

Principle Component Analysis (PCA) was used to find patterns and identify candidate food compounds with remedial effects against the impact of *C. difficile* toxins. For this, only the gene

expression effects of ToxA and B combined (so ToxAB) were used in this PCA strategy since the *in vivo* mechanism of action of CDI involves both ToxA and B affecting simultaneously. We combined the data obtained in this study with a compendium of Caco-2 specific gene expression data [22]. The dataset compendium contains differential gene expression data from food exposure studies performed in our lab and from public repositories. The compendium contains 73 experiments performed over 15 batches. The compendium was used together with the expression data from *C. difficile* toxins and toxoid studies here presented, to analyse similarities and/or contradictions between gene expression patterns that might be indicative for a remedial effect of food compounds on CDI.

In first instance, all differentially expressed genes induced by ToxAB and belonging to sumoylation pathway (17 genes) (Figure 3b), were used to find beneficial food substances using PCA. Sumoylation pathway was chosen since this pathway was found to be most significantly enriched among the differentially expressed genes in Caco-2 cells upon exposure to the *C. difficile* toxins.

The first two principal components (PCs) of the PCA are shown in Figure 4a. The two PCs account for 70% of the total variance in the data and the first 11 PCs explain 99% of the variability. In Figure 4a, we observe that ToxA and ToxAB are distinctly separated from toxoids based on sumoylation pathway genes. On the contrary, ToxB, which also showed less effect on Caco-2 cell integrity than ToxA and ToxAB, was in proximity to the toxoids in the PCA space. We hypothesized that food compounds closer to toxoids in the PCA space would probably have no or less beneficial effects while foods most distant from toxoids and toxins would be most efficient in countering the effects of toxins and thus prove to be beneficial for *C. difficile* toxin-induced disruption of intestinal integrity. Based on this hypothesis, blackcurrant (Ben Finlay) might potentially be the most counteracting and fibres (like GOS) least effective against toxin effects or at least against the sumoylation pathway activation. Compounds that are found to be in between toxoids and the toxins (ToxA and ToxAB), such as strawberry (Sabrina) and yellow onion, might be intermediately effective against the toxin induced TEER drop. White onion,

which is found closer to the toxoids in the PCA, was expected to be less effective in attenuating *C. difficile* cytotoxicity.

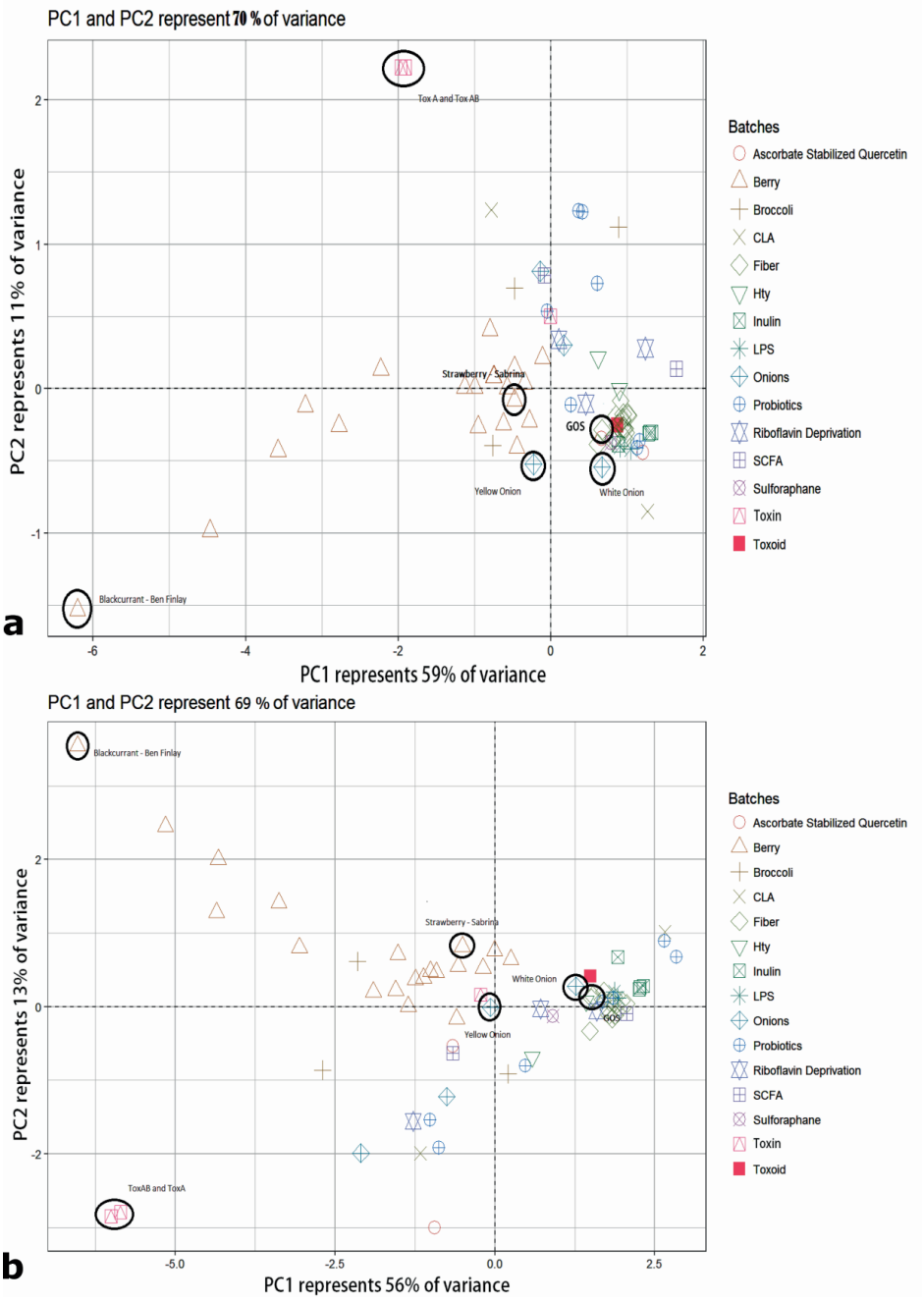


***Figure 4**: PCA plots based on a compendium of 15 batches of microarray data collected from 73 food related interventions on Caco-2 a) 17 genes in the sumoylation pathway were included in PCA b) Top 25 up and top 25 down regulated genes were considered for PCA while only 28 genes were used in PCA as not all 50 genes had differential expression values in all experiments in the compendium.*

PCA was also performed using top 25 up and top 25 down regulated genes induced by ToxAB in Caco-2 cells (Figure 4b). Differential expression values for 28 of the 50 genes (22 genes were not found in all 82 experiments) from other exposure experiments were obtained from the data compendium and used for PCA. In this case, the first two PCs explain 69 % of the total variability (Figure 4b). 99% of total variability in the data was explained by 17 PCs. Inspection of this PCA plot (Figure 4b) shows similar results as the analysis of the PCA plot based on genes within the sumoylation pathway (Figure 4a), thus reinforcing our hypothesis.

In addition to using sumoylation and top 50 up/down regulated genes, PCA was also performed using the full 1013 genes that were differentially expressed by ToxAB (Supplementary Figure S1). However, here two principal components did not show a clear separation of toxins, toxoids and other food compounds which could lead to predictions of beneficial food substances. In this case, the first two PCs explain 58% of the variance. The lack of clear separation can be due to the large number of considered genes. Many of them may represent a general response to stress conditions and as such they provide no specific response to the tested toxins.

Validation of beneficial effects by candidate food compounds on *C. difficile* toxin-induced intestinal integrity.

Food compounds that were expected to be the most effective, moderately effective and least effective against the toxin-induced cytotoxicity based on PCA, were chosen for experimental verification. These were blackcurrant (Ben Finlay cultivar, most effective), strawberry (Sabrina cultivar, moderately effective), yellow onion (moderately effective), white onion and GOS (least

effective). These food substances were verified for their beneficial effects by co-exposure of Caco-2 cells to the food substances and the *C. difficile* toxins.

The TEER values which represent the intestinal integrity of the Caco-2 monolayer are shown in Figure 5. The TEER values show that blackcurrant mostly diminished the drastic drop in TEER after 24 hours exposure to the toxins. Assuming that the TEER value of DMEM exposed Caco-2 cells (control) is 100%, then ToxA and ToxAB induced a TEER drop of 69% and 74% respectively, after 24 hours of exposure. However, in the presence of blackcurrant a drop of only 19% and 16% of control TEER values was observed to be induced by ToxA and ToxAB respectively (Figure 5). This was a significant counteractive effect of blackcurrant on toxin-disrupted TEER (p-value < 0.05). Next to this significant reduction in TEER after 24 hours exposure, blackcurrant also induced a substantial delay in TEER drop after 4, 5 and 6 hours of toxin exposure, even for ToxB.

Strawberry and yellow onion especially delayed the drop in TEER induced by the ToxA and ToxAB, as at time point 6 hours after exposure, significant differences in TEER (p-value < 0.05) could be detected. After 24 hours of exposure, the effects of strawberry and yellow onion were less strong than what was seen for blackcurrant, but still a significant beneficial effect was found for strawberry (but not yellow onion). It can be seen that this counteracting effect was weaker than that induced by blackcurrant. White onion still showed some significant beneficial effect on TEER (but less strong than strawberry and yellow onion), whereas GOS showed no significant effect on the *C. difficile* toxin-induced reduction in TEER (Figure 5).

Cytotoxicity by *C. difficile* toxins also includes translocation of the toxins over the intestinal barrier. As an additional test for confirmation of our PCA-based hypothesis, toxin translocation towards the basolateral compartment was studied by ELISA. The results obtained support the TEER measurements (Figure 6) as blackcurrant abolished the toxin translocation almost completely (p-value < 0.05). Strawberry and yellow onion showed a moderate preventive effect (p-value < 0.05) on toxin translocation, while white onion showed significant but weaker effect

compared to yellow onion and GOS did not have any significant effect on the translocation of ToxA and ToxB. Together these data indicate that especially blackcurrant, strawberry and yellow onions, but to a lesser extent maybe also white onion, may have a protective effect against the disruption of intestinal integrity induced by *C. difficile* toxins.

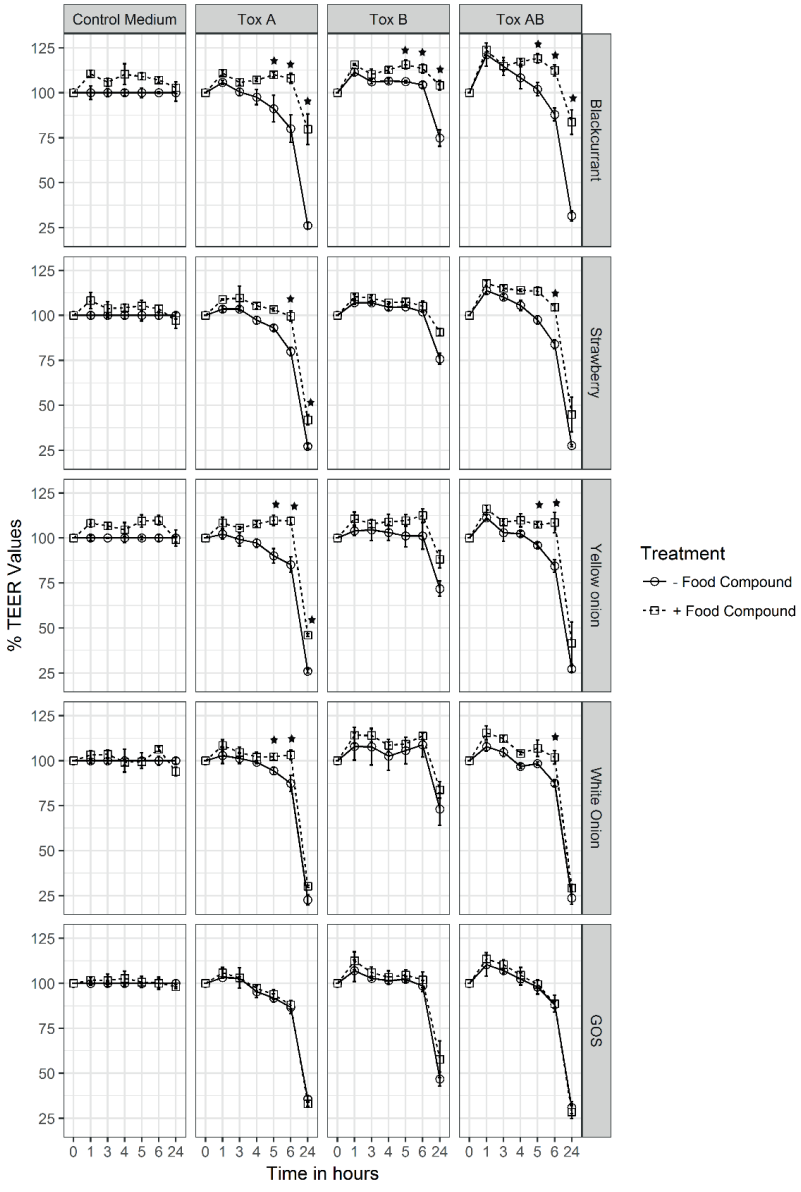


***Figure 5**: Relative TEER values measured before and after exposure of food compounds along with Toxin exposures (TEER values of control medium w/o food compounds were set to 100%). Toxins used include ToxA, ToxB and ToxA + ToxB together. Food compounds used are blackcurrant (Ben Finlay), strawberry (Sabrina), yellow Onion, white Onion and GOS. '*' indicates significant differences in TEER values between - and + food compounds (p-value < 0.05).*

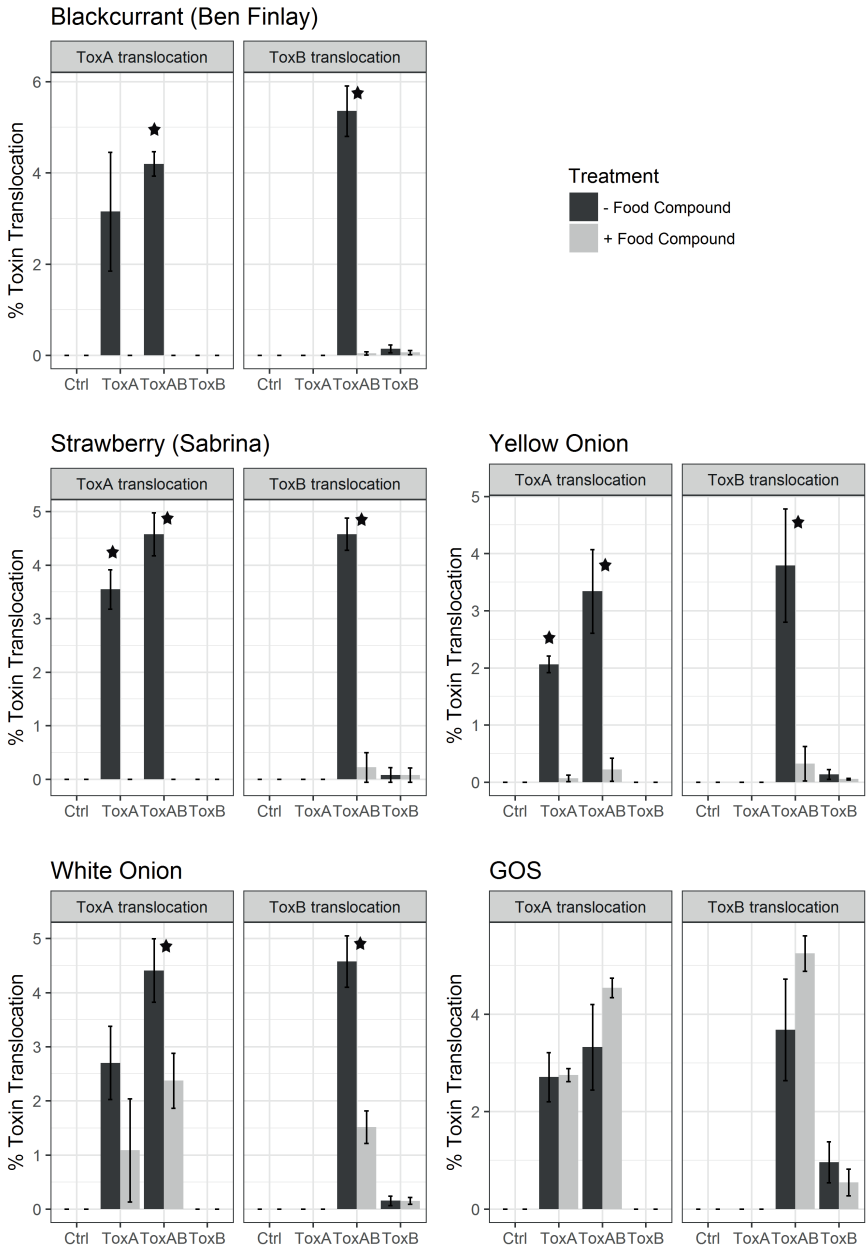


***Figure 6**: ToxA and ToxB translocation measured using ELISA. Measurements were made in the basolateral medium of Caco-2 cells with and without exposure of food compounds namely*

blackcurrant (Ben Finlay), strawberry (Sabrina), yellow onion, white onion and GOS. '' indicates significant results (p-value < 0.05).*

Sorting the food compounds on their counteractive effectiveness based on TEER values and toxin translocation, in the order of most effective to least effective, coincides with our PCA-based hypothesis. Thus, the efficiency of blackcurrant was found to be the strongest in counteracting toxin-induced cytotoxicity, while strawberry and yellow onion were moderately effective and white onion and GOS were least effective.

Discussion

A sizeable number of studies in the past have focused on the effects of *C. difficile* and its toxins on large intestine and cecum with studies involving physiological and gene expression studies [33–38]. However, recently, evidence is emerging on the effects of C. *difficile* on human small intestine and recurring CDI in hospitals [7]. To the best of our knowledge, our work is the first study focusing on the effects of toxins on transcriptional changes in a model of small intestinal-like enterocytes. Previous transcriptomics studies have been conducted on the effects of the toxins on cecal models, namely HCT-8 cells (human ileocecal cell lines) and mouse cecal cells [37,38]. Among the top upregulated genes we identified are RHOB, JUN, DUSP1, KLF6, GDF15 which were also found to be upregulated in toxin exposure studies involving mice cecum and HCT-8 cells [38]. This indicates that these are commonly activated genes by the *C. difficile* toxins. Among the most down regulated genes, ANKRD1 seems to be involved in apoptosis, whereas CCL15 belongs to the CC chemokine family and is a chemoattractant for neutrophils, monocytes, and lymphocytes [39,40].

Overall, we found that gene expression is most altered upon exposure to ToxA or ToxAB, whereas relatively little change was observed for ToxB. ToxB elicited differential expression of less than one third of the total number of genes differentially expressed by either ToxA or ToxAB.

Similarly, TEER measurements indicate that ToxB by itself has less impact on cellular monolayers, at least during the considered exposure time frame.

At pathway level sumoylation pathway processes seem to be significantly perturbed on all three exposures (ToxA, ToxB, ToxA + ToxB). SUMO (small ubiquitin-like modifier) are a family of small proteins having significant structural conservation with ubiquitin [41]. They have several isoforms in humans and some of the isoforms are found to be ubiquitously expressed in human organs while others are restricted to few organs. Sumoylation pathway is the post translational modification of some target proteins where the SUMO proteins attach and detach covalently to their targets and thus modify the functions of the respective target genes [42]. Sumoylation pathway is found to be an important target of bacterial infection [43,44] and although it is reported to play a role in bacterial infections in general [41], no previous study reports its activation by *C. difficile* toxins.

In addition to the Sumoylation pathway, cell and DNA repair pathways, tight junction signalling, and some cytokine activation pathways are significantly perturbed and have been previously mentioned in other studies [37,38,45]. p53 signaling pathway was also significantly perturbed on exposure of ToxAB and not on exposure of toxins individually. It might be that ToxA and ToxB have a synergistic effect as was suggested in previous toxins research [6]. Studies conducted on Caco-2 suggest that ToxB can cause more harm after reaching the basolateral side of a cell which it does not reach on its own but for which ToxA pave the way [6]. Cholecystokinin/Gastrin-mediated Signalling showed highly significant perturbation on exposure to ToxA and genes like RHOB, JUN and others belonging to this pathway, are significantly expressed. This was not surprising as Rho GTPases were identified before to be highly affected by *C. difficile* toxin exposure [31,32]. It should be noted that Rho and/or Jun activation have a substantial role in not only sumoylation pathway, but also for other significant pathways [37,38].

In the past, many transcriptomics based analyses were performed on Caco-2 cells exposed to different foods and other luminal factors. We have made use of this available dataset and

performed an integrative analysis to conceptualize the present experiment. The enhanced context provided by this compendium allowed the prediction of food compounds that might alleviate the effects of CDI. PCA is a common technique for dimensionality reduction [27] and has been used over the years for descriptive and inferential statistics in all kinds of scientific data analysis [27,46]. The PCA plots presented in this work were interpreted on the assumption that a compound, which is most distinct from toxins and toxoids based on transcriptomic effects in Caco-2 cells, might counteract the effects of harmful effects of the toxins. On the other hand, a compound that groups closer to the toxoids is expected to show an effect more similar to that of toxoids, implying it would have a minimal attenuating effect on the toxins or no effect. Food substances chosen based on this assumption proved (by experiments) to have the predicted relative efficiencies (so, blackcurrant - most effective, strawberry and yellow onion - moderately effective and white onion and GOS - least effective). In our study, we did not test compounds that were close to toxins, as their interpretation might be more ambiguous, as similar gene expression patterns and close proximity in the PCA plot indicate foods that trigger similar response as the toxins. It could be speculated that these compounds might enhance the effects of toxins by inducing the same pathways as the toxins, such as the induction of apoptotic processes or reduce the effects of the toxins by interfering with the toxin-induced pathways for instance pathways associated to defence mechanisms. This might be probed in future experiments. Another potential study would be to focus on the ToxA and ToxB breakdown by selecting specific proteases that block enzyme activity or by providing substrate for the toxins glycosylation activity to divert them away from intestinal epithelial cells.

The results of the integrative analysis were followed by experimental characterization of the selected target food compound. In this study we found that berries, such as blackcurrant and strawberries, can have attenuating effects on cytotoxicity induced by *C. difficile* toxins. Berries have been proven to be rich in flavonoids, particularly polyphenols and phenolic acids [47,48] which have been brought in relation with many type of health effects [49]. However, it is not proven whether the polyphenols are indeed responsible for the observed effect or whether there is

another mechanism involved. Follow up experiments with candidate bioactive components present in these foods could help select sources with the highest counteracting effects and such sources could then be further probed *in vivo* in animal studies.

Although our study did not show a counteracting effect of GOS on the cytotoxic effects induced by *C. difficile* toxins, this does not exclude the possibility that specific fractions of GOS can still have a beneficial effect. Previously Sinclair *et al.* found that specific fractions of GOS (especially DP6; fractionated by cation exchange chromatography) could inhibit the binding of cholera toxins to their effective receptor [50]. This inhibitory effect was not found with unfractionated GOS, which was used in our study.

Conclusion

In conclusion, we show that transcriptomics data can be used to identify beneficial food compounds. Our study, specifically indicates that blackcurrant (Ben Finlay) in particular and strawberry (Sabrina) and yellow onion may help to reduce the cytotoxic effects of *C. difficile* toxins on the small intestinal barrier. This result warrant further studies in animal and human models to prove its effectiveness *in vivo*.

Acknowledgements

We would like to thank Friesland Campina, Amersfoort, the Netherlands for providing us with galacto-oligosaccharides (GOS). The project was financial supported by the Dutch Ministry of Economic Affairs within the Systems Biology programme 'Virtual Gut', KB-17-003.02-021 and the TO2Flex programme.

References

1. Burke, K. E. & Lamont, J. T. Clostridium difficile Infection: A Worldwide Disease. *Gut Liver* **8,** 1–6 (2014).

2. Leffler, D. A. & Lamont, J. T. Clostridium difficile Infection. *N. Engl. J. Med.* **372,** 1539–1548 (2015).

3. DePestel, D. D. & Aronoff, D. M. Epidemiology of Clostridium difficile Infection. *J. Pharm. Pract.* **26,** 464–475 (2013).

4. Di Bella, S., Ascenzi, P., Siarakas, S., Petrosillo, N. & di Masi, A. Clostridium difficile Toxins A and B: Insights into Pathogenic Properties and Extraintestinal Effects. *Toxins* **8,** (2016).

5. Kasendra, M., Barrile, R., Leuzzi, R. & Soriani, M. Clostridium difficile Toxins Facilitate Bacterial Colonization by Modulating the Fence and Gate Function of Colonic Epithelium. *J. Infect. Dis.* **209,** 1095–1104 (2014).

6. Du, T. & Alfa, M. J. Translocation of Clostridium difficile toxin B across polarized Caco-2 cell monolayers is enhanced by toxin A. *Can. J. Infect. Dis.* **15,** 83–88 (2004).

7. Killeen, S., Martin, S. T., Hyland, J., O' Connell, P. R. & Winter, D. C. Clostridium difficile enteritis: A new role for an old foe. *The Surgeon* **12,** 256–262 (2014).

8. Navaneethan, U. & Giannella, R. A. Thinking beyond the colon-small bowel Involvement in clostridium difficile infection. *Gut Pathog.* **1,** 7 (2009).

9. Baumgart, D. C. & Carding, S. R. Inflammatory bowel disease: cause and immunobiology. *Lancet Lond. Engl.* **369,** 1627–1640 (2007).

10. Butler, M. et al. Early Diagnosis, Prevention, and Treatment of Clostridium difficile: Update. (Agency for Healthcare Research and Quality (US), 2016).

11. Navaneethan, U., Venkatesh, P. G. & Shen, B. Clostridium difficile infection and inflammatory bowel disease: Understanding the evolving relationship. *World J. Gastroenterol. WJG* **16,** 4892–4904 (2010).

12. Spigaglia, P. Recent advances in the understanding of antibiotic resistance in Clostridium difficile infection. *Ther. Adv. Infect. Dis.* **3,** 23–42 (2016).

13. Zhang, L. *et al.* Efficacy of Cranberry Juice on Helicobacter pylori Infection: a Double-Blind, Randomized Placebo-Controlled Trial. *Helicobacter* **10,** 139–145 (2005).

14. Vreeburg, R. A., van Wezel, E. E., Ocaña-Calahorro, F. & Mes, J. J. Apple extract induces increased epithelial resistance and claudin 4 expression in Caco-2 cells. *J. Sci. Food Agric.* **92,** 439–444 (2012).

15. Minekus, M. *et al.* A standardised static in vitro digestion method suitable for food – an international consensus. *Food Funct.* **5,** 1113–1124 (2014).

16. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5,** R80 (2004).

17. Lin, K. *et al.* MADMAX - Management and analysis database for multiple ~omics experiments. *J. Integr. Bioinforma.* **8,** 160 (2011).

18. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinforma. Oxf. Engl.* **19,** 185–193 (2003).

19. Irizarry, R. A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31,** e15 (2003).

20. Dai, M. *et al.* Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **33,** e175–e175 (2005).

21. Piccolo, S. R., Withers, M. R., Francis, O. E., Bild, A. H. & Johnson, W. E. Multiplatform single-sample estimates of transcriptional activation. *Proc. Natl. Acad. Sci.* **110,** 17778–17783 (2013).

22. Venkatasubramanian, P. B. *et al.* Use of Microarray Datasets to generate Caco-2-dedicated Networks and to identify Reporter Genes of Specific Pathway Activity. *Sci. Rep.* **7,** 6778 (2017).

23. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20,** 307–315 (2004).

24. Carvalho, B. S. & Irizarry, R. A. A framework for oligonucleotide microarray preprocessing. *Bioinformatics* **26,** 2363–2367 (2010).

25. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47 (2015).

26. Phipson, B., Lee, S., Majewski, I. J., Alexander, W. S. & Smyth, G. K. Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *Ann. Appl. Stat.* **10,** 946–963 (2016).

27. Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Philos. Transact. A Math. Phys. Eng. Sci.* **374,** (2016).

28. Saccenti, E., Hoefsloot, H. C. J., Smilde, A. K., Westerhuis, J. A. & Hendriks, M. M. W. B. Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics* **10,** 361–374 (2014).

29. Saccenti, E. & Timmerman, M. E. Considering Horn's Parallel Analysis from a Random Matrix Theory Point of View. *Psychometrika* **82,** 186–209 (2017).

30. Johnstone, I. M. On the Distribution of the Largest Eigenvalue in Principal Components Analysis. *Ann. Stat.* **29,** 295–327 (2001).

31. Lerm, M., Schmidt, G. & Aktories, K. Bacterial protein toxins targeting Rho GTPases. *FEMS Microbiol. Lett.* **188,** 1–6 (2000).

32. Voth, D. E. & Ballard, J. D. Clostridium difficile Toxins: Mechanism of Action and Role in Disease. *Clin. Microbiol. Rev.* **18,** 247–263 (2005).

33. Mitchell, T. J. *et al.* Effect of toxin A and B of Clostridium difficile on rabbit ileum and colon. *Gut* **27,** 78–85 (1986).

34. Hecht, G., Pothoulakis, C., LaMont, J. T. & Madara, J. L. Clostridium difficile toxin A perturbs cytoskeletal structure and tight junction permeability of cultured human intestinal epithelial monolayers. *J. Clin. Invest.* **82,** 1516–1524 (1988).

35. Riegler, M. *et al.* Clostridium difficile toxin B is more potent than toxin A in damaging human colonic epithelium in vitro. *J. Clin. Invest.* **95,** 2004–2011 (1995).

36. Buckley, A. M. *et al.* Susceptibility of Hamsters to Clostridium difficile Isolates of Differing Toxinotype. *PLOS ONE* **8,** e64121 (2013).

37. D'Auria, K. M. *et al.* Systems analysis of the transcriptional response of human ileocecal epithelial cells to Clostridium difficile toxins and effects on cell cycle control. *BMC Syst. Biol.* **6,** 2 (2012).

38. D'Auria, K. M. *et al.* In Vivo Physiological and Transcriptional Profiling Reveals Host Responses to Clostridium difficile Toxin A and Toxin B. *Infect. Immun.* **81,** 3814–3824 (2013).

39. Berahovich, R. D. *et al.* Proteolytic Activation of Alternative CCR1 Ligands in Inflammation. *J. Immunol.* **174,** 7341–7351 (2005).

40. Shen, L. *et al.* Overexpression of ankyrin repeat domain 1 enhances cardiomyocyte apoptosis by promoting p53 activation and mitochondrial dysfunction in rodents. *Clin. Sci.* **128,** 665–678 (2015).

41. Srikanth, C. V. & Verma, S. Sumoylation as an Integral Mechanism in Bacterial Infection and Disease Progression. 389–408 (2017). doi:10.1007/978-3-319-50044-7_22

42. Hay, R. T. SUMO. *Mol. Cell* **18,** 1–12 (2005).

43. David, R. Bacterial pathogenesis: Targeting SUMO. *Nat. Rev. Microbiol.* **8,** 386–386 (2010).

44. Ribet, D. & Cossart, P. SUMOylation and bacterial pathogens. *Virulence* **1,** 532–534 (2010).

45. Nusrat, A. *et al.* Clostridium difficile Toxins Disrupt Epithelial Barrier Function by Altering Membrane Microdomain Localization of Tight Junction Proteins. *Infect. Immun.* **69,** 1329–1336 (2001).

46. Kaya, I. E., Pehlivanlı, A. Ç., Sekizkardeş, E. G. & Ibrikci, T. PCA based clustering for brain tumor segmentation of T1w MRI images. *Comput. Methods Programs Biomed.* **140,** 19–28 (2017).

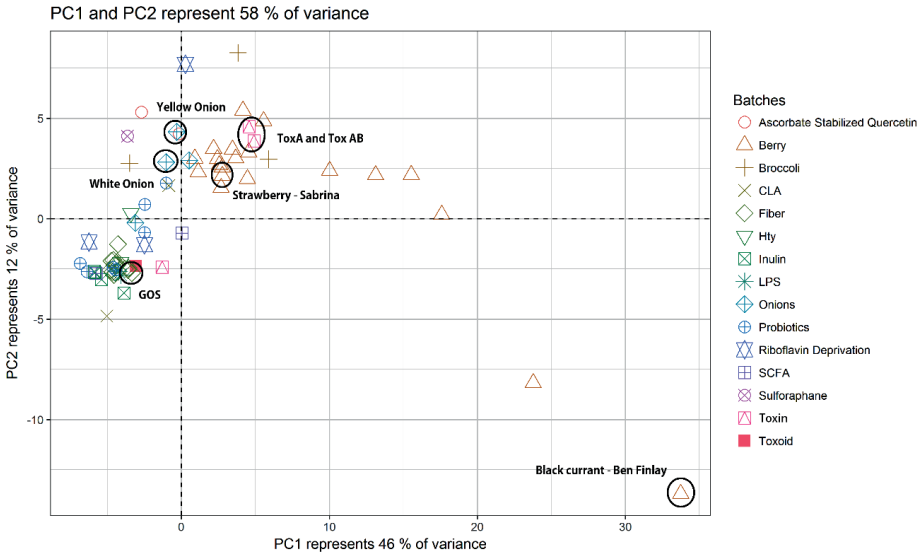
47. Brown, E. M. *et al.* Persistence of Anticancer Activity in Berry Extracts after Simulated Gastrointestinal Digestion and Colonic Fermentation. *PLOS ONE* **7,** e49740 (2012).

48. Oszmiański, J. *et al.* Analysis of Phenolic Compounds and Antioxidant Activity in Wild Blackberry Fruits. *Int. J. Mol. Sci.* **16,** 14540–14553 (2015).

49. Boeing, H. *et al.* Critical review: vegetables and fruit in the prevention of chronic diseases. *Eur. J. Nutr.* **51,** 637–663 (2012).

50. Sinclair, H. R., de Slegte, J., Gibson, G. R. & Rastall, R. A. Galactooligosaccharides (GOS) Inhibit Vibrio cholerae Toxin Binding to Its GM1 Receptor. *J. Agric. Food Chem.* **57,** 3113–3119 (2009).

Supplementary figures and tables



Supplementary Figure S1*: The PCA plotted with 1013 genes that were differentially expressed in Caco-2 after exposure to ToxAB*

Supplementary Table T1: *Table provides details of accession number of the microarray data used in this publication. 'In preparation' indicates data that are yet to be submitted to the public data repositories*

Experiment name	GEO / Array Express	Accession
Broccoli extracts that had been cooked for different lengths of time	Array Express	E-MEXP-1372
Caco-2 cells co-cultivated with B. animalis subsp. lactis BB-12	Array Express	E-GEOD-21930
Conjugated linoleic acid (CLA)	Array Express	E-GEOD-6518
Hydroxytyrosol (HTy) and hydroxytyrosyl ethyl ether (HTy-Et)	Array Express	E-GEOD-38833
Riboflavin depletion	Array Express	E-GEOD-15132
Bifidobacterium bifidum PRL2010 on gene expression in intestinal epithelial cells	Array Express	E-GEOD-21976
Ascorbate-stabilized quercetin	Array Express	E-GEOD-7259
Caco-2 co-culture with Lactobacillus casei and Bifidobacterium breve	Array Express	E-GEOD-37369
Sulforaphane (SF)	Array Express	E-MEXP-170
Onion experiment	GEO	GSE83893
Berry experiment	In preparation	From the lab of J. Mes
Dietary fibres	In preparation	From the lab of J. Mes
Probiotics	In preparation	From the lab of J. Mes
Clostridium Toxins	In preparation	From the lab of J. Mes
Short Chain Fatty Acid Experiments (SCFA)	In preparation	From the lab of J. Mes

A normalisation protocol to mitigate batch effects and allow comparison of the effects of different treatments on the same biological system

Prashanna Balaji Venkatasubramanian[1], Enrico Giampieri[2], Jurriaan Mes[1], Edoardo Saccenti[3], Gastone Castellani[2], Daniel Remondini[2]

Manuscript in final stages of preparation

Abstract

The availability of microarray experimental data has been growing since the establishment of public database repositories such as GEO and ArrayExpress. These databases include large amount of data from microarray experiments using Caco-2 cells, which is a well-established model of the intestinal epithelium. In addition to extracting useful insights into cell function, the Caco-2 microarray datasets can be used to build a classifier that can organise new experiments into different categories of treatments. However, combining different datasets is hampered by batch and platform effects. We provide a simple solution in cases where the sample sizes are low, meta-data is incomplete and where controls are present for each batch of experiments, a situation usually not commonly addressed by available batch correction methods. Our method is based on non-parametric transformation of treatment vs control values using ranks and log-odds.

Initially, the normalisation protocol was used with Caco-2 control only datasets exhibiting both batch and platform effects and then validated using an independent microarray dataset generated with samples from arthritis patients. Finally, the normalisation protocol was applied to microarray data from Caco-2 cells exposed to luminal factors and shown to reduce both batch and platform effects, allowing comparison of treatment effects.

Introduction

DNA microarray technology is an important watershed moment in the history of biological sciences [1–4]. It opened the path to simultaneous genome-wide gene expression analysis using RNA transcripts derived from cells and tissues of interest[2]. Gene Expression Omnibus (GEO) is a data repository that contains high throughput gene expression data, experimental attributes and methodologies. At the time of this writing, more than half the data were from expression profiling using microarrays (~51400 records) [4,5]. Since the dawn of high-throughput technologies, development of methods for analysing big data has become an important topic of research. Reanalysis and cross comparison of high-throughput data generated from different platforms and sources enabled developing models (*e.g.* connectivity map) that may provide newer biological insights [6].

High-throughput experiments can be designed in different ways based on the research questions, requirements and availability of resources for the study. Owing to differing sample sizes and varying design, experiments may be conducted in a single batch or in multiple batches. We categorized the experimental designs under three types: i) Studies with single experimental challenge (*e.g.* a disease condition) conducted over multiple batches containing both experiment and control assays in the same batch (Figure 1); ii) Studies with single experimental challenge (*e.g.* a disease condition) conducted over multiple batches containing experiment and control assays in different batches (Figure 1) iii) studies with multiple experimental challenges conducted over single or multiple batches with each batch having its own control experiments (*e.g.* studies with food related challenges).

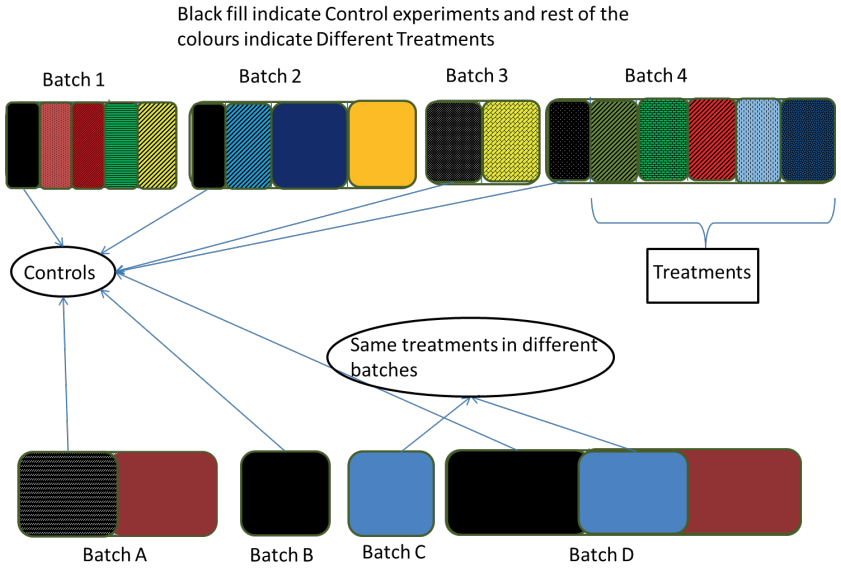


Figure 1*: Figure explains the experimental setup of microarray analysis. The top part of the figure explains the experimental setup in batches (batch 1, batch 2, batch 3 and batch 4) in case of cell cultures. The bottom half of the figure indicates diseased vs healthy samples experiments explained in batches (batch A, batch B, batch C, batch D).*

Studies involving patient samples are often designed based on the first two methods mentioned above with a relatively large sample number (at least >5) while cell culture experiments are designed with relatively much smaller sample sizes and designed based on the third method mentioned above. Reanalysis of these cell culture experimental datasets requires combining multi-experimental dataset that may not necessarily have similar experimental challenges between batches. For instance, food related experimental challenges on Caco-2 cells could be Caco-2 cells exposed to white onion, yellow onion and red onion in one batch and blueberry, blackcurrant and strawberry in another. Combining such datasets could offer detailed insights in to the functioning of Caco-2 cells under differing categories of treatment (*e.g.* onions, berries, *etc.*). Moreover, such datasets could further be used to develop classifiers that can categorise

new experiments into food classes based on the experimental genetic profiles, similar to the approach used in connectivity map [6].

One important challenge in fusing biological high-throughput data is the uncontrollable errors induced into the measurements due to external factors, known as batch effects [7]. Lazar and colleagues have provided a detailed analysis on definition and causes of batch effects [8]. Batch effects are a qualitatively different measurement errors based on the conditions in which experiments were conducted and are unrelated to the biological or scientific focus of the study [9]. These effects are induced by various factors, such as different technicians performing the experiments, difference in days in which each batch of studies were conducted, variation in time of the day of the experiments and the batch of reagents used, among others [7,9]. The batch effects are not only persistent in high-throughput data, but also in low-throughput experiments [9]. Batch effects interfere in analysis of expression data and may lead to false outcomes and predictions. Previous studies on batch effects show that standard normalization protocols do not remove batch effects[9].

***Table 1**: Table showing different batch correction methods along with their required sample sizes and the number of studies that they could be used for. Table reproduced from Lazar et al, 2013 [8].*

Method	Complexity	No. of samples	No. of studies	Flexibility	Additional info required	Computational time
BMC	Low	> 25	> 2	Low	No	Low
Gene standardization	Low	> 25	> 2	Low	No	Low
Ratio based methods	Low	> 25	> 2	Low	Yes	Low
Scaling relative to reference	Low	> 25	> 2	Average	No	Low
Empirical Bayes (ComBat)	High	> 5	> 2	High	No	Low
XPN	High	> 30	2	Low	No	High
DWD	Average	> 25	2	Low	No	Average
SVD-BR	Average	> 25	> 2	Average	No	Average
SVA	Average	> 25	> 2	Average	No	Average
RUV-2	Average	> 25	> 2	High	Yes	Average
Quantile discretization	Low	> 25	> 2	Low	No	Low
fRMA Barcode	Low	1	> 2	Low	No	Low

As the commonly used normalization protocols do not remove batch effects, several batch effect removal protocols have been introduced over the years (Table 1). Some of the commonly[10] used methods from the table are ComBat [7], surrogate variable analysis (SVA) [11], Distance weighted discrimination (DWD)[12] and Ratio based methods[13]. However, most of these methods have been tested on datasets that were derived from diseased vs healthy sample studies[7,14,9,15]. Additionally, most of these methods are reported to require a large number of samples (at least > 5) [8], which are generally a common place in studies involving diseases, particularly cancer. However, most luminal factor exposure studies that are available in public databases, use cell or tissue culture models and are usually conducted in biological triplicates [16]. Owing to the small sample sizes in food intervention studies, usage of popular batch removal techniques like Combat/ SVA could not be implemented [8]. In addition to these, ComBat allows the user to provide experimental metadata as covariates for better batch effect correction (for instance the control medium needs to be defined because it often varies between labs and experiments). However, all required metadata is often not available. Moreover, some of the commonly used batch effect methods (ComBat and Limma with batch effect removal option set true) have been reported to deflate group differences, when the study groups are not evenly distributed across batches [17].

In this article, we present a non-parametric method for cross comparison of different gene expression datasets, consisting of experiments conducted on same biological systems and collected over time. The method can be applied to experiments with smaller sample sizes, unevenly distributed study groups and without much need for all the metadata.

Methods

Batch effect removal protocol

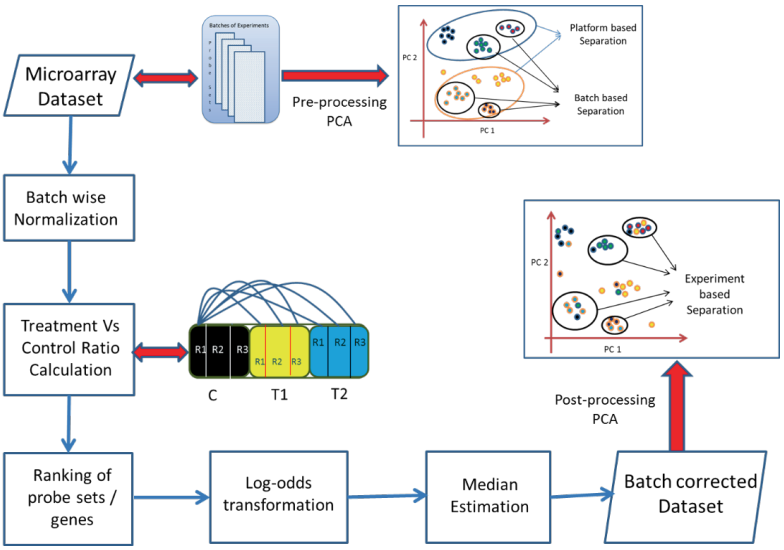


***Figure 2**: Workflow of the method described in the paper.*

The batch effect removal protocol is illustrated in Figure 2. Assuming that:

M indicates total experiments (*e.g.* onion experiment, berry experiment, *etc.*), each conducted in separate batches; M_y indicates each experiment; *C* indicates control exposure; *T* indicates total treatment exposure in each experiment; T_w indicates each treatment (*e.g.* yellow onion, red onion, *etc.*); *k* indicates the number of control replicates; *t* indicates the number of treatment replicates and *N* indicates the number of probesets, the normalization protocol is as follows

1. The microarray dataset for each experiment is normalized batch-wise using RMA (robust multi average) normalization protocol which results in an RMA normalized data matrix of dimension $N \; X \; M_y*(k+w*t)$. The rows represent the probesets and the columns represent control and experimental replicates.

2. The treatment (T_{wt}) vs control (C_k) ratio values (V_{wtk}) are estimated for each experiment. The ratios are estimated between all treatment replicates of each experiment against all control replicates.

$$V_{kwt} = \frac{T_{wt}}{C_k}$$

Where, *k* is the number of control replicates, *w* is the number of treatments and *t* is the number of treatment replicates. This step yields *Z* = *k* * *t* number of ratio values (*V*), *ie.* 3 control replicates and 3 treatment replicates yield *Z* = 9 (=3*3) ratio values (*V*) for each treatment *w*. This step will result in a ratio matrix of dimension *N X k*w*t*.

3. The *N* probesets are then ranked (R_{ij}) from 1 to *N* based on their V_{wtk}, where *i* indicates the probeset and j indicate the column (a total of *k*w*t* columns). Each treatment with *Z* number of treatment-to-control ratios will get Z different ranks for each probeset. This step will result in a rank matrix of dimension *N X k*w*t*.
4. The ranked values are then log-odds transformed for each ranked probesets.

$$Xi = \left(\frac{Rij}{N+1}\right) \Big/ \left(1 - \frac{Rij}{N+1}\right)$$

Where, X_i is the log-odds transformed value of probeset *i*, R_{ij} is the rank of each probeset *i* for each column *j*. The log-odds transformation converts the ranks which are then distributed between (0, 1). This step will again result in log-odds matrix of dimension *N x M*k*t*.

5. The median value of treatment vs control set (*k*t* values) for each probeset is estimated. This step will result in batch corrected matrix of dimension *N x M*.

Experimental data

Compendium of Caco-2 datasets

The compendium contains microarray data of Caco-2 exposure experiments collected from GEO and from in-house experiments[16] using affymetrix© microarray platform (Hgu133plus2,

Hugene 1.0 ST and Hugene 1.1 ST arrays). The experiments include exposure of food substances, pathogens and food compounds like sulforaphane, among others, on Caco-2 cells and in total there are 82 exposure experiments performed in 14 batches. Each batch consists of several exposure challenges (e.g, white onion, yellow onion and quercetin in onion experimental batch) and one control exposure (say, DMEM medium control).

Arthritis Datasets

Arthritis data was obtained from experiments conducted using cells cultured after extraction from Arthritis patients [18]. The dataset was taken from GEO with accession number GSE13837. The dataset consists of four batches of experiments conducted with Synovial fibroblasts tissue samples collected from Osteoarthritis (OA) and Rheumatoid Arthritis (RA) patients. The cells were cultured for several days and treated with TGF-β and TNF-α. RNA was harvested after different exposure times (0, 1, 2, 4 and 12 hours) and hybridized to Affymetrix© HG133plus2 arrays. The study was separated into different batches based on the date of array hybridisation. Of the four batches, 2 consisted of data from 2 patients (i.e. 2 samples) while the other two batches contained data from only one patient (i.e. 1 sample). In order to mimic the Caco-2 compendium experimental design, 1 RA only batch (2 patients) and 1 OA only batch (2 patients) [Batch 3 and Batch 4] were chosen as additional validation set for verifying the efficacy of the proposed batch effects mitigation protocol.

Caco-2 dataset pre-processing

As the Caco-2 datasets belonged to three different array platforms, a pre-processing step involving ID mapping was carried out. The dataset was corrected for probeset differences between HGU133plus2 arrays and Hugene arrays. HGU133plus2 arrays contained 54676 probesets while Hugene 1x ST arrays contained 32321 probesets. The Affymetrix© provided probeset conversion data for HGU133plus2 and Hugene 1x ST arrays were used and only the best matches were considered for ID mapping. 7362 probes in Hugene 1x ST arrays had unique

one-to-many mapping with 18668 probes from HGU133plus2. 28763 probesets of HGU133plus2 were found to be mapped to Hugene 1x ST arrays. For all arrays with multiple mapping in HGU133plus2 arrays the same expression values from Hugene arrays were used.

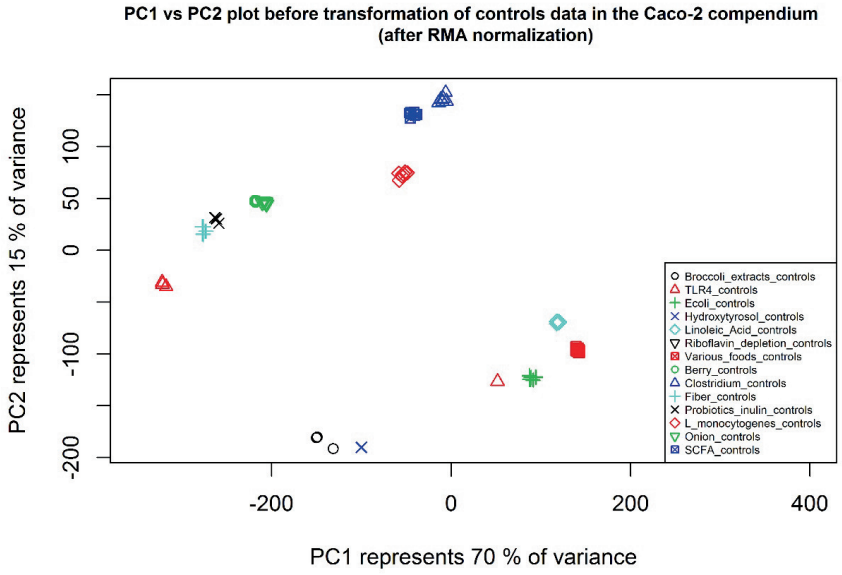
Estimation of Control vs Control differential expression values

To test the efficacy of the batch effect removal protocol proposed in the paper, it was initially applied to the Caco-2 dataset controls only. Controls array data from the Caco-2 data compendium were collected and the differential expression ratio between the replicates of control experiments were calculated against each other. For example, if a control experiment was conducted in triplicates (C_1, C_2, C_3), that would give rise to 3 sets of differential expression values (C_1/ C_2, C_1/ C_3 and C_2/ C_3). In other words, it would be based on mathematical combinatorics C_2^3.

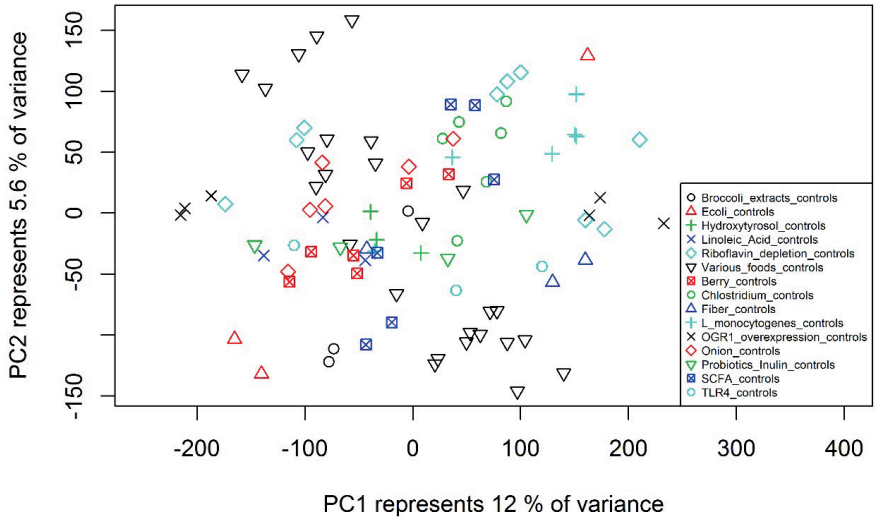
Results

Caco-2 Control data

As a validation of the protocol described here, the method was applied on a dataset that contained only the control experiments derived from the Caco-2 compendium. Figure 3a shows the clustering of array data from controls only, after RMA normalization. The arrays were found to be clustered into groups based on batches of experimental protocol. The controls data were then processed using the proposed methodology. The clustering of the processed data using PCA showed that there was no more separation based on the experimental batch. Figure 3b shows the results of PCA (first two principal components explaining 17% of variance). Figure 3c shows the PCA clustering with hulls plotted pairwise against the first three PCs. Given a group of points, hulls are lines that join some/ all of the given points to form a closed figure that encompasses all the points within its perimeter and most often the points connecting the hull lines lie far apart from each other.


PC1 vs PC2 plot before transformation of controls data in the Caco-2 compendium
(after RMA normalization)
PC2 represents 15 % of variance
PC1 represents 70 % of variance
-200
-100
0
50
100
-200
0
200
400
Broccoli_extracts_controls
TLR4_controls
Ecoli_controls
Hydroxytyrosol_controls
Linoleic_Acid_controls
Riboflavin_depletion_controls
Various_foods_controls
Berry_controls
Clostridium_controls
Fiber_controls
Probiotics_inulin_controls
L_monocytogenes_controls
Onion_controls
SCFA_controls

a.


PC1 vs PC2 plot after transformation of controls data in the Caco-2 compendium
PC2 represents 5.6 % of variance
PC1 represents 12 % of variance
-150
-50
0
50
100
150
-200
-100
0
100
200
300
400
Broccoli_extracts_controls
Ecoli_controls
Hydroxytyrosol_controls
Linoleic_Acid_controls
Riboflavin_depletion_controls
Various_foods_controls
Berry_controls
Chlostridium_controls
Fiber_controls
L_monocytogenes_controls
OGR1_overexpression_controls
Onion_controls
Probiotics_Inulin_controls
SCFA_controls
TLR4_controls

b.

c.

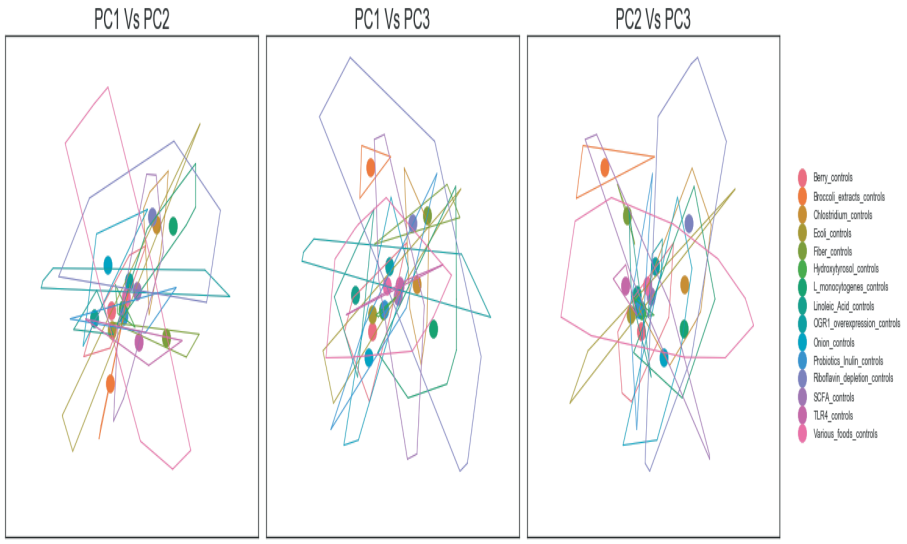
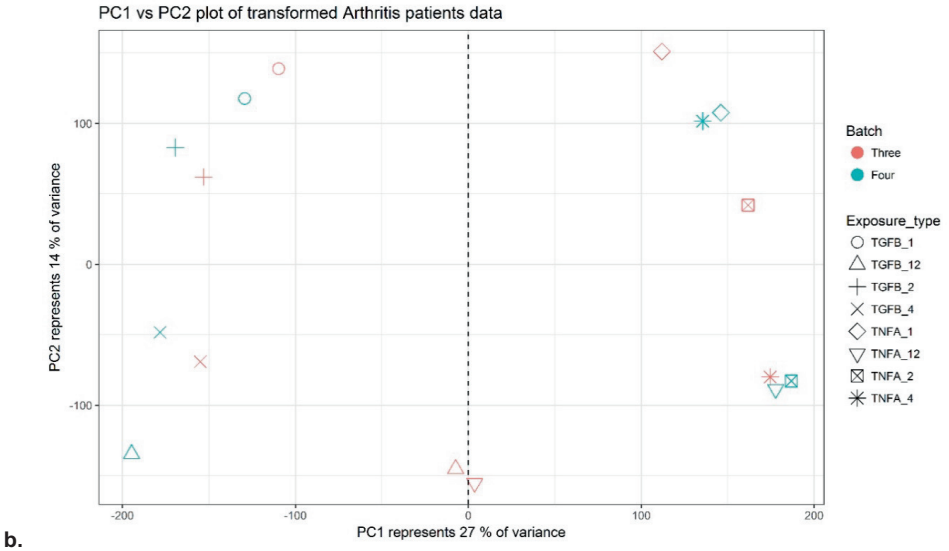


***Figure 3**: **a)** The first two Principal Components are plotted for raw controls data from the Caco-2 data compendium. Control data group together based on the batch in which experiments were performed. Additionally, another grouping of data can be observed based on experimental platform (old affymetrix arrays vs new affymetrix arrays) about a diagonal drawn across the two PCs. **b**) The first two Principal Components are plotted for transformed (using the method described in the paper) controls data from the Caco-2 data compendium. No grouping is observed among the control experiments which was as expected from controls data. **c)** The plots show pairwise comparison between the first three Principal Components plotted for batch corrected (using the method described in the paper) controls only data from the Caco-2 data compendium. The lines forming each polygon is constructed by joining extreme values of a data group (hulls) and all data points belonging to a group fall within the hull. The points indicate the centre of each group of data points. No clear grouping is observed among the control experiments as the data points of each experimental group overlap with each other.*

Arthritis data

The protocol was further validated using microarray dataset derived from Rheumatoid arthritis and Osteoarthritis patients. Figure 4a shows the plotted results of the Principal Component Analysis (PCA) of the arthritis array data based on the first two components after RMA normalization. The first two PCs together represent 51% of the total variance in the data. The arrays were found to separate clearly on PC1, into batches in which the experiments were performed. The data was later processed as mentioned in the batch correction method proposed here. A PCA was again performed and the first two PCs represented 42% of the variance (Figure 4b). The mitigation of batch effects was confirmed by the separation into two groups based on exposure to cells instead of the separation into batches in which experiments were conducted.

a.



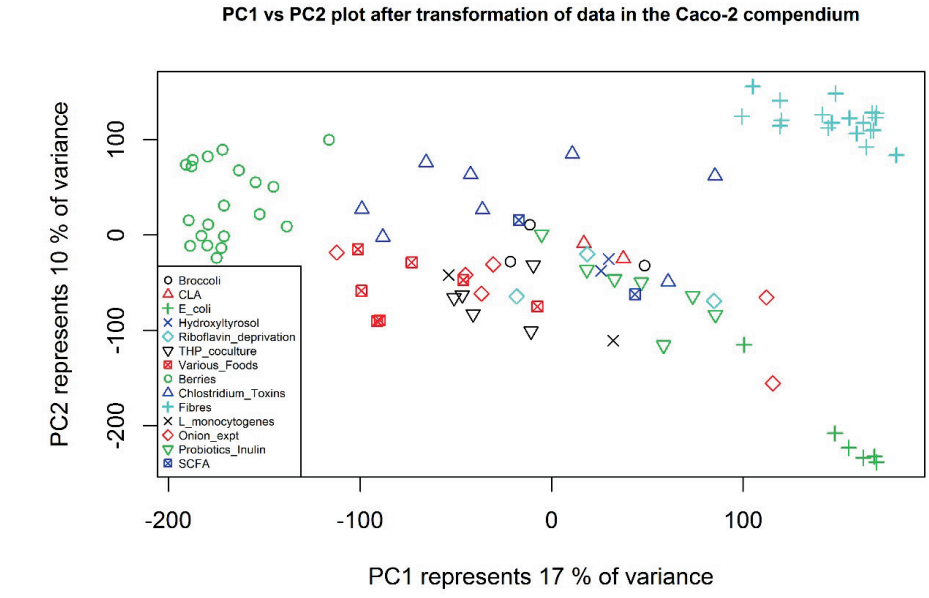
b.

***Figure 4**: **a**) The first two Principal Components are plotted for the raw RMA normalized arthritis patient sample data. Data from similar exposure experiments do not group together while a clear separation is observed between the two batches of experiments; **b)** The first two Principal Components are plotted for transformed (using the method described in the paper) arthritis patients data. Data from similar exposure experiments group together and a separation is observed between the two different experiments (TNF-α and TGF-β).*

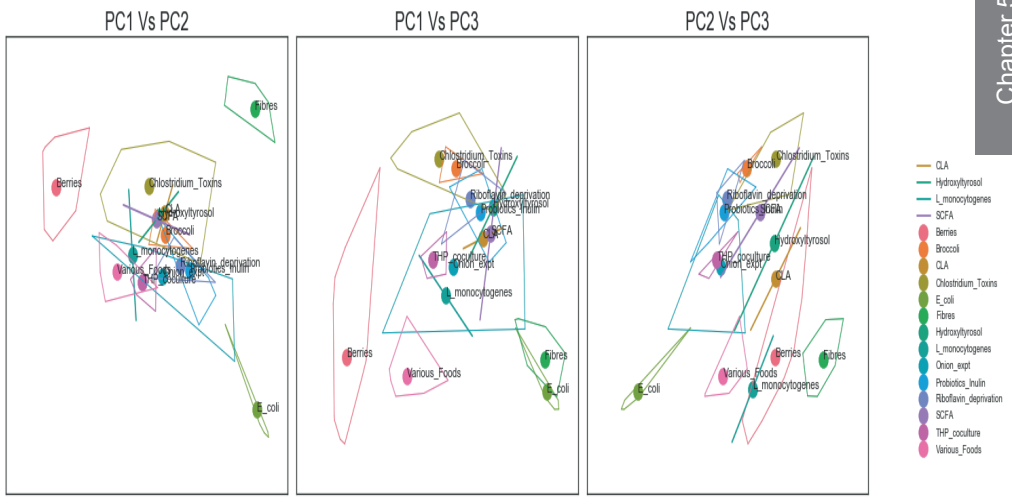
Application on Caco-2 compendium

The proposed non-parametric method was further applied on the complete Caco-2 compendium. The compendium consisted of 82 experiments conducted over 14 batches of experiments. The batch correction protocol mentioned here was applied on the complete Caco-2 data compendium and the PCA was estimated. Figure 5a and figure 5b show the PCA plots (first two principal components (PCs) explain 28% of variance) of the batch corrected Caco-2 data. In figure 5a, the first two PCs are plotted and the plots show that the experiments are separated into local groups which is what is expected from the data. There are no platform effects observed (old vs new arrays). None of the exposures are repeated in different batches

and the grouping could only be due to differences in gene expression induced by the differences in exposures. Figure 5b shows the PCA between the first three PCs in pairwise manner plotted using hulls and the points indicated in the centre of the polygons.



a.



b.

***Figure 5**: **a)** The first two Principal Components are plotted for transformed (using the method described in the paper) data from the Caco-2 data compendium. Grouping of data is observed based on the differences in exposure and not based on platform or batches. **b)** The plots show pairwise comparison between the first three Principal Components plotted for transformed (using the method described in the paper) controls data from the Caco-2 data compendium. The lines forming each polygon is constructed by joining extreme values of a data group (hulls) and all data points belonging to a group fall within the hull. The points indicate the centre of each group of data points. Some grouping is observed among the experiments due to differences in exposure. Some of the experimental groups overlapping with each other may have some similarity in gene expression.*

Discussion

Affymetrix© oligonucleotide microarrays are one of the most popular single channel microarray technologies and in the past vast amounts of data has been deposited public databases such as GEO [5]. Of these, HGU133 and HGU133_Plus2 arrays were widely used in the past while newer platforms like HuGene arrays are gaining popularity in recent years [19]. Therefore, we chose to test our protocol for mitigation of batch effects by combining data generated from HGU133_Plus2 arrays and the recent Hugene arrays. While data from other platforms have not been included in this study, being a non-parametric method, this method will be robust to integration studies involving gene expression data from different platforms, such as those from two channels arrays of agilent© technologies and from RNAseq experiments.

The method described in this study could be used in cases where there are few samples (> 2), large number of batches, the controls are different between the batches, platform effects are present and complete set of relevant metadata is unavailable. These are conditions in which the other commonly used batch effect correction methods are not effective[8,17]. The batch correction protocol here is a culmination of ratio based methods[20] and median rank protocols[8]. The intuition

behind this is that the batch effects are same in control and samples, since they were performed in the same experimental batch. This may however not apply to experiments with control and treatments being performed in different batches and therefore this may not be the optimal method in such cases. In this situation, methods like surrogate variable analysis and ComBat maybe better suited. While protocols like ComBat and SVA have been proven to be efficient in several benchmarking studies [8,10,13,21], most of these studies and methods have been proven only on datasets consisting of samples from disease studies. Moreover, the focus of these studies were to assess the impact of the disease on transcripts while the same data could have been used to understand other kind of biological questions which have been largely ignored [7–9,13,13,20]. For instance, the arthritis patient data could be used to either understand the molecular differences between osteoarthritis and rheumatoid arthritis samples or to study the molecular effects of the temporally varied exposure of certain stimulants (TNF-a and TGF-b) on arthritis patient samples (irrespective of the type of arthritis / gender of the subjects). While the benchmarking studies have focused on the former, we chose the latter as it resembles the studies in our data compendium.

We applied our batch correction protocol on Caco-2 controls only dataset as a validation step. Different types of growth medium at differing concentrations have been used as control experiments in the Caco-2 dataset and it was expected that even if the different growth media trigger significant gene expression changes, these changes should stand neutralised when the ratio is calculated against each other. Therefore, after the mitigation of batch effects using our protocol we predicted that control experiments will not form any grouping based on the minor differences in medium, which was reflected in our results (Figure 3b and Figure 3c). Moreover, it should be noted the Caco-2 dataset used in this study comprises of no common experiments except for some common controls. Therefore, the differences observed in the dataset (Figure 5b) were mostly due to the biological effects.

It should further be noted that the method described in this chapter, does not contain statistical measures (e.g. p-values) that enable identification of differentially expressed genes. While the

protocol mentioned will be suitable for analysis based on data fusion, other methods maybe better suited for studies that focus on identification of differentially expressed genes. After application of the batch effect mitigation protocol on the Caco-2 data compendium used in this chapter, it could further be exploited to develop a classifier model based on the gene expression profiles, similar to the connectivity map[6]. This Caco-2/ luminal factor exposure experiments specific connectivity map could be used to classify new Caco-2 exposure experiments and find luminal factors that are similar to each other.

In conclusion, we provide a non-parametric protocol that mitigates batch effects and allows integration of cross-platform gene expression data generated from different labs.

Acknowledgement

The project was financial supported by the Dutch Ministry of Economic Affairs within the Systems Biology programme 'Virtual Gut', KB-17-003.02-021 and by Wageningen Institute of Animal Sciences (WIAS) PhD travel grant, 2016.

References

1. Lodish, H. *et al.* DNA Microarrays: Analyzing Genome-Wide Expression. (2000).

2. Trevino, V., Falciani, F. & Barrera-Saldaña, H. A. DNA microarrays: a powerful genomic tool for biomedical and clinical research. *Mol. Med.* **13,** 527–541 (2007).

3. Bumgarner, R. DNA microarrays: Types, Applications and their future. *Curr. Protoc. Mol. Biol. Ed. Frederick M Ausubel Al* **0 22,** Unit-22.1. (2013).

4. Clough, E. & Barrett, T. The Gene Expression Omnibus database. *Methods Mol. Biol. Clifton NJ* **1418,** 93–110 (2016).

5. GEO Summary - GEO - NCBI. Available at: https://www.ncbi.nlm.nih.gov/geo/summary/?type=series. (Accessed: 1st September 2017)

6. Lamb, J. *et al.* The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science* **313,** 1929–1935 (2006).

7. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8,** 118–127 (2007).

8. Lazar, C. *et al.* Batch effect removal methods for microarray gene expression data integration: a survey. *Brief. Bioinform.* **14,** 469–490 (2013).

9. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11,** 733–739 (2010).

10. Chen, C. *et al.* Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. *PLOS ONE* **6,** e17238 (2011).

11. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28,** 882–883 (2012).

12. Huang, H., Lu, X., Liu, Y., Haaland, P. & Marron, J. S. R/DWD: distance-weighted discrimination for classification, visualization and batch adjustment. *Bioinformatics* **28,** 1182–1183 (2012).

13. Luo, J. *et al.* A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *Pharmacogenomics J.* **10,** 278–291 (2010).

14. Walker, W. L. *et al.* Empirical Bayes accomodation of batch-effects in microarray data using identical replicate reference samples: application to RNA expression profiling of blood from Duchenne muscular dystrophy patients. *BMC Genomics* **9,** 494 (2008).

15. Müller, C. *et al.* Removing Batch Effects from Longitudinal Gene Expression - Quantile Normalization Plus ComBat as Best Approach for Microarray Transcriptome Data. *PLoS ONE* **11,** (2016).

16. Venkatasubramanian, P. B. *et al.* Use of Microarray Datasets to generate Caco-2-dedicated Networks and to identify Reporter Genes of Specific Pathway Activity. *Sci. Rep.* **7,** 6778 (2017).

17. Nygaard, V., Rødland, E. A. & Hovig, E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* **17,** 29–39 (2016).

18. Wollbold, J. *et al.* Adapted Boolean network models for extracellular matrix formation. *BMC Syst. Biol.* **3,** 77 (2009).

19. Jaksik, R., Iwanaszko, M., Rzeszowska-Wolny, J. & Kimmel, M. Microarray experiments and factors which affect their reliability. *Biol. Direct* **10,** 46 (2015).

20. Liu, H.-C. *et al.* Cross-generation and cross-laboratory predictions of Affymetrix microarrays by rank-based methods. *J. Biomed. Inform.* **41,** 570–579 (2008).

21. Larsen, M. J., Thomassen, M., Tan, Q., Sørensen, K. P. & Kruse, T. A. Microarray-Based RNA Profiling of Breast Cancer: Batch Effect Removal Improves Cross-Platform Consistency. *BioMed Res. Int.* **2014,** (2014).

General Discussion

Research Aim:

The role of food in human health is enormous and there is a clear need to study the systemic impact of food on enterocytes, which provide a critical interface for absorption of nutrients while also playing a defensive role against colonisation and invasion by pathogens. Food is composed of many different compounds capable of provoking distinct effects in intestinal cells either alone or in combination. Besides food, enterocytes encounter a large variety of microorganisms [1–3] including metabolites, such as the short-chain fatty acids which can reach concentrations of 100 mM in the intestine. Over the years, many transcriptomics studies have investigated the impact of individual food substances or organism on enterocytes using monolayers of Caco-2 cells as a model system. With such a large collection of data available, there is an opportunity to analyse the transcriptomics responses of intestinal cells systematically, in order to understand how the different functions of enterocytes such as nutrient uptake, hormonal signalling, immune signalling and tissue renewal are affected by foods and food compounds.

This thesis can be seen as a culmination of two related research directions. The first was aimed at facilitating the integrated systemic analysis at the transcriptomic level using the following:

I. Generation of a Caco-2-specific ingestion-related compendium of transcriptomics data and a Caco-2 specific protein-protein interaction network (PPIN) (**Chapter-2**)
II. Development of a protocol to identify reporter genes in a pathway of interest for qPCR analysis using the data compendium above (**Chapter-2**)
III. Development of a protocol to overcome batch effects and allow for a low-level integration of data in the Caco-2 compendium. (**Chapter-5**)

The second research line, investigated enterocyte responses to *Clostridium difficile* toxins focusing on:

i) The impact of *C. difficile* toxins at both miRNA and mRNA level in colon-like Caco-2 cells (**Chapter-3**)

ii) The modulatory effect of food on the activity of *C. difficile* toxins on small intestinal like Caco-2 cells. This was achieved using the data from the compendium we have generated in the first research line, combined with transcriptomic data of Caco-2 cells exposed to *C. difficile* toxins. (**Chapter-4**)

In this chapter, the key findings of the previous chapters are discussed in the context of the broader research field, including the challenges associated with using systems biology approaches. Finally, the conclusions and future prospects for the research are discussed.

Key Research Findings:

1. Large amount of transcriptomics data from microarray experiments specific to Caco-2 studies are already available in public databases and it was demonstrated that these datasets could be used for selection of genes of interest (reporter genes) in order to study the effects of treatments towards a certain pathway of interest.
2. While complete metadata on Caco-2 exposure experiments are not available, the experiments are designed to contain a control treatment in each batch. Thus, batch effects can be handled efficiently by combining ratio-based methods and median rank scores.
3. Several miRNAs like miR-16-5p, miR-128-3p and miR-194-3p among others, may have a significant influence in gene expression regulation and biological processes that are regulated in enterocytes exposed to *C. difficile* toxins and might be important targets for therapeutic strategies.
4. Blackcurrant and other berries might help to mitigate the harmful effects of *C. difficile* toxins in small enterocytes.

Current approaches to the reanalysis of transcriptomics data

In the past two decades, the number of high-throughput experiments has increased exponentially [4] resulting in the rapid expansion of data in public database repositories such as GEO [5] and array express [6], among others. Most of these data are generated from targeted experiments, performed to answer specific questions related to the biological system under study. This has provided ample opportunities for combining datasets from multiple sources and reanalysing them. Reanalysis of publicly available transcriptomics data has been performed previously in multiple studies involving different tissues/ organisms and has proven to give novel insights from a systems perspective [7–9]. For example, Clarke and colleagues reanalysed 13 independent microarray datasets related to cystic fibrosis and other respiratory disorders, thereby identifying biomarkers common to cystic fibrosis and other similar respiratory disorders. They have also identified potential regulators of CFTR (cystic fibrosis transmembrane conductance regulator) gene using correlation studies which may lead to novel therapeutic targets [10]. Similarly, Kangaspeska and colleagues have reanalysed RNAseq data to identify fusion genes having multiple isoforms [7] while Sharma reanalysed transcriptomics data to establish the link between embryonic gene expression and offspring phenotypes related to altered metabolism [9]. In **chapter 2 and 4**, we have performed reanalysis of Caco-2 specific transcriptomics data to achieve two different goals (**Chapter 2** – to identify pathways specific reporter genes and **Chapter 4** – to identify therapeutic food substances that mitigate the effects of *C. difficile* toxins).

Moreover, reanalysis of data could lead to data fusion at different levels. In Connectivity map, Lamb and colleagues have fused data at probeset level (gene expression profile) and used non-parametric ranking based method to find functional association between diseases, genes and

small molecules [11]. The connectivity map could be used as a classifier to find novel functional associations between new drugs and diseases. A key challenge associated with data fusion approaches are batch effects. These are systemic effects that are dependent on the environment of the experiment and they are reflected in the gene expression data [12–14]. While Batch effects are unavoidable when experiments are not performed together, providing all the metadata related to an experiment would help scientists in making a better decision related to statistics when they set upon reanalysing datasets. While statistical models that mitigate batch effect are already in existence, each have their own limitations and thus do not suit all situations [15]. In **chapter 5,** we've developed an alternative method to address batch effects given a specific experimental design.

A data compendium built from different data sources could be exploited in many more ways that provided above. The data could be integrated with canonical biological networks such as a protein-protein interaction network (PPIN) [16], genome scale metabolic network [17], signalling network, or gene regulatory network [18], *etc*. This could in turn be used to generate novel insights by identifying hub genes (genes that interact with multiple genes and act as a common link between them) in a PPIN [16] for example, by performing flux balance analysis after integrating gene expression data with a metabolic network [17]. The data could also be superimposed in an exposure specific manner on a PPI network and such networks could further be used in multiplex network analysis in order to identify conserved modules, important hub genes and also study network growth via a diffusion framework [19]. In addition to this, knock-out and time series experiments in the future could enable dynamic modelling and in developing petri nets to study particular processes of interest [20,21].

Advantages of cell-specific dedicated networks and further advancement into experiment-specific networks

Expression of genes in perturbed biological systems may vary in multiple ways owing to the type of tissue under study [22], environmental factors including microbe-host interactions [23] or the specific cell culture conditions [24,25]. This in turn affects the proteins that are available in a cell and therefore canonical networks will need to be modified to represent interactions specific to cells under study. In **chapter 2**, we have provided a PPI network that is specific to the Caco-2 cell system which was derived using combined expression data from multiple Caco-2 exposure experiments. This network could be considered as the first step towards a Caco-2 specific network and could further be improved using data from additional experimental exposures. In addition to this, such networks could be made specific to certain types of exposure. Multiple networks of this type could be combined and studied to derive novel insights about cell specific conserved modules, as mentioned earlier.

The need for standardized network inference protocols

Over the past two decades, gene expression data has been extensively utilised for developing *ab-initio* gene regulatory networks and different protocols exist for the same purpose. Some of the popular protocols like ARACNE [26] and CLR [27] were mentioned in **Chapter 1**. While these protocols have been efficient in determining the regulatory network in prokaryotes using mutual information matrix with high accuracy, they have not been efficient in predicting regulatory networks in lower order eukaryotes such as *Saccharomyces cerevisiae*. This may be due to the greater complexity of gene regulation in higher organisms. To overcome this problem, Marbach and others have recommended using several network inference protocols together to achieve an accurate consensus network [28]. However, this strategy has not been proven in case of higher order eukaryotes like mammals and an efficient *ab-initio* protocol still remains elusive. Recently, Banf and Rhee have tried to work around this limitation by developing an algorithm called

'GRACE' [29]. The GRACE algorithm functions by combining *a priori* knowledge on regulatory networks with heterogeneous expression data using markov random fields to improve the accuracy of the predicted gene regulatory network. They have tested their algorithm by identifying regulatory networks in *Arabidopsis thaliana* and *Drosophila melanogaster.* While this protocol has been shown to work in cases of sparse information, a complete *ab-initio* network inference protocol remains a challenge for higher order eukaryotes and should be developed in the future.

Challenges in analysing expression data related to food perturbations

Studying the impact of food exposure on gene expression in enterocytes is important to understand the functional mechanism underlying the interaction between food and the host. Statistical measure of significance using standardised tests is generally carried out in such studies to prevent identification of false positives [30,31]. However, in case of food exposure, the change in expression of genes analysed by microarrays is usually small in terms of fold change. This often leads to very few true positives that are identified as statistically significant [32]. Using such small set of significant genes to perform gene set enrichment analysis and pathway over representation analysis could subsequently lead to false results (false positives). All these drawbacks highlight the need for an experimental technique that is more robust and sensitive for identification of regulated genes. RNA-sequencing technology might be the required solution, as it offers a good dynamic range of transcripts identification from low expression to very high expression.

Microarray vs RNAseq for food related exposure studies

Soon after the introduction of next generation sequencing techniques like the RNAseq, multiple studies have been conducted comparing the efficiency of using RNAseq against microarrays. Comparison studies have been conducted using different sources of RNAseq and microarray data, including *Saccharomyces cerevisiae* [33], rat liver samples [34], activated human T-cells [35],

human neuroblastoma [36] and with human fibroblast cells [37]. Despite the varying conditions in these studies, the general conclusion indicates that RNAseq offers several advantages over the use of microarray as an expression analysis platform. For instance, Wang and others (rat liver sample study) have concluded that RNAseq outperforms microarrays when validating differential gene expression by qPCR (93% concordance with qPCR in case of RNAseq against 75% concordance with qPCR in case of microarray) [34]. Zhao and colleagues (using activated T-cells data) have shown that RNAseq was superior in detecting transcripts with low abundance, differentiating biologically important isoforms and also offered a broad dynamic range for identification of differentially expressed genes with high fold change [35].

We also performed a cross comparison of results between array and RNAseq. We generated our own RNAseq data from Caco-2 cells exposed to *Clostridium difficile* toxins, which was utilized in **chapter 3**. Apart from RNAseq data, we also generated microarray (Hugene 1.1 ST array) data using the same RNA samples used for RNAseq. Data processing was performed as similar as possible for both techniques and only Ids common to both technologies were considered for further analysis (8018 common genes). We initially compared the fold changes of genes between RNAseq and microarray, without any statistical significance filtering (*i.e.* no p-value cut-off) and observed that the fold changes had a high correlation ($r^2 = 0.87$) (Figure 1).

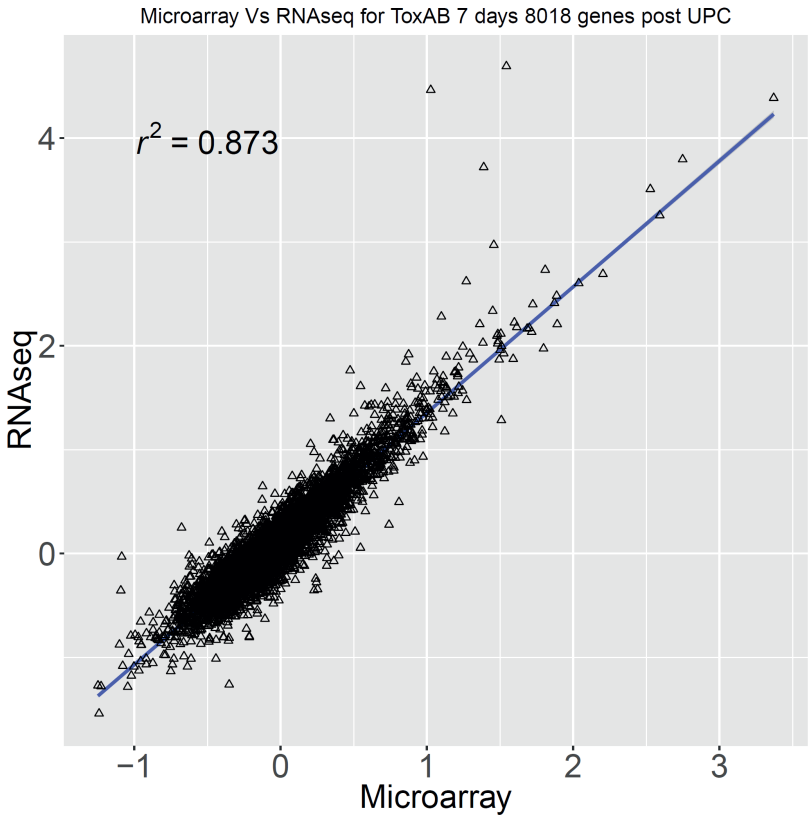


Figure 1: *Genes are plotted based on fold changes (FC). X-axis represents the microarray derived FC and y-axis represents the RNAseq derived FC. Only genes that were common to both microarray and RNAseq were chosen. The comparison shows a high degree of correlation.*

Next, only the fold changes of statistically significant (FDR p-value < 0.05) differentially expressed genes (DEG) from both RNAseq and microarray (n = 787) were compared against each other (Figure 2). The comparison showed a very high degree of correlation (r^2 = 0.94). This indicates that significant DEG in particular, show high similarity in expression changes between microarray and RNAseq.

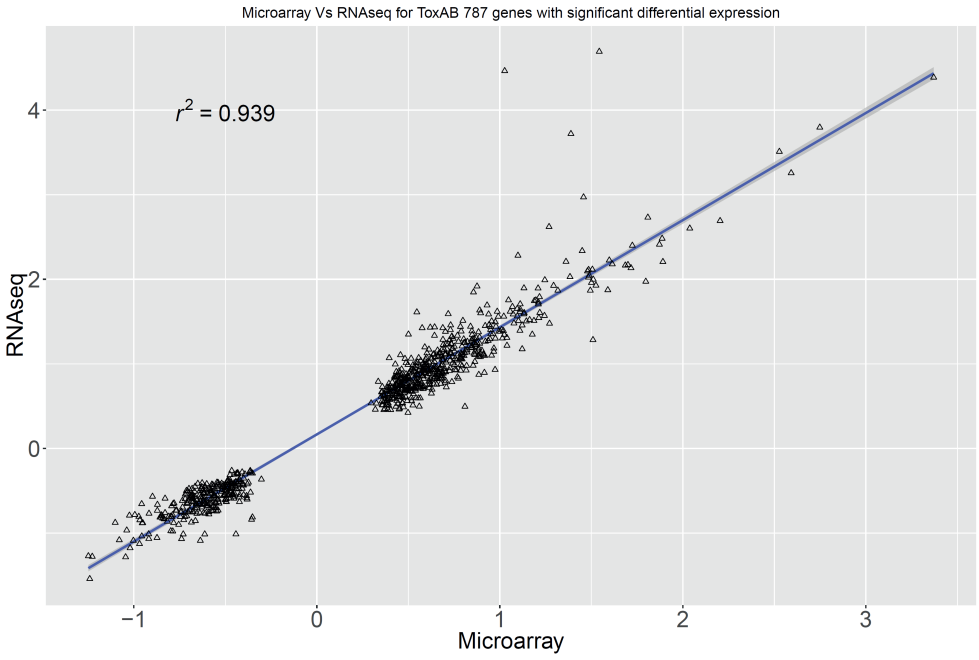


Figure 2: *Genes are plotted based on fold changes (FC). The genes are selected based on significance threshold (p-value < 0.05). Only genes that were common to both microarray and RNAseq were chosen. X-axis represents the microarray derived FC and y-axis represents the RNAseq derived FC. The comparison shows a very high degree of correlation.*

One important challenge in executing the above comparison was in mapping gene Ids used by RNAseq (Ensembl gene id) and arrays (Entrez gene Id) to identify common genes. Many entrez Ids were mapped to multiple ensembl Ids and vice versa. This would also be an issue when combining RNAseq data with array data in a compendium. Therefore, a common id that can cater to different kinds of technologies would be of great help for data fusion studies.

Identification of the relevant pathway databases

Once the significant DEGs are identified, they could be used for gene set enrichment analysis and pathway enrichment analysis. Pathway enrichment analysis using pathway databases is used often in current analysis. Khatri and colleagues have written a detailed review on the three

generations of pathway analysis approaches [38]. The first-generation pathway approaches were based on over-representation analysis while the second-generation were based on functional class scoring. Finally, the third-generation pathway databases are based on pathway-topology approaches. There are several outstanding challenges associated with pathway databases. Khatri and others have previously highlighted 5 challenges: 1. Low resolution knowledge bases; 2. Incomplete and inaccurate annotations; 3. Missing condition – and cell – specific information; 4. Inability to model and analyse dynamic responses; and 5. Lack of benchmarking [38].

The last point leads to the issues of identifying the gold standard against which other pathways could be benchmarked. A manually curated pathway database that considers experimental evidences for interaction between genes/proteins, like the ingenuity pathway analysis® (IPA , www.qiagen.com) which we have used in our analysis, could be considered for benchmarking. However, IPA comes with a cost and is not open-source. Among the freely available, open-source pathway databases, there are multiple pathway databases and a handful of them provide essentially the same information. For instance, Reactome - "*an open-source, open access, manually curated and peer-reviewed pathway database*" [39–41], KEGG – "*a database resource for understanding high-level functions and utilities of the biological system, from molecular-level information*" [42] and Wikipathways – "*an open, collaborative platform dedicated to the curation of biological pathways*" [43], all provide information about the signalling and metabolic pathways in multiple organisms and humans in particular. Although these three databases fulfil the same purpose, the pathway and sub-pathway names are not interchangeable.

In case of pathway analysis, it often leads to confusion for the uninitiated user to choose the optimal pathway database. One strategy could be to use all these four pathway databases simultaneously and check for consensus between them. However, due to lack of commonality between their names, this is difficult. The pathway commons is a great initiative, combining data from 21 pathways with Id mapping, validation of information and integration of data from these pathway databases [44]. This effort has resulted into more than 4000 pathways comprising of 1.3

million interactions spanning across 5 different types of interaction pathways (metabolic pathway, molecular interactions, signalling pathways, regulatory pathways and genetic interactions). This could be useful for users requiring open-source and freely available pathway database.

Future perspectives

Owing to the large number of high-throughput experiments conducted in the area of food and health, "foodomics" has recently been advanced as a discipline [45]. Foodomics deals with management of various food-related, high-throughput experimental data along with development and application of computational methods to advance the knowledge in food sciences with impact on human health and wellbeing as the primary objective. Being a very nascent discipline, the developments in foodomics are quite limited and will advance when more data becomes available. In this thesis, we have contributed to progress the knowledge in food transcriptomics. However, there is more scope for innovation as discussed below:

1. One of the major advantages of using microarray is the cost to benefit ratio. However, as mentioned earlier, RNAseq experiments offer more scientific benefits than microarrays for transcriptomics analysis. RNAseq is more sensitive than microarrays, allowing detection of low abundance transcripts and enables the detection of novel splicing isoforms [36]. While identification of splice isoforms has been reported in relation to tight junction regulation [46], such studies have not been reported in response to exposure of food to Caco-2. The possibility of existence of splice isoforms on exposure to luminal factors cannot be ignored. Therefore, it is of importance to probe for splice isoforms in relation to exposures to food and other intestinal luminal factors. Moreover, as the RNAseq technology improves, the costs of conducting these experiments will decrease [47]. Therefore, given the cost-benefit ratio, the foodomics as a discipline should migrate towards next generation sequencing technology. The newly available RNAseq

data from Caco-2 exposure experiments should be integrated with existing microarray data and should be used for data reanalysis. The RNAseq data, as mentioned already, offers more information compared to the microarray. It could be challenging to combine data from these two different platforms and will require new bioinformatics strategies. One way could be to focus on transcripts that are common to both technologies when it comes to data fusion. This strategy would, however, result in omission of the additional information provided by RNAseq.

2. The use of compendium data is growing as mentioned elsewhere in the thesis. The data could be used to generate gene regulatory networks, enabling identification of upstream regulators. The data could be combined with the batch correction methods prescribed in this thesis to develop classifiers that can profile future experiments into different classes based on current ones.
3. Integromics is the study of integration of different types of biological data [48] which is a sub-discipline of systems biology. In light of expansion of high-throughput data in foodomics, the next logical step would be to focus on integration of data of different types (*e.g.* genetic, transcriptomic, metabolic, proteomic, epigenomic, *etc.*). Such an integration would allow scientists to take a multi-pronged approach at understanding the impact of food on enterocytes. For instance, identification of a transcription factor involved in an important pathway is possible with transcriptomic data alone while identification of an important protein that plays a key role in the same pathway is possible with proteomic data alone. However, on integration of the two datasets, identification of the regulatory transcription factor and the exact status of the regulated gene product (*i.e.* protein) becomes simultaneously possible. This offers further insight into building a better model of the pathway. Intergromics should be carried out in a complimentary fashion to current research. While integromics offers new avenues, many challenges will exist and new tools must be developed to overcome the hurdles.
4. Multiplex networks are multilayer graphs with same nodes across the layers while the connections within layers (edges of each network layer) may vary across networks. The

interconnections between layers is usually the connection between a given node in a layer and its counterpart (same node) in another layer [19]. For instance, protein-protein interaction networks could be superseded with data from different Caco-2 exposure experiments and the resulting multiplex networks analysed for similarities (conserved network modules between exposures) and differences (modules that are unique to an exposure). Gomez and colleagues have provided a mathematical framework for analysing time dependent multiplex networks using a diffusion model [19]. Such frameworks could be applied to foodomics to check if diffusion models provide new vistas into food related network analysis.

5. While Caco-2 cells prove to be a useful model for studying enterocytes, the results have to be validated to avoid artefacts associated with a aneuploidy or chromosomal rearrangements in cancer cell lines. As such organoid models of intestinal tissues represent a potentially interesting model [49]. It has also been shown that organoid cultures from different parts of the intestine retain their location-specific expression of genes including transporters and receptors [50]. Multiple protocols have recently been proposed for culture and development of human intestinal organoids (HIO) [51,52]. While it is yet to be seen if such systems are reliable, they could be exploited to study the systemic effects of food substances. Another approach would be to use human intestinal biopsies suspended in an 'Ussing' chamber and exposed to luminal factors [53,54]. Several feasibility studies regarding the efficiency of 'Ussing' chamber as a tool for characterizing the intestinal absorption of drug and as a macromolecular permeability model have been carried out in the past and 'Ussing' chamber was found to be efficient [53,54]. However, a major limitation is that gut tissue rapidly undergoes necrosis and experiments have to be performed in short exposure time. Additionally, the collection of biopsies is an invasive procedure that needs to be performed in a hospital setting.

References

1. Tham, W. & Danielsson-Tham, M. L. *Food Associated Pathogens*. (CRC Press, 2013).

2. Wang, H. H. *et al.* Food commensal microbes as a potentially important avenue in transmitting antibiotic resistance genes. *FEMS Microbiol. Lett.* **254,** 226–231 (2006).

3. Zhang, C. *et al.* Ecological robustness of the gut microbiota in response to ingestion of transient food-borne microbes. *ISME J.* **10,** 2235–2245 (2016).

4. Bairoch, A., Cohen-Boulakia, S. & Froidevaux, C. Data Integration in the Life Sciences: 5th International Workshop, DILS 2008, Evry, France, June 25-27, 2008, Proceedings. (Springer, 2008).

5. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41,** D991–D995 (2013).

6. Kolesnikov, N. *et al.* ArrayExpress update--simplifying data submissions. *Nucleic Acids Res.* **43,** D1113-6 (2015).

7. Kangaspeska, S. *et al.* Reanalysis of RNA-Sequencing Data Reveals Several Additional Fusion Genes with Multiple Isoforms. *PLOS ONE* **7,** e48745 (2012).

8. Gilad, Y. & Mizrahi-Man, O. A reanalysis of mouse ENCODE comparative gene expression data. *F1000Research* **4,** (2015).

9. Sharma, A. Transcriptomic data reanalysis allows for a contribution of embryonic transcriptional change-induced gene expression reprogramming in transgenerational epigenetic inheritance. *Environ. Epigenetics* **2,** (2016).

10. Clarke, L. A., Botelho, H. M., Sousa, L., Falcao, A. O. & Amaral, M. D. Transcriptome meta-analysis reveals common differential and global gene expression profiles in cystic fibrosis

and other respiratory disorders and identifies CFTR regulators. *Genomics* **106,** 268–277 (2015).

11. Lamb, J. *et al.* The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science* **313,** 1929–1935 (2006).

12. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8,** 118–127 (2007).

13. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11,** 733–739 (2010).

14. Chen, C. *et al.* Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. *PLOS ONE* **6,** e17238 (2011).

15. Nygaard, V., Rødland, E. A. & Hovig, E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* **17,** 29–39 (2016).

16. Li, M., Wu, X., Wang, J. & Pan, Y. Towards the identification of protein complexes and functional modules by integrating PPI network and gene expression data. *BMC Bioinformatics* **13,** 109 (2012).

17. Jensen, P. A. & Papin, J. A. Functional integration of a metabolic network model and expression data without arbitrary thresholding. *Bioinformatics* **27,** 541–547 (2011).

18. Cline, M. S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* **2,** 2366–2382 (2007).

19. Gómez, S. *et al.* Diffusion Dynamics on Multiplex Networks. *Phys. Rev. Lett.* **110,** 28701 (2013).

20. Durzinsky, M., Wagler, A. & Marwan, W. Reconstruction of extended Petri nets from time series data and its application to signal transduction and to gene regulatory networks. *BMC Syst. Biol.* **5,** 113 (2011).

21. Solti, A., Vana, L. & Mendling, J. Time Series Petri Net Models. in *Data-Driven Process Discovery and Analysis* 124–141 (Springer, Cham, 2015). doi:10.1007/978-3-319-53435-0_6

22. Breschi, A. *et al.* Gene-specific patterns of expression variation across organs and species. *Genome Biol.* **17,** 151 (2016).

23. Hooper, L. V. *et al.* Molecular Analysis of Commensal Host-Microbial Relationships in the Intestine. *Science* **291,** 881–884 (2001).

24. Choi, J. K. & Kim, S. C. Environmental Effects on Gene Expression Phenotype Have Regional Biases in the Human Genome. *Genetics* **175,** 1607–1613 (2007).

25. Guo, G. *et al.* Serum-Based Culture Conditions Provoke Gene Expression Variability in Mouse Embryonic Stem Cells as Revealed by Single-Cell Analysis. *Cell Rep.* **14,** 956–965 (2016).

26. Margolin, A. A. *et al.* ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics* **7,** S7 (2006).

27. Faith, J. J. *et al.* Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* **5,** e8 (2007).

28. Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nat. Methods* **9,** 796–804 (2012).

29. Banf, M. & Rhee, S. Y. Enhancing gene regulatory network inference through data integration with markov random fields. *Sci. Rep.* **7,** (2017).

30. Cui, X. & Churchill, G. A. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.* **4,** 210 (2003).

31. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47 (2015).

32. Tang, Y. Non-digestible polysaccharides to support the intestinal immune barrier: in vitro models to unravel molecular mechanisms. (Wageningen University, 2017).

33. Nookaew, I. *et al.* A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Res.* **40,** 10084–10097 (2012).

34. Wang, C. *et al.* The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nat. Biotechnol.* **32,** 926–932 (2014).

35. Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K. & Liu, X. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLOS ONE* **9,** e78644 (2014).

36. Zhang, W. *et al.* Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biol.* **16,** 133 (2015).

37. Trost, B. *et al.* Concordance between RNA-sequencing data and DNA microarray data in transcriptome analysis of proliferative and quiescent fibroblasts. *Open Sci.* **2,** 150402 (2015).

38. Khatri, P., Sirota, M. & Butte, A. J. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLOS Comput. Biol.* **8,** e1002375 (2012).

39. Milacic, M. *et al.* Annotating cancer variants and anti-cancer therapeutics in reactome. *Cancers* **4,** 1180–1211 (2012).

40. Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res.* **42,** D472-477 (2014).

41. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **46,** D649–D655 (2018).

42. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45,** D353–D361 (2017).

43. Slenter, D. N. *et al.* WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* **46,** D661–D667 (2018).

44. Cerami, E. G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* **39,** D685-690 (2011).

45. Capozzi, F. & Bordoni, A. Foodomics: a new comprehensive approach to food and nutrition. *Genes Nutr.* **8,** 1–4 (2013).

46. Clayburgh, D. R. *et al.* A Differentiation-dependent Splice Variant of Myosin Light Chain Kinase, MLCK1, Regulates Epithelial Tight Junction Permeability. *J. Biol. Chem.* **279,** 55506–55513 (2004).

47. Zhang, J., Chiodini, R., Badr, A. & Zhang, G. The impact of next-generation sequencing on genomics. *J. Genet. Genomics Yi Chuan Xue Bao* **38,** 95–109 (2011).

48. Venkatesh, T. & Harlow, H. B. Integromics: challenges in data integration. *Genome Biol.* **3,** reports4027.1-reports4027.3 (2002).

49. Fatehullah, A., Tan, S. H. & Barker, N. Organoids as an *in vitro* model of human development and disease. *Nat. Cell Biol.* **18,** 246–254 (2016).

50. Middendorp, S. *et al.* Adult stem cells in the small intestine are intrinsically programmed with their location-specific function. *Stem Cells Dayt. Ohio* **32,** 1083–1091 (2014).

51. Miura, S. & Suzuki, A. Generation of Mouse and Human Organoid-Forming Intestinal Progenitor Cells by Direct Lineage Reprogramming. *Cell Stem Cell* **21,** 456–471.e5 (2017).

52. Múnera, J. O. & Wells, J. M. Generation of Gastrointestinal Organoids from Human Pluripotent Stem Cells. *Methods Mol. Biol. Clifton NJ* **1597,** 167–177 (2017).

53. Wallon, C., Braaf, Y., Wolving, M., Olaison, G. & Söderholm, J. D. Endoscopic biopsies in Ussing chambers evaluated for studies of macromolecular permeability in the human colon. *Scand. J. Gastroenterol.* **40,** 586–595 (2005).

54. Rozehnal, V. *et al.* Human small intestinal and colonic tissue mounted in the Ussing chamber as a tool for characterizing the intestinal absorption of drugs. *Eur. J. Pharm. Sci. Off. J. Eur. Fed. Pharm. Sci.* **46,** 367–373 (2012).

English Summary

The connection between food and human health needs no introduction, but apart from food, there are multiple other luminal factors that alone or in combination with others can impact on enterocyte function and thus play a role in health. These interactions maybe at different levels, but primarily at the molecular level. The intestinal epithelial cell line Caco-2 has been used extensively as a model to investigate molecular interactions between different luminal factors and enterocytes. More recently attention has been given to interactions of more complex food extracts with the aim of gaining more insights into the combined effects of food components, as would occur *in vivo*.

In **chapter 1**, The importance of studying enterocytes, food interactions and the Caco-2 as an enterocyte model system is discussed. Previous research on *Clostridium difficile* toxins and their interaction with enterocytes as well as systems biology and related approaches are described in detail. The chapter ends with a brief introduction on the key research questions and the content of the chapters which follow.

In this thesis, we have generated a data compendium by collecting transcriptome data (largely from microarray experiments) pertaining to Caco-2 cells exposed to luminal factors from in-house experiments and public databases (**Chapter 2**). Initially, the data compendium was used to develop a Caco-2 specific protein-protein interaction network. Then, we addressed the issue of identifying pathway specific reporter genes for qPCR experimental validation. To this end, we developed a statistical method called differential expression correlation analysis (DECA), which is designed to mine knowledge from the Caco-2 data compendium. The method utilises differential expression values of genes in the compendium combined with limited knowledge of pathways of interest to identify reporter genes. This method was further used to predict genes that belonged to AhR and Nrf2 mediated stress response pathways and was experimentally validated using Caco-2 cells exposed to coffee extracts.

In **Chapter 4**, the Caco-2 data compendium was utilized for identification of food substances that may mitigate the effects of *C. difficile* toxins on small intestinal enterocytes. This was combined with Caco-2 microarray data obtained from Caco-2 cells exposed to *Clostridium difficile* toxins (toxin A and toxin B) and toxoids. The identification of possibly beneficial foods was carried out using multivariate techniques such as principal component analysis (PCA). Blackcurrant of Ben Finlay cultivar was found to be the most beneficial food among the food substances used in the compendium and was experimentally verified. It was found to help maintain the epithelial barrier and also in preventing the translocation of *C. difficile* toxins from the apical side to the basolateral side. Additionally, we also tested the efficacy of strawberry (Sabrina), yellow onion, white onion and Galacto-oligosaccharides (GOS) and found, in accordance with PCA results, that while strawberry and yellow onion were moderately effective against the toxin translocation, white onion and GOS had almost no effect.

We delved further into an investigation of the impact of *Clostridium difficile* toxins on the miRNA expression of the colonic enterocytes and probed the role of miRNAs in regulating toxin-induced changes in mRNA expression (**Chapter 3**). miRNA-mRNA interaction was studied with the help of public database, miRWalk 2.0 and the network analysis tool, Cytoscape. We performed pathway analysis with the data obtained and found a role for miRNAs in several pathways that are affected by *Clostridium difficile* toxins in Caco-2 cells.

Finally, to enable data fusion of experiments that have low and varying sample sizes, we developed a batch effect mitigation protocol (**Chapter 5**). The method is a combination of ratio-based methods and median rank scores. The method was tested on the controls-only data derived from the Caco-2 compendium and was shown to mitigate batch effects. It was further tested on arthritis patient sample data, applied to the Caco-2 compendium data and shown to be efficient at batch effect mitigation.

In **chapter 6**, the results of the thesis are discussed including the limitations and future perspectives for the advancements in the field of foodomics.

Samenvatting

Het verband tussen voeding en gezondheid in mensen behoeft geen introductie. Een goed werkend darmepitheel is daarbij onontbeerlijk. De enterocyten die het grootste deel van het darmepitheel uitmaken zijn daarbij heel belangrijk. Behalve voedsel zijn er ook andere luminale factoren in de darm die alleen of in combinatie met andere componenten invloed kunnen hebben op de functie van de enterocyten en dus een rol spelen bij de gezondheid. In vivo onderzoek naar de interactie tussen luminale factoren en darm epitheliale cellen is moeilijk en zeer onprettig voor deelnemers aan het onderzoek. De epitheliale darmcellijn Caco-2 wordt daarom veel gebruikt als model om dit soort interacties te onderzoeken, waarbij vaak gebruik gemaakt wordt van transcriptomics (genexpressieanalyse) om de reactie van cellen te analyseren. In toxicologisch en farmacologisch onderzoek ligt de focus vaak op een of enkele stoffen, terwijl in het voedingsonderzoek recent steeds meer aandacht is voor de interacties van complexere voedingsextracten en -producten met als doel meer inzicht te krijgen in de gecombineerde effecten van voedsel, zoals ook de exposure in de darmen is.

In **hoofdstuk 1** wordt het belang van het bestuderen van enterocyten, interactie tussen voedsel en het darmstelsel en het Caco-2 als een enterocyten modelsysteem besproken. Tevens wordt eerder onderzoek naar Clostridium difficile (C. difficile) toxines en de interactie met enterocyten beschreven, evenals systeembiologie en verwante mathematische benaderingen van onderzoek. Het hoofdstuk eindigt met een korte inleiding over de belangrijkste onderzoeksvragen en de inhoud van de hoofdstukken die volgen.

In dit proefschrift hebben we een data compendium gegenereerd door transcriptoom data te verzamelen van Caco-2 cellen blootgesteld aan allerlei luminale factoren. Deze data waren afkomstig van eerder uitgevoerde experimenten bij Wageningen - Food & Biobased Research of afkomstig van openbare databases (**hoofdstuk 2**). Aanvankelijk werd het compendium gebruikt om een Caco-2/enterocyte-specifiek genexpressie- en interactienetwerk te

ontwikkelen. Vervolgens hebben we de data gebruikt voor het identificeren van responderende pathway-specifieke reportergenen in Caco-2, welke zijn gevalideerd in nieuwe Caco-2 experimenten met qPCR. Voor het selecteren van dergelijke reportergenen hebben we een statistische methode ontwikkeld, differentiële-expressiecorrelatieanalyse (DECA) genaamd, die is ontworpen om kennis uit het Caco-2-data compendium te ontginnen. De methode om reportergenen te identificeren maakt gebruik van differentiële expressiewaarden van genen in het compendium in combinatie met kennis van de genen en pathways van interesse. Deze methode hebben wij toegepast om genen te voorspellen die behoren tot door AhR en Nrf2 gemedieerde responsen en dit is vervolgens experimenteel gevalideerd met behulp van Caco-2-cellen die waren blootgesteld aan koffie-extracten.

In **hoofdstuk 3 en 4** werd het onderzoek verder toegespitst op de schadelijke effecten van C. difficile toxines op enterocyten van de dunne darm. Genexpressieanalyses (met behulp van microarrays) werden uitgevoerd om de reactie van Caco-2 blootgesteld aan C. difficile-toxinen (toxine A en toxine B) en toxoïden te onderzoeken. Deze genexpressie-gegevens werden gecombineerd met het bovengenoemde data compendium met als doel luminale kandidaat-factoren te identificeren met een mogelijk heilzame werking tegen de negatieve impact van deze toxines (**hoofdstuk 4**). De analyses werden uitgevoerd met behulp van multivariaat technieken zoals principale component analyse (PCA). Zwarte bessen, hier onderzocht op basis van cultivar Ben Finlay, bleken de beste kandidaat te zijn binnen de beschikbare dataset. Het neutraliserende effect van het zwarte bes extract op de schadelijke impact van de toxines, kon ook experimenteel geverifieerd worden in Caco-2 cellen. Specifieke voedingsproducten lijken dus de epitheliale cellen te kunnen beschermen en translocatie van C. difficile toxines van de apicale zijde naar de basolaterale zijde (deels) te kunnen voorkomen. Daarnaast hebben we ook de heilzame werking van aardbei (Sabrina), gele ui, witte ui en galacto-oligosacchariden (GOS) getest. In overeenstemming met de PCA-resultaten, werd gevonden dat aardbeien en gele ui matig effectief waren tegen de toxine-geïnduceerde effecten en dat witte ui en GOS bijna geen effect hadden.

In **hoofdstuk** 3 doken we verder in de interactie tussen C. difficile toxines en enterocyten door de genexpressieveranderingen met behulp van RNAseq te onderzoeken. Tevens werd de rol van microRNA's (miRNA's) bij het reguleren van toxine-geïnduceerde veranderingen in mRNA-expressie onderzocht. miRNA-mRNA-interactie werd bestudeerd met behulp van de openbare database miRWalk 2.0 en de netwerkanalysetool Cytoscape. We voerden pathway-analyses uit met de verkregen data en vonden een rol voor miRNA's in verschillende routes en pathways die worden beïnvloed door C. difficile-toxinen in Caco-2-cellen.

Genexpressie onderzoek, gecombineerd vanuit verschillende (cel-)experimenten en onderzoek uit verschillende laboratoria, kent vele biologische en technische variabele factoren waardoor het vaak moeilijk is kleine effecten uit een dergelijke data compendium te selecteren. Om toch dergelijke analyses mogelijk te maken hebben we een batch-effect-mitigatie-protocol ontwikkeld (**hoofdstuk 5**). De methode is een combinatie van ratio-gebaseerde methoden en mediane scores. De methode is getest op een steekproef-set afgeleid van het Caco-2-compendium en bleek inderdaad batcheffecten te verminderen. De methode werd verder getest op een dataset van artritispatiëntmonsters waarbij aangetoond werd dat de methode efficiënt is bij partiële effectmitigatie.

In **hoofdstuk 6** worden de resultaten van het proefschrift besproken, inclusief de beperkingen en toekomstperspectieven voor de vooruitgang op het gebied van 'omics' technologieën in het voedings- en gezondheidsonderzoek (foodomics).

Acknowledgements

My PhD friends often told me that the PhD thesis is like a baby and the time of research is the gestation period (quite a long one!!). I do not agree to this view since to me, a PhD thesis is a dot in the painting of human knowledge. Moreover, a thesis in biology is like a dot that lies on the outline of the painting. The outline helps the humans to know better about lives (including their own) and sets the basis for the colours (other research) that make the painting lively. To make this one dot, I needed ample amount of help from many fellow humans without whom I could not have reached the canvas, nor have held my brush to make the dot. In sanskrit, one of the classical languages of the world, extracted from the Taitriyo Upanishad, there goes a saying:

मातृदेवो भव, पितृदेवो भव, आचार्यदेवो भव

The above saying explains that one starts with paying their respects to their mother, followed by their father and finally the teachers. To this list, I would also like to include my friends and kins.

I would like to begin my acknowledgement note starting with my parents, without whom I wouldn't have made it till here. My mom has always availed me with the best care and my dad has provided me with ample support just as the thirukkural (estimated to be written in early 1st century B.C.E.) goes:

தந்தை மகற்காற்று நன்றி அவையத்து

முந்தி இருப்பச் செயல்

which means "The benefit which a father should confer on his son is to enable him to precede in the assembly of the learned". It was my dad's support and encouragement that helped me dream big. I do not have enough words to thank my parents or show my respect and gratitude. Next comes my sister (**Akshaya**) and my fiance (**Sindhuja**), who have always been my greatest friend and companion at all times. My conversations with **Akshaya** helped me get out of my mental stress all these years. It was fun acting like the big brother with her. The support provided by **Sindhuja** during the preparation of this thesis is immeasurable. I am eagerly looking forward to our wedded life and the years to follow. I am sure I've many more years to show my love, respect and gratitude towards all of them and hence would like to leave it here. Next, my **grandparents'** care has been of tremendous importance and I am extremely grateful for their immense role in building up my character. I particularly learnt the art of perseverance and courage to face challenges from them. I would like to thank my entire family (my aunts (**athais**

and chithi), my uncles (**chithappa and athimbers**) and my cousins), who have nourished me and encouraged me all these years. I have the pleasure of having spent ample amount of time in each of their houses and learnt a great deal from them. I would particularly like to thank my aunt (**Savithiri**) and my uncle (**Kalyanram**) for their immense support during all these years, especially during my graduation years. Although this uncle of mine, is no more mortal with us, I'll always remember his inspiring speeches and all the philosophical and political discussions that I had with him. I would also like to thank my other aunts **Vimala** and **Meenakshi**, along with my uncle **Ramakrishnan** for their immense role during my childhood years and beyond. Their contributions to my wellbeing is undeniable and have equally helped me shape my character. Another key person that I would like to thank is my cousin, **Karthik**. He is the first scientist from my dad's side and has been a big inspiration. He has helped me in many ways all these years, including his monetary help for my TOEFL exam, recommending Europe as a fine destination for doctoral studies and helping me with finding a job after my graduation. I am grateful to you for all these and many more to come, anna. I would also like to thank my other cousins, **Kirthika and Keerthi** with whom I spent many formative years together. I am equally thankful to **Meghana and Miliind** for all the fun we had together, whenever we could catch up.

Further, I would like to thank my gurus, beginning with the ones in the Netherlands. I would take this opportunity to thank my thesis advisors, **Dr. Jurriaan Mes, Dr. Edoardo Saccenti, Dr. Nicole de Wit and Prof. Dr. Jerry Wells**. **Jurriaan** has been very supportive and has helped me think better through his critical analysis of my ideas. He has always had his doors open for all my questions and never made me feel the hierarchy. I have always felt like working with a friend, under him. **Edoardo** has been my go to for any statistical and computational biology issues while **Nicole** had been supporting me with answering my biological questions. Both have been of immense help for completing my papers and my thesis. I am very grateful to **Jerry** for his support and his inputs during our monthly meetings. I cannot thank **Maria** (**Dr. Maria Suarez-Diez**) enough, for all her contributions towards making my research fruitful. She has helped and guided me in times of extreme needs and has been a key pillar for this thesis coming to a fruition. To me, she was the scientist that always had a solution. Her calmness in the face of crisis of (sometimes) my creation was her hallmark signature. I cannot thank enough, my Italian collaborators, **Prof. Dr. Gastone Castellani, Dr. Daniel Remondini and Dr. Enrico Giampieri**, who were kind enough to host me and guide me with their expertise in statistical physics. I was able to gain lot of new knowledge in statistics, particularly from **Enrico** and was highly impressed and motivated by his knowledge. I would also like to thank him for all the engaging conversations we had over my stay there and especially for introducing me into Wing Chun kung-fu.

I would like to thank the thesis committee members (my esteemed opponents in particular) for taking their valuable time to read through this thesis and evaluate it. I would also take this opportunity to thank every teacher who has ever taught me, right from my childhood to the present day. I do not know for sure if they envisaged that I would end up becoming a scientist, but, I hope they are proud of their teaching outcome. I would particularly like to remind myself of the contributions in inspiring me by **Prof. Dr. K. R. Pardasani, Dr. Arijit Bhattacharyay, Dr. Arnab Ghosh, Dr. Patrick Nelson, Dr. Sairam Kalapatapu, Prof. Dr. Santiago Schnell, Mr. Rajini Kumar and Dr. P. K. Dash**. Some of them may not recognize me or may not know their influences towards inspiring me, but I always feel grateful to them for their inspirations in their own unique ways.

No research goes on without the lab and for a good work output, the lab environment needs to be positive, cohesive and happy. I would like to thank all the members of FQHE (Jurriaan's lab) for their immense contribution in keeping up the positive attitude and environment in the lab. I would like to thank all my lab mates (**Yong-fu, Lonnekke, Marit** in particular) for making my time in the lab comfortable, engaging in interesting conversations and helping me integrate into the lab. I thank the lab technicians (**Renata, Els, Shanna and Monic** in particular) for their help in conducting experiments and helping me ease into cultural aspects of Netherlands and handle the lab affairs. I cannot forget to thank **Coen**, for his wonderful wisdom and jovial nature that kept even difficult things going smooth in the lab. His sarcasm is definitely infectious and I did enjoy watching cricket matches (explaining him along the way) and skating events (winter Olympics). I am pretty sure he doesn't remember the rules of cricket, but hey, at least he tried!! I will not forget participating in the We-day events. I am thankful to **Harry and Judith** for the energy they brought with them during these events! Had so much to learn from them. I would also like to thank the members of IPOP Systems Biology Virtual-Gut team (**Neeraj, Anupama and Bastiaan**) for all the interesting discussions at coffee breaks.

I believe that a sound achievement requires a sound mental health which in fact requires a happy environment. My friends have had two important contributions to my life. They have always been my instructors and my happiness keepers. Every discussion I've had with each of them are wisdom for life and needless to say, if not for them, my life would not have been as fun as it had been. Although, I wish to thank each of my friends personally, I would like to mention a few that have particularly helped me come this far. As before, I would like to start with the people in the Netherlands and would like to thank the two of my oldest housemates, **Soumya** (also my paranymph) and **Neeraj**. It was a pleasure cooking with them and discussing varied topics, extending from science and philosophy to politics and sports. I would like to thank them for their immense help in all walks of my life. I am also thankful to their better halves', **Raka and**

Arathi. I thank you two for all those interesting discussions we always held during our catch-ups. Thank you, **Raka**, in particular, for proof-reading my thesis chapters and thank you, **Arathi**, for helping me through many things in my life and irritating me in things beyond. If not for those irritations, life would have certainly been boring.

Next, I would like to thank the "Tamil Gang". This includes **Rohan**, **Bharath**, **Kaushik**, **Surabhi**, **Janardhan**, **Nitin**, **Moses**, **Ananth**, **Vani, Athul** and **Adhitya**. I should say, I thoroughly enjoyed the company of these guys and they certainly made me feel younger. There have been countless dinners, board games and discussion with these people, all of which helped me ease into my life during the course of my PhD. Thank you, **Rohan**, for all those sporty adventures we ventured into together and I wish you the best for the years to come. Thank you, **Kaushik**, for those amazing discussions I had with you during the coffee breaks at Axis - Z. It had always been a pleasure discussing ideas with you and I wish you the best for your PhD. Thank you **Janardhan**, for the interesting political updates you provided me with and I wish you the best for finding the suitable PhD course. **Bharath**, thank you for making me fatter with your cooking and thank you for teaching me (in a very weird way) how to run on a treadmill!! Thank you, **Nitin, Surabhi and Vani**, for all those nice cooking experiences together and thank you **Nitin** for your insights into music that I never had before.

I would then like to thank my present housemates **Tanya** and **Abhirup** for their immense help in many things in my life. **Abhirup** has stood through many things in my life and I would like to personally thank him for all that. Then I take this special opportunity to thank **Dr. Suraj Jamge**. **Suraj**, I still remember you telling me that I will cross over the finish line in a matter of few weeks and I'm happy that I could. I would also like to thank you for introducing me into SCUBA diving and board games. You, **Rohan** and **Sacha** have been such an inspiration for me in the area of sports. Then, I would like to thank "the pariah gang" which includes **Liana, Kamesh, Artemis, Katarina, Yasmeen, Tom and Jess**. Their contribution in making me feel at home in the Netherlands is immense. I have always enjoyed telling stories on Hinduism to **Liana** and with her questions she has made me think farther than what I have thought. I totally owe her a good portion of my knowledge on Hindu philosophy. **Kamesh**, all those squash games were certainly a joy to play against you and beat you! Not anymore, I think! Thank you, 'the pariahs' for all those intense political and cultural discussions we've had and for all those amazing recipes from across the globe that you made me eat.

Next, I would like to thank 'the climbers gang', for being my first set of friends from foreign soil. Thank you, **Sacha, Suvi, Victor, Kim, Rita and Monica**. I completely relished that dinner we had together after our first outdoor climb and will never forget that day. I was always happy to dine with you guys, but sad that we couldn't climb much after that. I am extremely happy that

we are in touch till this day and hope to continue to do so. **Sacha**, thanks for everything. You've been extremely amusing to hang out with and a crazy inspiration for me. I still cannot believe that you have travelled half the World already and that you're taking a group of tourists into India as a guide (total sarcasm, you see!!). I will not forget staying at your house during my adventures in that seed company (irony) of my choosing. Your house was like an oasis in the middle of a desert. Thank you, **Monica,** for leading the way into PhD and I will always remember that trip to Norway with you. Best wishes to **Victor**, **Suvi** and **Rita** for your respective studies and my best wishes to **Suvi - Eric** and **Victor - Kim** for a happy future together.

I would finally like to thank my friends from India, **Anand, Naveen, Vineel, Thalai (Prabhakaran), Manoj and Poochi (Ram prasad)** for sticking out for me in all times of needs. You guys made my B. Tech. life amazing and were patient enough to listen to my rants there after (particularly during my MS). I know you guys are there for life and would like to keep it going for all the years to come. I would also like to thank **Karthik, Vibin, Manjunath and Rajarajan** for staying in touch through these years. I am extremely thankful to **Karthik and Vibin** for all the wisdom over the years. I have always been inspired by the way **Karthik and Rajarajan** thought through things and the way **Vibin** handled issues. I've learnt so much from you all and have applied it over the years. I would also like to thank **Sanjeev and Madhuri,** the closest of my friends from my class during under graduation days, for all those happy memories during college and the interesting rants about our lives we exchanged over the years after. I know that all you guys are for a lifetime and am happy to have stayed in touch with you all these years. I would also like to thank my childhood friends **Aadhith, Harish, Karthik, Ramki and Ajay** for their involvement in all walks of my life for all these years. I wish to stay in contact with you guys more often. I am particularly thankful to **aadhith** for all our memories over the years. I cannot forget the time I spent with your family (**Sankaran** uncle, **Abhirami** and **Gowri** aunty, in particular) and hope to continue to do so. I am extremely thankful to your parents for their encouragements!

I would finally like to thank all those who have helped me succeed in every small way possible and conclude this section. I'm sorry if I have missed someone in this list! It was not at all intentional!!

About the author

Prashanna Balaji Venkatasubramanian was born in Chennai in India and grew up in Trichy. He completed his first ten years of schooling in Trichy and finished his high school at FiiT-Jee, Hyderabad. In 2007, he joined the Bachelor of Technology program, with specialization in Bioinformatics at Maulana Azad National Institute of Technology, Bhopal. He graduated as an engineer in 2011 with his major thesis on application of Flux Balance Analysis in Melanogenesis pathway.

He started his Master of Science in Bioinformatics at the University of Michigan, Ann Arbor in the Fall of 2011. He graduated the MS by coursework with focus on Complex Systems and Systems Biology in 2012. Later in May 2013, he started his PhD at Wageningen University and Research with focus on interdisciplinary research in Systems Biology and food sciences in the group of Dr. Jurriaan Mes. During his PhD, the author developed methods to analyze large transcriptomics data collected from Caco-2 exposure experiments. The author also investigated and identified foods (berries) that may mitigate the ill effects of Clostridium difficile on enterocyte barrier integrity using Caco-2 as a model system.

Aside from the research, travel and adventure sports are his major interests. He plays football, squash and floorball. Moreover, he climbs for sport and is also interested in SCUBA diving. Currently he is working as a postdoctoral researcher in Systems Vaccinology at the Comparative Genomics group in Radboud UMC, Nijmegen, headed by Prof. Martijn Huijnen. The author's work lies in the area of predicting biomarkers associated with Bordetella pertussis infection for further vaccine development.

EDUCATION AND TRAINING

Basic Courses	**year**
WIAS Introduction Course, 2014	2014
Course on philosophy of science and/or ethics	2013
Basic Courses Credits	***3***

Scientific Exposure: International conferences, seminars and workshops	**year**
Basics of parameter estimation, Wageningen held for a day	2013
ISGSB (Poster), Durham, UK held for 5 days	2014
Foodomics (Poster), Cesena, Italy held for 2 days	2015
SBNL, Lunteren held for 2 days	2016
SBNL, Lunteren held for 2 days	2017
Talk at TU Dresden dermatology institute (2018 - Dresden, Germany)	2018
Scientific Exposure credits	***8***

In-Depth Studies	**year**
Algorithms for Biological Networks, Delft University of Technology(2nd edition)	2014
Analysis of Microarray and RNAseq data using R, Erasmus MC, Rotterdam	2013
Quantitative and Predictive Modelling	2015
ECAS course on Statistical Analysis of Network Data, Herrsching, Germany	2015
PhD students' discussion groups	2013-2017
BRD-31306 Systems and Control theory	2014
In-Depth Studies credits	***13***

Professional Skills Support Courses	year
Course Techniques for Scientific Writing	2015
Course Techniques for Scientific Writing	2015
PhD Competence assessment or Job assessment	2014
Professional Skills Support Courses Credits	***3***

Research Skills Training	**year**
Preparing PhD research proposal	2013
External training period (Sept - Nov 2016, 3 months at Bologna, Italy)	2016
Research Skills Training Credits	***8***

Didactic Skills Training	**year**
Molecular Systems Biology (43 hrs including exam supervision)	2015
Molecular Systems Biology (43 hrs including exam supervision)	2016
MSc major thesis (Architha Ellappalayam, MSc Bioinformatics)	2016
BSc thesis (Judith Cantó, BSc)	2017
Didactic Skills Training credits	***6***

Education and Training Total Credits	**40**

List of Publications

Peer reviewed publications

Venkatasubramanian, P. B. et al. Use of Microarray Datasets to generate Caco-2-dedicated Networks and to identify Reporter Genes of Specific Pathway Activity. Scientific Reports 7, 6778 (2017).

Prashanna Balaji V., Anvita Gupta Malhotra and Khushhali Menaria. Flux Balance Analysis of Melanogenesis Pathway. International Journal of Soft Computing and Engineering (IJSCE). March 2012; 2(1): 162-170

Expected publications

Venkatasubramanian P. B., Els Oosterink, Monic Tomassen, Maria Suarez-Diez, Edoardo Saccenti, Jurriaan Mes, Nicole de Wit; Identification of food compounds to attenuate the cytopathic effects of Clostridium difficile toxins using transcriptome data (Submitted to BMC Genomics).

Venkatasubramanian P. B., Renata Ariens, Els Oosterink, Edoardo Saccent, Jurriaan J. Mes and Nicole de Wit; Exploring the role of miRNAs in regulation of the Caco-2 cell transcriptional response to Clostridium difficile toxins (Manuscript in final stages of preparation).

Venkatasubramanian P. B., Enrico Giampieri, Jurriaan Mes, Edoardo Saccenti, Gastone Castellani, Daniel Remondini; A normalization protocol to mitigate batch effects and allow comparison of the effects of different treatments on the same biological system (Manuscript in final stages of preparation).

G Toydemir, L Loonen, **P B Venkatasubramanian**, E Capanoglu, J Wells, N De Wit, J Mes. Coffee-mediated in vitro activation of coupled AhR-Nrf2 pathways: Altered effects of single coffee and combined benzo[α]pyrene/coffee treatments. (Manuscript in final stages of preparation).

Conferences (oral and poster presentation)

Venkatasubramanian P. B., Enrico Giampieri, Jurriaan Mes, Gastone Castellani, Daniel Remondini. Cross comparison of large microarray dataset: Batch effects problems and solutions. Oral presentation at: BioSB 2017, Dutch Bioinformatics & Systems Biology conference; 2017 Apr 4-5; Lunteren, Netherlands

Venkatasubramanian P. B., Suarez-Diez M, Saccenti E, Mes J. Caco-2 specific Gene targets Identification. Poster session presented at: Food to life, 4th International Conference on Foodomics; 2015 Oct 8-9; Cesena, Italy

Venkatasubramanian P. B., Suarez-Diez M, Saccenti E, Mes J. Caco-2 Protein-Protein network development. Poster session presented at: From cell to organism, ISGSB; 2014 Sept 2-5; Durham, UK

Prashanna Balaji V. and Neel Mehta. Riboswitches as a novel drug target. Poster session presented at: APBC2010, The eighth Asia Pacific Bioinformatics Conference, 2010 Jan 18-21; Bangalore, India

The research described in this thesis was financially supported by the Dutch Ministry of Economic Affairs within the Systems Biology programme ‘Virtual Gut’, KB-17-003.02-021. Financial support from the Wageningen Institute of Animal Sciences for printing this thesis is gratefully acknowledged.

Cover design by the author and Tatiana Blokhina

Printed by ProefschriftMaken || DigiForce