
Bioinformatic pipelines to reconstruct and analyse intercellular and host- microbe interactions affecting epithelial signalling pathways

Lejla Potari-Gul

A thesis submitted for the degree of Doctor of Philosophy

University of East Anglia
Earlham Institute

Supervisors:

Dr. Tamas Korcsmaros

Dr. Janette Jones

Dr. Neil Hall

United Kingdom

July 2022

©This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that use of any information derived there-from must be in accordance with current UK Copyright Law. In addition, any quotation or extract must include full attribution.

Abstract

The epithelium segregates microorganisms from the immune system through tightly connected cells. The epithelial barrier maintains the integrity of the body, and the microbiome influences this through host-microbe interactions. Therefore its composition has an impact on the host's physiological processes. Disruption in the microbiome composition leads to an impaired epithelial layer. As a consequence, the cell-cell interactions between the epithelium and immune cells will be altered, contributing to inflammation. In this thesis, I examined the interconnectivity of the microbiome, epithelium and immune system in the gastrointestinal tract focusing on the oral cavity and gut in healthy and diseased conditions.

I combined multi-omics data with network biology approaches to develop computational pipelines to study host-microbe and cell-cell connections. I used network propagation algorithms to reconstruct intracellular signalling and identify downstream pathways affected by the altered microbiome composition or cell-cell connections. I studied inflammation-related conditions in the oral cavity (periodontitis) and gut (inflammatory bowel disease (IBD)) to reveal the contribution of interspecies and intercellular interactions to diseases. I inferred host-microbe protein-protein interaction (HM-PPI) networks between healthy gum-/periodontitis-related bacteria communities and epithelium, and found altered HM-PPIs during inflammation. I connected the epithelial cells to dendritic cells and identified the Toll-like receptor (TLR) pathway as a potential driver of the inflammation in diseased gingiva. While in the oral cavity I focused on complex microbial communities and their impact on one cell type, I discovered the direct effect of gut commensal bacteria on several immune cells in IBD. This study observed the cell-specific effect of *Bacteroides thetaiotaomicron* on TLR signalling.

The pipelines I developed offer potentially interesting connections that aid detailed mechanistic insight into the relationship between the microbiome, epithelial barrier and immune system. These systems-level analysis tools facilitate the understanding of how microbial proteins may be of therapeutic value in inflammatory diseases.

Access Condition and Agreement

Each deposit in UEA Digital Repository is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the Data Collections is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form. You must obtain permission from the copyright holder, usually the author, for any other use. Exceptions only apply where a deposit may be explicitly provided under a stated licence, such as a Creative Commons licence or Open Government licence.

Electronic or print copies may not be offered, whether for sale or otherwise to anyone, unless explicitly stated under a Creative Commons or Open Government license. Unauthorised reproduction, editing or reformatting for resale purposes is explicitly prohibited (except where approved by the copyright holder themselves) and UEA reserves the right to take immediate 'take down' action on behalf of the copyright and/or rights holder if this Access condition of the UEA Digital Repository is breached. Any material in this database has been supplied on the understanding that it is copyright material and that no quotation from the material may be published without proper acknowledgement.

Acknowledgements

I would like to thank all the people who contributed to my thesis in the last five years. I got much scientific and scientific support from my supervisors, colleagues, friends and family.

From a scientific perspective, I am grateful for the Korcsmaros group, especially for my primary supervisor Tamas Korcsmaros for all the support, encouragement and advice that I have got not only during the PhD but in the last eight years. Particularly, I would like to thank Dezso Modos and Marton Olbei for their great advice and meetings that helped me to finish my thesis. I am thankful to Denes Turei, David Fazekas and Matthew Madgwick for helping me to learn and improve my Python and R programming skills. Additionally, I appreciate the contribution of Isabelle Hautefort, Martina Poletti, Polina Kornilova, Luca Csabai, Balazs Bohar, and also former group members, Amanda Demeter, Padhmanand Sudhakar to the project. It was a pleasure to work with all of you over the last 5 years.

I would like to acknowledge my secondary supervisor, Neil Hall, for his aid in the project and for establishing the connection with the iCASE partner. The project could not have been carried out without the support of Unilever R&D in Liverpool. Special appreciation for my iCASE supervisor, Janette Jones, who took care of me during the 3 months internship in person and even virtually due to the pandemic. I would like to thank Shruti Malviya to coordinate the virtual internship and meetings in the last year of the project. I am appreciative to Jenny Pople, Andrew Cawley, Fei-Ling Lim, Barry Murphy and Richard Skinner but also to Rebecca Ginger and Patrick Warren, as former members of Unilever, for all the meetings, advice and scientific assistance that improved the project.

I am also thankful to Julio Saez-Rodriguez and his research group at the BioQuant Institute in Heidelberg to have me for 6 weeks and for helping me to learn network modelling. Last but not least, I am grateful to the Powell group at the Imperial College London for the meetings and for establishing connections for future collaborations.

From a personal perspective, I would be remiss in not mentioning my parents, brother, husband, son and friends. They kept me motivated, encouraged and believed me, even from far away they were always there to listen to my complaints and supported me in the difficult times. I could not have done it without you.

I am grateful for the NRP Graduate School Office to help with my questions, when I felt lost in the paperwork, for the administrative assistance of the Business Support at the Earlham Institute and finally, for the support of the Biological Sciences Research Council (BBSRC) for funding to undertake this PhD.

Table of contents

Abstract	2
Acknowledgements	3
Table of contents	4
List of abbreviations	7
List of Figures	10
List of Tables	12
List of Supplementary materials	13
List of peer-reviewed publications	14
Chapter 1: General introduction	16
1.1 <i>Preface</i>	16
1.2 <i>Gastrointestinal tract</i>	18
1.2.1 Oral cavity	19
1.2.2 Intestine	20
1.3 <i>The human microbiome</i>	27
1.4 <i>External signals affecting the epithelium</i>	30
1.4.1 Cell-cell interactions	30
1.4.2 Host-microbe interactions	34
1.5 <i>Molecular background of inflammation</i>	35
1.5.1 Cytokine secretion	36
1.5.2 Pathogen-associated inflammation	37
1.6 <i>Omics data</i>	40
1.6.1 Host omics data	40
1.6.2 Meta-omics approach	43
1.7 <i>Data processing and modelling</i>	48
1.7.1 Network biology approaches	48
1.7.2 Databases and tools	53
1.8 <i>Aims and Objectives</i>	58

Chapter 2 - Development of <i>in silico</i> approaches to study intercellular communication	59
2.1 Introduction	59
2.2 Methods	62
2.2.1 Identifying intercellular interactions among different cell types	62
2.2.2 Single-cell data processing	63
2.2.3 RNA-seq data filtering	64
2.2.4 Reconstructing a cell-cell interaction network	64
2.2.5 Building up ligand-receptor interaction networks between myofibroblasts and regulatory T cells	65
2.3 Results	66
2.3.1 Semi-automated pipeline to build cell-cell interactomes	66
2.3.2 Analysing intercellular interactions in healthy and diseased colon	66
2.3.3 Effect of myofibroblasts on regulatory T cells	68
2.4 Discussion	71
Chapter 3: Discovering the effect of the human microbiome on host cell signalling	74
3.1 Introduction	74
3.2 Methods	77
3.2.1 Location analysis of proteins	78
3.2.2 Host-microbe protein-protein interactions	79
3.2.3 Network propagation algorithms	81
3.2.4 Gene enrichment and overrepresentation analysis	83
3.3 Results	84
3.3.1 Host-microbe interactome	85
3.3.2 Network diffusion modelling	87
3.3.3 Functional analysis	88
3.4 Discussion	88
Chapter 4: Analysing the cell-type specific effect of bacterial outer membrane vesicles on the immune system	90
4.1 Introduction	90
4.2 Methods	92
4.2.1 Experimental analysis of BEV proteins	93
4.2.2 Single-cell transcriptomic datasets analysis	93
4.2.3 Analysis of bulk transcriptomic data	93
4.2.4 Cell-type specific Bt BEV - human interactome	94
4.2.5 Functional analysis of Bt BEV protein targets	94
4.2.6 Cell-type specific signalling pathway analysis	95
4.2.7 <i>In vitro</i> validation of <i>in silico</i> findings	95
4.3 Results	95
4.3.1 Reconstructing a BEV - human interactome	95
4.3.2 Functions of the human target proteins	97
4.3.3 Effect of Bt BEVs on the Toll-like receptor pathway of immune cells	97

4.3.4 Role of Bt BEV proteins in TIRAP-mediated TLR signalling	100
4.3.5 Inhibition of TLR4 signalling pathway diminishes monocyte activation by Bt BEVs	101
<i>4.4 Discussion</i>	104
Chapter 5: Predicting the effect of the oral microbiome to the host in healthy and in inflamed conditions	107
<i>5.1 Introduction</i>	107
<i>5.2 Methods</i>	113
5.2.1 Collection of bacterial proteins	114
5.2.2 Single-cell transcriptomic analysis	114
5.2.3 RNAseq data filtering	114
5.2.4 Inferring a host-microbe interaction network	115
5.2.5 Functional analysis of microbe-targeted human proteins	115
5.2.6 Downstream network modelling	115
5.2.7 Reconstructing an epithelial cell - immune cell interaction network	116
<i>5.3 Results</i>	116
5.3.1 <i>An in silico</i> host-microbe protein-protein interaction network	116
5.3.2 Functional analysis of host target proteins	119
5.3.3 Downstream signalling network modelling	120
5.3.4 Interaction between epithelial and dendritic cells	122
<i>5.4 Discussion</i>	126
Chapter 6: Perspectives and final discussion	129
References	135
Appendix 1: Supplementary material	172
Appendix 2: Peer-reviewed publications	176

List of abbreviations

3did - three-dimensional interacting domains

AA - amino acid

ADAN - interActions of moDular domAiNs

AJ - adherens junction

APC - antigen-presenting cell

ASV - amplicon sequence variant

BCR - B cell receptor

BEV - bacterial extracellular vesicle

Bt - Bacteroides thetaiotaomicron

CARNIVAL - CAusal Reasoning pipeline for Network identification using Integer VALue programming

CD - Crohn's disease

CLR - C-type lectin receptor

DAMP - danger-associated molecular pattern

DB - database

DC - dendritic cell

DDI - domain-domain interaction

DEG - differentially expressed gene

DMI - domain-motif interaction

EGF - epidermal growing factor

ELM - Eukaryotic Linear Motif database

EMBL - European Molecular Biology Laboratory

EV - extracellular vesicle

FDR - false discovery rate

FPKM - fragments per kilobase of gene model per million mapped reads ratio

GE - gingival epithelium

GEO - Gene Expression Omnibus

GI tract - gastrointestinal tract

GJ - gap junction

GO - Gene Ontology

GPL - general public license

GSA - gene set analysis

GSEA - Gene Set Enrichment Analysis

GSOA - GEne Set Overrepresentation Analysis

GSVA - Gene Set Variation Analysis

HM-PPI - host-microbe protein-protein interaction
HMI - host-microbe interaction
HMP - Human Microbiome Project
HOMD - Human Oral Microbiome Database
HT- high throughput
IBD - intestinal bowel disease
ID - identifier
IDR - intrinsically disordered region
IFNG - interferon-gamma
IL - interleukin
ILC - innate lymphoid cell
ILP - integer linear programming
ISC - intestinal stem cell
JAM - junctional adhesion molecules
JE - junctional epithelium
KEGG - Kyoto Encyclopaedia of Genes and Genomes
LMPID - Linear Motif mediated Protein Interaction Database
LOS - lipooligosaccharide
LPS - lipopolysaccharide
LRI - ligand-receptor interaction
MAPK - mitogen-activated protein kinase
ML - machine learning
MS - mass spectrometry
NCBI - National Center for Biotechnology Information
NHR - nuclear hormone receptor
NK - natural killer cell
NLR - NOD-like receptor
NMR - nuclear magnetic resonance
OP - OmniPath
OSE - oral sulcular epithelium
OTU - operational taxonomical unit
PAMP - pathogen-associated molecular pattern
PBD - phosphatidyl-inositol binding domain
PBS - phosphate-buffered saline
PD - periodontitis
PDB - Protein Data Bank
PIANO - Platform for Integrated Analysis of Omics data

PPI - protein-protein interaction
PRR - pattern recognition receptors
PSI-MI - proteomics standards initiative - molecular interaction
PTM - post-translational modification
pw - pathway
RLR - RIG-like receptors
RNAseq - RNA sequencing
rRNA - ribosomal RNA
RTK - receptor tyrosine kinase
sc - single-cell
SLiM - short linear motif
SLK3 - Signalink3
SVM - support vector machine
Tc - cytotoxic T cell
TF - transcription factor
TG - target gene
TGF β - transforming factor-beta
Th - helper T cell
TieDie - Tied Diffusion through Interacting Events
TIR-domain - Toll/interleukin-1 receptor/resistance protein domain
TIRAP - TIR domain containing adaptor protein
TJ - tight junction
TLR - Toll-like receptor
TNF-a - tumor necrosis factor alpha
TPM - transcript per million
Treg - regulatory T cell
UC - ulcerative colitis
UMI - unique molecular identifier
UMLS - unified medical language system
UniProtKB - UniProt Knowledgebase
Y2H - yeast 2 hybrid

List of Figures

Figure 1.1. Overall view of the PhD project.

Figure 1.2. Gingiva in periodontal health and disease conditions.

Figure 1.3. Lieberkühn-crypts in the healthy and inflamed colon focusing on cell types analysed in the thesis.

Figure 1.4. Intercellular interactions between epithelial cells.

Figure 1.5. Immune response in healthy condition and during pathogen-associated inflammation.

Figure 1.6. Uniform manifold approximation and projection (UMAP) of cell clusters after cell type identification using single-cell RNAseq data from the oral cavity.

Figure 1.7. Introduction to networks and graph-based pathfinder algorithms.

Figure 2.1. Workflow for analysing intercellular interaction and their downstream effect.

Figure 2.2. Pairwise comparison of cell-cell interactions.

Figure 2.3. Condition-specific connections between myofibroblast ligands and Treg cell receptors in ulcerative colitis and healthy control.

Figure 2.4. Intercellular connections and their downstream effect in UC compared with healthy control.

Figure 3.1. Workflow of host-microbe interaction prediction.

Figure 3.2. Brief comparison of TieDie and CARNIVAL network propagation algorithms.

Figure 3.3. Comparison of motif targeting domains between ELM and 3did.

Figure 4.1. Computational workflow to analyse cell-type specific effects of BEVs.

Figure 4.2. Interactions of 48 BEV proteins with various human cells.

Figure 4.3. Expression of TLR pathway members in the A, selected five cell types and B, THP-1 monocytes highlighting the potential BEV targets.

Figure 4.4. Structural details about TLR4 - Bt BEV protein and TIRAP - Bt BEV protein interactions.

Figure 4.5. Inhibition of TLR4 and TIRAP signalling pathways abrogates THP1-Blue cells activation by Bt BEVs.

Figure 5.1. Bacteria representing the human core oral microbiome.

Figure 5.2. Computational workflow to analyse the effect of the gingival microbiome on epithelial and immune cells in periodontal health and disease.

Figure 5.3. Host-microbe interactions in A) healthy and B) severe periodontitis conditions predicted by MicrobioLink2.

Figure 5.4. Functional analysis of bacteria targeted human proteins.

Figure 5.5. Signalling network in epithelial cells focusing on the downstream effect of bacteria.

Figure 5.6. Overlap between functions across bacteria-affected membrane proteins, intermediate proteins, TFs and expressed genes translated to ligands.

Figure 5.7. Overlap of epithelial cell secreted ligands and DC expressed receptors in healthy and periodontitis conditions.

Figure 5.8. Condition-specific connections between epithelial cell ligands and DC receptors in healthy control and severe periodontitis.

List of Tables

Table 1.1. Summary of the thesis.

Table 2.1. Number of expressed genes in cell types.

Table 3.1. *In silico* PPI prediction approaches.

Table 5.1. Bacterial clusters in subgingival plaque described by Socransky *et al.* (1998).

Table 5.2. List of bacterial strains analysed in the study.

Table 5.3. Bacterial Pfam domains targeting SLiMs on human proteins.

Table 5.4. Top 10 signalling pathways represented among the receptors and their first neighbours.

List of Supplementary materials

Supplementary Table 2.1. Number of condition-specific intercellular PPIs.

Supplementary Table 2.2. Top ten overrepresented pathways in upstream Treg signalling network.

Supplementary Table 5.1. Number of bacterial proteins derived from UniProt Proteome.

List of peer-reviewed publications

Peer-reviewed journal articles published during my PhD (2017-2022). Those covered in Chapters 2 and 4 are reproduced in Appendix 2.

Chapter 2:

- Türei, D., Valdeolivas, A., **Gul, L.**, Palacio-Escat, N., Klein, M., Ivanova, O., Ölbei, M., Gábor, A., Theis, F., Módos, D., et al. (2021). Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Mol. Syst. Biol.* *17*, e9923. <https://doi.org/10.15252/msb.20209923>.

Chapter 4:

- **Gul, L.**, Modos, D., Fonseca, S., Madgwick, M., Thomas, J.P., Sudhakar, P., Booth, C., Stentz, R., Carding, S.R., and Korcsmaros, T. (2022a). Extracellular vesicles produced by the human commensal gut bacterium *Bacteroides thetaiotaomicron* affect host immune pathways in a cell-type specific manner that are altered in inflammatory bowel disease. *J. Extracell. Vesicles* *11*, e12189. <https://doi.org/10.1002/jev2.12189>.

Others (not represented in the thesis):

- Jacomin, A.-C., **Gul, L.**, Sudhakar, P., Korcsmaros, T., and Nezis, I.P. (2018). What we learned from big data for autophagy research. *Front. Cell Dev. Biol.* *6*, 92. <https://doi.org/10.3389/fcell.2018.00092>.
- Jones, E.J., Matthews, Z.J., **Gul, L.**, Sudhakar, P., Treveil, A., Divekar, D., Buck, J., Wrzesinski, T., Jefferson, M., Armstrong, S.D., et al. (2019). Integrative analysis of Paneth cell proteomic and transcriptomic data from intestinal organoids reveals functional processes dependent on autophagy. *Dis. Model. Mech.* *12*. <https://doi.org/10.1242/dmm.037069>.
- Sudhakar, P., Jacomin, A.-C., Hautefort, I., Samavedam, S., Fatemian, K., Ari, E., **Gul, L.**, Demeter, A., Jones, E., Korcsmaros, T., et al. (2019). Targeted interplay between bacterial pathogens and host autophagy. *Autophagy* *15*, 1620–1633. <https://doi.org/10.1080/15548627.2019.1590519>.
- Ölbei, M., Thomas, J.P., Hautefort, I., Treveil, A., Bohar, B., Madgwick, M., **Gul, L.**, Csabai, L., Modos, D., and Korcsmaros, T. (2021b). CytokineLink: A Cytokine Communication Map to Analyse Immune Responses-Case Studies in Inflammatory Bowel Disease and COVID-19. *Cells* *10*. <https://doi.org/10.3390/cells10092242>.
- Ölbei, M., Hautefort, I., Modos, D., Treveil, A., Poletti, M., **Gul, L.**, Shannon-Lowe, C.D., and Korcsmaros, T. (2021). SARS-CoV-2 Causes a Different Cytokine Response Compared to Other Cytokine Storm-Causing Respiratory Viruses in Severely Ill Patients. *Front. Immunol.* *12*, 629193. <https://doi.org/10.3389/fimmu.2021.629193>.
- Treveil, A., Bohar, B., Sudhakar, P., **Gul, L.**, Csabai, L., Ölbei, M., Poletti, M., Madgwick, M., Andrighetti, T., Hautefort, I., et al. (2021). ViralLink: An integrated workflow to investigate the effect of SARS-CoV-2 on intracellular signalling and regulatory pathways. *PLoS Comput. Biol.* *17*, e1008685. <https://doi.org/10.1371/journal.pcbi.1008685>.

- Csabai, L., Fazekas, D., Kadlecsek, T., Szalay-Bekő, M., Bohár, B., Madgwick, M., Módos, D., Ölbei, M., **Gul, L.**, Sudhakar, P., et al. (2022). Signalink3: a multi-layered resource to uncover tissue-specific signaling networks. *Nucleic Acids Res.* 50, D701–D709. <https://doi.org/10.1093/nar/gkab909>.
- **Gul, L.**, Korcsmaros, T., and Hall, N. (2022). Flaviviruses hijack the host microbiota to facilitate their transmission. *Cell* 185, 2395–2397. <https://doi.org/10.1016/j.cell.2022.06.019>.
- Demeter A, Jacomin A-C, **Gul L**, et al. (2022) Computational prediction and experimental validation of Salmonella Typhimurium SopE-mediated fine-tuning of autophagy in intestinal epithelial cells. *Front Cell Infect Microbiol.* 12:834895. <https://doi.org/10.3389/fcimb.2022.834895>

Chapter 1: General introduction

1.1 Preface

Humans are colonised by complex microbial communities comprising viruses, archaea, bacteria and eukaryotes. The composition of the microbiome has an impact on the host's physiological processes, hence the communication between the microbe and host is crucial to maintaining homeostasis. Accordingly, disruption in the community composition potentially leads to increased inflammation and the pathogenesis of diseases, such as periodontitis in the oral cavity or inflammatory bowel disease (IBD) in the gut ^{1,2}. Currently, there are correlation-based approaches to study the interplay between host processes and the microbiome (e.g., blood biomarkers ³). These studies reveal significant associations between microbial taxa and host factors, such as the level of cytokines ^{3,4}. The limitation is that this approach can not detect the effect of microbial strains on host signalling at the molecular level due to the complexity and cross-talk of biological processes.

This iCASE PhD scholarship was supported by Unilever, the industrial collaborator of the project. Together, we aimed to provide mechanistic insights into the beneficial and harmful effects of the healthy and unhealthy microbiota that facilitates the product design and improves consumer experience, mitigates any negative effects and enhances the product use benefits in the marketplace. I undertook a three months internship at the company, where I further built my research and personal skills and competencies. I participated in meetings with my industrial supervisor and her team extending my knowledge about metabolic models and network modellings. The internship contributed to my personal qualities by working in a professional industrial environment and practising presentation/communication and time management skills. The project's output for Unilever covered a methodology development (Chapter 3) and a case study by analysing public data from the oral cavity (Chapter 5).

This introductory chapter explores the background theories and literature relating to the current knowledge about the human microbiome and its interaction with the host. Chapter 2 and 3 present interdisciplinary workflows developed by myself and colleagues to study intercellular and host-microbe interactions and their downstream effect on host cellular processes. The following two chapters (Chapters 4 and 5) demonstrate these workflows with case studies. Chapter 4 is a case study for the host-microbe interaction pipeline analysing single-cell transcriptomic data and proteomic profiling of bacterial extracellular vesicles. This study

focuses on the effect of *Bacteroides thetaiotaomicron* (Bt) BEVs on immune cells in the healthy and inflamed colon. Chapter 5 combines the developed pipelines to analyse connections between the gingival microbiome and epithelium and cell-cell communication between epithelial and immune cells in healthy and periodontal conditions. Finally, Chapter 6 discusses overall conclusions of the thesis and gives future directions. The general structure of the thesis is presented in Figure 1.1 and Table 1.1.

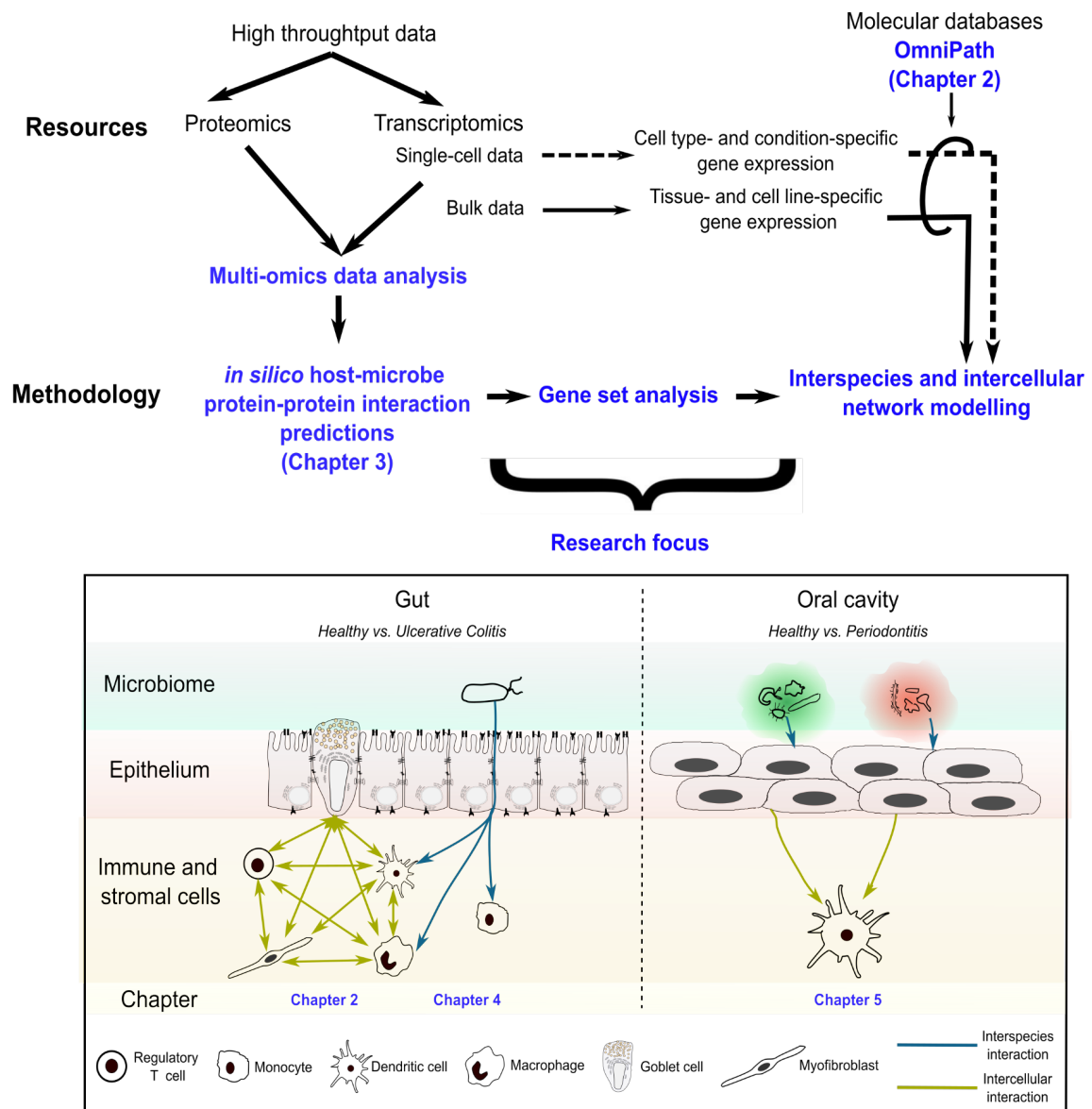


Figure 1.1: Overall view of the PhD project. I highlighted those resources and methods that I carried out by myself.

Table 1.1: Summary of the thesis			
<i>Chapter</i>	<i>Type of the interaction</i>	<i>Object</i>	<i>Aim of the project</i>
Chapter 2	Intercellular interaction	Myofibroblasts - regulatory T cells	Rewiring of stromal and immune cell interactions in UC
Chapter 4	Microbe - host interaction	Bt - immune cells	Role of Bt BEVs on immune cells focusing on the TLR pathway in healthy and UC conditions
Chapter 5	Microbe - host interaction	Microbiota - epithelial cells	Understanding the effect of healthy and periodontitis-related microbiome composition on epithelial cell signalling
Chapter 5	Intercellular interaction	Epithelial cells - dendritic cells	Effect of disturbed epithelial signalling on DCs during severe periodontitis

1.2 Gastrointestinal tract

The gastrointestinal tract (GI tract) involves organs of the digestive system between the oral cavity and anus. The human GI tract is approximately 7m long, and its total surface is around 300 m² with multiple levels of invaginations ⁵. This structure facilitates the main functions of the GI tract - digestion and nutrient absorption - but as later chapters will introduce, immune homeostasis maintenance is also among its main objectives ^{6,7}. Maintaining this balance of defending against pathogenic organisms and ensuring the commensals' environment is challenging. However, epithelial barriers via tightly connected cells, secreted antimicrobial peptides and specialised immune responses enhance this defence mechanism ⁷.

In the thesis, I observed the disrupted homeostatic functions of the epithelial layer in the oral cavity and the gut; therefore the following two sub-chapters introduce their anatomical structures and highlight the crucial cell types which are exposed to external stimuli (i.e., intercellular or interspecies interactions).

1.2.1 Oral cavity

The oral cavity is the entrance to the digestive system, including lips, buccal mucosa, tongue, gingiva, teeth, hard palate, retromolar trigone (area behind wisdom teeth) and salivary glands⁸. The gingiva surrounds and protects the teeth; it is covered with a layer of stratified squamous epithelium, which acts as the first barrier of defence against pathogens. This layer consists of keratinocyte stem cells (1-10%), transit-amplifying cells (~50%) and postmitotic differentiating cells in early-stage keratinisation (~40%)⁹. The epithelial cells establish three layers: junctional epithelium (JE), oral sulcular epithelium (OSE) and gingival epithelium (GE)¹⁰ [Figure 1.2].

The non-keratinized JE lies at the base of the gingival sulcus. It is directly connected with the tooth surface by several intercellular interactions (e.g. hemidesmosomes, desmosomes, adherens junctions, and gap junctions). Here, the cells are flat with loose cellular junctions, abundant in organelles and have large nuclei¹¹. By expressing cytokines and chemokines, JE can indirectly control the microbes through recruiting immune cells^{12,13} [Figure 1.2].

OSE is an intermediate area between the junctional and gingival epithelium. OSE and JE interact with the subgingival microbiome; hence, they play a crucial role in the immune response¹⁴ [Figure 1.2].

GE is the keratinised external layer of the gingiva¹⁵. In contrast to the JE, cells in the GE are tightly arranged polygons with less intercellular space and round nuclei in the centre¹¹. I focused on JE and OSE in the thesis because GE is not in direct contact with the subgingival microbiome [Figure 1.2].

Periodontal health Periodontal disease

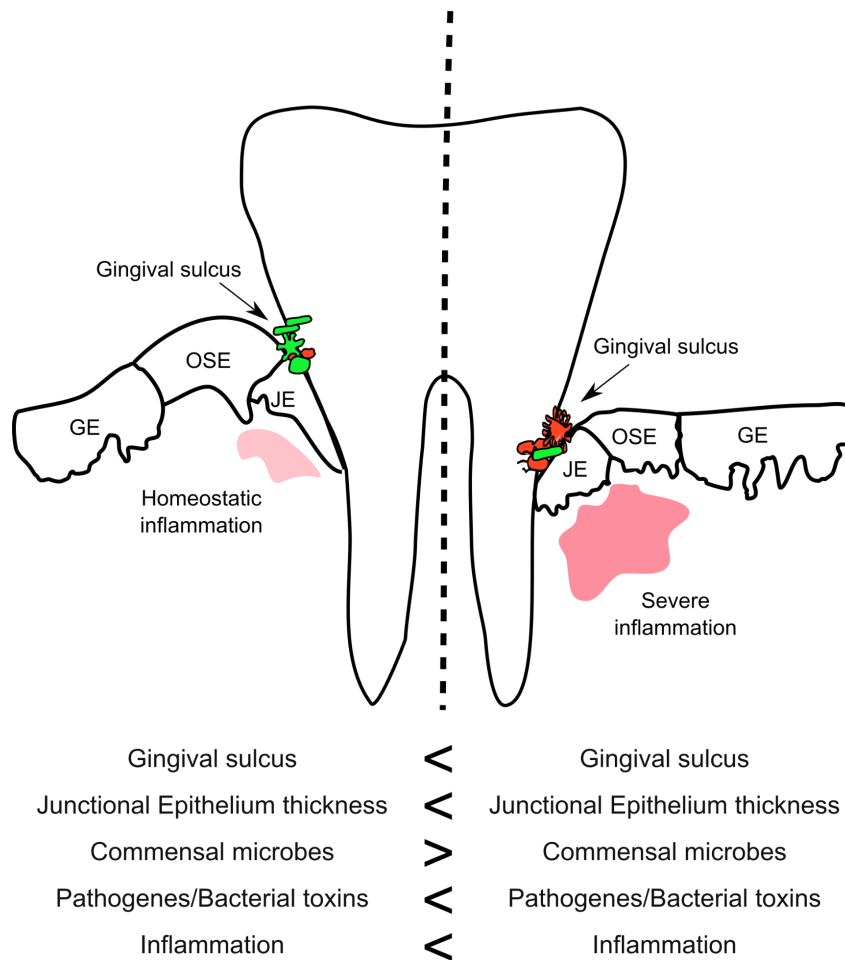


Figure 1.2: Gingiva in periodontal health and disease conditions. GE - gingival epithelium, OSE - oral sulcular epithelium, JE - junctional epithelium. Commensal bacteria are highlighted by green while pathogens appear by red colour. The figure was drawn by myself.

1.2.2 Intestine

The intestinal tract has two main sections: the small intestine and the large intestine. It consists of cells deriving from three main cell lines - epithelial, immune and stromal cells - organised in four layers: mucosa, submucosa, muscularis and serosa. The mucosa involves the epithelial layer, lamina propria and muscularis mucosae - a thin muscular layer. The submucosa is mainly a thick connective tissue layer with blood and lymph vessels, and neurons of the enteric nervous system. Muscularis is a thick layer of smooth muscle, while serosa (or *visceral peritoneum*) is the outer layer surrounding the gut ¹⁶.

Small intestine and colon differ at many points regarding their structure and cell types. In the thesis I worked with gene expression data of colon cells, therefore this part of the intestine is in the centre of the thesis. Following the small intestine where the enzymatic digestion of the food takes place, the colon absorbs the remaining water and ions. Here, the digestion is slower and commensal microbes are responsible for it instead of enzymes. Regarding its anatomical structure, the colon lacks in microvilli, is enriched in Goblet cells and decreased in Paneth cells, and also includes deeper crypts compared to the small intestine.

I focus primarily on the epithelial and secondarily on the immune cells in the colon, therefore, I would like to introduce the common cell types in the epithelial layer and the immune system.

Epithelial cells

Epithelial cells build up the mucosal layer, which serves as a physical barrier, and they defend against pathogens and secrete and absorb molecules. In the intestine, to increase the surface area for absorption, the intestine developed Lieberkühn-crypts, which are fold-like invaginations of the epithelium¹⁷. While the crypts are more expressed in the small intestine, in the colon these structures are less deep and include tubular pits, which increase in depth towards the rectum¹⁸.

Intestinal stem cells (ISC) are multipotent adult stem cells capable of renewing themselves and differentiating into a limited number of gut cells. There are around six ISCs at the bottom of each Lieberkuhn-crypt. The connection of ISCs with other epithelial and mesenchymal cells subserve the homeostatic behaviour¹⁹. Stem cells can renew themselves by an active Wnt signalling pathway. Through the Notch pathway, these multipotent cells can also differentiate into short-living transit-amplifying cells that are rapidly proliferating - but not differentiating - cells with a limited number of cell cycles²⁰ [Figure 1.3]. Alteration of the WNT and Notch pathways can lead to the malfunction of differentiation causing diseases, such as cancer²¹ or inflammatory bowel disease (IBD)²².

Short-living transit-amplifying cells differentiate into progenitor cells that are also multipotent. However, these progenitors quickly differentiate further to more specialised cell types. In the intestine, there are two types of progenitor lineages: absorptive and secretory. These lineages determine two main cell types. Absorptive cells (enterocytes) require an active Notch pathway in the progenitors, while in secretory progenitors (producing goblet, enteroendocrine and tuft cells), the WNT pathway is triggered^{23,24} [Figure 1.3].

Absorptive cells

Enterocytes, the most common cell type in the intestinal epithelium, build up a polarised monolayer with tight cell-cell adhesive interactions [Figure 1.3]. Their function is related mainly to absorption - microvilli on the apical surface increase their membrane surface -but they also play an essential role in host-microbe interactions ^{25,26}. As part of the intestinal epithelium, enterocytes can establish a direct connection with the microbiome, such as detoxifying bacterial toxins ²⁷. During inflammation, their junctional interactions are disturbed, leading to an impaired barrier function ²⁸.

Secretory cells

The bottle-shaped tuft cell is a rare cell type in the intestinal epithelium. It has characteristic microvilli at its apical side. Tuft cells are also connected to the nervous system by expressing acetylcholine ²⁹. Although their functions are less studied, they regulate intestinal epithelial cell response to injury. While these cells have been connected to inflammation-related diseases, their role in inflammation remains unclear. Experiments in mice show that the marker receptor - Dclk1- mediates epithelial repair responses, a process that dysfunctions during induced colitis ^{30,31}.

Enteroendocrine cells are part of the secretory system containing either large dense-core vesicles or smaller synaptic-like microvesicles ³². They secrete a wide range of peptide hormones, but also sense microbial metabolites and release cytokines in response ³³. Duodenum and terminal ileum express the most enteroendocrine cells, and studies show that these cells are strongly affected during Crohn's disease ³⁴.

Goblet cells derive from the secretory progenitor cells and undergo a maturation process. Therefore mature and immature forms can be distinguished. While immature cells are placed in the middle part of the crypt and the vesicle system is less developed, the mature goblet cells are at the top of the crypts. They secrete mucin, antimicrobial proteins, chemokines and cytokines to strengthen the barrier between the gut lumen and epithelial surface ^{22,35,36} [Figure 1.3]. Interestingly, in colon not only the number of these cells is increased but also the epithelial layer is covered by a two-layered mucin layer ³⁷. Several studies support that goblet cells are affected in IBD. For example, in ulcerative colitis, the number of these cells is often reduced. It is still a question whether there is an impaired differentiation or apoptosis is increased in the mature/immature cells ²².

M cells are the direct neighbours of enterocytes in both small and large intestines. The cells are specialised epithelial cells with several characteristic features, such as a lack of apical microvilli and the appearance of a basolateral pocket that usually contains a B lymphocyte. However, T cells and myeloid cells may also be present [Figure 1.3]. This connection with the immune system shows the contribution of these cells to normal immune surveillance³⁸. The primary role of M cells is to deliver microbial antigens to gut-associated lymphoid tissue for efficient mucosal and systemic immune responses³⁹. M cells' behaviour depends on cytokine molecules that also influence inflammation in the gut, hence the role of these cells is significant in IBD⁴⁰.

Immune cells

The immune system is outstandingly important in the gut to maintain the homeostatic state. Several factors are responsible for normal immunity, including the gut microbiome with high priority⁴¹. Similarly to the epithelial cells, immune cells also derive from a multipotent stem cell - called hematopoietic stem cell. Progenitor cells determine the myeloid and lymphoid cell lines: differentiated myeloid cells are in the blood while lymphoid cells mediate the production of immunity⁴².

Myeloid cells

Monocytes derive from myeloblast cells and can differentiate into macrophage and dendritic cell subpopulations in the blood. During inflammation, monocytes go through the endothelial cells and differentiate into anti- and pro-inflammatory macrophage subsets in the tissue. Macrophages are heterogeneous cells rapidly adapting to the changes in the microenvironment⁴². Besides their main phagocytic activity, as a professional antigen-presenting cell type (APC), macrophages also play a role in the maintenance of T cell subpopulations, clearance of apoptotic cells, and maintenance of epithelial barrier integrity⁴³.

Dendritic cells' (DCs') main role is in T cell response via their APC activity. Classical DCs are divided into two subsets: DC1 has CD8 α ⁺ and CD103⁺ on their surface while DC2s are characterised by CD11b⁺ and CD172a⁺^{44,45}. Experiments show that in the inflamed intestine, the number of CD103⁺ DC1 cells is reduced. Based on studies, intestinal inflammation causes the malfunction of DCs that leads to dysregulated T cell responses and tissue damage⁴⁶.

Lymphoid cells

B cells are a core part of the adaptive immune system. Antigen-activated B cell receptors (BCRs) initiate B cell differentiation to plasma cells. Beside the adaptive immune response, B cells are professional APCs and contribute to effector T-cell activation. Not surprisingly, B cells show abnormalities in inflammatory diseases, including the aberrant expression or function of key signalling molecules and cytokines, as well as perturbations in the development of B cell subsets ⁴⁷.

Similarly to B cells, T cells are also a key part of the adaptive immune response. Intestinal T cells have two subgroups, conventional and non-conventional T cells. Conventional T cells derive from CD4⁻CD8⁻ progenitors in the thymus and develop into CD4⁺ or CD8⁺ T cells. These cells subsequently migrate to peripheral lymphoid organs, such as lymph nodes, where they encounter antigens and acquire an activated effector phenotype that drives their migration to the gut. CD4⁺ helper T cells have a CD4 glycoprotein at their surface. They become activated by binding the MHC II complex expressed on the surface of APCs and through rapid proliferation, they differentiate into several subpopulations (Th1, Th2, Th17, and Treg). In contrast, CD8⁺ cytotoxic T cells recognise a short part of the MHC I complex leading to cytokine expression and apoptosis triggering ^{48,49}. In IBD, patients have normal amounts of CD4⁺ T cells and CD8⁺ T cells, however, their activation is different to the normal condition ⁵⁰. It manifests in an increased expression of major lymphocyte activation antigens, such as interleukin-2 receptor, transferrin receptor and 4F2, on the cell surface ⁵¹.

Innate lymphoid cells (ILCs) are a heterogeneous group of immune cells dividing into five main groups: ILC1, ILC2, ILC3, natural killer (NK) and lymphocyte tissue-inducer cells ⁵². Without their antigen receptors, they sense the changes in the environment by cytokine receptors. In the intestinal mucosa, ILCs block pathogen infection by secreting IFN-gamma but also promote IBD and cancer through IFN- γ , IL-17 and IL-22 expression ⁵³.

In contrast to CD8⁺ T cells, NK cells do not require antigen presentation for cytokine secretion. However, they also have a cytolytic function that destroys the target cell ⁵⁴. In the intestinal tract, NK cells can trigger inflammation through several signalling pathways (e.g. Toll-like receptor (TLR) signalling) ⁵⁵. During intestinal inflammation, NK cell-related cytokine secretion is increased ⁵⁶.

Stromal cells

Intestinal stromal cells are part of the mesenchymal compartment. The stromal cells (fibroblasts, myofibroblasts, pericytes, endothelial cells, and smooth muscle cells) are connected to the epithelial and immune cells. These cells have common characteristics such as abundant collagen production, expression of vimentin and α -smooth muscle actin filaments, and a lack of surface CD45 expression⁵⁷. Evidence shows that stromal cells are strongly influenced by intestinal inflammation⁵⁸.

Fibroblasts are localised close to the basolateral surface of epithelial cells⁵⁹ [Figure 1.3]. These cells are responsible for establishing the extracellular matrix by secreting collagen and fibronectin molecules. Fibrosis is a well-known complication of intestinal inflammation caused by mesenchymal cells, such as fibroblasts, that secrete an immoderate amount of extracellular matrix⁶⁰.

Myofibroblasts are subepithelial cells in the intestine sharing features of fibroblasts and smooth muscle cells⁵⁹. Mifflin *et al.* described strict criteria of being a myofibroblast; based on their definition, myofibroblasts are '*spindle-shaped or stellate cells that are α -SMA positive, vimentin positive, smooth muscle myosin negative but non-smooth muscle myosin positive, fibronectin positive, and very weakly positive or negative for desmin*'⁶¹. Myofibroblasts are also responsible for fibrosis during intestinal inflammation⁶².

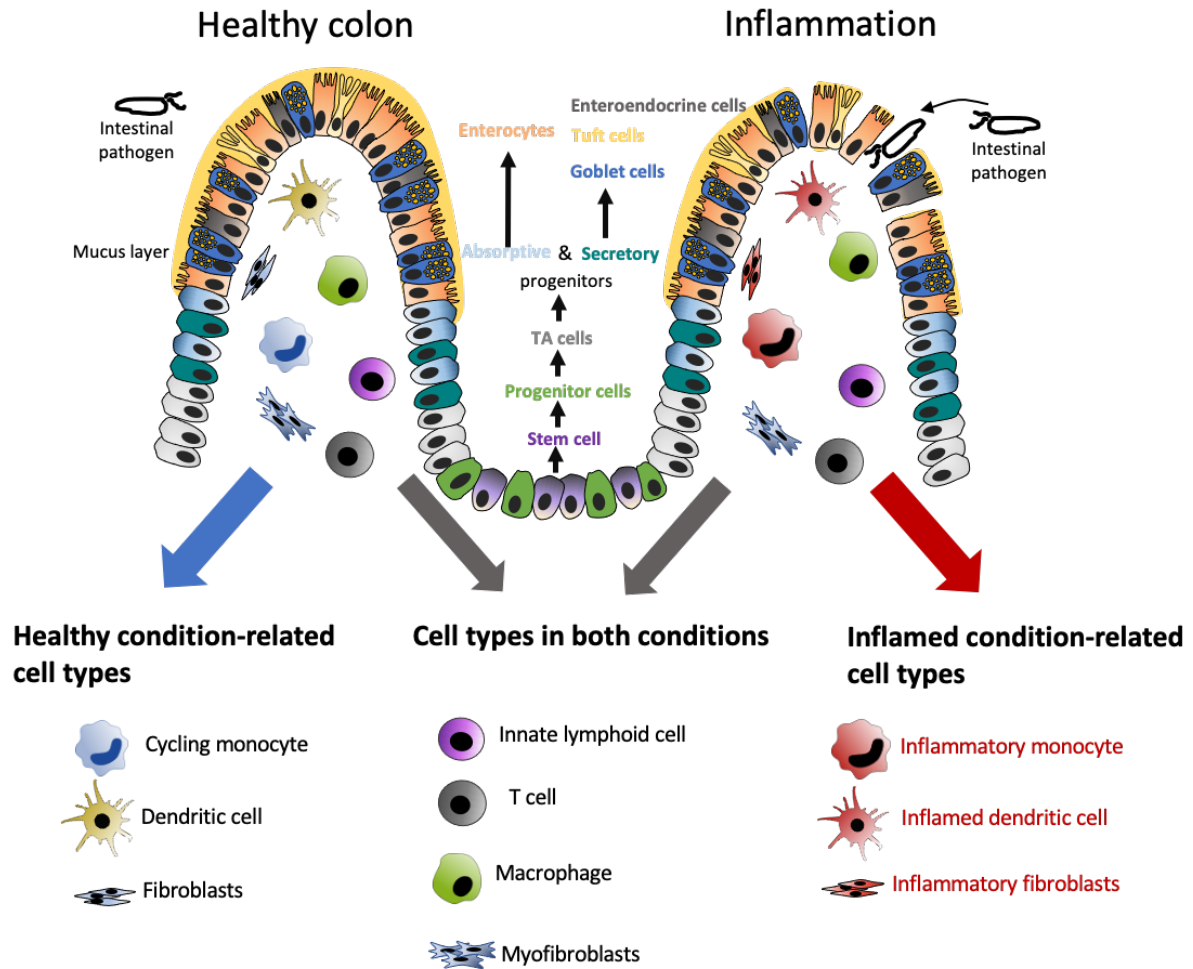


Figure 1.3: Lieberkühn-crypts in the healthy and inflamed colon focusing on cell types analysed in the thesis. In healthy condition, intestinal epithelial cells establish a tightly connected barrier and the differentiated cells secrete mucus and antimicrobial peptides to inhibit the direct interaction between pathogens and the host immune system. In contrast, the continuous layer is disrupted during inflammation which allows microbes to reach immune cells and lead to cytokine secretion. I modified the original figure drawn by Isabella Hautefort, member of our research group.

1.3 The human microbiome

Around 38×10^{12} microorganisms inhabit our tissues and organs, while a human adult body consists of 30×10^{12} cells^{63,64}. It was estimated in 2012 that there were approximately 10,000 different bacterial species that made up the human microbiome (source: <https://www.nih.gov/news-events/news-releases/nih-human-microbiome-project-defines-normal-bacterial-makeup-body>). However, the exact number of strains is difficult to determine as it's constantly changing and varies from person to person. The number of genes expressed in the human microbiome is estimated to be around 246 millions (source: <https://microbiomepost.com/how-many-genes-make-up-the-human-microbiome/>). A study from 2010 has shown that the genetic diversity of the microbiome is much greater than that of the human genome, with the estimated number of microbial genes in the human body being around 150 times greater than the number of human genes⁶⁵. Due to the development of omics technology, this number is probably much higher currently.

The composition of the microbiome differs between organs/tissues and also between individuals as these living communities have adapted to distinct environments⁶⁴. The Human Microbiome Project was the first interdisciplinary effort to describe microbial communities, initially analysing samples from 300 healthy adults, including 18 body sites⁶⁴. The appearance of omics data [details in section 1.6] boosted individual microbiome analysis that has extended our knowledge about the communities.

The most diverse microbial communities inhabit the gut, oral cavity and skin^{66,667,68,66}. The skin is our largest organ therefore the microbiome composition has an essential role to defend against pathogens and the maintain homeostasis of skin cells⁶⁹. Around 1000 bacteria species inhabit the skin besides fungi, archaea, viruses, and mites are also represented in the community⁷⁰. The diversity and composition of the skin microbiome can vary greatly depending on factors such as age, genetics, diet, and environment. While in the gut the microbiome stabilises around 3 years of age⁷¹, the microbial community of the skin changes during time, especially during puberty when lipophilic organisms are enriched on the skin⁶⁹. As Unilever is interested in the effect of the microbiome on scalp, I investigated the scalp microbiome in more detail, although due to confidential reasons, the thesis doesn't include data analysis related to this microbiome - tissue interactions. The scalp microbiome is less discovered; currently, there are 89 articles related to the 'scalp microbiome' keywords in PubMed (September 2022). The community of microorganisms differs on the scalp compared to the skin. The microbiome of the skin is more affected by the different external factors like

moist, dry and sebaceous microenvironments⁶⁹. On the scalp, then microbiome diversity is lower and mainly the *Malassezia*, and *Propionibacterium* and *Staphylococcus* taxa are enriched^{72,73}.

The gut microbiome is currently a hot topic in the research since the appearance of meta-omics data [details in Chapter 1.6.2]. While in 2013 the literature described around 300 to 500 bacterial species⁷⁴, a study published in 2019 described around 8000 strains⁷⁵. The gut microbiota is a complex community of commensal bacteria, fungi, and viruses in the human intestinal system. The composition of the microbiome is influenced by many environmental factors (e.g. diet or use of antibiotics)⁷⁶. In the past years, the definition of the 'core' gut microbiome has been changed: while earlier the core community was defined by microbial taxa which overlap among people, currently, researchers believe that the core microbiome could be defined by genes and/or metabolic capabilities that can be explored by meta-omic data [details in Chapter 1.6.2]. The disturbance of the core microbiome leads to changed regulation of host cellular processes⁷⁷.

The oral microbiome is the second-largest microbiome in humans, containing a complex community of a vast spectrum of species from bacterial, viral, fungal and protozoan taxa. As an open community, it does not have a permanent structure because there are several factors which influence its composition, such as food or the condition of teeth⁶⁸. While in 2010 around 600 species were known in the oral cavity⁷⁸, currently ~ 700 species have been described as a result of the omic data revolution⁶⁸ [details in Chapter 1.6]. Around 96% of the species belongs to the phyla of *Firmicutes*, *Actinobacteria*, *Proteobacteria*, *Fusobacteria*, *Bacteroidetes*, and *Spirochaetes* in the Bacteria domain^{79,80}.

Healthy microbiota - consisting of microbes that colonise the host in normal circumstances and do not usually cause disease - is important for maintaining homeostasis. Disruption of the microorganism communities leads to dysbiosis when the equilibrium state of commensal and harmful pathogens is disturbed⁸¹. Therefore dysbiosis is not necessarily associated with the appearance of new pathogens, in most cases, there are disease-associated bacteria with an increased abundance compared to the healthy condition⁸². There are several factors that can contribute to dysbiosis, including poor diet, antibiotic use, chronic stress, and exposure to toxins or pollutants. Dysbiosis is also seen in certain medical conditions such as diabetes, autoimmune diseases, and cancer⁸³.

When the balance of beneficial and harmful bacteria is disrupted, this can lead to inflammation in the gut. The gut microbiome plays a key role in maintaining the integrity of the gut lining and regulating the immune response, so when the balance is disrupted, it can lead to an overactive immune response and inflammation through altered signalling pathways⁸⁴. Inflammation in the gut can also lead to dysbiosis, as the inflammatory response can damage the epithelial layer and alter the environment for the gut microbiome. This can make it difficult for beneficial bacteria to survive and thrive, while allowing harmful bacteria to overgrow⁸⁴. Dysbiosis has been linked to various inflammatory conditions such as inflammatory bowel disease (IBD), autoimmune diseases, allergies, and metabolic disorders. Anti-inflammatory diet and some probiotics can help to reduce inflammation and restore the balance of the gut microbiome⁸⁵.

A dysbiotic community is usually characterised by a reduced diversity of the microbiome, this has been associated with many diseases, such as IBD in the gut, periodontitis affecting the gingiva or eczema on the skin⁸⁶. Interestingly, while the number of taxa is decreased during dysbiosis, the variability of microbes is increased⁸⁷. As the 'Anna Karenina principle' says, healthy microbiomes are similar to each other while the disease associated microbiomes differ from each other^{88,89}. Another assumption is that the dysbiosis potentially causes dysanaerobic processes based on the oxygen hypothesis. This rule is based on the observation that in the dysbiotic intestinal microbiome, the obligate anaerobic taxa are shifted to facultative anaerobic species⁹⁰.

In the gut, a dysbiotic microbiome may be characterised by (1) an overgrowth of pathogenic bacteria such as *Escherichia coli*, *Clostridium difficile*, or *Salmonella*⁹¹; (2) a decrease in the abundance of beneficial bacteria such as *Lactobacillus* and *Bifidobacterium*⁹²; (3) an increase in the ratio of Firmicutes to Bacteroidetes, this ratio is usually associated with obesity and metabolic disorders⁹³; (4) an increase in proinflammatory bacteria and a decrease in anti-inflammatory bacteria⁹⁴.

Microbiologists distinguish two groups of bacteria based on the membrane structure: Gram-positive bacteria consist of a thick peptidoglycan layer and periplasm⁹⁵. In contrast, the cell wall in Gram-negative bacteria has three layers: the outer membrane, peptidoglycan layer and periplasm⁹⁶. The diverse outer coat infers an altered communication with the host cells and tissues. Although microbes in both categories can produce small, nano-sized extracellular vesicles to transport bioactive molecules to the host cells, the composition differs between them. Gram-negative bacteria secreted vesicles consist of lipopolysaccharide (LPS), in contrast, Gram-positive vesicles contain lipoteichoic acid⁹⁷ [for details, see Chapter 4].

The gastrointestinal tract is exposed to several factors, such as diet or smoking that have an impact on tissue homeostasis. Therefore, the microbiome has a crucial role in the maintenance of physiological conditions ^{67,68}. In general, studies analyse the oral cavity and gut separately, however, they are not only linked physically but also share microbes. The ‘oral – gut microbiome axis’ expression describes the oral-to-gut and faecal-to-oral translocation of microbes ⁹⁸. In healthy condition, the overlap is less between the communities due to functional barriers, such as gastric acid or bile. Nevertheless, an impaired oral-gut barrier leads to the translocation of microbes and contributes to a diseased condition ⁹⁸. Further details about the gut and oral microbiome are described in Chapter 4 and Chapter 5.

1.4 External signals affecting the epithelium

The tightly connected epithelium is exposed to both host cell factors and the microbiome, therefore intercellular (cell-cell) and interspecies (cell-microbe) interactions are crucial in the epithelial layer ^{99,100}.

1.4.1 Cell-cell interactions

Cell-cell interactions are essential for growing and differentiating multicellular organisms by transducing the signal from cell to cell ¹⁰¹. These interactions are specific and highly regulated due to their significant impact on physiological processes ¹⁰². Disruption of the intercellular interactions affects the homeostatic processes and leads to diseases. Understanding the mechanisms of cell-cell interactions is crucial to the development of new therapies and treatments for a wide range of medical conditions ¹⁰¹. Despite its importance, the molecular background is less well described due to the lack of data. With the increasing amount of high-throughput data (genomics, transcriptomics, proteomics, etc - details in Chapter 1.6) available, bioinformatics tools such as network analysis, machine learning and computational modelling can be used to infer cell-cell interactions ^{103–111}.

NicheNet ¹¹⁰ is a computational tool, available as an R package, to model cell-cell communication. It uses network-based approaches to identify key players in a particular pathway or biological process, based on the analysis of large-scale genomic and proteomic data. The core of the algorithm is based on the principle that genes that are functionally related tend to be co-expressed or co-regulated, meaning that they are often active at the same time and in the same tissue or cell type. NicheNet uses this principle to identify functional relationships between genes by analysing patterns of co-expression in various experimental data sets such as gene expression microarrays, RNA-seq, or proteomics data. It requires ligand–receptor, signalling and gene regulatory networks as input then infers a weighted network prioritising source cells based on their ligands' effect on target cell gene expression. NicheNet was the first pipeline which explored the downstream response in the target cell. Like any computational tool, NicheNet has certain limitations that should be considered when interpreting the results: (1) the quality and quantity of the input data can greatly affect the accuracy and reliability of the results; (2) the results are specific to the biological context in which the analysis is performed, meaning that the predictions may not be valid in other contexts or cell types; (3) as the predictions made by NicheNet are based on statistical association, there is a risk of false positives, meaning that some interactions or functional relationships predicted by the algorithm may not be biologically relevant; (4) NicheNet does not provide information about the directionality of these interactions, meaning that it cannot distinguish between activating and inhibitory interactions ^{110,112}.

CellphoneDB ¹¹¹ is a computational tool that predicts ligand-receptor interactions by analysing the structural and functional properties of proteins. The tool uses a structural alignment method to identify similar binding pockets in proteins, which are then used to predict potential ligand-receptor interactions. It also uses functional annotation information, such as gene ontology terms and enzyme commission numbers, to identify proteins that are likely to be involved in similar biological processes and therefore more likely to interact. The predictions are ranked based on the basis of their total number of significant P values across the cell populations ¹¹¹. The main limitations of the tool are that (1) it does not include all of the possible ligand-receptor interactions, therefore analysing cell-cell interactions can be misleading; (2) the statistical method that calculates p-values and ranks the interactions is based on the Importance of the PPI in the downstream signalling in the target cell, therefore a non-significant interaction does not mean that the LRI is not present, but it is not highly specific between the source and target cells ¹¹¹.

LIANA (Ligand-receptor Analysis frAmework) ¹⁰⁹ is a bioinformatics tool for the analysis of ligand-receptor interactions. It integrates 16 intercellular resources and 7 methods (including CellphoneDB) to analyse large-scale genomic and proteomic data and infer ligand-receptor interactions from it. The algorithm combines different scores to rank the intercellular interactions with the Robust Rank Aggregation algorithm ¹¹³ Compared to NicheNet that analyse intercellular interactions and their downstream effect, LIANA is specialised to PPIs between cells. Nevertheless, the two approaches are not mutually exclusive, these tools discover cell-cell interactions from different points of view ¹⁰⁹.

In the thesis, I distinguished two major cell-cell interaction types while exploring intercellular interactions: cell-cell junctions and cell-cell communication through ligand-receptor interactions. While junctional interactions support a structural and physical cell-cell interaction, ligand binding to the complementary receptor triggers signal spreading through the cell mediating intercellular communication ¹¹⁴ [Figure 1.4].

Structural cell-cell junctions

Adhesive cell-cell interactions are mediated by adherens junctions, gap junctions, tight junctions and desmosomes ⁹⁹: Adherens junctions (Ajs) are essential in the development and tissue homeostasis, cells are connected through molecules which are anchored to actin filaments in the cytoplasm [Figure 1.4]. Ajs help to polarise the epithelial cells and distinguish the apical and basolateral membranes, besides these molecular complexes link the adjacent cells tightly in the intestinal epithelium, therefore, ensuring its barrier function ^{100,115}. Impaired Ajs, which cause incompletely polarised epithelial cells, characterise both CD and UC conditions ¹¹⁶.

Tight junctions (TJs) bind cells only in epithelium and endothelium [Figure 1.4] and give polarity to the cells by separating the upper and lower part. In contrast to Ajs, this junctional complex contributes to a semipermeable barrier through that small molecules (ions, solutes) can pass. TJ complexes control proliferation and differentiation. Disruption of these intercellular structures causes impaired barrier function and enhanced inflammatory cytokine secretion, leading to inflammation-associated diseases ^{99,100}.

Gap junctions (GJs) are essential for growing, developing and maintaining homeostatic functions. Molecules establishing GJs – called connexins – are transmembrane proteins appearing in clusters and facilitate the nutrient and solute passing into the intercellular space [Figure 1.4]. During intestinal inflammation, connexin expression is reduced and re-organised from the apical side to the basolateral membrane. These findings suggest a hypothesis that intercellular communication is more intense between epithelial cells ¹¹⁷.

Desmosomes provide mechanical connections between cells rather than controlling the solute transport ⁹⁹. In the cytoplasm, intermediate filaments bind to the cell surface part of the molecular complex through desmoplakin – an intermediate filament binding protein ¹¹⁸. Altered desmosome structures contribute to the IBD pathology affecting the integrity of the epithelium during intestinal inflammation ¹¹⁹. Hemidesmosomes look like half a desmosomes and also facilitate cell adhesion, however, these multiprotein complexes mediate interactions between the cells and the basal cell membrane in contrast to desmosomes [Figure 1.4].

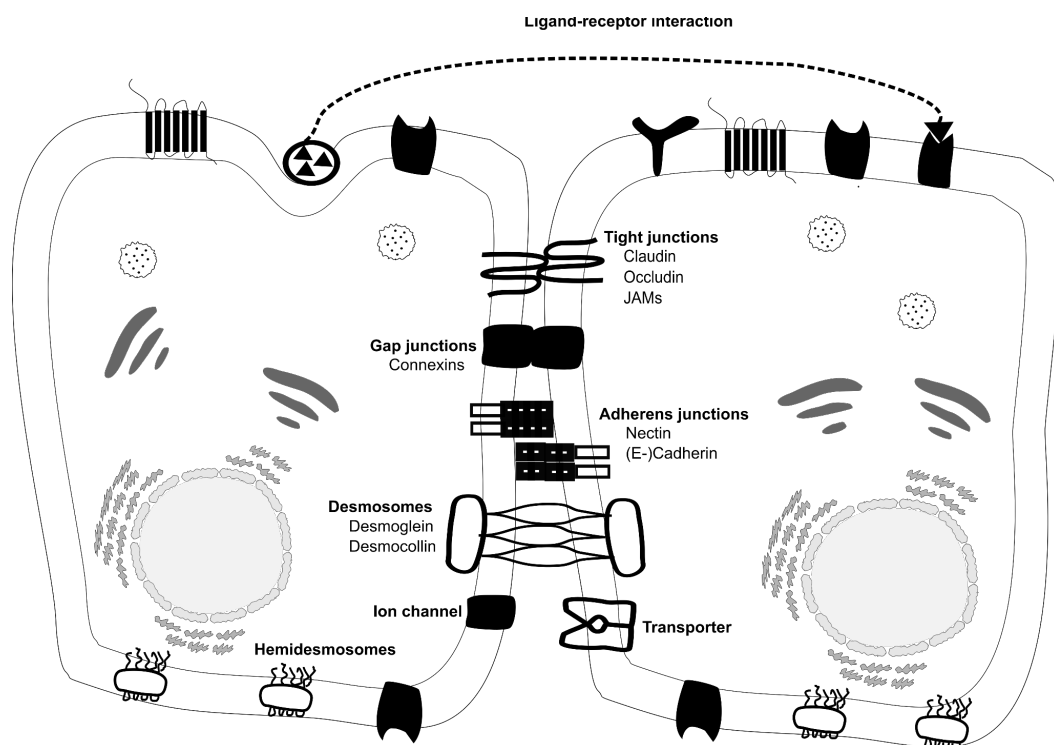


Figure 1.4: Intercellular interactions between epithelial cells. Tight junctions, gap junctions, adherens junctions and (hemi)desmosomes represent the structural connections, while ion channels, transporters and ligand-receptor interactions contribute to the cell-to-cell communication. JAMs – Junctional adhesion molecules. The figure was drawn by myself.

Intercellular communication by ligand-receptor interactions

Ligand-receptor interactions (LRIs) constitute a significant class of intercellular interactions where the source ligand (small molecule, short DNA/RNA or protein) is recognised by a receptor protein. During my PhD, I analysed protein-protein interactions that are the most complex LRIs. The distance is large between molecules; besides, precise orientation and strong physical binding are needed between the molecules ¹²⁰. The connection between a ligand and receptor is based on precise pattern recognition, and selective molecule binding that triggers downstream activation of protein cascades. It was known decades ago that competition is also essential in terms of LRIs because one receptor can bind multiple ligands ¹²¹. Binding affinity determines the strength of the ligand-receptor connection, therefore, a low-affinity interaction can be replaced in the presence of a high-affinity link ¹²².

Ligands and receptors are upstream in the signalling pathways, thus their expression shows more cell specificity than intracellular proteins ¹²³. Impaired LRIs cause altered downstream signalling paths but as a potentially severe consequence leads to activation of other signalisation events ^{124,125}.

Although intercellular protein-protein interactions (PPIs) play an essential role in cellular behaviour, cell-cell connections were less studied till the early 2000s. The limitations were on the one hand technical – no data about individual cells – on the other hand, there was no integrated database that included information about intercellular components and their interactions.

1.4.2 Host-microbe interactions

Host-associated microbial communities are in continuous interaction with the host cells. Modulation of host processes is required for homeostasis ¹²⁶. However, these interspecies interactions have evolved. Microorganisms have adapted to their host and mimic the structure, sequence, motif and interface of many host proteins, facilitating host-microbe interactions (HMIs) and increasing their influence on host processes ^{127,128}.

Proteins do not have a rigid structure, their binding sites are dynamic and shared by various interaction partners. Even the same target protein can mediate diverse downstream signalling events based on their actual binding partner ¹²⁹. This molecular evolutionary strategy leads to competition with host proteins. Generally, a microbe's purpose is to benefit from the

interaction with the host (e.g., by hijacking or evading the immune system), therefore, molecular mimicry is commonly used for immune system-related signalling receptors^{127,128}. Franzosa and Xia introduced the term 'evolutionary arms race', which means microbes mimic the host proteins, and these target structures evolve to avoid interaction with specific microorganisms¹²⁸.

Regarding the effect of bacteria in host signalling, these microorganisms can secrete ligands recognised by host receptors¹³⁰. Scientists have demonstrated that studying these secreted molecules can help us understand diseases and design drugs^{131–133}.

In recent years, the appearance of high-throughput experiments has led to increased knowledge about human HMIs, and more and more studies have included experimental evidence on molecular mechanisms in HMIs^{126,134–139}. The challenge in this field is to analyse large microbial community data and discover the collective effect of the microbes on cell type levels in humans.

1.5 Molecular background of inflammation

Inflammation is triggered by several infectious (e.g., pathogenic organisms) and non-infectious factors (damaged cells, toxins, burn, etc.). These components lead to tissue damage or diseases in the human body. Although inflammatory response depends on the type of the initial factor and the location in the body, the key events are similar: Firstly, receptors on the cell surfaces or in the cytoplasm sense the stimulus. Triggered receptors lead to the activation of inflammatory pathways. As a result, the transcriptional program of the cells is changed, and the expressed inflammatory markers (e.g. cytokines, chemokines) facilitate inflammatory cell recruitment¹⁴⁰.

1.5.1 Cytokine secretion

Cytokines are small molecules (<40 kDa) secreted by various cell types (epithelial, immune and also fibroblast cells), playing an essential role in immune response ¹⁴¹. Cytokines can stimulate each other, but there are also inhibitory relationships among them to suppress positive feedback loops of the inflammatory chemokines. The seriousness of diseases is often associated with the inflammation-activated cytokine storm ^{141,142}.

There are two main groups of cytokines: pro-inflammatory and anti-inflammatory molecules. Proinflammatory cytokines facilitate the inflammatory processes therefore contributing to inflammatory diseases. Conversely, anti-inflammatory cytokines suppress inflammation by responding to the effect of pro-inflammatory molecules ¹⁴³. While the definition of the two groups is self-explanatory, many cytokines, such as IL-6 or IL-8, have anti- and pro-inflammatory effects depending on the environment (e.g. location, nearby cytokines) ¹⁴⁴. Cytokines expose their effect by binding to cytokine receptors, emphasising the importance of intercellular communication in immune response ¹⁴⁵.

The importance of cytokines in oral health is a well-studied topic. In the saliva, several types of these small molecules (e.g. IL-1 β) are in contact with the oral mucosa and gingiva ¹⁴⁶. Microbiome shift contributes to the imbalanced cytokine expression and leads to periodontal diseases ¹⁴⁷. Studies show that the cytokine expression profile differs in gingival inflammation: IL-1 β , IL-6, IL-33 and IL-18 are upregulated ¹⁴⁸⁻¹⁵⁰, while IL-11 (anti-inflammatory cytokine)/IL-17 (pro-inflammatory cytokine) ratio is significantly decreasing ¹⁵¹ suggesting the importance of pro-inflammatory cytokines in the disease.

The gut also responds with a disturbed cytokine secretion profile to the dysbiotic state. Figure 1.5 compares these patterns in the healthy and diseased intestines through the example of pathogen-associated inflammation.

1.5.2 Pathogen-associated inflammation

Ligand-receptor interactions are essential in infectious factor-activated innate immune response. This intercellular communication is established between pathogen-associated molecular pattern (PAMP) carrying molecules (glycans, bacterial proteins, nucleic acids) and pattern recognition receptors (PRRs). LPS is one of the most popular Gram-negative bacteria secreted glycan that contains short, conserved PAMPs. PRRs are expressed by innate immune system-related cell types, including epithelial and immune cells¹⁵². These receptors are either on the cell surface, such as some Toll-like receptors (TLR1, TLR2, TLR4, TLR5, TLR6, TLR11), NOD-like receptors (NLRs), RIG-like receptors (RLRs) and C-type lectin receptors (CLRs)¹⁵³, or intracellular (TLR3, TLR7, TLR8, TLR9)¹⁵⁴.

As a first step of the pathogen-associated inflammation, pathogens cause the contraction of epithelial cells through their secreted endotoxins. Disrupted structural cell-cell interactions lead to gaps on the cell layer that serve as an entrance for microbes to affect deeper tissue layers and contact the immune cells¹⁵⁵. The secreted pathogenic molecules reach the blood vessels and burst the continuous endothelial layer. Endotoxins force the endothelial and immune cells in the blood to express selectins. This family of cell adhesion molecules facilitates anchoring cells to the endothelial layer. Gaps on the surface establish a direct connection between the pathogen and immune cells due to the infiltration of immune cells into the tissue area¹⁵⁶.

Professional APCs (DCs, macrophages and B cells) bind the foreign antigen and present it to naive helper T cells (Th) - which have not met that specific antigen yet - through their MHC-II complexes. Immature Th cells differentiate into diverse subpopulations based on the nature of the presented antigen. While recognising commensal bacteria leads to increased amounts of immunosuppressive cells (Th2, Treg), pathogens induce Th1 and Th17 cell expressions¹⁵⁷. The immune system activator Th subsets express pro-inflammatory cytokines to attract CD8⁺ T cells. Cytotoxic T cells bind the MHC-I complex, appearing on every cell surface. The cells on which this protein complex involves pathogenic antigens, CD8⁺ T cells induce the apoptosis of the infected cells¹⁵⁸ [Figure 1.5].

In summary, the intestinal epithelium is exposed to microbial attacks, therefore PRR-mediated inflammation is outstandingly important in the cells, therefore almost every kind of PRR is expressed by cells in the gut¹⁵⁹. Following the PRR activation, inflammatory signalling is induced and leads to pro-inflammatory cytokine and chemokine secretion. The downstream effect of PRR activation is to restore the damaged mucosal barrier¹⁵⁹.

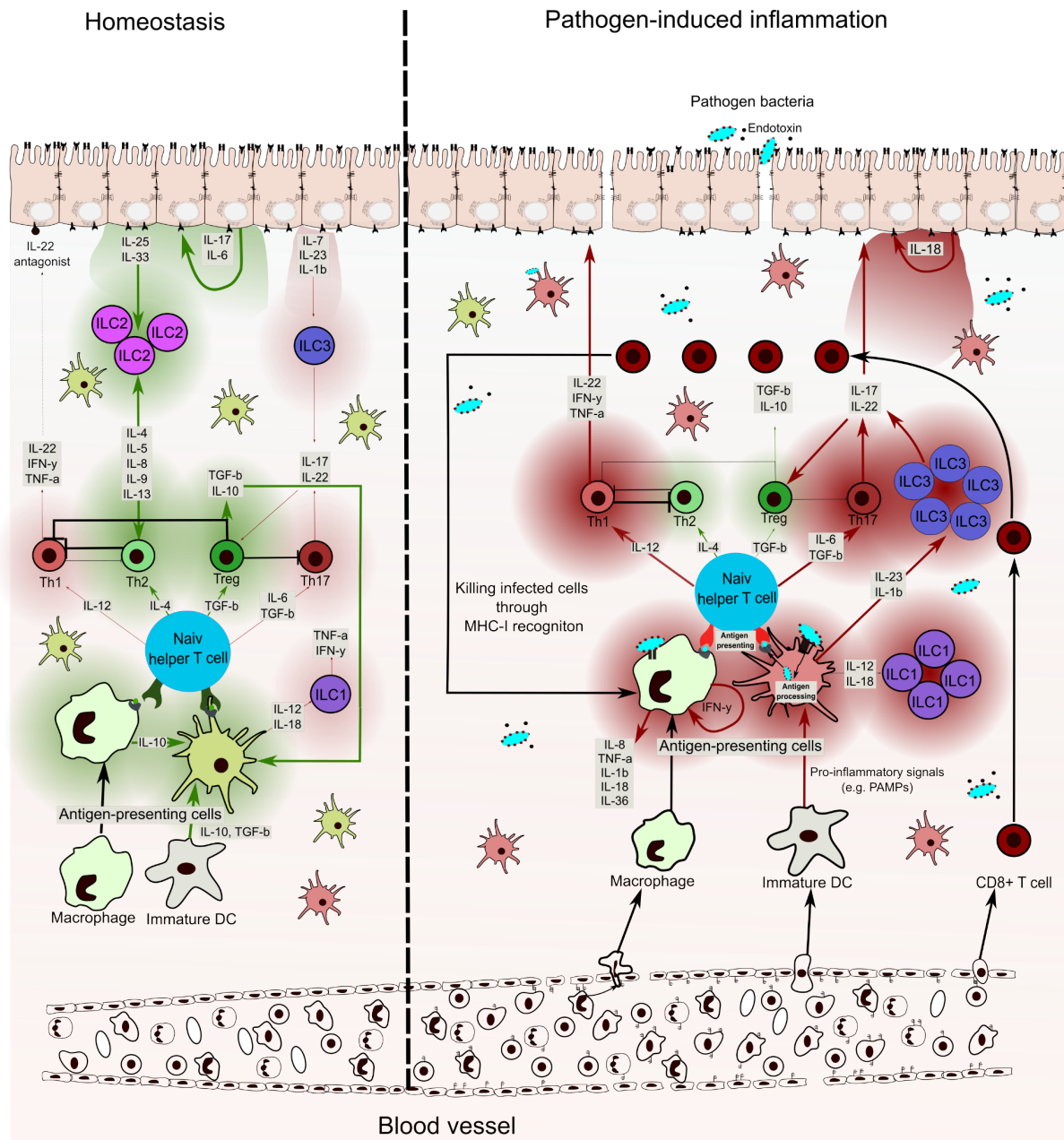


Figure 1.5: Immune response in healthy condition and during pathogen-associated inflammation. The majority of immune cells secrete anti-inflammatory cytokines (highlighted by green) in a homeostatic condition. In contrast, pathogen-derived endotoxins enhance the migration of immune cells, including CD8+ T cells, a major member of immune system responsible for inflammation, from the blood vessel into the tissue that leads to an increased pro-inflammatory cytokine expression (highlighted by red). The naive helper T cells differentiate into diverse subpopulations depending on the signals. While in healthy tissue Th2 and Tregs dominate, Th1 and Th17 cells are overrepresented during pathogen-induced inflammation. This figure was inspired by the following articles^{53,157,160–164} and was drawn by myself.

In the following, I would like to introduce the TLR signalling in detail, because it has been closely related to the projects in which I was working during the PhD.

TLR signalling is one of the innate immunity pathways recognising the extracellular pathogens via PAMPs. Based on the manual curations in the SignalLink3 database, there are 203 proteins involved in the TLR signalling excluding the regulators of the pathway ¹⁶⁵. Receptors and their adaptor proteins (molecules that bind to signalling components resulting in protein complexes instead of mediating specific PPIs) give the main characteristics of a signalling pathway which determine the downstream protein activation. Regarding the TLR pathway, there are ten TLRs (TLR1-10), and five TIR-domain-containing adaptors (MyD88, TRIF, TRAM, SARM, TIRAP) ¹⁶⁶ in human with diverse functions and interaction partners. The receptors, as PRRs in general, are expressed not only on antigen-presenting cells but also on most cell types in the epithelium. In a normal condition, TLR signalling is less active due to decreased receptor and increased receptor inhibitor expression ^{159,167}. TLR pathway regulates both inflammatory and anti-inflammatory responses, disruption of this balanced state results in dysregulated inflammation or abnormal epithelial regeneration. Its main role is to control cytokine secretion, therefore, influencing the appearance of regulatory T cell subpopulations. The impaired signalling causes uncontrolled gastrointestinal inflammation. Studies show that PRR-related gene mutations assist in the development of idiopathic IBD ^{168,169}. TLR signalling has a dual role, it can activate inflammation in the gut but it is also responsible for stopping it and repairing the epithelium in IBD ¹⁵⁹.

1.6 Omics data

Omics data and technologies are large high-throughput (HT) or large-scale assays that measure different kinds of molecules from biological samples. Based on the examined molecular object (highlighted in brackets) the main omics areas are: genomics (genome), epigenomics (epigenome), transcriptomics (transcriptome), proteomics (proteome), metabolomics (metabolome), microbiomics (microbiome), lipidomics (lipidome)¹⁷⁰. Since the past decade, these large-scale datasets have been dominating the biological data generation field, because omics data gives an insight into biological processes on a systems-level¹⁷¹. Single-omics data measures one molecular object (e.g. protein abundance or gene expression) while multi-omics approaches cover not only data coming from the same samples at the same time, but also describe the combination and integration of single-omics datasets. Analysis of clinical samples by new technologies has expanded our knowledge about the molecular background of a wide spectrum of disorders^{172–177}.

The advantage of omics data generation is that the complex set of information gives a more precise and realistic insight into biological processes. However, it is difficult to store and handle big data. Also, large-scale methods increase the false positive rate in datasets compared to small-scale experiments¹⁷⁸. Not only the generation of omic data is challenging but also their analysis. The appearance of a new data type always infers the development of new computational pipelines^{177,179–181}.

1.6.1 Host omics data

Host omics data reveal the cellular processes from different aspects depending on the data type. Genomics analyses the genome - the total amount of DNA in a cell - of the organism and reveals functional information implied in the DNA sequence. It reveals genetic diversity and genomic variation and can also highlight mutations in the nucleic acid sequence¹⁸². Based on estimations, there are around 24,000 protein-coding genes overall in humans¹⁸³, but genetic information differs among people. The reference genome is the standard DNA sequence that derives from multiple donors and represents the pan-human genome¹⁸⁴. Researchers use this standard to align and assemble genome sequence data. Due to the continuous improvement of assembling techniques, the reference sequence has been updated with time. The most current version is the GRCh38.p14 published by the Genome Reference Consortium in February 2022¹⁸⁵.

Transcriptomics measures the total amount of RNA in cells and infers gene activities in the organism. The two major approaches to discovering transcriptional profiles of tissues and cells are microarray and RNA sequencing (RNAseq) techniques. Firstly the microarray assay was invented in the 1990s, this technique is based on hybridisation to predefined transcripts. In contrast, RNAseq can describe the whole transcriptome without prior assumptions of what sequences are present^{186,187}. The identification of new transcripts and other advantages (such as exploring allele-specific expression and splice junctions, independence from genome annotation for prior probe selection) has meant that microarrays have been replaced by RNAseq¹⁸⁷.

Regarding the sequencing approach, there are two different approaches: bulk and single-cell (sc) profiling. Bulk sequencing is a large-scale analysis of cell lines or tissues, it describes an average expression of genes across thousands of cells. Its advantages are the cost-effectiveness and the ability to reveal the altered molecular background of compared conditions (e.g. healthy vs diseased).

Single-cell sequencing is a relatively new methodology, it was used first in 2009 on mouse cells by Tang *et al*¹⁸⁸. This approach gives a high-resolution insight into the tissue composition by detecting the RNA content of samples at the individual cell level. Based on the individual gene expression profile, dimensionality reduction algorithms facilitate the clustering of cells and use markers to distinguish cell subpopulations [Figure 1.6]. There are several existing algorithms (such as UMAP¹⁸⁹, t-SNE¹⁹⁰, IsoMap¹⁹¹ or DiffusionMap¹⁹²) that use different approaches to reduce the dimensionality and facilitate the understanding of cell clusters visualised on diagrams.

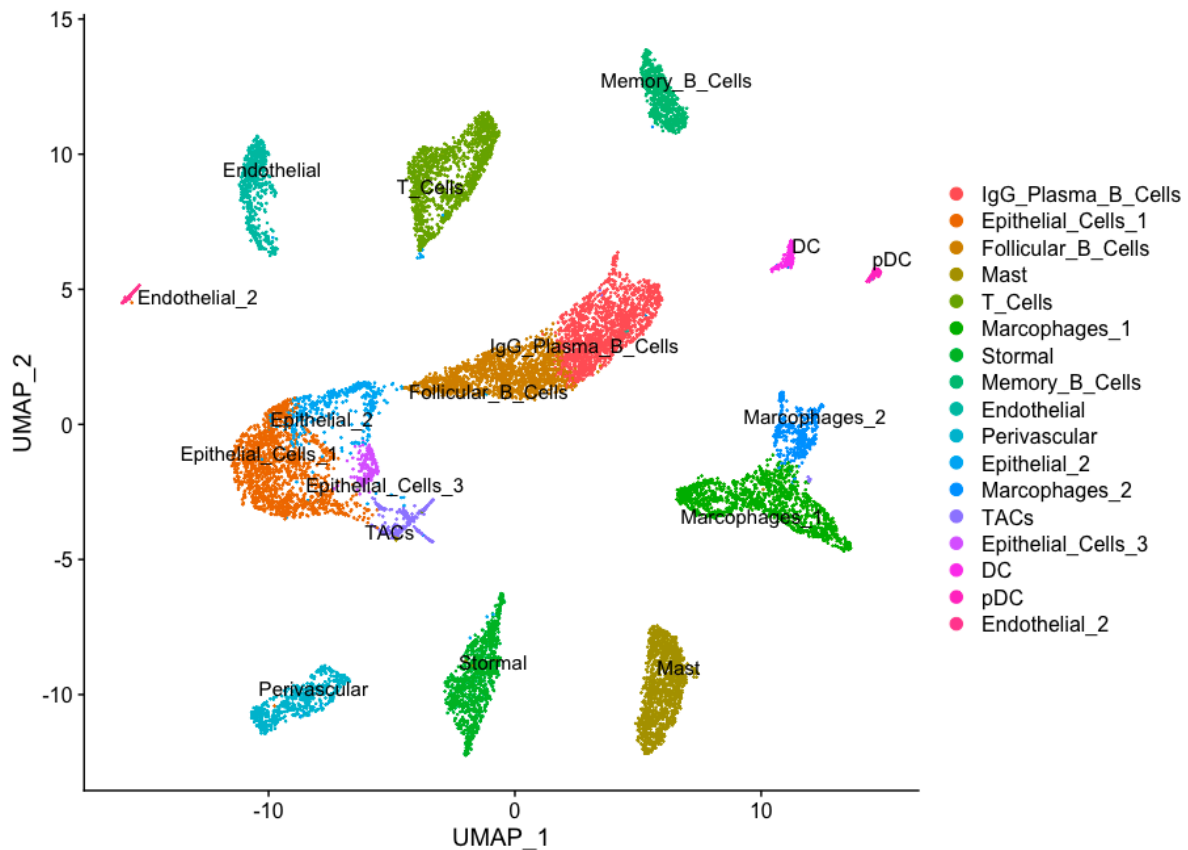


Figure 1.6: Uniform manifold approximation and projection (UMAP) of cell clusters after cell type identification using single-cell RNAseq data from the oral cavity. The figure has been created by Matthew Madgwick processing a public dataset ¹⁹³.

Single-cell approaches are becoming more popular in clinical research. In contrast to bulk RNA analysis, it gives an insight into the cell type and condition-specific gene expression patterns that enables biomedical researchers to better understand the molecular background of disorders ¹⁹⁴ [Figure 1.6]. Due to the complex analysis, data storage requires more space, and computational analysis is more time-consuming than the bulk approach ¹⁹⁵. Details about analysing single-cell transcriptomics are described in Chapters 2 and 3.

The level of detected transcripts does not always correlate with the amount of proteins in samples due to translational regulation. The term ‘proteomics’ was used first in 1995 to describe the analysis of protein content in samples ¹⁹⁶. Due to differences in gene expression patterns, the set of proteins also differs between cells, conditions or individuals. The first step of proteomic profiling is protein extraction from collected samples. Gel electrophoresis facilitates the separation of proteins based on molecular mass and isoelectric points. The next step includes the enzymatic digestion of proteins, the resulting peptides are analysed by mass spectrometry (MS). MS is an essential tool to detect the molecular weight of proteins, completing the analysis with tandem mass spectrometry (two or more mass analysers are

coupled) which then enables peptide sequencing. Finally, computational analyses are used for the identification of proteins based on sequenced peptides. The challenges in the field of proteomics are the following: (1) separating peptides/proteins in the gel can exclude potential candidates that have extreme weight or isoelectric points; (2) proteins with low abundance are may excluded from the analysis, (3) analysis of proteins with lipophilic features (e.g. membrane proteins) ^{197,198}.

Proteins influence the phenotype of cells through their participation in biological processes, therefore, exploring interactions between proteins is crucial in studying cellular behaviour. PPI detection methods (often also called interactomics) have been expanded over the last decade with the appearance of HT screens. Based on the environment, these techniques are grouped into *in vitro*, *in vivo* and *in silico* categories. The main experimental approaches are the tandem affinity purification-mass spectroscopy, affinity chromatography, coimmunoprecipitation, protein microarrays, protein-fragment complementation, phage display, X-ray crystallography, NMR spectroscopy, Yeast 2 hybrid (Y2H) and synthetic lethality [details in Chapter 3] ¹⁹⁹. The *in silico* methods are using computational algorithms to infer PPIs. These tools gain information from *in vivo* and *in vitro* experiments, and predict new potential connections based on these interactions [details in Chapter 3]. There are four possible outcomes of the predictions cases if the interaction is predicted *in silico* and experimentally verified, the result is a true-positive; if experimental evidence was not found then it is a potential false-positive result though future, targeted experimental tests may be needed to verify the PPI's existence in a given living system. Similarly, if two proteins are not connected computationally and there is no evidence for the PPI a true negative prediction occurs, however, if they are found to interact, it is a false-negative prediction outcome ²⁰⁰.

1.6.2 Meta-omics approach

Microbial communities have been studied for decades to understand the complex relationships and interactions between organisms that share the same ecological niche and the function of the community as a whole ²⁰¹. Recent technological advances, including the development of large-scale omics methods, make such approaches possible, where mixed microbial communities are considered as one meta-organism ²⁰².

It is essential to determine the microbial composition in the host because dysbiosis-related diseases (e.g. diabetes, periodontitis, IBD) affect more and more people worldwide. Due to the importance of the topic, the number of microbiome analyses has been rapidly increasing. Toh and Allen-Vercoe revealed that - based on PubMed - around 500 articles included the term 'Human microbiome' in 2008, reaching almost 4000 in 2013²⁰³. At the end of 2021, this number reached 21,594 (source: PubMed - <https://pubmed.ncbi.nlm.nih.gov/>).

The first observation of bacteria was by Leeuwenhoek in 1673. The first artificial bacteria culture, established 200 years later by Louis Pasteur²⁰⁴, let researchers discover a broad spectrum of microbes²⁰⁵. The evolution of microbiology led to a paradigm shift that promoted culture-independent approaches, allowing researchers to explore microbial communities' role in human diseases²⁰⁶.

Culture-dependent methods have several drawbacks, the most important being that there are bacteria that cannot be cultivated in artificial media due to the lack of knowledge about their metabolism and physiological requirements²⁰⁷. The first study describing viable but non-culturable microorganisms, was published in the early 1980s²⁰⁸. This finding established a new direction in microbiology, a sequence-based approach to studying complex microbial communities, firstly using 16S rRNAs^{209,210} and then completing the analysis with whole-genome profiling²¹¹.

Meta-omics data, including metagenomics, metataxonomics, metatranscriptomics and metaproteomics, describes microbial community composition, expressed genes, proteins and metabolic pathways²¹². Each meta-omics data type (layer) reveals a different aspect of host-microbe interactions²¹³. The following paragraphs describe these features and their relevance in HMIs.

High-throughput sequencing-based on methods to study microbial communities

The isolation of microbial genes and genomes from biological samples has extended our knowledge about microbiome composition, especially about unculturable taxa. Metataxonomics and metagenomics provide information about microbiome composition while metatranscriptomics describes regulation in the microbial community.

Metataxonomics explores the diversity of microbiome communities, it reveals the abundance of microbial taxa and also highlights global composition differences between samples. A metataxonomic workflow consists of the following steps: sample taking, DNA extraction, amplicon sequencing of phylogenetic markers (16S rRNAs), processing sequences, taxonomic analysis and comparative analysis²¹⁴.

16S RNA genes are highly conserved across microbial taxa consisting of conserved and variable regions²¹⁵. The choice of primer is essential for marker gene amplification. It should cover most bacterial species using universal primers, but some species remain unresolved. A short sequence, called a barcode, is added to the 5' end of the primers to identify the different samples during the analysis²¹⁶.

Following the amplicon sequencing, quality control steps filter the reads. Reads are nucleotide sequences and depending on the sequencing method, they can be short or long. Quality control software, such as QIIME2²¹⁷ or Mothur²¹⁸, trims the end of the sequences and removes duplicated and low complexity reads. These steps facilitate the selection of high-quality reads without human DNA contamination²¹⁹. Taxonomic profiling with 16S rRNA results in a species-level identification of the microbiome community composition. During the analysis, the processed reads are mapped to reference gene sequences using public databases (e.g. Silva²²⁰ or GreenGenes²²¹). The output is an operational taxonomical unit (OTU) or amplicon sequence variant (ASV) table. OTUs describe sequences extremely similar to each other, represented by consensus sequences from clustering analysis. In contrast, ASVs reveal a single exact sequence with high confidence²²². Finally, alpha and beta diversities are measured to determine and compare microbiome compositions. Bioinformatics tools, such as PICRUST2²²³ or Tax4Fun²²⁴, carry out a functional analysis. The biggest advantage of this methodology is the fast and cost-effective 16S rRNA sequence analysis. Also it can examine correlations between the microbiome community composition and the host condition. However, metataxonomics gives information about taxonomical composition at low resolution and does

not reveal the absolute quantity of microbes ²²⁵. Metagenomics analyses whole genome nucleotide sequences isolated from complex microbiomes ²¹². The workflow is similar to metataxonomic protocols, however this approach analyses the whole genome instead of the marker gene. Following the DNA extraction from samples, shotgun sequencing randomly creates short reads. Assembling the filtered reads into larger constructs, called contigs, can be done by mapping them to a reference genome or using *de novo* assembling methods to identify new genomes. These steps enable gene detection, their functional annotation and finally to taxonomic analysis. The advantage of metagenomics is that it can identify microbes at the strain level. Also, using *de novo* assembling new pathogens can be identified from samples. Nonetheless, gene/genome identification does not give details about gene expression, only the presence of genes is obtainable, besides the analysis is costly and *de novo* assembly is time-consuming and requires a robust computational background ²²⁶.

With these new methodologies, the number of identified microbes has steeply increased. For instance, in the case of the oral cavity, this number has jumped from ~280 different bacteria identified by culture-dependent methods to 700 species ⁸⁰. There are several metataxonomic studies describing the oral microbiome ^{80,227-229}. The first study was published in 1995 about *Haemophilus parainfluenzae* ²³⁰. In the last twenty years, researchers established databases to store reference genomes based on meta-omics experiments. National Center for Biotechnology Information (NCBI), Human Microbiome Project (HMP) ⁶⁴ and Human Oral Microbiome Database (HOMD) ⁷⁸ are the main sources of oral microbiome data. HOMD (<http://homd.org>) is the currently most comprehensive resource which involves core taxa from the literature and 16S rRNA sequences obtained in their laboratory or from GenBank ⁷⁸.

Metatranscriptomics analysis explores the microbial RNA content of samples. This approach provides information on the regulation and expression profiles of complex microbiomes ²¹². Analysing the meta-RNA gives a more detailed insight into the interactions between microbes and between microbes and the host. The workflow consists of experimental steps (sample collection, bacterial extraction, RNA purification and sequencing) and computational data analysis: raw data pre-processing, *de novo* assembly, taxonomic analysis, functional annotation and differential expression analysis. The upstream part of the analysis is very similar to the previous two meta-data analysing approaches. Assembling the high-quality reads into putative transcripts helps to identify the taxonomic composition of the microbial community. Functional annotation is one of the most important steps in the metatranscriptomics pipeline because it infers the functional activity of the microbiome. Differential expression analysis is optional, but it enhances the understanding of an altered condition compared to the control ²³¹. The advantages of metatranscriptomics are that it

captures only living organisms, and due to *de novo* assembly, it does not require reference data. Also, it can compare different communities and their activities²³². However, there is no information about translational and post-translational modifications.

There are challenges in the metatranscriptomics area: on the one hand, the analysis requires many data points/reads by short-read sequencing technologies; on the other hand, longer reads would help the assembling and taxonomical/functional annotations. Completing metagenomics and metatranscriptomics with other approaches (e.g. metaproteomics, metabolomics) can improve the insight into the composition and function of the microbial community²¹³.

Metaproteomics methods to study microbial communities

Metaproteomics describes the protein content of the microbial community in a given sample. The term was used first in 2004 by Rodriguez-Valera²³³ analysing environmental samples. Proteins play the most important role in cellular functions, therefore measuring their abundance correlates with microbial activity²³⁴.

A general metaproteomic workflow consists of sampling, protein extraction and purification, separation of microbial proteins and digesting them into peptides and then mass spectrometry analysis. Databases provide information for the taxonomic analysis or *de novo* peptide sequencing that can discover new proteins. Finally, data interpretation helps to identify pathways and infer information about system functioning^{202,234}.

There are several advantages compared to HT sequence-based methods: metagenomic and metataxonomic data do not provide any insights into microbial activity, and also, data typically include numerous genes with unknown functions (Ram et al., 2005). Besides, metatranscriptomics does not allow translational regulation to be considered²⁰².

More and more metagenomic and metataxonomic data are becoming available, however, only a small number of metaproteomic studies have been reported. There are experimental and data analysis-related challenges in this field. In the 'Host omics' section, I mentioned the major limitations, however, metaproteomics includes additional computational challenges. Due to a lack of complete bacterial sequences (experimentally cultured, sequenced and characterised strains), mass spectrometry data analysis is challenging, and peptides can be mapped to a variety of homologous proteins from different species^{201,235}.

1.7 Data processing and modelling

Discovering the role of the microbiome in host signalling requires integrating data from different sources. Omic data is highly interconnected, each approach explores the samples from a different point of view. Modelling cellular behaviour by multi-omics data analysis requires systems-level representation and analysis to facilitate the understanding of complex, interconnected processes²³⁶. Systems biology aims to integrate and model complex biological processes and their interactions. Instead of focusing on one object in an experiment, it gains a holistic view of cellular processes in response to external stimuli^{236,237}.

1.7.1 Network biology approaches

Appearance of HT and omics technologies in the past two decades has led to large data generation and rapid development of the computational biology area. Using networks facilitates the representation and visualisation of large data. There are two main directions of network modelling: (1) static networks can represent and integrate small-scale and HT data sets, but the objects and their interactions are not changing, (2) in dynamic models, the network structure changes over time, the approach is used for computational simulations and mathematical modelling^{238,239}.

Graph theory and network science

Networks describe pairwise connections between organisms or objects. The entities in the network are called 'nodes' and interactions between them are 'edges' [Figure 1.7]. Graph theory is part of mathematics and computer science but it is also applied in several other areas of science, e.g. physics, sociology, and medicine. The definition of the network (theory) is similar to graph (theory) but not the same. Graph theory is often described as the mathematical foundation of network science²⁴⁰. Also, the terminology differs between the two objects: a point is called 'vertex' in graphs but 'node' in the network. Similarly, there are 'edges' between the vertices in a graph and 'edges' or 'links' between the nodes in a network. In biological networks, nodes can be different kinds of molecules (e.g. RNAs, genes, proteins), organisms or pathways, interactions can be physical relationships, associations or even regulatory connections²⁴¹.

In a network, links can be directed or undirected. The first term describes a connection between source and target nodes (e.g. regulation), while in the second case, interactions do not have directions between the nodes (e.g. co-expression). Edges also can be unweighted or weighted. An unweighted network represents equal connections between entities; edges in a weighted network are measured by weight (e.g. strength of an interaction) ²⁴¹. Besides holistic data analysis, networks are usually used for data visualisation. Node (size, colour, shape, label, etc), edge (thickness, colour, etc) and network (layout) attributes facilitate understanding patterns in large data sets ²⁴² [Figure 1.7].

Network topology refers to the arrangement of nodes and edges and gives information about networks' sub-structures. In terms of network analysis in the thesis, the most important topological parameters are degree, hub, and shortest path. The degree of a node is the number of interactions a node has in the network. Unlike the average in the graph, nodes with much higher connectivity are called hubs. These points have a huge impact on the network, removing hubs from the network leads to disconnected graphs ^{241,243,244}. Translating it to the field of biology, the mutation or deletion of these genes/proteins often leads to a lethal phenotype (e.g. knock out of chaperon proteins) ^{245,246}. The shortest path measures the minimal number of edges which connect nodeA to nodeB ^{243,244}, it is equal to the functional distance between two molecules ²⁴⁷.

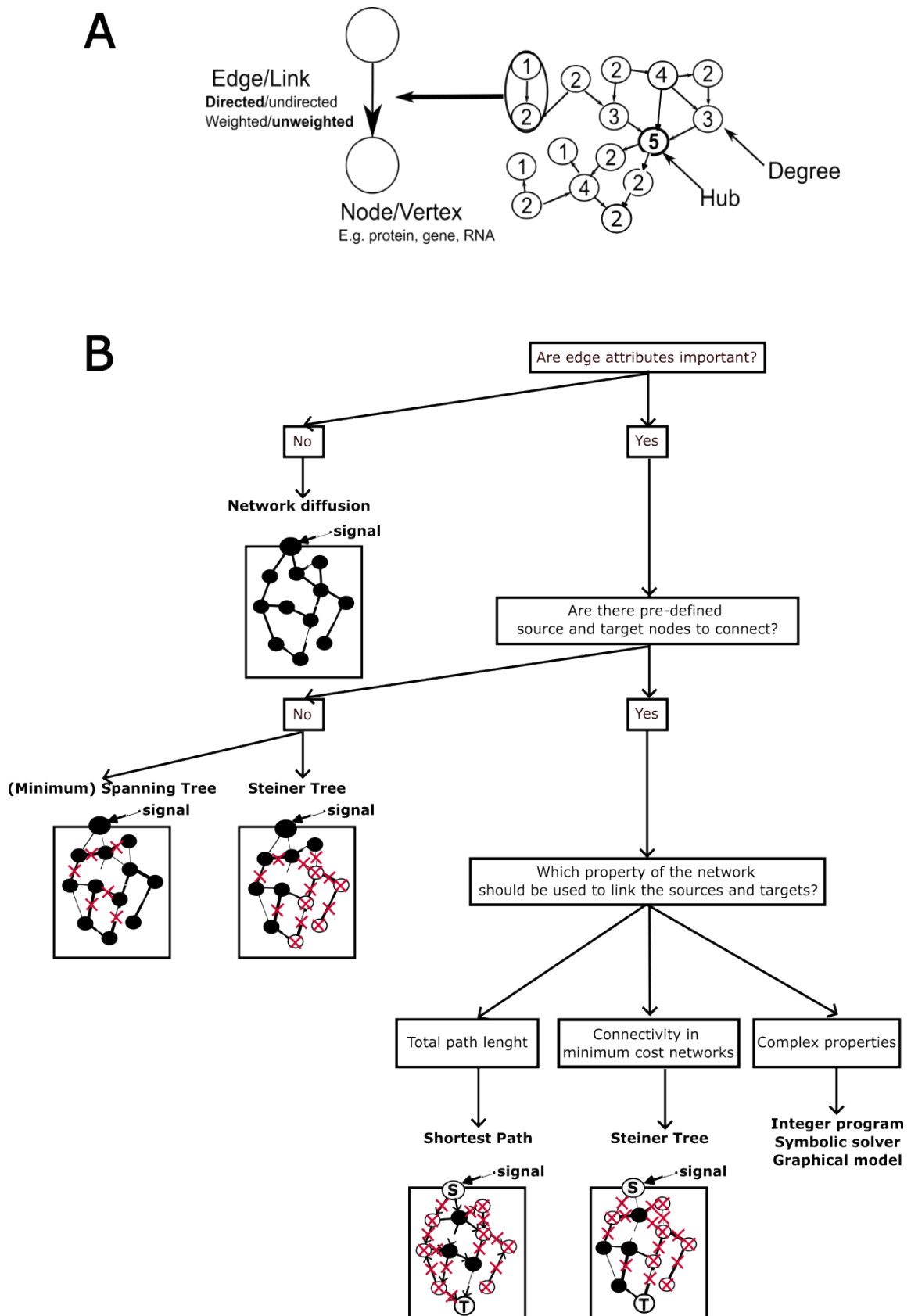


Figure 1.7: Introduction to networks and graph-based pathfinder algorithms. A, Structure of a directed molecular network highlighting its topological parameters used in the thesis. Numbers in the circles show the degree number. B, Classifying the key pathfinding algorithms discussed in the thesis (S – source node, T – target node). The figure was drawn by myself.

Computational algorithms are able to identify paths in the network using different topological parameters. As large hairball networks are difficult to be analysed, the path finding algorithms filter to the enriched subnetworks in the whole graph. Although there are several algorithms to solve this issue, the approaches are different based on the type of the graph and its parameters.

Network diffusion algorithms propagate information through the network based on the connections between nodes²⁴⁸. One example of a network diffusion algorithm is the Random Walk with Restart algorithm. This algorithm simulates a random walk on the network, where at each step, the algorithm has a probability of moving to a neighbouring node or staying at the current node. This process is repeated many times, with the goal of reaching a steady state in which the probability of being at each node is proportional to the degree of connectivity of that node²⁴⁹. Network diffusion algorithms are used by TieDie²⁵⁰, NBS²⁵¹ and NBS2²⁵², mND²⁵³ and many other tools (the full list is available in this review²⁴⁸).

In contrast, there are algorithms that handle weighted graphs and look for the optimal subnetwork including the edge attributes. The spanning tree algorithm creates a loop-free subgraph including the selected nodes, the minimum spanning tree aims to connect the vertices through edges with the minimum weight. [This algorithm is often used in neuroscience analysing the connectivity of the brain](#)²⁵⁴. [PHYLOViZ](#)²⁵⁵ [and CySpanningTree](#)²⁵⁶ [tools are examples for the usage of the spanning tree algorithm.](#)

There are a group of algorithms that connect pre-determined start (source) and end (target) points in the network. The shortest path algorithm uses the weights of the edges to find the path that minimises the total distance. This method is often used in molecular networks, for instance for inferring regulatory networks²⁵⁷. The shortest path is often used to estimate the functional distance between two molecules and identify functional clusters in the network²⁴⁷. CARNIVAL²⁵⁸, PesCa²⁵⁹ and PathExt²⁶⁰ are examples for shortest path using tools.

The Prize-collecting Steiner Forest (PCSF) algorithm infers a subnetwork including most of the selected nodes (terminals; e.g. expressed genes in a transcriptomics dataset) from the network connecting them with the minimum weights of edges. The novelty of the algorithm is that the nodes which were not selected, so called Steiner nodes, can establish a bridge between terminals that are not directly connected. The algorithm gives penalties for the following nodes: Steiner nodes, hub nodes and terminals that can't be connected to the subnetwork. Hub nodes are misleading in the graph as these points (or molecules in a biological network) usually have diverse connections (or functions) and disturb the simplicity of the graph. The goal of PCSF is to minimise the cost and maximise the prize in the subnetwork therefore that subnetwork tries

to include the most terminals and less hub and Steiner nodes connected by the least weighted edges ²⁶¹. Such tools are the web-based OmixIntegrator (<http://fraenkel-nsf.csbi.mit.edu/omicsintegrator/>) ²⁶² and the PCSF R package ²⁶¹.

Biological networks

Biological networks represent relationships between molecules and organisms but also between regulatory, signalling and metabolic pathways. In contrast to experiments, these models reveal complex interactomes and patterns of biological systems. The main aim of these models is to integrate, analyse and visualise complex data ²⁶³. This thesis focuses specifically on molecular networks to discover cellular behaviour on systems-level under different conditions.

The appearance of transcriptomic data and network biology approaches inferred new methodologies to analyse gene regulation on systems level. Gene regulatory networks consist of genes and DNA/RNA or protein molecules connected through regulatory interactions. These models can be used for (1) causal mapping of molecular interactions between transcription factors (TF) and their target genes (TGs), (2) guiding experimental design by highlighting potentially important regulatory interactions, (3) identifying biomarkers, (4) comparing the regulatory profile of diverse conditions (e.g. healthy vs. diseased), (5) drug design ²⁶⁴.

This thesis focuses on the transcription factors (TFs) regulating different genes in inflamed conditions and connects them to the activated signalling by the bacteria. TFs are proteins including DNA binding domain which binds to a specific DNA element (enhancer, silencer or promoter region) and enhances or inhibits gene transcription depending on the binding region. Based on a study, there are ~1600 potential TFs in the human genome ²⁶⁵. Altered TF - target gene interactions disrupt the normal gene expression pattern resulting in disorders ²⁶⁶.

Cross-talk between signalling pathways coordinates biological processes in the cell. Signals flow through molecular interactions such as protein-protein or metabolic interactions and biochemical reactions. The aim of signalling networks is to understand the communication system that controls cellular behaviour in different environments and conditions ²⁶³.

This thesis focuses on the role of altered signalling in inflammation compared to a healthy condition through analysing PPI networks. PPI analysis is a major focus of systems biology due to the pivotal role of proteins in cellular behaviour ²⁶⁷. The global human protein

interactome describes all the PPIs which are currently known by experiments or *in silico* predictions. This large network highlights that proteins can have diverse sets of interactors but currently, it is less studied how these PPIs vary in time and differ between tissues or cell types²⁶⁸.

1.7.2 Databases and tools

Molecular databases (DBs) are structured sets of different kinds of data essential for computational biology. These collections involve experimentally verified and computationally predicted information about molecules and their interactions in different organisms. The number of biological databases has been steeply increasing, based on articles, there were 281 molecular databases in 2001²⁶⁹, while this number was between 500 - 1000 in 2003²⁷⁰. With the appearance of HT experiments and omics data, this number rose to more than 1700 based on an analysis in 2018 which explored the published DB articles in *NAR (Nucleic Acids Research)* journal²⁷¹.

HT screens and omics approaches expanded the knowledge about the existing genes, proteins and their features (e.g., sequence, structure). From 2004 - when UniProt was published - until 2015 around 90 000 000 protein sequences have been described. Based on Chen *et al*, gene and protein DBs can be grouped into the following sets: sequence databases, 2D gel databases, 3D structure databases, chemistry databases, enzyme and pathway databases, family and domain databases, gene expression databases, genome annotation databases, organism-specific databases, phylogenomic databases, polymorphism and mutation databases, protein-protein interaction databases, proteomic databases, PTM databases, ontologies and specialised protein databases²⁷².

It was reported in 2005 that although there are numerous biological DBs, a high percentage of them are not up-to-date due to a lack of stable funding for these projects²⁷³, due to the steeply increasing number of DBs, the situation is even worse in 2022. In the following sections, I would like to introduce the main DBs/tools involved in the development of workflows presented in this thesis.

Sequence databases

UniProt Consortium is a central resource for protein sequences and annotations. From its four databases (UniProt Knowledgebase (UniProtKB), UniRef, UniParc and Proteomes), UniProtKB and Proteomes have been used in the projects. UniProtKB combines reviewed/curated Swiss-Prot entries with the unreviewed TrEMBL identifiers (IDs) that are annotated by automated systems. Currently, there are more than 65 million UniProtKBs in the database, an increase of >50% in just 2 years. Every protein in the database has a profile where its annotations are available. This knowledge consists of the protein sequence, function, taxonomy, subcellular location, post-translational modification (PTM), expression, interactions and structure by collecting external databases (e.g. Gene Ontology database (GO), Pfam) and literature evidence ²⁷⁴.

UniProt Proteomes consists of 20 125 reference proteomes and 327 987 non-reference proteomes ²⁷⁴. Regarding their distribution in superkingdoms, there are 238 208 proteomes in Bacteria, 103 543 proteomes in Virus, 3 172 proteomes in Archaea and 3 189 proteomes in Eukaryota as of 30/08/2021 (source: www.uniprot.org/proteomes/).

Protein structure databases

Technological improvements established an increased number of experimental methods which detect protein structures, such as X-ray crystallography, Nuclear magnetic resonance (NMR) spectroscopy, or cryo-electron microscopy ²⁷⁵. Nevertheless, the number of *in silico* structure prediction algorithms (e.g. homology-based prediction by BLAST ²⁷⁶) is steeply raising which led to an explosive growth of known protein structures.

The smallest structural unit is the motif, a short, conserved amino acid sequence associated with distinct functions of proteins. Short linear motifs (SLiMs) - sub-sequences of usually 3 to 20 amino acids - are essential for dynamic PPIs therefore they have an important role in signalling ^{277,278}. Eukaryotic Linear Motif (ELM - <http://elm.eu.org/>) is a computational resource for SLiM collection. The database annotates experimentally verified motifs and arranges them into classes based on the functions which SLiMs mediate. Motifs are flexible patterns, there is no need to know the whole sequence, usually regular expressions describe SLiMs ²⁷⁹.

Domains are tertiary structural components which are often functional units on their own. These compact folded components are key regulatory participants of signalling²⁸⁰. Pfam DB - developed by the European Molecular Biology Laboratory (EMBL) - is the largest domain collection; there were 19 632 entries derived from multiple organisms in the DB in December, 2021²⁸¹.

Regarding the tertiary or quaternary structural levels, there are DBs which give information about the 3D structure of proteins. Protein Data Bank (PDB) is a central resource that collects information about experimentally verified 3D structures of large biomolecules²⁸². Besides, there are other DBs, like ModBase²⁸³, SCOP²⁸⁴, SWISS-MODEL Repository²⁸⁵, that infer protein structures by comparative, evolutionary or homology modelling.

Protein-protein interaction databases

PPI DBs collect outcomes from small and large-scale experiments [details in Chapter 3] but some of the resources integrate data from *in silico* predictions as well²⁸⁶. Most DBs use the standardised PSI-MI (proteomics standards initiative - molecular interaction) format to store interaction data²⁸⁷. This XML-based data type unifies details about experiments to avoid overlapping information deriving from diverse databases.

Currently, the most popular, frequently updated PPI DBs are STRING²⁸⁸, IntAct²⁸⁹ and BioGrid²⁹⁰. Bajpai *et al* collected 375 PPI resources and selected the top 16 databases for comparative analysis²⁹¹. Among the examined parameters, there are the number of total PPIs, experimentally verified interactions and exclusive interactions. The study concluded that STRING is the most ideal resource to collect the most interactions, also, this database contains the most information about experimentally verified links.

In an ideal case, every database should contain the same information using the same publications, but there are differences in curation efforts. Also, there is a long list of protein or gene IDs which are used by molecular databases, such as protein name, gene symbol, Uniprot ID, Ensembl ID, Gene ID, Refseq ID. Mapping the IDs links the databases, but it is not a simple process, because of the different versions and redundancy of IDs²⁹². There are many algorithms which help to solve this problem using different approaches²⁹²⁻²⁹⁶ although it is important to keep it in mind that manual curation causes an initial difference among data repositories.

Not only the ID mapping causes differences in data curation but also DBs often filter or rank the interactions based on scores estimated by a diverse set of parameters. For instance, STRING DB offers confidence scores measured by the type of interaction evidence (text mining, experiment, data in another database, co-occurrence, co-expression, etc); and transferred scores when an interaction has been described in another organism and through homologue/orthologue prediction the two proteins are connected in the species ²⁸⁸. Both scores have a value between 0 (two proteins are not interacting) and 1 (two proteins are likely to interact).

The structure and content of PPI DBs have been improved in the last decades. Although there are still differences between them, using a standardised format and the hierarchical annotation of interactions instead of filtering facilitate the data integration. During my PhD, I have been involved in the development of the OmniPath database, therefore a detailed section (Chapter 2) describes this molecular interaction resource.

Pathway databases

Biological pathways include interactions between molecules that facilitate the signal spread through the cell. Pathway DBs contain two main types of information, a list of pathway members and/or interactions between molecules.

Reactome is a freely available DB, which contains manually curated data about signalling and metabolic molecules and their relation to pathways in multiple species. Regarding the human organism, it contains 10720 proteins, in 2546 pathways. ReactomeDB describes not only pathways but also splits them into reactions. Currently 13890 reactions exist in the database as of 03/12/2021. The database has R and Python packages to use its data automatically but it is available through a website (<https://reactome.org/>) where graphical views are available for each reaction ^{297,298}.

KEGG (Kyoto Encyclopaedia of Genes and Genomes) is a large integrative biological resource which consists of 16 databases. KEGG Pathway (<https://www.genome.jp/kegg/pathway.html>) - developed in 1995 - has collected manually curated reference paths and computationally predicted organism-specific paths ²⁹⁹. Compared to Reactome, it contains information about only 540 pathways (last updated: 24 March 2022), also KEGG uses more broad terms to describe pathways. All in all, currently, KEGG is less suitable for pathway analysis but still works for enrichment analysis.

Signalink3 (SLK3 - <http://signalink.org/>), developed by our group, is currently one of the largest signalling pathway resources. In contrast to ReactomeDB and KEGG Pathway, Signalink3 contains information about pathway regulators on diverse levels (transcriptional, post-translational, etc) in humans and other popular model organisms. The database has details about 13 pathways: RTK (Receptor Tyrosine Kinase, containing all MAPK and Insulin subpathways), TGF- β , Wnt, Hedgehog, JAK/STAT, Notch, NHR (Nuclear Hormone Receptor), B- and T-cell receptor, Hippo, Toll-like receptor and innate immune pathways. Currently (December 2021) there are 17,918 proteins and more than 700 000 interactions between signalling molecules ¹⁶⁵.

Ontology databases

Ontologies describe and classify the context of a biological entity (interaction, protein, etc) thereby facilitating the data analysis and giving a focus for studies. These terms include diseases, developmental stages, molecular functions, location, anatomy, pathways, etc. While these annotations contribute to the context-specific analysis, from a DB infrastructure point of view, it is challenging to handle and standardise ontologies ³⁰⁰. The Unified Medical Language System (UMLS) addressed this issue by developing standardised biomedical terminology for annotations ³⁰¹. UMLS integrates ontologies from several databases, such as OMIM ³⁰², NCBI Taxonomy ³⁰³ and GO ³⁰⁰.

GO - developed by Gene Ontology Consortium - annotates genes and their products in a tree-like structure where parent and child categories are represented in a hierarchical way. The ontologies are grouped into three sets: molecular functions, cellular location and biological processes. GO is not species-specific therefore the database enables cross-species comparisons ^{304,305}.

1.8 Aims and Objectives

This chapter introduced the main areas of biology and bioinformatics that are covered by the PhD project. I highlighted the current challenges in the existing methods, such as analysing large microbiome data and exploring cell type and condition-specific host-microbe interactions. These gaps aimed to be addressed by new technologies, such as omics approaches, including single-cell sequencing) and systems biology methods. Therefore, the primary research aim of this iCASE PhD project was to establish computational pipelines to predict host-microbe interactions and their cellular effects based on multi-omic data analysis using network biology approaches. The following objectives have been defined to achieve the goals of this PhD project:

1. Computational analysis of intercellular communication using single-cell transcriptomics data.
2. Distinguish and list healthy and inflammation-related bacterial strains of the gastrointestinal tract, and predict their condition-specific interactions with the host using multi-omics data.
3. Functional analysis of the microbiome targeted host proteins to reveal the processes directly affected by bacteria.
4. Development of standardised, semi-automatic bioinformatics pipelines to enable reusability, and make in silico interspecies and intercellular analysis accessible for researchers without strong computational background.

Chapter 2 - Development of *in silico* approaches to study intercellular communication

2.1 Introduction

In multicellular organisms, cells are interacting with their environment and also with each other through a vast spectrum of molecules that ensures the growth and differentiation by spreading the signal from cell to cell. In junctional interactions, cells are physically connected *via* various structural complexes while cells interact through chemical signals in cell-cell communication [details in Chapter 1.3.1].

Both junctional interaction and cell-cell communication are crucial in the epithelial layer, these connections allow the cells to grow, differentiate and proliferate properly. This coherent surface establishes a barrier that separates the outer environment, the living space of external microbes, from the internal milieu, including stromal and immune cells. Hence, the interaction between epithelial and hematopoietic cells contributes to tissue homeostasis. Inflammation causes impaired cell-cell interactions hence disrupting the continuous layer that allows immune cell infiltration. For instance, the malfunction of tight junction structures leads to altered cytokine secretion, resulting in new cell-cell interactions^{99,117} *via* immune mediators. This chapter focuses on the development of a semi-automated pipeline that infers cell-cell interaction networks from single-cell transcriptomics data. In a case study, I explored altered cell-cell interactions in inflammatory bowel disease (IBD).

The knowledge about intercellular communication is scattered across different resources. Despite its importance, the molecular background is less discovered due to the lack of data. As mentioned in details in Chapter 1.6.2, there are existing methods to connect cells using predictions [REF] or combine experimentally verified knowledge with computational pipelines [REF]. However, the effect of the altered intercellular interactions on downstream signalling is less discovered. This gap has been addressed with the combination of single-cell omics and network biology approaches that provide an insight into the gene expression and molecular interactions of individual cells [details in Chapter 1.6.2]¹¹⁴.

Signalling databases are crucial for omics data analysis [details in Chapter 1.7]. OmniPath is an integrated, literature-curated resource for signalling pathways. The first version was published in 2016 and consisted of 27 popular interaction resources describing the human interactome. Not only protein-protein interactions (PPIs) were represented in the database, but OmniPath also provided rich annotations on the properties of proteins, including function, localisation, and role in diseases.

In a collaborative project with Julio Saez-Rodriguez's group in Heidelberg, we updated OmniPath in 2020, this time combining over 100 resources into one single database. The new version covers the interactions and role of proteins in signal transduction and also transcriptional and post-transcriptional regulations. Besides, OmniPath became available for mice and rats via homology translation and includes information about intercellular signalling.

There are existing databases describing ligand-receptor or junctional interactions, also there are resources that highlight intercellular protein annotations. The novelty of OmniPath is, firstly, the data integration that reveals new potential cell-cell interactions through merging annotations and existing PPIs. Secondly, OmniPath is accessible via the web service at <https://omnipathdb.org/>, as a Cytoscape plugin ³⁰⁶, and packages in R/Bioconductor (*OmnipathR*) and Python (*pypath*), providing convenient access options for both computational and experimental scientists ¹⁰⁷.

I contributed to the computational development of the '*pypath*' Python module, and carried out a quality control check of the intercellular interactions and annotations using the literature. I also demonstrated the capabilities of the new OmniPath through the implementation of a case study about intercellular communication in ulcerative colitis (UC).

Inflammatory bowel disease

IBD describes disorders that cause chronic inflammation in the gastrointestinal tract. Its symptoms range from mild (e.g. fatigue) to severe (e.g., abdominal pain and blood in the stool). The number of people suffering from IBD has increased steeply in the last decade. In the past, IBD has mainly affected developed, Western countries (based on studies in 2018, the highest prevalence was in North America ³⁰⁷), while today studies show that IBD is more and more prevalent in more recently industrialised countries, such as China and India ³⁰⁸. In 2020, Based on the analysis carried out by The Global Burden of Disease Study published that in 2017 around 3.9 million females and 3 million males were living with IBD ³⁰⁷. IBD is a multifactorial

disorder, several external (environment, diet, age, etc) and internal (genetic background, microbiome, immune-mediated tissue damage, etc) factors influence its emergence ³⁰⁹.

The two major forms of IBD are Crohn's disease (CD) and ulcerative colitis (UC). In CD, inflammation can affect the small or large intestine and can be continuous or involve multiple segments (skip lesions) ³¹⁰. UC affects the colon and the rectum, compared to CD, inflammation appears in the mucosal layer avoiding the deeper submucosal layers. While IBD is not curable, various treatments can reduce the symptoms and ensure remission ³¹¹.

A recent single-cell study ³¹² revealed that intercellular connections were changed in UC. The altered ligand-receptor connection affected the dynamic of cell populations. For example, the elevated level of IL-18 cytokine in inflamed enterocytes led to an increased amount of Treg cells due to IL-18 receptor expression on their surfaces ³¹². The triggered receptors on immune cells mediate pro-inflammatory cytokine expression causing an amplified inflammatory response in the gut and leading to an imbalanced immune response ³¹³. The limitation of their approach is that the authors focused on cell type- and condition-specific LRIs. However, the developed intercellular interaction pipeline discovers adhesive interactions as well without restricting the analysis to LRIs between cell markers and differentially expressed genes.

This chapter focuses on an *in silico* pipeline that establishes ligand-receptor interaction networks combining single-cell transcriptomics and network resources. The public data analysis expands the interactions to junctional connections between cells and identifies gaps in our knowledge about cellular communication in inflammation. The intercellular interaction pipeline and the case study were published in *Molecular Systems Biology* ¹⁰⁷.

2.2 Methods

The intercellular interaction pipeline discovers intercellular rewiring between diverse cells using single-cell RNAseq data from healthy and diseased conditions. Transcriptomic data describes a list of genes with expression values. There is a need for a reliable network resource that describes potential PPIs to infer cell-cell connections. The pipeline builds up contextualised networks by combining the two kinds of information to highlight the cell type-specific signalling. OmniPath provides inter- and intracellular interactions and protein annotations to infer cell-specific signalling networks [Figure 2.1].

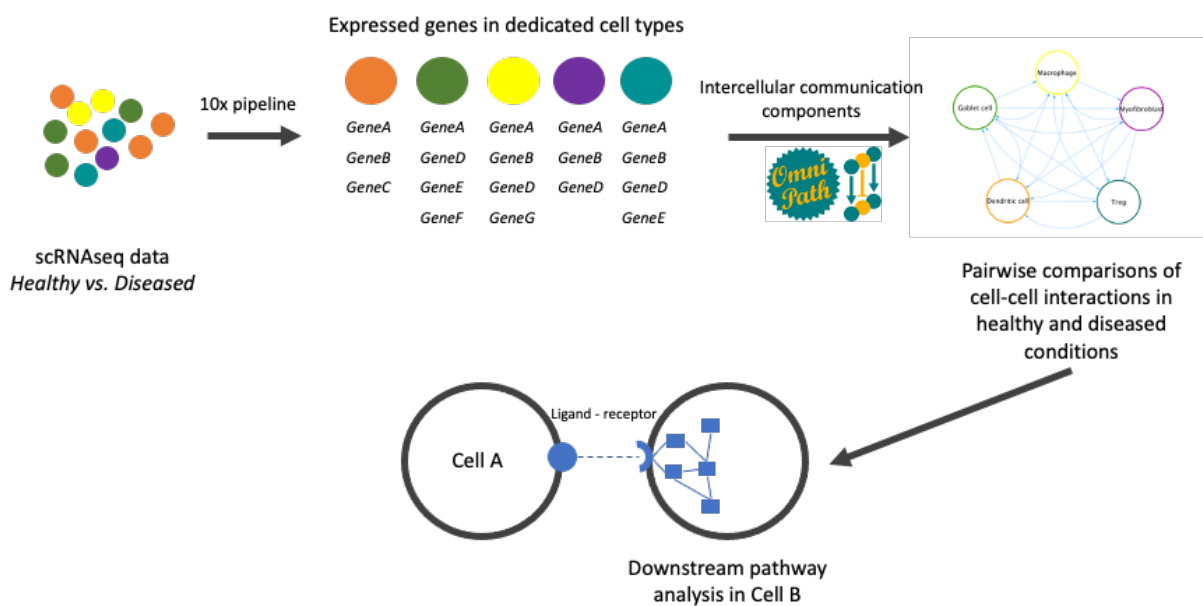


Figure 2.1: Workflow for analysing intercellular interaction and their downstream effect.

2.2.1 Identifying intercellular interactions among different cell types

I downloaded all intercellular interactions from OmniPath using the OmniPathR R package (version 3.15 on Bioconductor) and filtered the interactors based on their subcellular locations in OmniPath. OmniPath collects information from many resources including Gene Ontology DB ³⁰⁴, UniProt ²⁷⁴, Human Protein Atlas ³¹⁴, LOCATE ³¹⁵, ComPPI ³¹⁶ and a literature collection ¹²³ (more details about the script collecting location information can be found here: https://github.com/saezlab/pypath/blob/3820c3a28c13ce701f1d2b5f9ac6e00834c757da/pypath/core/intercell_annot.py). I discarded extracellular matrix proteins and regulators of intercellular proteins (ligand-, receptor- and matrix adhesion regulators) as these molecules usually appear in the cytosol. Therefore I focused on membrane-based or secreted proteins (membrane-based or secreted ligands, membrane-based receptors, tight junctions, gap

junctions, desmosomes or other adhesions, ion channels, transporters and cell surface or secreted enzymes). Importantly, because proteins are multi-functional molecules, some interactions were duplicated due to the diverse protein annotations (e.g. A2M protein has both adhesive molecule and receptor annotations).

2.2.2 Single-cell data processing

I analysed a publicly available single-cell RNA-seq published by Smillie *et al*³¹² to explore interactions between cells in the intestinal tract. The dataset was utilised in the study due to several key factors. Firstly, it was deemed representative of the research problem, with a focus on inflammatory bowel disease. Secondly, the sample size was substantial, with biopsies collected from 18 patients with the disease and 12 healthy individuals. Additionally, the research methods and techniques used in the dataset were highly relevant, as it was the first (and only) available dataset at the time of analysis (in 2019) that examined gene expression patterns at the cellular level, comparing samples of healthy, non-inflamed, and inflamed ulcerative colitis. The processed scRNAseq data included 51 cell types from epithelial, immune and stromal cell lineages. Finally, the dataset was highly accessible, with both raw and processed data available for public use. Matthew Madgwick, a PhD student in our group, developed an internal pipeline, called ScOmix, to analyse transcriptomic data and processed the published raw single-cell RNAseq dataset (available at Single Cell Portal under SCP259 ID) using the original parameters from the article³¹². Briefly, the Cell Ranger pipeline³¹⁷ was used for processing single-cell RNAseq data prior to analysis according to the instructions provided by 10x Genomics. The resulting FASTQ files were aligned to the human reference genome GRCh38/hg19 and subsequently filtered and count files generated for each sample.

The gene expression matrices of healthy, non-inflamed and inflamed samples were integrated together for cell annotation and direct comparisons. Then entries with a few genes were filtered to remove any dead or dying cells from the data. To account for differences in sequencing depth across samples, expression values were normalised for total Unique Molecular Identifiers (UMIs) per cell and the counts were log-transformed. The highly variable genes were selected for downstream clustering to confirm that the clusters matched the original annotations.

Output files described the average expression of genes under healthy, non-inflamed and inflamed UC. I selected the healthy and non-inflamed UC conditions, to study the effect of intercellular interactions on cellular behaviour.

2.2.3 RNA-seq data filtering

I filtered the average gene expression matrix to discard the lowly expressed genes because they frequently derive from the technical or biological noise of the experiment. In general, z-score transformation helps to standardise data across a wide range of values³¹⁸. The z-score (also known as a standard score) is a measure of how many standard deviations an observation or data point is from the mean of a distribution. It is calculated by subtracting the mean of the distribution from an individual data point, and then dividing by the standard deviation of the distribution. A z-score can describe how unusual a data point is within a distribution. A z-score of 0 indicates that the data point is exactly at the mean of the distribution. A z-score of +2 or -2 would indicate that the data point is two standard deviations away from the mean.

Hart *et al* published a z-score-based normalisation method that determines which genes were expressed using a comparison between expressed genes and active promoters³¹⁹. While the authors applied it for FPKM data (Fragments Per Kilobase of gene model per Million mapped reads ratio), I adapted their methodology and used it for log₂-based expression values instead. I kept genes where the z-score was greater than -3, a cut-off suggested by the authors³¹⁹. This value includes those genes where the expression value is higher than three times the standard deviation below the mean.

2.2.4 Reconstructing a cell-cell interaction network

I implemented a Python script to build up cell-cell interaction networks based on a predefined list of selected cell types. In the case study, I selected five cell types from the processed single-cell dataset: goblet cells, myofibroblasts, DCs, Tregs and macrophages. These cells have a crucial role in intestinal homeostasis and are involved in UC pathogenesis^{320–324}. I combined the intercellular interactions from OmniPath with cell-specific gene expression patterns derived from the single-cell transcriptomic dataset and examined all possible connections of cells by pairwise comparisons. The focus of this study was on rewired cell-cell interactions during UC, therefore, I selected the condition-specific interactions between cells. I defined condition-specific interactions by their exclusive appearance either in the healthy or in the diseased state. The extent of the condition-specificity was measured by the number of unique intercellular PPIs in healthy and UC samples [Figure 2.2].

2.2.5 Building up ligand-receptor interaction networks between myofibroblasts and regulatory T cells

I analysed the cellular communication between myofibroblasts and Tregs in more detail focusing on ligand-receptor interactions. I grouped the similar ligands (e.g., CCL2 and CCL3 = CCLs) and merged the connections within groups. Although this misses the different effects of paralog ligands, it results in a simplified LRI network that highlights the main ligand-activated downstream signalling. I assigned pathways to the receptors defined in the SignalLink3 database ¹⁶⁵ to improve biological insight and visual clarity [Figure 2.3]. One receptor can be part of several pathways, hence I selected the most relevant one using knowledge from the literature. Differences between conditions have been visualised by the Circos R package ³²⁵.

I established a Treg-specific signalling network using intracellular interactions from the OmniPath *Cytoscape* application ³⁰⁶ and single-cell data, limiting the large PPI network to genes expressed in Tregs. I focused on the upstream part of the triggered pathways by ligands, therefore the receptors and their first two neighbours (proteins, that the receptor can reach in two steps) were selected for a pathway enrichment analysis using the online interface of the Reactome pathway knowledgebase ²⁹⁷ (<https://reactome.org/>) with its default settings (hypergeometric test, Benjamini–Hochberg FDR correction, the human genome as the universe gene set).

2.3 Results

2.3.1 Semi-automated pipeline to build cell-cell interactomes

The primary workflow focuses on ligand-receptor interactions (LRIs) between source and target cells in healthy and diseased conditions. It consists of two parts: building up the cell-cell interaction network (Python script) and visualising the LRIs on circos plots in R. The inputs are (1) intercellular interactions (built in table, derived from OmniPathR), (2) a processed single-cell transcriptomic dataset describing the average gene expression (user-provided) and (3) a list of cells that will be connected and compared in healthy and diseased condition (user-provided). Currently, the pipeline is able to handle the gene expression data in a fixed table format describing the genes in the first column and the cell types and condition in the further columns. Therefore the pipeline is not sensitive to how the user pre-processed the transcriptomic dataset.

The pipeline can be downloaded from GitHub (https://github.com/korcsmarosgroup/uc_intercell). Following the cloning of the repository enables the user to run the pipeline. The intercellular interactome can be built up from the Terminal, using the following command: *python intercell_pipeline.py --scdata 'path to the single-cell transcriptomics' --cells 'list of interacting cells'; while the visualisation takes place in RStudio, running the `circos_LRI.R` script.*

2.3.2 Analysing intercellular interactions in healthy and diseased colon

I analysed the filtered average gene expression matrix from single-cell data in healthy and non-inflamed UC samples [Table 2.1] and combined them with intercellular interactions to build up the cell-cell interaction networks. From the 22,550 PPIs derived from OmniPath, I discarded regulators - as these molecules mostly appear in the cytosol - and extracellular matrix proteins - due to focusing on direct cell-cell interactions - which resulted in 22,283 PPIs between 1800 source proteins and 2074 target proteins connecting cells. I combined the expressed genes from the five cell types with the general intercellular PPIs to observe the cell-cell connections.

Focusing on the differences in cell-cell interactions between the healthy and diseased colon, I filtered the potential intercellular PPIs to condition-specific connections (PPI represented only in healthy or diseased conditions). Although each cell could potentially bind to each other, the type of communication was divergent based on the results. I found significantly fewer

intercellular PPIs in UC networks. Interestingly, the number of interactions targeting myofibroblasts or Tregs remains the same in both conditions [Figure 2.2, Supplementary Table 2.1]. Supplementary Table 2.1 indicates the difference in the number of PPIs between the conditions. The outcome of the analysis shows that LRIs and adhesion connections are dominating in both conditions between cells, probably due to the high number of these PPIs in OmniPath (9439 adhesive and 9565 LRIs). In contrast, I found 76 PPIs between the cells that describe tight junction, desmosome and gap junction connections. These results indicate that intercellular communication varies among cells, moreover the analysis suggests that cell-cell interactions are potentially weaker in UC.

Table 2.1: Number of expressed genes in cell types		
Cell type	Healthy colon	UC
Goblet cell	13 744	12 561
Myofibroblast	11 884	12 135
Dendritic cell	10 558	7 501
Regulatory T cell	11 881	11 609
Macrophage	14 225	14 092

Based on the results, intercellular communication appears different between the five cell types in UC condition compared to healthy cell-cell interaction networks. Macrophages, Tregs, goblet cells and myofibroblasts target the dendritic cells with significantly more interactions in healthy condition. In contrast, the focused targets are the Tregs in UC. The other four cell types express more ligands and adhesive molecules that reach the membrane-based target proteins on Tregs' surface in diseased condition [Figure 2.2].

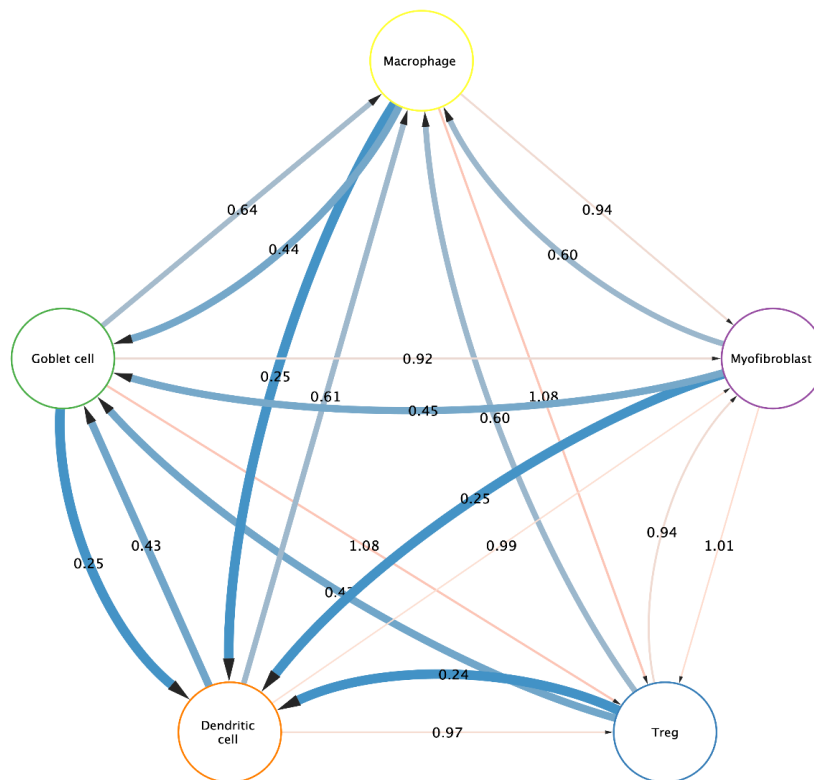
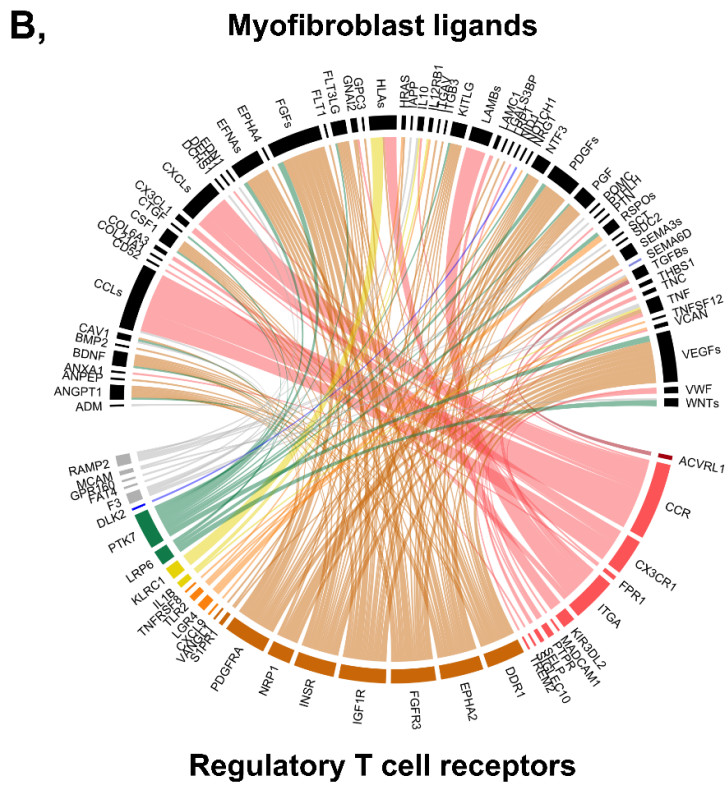
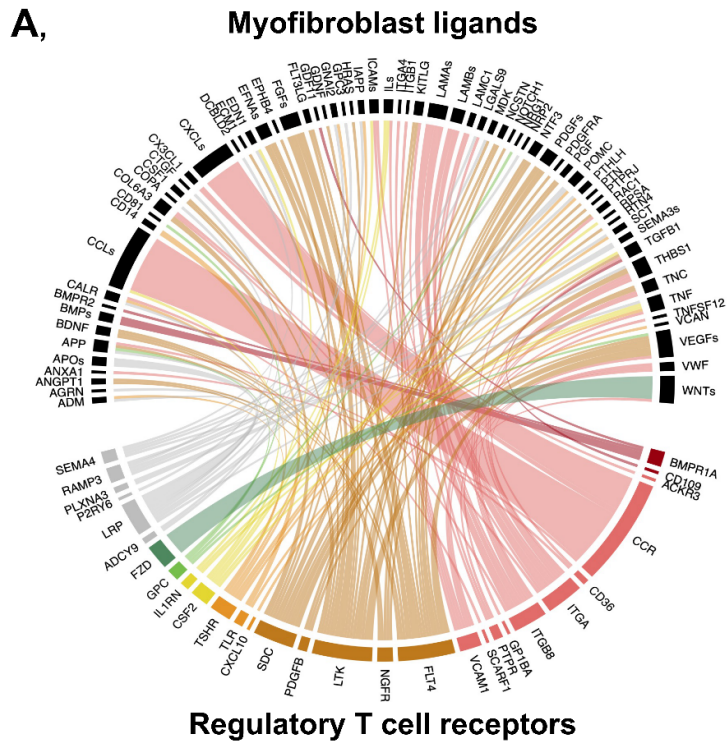


Figure 2.2: Pairwise comparison of cell-cell interactions. The thickness and colour of the edges indicate the ratio of cell-cell interactions in UC relative to those in a healthy state. The labels on the edges display the exact rate. The blue-red colour scale highlights the differences between conditions, with blue indicating an increased number in healthy samples, and red indicating a shift in the ratio towards the UC condition. Network is visualised in Cytoscape³²⁶.

2.3.3 Effect of myofibroblasts on regulatory T cells

I chose the interaction between the myofibroblast and Treg for further analysis. The reason for highlighting this cell-cell interaction was that the number of the interactions between cells remained similar in both conditions (472 PPIs in healthy colon, 478 PPIs in UC colon), however, the function of the corresponding proteins found to be altered during the disease. The analysis revealed 208 LRIs in healthy- and 304 LRIs in diseased colon. The latter shows a ~30% increase of the annotated communication in cellular communication in UC. At the protein level, the 208 LRIs occurred between 32 ligands and 41 receptors, while 36 ligands and 41 receptors established the 304 disease-related LRIs. Figure 2.3 depicts circos plots highlighting interactions between myofibroblast and Treg. As the analysis focused on condition-specific interactions, these results revealed that there could be unique connections in both states. Although the number of ligands and receptors is similar, the raised amount of LRIs supposes a more active cellular communication in UC.



Receptor pathway: ■ TGF-beta ■ Immune ■ RTK ■ TLR ■ Other
■ Hedgehog ■ WNT ■ Notch ■ JAK/STAT

Figure 2.3: Condition-specific connections between myofibroblast ligands (upper semicircles, black) and Treg cell receptors (lower semicircles, coloured by pathways) in A, ulcerative colitis and B, healthy control. Immune—innate immune response, RTK—receptor tyrosine kinase, TLR—Toll-like receptor. Circos plots were created by using the ‘*circlize*’ R package ³²⁵.

I analysed the role of the target receptors on Treg cell surfaces in downstream signalling pathways. I found that all pathways derived from SignaLink3 database (TGF-beta, innate immune response, receptor tyrosine kinase, Toll-like receptor, Hedgehog, WNT, Notch and JAK/STAT pathways) were affected at some level by myofibroblast ligands. The ligands had an impact on all of the pathways in both conditions, however, different receptors driving the same signalling were targeted on T cells. The distinct upstream interactions potentially cause varying downstream signalling in Treg cells.

I built up a Treg-specific signalling network for each condition based on the scRNAseq-derived contextualised interactome to analyse the downstream effect of altered LRIs. I created a subnetwork including the targeted receptors and proteins within two steps (interactions) from the receptors. The filtered network consisted of 835 proteins in healthy and 1971 proteins in non-inflamed UC condition. This potentially suggests more tight regulation of the Treg cells by the myofibroblasts in UC but there is also a chance that there are more understudied processes in terms of healthy data. According to Reactome, MAPK, TLR6/2 and TLR7/8 pathways were enriched among the 835 proteins in the healthy colon, while in samples from UC patients TLR4 and TLR3 pathways were overrepresented ($p < 0.05$; $FDR < 0.05$) [Figure 2.4, Supplementary Table 2.2]. Based on the results, there is a potential shift towards inflammation-related pathways during the disease.

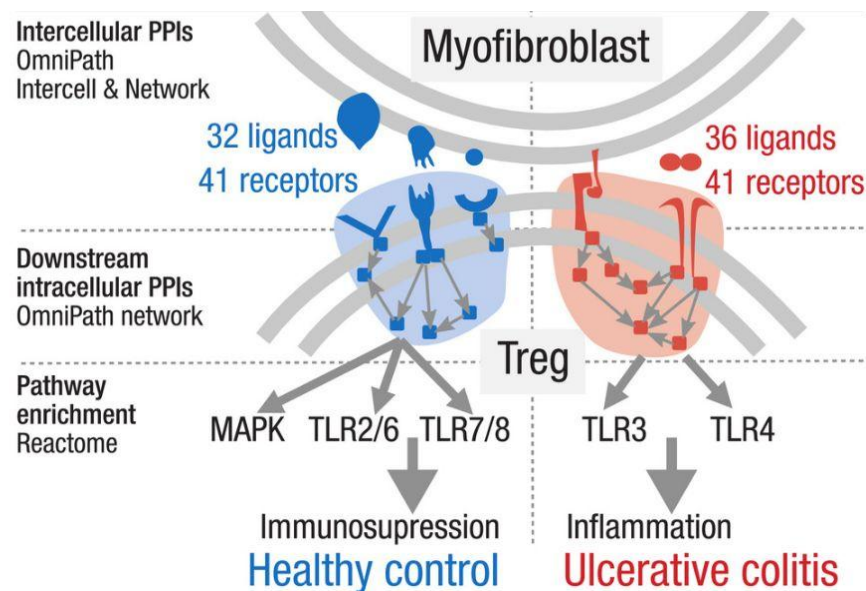


Figure 2.4: Intercellular connections and their downstream effect in UC compared with healthy control. Condition-specific ligand–receptor connections between myofibroblasts and regulatory T cells trigger an immunosuppressive versus inflammatory signalling in T cells, in healthy and UC, respectively.

2.4 Discussion

In this chapter, I introduced an *in silico* analysis that infers cell-cell networks using OmniPath, an integrated resource for intra- and intercellular interactions. The semi-automated pipeline consists of two main steps, (1) building up an intercellular network in Python and (2) visualising the LRIs using R. The algorithms offer a solution for biologists with limited programming skills as the requirements are a single-cell dataset and a list of cells to connect. The output offers a potential overview about the key differences in cell-cell interactions between conditions visually on a circos plot. Although the intercellular interaction pipeline uses OmniPath, a new and integrated resource to study inter and intracellular signalling, it relies heavily on one database including only the known interactions, besides it does not prioritise interactions. In the future, we plan to integrate scores which contribute stronger to indirect causal relationships. Also, we are going to include text mining approaches to extend the interaction annotations to conditions, therefore reducing the number of false positive PPIs. Another limitation is that the pipeline processes gene expression data but infers PPI networks. Combination of transcriptomic data with (phospho)proteomics could solve this issue. Although phosphoproteomics is popular among bulk data, there are studies that describe single-cell approaches to identify the cellular signalling on individual cell level ^{327–329}. Due to the focus on functional proteins and their interactions, this omics data would be the most suitable for the accurate estimation of pathway activity. As the current circos plot can handle one signalling (edge colour), if a receptor attends in more than one process, the user should define manually which one is the most important in terms of the study. This can be done with a shorter list, like in this case study, but having a large list of LRIs, it remains challenging. Currently, a potential solution is to create multiple plots highlighting each pathway and the related interactions separately. Finally, the annotation of receptor proteins is not complete, I used OmniPath including data from CellPhoneDB ¹¹¹, Guide2Pharma (<https://guide2pharma.com/>), HPMR ³³⁰, Gene Ontology ³⁰⁴, and two literature-derived resources ^{123,331}, hence the analysis can miss potentially important signalling pathways. Despite these limitations, the intercellular interaction pipeline gives a new insight into context-specific cellular interaction. The scripts and examples are accessible through GitHub (https://github.com/korcsmarosgroup/uc_intercell).

In a case example, I discovered public single-cell transcriptomic data from colon samples deriving from healthy and UC patients and selected five cell types representing epithelial, stromal and immune cell populations. I pointed out a possible altered cellular communication and adhesive structures in non-inflamed UC compared to healthy colon. Ligand-receptor interactions and their downstream effect were explored between myofibroblasts and Tregs, two cell types that play crucial roles in UC pathogenesis ^{323,332}.

Although studying intercellular communication on the systems- level is a relatively new approach - due to the requirement of single-cell omic data - there are other methods and pipelines which analyse relationships between cells ^{312,333-335}. Smillie *et al* not only published the scRNAseq datasets used in this study but also discovered cell-to-cell interactions in the collected samples. My motivation to use the same dataset and carry out a similar analysis was to show how OmniPath can provide a potentially more precise insight into physical cell-cell interactions and cellular communication. The main focus of the authors was on the rewiring of cell-cell interactions through ligand-receptor connections in UC. They used the FANTOM5 database ¹²³ as a source of LRIs. Additionally, instead of considering all of the expressed ligands and receptors they only selected the cell markers and differentially expressed genes between conditions. This filtration step resulted in fewer LRIs and more compact networks where not every cell was connected to the others. Based on their analysis, in healthy condition, DCs and T cells are described as hub nodes in the network, while in non-inflamed UC interactions were enriched between epithelial cells and fibroblasts and T cells ³¹².

Their results are not contradicting the output of the workflow I presented in this chapter, however their methodology highlights cell-cell interactions in a different point of view. By limiting the analysis to markers and DEGs, the output focuses on differences at cell type level while I explored the rewiring at molecular level. In contrast to their analysis, my workflow (1) also explores cell adhesion structures; (2) uses OmniPath as an integrated resource of intercellular interactions and protein annotations providing a larger coverage of the known LRIs; (3) discovers and assesses the affected pathways downstream in the target cells.

I compared the five cell types with each other, and found a key difference between the intercellular interaction networks deriving from healthy and UC conditions. Based on the outcome of the pipeline, cells are tightly connected to DCs in healthy condition, while in UC condition this tendency shifts in the direction of regulatory T cells. Because there is no experimental validation for these results, the findings are only assumptions. However, this hypothesis is not contradictory to the currently available information in the literature: DCs are professional antigen-presenting cells that recognise surface molecules on other cells through diverse cell-cell interactions, therefore maintaining normal immune response ³³⁶; Tregs are immunosuppressive in general, however, in IBD patients their phenotype and gene expression pattern are altered ³²³ which potentially lead to more intense communication with other cell types in the gut.

The analysis indicated a potential increase in ligand-receptor interactions between myofibroblast and Treg during UC that caught my attention. Therefore I carried out a downstream pathway analysis in Treg that showed overrepresented MAPK and TLR signalling in the disease. Based on the literature, the enriched MAPK signalling pathway plays a key role in the immunosuppressive function of induced Tregs in healthy conditions³³⁷. The also health-related TLR2 and TLR7 pathways facilitate the maintenance of Th17 (pro-inflammatory T cell subpopulation) and Treg balance and increase the immune suppression function of Treg^{338,339}. In contrast, in non-inflamed UC, the overrepresented TLR4 and TLR3 pathways contribute to inflammatory cytokine expression³⁴⁰⁻³⁴². This evidence supports the fact that in healthy condition, regulatory T cells protect against inflammation, while in non-inflamed UC this starts to deteriorate partially by the myofibroblasts.

The pipeline gives a detailed insight into the rewired intercellular interactions during inflammation, but some limitations need to be improved in the future. Most importantly, the analysis relies on data coming from OmniPath. Although the resource includes the major molecular databases (such as IntAct, BioGrid) and keeps them updated, the results rely on the available interactions and intercellular annotations. Besides, the workflow handles transcriptomic datasets, which give gene expression information, while the inter- and intracellular networks describe connections between proteins. I assumed all genes expressed translate to proteins. Also, the case study and its conclusions are based on one scRNAseq dataset, but the 10X approach could miss some potentially expressed genes. Hence, the finding that there is a weaker communication between cells in UC could be deceptive due to the lower read depth in single-cell sequencing¹⁹⁴. Besides the biases in the data analysis, conclusions about the type of intercellular interactions can be misleading. Usually, databases provide information about ligand-receptor and adhesive interactions, less interactions describe tight junctions, gap junctions, desmosomes and ion channels. I tried to overcome this limitation by focusing on the altered ratio and differences between ligand-receptor and adhesive interactions and ignoring small interaction categories.

In conclusion, we established an integrated resource, called OmniPath, in a collaborative project with the Saez group that details inter- and intracellular interactions collected from more than 100 sources. I built a computational workflow combining OmniPath with public omic data to address gaps in the current knowledge about context-specific cellular communication. My colleagues are now teaching the use of my workflow to the future generations of systems biologists in training courses at the EMBL European Bioinformatics Institute.

Chapter 3: Discovering the effect of the human microbiome on host cell signalling

3.1 Introduction

Systems microbiology uses analyses of omics data to understand the interactions between microbial cells or communities and their host. During the evolution of host-microbe interactions, three main directions have emerged: infection, colonisation and commensalism. Infection describes a process when pathogenic microbes enter the host and start to replicate potentially leading to diseases³⁴³. Colonisation describes the presence of microbes in the host without causing damage or disease³⁴⁴. The term 'commensalism' has been used in literature to define multiple processes, such as '*The ability (of a microorganism) to live on the external or internal surfaces of the body without causing disease*'³⁴⁵. The current terminology says that commensal bacteria do not induce damage after colonising the host however they can elicit an immune response³⁴⁶. When both the microbes and the host benefit from the interaction the connection is mutualistic. However, depending on the environment (e.g impaired immune activity or altered microbiome), commensal bacteria can turn into pathogens causing damage and potentially disease in the host^{346,347}.

Rewired host-microbe interactions can lead to disease in the host. Hence, understanding the detailed cellular communication between the microbes and the host has become a significant research area. While earlier studies focused on the role of single microbial strains in disease³⁴⁸ in the last decades, the integration of systems biology approaches [details in Chapter 1] drastically improved the host-microbe interaction detection methods. Techniques have been shifted from *in vivo* / *in vitro* experiments toward *in silico* predictions. Chapter 1 describes the experimental approaches shortly, Table 3.1 summarises the major types of *in silico* algorithms with examples.

Table 3.1: <i>In silico</i> PPI prediction approaches		
Method	Description	Resources/tools
<i>Orthology-based approaches</i>	Orthologous proteins share similar sequences, therefore, an experimentally found PPI can be predicted in another organism using sequence alignment	POINT ^{349–354} , pathDIP ³⁵⁰ , IsoBase ³⁵¹ , InParanoid ³⁵² , IID ³⁵³ , Singh <i>et al</i> ³⁵⁴
<i>Gene expression profile based approaches</i>	Genes belonging to the common expression-profiling clusters are more likely to interact with each other ²⁶⁷	Enright <i>et al</i> ³⁵⁵
<i>Phylogenetic profile-based approaches</i>	Inferring PPIs based on the evolutionary history of proteins ²⁶⁷	COG ³⁵⁶
<i>Domain-domain interaction-based approaches (Structural composition-based prediction)</i>	The general assumption is that domainA binds to domainB, then proteinA carrying domainA interacts with proteinB having domainB ³⁵⁷	PPIDomainMiner ³⁵⁸
<i>Domain-motif interaction-based approaches (Structural composition-based prediction)</i>	Similarly to DDI, the known interaction between structural components is used to infer connections between proteins. Further details in Chapter 2.	LMPID ^{359,360}
<i>Machine learning algorithms</i>	Machine learning combines several protein features (e.g. amino acid composition, hydrophobicity profile) to predict the probability of the interaction. It requires true-positive and true-negative interactions for the training used to train the algorithm. It is currently the most powerful approach for PPI prediction ²⁰⁰ .	InterSPPI ³⁶¹

Revealing cross-species interactions is challenging due to obtaining multi-omic datasets from the same sample³⁶², therefore there is a need for computational pipelines to overcome this problem. *In silico* algorithms attempt to predict interactions between molecules at a systems level. There are several approaches, a common characteristic is that all of them look for similarities between molecules and interactions (e.g. similar sequences, expression patterns, structural composition or evolutionary history)³⁶³.

The thesis focuses on detecting host-microbe PPIs, but microbes interact with the host through metabolite- and RNA-mediated interactions as well. . The following paragraph details the structural-based domain-motif interaction (DMI) -based PPI detection method.

Proteins do not have a simple linear shape, these macromolecules are organised in complex 3D structures. Knowledge about structural properties can improve the PPI networks ³⁶⁴. Primary structure describes the proteins on the amino acid (AA) level. Short linear motifs (SLiM) are short amino acid sequences (3-20 Aas) containing conserved positions. Secondary protein structure (e.g alpha-helices and beta-pleated sheets) highlights smaller organised parts of the molecules, while the tertiary structure is a 3D construction that is folded into functional units, called domains ³⁶⁵. Regarding the tertiary structure of the proteins, globular and fibrous constructions can be distinguished. Fibrous proteins are long-shaped molecules consisting of repetitive amino acid sequences that are less sensitive to changes in the environment, such as pH or temperature. Usually, these proteins have a structural role (e.g. actin or collagen). Globular proteins are more compact and built up from irregular amino acid sequences. In contrast to fibrous protein, they have functional roles (e.g. enzymes) ³⁶⁶.

The irregular amino acid sequence in globular proteins is described as unstructured and flexible regions without regular structure ³⁶⁶. These intrinsically disordered regions (IDRs) play a pivotal role in the host-microbe interactions. IDRs are determined based on the primary protein sequence composition by identifying parameters, such as disorder-promoting hydrophilic features and charged amino acids. Most of the prediction tools use machine learning algorithms to combine these features and determine potential IDRs ³⁶⁷⁻³⁷³.

SLiMs on IDRs provide binding sites for domains and the established PPI plays an important role in signalling pathways ³⁷⁴. Currently, resources describing DMIs or even SLiMs are limited. Eukaryotic Linear Motif (ELM) resource ³⁷⁵, Linear Motif mediated Protein Interaction Database (LMPID) ³⁵⁹, interActions of 76ocatio domAiNs (ADAN) ³⁶⁰, and the database of three-dimensional interacting domains (3did) ³⁷⁶ provide structural and interaction-related information.

There are existing studies and methods which use DMIs to infer PPI, all of them are based on structural information derived from the ELM and/or 3did resources. Zhang *et al* discovered DMI-based PPIs between grass carp and grass carp reovirus ³⁷⁷; Halehalli and Nagarajaram established a workflow to study viral-human PPIs ³⁷⁸; Evans *et al* discovered human – HIV PPIs and I was also contributing to a study to reveal PPIs between bacterial pathogens and human autophagy proteins ³⁷⁹.

All of these studies focus on discovering various types of host-microbe interactions, however they lack details regarding the tissue/cell and condition specificity of those interactions. The workflow, presented in this chapter, aims to address this limitation and fill this crucial

knowledge gap in the field. It combines structural *in silico* PPI detection methods with host omics data analysis and predicts the effect of extracellular microbes on host signalling pathways in tissue- and cell type-specific ways. Moreover, I reconstructed tissue/cell type-specific intracellular signalling to analyse the downstream signalling effects of the microbes. In addition, completing the network modelling with functional analysis gives an overview of the potential changes in the signalling flow and cross-talks between pathways due to diverse host-microbe interactions.

3.2 Methods

The original host-microbe interaction pipeline has been published as ‘MicrobioLink’ in 2020³⁸⁰. However, during my PhD, I started to work on a newer version, called MicrobioLink2. In the following sections, I would like to introduce the existing workflow and highlight the improvements in MicrobioLink2. A detailed description of the practical application of the algorithm, as well as information on its ease of use and input requirements are described in Chapter 4 and Chapter 5.

As the pipeline potentially will be used for commercial purposes by the industrial partner, I checked the licences for the tools used in the pipeline [Figure 3.1]. In the project, I did not use any resource which was not been updated in the last 10 years or the website was not available 381–385.

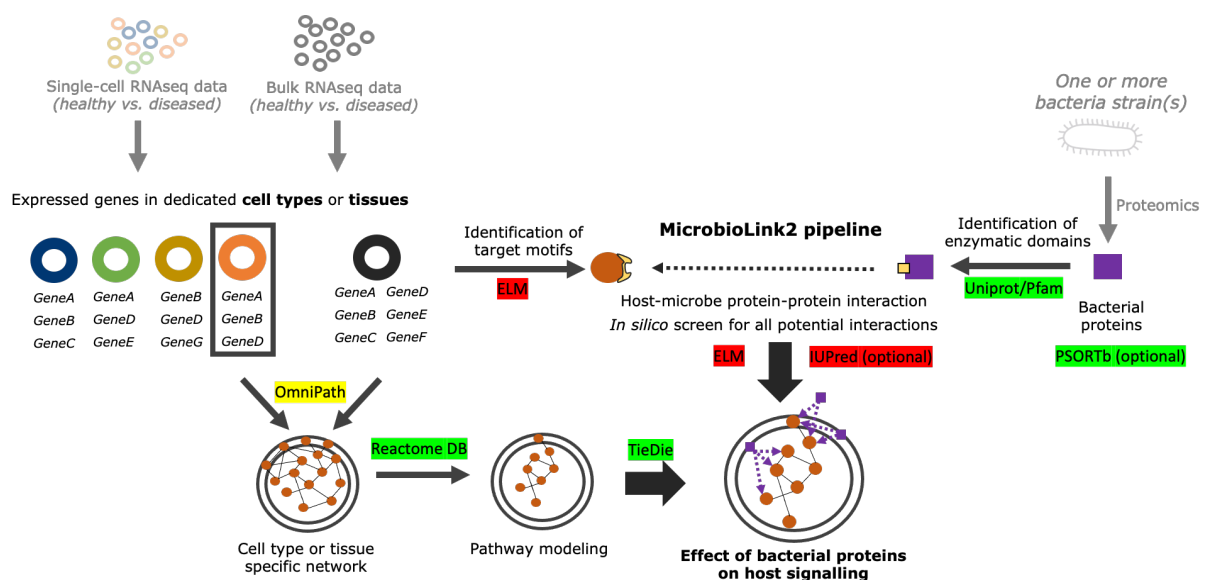


Figure 3.1: Workflow of host-microbe interaction prediction. Pre-processing of host and microbial proteins (highlighted by transparent colours) is not included in the pipeline. Resources with green background: no restriction for commercial use. Yellow background: some resources require a licence for commercial users. Red background: commercial use requires a licence.

3.2.1 Location analysis of proteins

The pipeline focuses on extracellular bacteria that interact with the host cell membranes and mediate their effect through ligand-receptor interactions or disrupting the cell adhesion structures^{386,387}. Hence, secreted and membrane-based proteins are essential for bacteria to contact and communicate with the membrane-based host proteins³⁸⁸. I implemented a subcellular location analysis of microbial and host proteins into the workflow, however, it is an optional filtration step for bacteria due to the low number of the microbial proteins with known or predicted annotation.

Microbial proteins

I used the PSORTb computational tool (v.3.0.3) for the subcellular location analysis of bacterial proteins. PSORTb is available both online (<https://www.psort.org/psortb/>) and as a Docker container³⁸⁹. To avoid the installation of docker service by the users, in MicrobioLink2, I built in the original data from the database into the pipeline which contains microbial proteins and their predicted subcellular location scores³⁹⁰. Depending on the type of the analysis, using microbial locations is optional in the pipeline. PSORTb is licensed by the GNU General Public License v2.0 and is available for everyone both in academia and industry.

Due to the different membrane structures of bacteria, the prediction is different in Gram-positive bacteria, where the microbe has a thin inner plasma membrane and an outer thicker peptidoglycan cell wall compared to Gram-negative strains which have a thin plasma membrane, peptidoglycan layer and an outer membrane. Depending on the membrane structure, the locations can be cytoplasm, cytoplasmic membrane, periplasm, outer membrane/cell wall and extracellular space³⁹⁰.

The location prediction algorithm of PSORTb consists of multi-analytical modules including SCL-BLAST & SCL-BLASTe, Support Vector Machines (SVMs), Motif & Profile Analysis, Outer Membrane Motif Analysis, ModHMM, and analysis of signal peptides. *SubCellular Localization BLAST* reveals the homolog of proteins with known subcellular locations using Blast-P search³⁹¹ (comparing a protein query to SCI-BLAST database). *Support Vector Machine Modules* are machine learning-based methodologies which help the algorithm to assign proteins to locations based on positive and negative training sets [see Chapter 1.7.2 for more details]. *Motifs analysis* is based on specific motifs which determine the location of proteins in the cells. Gram-negative bacteria have unique beta-barrel proteins in their outer membranes. PSORTb

collected over 250 motifs which characterise these structures and therefore decides whether the query protein is part of the outer membrane structure or not. *ModHMM* identifies transmembrane proteins by the hidden-Markov model (a statistical method to predict the sequence of unknown variables based on a set of observed features). Finally, *signal peptides* are a specific part of the protein sequence which determine the subcellular location of the protein³⁹⁰.

PSORTb uses RefSeq IDs in the prediction files, therefore I implemented an ID translation script, which maps RefSeq to UniProt ID. Also, the script downloads the corresponding protein sequence for the RefSeq ID. As an input, the user should provide a table including the organism ID of the bacteria of interest and its Gram-state for prediction. The output of this step is a list of membrane-based or secreted microbial proteins annotated with strains in which these molecules are expressed. Although this step is optional in the pipeline, it can give a more focused interactome between the microbiome and human tissues/cells.

Human proteins

Subcellular location of human proteins derived from OmniPath (<https://omnipathdb.org/>) which is a curated, regularly updated resource merging the main molecular databases in the field [details in Chapter 3]. OmniPath collects information from many resources including Gene Ontology DB³⁰⁴, UniProt²⁷⁴, Human Protein Atlas³¹⁴, LOCATE³¹⁵, ComPPI³¹⁶ and a literature collection¹²³ (more details about the script collecting location information can be found here: https://github.com/saezlab/pypath/blob/3820c3a28c13ce701f1d2b5f9ac6e00834c757da/pypath/core/intercell_annot.py). To download plasma membrane proteins I implemented an R script using the OmniPathR package¹⁰⁷.

3.2.2 Host-microbe protein-protein interactions

In Microbiolink2, I modified an existing DMI-based PPI prediction algorithm from the previous version that connects bacterial domains to SLiMs on human protein sequences and it reduces the number of false-positive PPIs through IDR prediction on host binding sites.

The knowledge of the existing bacterial domain structures was limited³⁹² till 2021 when AlphaFold2 was published³⁹³. This artificial intelligence-based method exploded the field of structural biology by predicting more than 200 million protein structures, including bacterial proteins. The current version of Microbiolink uses InterPro³⁹⁴ for structural analysis, which

provides an automated sequence analytical program for domain prediction called InterProScan³⁹⁵. It includes several protein signature recognition methods to identify Pfam domains based on the FASTA sequence. SLiMs derived from the ELM database³⁹⁶, to avoid too general motifs, only SLiMs with a length greater than two amino acids were used in the analysis.

Bacterial domains are able to bind SLiMs on the target protein sequence. The idea behind the DMI algorithm is that proteins that carry these structural components are able to establish directed PPI networks where the source is the bacterial protein and the target is the human protein.

The first version of MicrobioLink used experimentally verified domain-motif interactions from the ELM database (data from 2013). In the new MicrobioLink2 pipeline, I updated ELM (data from 2021) and also added 3did³⁷⁶ to the resources [Figure 3.2]. ADAN³⁶⁰ was not updated since 2009 therefore I did not use it for predicting PPIs. Although LMPID³⁵⁹ was published in 2015, the database does not contain Pfam IDs and regular expressions for motifs, therefore I did not implement it into the workflow. Merging databases is challenging because resources use a diverse set of IDs to describe interactions, hence the first step is mapping them to a common identifier. In MicrobioLink2, the motifs are represented by regular expressions and the domains are described by Pfam IDs.

Analysing motifs allows the implementation of a quality control step into the pipeline that reduces the number of false-positive interactions in the networks. Among the available IDR prediction tools, I chose IUPRED, because originally, MicrobioLink uses this tool (version 1). The algorithm discarded those motifs which appeared out of disordered regions because these parts of the proteins are rigid and there is less chance that they can be caught by domains³⁷³. The tool uses scores based on two methods (IUPred and ANCHOR2) to measure residue-level energy terms. The energy terms correlate with how intrinsically disordered the protein region is. Higher disordered regions are more accessible for the bacterial domain. Two cut-off values (IUPred > 0.5 and ANCHOR2 > 0.4 - defined in the source article) were set up to select human motifs which are presented out of globular domains and at an intrinsically disordered protein region³⁷³. Both IUPred and ANCHOR2 scores represent the probability of a given residue being part of a disordered binding region, but ANCHOR2 provides two additional methods to estimate the energy associated with interaction with a globular protein³⁷³.

Here, I was working on the integration of IUPRED (version 2) into the pipeline. The tool requires as input the result of the DMI-based PPI interaction list and gives back a score describing how many AAs are on a disordered region. It accepts those motifs as 'disordered motifs' where a maximum of one amino acid is out of the IDR. Consequently, the output of the analysis gives back a reduced list of PPIs between bacterial domains and human motifs.

3.2.3 Network propagation algorithms

I discovered two different network propagation approaches in detail during my PhD. Both CARNIVAL²⁵⁸ and TieDie²⁵⁰ implement causal network approaches to model the signal flow between the bacteria-affected human proteins and genes from host transcriptomics datasets. However, there are differences in the tools regarding the diffusional algorithm, input data and the focus of the algorithms.

CARNIVAL (CAusal Reasoning pipeline for Network identification using Integer VALue programming) is available as an R package. It requires several input files including the start point where the signal comes from (triggered receptor in the case of the pipeline), endpoints of the signal (differentially expressed genes (DEGs)) and a directed signalling network which came from the OmniPath database³⁹⁷. CARNIVAL provides a subnetwork which reflects the transcriptional footprint of samples. CARNIVAL uses DoRoThEA³⁹⁸ and VIPER³⁹⁹ to directly predict transcription factor (TF) activity from the gene expression dataset, as it skips the additional interaction step between regulatory TF - target gene (TG) in the signalling network.

DoRoThEA is a source of TF - TG interaction, these regulatory interactions come from (1) the literature, (2) ChIP-seq experiments, (3) TF binding motif predictions and (4) inference from gene expression data. It consists of five categories based on the reliability of the interaction: Category A (highest confidence) includes interactions from at least two literature curated resources or described in one study but proved by the previously mentioned (2)-(4) methods. Category B involves interactions supported by a curated resource AND ChIP-seq data; or detected by methods (1), (3), (4); or detected by methods (2), (3), (4). Category C contains interactions either from curated resources AND TF binding motif prediction; or from ChIP-seq data AND TF binding motif prediction. Category D involves interactions deriving from one curated resource or from ChIP-seq data. Category E (least reliable) describes interactions from method (3) or (4). I used TF - TG interactions with the confidence score A, B or C in the analysis to create a focused, more reliable signalling network.

The VIPER package calculates TF activity using DEG fold change values from the host omics dataset and regulatory interactions from DoRoThEA. It ranks the TFs based on their activity on DEG regulation, and gives only the top 50, as a default setting, to the network diffusion algorithm. This means that CARNIVAL builds up causal networks that highlight the altered signalling between conditions.

CARNIVAL uses *cplex*, software to solve integer linear programming (ILP) problems, for causal reasoning to integrate information from TF and signalling pathway activity scoring. *Cplex* is free for students and academic workers however for commercial purposes researchers have to pay for the program and therefore can not be included in the public workflow. There are alternatives of *cplex*, like Gurobi (<https://www.gurobi.com/>) and CBC-COIN (<https://github.com/coin-or/Cbc>) solvers or IpSolve (<http://lpsolve.sourceforge.net/5.5/>) as network optimisers. Gurobi requires a licence for using it in industrial research, CBC-COIN is freely available for everyone, but its performance is much lower compared to the previous two algorithms. IpSolve is an R package and could replace *cplex* but only for small networks due to the time-consuming analysis.

A limitation for CARNIVAL is that in most cases describing the shortest path does not give a realistic insight into the signalling, also, it relies on the prior annotated pathways. However, it provides a more compact network to analyse the transcriptional footprint of samples and highlights the key differences between conditions (e.g. healthy vs. diseased) by analysing DEGs.

TieDIE (Tied Diffusion through Interacting Events) describes a sub-network from general signalling networks focusing on the signal transduction between the perturbation point and the expressed genes (not necessarily DEGs)²⁵⁰. In general, diffusion-based network propagation algorithms do not take into account the causal parameters, such as the effect of the interactions. TieDIE solved this particular issue by focusing the diffusion process on causally coherent parts of the network. Instead of using *cplex*, TieDIE computed a diffusion kernel module generated by *scipy* (Python package)⁴⁰⁰ to explore the flow of the signal in the general network. The tool is under the GNU General Public License (GPL) v3.0 and is available for everyone both in academia and industry.

The input files for the tool are similar to CARNIVAL: a list of perturbation points and (differentially) expressed genes and a signalling network that connects the start and end points. I used the core (literature curated) interactions from OmniPath as a signalling network. These PPIs are directed but the effect is not always known. Because TieDie requires this information,

I set up the 'unknown' effect to 'stimulatory' interaction. The reason for that is the 'inhibitory' effect in TieDie describes perturbation/mutation in the connection between proteins because the algorithm itself was developed using cancer cell lines ²⁵⁰. Therefore it is better to consider the unknown effect as a positive connection rather than losing them from the networks.

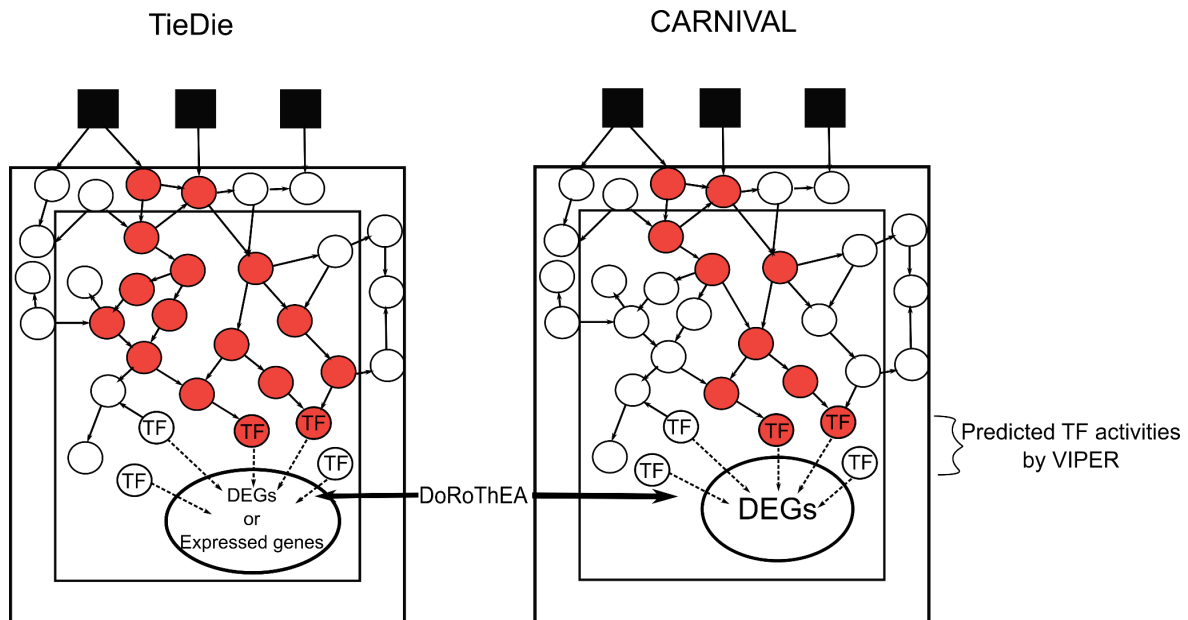


Figure 3.2: Brief comparison of TieDie and CARNIVAL network propagation algorithms. Black boxes highlight the microbial protein, and circles reveal the human proteins. Coloured circles are part of the downstream signalling subnetwork that mediates the effect of the upstream HMIs. While TieDie infers a larger subnetwork, including all possible paths between the bacteria-affected human proteins and TFs regulating the expression of genes, CARNIVAL focuses on the shortest path between the membrane-based protein and those TFs which drive differences in gene expression comparing the conditions.

3.2.4 Gene enrichment and overrepresentation analysis

Both CARNIVAL and TieDIE reconstruct signalling networks which give insight into the signal transduction in specific tissues or cells also highlighting differences between healthy and diseased conditions. Due to the complexity of the networks and the cross-talk of signalling pathways, the affected biological functions are not detectable by modelling the downstream intracellular signalling itself. I used GSEA (GORilla) and GSOA (PIANO) tools to interpret the output of the network modelling in an unbiased manner.

GOrilla is a web-based GO annotation analysis tool (<http://cbl-gorilla.cs.technion.ac.il>) highlighting the overrepresented GO terms among the genes or proteins provided by the user⁴⁰¹. This approach reveals what processes are affected in the network and also there is a possibility to provide background gene sets (complete genome or a customised larger set of genes) to reveal the enriched functions compared to the functions in the background set. An important note is that the p-value does not include the multiple hypothesis correction on the number of tested GO terms.

PIANO (Platform for Integrated Analysis of Omics data) includes 11 gene set analysis (GSA) calculation methods to perform gene set enrichment analysis and visualise results interactively. The extended statistical approaches give flexibility to the algorithm, therefore it accepts expression values, p-values, t-values and even fold change values as input for the GSEA. PIANO ranks the input values based on gene set statistics and gives back a list with enriched signalling pathways or other biological processes. The tool is available as an R package⁴⁰².

3.3 Results

Because this chapter describes the development of a novel methodology, the result is the pipeline itself. In this section, I would like to provide a user guide describing the inputs and outputs for MicrobioLink2 and also highlighting the automated and manual steps in the workflow.

I implemented two versions of the workflow depending on its usage: (1) for people in academia the whole pipeline is available, (2) for commercialising purposes a licence is required from EMBL, which institute provides the ELM database; besides a limited number of interactions and annotations can be used from OmniPath (35 686 interactions instead of the 40 014 PPIs that are in the academic version) and the quality control step left out due to the lack of industrial licence from IUPred.

3.3.1 Host-microbe interactome

The *in silico* host-microbe interaction prediction is a semi-automated workflow written in Python. The input files are separated into user-provided and hard-coded files. The user has to upload the following inputs: (1) a list of bacterial proteins or UniProt Proteome IDs, (2) a list of host genes/proteins of interest or a processed host transcriptomic/proteomic dataset where the required format is either a list of genes/proteins or a matrix describing the expression/abundance of the molecules. Compared to MicrobioLink which accepts bulk transcriptomics, the new version is able to handle single-cell transcriptomics data as well.

The DMI table and the SLiM patterns derive from ELM and 3did, also resources for the IDR prediction by IUPred are provided and implemented in the script because these standard files are part of the DMI-based PPI prediction algorithm and the quality control afterwards.

MicrobioLink includes 258 DMIs from ELM (data from 2013). However, the latest version of the interactions describes 354 interactions, I updated it in MicrobioLink2. The new table not only expanded the list of interactions but also removed 13 DMIs due to motif ID changes, therefore the new pipeline describes connections with the most updated Pfam and ELM motif IDs. Also, the new pipeline includes extra 985 interactions derived from the 3did database. Interestingly, there was no overlap between ELM and 3did in terms of the motif regular expressions resulting in a lack of common DMIs.

The bottleneck of this analysis is the number of described domains that have known target motifs. MicrobioLink includes 114 domains while MicrobioLink2 contains 277 domains [Figure 3.3].

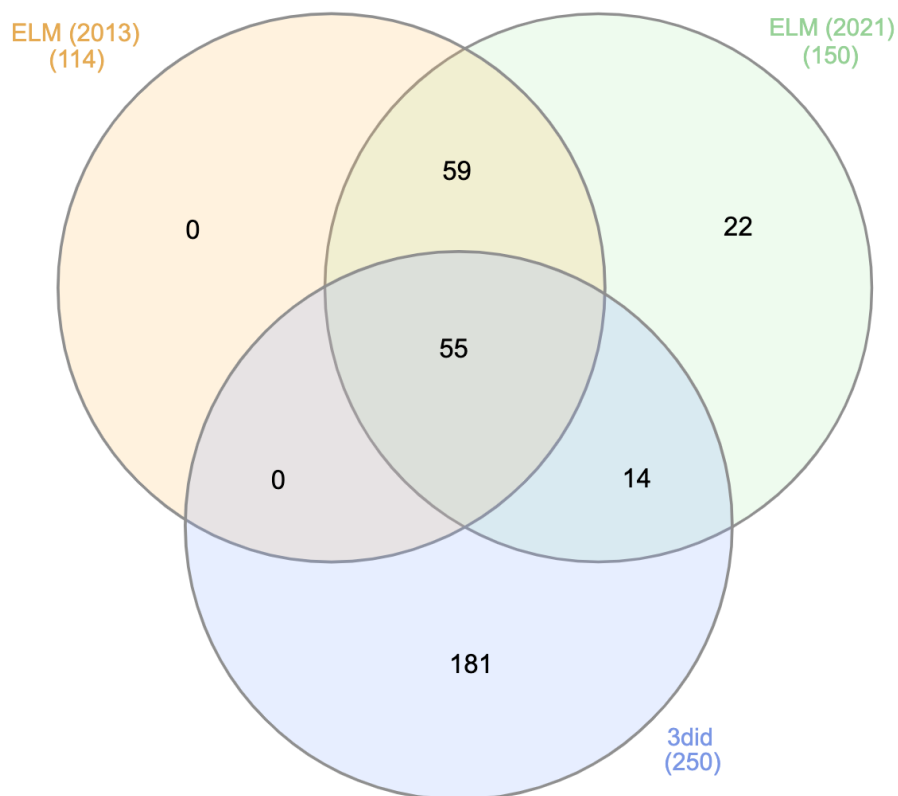


Figure 3.3: Comparison of motif targeting domains between ELM (old version - ELM (2013) - in MicrobioLink and the new one - ELM (2021) - in MicrobioLink2) and 3did.

Firstly, the script downloads the domain structure of bacterial proteins and the .FASTA sequence file of human membrane proteins. If there is an available metaproteomic dataset that describes protein abundance in the microbial community it can be used directly for the pipeline. However, in the lack of metadata, the algorithm accepts a list of UniProt Proteome IDs and downloads the whole proteome from the database. In this case, the input file consists of two columns, one with the bacteria strain and the other one with the proteome ID. Also, there is a possibility to analyse another condition (e.g. diseased microbial community) in parallel. The script accepts an optional parameter, the column that describes the proteome ID of the bacteria strains in the other condition. There is an opportunity to assign a location for bacterial proteins or even filter the dataset based on subcellular location to avoid the large host-microbe interactome. However, this step is not part of the core workflow, it should be run separately.

Regarding the host side, there is a need for processed transcriptomic or proteomic datasets. The required format is an average gene count or protein abundance matrix that includes the description of genes (by gene symbols) or proteins (by UniProt IDs) in rows, the cell or tissue type with the condition in columns and the average expression or abundance value in the cells. The algorithm downloads only the sequences of the potential membrane-based proteins filtered by annotations in OmniPath.

The inputs are ready for the interaction prediction (and the quality control in the academic version). IUPred is not available for commercialising purposes, therefore it is available in a separate script. The output of the prediction is a table describing details about host-microbe interactions including the interacting proteins (UniProt IDs), the interacting bacterial domain (Pfam ID) and human motif (exact position in the protein sequence and its length) pairs. This interaction table can be used as an input for Cytoscape to visualise the interactome manually.

3.3.2 Network diffusion modelling

The second part of the analysis consists of another semi-automated workflow implemented in R and Python that constructs the downstream intracellular network. Although the heat propagation algorithms use different methods to build up downstream signalling network, their input files are similar: (1) list of perturbation points (bacteria targeted human proteins from the prediction), (2) endpoint for the signal spread (TFs), (3) contextualised regulatory interactions from DoRoThEA, (4) (differentially) expressed genes, (5) contextualised PPI network with directed edges. By contextualised data, I mean interactions between expressed genes/proteins from the input host transcriptomics/proteomics dataset.

Using CARNIVAL the most important output is the final network visualised by GraphViz⁴⁰³, besides it creates files for each network model created by cplex. In contrast, TieDie makes a report.txt file about network statistics and a summary of the analysis. Also, it gives information about the interactions and about the heat of each node which parameter quantitatively describes the strength of the signal spread through the downstream proteins. For further visualisation, the interaction and node annotation files can be imported into Cytoscape.

3.3.3 Functional analysis

For the gene set enrichment analysis (GSEA), users should provide two gene sets to analyse the statistical significance of the overrepresented pathways. Using CARNIVAL automatically results in GSEA by PIANO. However, for the TieDie output network, I inserted a manual functional analysis by GOrilla tool to extend the scope of GSEA tools. Here, the observed gene set includes the nodes that appear in the output networks from TieDie, and the background gene set contains all the genes in the whole contextualised PPI network.

3.4 Discussion

I have developed an *in silico* host-microbe interaction prediction method to analyse the direct effect of microbes on host cells and tissues by inferring a host-microbe PPI network and exploring the indirect effect of bacteria on the downstream signalling pathways using transcriptomic data from the host.

The workflow is based on our previously published MicrobioLink pipeline³⁸⁰ however I modified several parts of the tool and established MicrobioLink2. I updated databases involved in the previous version and also contributed to the development of OmniPath database which increased the reliability of the output data [details in Chapter 2]. There are three novelties of the work presented in this chapter. Firstly, the workflow is able to handle UniProt Proteome IDs and gain all proteins automatically instead of requiring a list of microbial proteins. Secondly, I included CARNIVAL as a network propagation method - complementing TieDie - to have an insight into the transcriptional footprint of the data as an effect of microbes by inferring the shortest path between the bacteria-affected receptors and the differentially expressed genes in tissues or cells. Finally, not only bulk but single-cell data can be also used as input host transcriptomic profile for the workflow.

This study is a gap-filling approach among plenty of host-microbe interaction prediction resources. While most of them^{404,405} focus on the functional relevance of microbes on host tissues or focus on one specific microbe, here, I propose a computational tool to explore HMIs (where the domain structure is available) on molecular level including the advantage of following the signal from the triggered receptors down to the transcriptional changes in the host. Also, including single-cell transcriptomics offers a new perspective for HMI analysis and establishes the cell and condition-specific analysis of microbes on host signalling.

The thesis presents two use case examples of the workflow: Chapter 4 explores the cell and condition specificity of the gut commensal *Bacteroides thetaiotaomicron* in healthy intestine and during ulcerative colitis focusing on immune cells. Chapter 5 presents a use case of the pipeline by analysing the effect of the disrupted microbiome on epithelial cells in the oral cavity during periodontitis.

The pipeline includes several limitations. Predicting bacterial-human PPIs is challenging due to the lack of knowledge regarding the motifs bound by bacterial domains. Using the ELM and 3did databases limits the results to those domains which occur in Eukaryotes and misses the bacteria-specific tertiary structures. This gap could be addressed by using protein structures from the AlphaFold2 (AF2) resource. AF2 is a software that predicts the 3D structure of a protein based on its amino acid sequence using deep learning and sophisticated algorithms³⁹³. By utilising the tool that deduces the structure of proteins from bacteria, we can expand the list of potential domains present in these proteins, thereby improving the precision of our predictions regarding the interactions between the host and microbe. Also, the effect (activator or inhibitor) of bacterial proteins is unknown on host proteins. These restrictions can be solved by integrating manually curated information about bacterial domain-binding motif connections into the HMI prediction. Another issue is that I assume every transcript, expressed in transcriptomics, translates to functional protein which is not true. Post-transcriptional and post-translational modifications affect the mature RNA structure and the translated protein activity. Analysing proteomics and transcriptomics from the same samples could improve the model. Finally, I would like to include other pathway finding methods in the downstream analysis like the Prize-collecting Steiner Forest (PCSF) algorithm to include genes/proteins missed by the detection platform [details in Chapter 1.7.1].

In the future, I plan to upgrade the current methodology by integrating bacterial metabolite-protein interactions into the model and analysing the effect of small molecules on host signalling. Also, I would like to connect the bacterial metabolite-affected human receptors to the human metabolic network (e.g Recon3D⁴⁰⁶) to map which host metabolic pathways are affected and how the host metabolite secretion is differing under a disordered condition. Discovering the effect of the altered microbial metabolites in dysbiotic communities on host processes could provide a complementary, host systems biology interpretation to the existing community modelling efforts.

Chapter 4: Analysing the cell-type specific effect of bacterial outer membrane vesicles on the immune system

4.1 Introduction

The gut microbiome has been linked to a variety of health conditions, research suggests that an imbalance in the gut microbiome, called dysbiosis, may contribute to the development of certain diseases such as IBD ⁴⁰⁷. However, it is challenging to describe the IBD-associated microbiome because the composition differs person by person. In IBD patients the intestinal tract is colonised by a reduced number of *Firmicutes* species and there is an increase in *Bacteroidetes* ⁴⁰⁸, *Proteobacteria* and *Actinobacteria* species ⁴⁰⁹. Studies show a lower level and taxon diversity of *Bacteroides* in UC patients compared to a control group ^{407,410}.

The *Bacteroides* taxon is one of the most common groups of bacteria in the intestine (with 25-50% average abundance) ^{411,412}. Interestingly, some of these Gram-negative anaerobic microbes are able to act as commensals in the intestine, however others, outside the gut, can be harmful pathogens (e.g. *Bacteroides fragilis*, *Bacteroides thetaiotaomicron*) ⁴¹³. This commensal - pathogen conversion is due to a typical large genome of the members of the *Bacteroides* taxon. These species can easily turn on and turn off a couple of genes (mainly metabolic pathway-related ones) to adapt to the actual environments ⁴¹⁴. Bacteria in the *Bacteroides* taxon produce extracellular vesicles that are known to play key roles in intercellular communication ^{415,416}. In particular, bacterial extracellular vesicles (BEVs) are 20–500 nm-sized and have spherically bilayered structures. BEVs are released by intestinal bacteria into the gut lumen to mediate cross-kingdom interactions with host cells resulting in modulation of host signalling pathways ⁴¹⁷. BEVs produced by Gram-negative bacteria are mainly composed of phospholipids, lipopolysaccharides (LPS), peptidoglycan, outer membrane proteins and periplasmic content and also include some inner membrane and cytoplasmic fractions ⁴¹⁵.

The LPS structure of *Bacteroides* shows unique parts compared to other bacteria (e.g. *Escherichia coli*)⁴¹⁸ that causes a taxon-specific immune response in the host. A modified structure of LPS, called lipooligosaccharide (LOS), has penta-acylated and monophosphorylated lipid A that does not promote pro-inflammatory responses in immune cells^{419,420}.

Bacteroides thetaiotaomicron (Bt) is a prominent Gram-negative anaerobe in the *Bacteroides* taxon residing in the caecum and colon of most or all animals. Bt BEVs can access and transmigrate across boundary epithelial cells using different routes⁴²¹, interact and modulate the mucosal immune system and disseminate more widely *via* the bloodstream^{422–426}.

Bt is one of the potential next-generation probiotics⁴²⁷. To restore dysbiosis in the gut, researchers analyse the effect of probiotics as a potential treatment for diseases^{428–432}. Probiotics are living bacteria promoting health benefits that are able to repair the disrupted mucosal layer and restore the bacterial equilibrium state⁴³². As an essential gut symbiont, Bt has a well-studied anti-inflammatory effect in the gut^{433–435} affecting both epithelial cells (causing increased goblet cell differentiation⁴³⁴) and immune cells (e.g DCs, T cells⁴²⁵). Bt is able to selectively antagonise transcription factor NF-kappaB in the host cells, therefore decreasing the secretion of IL-8, TNF- α , and IL-1 β and attenuating inflammation⁴³³. Studies on DSS-induced colitis in mice showed the relevance of Bt in IBD. Bt strongly induces the maturation of the colon immune system including Treg pathway activation reducing Th1, Th2 and Th17 cytokines and increasing the expression of *IL-10*, *TGF β* and *PDCD1* genes⁴³⁶.

In this chapter, I present how I used the established host-microbe interaction workflow [detailed in Chapter 3] to explore the role of BEVs derived from the gut commensal Bt on immune cells in healthy and UC colon. The following case study was published in *Journal of Extracellular Vesicles*⁴³⁷.

4.2 Methods

The project consists of wet lab experiments - carried out by Simon Carding's group at the Quadram Institute - and computational data analysis by Matthew Madgwick and myself [Figure 4.1]. The isolation, purification and proteomic analysis of Bt BEVs were performed by Regis Stentz (QIB). Imaging was done by Catherine Booth (QIB). Sonia Fonseca (QIB) was responsible for the experimental verification of *in silico* findings. Processing of raw single-cell data was carried out by Matthew Madgwick. All the other computational analyses and interpretations were executed by myself.

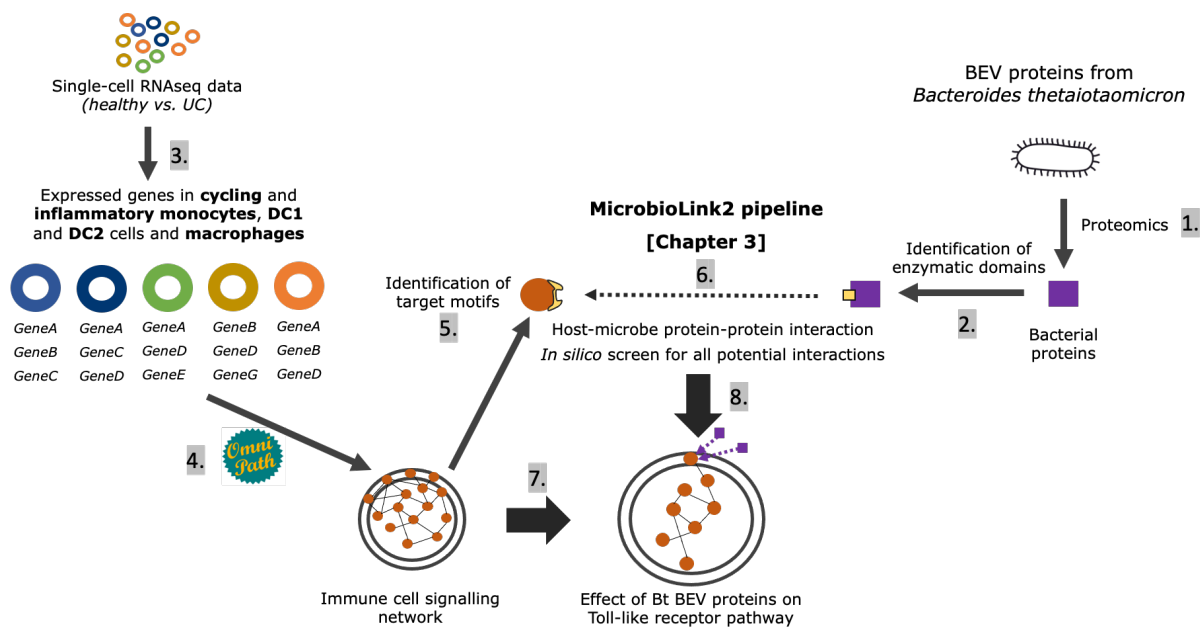


Figure 4.1: Computational workflow to analyse cell-type specific effects of BEVs. Numbers indicate the sequence of the main steps: 1, Extraction of BEV proteins from the proteomic dataset 2, Identification of bacterial domains using the Pfam database 3, Processing the raw single-cell transcriptomics from human colon 4, Creating cell type-specific networks using PPIs from OmniPath¹⁰⁷ 5, Identification of SLiMs on human proteins using the ELM database 6, Predicting protein-protein interactions (PPIs) between BEV and host proteins in each cell-type separately by MicrobioLink2 7, Reconstruction of Toll-like receptor pathway using Reactome database²⁹⁷ 8, Combining cell-specific signalling with BEV targeted human proteins.

4.2.1 Experimental analysis of BEV proteins

Regis Stentz and Sonia Fonseca were working on the experiments that resulted in a list of BEV proteins that I used as an input for the MicrobioLink2 pipeline. The experimental protocol consisted of the following steps: isolations and characterisation of Bt BEVs, proteomic analysis of the vesicles and checking of the structure of BEVs by transmission electron microscopy. Details about the experiments can be found in the original article ⁴³⁷.

4.2.2 Single-cell transcriptomic datasets analysis

The same public study ³¹² has been used for the project that has been described in Chapter 2. While in that study epithelial, immune and stromal cells were analysed, here, I filtered the data for only the following immune cells: cycling monocytes, inflammatory monocytes, macrophages, DC1 (healthy mucosa-related subset) and DC2 (inflammation-related subset). For these cell populations, I used those that were from healthy or non-inflamed UC conditions. Further information about the single-cell data processing is described in Chapter 2.

4.2.3 Analysis of bulk transcriptomic data

I processed two public bulk RNAseq datasets to model the effect of Bt BEVs on the THP-1 monocytes - the cell line that was used for the experimental verification of the *in silico* results. I collected pre-processed datasets from Gene Expression Omnibus (GEO) (GSE132408 and GSE157052) that described gene expression in healthy condition. Due to the different protocols of the two studies, I normalised the datasets using the DESeq2 R package. Also, GSE132408 used gene symbols while GSE157052 described geneIDs, hence I unified them to gene symbols using Uniprot ID conversion tool ²⁹², and kept only genes which were detected in both experiments. I filtered the expressed genes with the Z-normalisation method (cutoff > -3) [details in Chapter 2].

4.2.4 Cell-type specific Bt BEV - human interactome

I explored the effect of BEV proteins on different cell types based on host-microbe PPI networks using MicrobioLink2 introduced in Chapter 3. The assumption was that a BEV protein can bind to a human protein if a BEV protein domain targets an amino acid motif on the host protein based on the ELM database ²⁷⁹. First, I downloaded the sequence of BEV proteins detected in the proteomics analysis from the Carding lab and of the human proteins, which were translated from genes in the single-cell transcriptomics using the Uniprot database ⁴³⁸. I connected the two sets of proteins with the MicrobioLink2 pipeline [details in Chapter 3].

4.2.5 Functional analysis of Bt BEV protein targets

I performed gene set overrepresentation analysis by GOrilla ⁴⁰¹ to highlight the main functions affected by Bt BEVs. The observed input gene set consisted of the Bt BEV targeted human proteins while the background set described all expressed genes in the examined cell type under healthy or UC condition. An annotation was significantly overrepresented among the Bt targets if the p-value was less than 10^{-3} and the FDR q-value calculated by Benjamini and Hochberg method was less than 0.05. The output of the functional analysis describes a list of processes affected in each cell type by Bt. Due to the complexity and difficulties in data interpretation, I used REVIGO to reduce the dimensionality of the annotations and identify significant differences among functions ⁴³⁹. simRel scores were applied to measure the GO semantic similarity. To visualise the functional overlap among cell types and conditions, I used the InteractiVenn web-based tool ⁴⁴⁰.

4.2.6 Cell-type specific signalling pathway analysis

The TLR signalling pathway is complex and encompasses nine receptors and numerous downstream components. I obtained the pathway members from Reactome²⁹⁷. I combined the host-microbe PPI prediction results with log₂ expression values from scRNAseq and bulk RNAseq datasets (monocytes, dendritic cells, macrophages, THP-1 cells). This allowed me to understand cell type-specific gene expression patterns and their impact on interspecies interactions. I calculated the difference in gene expression between two states (expression in UC condition – expression in healthy condition), and visualized the results using heatmaps created in Python, in order to compare the patterns under varying conditions.

4.2.7 *In vitro* validation of *in silico* findings

Sonia Fonseca performed the experimental validation; the exact protocol is described in the original article⁴³⁷. Briefly, she handled THP-1 monocytes with NF-κB reporter constructions to highlight the activation of the TLR pathway under diverse conditions. The cells were exposed to Bt BEVs, *E. coli* LPS and phosphate-buffered saline (PBS) as a control to explore the effect of LPS- and LOS-coated bacterial vesicles on the signalling. The TLR pathway was inhibited by CLI-095 (a TLR4 inhibitor) and a TIRAP inhibitor which enabled us to study the TLR4 and TIRAP-mediated activation of the TLR pathway.

4.3 Results

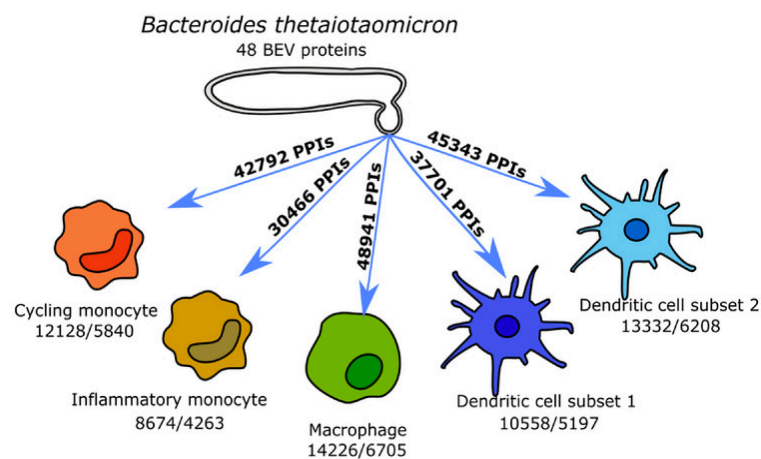
4.3.1 Reconstructing a BEV - human interactome

We combined experimental approaches with *in silico* analysis to reveal the effect of Bt BEV proteins on signalling pathways in human cells. Following the isolation and purification of BEVs, a proteomic analysis unveiled 2068 proteins in the bacterial vesicles. I analysed single-cell RNAseq data highlighting the expressed genes in the selected five cell types: cycling monocyte, inflammatory monocyte, DC1, DC2 and macrophage. Figure 4.2 shows the predicted number of PPIs between the BEV and human proteins. Although RNAseq data describes genes, I inferred the protein-protein interaction (PPI) network by assuming that all expressed genes were translated into functional proteins.

I found 48 BEV proteins which were able to bind target sequences on human proteins. Most of them (43 out of 48) were hubs in the network, each contacting hundreds of human proteins separately due to their enzymatic nature. These 43 proteins are hydrolases, proteases, and other catabolic enzymes without a specific cleavage site. The rest of the five BEV proteins interact with a human polymerase.

Around half of the expressed genes in human cells were potentially able to connect to bacterial proteins in every cell type [Figure 4.2]. There was no difference among the interacting bacterial proteins, the same 48 proteins were included in the PPI networks in both conditions. However, I found human proteins which interacted with the BEVs only in one of the conditions (healthy or UC) or in a few cell types. This outstandingly high ratio of host targets shows the need for a specific focus on the data instead of analysing the whole interactome.

a, Healthy condition



b, Ulcerative colitis

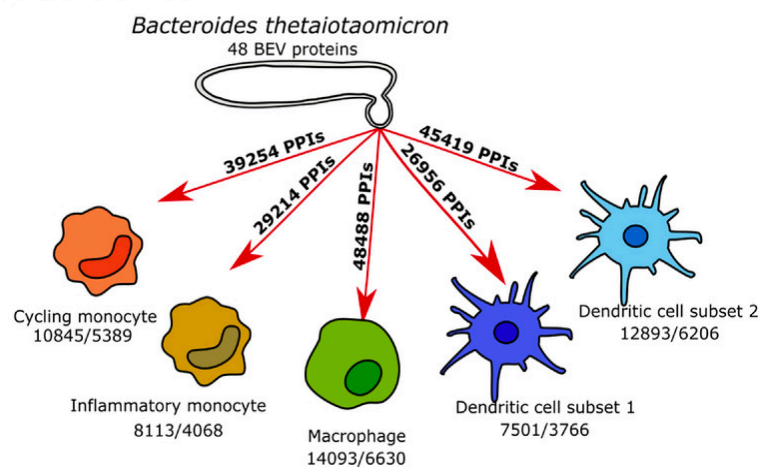


Figure 4.2: Interactions of 48 BEV proteins with various human cells. Monocytes, macrophages and dendritic cells in healthy (a) and UC (b) conditions interacting with BEV proteins. The number of expressed genes/number of interacting proteins is highlighted for each cell type. The figure was drawn by myself.

4.3.2 Functions of the human target proteins

The described BEV - immune cell interactome highlights which proteins are directly targeted by the Bt BEVs, however it does not give information about the affected processes themselves. I explored the function of the BEV targets using the GOrilla enrichment analysis tool and then compared the differences among the selected immune cells. Around 60% of annotations in all the cell types were overlapping between cells related to basic cellular processes, such as metabolic pathways and chromatin organisation. Although in a smaller ratio, I found bacteria-targeted processes appearing in one of the cell types. In inflammatory monocytes, Bt BEV proteins affect apoptosis and myeloid cell differentiation, while in cycling monocytes, proliferation-related functions were enriched in both conditions. Interestingly, in the healthy condition, among the BEV-affected processes in DC1 cells, somatic diversification of immune receptors and B cell apoptosis were uniquely over-represented. In contrast, vesicle fusion, negative regulation of apoptotic signalling pathways, and the intracellular steroid hormone receptor signalling pathway were found as uniquely affected functions in UC. Regarding DC2 cells, there were only 11 cell-specific annotations in the healthy state that did not relate to specific functions, whereas in UC, 35 unique annotations affected the cell cycle. Finally, in macrophages, epidermal growth factor (EGF) receptor and regulation of TGF β receptors were involved in the healthy state and RAS protein signal transduction in UC.

4.3.3 Effect of Bt BEVs on the Toll-like receptor pathway of immune cells

As introduced in Chapter 1, the TLR pathway is important in regulating inflammatory processes. In addition, there are publications that support how Bt affects the TLR pathway⁴²⁵. I explored the impact of the BEV proteins on the TLR pathway in detail to reveal potential condition-specific key signalling components. Cell types have different gene expression profiles, therefore I have analysed the TLR pathway in diverse cells under healthy and UC conditions.

In general, results show interactions mainly between Bt BEVs and downstream TLR pathway members, receptors are less likely to be a target for the bacterial vesicles. The heatmap highlights another common feature, the transcription factors did not show a different expression pattern between conditions, besides, all of them are potential BEV targets [Figure 4.3]. I made pairwise comparisons analysing different subpopulations of cell types, such as dendritic cells (DC1 vs DC2) and monocytes (cycling vs inflammatory) to identify cell- and condition-specific parts of the pathway.

Dendritic cells show exciting examples for condition- but also cell type-specificity. DC1, a DC subset dominating in healthy condition, includes less active signalling during UC because most receptors (*TLR1*, *TLR2*, *TLR3*, *TLR7*) and many downstream pathway components are expressed only in healthy state. Among the condition-specific proteins, 16 BEV protein targets suggest a diverse effect of Bt BEVs on the TLR pathway in DC1 cells in healthy condition compared to UC. In contrast, in DC2s (inflammation-related DC population) almost the whole TLR signalling is equally active between states, including the TLR4 receptor, which was not found in the DC1 cells [Figure 4.3].

In monocytes, the TLR signalling shows differences between conditions rather than subpopulations. Unlike the DCs, here, the inflammatory subtype includes UC-specific (12 genes) and healthy condition-specific (17 genes) gene expression suggesting signalling rewiring. TLR7 and TLR10 trigger signalling in a healthy state, while in non-inflamed UC, TLR4 and TLR5 can be found uniquely. In cycling monocytes, the signalling is balanced between the conditions. 11 genes are expressed condition specifically, although the BEV targeted TLR4 is strongly expressed in UC [Figure 4.3].

Experimental validation was carried out on monocytes derived from THP-1 cell line, therefore I analysed public bulk transcriptomic data to predict BEV - TLR pathway interactions in THP-1 monocytes. Results overlap with the output of the cycling monocyte scRNAseq data analysis, however, here I found more potential BEV-interacting proteins (*PELI2-3*, *IRAK2*, *DNM1*, *RPS6K2*, *MAPK11*) [Figure 4.4].

Macrophages show less condition specificity in terms of TLR pathway member expression. *MAPK10* and *MAPK11* are related to healthy, and *PELI2* is related to UC condition, otherwise, genes are equally expressed between states [Figure 4.3]. Although the number of expressed receptors is the highest in this cell type, only TLR4 is predicted to be targeted by BEV proteins.

4.3.4 Role of Bt BEV proteins in TIRAP-mediated TLR signalling

TLR4 has been identified as a potential target for the Bt BEV proteins. TLR4 usually binds LPS molecules rather than bacterial proteins⁴⁴¹. However, the interactions I found were related to the intracellular TIR domain of TLR4 that could promote the effect of BEV proteins in the cytosol. BT_2239 is a carboxyl-terminal protein expressed by Bt carrying three domains (Pfam domains from UniProt): a PDZ- and two peptidase S41 family domains. The PDZ domain binds to a C-terminal motif (833-839 AAs) on the cytoplasmic TIR domain on TLR4 [Figure 4.4]. Although there is no exact information about the binding sites of the S41 peptidases, these domains recognise tripeptides at the C-terminal end of proteins⁴⁴². I assume that these known and supposed interactions could influence the downstream part of the TLR4 pathway.

Based on the *in silico* prediction, Bt BEV proteins potentially bind to TIRAP. TIRAP is an adaptor protein for TLR2 and TLR4. It is responsible for driving the signal into the Myd88-dependent direction, which causes pro-inflammatory cytokine secretion (e.g. TNF- α , IL-6). Activation of TIRAP results in the induction of MAPK signalling and NF- κ B-mediated gene transcription⁴⁴³. The literature describes an altered MYD88-dependent TLR4 pathway due to interaction with Bt⁴⁴⁴. Also, knock-out of the protein leads to a substantial decrease in TNF- α secretion. Surprisingly, the analysis revealed four potential domains expressed by 19 BEV proteins which can bind to TIRAP at different target motifs all over the protein sequence [Figure 4.4].

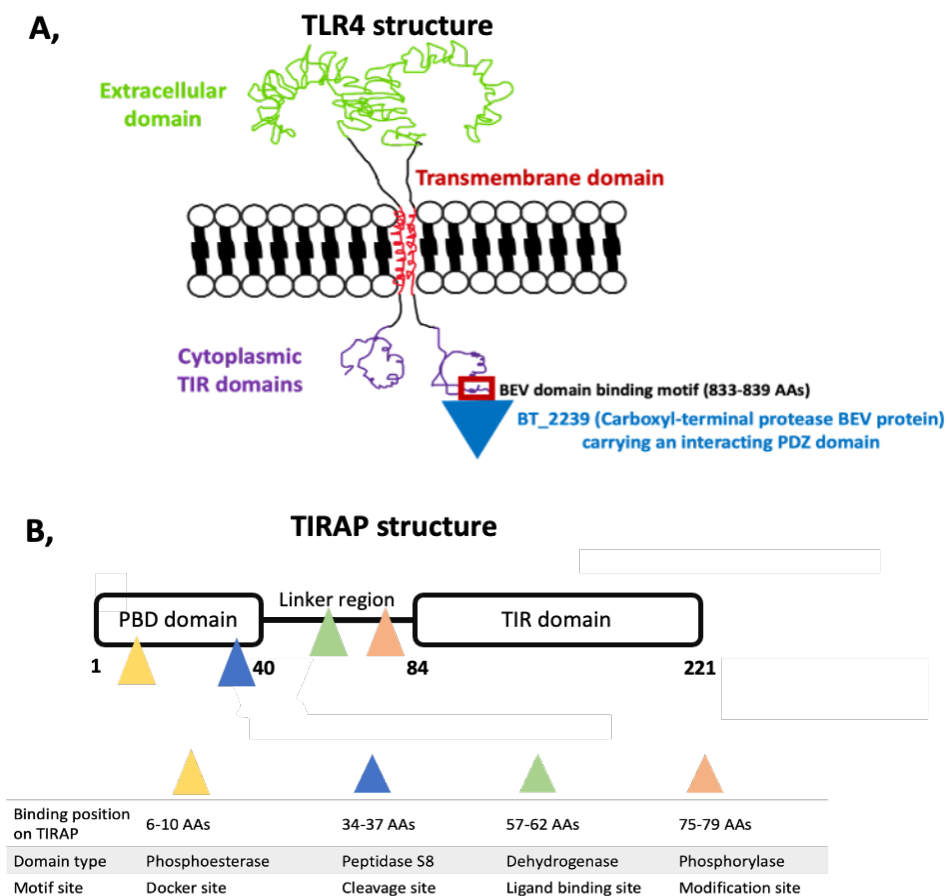


Figure 4.4: Structural details about A, TLR4 - Bt BEV protein and B, TIRAP - Bt BEV protein interactions. The *in silico* host-microbe PPI prediction revealed an interaction between the PDZ domain of a bacterial carboxyl-terminal protease and a motif on the intracellular TIR domain of TLR4. Interestingly, the TLR2/4 adaptor TIRAP is bound by four bacterial domains based on the prediction characterised by diverse functions. TIR – Toll/interleukin-1 receptor/resistance protein domain; PBD - Phosphatidyl-inositol binding domain. The figure was drawn by myself.

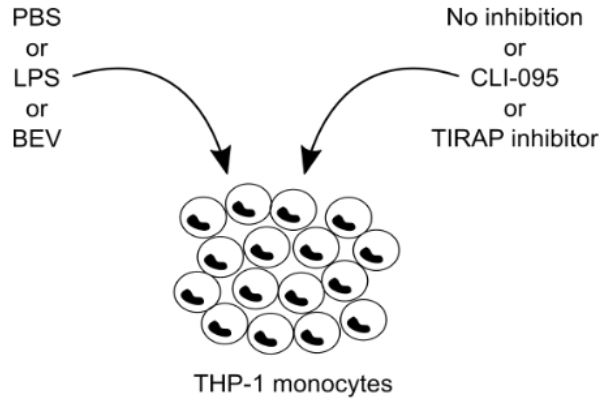
4.3.5 Inhibition of TLR4 signalling pathway diminishes monocyte activation by Bt BEVs

I found TLR4 to be the only receptor predicted to be targeted by BEV proteins in monocytes, macrophages and DCs. Coats *et al* validated the interaction between Bt and TLR4 experimentally, they found the lipidA component of LOS on Bt triggers different TLR-response compared to LPS on *E.coli*'s surface⁴⁴⁵. Sonia Fonseca from the Carding lab examined the effect of BEVs on the receptor measured by NF- κ B activation in BEV-THP-1 monocyte co-cultures in the presence or absence of CLI-095.

Results show that increasing BEV concentration (3×10^7 - 3×10^9 /ml) enhances NF- κ B activation compared to control with PBS in monocytes. Adding CLI-095 inhibitor results in a decrease of the transcription factor activity with the highest level of inhibition (~37%) seen at the lower dose of BEVs (3×10^7). In contrast, applying the inhibitor during the lack of BEVs has no significant inhibition ($P > 0.05$) of NF- κ B activation [Figure 4.5]. All in all, the incomplete inhibition of NF- κ B by the TLR4 inhibitor offers a TLR4-independent effect of BEV proteins on NF- κ B activation.

Signalling networks identified BEV-interacting downstream TLR pathway components that support the experimental result from TLR4 inhibition. Therefore Sonia repeated the experiment and used a TIRAP inhibitor instead this time. Here, she found a significant ($P < 0.01$) reduction of NF- κ B activation (37.5%) at 3×10^8 BEVs/ml concentration but no significant effect using higher dose of bacterial vesicles [Figure 4.5].

A,



	PBS	LPS	BEV
No inhibitor	~ no NF-κB activation	significantly increased NF-κB activation	significantly increased NF-κB activation
CLI-095	~ no NF-κB activation	significantly decreased NF-κB activation	no significant changes
TIRAP inhibitor	~ no NF-κB activation	significantly decreased NF-κB activation	significantly decreased NF-κB activation

B,

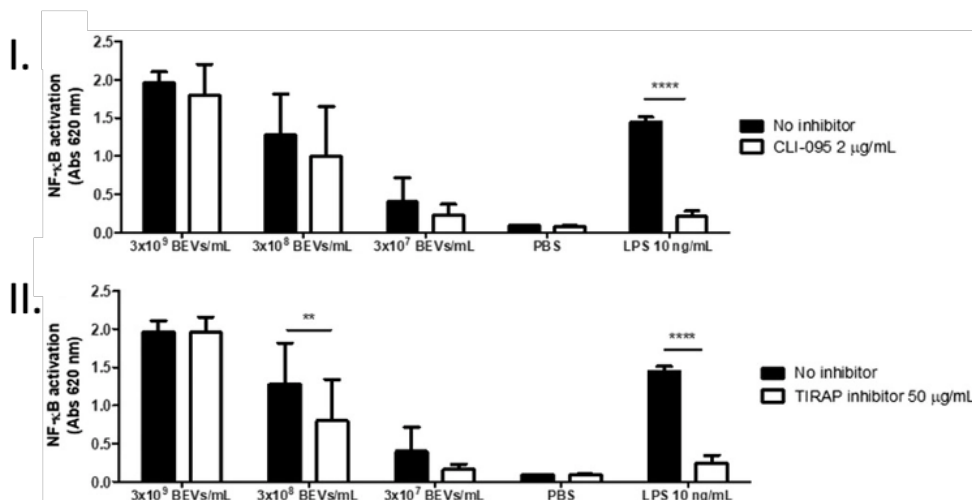


Figure 4.5: Inhibition of TLR4 and TIRAP signalling pathways abrogates THP1-Blue cells activation by Bt BEVs. A, Schematic view of the experiment; B, Experimental validation of I. TLR4 - BEV and II. TIRAP - BEV interactions. NF-κB activation was assessed using different doses of BEVs in 5×10^5 THP1-Blue cells/ml in the presence or absence of the TLR4 inhibitor CLI-095 (I.) or TIRAP inhibitor (II.) and by measuring absorbance at 620 nm after incubation with the colourimetric assay reagent Quanti-Blue. LPS from *E. coli* was used as a positive control and PBS as a negative control. Data are presented as mean \pm SD ($n = 9$). Significant differences were determined by using two-way ANOVA followed by Bonferroni's multiple comparison post hoc test. ** ($P < 0.01$), **** ($P < 0.0001$). Part A was drawn by myself while part B was created by Sonia Fonseca.

4.4 Discussion

In Chapter 4, I have shown a use case example for the host-microbe interaction pipeline described in Chapter 3 that reveals the cell type-specific effect of a gut commensal bacteria - *Bacteroides thetaiotaomicron* - upon ulcerative colitis. Due to the recent appearance of single-cell omics data, there is a lack of knowledge on cell type-specific effects of microbes, especially in a diseased condition.

Here, I was focusing on proteins in bacterial vesicles due to their significant impact on cross-species interactions ⁴⁴⁶. As Bt is a potential therapeutic agent in IBD ⁴⁴⁷, it is important to understand how BEV proteins are able to interact with and alter the signalling in the immune cells thus controlling inflammatory processes.

The diverse gene expression profile of immune cells enabled BEV proteins to establish immune cell-specific interactions. Hence, I selected cycling monocytes, inflammatory monocytes, DC1s, DC2s, and macrophages in both the healthy and the non-inflamed UC colon to reveal the differences and the effect of host-microbe interactions. The *in silico* prediction revealed large interspecies interactomes in all five cases. Although the participants differed between cell types, I did not find differences among the targeting BEV proteins. The majority of the bacterial proteins belong to the diverse groups of catabolic enzymes and establish non-specific interactions.

Functional analysis of BEV-targeted human proteins revealed cell type-specific differences, such as overrepresented cell division in cycling monocytes in the healthy condition. These monocytes circulate in the blood and then migrate and differentiate into macrophages in various tissues. For a homeostatic state, it is necessary to maintain a pool of macrophages by proliferating cycling monocytes ⁴⁴⁸. In contrast, during UC, DNA repair is strongly affected in the same cell type. Here, the literature supports the fact that a higher level of oxidative DNA damage characterises the mucosal layer during a severe UC ^{449–453}. Therefore this finding promotes that Bt BEV proteins can potentially affect DNA repair thus contributing to the treatment of the disease.

Inflammatory monocytes dominate during an inflamed condition, however, they are represented in the healthy colon as well but in a reduced amount. Here I found that BEV proteins are likely to connect to proteins that are involved in apoptotic processes regardless of the condition.

Interestingly, in DC1 cells the somatic diversification of the immune cell receptors is affected by the bacterial proteins. This process increases the specificity of these proteins, such as TLRs, recognising a wider range of molecular patterns ⁴⁵⁴. In UC, however, proteins responsible for vesicular transport are targeted dominantly by Bt BEV proteins. DCs secrete many kinds of cytokines, therefore altered vesicular transport can lead potentially to inflammation and modulation of the immune system ⁴⁵⁵.

EGF signalling plays a key role in macrophage activation, which cells are essential to control inflammation. In healthy condition, BT BEV proteins target some members of the pathway, therefore, leading to a potential change in the output in the cells ⁴⁵⁶. Analysing bacterial targets in diseased samples revealed an enriching effect on Ras-mediated signalling. Although I have not found relevant information about the role of Ras in gut macrophages, a study highlighted that Ras forces macrophages to pro-inflammatory cytokine production, therefore, contributing to breast cancer ⁴⁵⁷.

The TLR pathway plays an important role in bacteria recognition, including Bt ^{458–460}, however the exact molecular background and cell type specificity are less studied. The current analysis contextualised the pathway in five cell types, and offered potential key signalling points that differ between cells or conditions. I could identify only the TLR4 receptor on the cell surface interacting with Bt proteins. Other target proteins in the TLR pathway are part of the downstream signalling network in the cytosol that assumes the intracellular presence of BEVs. The intracellular uptake of BEVs has been supported by the literature as well ⁴²¹. The TLR4 receptor shows cell type and condition-specific expression based on the analysis. TLR4 is not expressed in DC1 cells and healthy inflammatory monocytes, but shows a unique expression in inflammatory monocytes in samples from UC patients.

This finding encouraged me to look into further details in TLR4 - BEV protein interactions and to analyse the upstream part of the TLR4 pathway. Based on the *in silico* prediction, a bacterial carboxyl-terminal protease (BT_2239) is predicted to bind the receptor. In more detail, a PDZ domain catches a short motif - between 833-839 amino acid positions - at the end of the host protein's intracellular TIR domain. The PDZ domain typically binds to the C-terminal residues of target proteins, helping to organise and regulate the activity of signalling complexes ^{461,462}. Besides the PDZ domain, this Bt protein has two other S41 family peptidase domains but ELM does not contain information about the target motifs of these structural units.

The experiments carried out by the Carding group at the QIB validated this finding. The TLR4 inhibitor CLI-095 and the TIRAP inhibitor lead to an incomplete inhibition of the pathway in the presence of Bt BEVs compared to the LPS treated monocytes. I assume that the vesicular proteins from Bt can potentially interact with downstream pathway components and support the activation of the pathway even if the receptor and its adaptor protein is blocked.

It seems that TIRAP is an important target for the BEV proteins due to the 19 potential bacterial TIRAP interactors that I found. Interestingly, the four domain binding sites along the protein sequence are targeted by diverse enzymatic domains including phosphoesterase, phosphorylase, peptidase and dehydrogenase activities. Moreover, the adaptor protein shows condition-specific expression in inflammatory monocytes and DC1 cells. Based on these results, I assume that the presence of TIRAP in one of the conditions establishes an important interspecies connection between the Bt and human proteins. Because the adaptor is tightly connected to TLR4⁴⁶³, the Bt targeted cytoplasmic TIR domain on the receptor can alter the connection between TLR4 and TIRAP which could lead to altered downstream signalling resulting in disrupted pro-inflammatory cytokine secretion. All together, TIRAP could be a relevant candidate for further research in IBD treatment.

Although 2048 microbial proteins were detected in the proteomic analysis, the low number of predicted potential interactors (48) reveal the limitation of the pipeline in terms of the known structural information from bacterial proteins - discussion in Chapter 3 describes the future solution for this issue. This analysis is not suitable for depicting processes specific to a cell type or condition due to a large number of BEV interacting proteins in each cell type, therefore the output focuses mainly on common processes. A more fine-grained workflow can be achieved by involving gene expression values, and not only the presence or absence of a gene's expression when establishing condition-specific differences.

Despite the limitations described in Chapter 3, the established host-microbe interaction pipeline combines gap-filling approaches, such as structural PPI prediction and network analysis, which highlight the importance of condition and cell specificity. I not only identified new potential therapeutic targets for IBD treatment but also revealed the background of biological processes on the molecular interaction level.

Chapter 5: Predicting the effect of the oral microbiome to the host in healthy and in inflamed conditions

5.1 Introduction

The oral microbiome plays an important role in maintaining oral health. The colonisation of the oral cavity begins at birth. The first invaders are aerobic bacteria, such as *Streptococcus* (particularly *S. salivarius*), *Lactobacillus*, *Actinomyces*, *Neisseria* and *Veillonella* species. When the first tooth breaks through the gingiva, new strains inhabit the mouth resulting in a more diverse community as anaerobic organisms are able to appear in deeper layers of the gum. With tooth loss, the microbiota starts to become similar to the birth stage⁶⁸ indicating the importance of teeth in determining the oral microbiota.

Description of the healthy oral microbiota is difficult because the mouth is an open system, and is frequently exposed to exogenous bacteria in food, water, and air. Therefore studies separate the 'core' microbiome [Figure 5.1] that includes the most common taxa appearing among people from the variable microbiome characterising individuals depending on their lifestyle.

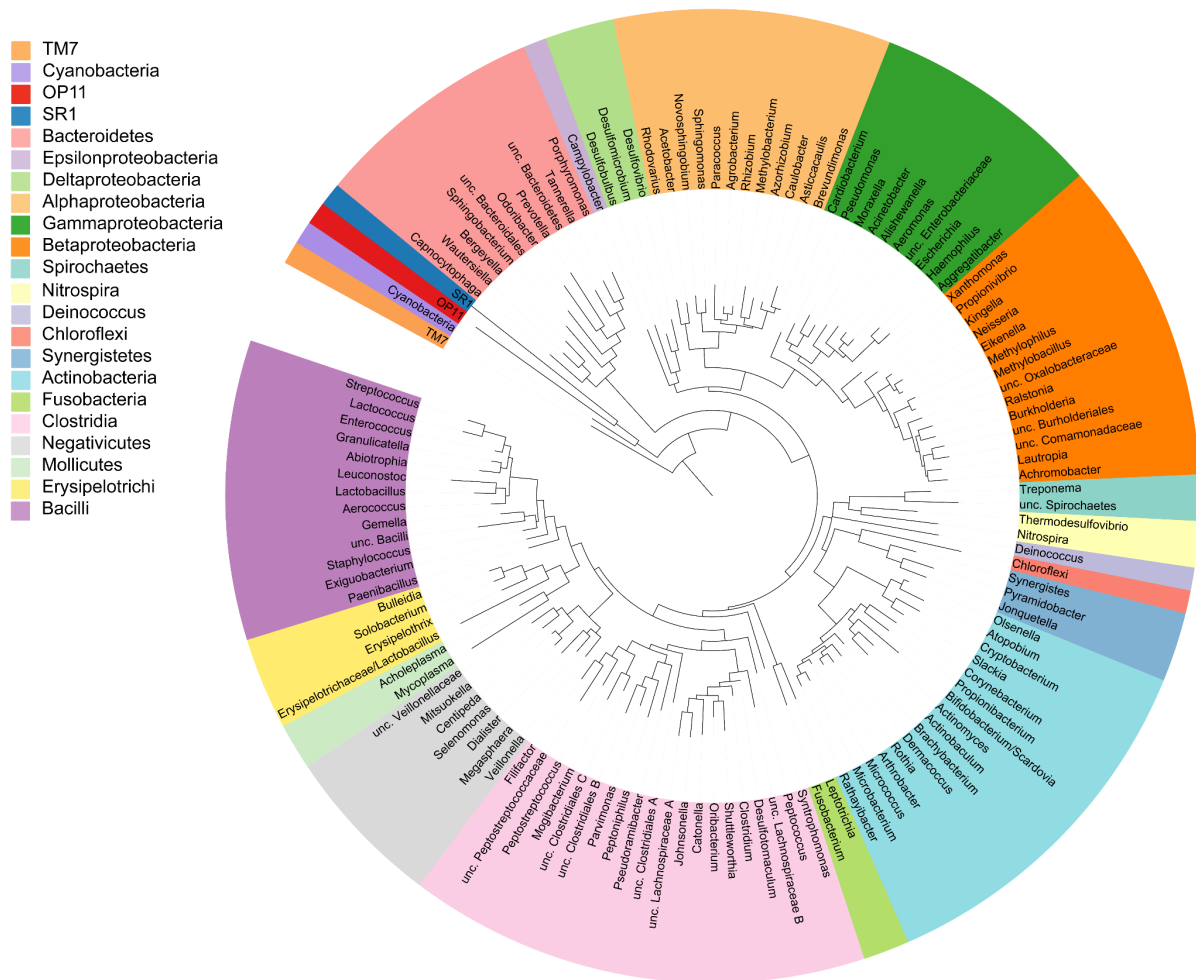


Figure 5.1: Bacteria representing the human core oral microbiome. The phylogenetic tree reveals the bacteria in the healthy oral cavity at the genus level. Source of the figure: ⁴⁶⁴

Various microbial communities are represented in the oral cavity (e.g. tongue or tonsil microbiome), the thesis focuses on the bacteria inhabiting the gingiva, which covers and protects the ligament and the neck of the tooth. Based on anatomical location the microbiome is divided into two parts: the supragingival plaque, which covers the enamel and root surface, contains Gram-positive rods and cocci bacteria (e.g. *Streptococcus mutans*, *Streptococcus salivarius*, *Streptococcus mitis*, *Lactobacillus*), appeared to be forming a tightly adherent band. Subgingival plaque is frequently characterised by anaerob Gram-negative species (*Actinobacillus*, *Campylobacter* spp, *Fusobacterium nucleatum*, *Porphyromonas gingivalis*) located adjacent to the epithelial lining of the pocket ^{82,465}.

Subgingival bacteria organise into different complexes defined by Socransky [Table 5.1]. The standardised name of the complexes has been derived from the original clustering analysis colouring, meaning that bacteria in the same colour group are similar to each other ⁴⁶⁵.

Purple complex, green complex and yellow complex characterise the early state. These bacteria facilitate the presence of the Gram-negative bacteria clusters (orange and red complexes). The orange complex consists of several bacteria that enable the appearance of the red complex ⁴⁶⁵⁻⁴⁶⁸. The red complex is usually found in the deeper periodontal pocket because these species form a separate community where interspecies interactions and metabolic cross-dependency are extremely strong ⁴⁶⁹.

Table 5.1: Bacterial clusters in subgingival plaque described by Socransky et al. (1998)				
Purple complex	Green complex	Yellow complex	Orange complex	Red complex
<i>Actinomyces odontolyticus</i>	<i>Capnocytophaga gingivalis</i>	<i>Streptococcus mitis</i>	<i>Fusobacterium nucleatum</i>	<i>Porphyromonas gingivalis</i>
<i>Veillonella parvula</i>	<i>Capnocytophaga ochracea</i>	<i>Streptococcus sanguis</i>	<i>Prevotella intermedia</i>	<i>Tannerella forsythensis</i>
<i>Actinobacillus actinomycetemcomitans</i> (serotype b)	<i>Capnocytophaga sputigena</i>	<i>Streptococcus oralis</i>	<i>Prevotella nigrescens</i>	<i>Treponema denticola</i>
<i>Selenomonas noxia</i>	<i>Campylobacter concisus</i>	<i>Streptococcus gordonii</i>	<i>Peptostreptococcus micros</i>	
<i>Actinomyces naeslundii</i>	<i>Eikenella corrodens</i>	<i>Streptococcus intermedius</i>	<i>Campylobacter rectus</i>	
	<i>Actinobacillus actinomycetemcomitans</i> (serotype a)		<i>Campylobacter showae</i>	
			<i>Campylobacter gracilis</i>	
			<i>Eubacterium nodatum</i>	
			<i>Streptococcus constellatus</i>	

Regardless of the location, microbiome composition has a huge impact on tissue homeostasis, alteration of the community composition infers a dysbiotic condition. Several external (oral hygiene, diet) and internal (autoimmune disease, immunodeficiency disorders) factors can disturb the healthy microbiome ⁴⁷⁰. The appearance of pathogens leads to enterotoxin secretion, which molecules alter the permeability of the epithelium. Host-microbe interactions are crucial for the regulation of physiological processes; alteration (rewiring) of these interspecies connections leads to inflammation in the host [details in Chapter 1.4.2] ⁴⁷¹. A serious consequence of the disrupted equilibrium state is that bacteria are able to enter the bloodstream and cause diseases, such as gingivitis and periodontitis - the two main disorders of the gum ¹⁴⁷.

While gingivitis refers to the mild, easily reversed inflammation of gum (with a prevalence in adults of over 90%) ⁴⁷², chronic periodontitis is a result of untreated gingivitis, which is a polymicrobial attack that destroys the periodontal ligament and supporting marrow that surrounds the teeth ⁴⁶⁶. Clinical studies revealed that chronic periodontitis is associated with several systemic diseases (diabetes, cardiovascular diseases, cancer) ⁴⁷³⁻⁴⁷⁵.

Van Dyke *et al* published a model which describes how the healthy gum becomes inflamed in four stages: Firstly, Gram-negative bacteria replace the Gram-positives (stage 0). This shift causes inflammation in the gingiva (stage 1), if it alters the subgingival microenvironment there will be a polymicrobial emergence (stage 2). Till this point, the process can be reversed by external and internal factors. Lack of treatment leads to impaired inflammatory processes and tissue damage resulting in deeper pockets by the tooth (stage 3). This early periodontitis turns to late-stage periodontitis (stage 4) when inflammation-mediated dysbiosis affects the gum ¹.

In some cases, the dysbiosis starts without clinical signs - especially in older people -, therefore samples from a healthy patient do not necessarily mean a healthy microbiome ⁴⁷⁶. However, a few taxa have been strongly associated with periodontal health, such as *Actinomyces* and *Streptococcus* species ^{477,478}, but the majority of bacteria inhabit both the healthy and diseased gingiva (e.g. *Fusobacterium nucleatum*, *Veillonella parvula*, *Streptococcus oralis*, *Streptococcus intermedius* and *Streptococcus anginosus*) ^{468,477}.

During the inflammation of the gum, the supragingival microbes expand to the subgingival area, therefore increasing the presence of anaerobic bacteria in the plaque. When the gingiva becomes inflamed, bacteria composition shifts from Gram-positives to Gram-negatives ⁴⁷⁹ leading to the dominance of red and orange complex members ⁴⁶⁵.

Defining the gingivitis-associated pathogens is difficult due to the mild symptoms, also patients do not visit doctors at this stage. Therefore, microbiome composition often overlaps with microbes in healthy samples, if gingivitis is at an early stage, but also with periodontitis, when the gum is not treated ⁴⁶⁵. In the thesis, I use the term 'periodontitis' to describe the inflammation in the gingiva.

Currently, there are studies which describe experiments or computational pipelines to analyse HMIs in the oral cavity ^{138,480-484}. However, there are limitations in terms of data quality, which means most of these approaches are working with a few microbes and exploring their effect on a tissue or cell line. With the appearance of meta-omics and single-cell transcriptomics data this gap has been addressed and I could establish a workflow during my PhD which aims to predict the effect of complex microbial communities on host signalling at the cell type level.

Instead of focusing on the whole oral microbiome, I explored the subgingival microbiome and its role in inflammation. Analysing publicly available datasets facilitated understanding the composition of the microbiome on higher taxonomic levels. Studies highlighted that *Firmicutes*, *Tenericutes*, *Proteobacteria*, *Actinobacteria*, *Bacteroidetes* and *Fusobacterium* taxa characterise mostly healthy gum ^{467,477,485} while members in orange and red complexes dominate in periodontitis [details in Chapter 1]. In this use case study, I was focusing on a limited list of strains that appear dominantly in healthy gum and during periodontitis [Table 5.2].

Table 5.2: List of bacterial strains analysed in the study

Condition	Strain	Role in the gingival microbiota
Healthy gum	<i>Streptococcus sanguinis</i> SK36	Gram-positive facultative anaerobes, one of the first appearing bacteria which help to colonise the gingiva. <i>S. sanguinis</i> stimulates the epithelial layer to express IL-8 and β -defensins to defend against periodontitis-associated pathogens ⁴⁸⁶ .
	<i>Haemophilus parainfluenzae</i> ATCC 33392	Gram-negative facultative anaerobe bacteria, one of the most abundant species in the healthy supragingival plaque interacting often with <i>Streptococcus</i> species (especially with <i>S. australis</i> , <i>S. infantis</i> , <i>S. pneumoniae</i> , <i>S. oralis</i> and <i>S. mitis</i>) ⁴⁸⁷ .
	<i>Lautropia mirabilis</i> ATCC 51599	Gram negative facultative anaerobe bacteria contribute to the healthy gingival microbiome but are dominant in mild inflammation affected gum microbiome ⁴⁸⁸ .
	<i>Veillonella parvula</i> ATCC 10790	Gram-negative anaerobe bacteria, also being an early coloniser along with <i>Streptococcus sanguinis</i> , facilitates the colonisation of the orange and red complex members in the advanced state of the inflammation ⁴⁶⁵ .
Periodontitis	<i>Porphyromonas gingivalis</i> ATCC BAA-308	Gram-negative anaerobe bacteria in the red complex inducing cytokine expression (IL-6, IL-8) in the host epithelial cells and contributes to severe inflammatory processes ⁴⁸⁹ .
	<i>Treponema denticola</i> ATCC 35405	Gram-negative anaerobes, which are strongly connected to the other two bacteria in the red complex (<i>P. gingivalis</i> , <i>T. forsythia</i>).
	<i>Tannerella forsythia</i> ATCC 43037	Gram-negative anaerobe bacteria, which secrete virulence factors having an influence on microbial community composition, therefore, leading to dysbiotic state and causing inflammation in the host ⁴⁹⁰ .
	<i>Filifactor alocis</i> ATCC 35896	Gram-positive anaerobe bacteria responsible for inflammation-related processes in the host. It is not only a potential new member of the red complex but also interacts with the core member <i>Porphyromonas gingivalis</i> ⁴⁹¹ .

Overall, I aimed to discover the effect of bacteria not only on the target cell's signalling but also on intercellular interactions. Therefore I combined the MicrobioLink2 [details in Chapter 3] and the intercellular interaction pipeline [details in Chapter 2] to reveal interspecies interactomes and their effect on cell-cell connections in healthy gum and in severe periodontitis.

5.2 Methods

I built up a case study based on the computational pipelines that have been described in Chapter 2 (intercellular interaction pipeline) and 3 (MicrobiLink2 pipeline). In this section, I would like to highlight the novelty and importance of the algorithms through analysing a public dataset [Figure 5.2].

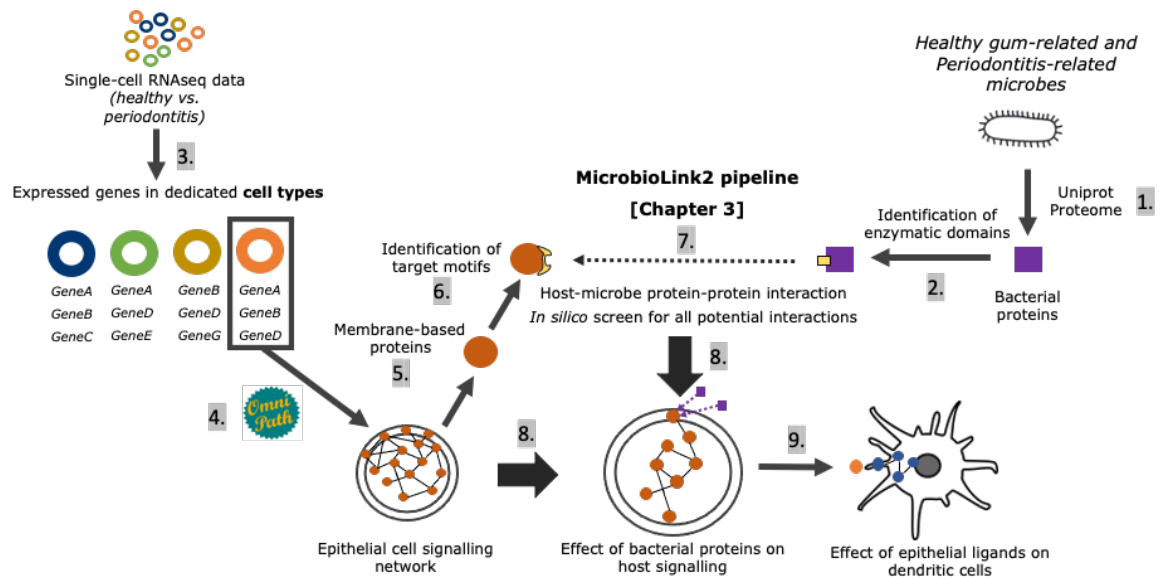


Figure 5.2: Computational workflow to analyse the effect of the gingival microbiome on epithelial and immune cells in periodontal health and disease. Numbers indicate the sequence of the main steps: 1, Downloading the proteome of periodontal health- and disease-related bacteria. 2, Identifying the domain structure using the Pfam database. 3, Processing single-cell RNAseq data from the gingiva, 4, Creating epithelial cell-specific network using the list of expressed genes combined with protein-protein interactions from OmniPath¹⁰⁷, Selection of membrane-based proteins using OmniPath. 6. Identifying SLiMs on membrane-based proteins using ELM database. 7, Predicting protein-protein interactions (PPIs) between microbial and host proteins using MicrobiLink2. 8. Building up a downstream signalling network to follow the signal from the bacteria-targeted membrane proteins till the expressed genes which will be translated to ligands. 9. Building up intercellular interaction network between epithelial cell secreted ligands and receptors on DC's surface using ligand-receptor interactions from OmniPath.

5.2.1 Collection of bacterial proteins

I selected a limited number of strains that are dominant in the healthy and diseased oral cavity based on an internal discussion with Unilever. Due to the lack of proteomics experiments, I downloaded all the proteins from UniProt Proteome²⁷⁴ for each strain. In the thesis, I have used reference proteomes that have been selected by the research community or by computational clustering filtering to the best-annotated proteomes in Uniprot. Using the PSORTb tool⁴⁹², locations have been predicted for each protein, but due to the high number of 'unknown' and multiple location annotated proteins, I decided not to filter the bacterial proteins based on their place in the cell.

5.2.2 Single-cell transcriptomic analysis

I analysed a publicly available study published by Caetano *et al*¹⁹³. Samples were taken from the buccal gingival margin region from four patients (two healthy, one with moderate periodontitis and one with severe periodontitis). Matthew Madgwick processed the published raw dataset (GSE152042) with the parameters defined in the original article. In general, errors in omic data analysis can arise from various sources such as poor quality control, data processing or statistical methods. The developed in-house pipeline is aware of the potential sources of bias and tries to minimise them by using appropriate quality control measures, and bioinformatics methods [details in Chapter 2].

Output files described the normalised count values for each gene in each cell and the average expression of genes under healthy, mild and severe periodontitis. I selected the healthy and severe periodontitis conditions to study the effect of microbes on host signalling.

5.2.3 RNAseq data filtering

To filter the processed RNAseq dataset, I used the same z-score normalisation method as described in Chapter 2, however, I added another gene expression filtration method for the data. Single-cell transcriptomics measures the gene expression in each individual cell in the sample and calculates an average of expression values counting with all of the cells. I discarded those genes which were expressed in less than 10% of cells clustered in a cell type in a specific condition. This method facilitates discarding technical or biological issues, such as differences between samples or lack of gene expression, coming from the experiment.

5.2.4 Inferring a host-microbe interaction network

I selected the epithelial cells from the single-cell dataset as the first layer that interacts with the bacterial community. The scRNAseq identified three different subpopulations of epithelial cells based on their differentiation states and markers: basal cells expressing *HOPX*, *IGFBP5* and *LAMB3*; proliferating basal cells expressing *MKI67* and *TOP2A*; and mature cells expressing *KRT1*, *KRT8*, *LAT* and *PTGER* marker genes. Because I was interested in the interactions between microbes and human proteins on the surface of the gingival layer, I selected the mature epithelial cells for the analysis. Assuming that every gene - which passed the filtration criteria - is translated to functional active proteins, I selected the membrane-based candidates using OmniPath and downloaded their sequences from UniProt. The established MicrobioLink2 pipeline was used to connect the healthy- and severe periodontitis-related microbiome dominant strains to the expressed human proteins.

5.2.5 Functional analysis of microbe-targeted human proteins

The *in silico* prediction highlighted the potential bacteria-affected human membrane proteins. I carried out a functional analysis to reveal their role in biological processes using the GOrilla web-based tool (<http://cbl-gorilla.cs.technion.ac.il/>)⁴⁰¹ [details in Chapter 3].

5.2.6 Downstream network modelling

Network propagation algorithms help to connect the perturbation points (host proteins which are in contact with the microbial proteins) to the (differentially) expressed genes through PPIs, and give a detailed insight into the signal spreading. I used TieDie²⁵⁰ to look at the signalling pathways affected indirectly by microbes by binding to the cell surface proteins [details in Chapter 3].

In this use case, I modelled two networks, one for the healthy condition and one for severe periodontitis. The reason for not using differentially expressed genes is that I aimed to reveal signalling processes in the two conditions separately, not only focusing on differences but including overlapping functions as well. The final network described the signal spread in the following order: bacterial protein → human targets → signalling pathway → transcription factor → target gene expression. To avoid large interactomes, I was focusing on the effect of host-

microbe interactions on ligands secretion, therefore as endpoints, I selected those expressed genes which are translated to ligands using annotations from the OmniPath¹⁰⁷.

I analysed which pathways could be potentially activated by upstream host-microbe interactions. I repeated the same functional analysis as in the case of bacteria-targeted human proteins, however the input protein list contained members of the downstream signalling network.

5.2.7 Reconstructing an epithelial cell - immune cell interaction network

I connected the epithelial layer to dendritic cells through ligand-receptor interactions (LRIs) to analyse the effect of the altered microbiome composition in the subgingival plaque on immune cells. Details about the intercellular interaction workflow are described in Chapter 2.

As a final step, to have a look not only at the pathways but also at the downstream processes which have been affected by the epithelial ligands, I created a dendritic cell-specific signalling network for both the healthy and diseased conditions to follow the signal spread in the cytosol as well. I selected the receptors which were in connection with ligands and their first neighbours - the proteins which they are interacting with - and created a subnetwork. This time, I used the Reactome database instead of the GO term-specific GOrilla tool. I looked for enriched pathways which were reached by the receptor using the default background settings in Reactome (curated entities in the database).

5.3 Results

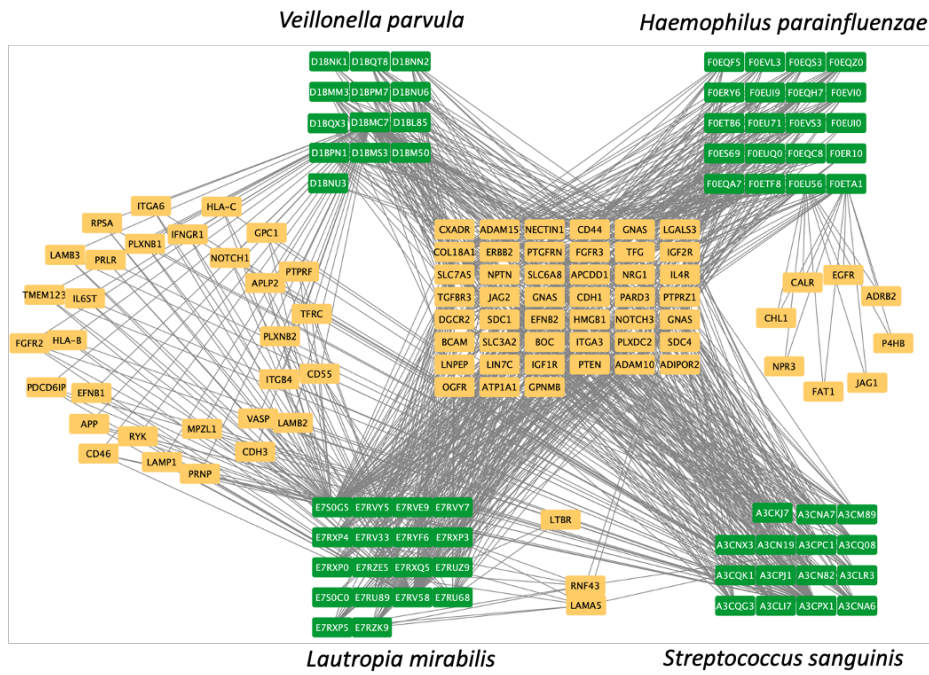
5.3.1 *An in silico* host-microbe protein-protein interaction network

I downloaded the bacterial proteomes from each condition to identify the proteins (~2000-3000 protein/strain) and their domains [Supplementary Table 5.1]. Meanwhile, 3344 genes were described in healthy epithelial cells and 3916 genes in severe periodontitis samples, although the selection of membrane-based proteins reduced their number. The *in silico* prediction identified 921 HMIs in healthy and 91 HMIs in diseased condition [Figure 5.3]. I found 8 domains out of 831 in health-related and 13 out of 1577 domains in periodontitis-related bacterial proteomes which can cause a significantly smaller size of the diseased network

[Table 5.3]. These results suggest that commensal and pathogenic proteins may have specific domain structures which can not be found in the ELM and 3did databases.

Table 5.3: Bacterial Pfam domains targeting SLiMs on human proteins			
<i>Healthy condition</i>		<i>Periodontitis</i>	
Pfam Accession	Domain name	Pfam Accession	Domain ID
PF00149	Calcineurin-like phosphoesterase domain	PF00149	Calcineurin-like phosphoesterase domain
PF00899	ThiF-family domain	PF00899	ThiF-family domain
PF00533	BRCT domain	PF00533	BRCT domain
PF00082	S8 peptidase domain	PF00082	S8 peptidase domain
PF00389	D-isomer specific 2-hydroxyacid dehydrogenase, catalytic domain	PF00389	D-isomer specific 2-hydroxyacid dehydrogenase, catalytic domain
PF01048	Phosphorylase domain	PF00675	M16 peptidase domain
PF00535	Glycosyl transferase family 2 domain	PF00535	Glycosyl transferase family 2 domain
PF00595	PDZ-domain	PF01048	Phosphorylase domain
		PF00515	Tetratricopeptide domain
		PF00089	Trypsin domain
		PF01344	Kelch domain
		PF00498	FHA (Forkhead-associated) domain
		PF00069	Protein kinase domain

A,



B,

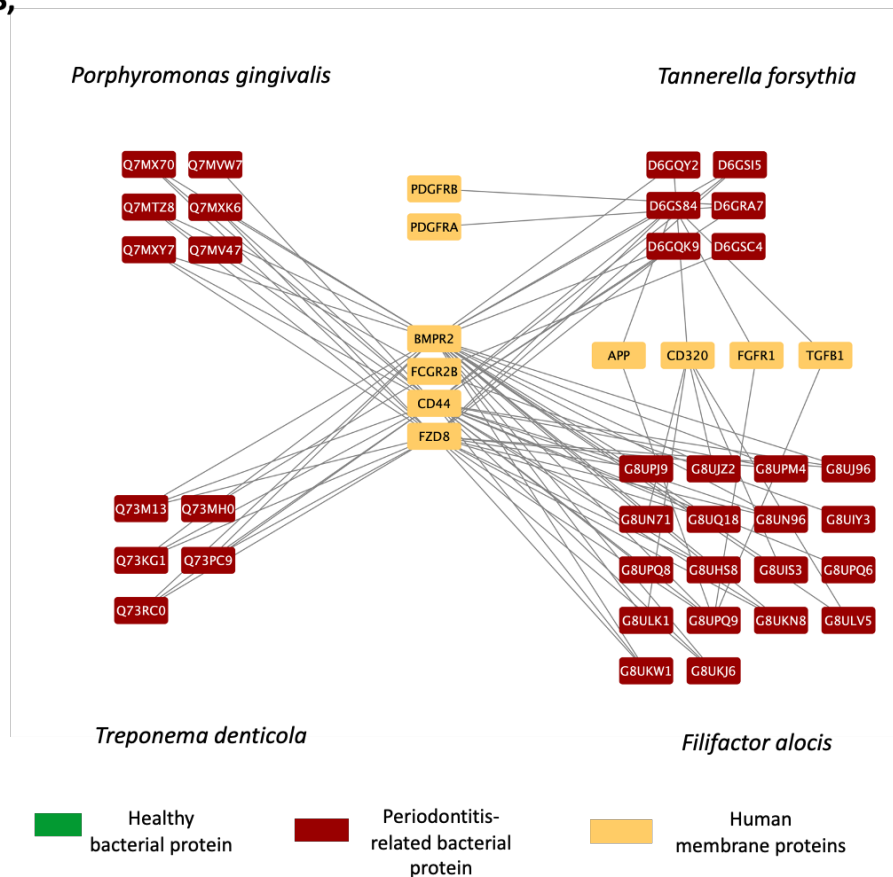


Figure 5.3: Host-microbe interactions in A) healthy and B) severe periodontitis conditions predicted by MicrobioLink2. Healthy (green) and diseased (red) bacterial proteins are grouped by strains and linked to the membrane-based human proteins highlighted by yellow. The networks were created with Cytoscape³²⁶.

5.3.2 Functional analysis of host target proteins

The GOrilla tool highlighted several biological processes among bacteria-targeted membrane proteins although the results were difficult to analyse due to the redundant annotations in the database. Therefore I visualised the GO terms in REVIGO ⁴³⁹. The REVIGO tool organises the annotations and removes the redundant terms, therefore, facilitating to identify of overlapping functional categories, such as metabolic or biosynthetic processes, and chromatin organisation but also reveals differences between conditions, such as Notch signalling in healthy gum or regulation of MAPK cascade in severe periodontitis [Figure 5.4].

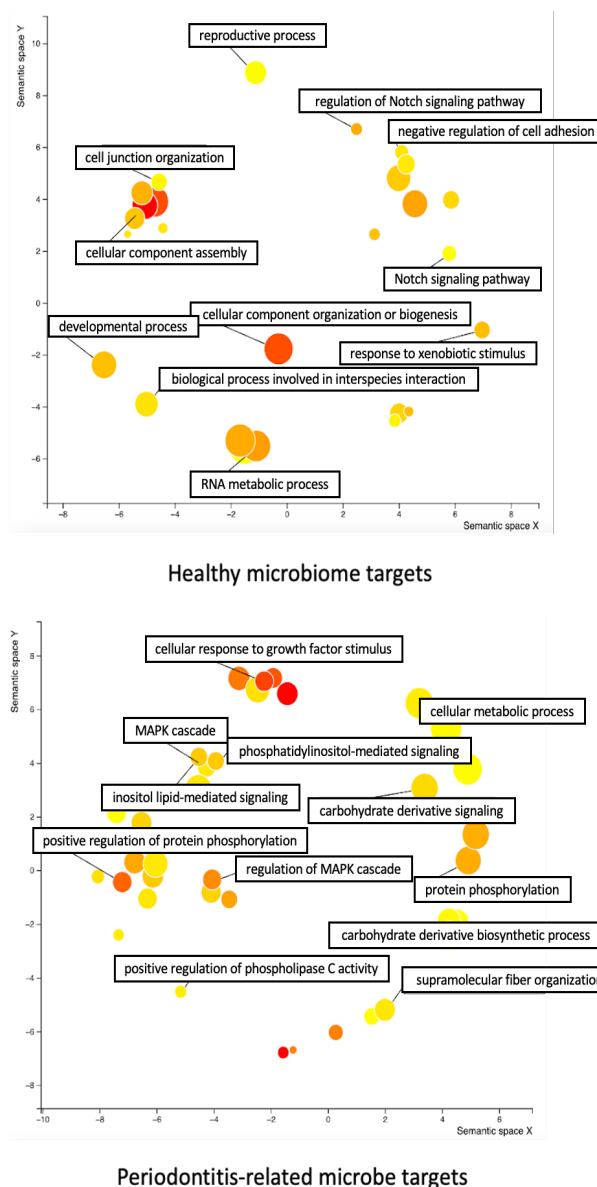


Figure 5.4: Functional analysis of bacteria targeted human proteins. The size of the points is equal to the number of proteins involved in the function, the colour represents the log10 p-value (red- lowest value, yellow - highest value). The diagrams were created with the REVIGO tool ⁴³⁹.

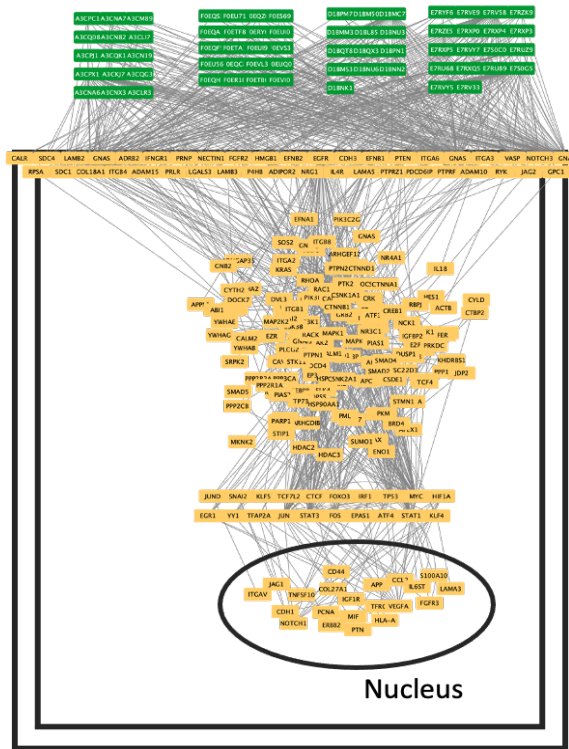
5.3.3 Downstream signalling network modelling

I built up the signalling networks of healthy and periodontitis-affected epithelial cells using TieDie. The input for the algorithm included (1) bacteria-affected membrane-based receptors potentially translated from genes identified from scRNAseq dataset (86 genes from healthy and 10 genes from diseased condition); (2) directed PPI network from OmniPath including 47925 interactions; (3) expressed genes in healthy and periodontitis samples (110 genes from healthy and 151 genes from diseased condition). The networks consisted of five different types of nodes: bacterial proteins, human membrane proteins, intermediate signalling proteins, transcription factors (TFs) and expressed genes that are potentially translated to ligands. Interestingly, although the number of bacteria-affected proteins was significantly less in periodontitis (5), the number of proteins in the intermediate network was similar to the healthy condition [Figure 5.5]. Besides, comparing the edges, I found a ~50% decrease in the number of connections in inflamed condition. These findings assume that the triggered signal by HMIs is less scattered and specific pathways were activated during periodontitis.

To establish statistical evidence, a randomised network analysis was conducted to examine the number of nodes and edges. Initially, 500 sets of five membrane proteins were randomly selected from the OmniPath database. TieDie was then run using the same intracellular network and downstream input utilised in the periodontitis analysis. The distribution of total node/edge counts among the 500 networks was visualised and the mean and standard deviation of the attributes were calculated to obtain the z-score (as described in Chapter 2.2.3).

The results showed that the average number of nodes in the random networks was 65 (standard deviation = 17.8) and the average number of interactions was 201 (standard deviation = 74.6). This indicated that the periodontitis network, consisting of 158 proteins (not including the 19 downstream genes), had significantly more proteins compared to a network connecting random 5 proteins with the same downstream genes (z-score = 5.22). However, the edge number analysis revealed that the original periodontitis network, with 663 interactions, was still more connected than the random networks (z-score = 6.24).

A,



Bacterial proteins (66)

Bacteria-targeted membrane proteins (41)

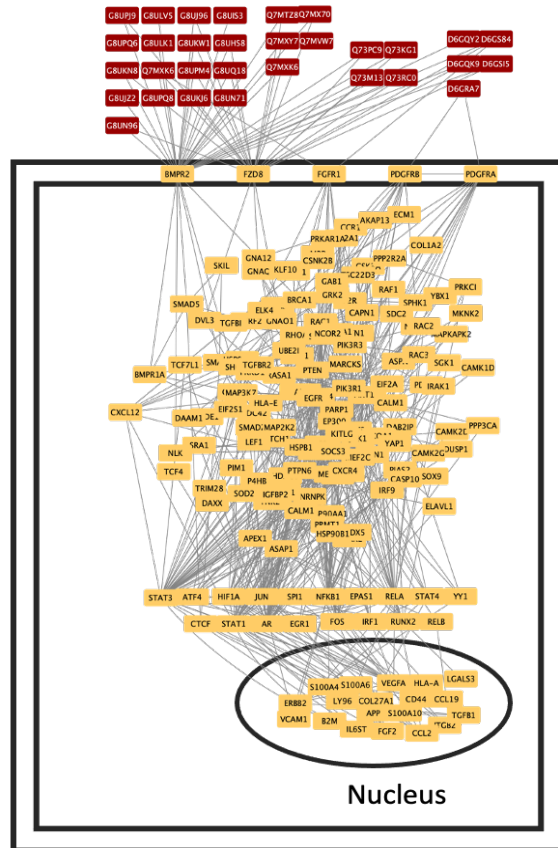
Intermediate protein-protein interaction network (140)

Transcription factors (20)

Genes translated to ligands (21)

- Healthy bacterial protein
- Human membrane protein

B,



Bacterial proteins (31)

Bacteria-targeted membrane proteins (5 + 3 as expressed ligands)

Intermediate protein-protein interaction network (135)

Transcription factors (18)

Genes translated to ligands (19)

- Periodontitis-related bacterial protein
- Human membrane protein

Figure 5.5: Signalling network in epithelial cells focusing on the downstream effect of bacteria in A, periodontal health and B, during severe periodontitis. The figures show the output of the TieDie algorithm connecting the upstream perturbation points to the expressed genes.

I revealed the pathways/functions in which the downstream components play a role using GOrila. While ~70% (813 annotations) of the processes did overlap between the conditions, I found intriguing differences in the rest of 30%, such as negative regulation of T cells and regulation of B cell activation in healthy condition and cytokine-mediated signalling (IL-12, IL-6, IL-23) and TLR signalling in diseased condition-specific networks [Figure 5.6]. In the healthy network, out of the 3344 expressed genes, 222 are represented in the subnetwork. Similarly, in the diseased network, out of the 3916 expressed genes, 177 are represented. It is important to note that revealed annotations are primarily focused on the bacteria-targeted proteins and their impact on ligand secretion, as only approximately 7% of the molecules have been found in the subnetworks.

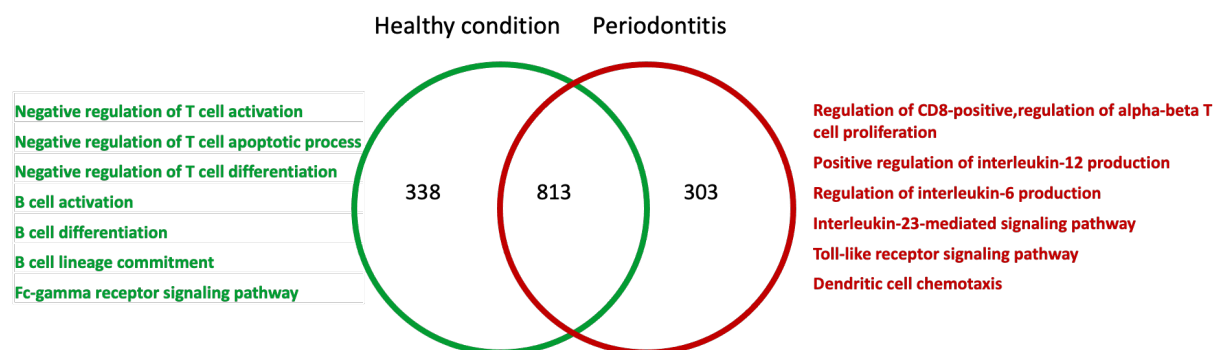


Figure 5.6: Overlap between functions across bacteria-affected membrane proteins, intermediate proteins, TFs and expressed genes translated to ligands.

5.3.4 Interaction between epithelial and dendritic cells

Epithelial cells are able to secrete immune system modulatory cytokines⁴⁹³, and have an impact on DCs⁴⁹⁴ therefore, I explored the interplay between the two cell types. As the results show, the epithelial cells express different sets of ligands under diverse conditions. The inflammation-related *TGFB1* and *CCL19* were expressed only during severe periodontitis and were not found among genes in the healthy cells. Also, the functional analysis of bacteria-targeted human proteins identified the Notch pathways were uniquely affected in healthy condition, here I found that the *NOTCH1* is expressed only in healthy state. Having a closer look at the receptors on the DC's surface also revealed condition specificity as I identified 60 receptors from healthy and 48 receptors from diseased cells [Figure 5.7].

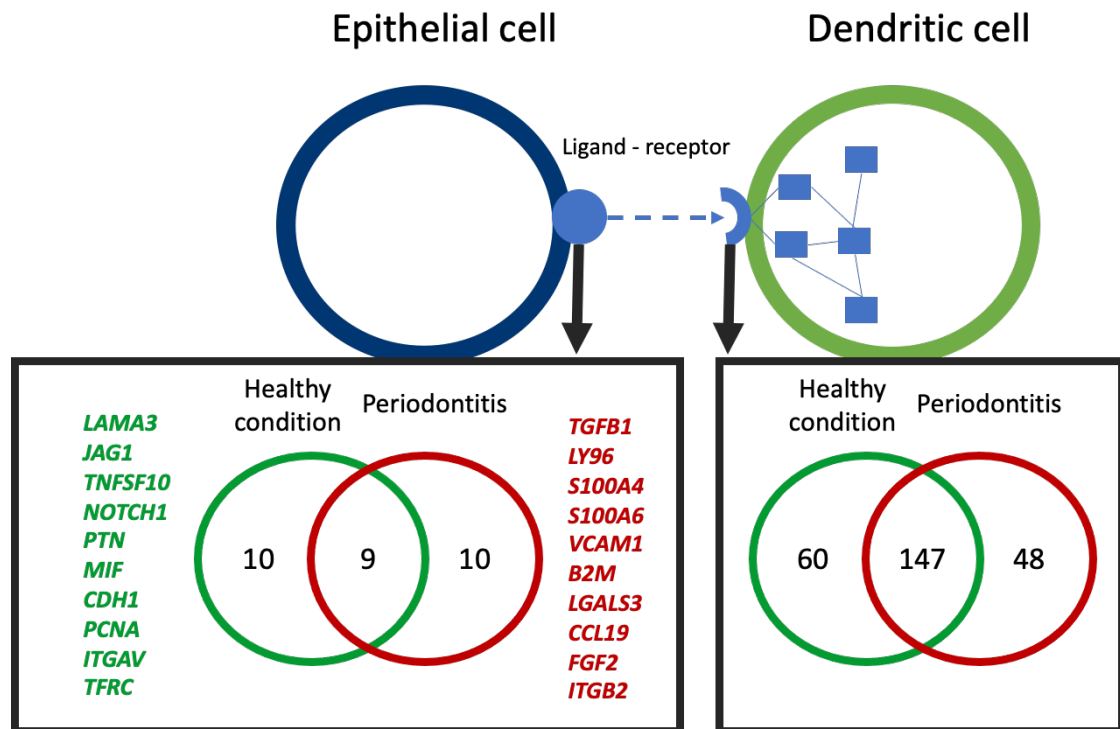


Figure 5.7: Overlap of epithelial cell secreted ligands and DC expressed receptors in healthy (green) and periodontitis (red) conditions. The condition specific ligands (genes coding them) are listed on the left side.

I connected the ligands to the receptors which resulted in 313 LRIs in the healthy gum and 328 LRIs in periodontitis. I visualised the interactions on a circos plot [Figure 5.8]. In OmniPath, some proteins have both receptor and ligand annotations, thus, I discarded these multifunctional points when creating the plot.

I compared the receptors on the target cell surface not only by their presence or absence but also by their role in signalling pathways (Innate immune system-related-, JAK/STAT-, Notch-, Receptor Tyrosine Kinase (RTK)-, WNT- and TLR signalling) using information from SignalLink3. I found that there is no difference on the pathway level, however, the signalling components vary between conditions. Most of the LRIs have an effect on TLR signalling, while only a few are related to the Notch pathway [Figure 5.8].

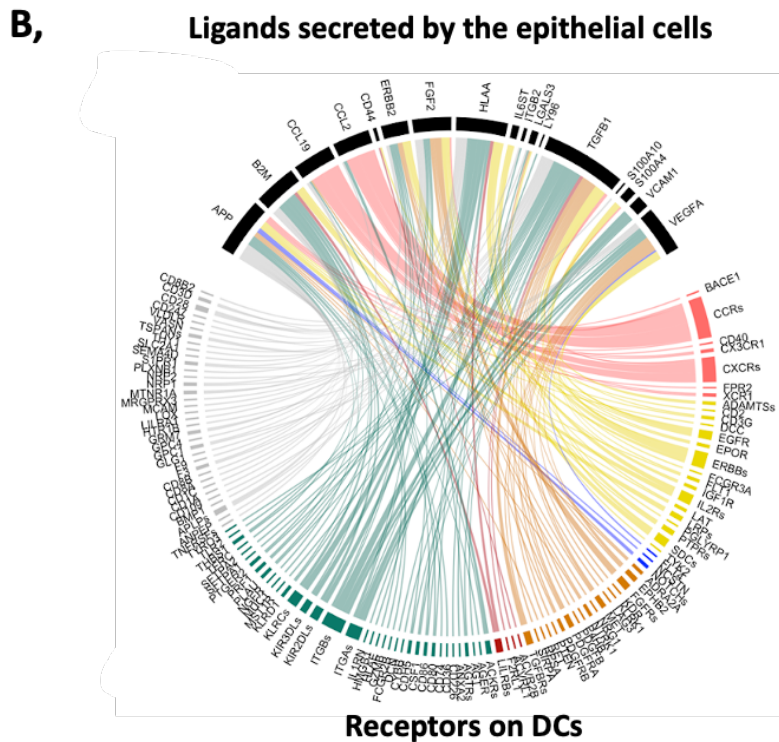
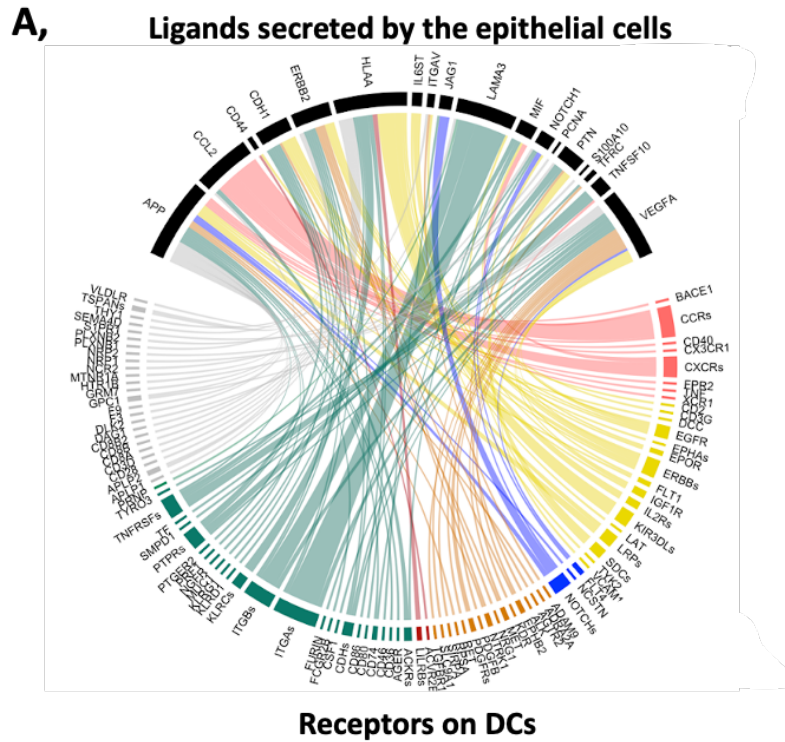


Figure 5.8: Condition-specific connections between epithelial cell ligands (upper semicircles, black) and DC receptors (lower semicircles, coloured by pathways) in A, healthy control and B, severe periodontitis. Immune—innate immune response, RTK—receptor tyrosine kinase, TLR—Toll-like receptor. Circos plots were created by using the 'circlize' R package³²⁵.

As a complementary analysis, I examined the changes in the perturbed signalling by the Reactome database, but I extended the analysis to the first neighbours of the receptors (ie., the direct protein interactors of the receptors) [Table 5.4 ($p < 0.05$, FDR = 5.11E)]. Results suggest that most of the signalling overlap between conditions but there are condition-specific differences, such as Notch signalling in healthy gum and death receptor signalling and MyD88-independent TLR4 cascade in severe periodontitis.

Table 5.4: Top 10 signalling pathways represented among the receptors and their first neighbours	
Healthy condition	Periodontitis
Signalling by CSF3 (G-CSF)	Signalling by CSF3 (G-CSF)
Inactivation of CSF3 (G-CSF) signalling	MyD88-independent TLR4 cascade
Constitutive Signalling by NOTCH1 PEST Domain Mutants	TRIF(TICAM1)-mediated TLR4 signalling
Signalling by NOTCH1	Interleukin-3, Interleukin-5 and GMCSF signalling
Interleukin-3, Interleukin-5 and GMCSF signalling	Toll-Like Receptor 4 (TLR4) Cascade
TRIF(TICAM1)-mediated TLR4 signalling	Interleukin-4 and Interleukin-13 signalling
Interleukin-4 and Interleukin-13 signalling	Death Receptor Signalling
Toll-Like Receptor 3 (TLR3) Cascade	Signalling by Interleukins
VEGFA-VEGFR2 Pathway	Toll-like Receptor Cascades
MyD88-independent TLR4 cascade	Cytokine Signalling in Immune system

5.4 Discussion

The microbiome plays an important role in homeostatic processes in the host therefore the altered community composition leads to differences in host signalling⁴¹. Currently, there are no studies which explore the role of complex microbiota on epithelial cell signalling and also infer cell-cell signalling networks between the epithelial and immune cells to explore the role of altered intercellular communication during inflammation. In this chapter, I presented a use case of the developed host-microbe and intercellular interaction pipelines to highlight the role of the microbiome in subgingival plaque on host inflammatory processes. Cell type- and condition-specific gene expression profiles lead to rewired protein-protein interactions between microbes and the host.

I analysed a publicly available single-cell transcriptomic dataset and combined it with network resources to establish a host-microbe interactome between a limited list of bacteria dominantly appearing in healthy condition or severe periodontitis and epithelial cell from the marginal part of buccal gingiva. Besides, I created cell-cell interactomes focusing on LRIs between epithelial and dendritic cells to show the indirect role of the altered microbiome on immune system modulation during severe periodontitis.

In general, the host-microbe interaction prediction revealed a potentially decreased amount of PPIs in the diseased state. In a healthy state, host-microbe interactions are often beneficial to both the host and the microbe. The microbe may help to maintain a balance of the host's gut microbiome, for example, by competing with other microbes for resources or by producing molecules that modulate the host's immune response. In contrast, in a diseased state, the host's immune system may respond more strongly to the presence of the microbe, leading to inflammation and tissue damage. This increased immune response can disrupt the normal interactions between the host and microbe, making it difficult for the microbe to survive in the host. Additionally, the microbe itself may produce toxins or other virulence factors that contribute to the disease state, further disrupting host-microbe interactions. Also, the microbe may avoid the host immune system by mutating, changing surface proteins, and hiding inside host cells. Therefore microbes may evade detection and reduce host-microbe interactions and in a diseased state, the number of host-microbe interactions is less than in the healthy state.

The results highlighted several already published responses, such as the activated MAPK cascade during periodontitis leading to cytokine secretion^{495,496} and the central role of the TLR pathway upon infection and inflammation^{497,498}, but I also found surprising outcomes of the

analysis. Firstly, the established host-microbe interaction networks showed differences in terms of the number of their host targets. I found a low number of membrane-based targets (8) during periodontitis due to the limited number of interacting bacterial proteins (31). Because pathogenic bacteria have more potential domains to reach the host proteins I assume that some pathogen-specific domains can not be found in the ELM or 3did database, therefore these structural elements can not be part of the prediction⁴⁹⁹. Secondly, the Notch signalling came up several times during the analysis, as a pathway affected by the healthy microbes. Notch pathway is important in cell differentiation and essential for bone development. Recently, researchers identified the altered Notch signalling contributing to severe periodontitis. Experiments show that a lower level of *NOTCH1* is related to periodontitis in patients^{500,501}. The analysis of expressed ligands by epithelial cells supported this statement by identifying NOTCH1 as a healthy condition-specific ligand. Thirdly, functional analysis of the microbiome triggered downstream signalling showed that B cell activation and differentiation is affected and T cell activation is negatively regulated in healthy condition while in periodontitis the proliferation of CD8+ alfa-beta T cells is enhanced in epithelial cells. This T cell subpopulation expresses the alfa and beta chains of the T cell receptors and is responsible for MHC-I complex recognition [details in Chapter 1] and for the elimination of malignant/infected cells⁵⁰². In terms of the affected B cell differentiation and activation, studies show that the amount of B cells is low in healthy gum and also mostly memory B cells are represented⁵⁰²⁻⁵⁰⁷. There is no information about the effect of proteins on B cell signalling but based on the literature the B cell activation pathway should be negatively regulated and the differentiation shifted towards memory B cell production.

Furthermore, several cytokine-related pathways were found among the affected proteins in periodontitis. Although I identified proteins playing a role in the positive regulation of IL12 secretion, this cytokine is expressed exclusively by immune cells. Because cytokine signalling is a complex network consisting of pathways which are cross-talking, potentially those proteins have been highlighted here which play a role in other cytokine secretion pathways, such as IL6 or IL23 expression. Experiments support the fact that IL6 and IL23 expression by epithelial cells is enhanced in gingiva when *Porphyromonas gingivalis* - a member of the red complex - is presented in the microbiome^{482,508}.

The functional analysis highlighted that DC chemotaxis is among the periodontitis-specific processes triggered by HMI downstream that assumes an altered communication between the gingival epithelium and DCs. This finding drove me in the direction of observing a potential altered communication between epithelial and dendritic cells during periodontitis.

The migration of these immune cells is usually caused by the interaction of chemokines and DC receptors. Such an interaction takes place between CCL19 chemokine and the CCR7 receptor which has been identified in the intercellular communication analysis [Figure 5.8] ⁵⁰⁹. Although results show that the number of LRIs did not change between conditions, the type of intercellular communication was altered.

I used two resources, SignaLink3 and Reactome, to analyse the role of epithelial ligand-affected receptors (and their first neighbours in Reactome analysis) in different signalling pathways. Both of the examinations revealed TLR signalling, as the most affected pathway in the intercellular network. Not surprisingly, TLRs are important receptors on the surface of DCs, especially during pathogen infection, controlling cytokine secretion ⁵⁰⁹. One of the most important differences among the affected receptors contributing to TLR signalling activation is the healthy gum-related expression of the CD46 co-receptor. A recently published study highlights the role of this receptor in the downregulation of CXCL-10 inflammatory chemokine in DCs ⁵¹⁰ which shows the control of the host inflammation processes in a homeostatic state. In contrast, analysis of periodontitis-derived samples revealed the unique expression of killer-cell immunoglobulin-like receptors (KIR) on the surface of DCs. KIRs are able to sense pathogens and activate cytokine expression usually on the surface of Natural Killer cells, but literature provides information about its expression in DCs as well ⁵¹¹.

Whilst providing new and potentially important insights into the altered microbiome composition and its effect on inflammation, the analysis has several limitations: (1) the lack of metadata resulted in the examination of the whole proteomes in each bacterial strain, (2) at the time of the analysis only one scRNAseq dataset was available to explore gene expression in healthy and severe periodontitis derived gingival cells, (3) the number of patients was low, only four. The detailed limitations of the intercellular interaction and MicrobioLink2 pipelines are described in Chapter 2 and Chapter 3.

Despite these challenges, the use case provides a deeper insight into the effect of the altered microbiome on host immunity at the protein level. In addition to predicting the affected host processes supported by the literature, I was able to reveal the molecular background and the key points in the signalling networks which facilitates the identification of new targets for experimental validation.

Chapter 6: Perspectives and final discussion

The epithelial layer establishes a tightly connected barrier separating the microbes from the body, including most of the immune cells. To maintain homeostasis, these epithelial cells communicate with the nearby immune cell populations. Understanding the interspecies host-microbe and the epithelial-immune intercellular interactions is crucial because their altered interactions lead to inflammation in host tissues. During my PhD, I aimed to develop computational workflows that examine interactions between the microbiome, the epithelium and the immune system. I chose inflammation-related diseases in the gastrointestinal tract as case studies for *in silico* analyses.

Intercellular interactions are essential for developing and growing multicellular organisms. It is well-studied that the communication between epithelial and immune cells coordinates responses to maintain homeostasis and prepare host defence⁵¹². Nevertheless, recent studies revealed that fibroblasts are also important components in immune cell regulation and modulate locale immune response⁵¹³.

In Chapter 2, I presented an *in silico* intercellular interaction workflow to explore cell-cell interactions in healthy and diseased conditions. As part of this project, I worked on the update of the OmniPath database, a resource contributing to understanding cell-cell signalling at the molecular interaction level. Researchers can ask fundamental questions about cellular communication or physical cell-to-cell interactions and address them by using OmniPath combined with single-cell data analysis. To demonstrate the LRI pipeline, I analysed public single-cell transcriptomic data³¹² from healthy and ulcerative colitis (UC) patients. UC is a subtype of inflammatory bowel disease (IBD) where the colon and rectum become inflamed. Cell-cell interactions are rewired during the disease³¹², however, a limited number of studies share proof of the altered intercellular communications. The developed LRI pipeline revealed essential information about cell-cell connections in disease, such as the shift of target cells from the dendritic cells (DCs) in the healthy colon to regulatory T cells (Tregs) in diseased samples. Also, the focused myofibroblast-Treg interaction analysis showed the central role of the target cells to switch between pro- and anti-inflammatory signalling pathways depending on the interacting myofibroblast's ligands.

The microbiome plays an essential role in homeostatic processes too. The microbial community consists of both commensal and harmful microbes but these species are often in an equilibrium state in healthy conditions. Studies show that dysbiosis disturbs this balance and contributes to inflammation and the appearance of diseases, such as gingivitis in the oral cavity or inflammatory bowel disease in the gut ⁸³. However, the current knowledge about the molecular details of how pathogens modulate inflammation-related pathways is still limited ⁵¹⁴. In Chapter 3, I presented MicrobioLink2, an *in silico* host-microbe interaction prediction algorithm that facilitates the understanding of cross-species interplays and their downstream effect on host signalling including inflammatory processes. This integrated approach is also capable of pointing out key microbial inferences, and cellular pathways transmitting normal microbial signals. The structural composition-based approach highlights the exact bacterial domains and their target motif on host proteins that gives a detailed insight into the mechanism of the protein-protein interactions.

Although there has been a steeply increasing amount of data in IBD research, patients still suffer from life-long symptoms. Current therapies aim to keep patients in a clinical remission state by suppressing the symptoms. The drawback of these treatments is that patients' intestinal tract is still exposed to inflammation that could lead to long-term problems, such as colon cancer ⁵¹⁵. The knowledge about gut microbiome composition is expanding due to the elevated number of meta-omic datasets but also host response becomes more understandable with the use of single-cell analysis. The established pipelines attempt to predict new therapeutic targets to treat IBD patients.

Interspecies interactions are crucial for the initiation and progression of periodontal diseases. Pathogens secrete proteases and endotoxins to destroy the extracellular matrix and trigger an inflammatory response. There is direct evidence for the contribution of the altered microbial film around the teeth to pro-inflammatory cytokine secretion by gingival cells ¹.

The pathomechanism of the two diseases is different, but the effect of interspecies connections plays a fundamental role in inflammatory processes ^{516,517}. Dysregulation of the cellular behaviour in epithelial cells results in altered cytokine secretion and potential infiltration of bacteria into the lamina propria by disrupted cell-cell interactions. Both processes generalise inflammatory response resulting in high levels of pro-inflammatory cytokines ^{482,518,519}. In the early 1990s, researchers described that patients suffering from IBD have periodontitis with a higher prevalence ⁵²⁰. The co-occurrence of the disease is high, and the gum inflammation is more severe in IBD patients ⁵²¹. Periodontitis and IBD are multifactorial diseases, sharing factors involved in the pathogenesis (e.g. smoking, diet), microbiological impact, and immuno-

inflammatory response. Both diseases are characterised by a shift to Gram-negative bacteria in the microbiome ⁵²². Besides, a few microbes (*Campylobacter rectus*, *Porphyromonas gingivalis* and *Tannerella forsythia*) appear in inflamed gingiva and are enriched in IBD ^{523,524}. Several papers explore the connection between oral and gut inflammation ^{520,521,525–529}, but still, there are conflicting results, indicating a complex, personalised pathomechanism of diseases.

Case studies revealed the central role of the Toll-like receptor (TLR) pathway during inflammation from different perspectives. In fact, it has been known from the literature years ago, but the pipeline gave an insight into the cell and condition specificity of the pathway. In Chapter 2, I revealed that different parts of the signalling were enriched in Tregs under diverse conditions. In healthy colon the TLR2/6 and TLR7/8 signalling, while in UC patients the TLR3 and TLR4 receptors-mediated signalling were enriched. In Chapter 4, I modelled the TLR pathway in several immune cell (sub)populations under healthy and UC conditions in presence of gut commensal bacteria. This approach introduced the TLR signalling on the level of molecular interactions. I combined the signalling network with single-cell data to observe the expression of pathway members in various immune cell types and showed altered interactions with the bacteria. Finally, in Chapter 5, I examined the affected cellular pathways in epithelial cells and DCs in the gingiva. Results showed that the TLR pathway is affected by HMIs on the cell surface of epithelial cells during periodontitis. Not surprisingly, in DCs TLR pathway was triggered by epithelial ligands in both conditions, however diverse sets of receptors were activated.

The appearance of meta-omics and single-cell transcriptomic data allowed the implementation of multi-omics data analysis pipelines. To take advantage of the information about microbiome composition details, I could fill a gap in the current knowledge to better understand the pathomechanisms of bacterial communities on host cells and highlight differences in cell type levels. Hopefully, there will be more paired multi-omics data available soon, describing both meta- and host data from the same samples that can be analysed seamlessly and efficiently with the developed pipelines. Contributing to OmniPath established a new direction for my project and gave me a strong base for intercellular analysis. All in all, in my PhD projects, I predicted interactions and effects that got validated from existing literature, such as the central role of TLR pathway in inflammation. However, I also revealed cell-specific differences in inflammation, such as the healthy condition-related expression of TIRAP adaptor protein - a potential new therapeutic target for IBD treatment - in one of the DC subpopulations.

The prediction of protein-protein interactions has been improved in the last few years as more and more machine learning based approaches came to light. These approaches use computational models trained on large sets of known PPIs to predict new interactions. There are several categories of machine learning approaches used in PPI prediction, including deep learning methods that use neural networks to model the sequence, structure, or both of the interacting proteins (e.g. DWPPI tool ⁵³⁰). They have shown to be very effective models, achieving high accuracy and outperforming traditional methods ⁵³⁰. The field is constantly evolving and new methods are being developed and tested to improve the performance of PPI prediction.

AlphaFold2 is a protein structure prediction algorithm developed by the European Molecular Biology Laboratory and the University of Washington. It uses deep learning techniques to predict the 3D structure of a protein from its amino acid sequence. AlphaFold2 was announced in 2018 as a significant improvement over the original AlphaFold algorithm, achieving near-experimental accuracy in many cases ³⁹³. The algorithm has been used in a number of research studies and has the potential to aid in drug discovery and the design of new biomaterials ^{531–533}.

While the original aim of AlphaFold is to predict 3D protein structures, bacterial domains can also be inferred with the algorithm by uploading the bacterial protein sequence to the webservice (<https://alphafold.org/>). It's important to keep in mind that the accuracy of the prediction will depend on the specific input, and the quality of the prediction may vary for different bacterial domains. Therefore, it's recommended to validate the predictions using experimental methods, if available.

I have plans to improve MicrobioLink2 in the near future by extending the model with predicted bacterial domains coming from AlphaFold to increase the number of potential host-microbe PPIs. Besides, I would like to include the detection of cross-species interplay to bacterial metabolite-human protein interactions in MicrobioLink, and focusing more on the role of small molecules on host cell receptor activation. The microbiome is a dynamic community, bacteria secrete metabolites to 'communicate' with each other therefore facilitating co-occurrence or modulating competition between strains ⁵³⁴. A dysbiotic condition leads to altered microbiota composition, which has an effect on the robustness and connectivity of microbial interaction networks ⁵³⁵. Network fragility modelling reveals the association between microbes in a community based on meta-omics (metataxonomics, metagenomics) analysis ⁵³⁶. Fragility measures how easy it is to disrupt the network and how coherent is the connection between the bacterial species/strains ⁵³⁷. On the one hand, altered metabolic secretion - as an outcome

of dysbiotic communities - affects differently the human tissue/cells which are inhabited by the community ⁵³⁸. On the other hand, the perturbed host cell signalling may lead to altered metabolite secretion which, reflecting the changes, interacts with the microbiome ⁵³⁸. Based on these assumptions, there is a potential and exciting connection point between the bacteria-bacteria and host-microbe interactions that I would like to discover later.

Microbiome analysis became a hot research area recently, as researchers found that the disrupted community potentially leads to diseases. Recently, I had the opportunity to write a preview article to *Cell* about a very interesting article ⁵³⁹ examining the role of the skin microbiome in vector-borne disease transmission ⁵⁴⁰.

The main outputs of my PhD work are the following:

- Established workflows to analyse single-cell data and build up cell type- and condition-specific networks
- Making an impact on cell-cell connection analysis by the development of the intercellular interaction pipeline and contributing to OmniPath, a gap-filling resource to study the intercellular interplay
- Developed the MicrobioLink2 pipeline that examines host-microbe interactions from a new perspective including the downstream effect of complex microbiomes on host signalling
- The developed workflows have already been used within my research group for current and future projects.

During my PhD, I published the updated OmniPath and intercellular interaction pipeline in *Molecular Systems Biology* [Chapter 2] ¹⁰⁷. The case study in Chapter 4 appeared in the *Journal of Extracellular Vesicles* beginning of this year ⁴³⁷. Both of these journals are the premier journals in their respective fields. We were recently invited to submit the MicrobioLink2 pipeline to Cell Press's *STAR Protocol* journal.

The COVID-19 pandemic strongly affected my research between 2020 - 2022. Our research group established several side projects to study the effect of the virus on human signalling pathways. We developed the ViralLink pipeline published in *PLoS Computational Biology* ⁵⁴¹, studied the effect of Sars-CoV-2 on epithelial-immune cell interactions appeared in *npj Systems Biology and Applications* ⁵⁴², and cytokine expression in *Frontiers in Immunology* ¹⁴²,

and finally, we established a cytokine communication map, called CytokineLink published in *Cells*, to model cytokine-cytokine interactions ⁵⁴³. I contributed to these projects with the MicrobioLink2 and the intercellular interaction pipelines, therefore there is no separate chapter for these studies. I hope that these articles will reach other research communities and the pipelines will be used/improved by them as well.

I contributed to autophagy-related publications during the first two years of my postgraduate studies. I had the possibility to co-work on a review in *Frontiers in Cell and Developmental Biology* about available databases, and resources in the field of autophagy research ⁵⁴⁴. Later, I published as a joint-first author my Master's thesis in *the Disease Models and Mechanisms journal* about proteomic data analysis derived from organoids exposed to impaired autophagy compared to control systems ⁵⁴⁵. Also that year, we examined the effect of bacterial pathogens on the autophagy process and published it in the *Autophagy journal* ³⁷⁹. Finally, I was involved in the development of the SignaLink3 database published in the Database issue of *Nucleic Acids Research* ¹⁶⁵.

As an iCASE PhD candidate, I worked together with Unilever, the industrial collaborator of the PhD. They were interested in host-microbe interactions in healthy and inflamed gingiva and scalp. Due to the lack of public microbiome and host transcriptomics data from the scalp, I focused on data analysis in the gingiva and provided the MicrobioLink2 pipeline for internal commercial purposes at Unilever. Following a handover session during my placement, Unilever is capable of running the pipelines with their confidential data. This was a key objective in the original iCASE project agreement.

In conclusion, the thesis provides methodological and biological advancement in the field of cell biology and cellular microbiology. The developed pipelines give mechanistic insight into host-microbe interactions and their effect on epithelial and immune cell signalling, including the context of inflammation-related diseases. The case examples reveal the high connectivity of factors that have an effect on inflammation and an outstanding need for such computational analysis and *in silico* workflows. Due to a lack of experimental validations, my aim was not necessarily to highlight potential new signalling pathways in inflamed conditions. I aimed to explore the molecular background of the currently known implications of pathogen-associated inflammation and extend it to the individual cell level. The highlighted limitations identified the project's next steps and future directions. The improved pipeline should lead to a better understanding of homeostasis and drive the development of targeted approaches for preventing and treating dysbiosis-related disorders such as periodontitis and IBD.

References

1. Van Dyke, T. E., Bartold, P. M. & Reynolds, E. C. The nexus between periodontal inflammation and dysbiosis. *Front. Immunol.* **11**, 511 (2020).
2. Buttó, L. F. & Haller, D. Dysbiosis in intestinal inflammation: Cause or consequence. *Int. J. Med. Microbiol.* **306**, 302–309 (2016).
3. Manor, O. *et al.* Health and disease markers correlate with gut microbiome composition across thousands of people. *Nat. Commun.* **11**, 5206 (2020).
4. Jackson, M. A. *et al.* Gut microbiota associations with common diseases and prescription medications in a population-based cohort. *Nat. Commun.* **9**, 2655 (2018).
5. Bengmark, S. Ecological control of the gastrointestinal tract. The role of probiotic flora. *Gut* **42**, 2–7 (1998).
6. Kim, Y. & Pritts, T. A. The Gastrointestinal Tract. in *Geriatric trauma and critical care* (eds. Luchette, F. A. & Yelon, J. A.) 35–43 (Springer International Publishing, 2017). doi:10.1007/978-3-319-48687-1_5.
7. Mason, K. L., Huffnagle, G. B., Noverr, M. C. & Kao, J. Y. Overview of gut immunology. *Adv. Exp. Med. Biol.* **635**, 1–14 (2008).
8. PDQ Adult Treatment Editorial Board. Lip and Oral Cavity Cancer Treatment (PDQ®): Patient Version. in *PDQ cancer information summaries* (National Cancer Institute (US), 2002).
9. Calenic, B. *et al.* Magnetic separation and characterization of keratinocyte stem cells from human gingiva. *J. Periodont. Res.* **45**, 703–708 (2010).
10. Takahashi, N. *et al.* Gingival epithelial barrier: regulation by beneficial and harmful microbes. *Tissue Barriers* **7**, e1651158 (2019).
11. Jiang, Q., Yu, Y., Ruan, H., Luo, Y. & Guo, X. Morphological and functional characteristics of human gingival junctional epithelium. *BMC Oral Health* **14**, 30 (2014).
12. Schroeder, H. E. & Listgarten, M. A. The gingival tissues: the architecture of periodontal protection. *Periodontol. 2000* **13**, 91–120 (1997).
13. Luke, D. A. The structure and functions of the dentogingival junction and periodontal ligament. *Br. Dent. J.* **172**, 187–190 (1992).
14. Dabija-Wolter, G., Bakken, V., Cimpan, M. R., Johannessen, A. C. & Costea, D. E. In vitro reconstruction of human junctional and sulcular epithelium. *J. Oral Pathol. Med.* **42**, 396–404 (2013).
15. Groeger, S. & Meyle, J. Oral mucosal epithelial cells. *Front. Immunol.* **10**, 208 (2019).

16. Peterson, L. W. & Artis, D. Intestinal epithelial cells: regulators of barrier function and immune homeostasis. *Nat. Rev. Immunol.* **14**, 141–153 (2014).
17. Clevers, H. The intestinal crypt, a prototype stem cell compartment. *Cell* **154**, 274–284 (2013).
18. Abdul Khalek, F. J., Gallicano, G. I. & Mishra, L. Colon cancer stem cells. *Gastrointest. Cancer Res.* S16-23 (2010).
19. Gersemann, M., Stange, E. F. & Wehkamp, J. From intestinal stem cells to inflammatory bowel diseases. *World J. Gastroenterol.* **17**, 3198–3203 (2011).
20. Rangel-Huerta, E. & Maldonado, E. Transit-Amplifying Cells in the Fast Lane from Stem Cells towards Differentiation. *Stem Cells Int.* **2017**, 7602951 (2017).
21. Krausova, M. & Korinek, V. Wnt signaling in adult intestinal stem cells and cancer. *Cell. Signal.* **26**, 570–579 (2014).
22. Gersemann, M. *et al.* Differences in goblet cell differentiation between Crohn's disease and ulcerative colitis. *Differentiation.* **77**, 84–94 (2009).
23. Noah, T. K., Donahue, B. & Shroyer, N. F. Intestinal development and differentiation. *Exp. Cell Res.* **317**, 2702–2710 (2011).
24. Hou, Q., Ye, L., Huang, L. & Yu, Q. The Research Progress on Intestinal Stem Cells and Its Relationship with Intestinal Microbiota. *Front. Immunol.* **8**, 599 (2017).
25. Miron, N. & Cristea, V. Enterocytes: active cells in tolerance to food and microbial antigens in the gut. *Clin. Exp. Immunol.* **167**, 405–412 (2012).
26. Overeem, A. W., Posovszky, C., Rings, E. H. M. M., Giepmans, B. N. G. & van IJzendoorn, S. C. D. The role of enterocyte defects in the pathogenesis of congenital diarrheal disorders. *Dis. Model. Mech.* **9**, 1–12 (2016).
27. Shifrin, D. A. *et al.* Enterocyte microvillus-derived vesicles detoxify bacterial products and regulate epithelial-microbial interactions. *Curr. Biol.* **22**, 627–631 (2012).
28. Gassler, N. *et al.* Inflammatory bowel disease is associated with changes of enterocytic junctions. *Am. J. Physiol. Gastrointest. Liver Physiol.* **281**, G216-28 (2001).
29. Schneider, C., O'Leary, C. E. & Locksley, R. M. Regulation of immune responses by tuft cells. *Nat. Rev. Immunol.* **19**, 584–593 (2019).
30. Yi, J. *et al.* Dclk1 in tuft cells promotes inflammation-driven epithelial restitution and mitigates chronic colitis. *Cell Death Differ.* **26**, 1656–1669 (2019).
31. Steele, S. P., Melchor, S. J. & Petri, W. A. Tuft cells: new players in colitis. *Trends Mol. Med.* **22**, 921–924 (2016).

32. Gunawardene, A. R., Corfe, B. M. & Staton, C. A. Classification and functions of enteroendocrine cells of the lower gastrointestinal tract. *Int. J. Exp. Pathol.* **92**, 219–231 (2011).
33. Worthington, J. J., Reimann, F. & Gribble, F. M. Enteroendocrine cells-sensory sentinels of the intestinal environment and orchestrators of mucosal immunity. *Mucosal Immunol.* **11**, 3–20 (2018).
34. Moran, G. W., Pennock, J. & McLaughlin, J. T. Enteroendocrine cells in terminal ileal Crohn's disease. *J Crohns Colitis* **6**, 871–880 (2012).
35. Nowarski, R. *et al.* Epithelial IL-18 Equilibrium Controls Barrier Function in Colitis. *Cell* **163**, 1444–1456 (2015).
36. Knoop, K. A. & Newberry, R. D. Goblet cells: multifaceted players in immunity at mucosal surfaces. *Mucosal Immunol.* **11**, 1551–1557 (2018).
37. Ermund, A., Schütte, A., Johansson, M. E. V., Gustafsson, J. K. & Hansson, G. C. Studies of mucus in mouse stomach, small intestine, and colon. I. Gastrointestinal mucus layers have different properties depending on location as well as over the Peyer's patches. *Am. J. Physiol. Gastrointest. Liver Physiol.* **305**, G341-7 (2013).
38. Dillon, A. & Lo, D. D. M cells: intelligent engineering of mucosal immune surveillance. *Front. Immunol.* **10**, 1499 (2019).
39. Ohno, H. Intestinal M cells. *J. Biochem.* **159**, 151–160 (2016).
40. Bennett, K. M. *et al.* Induction of Colonic M Cells during Intestinal Inflammation. *Am. J. Pathol.* **186**, 1166–1179 (2016).
41. Wu, H.-J. & Wu, E. The role of gut microbiota in immune homeostasis and autoimmunity. *Gut Microbes* **3**, 4–14 (2012).
42. Fang, P. *et al.* Immune cell subset differentiation and tissue inflammation. *J. Hematol. Oncol.* **11**, 97 (2018).
43. Bain, C. C. & Schridde, A. Origin, differentiation, and function of intestinal macrophages. *Front. Immunol.* **9**, 2733 (2018).
44. Williams, M. *et al.* Dendritic cells, monocytes and macrophages: a unified nomenclature based on ontogeny. *Nat. Rev. Immunol.* **14**, 571–578 (2014).
45. Stagg, A. J. Intestinal dendritic cells in health and gut inflammation. *Front. Immunol.* **9**, 2883 (2018).
46. Sun, T., Nguyen, A. & Gomerman, J. L. Dendritic cell subsets in intestinal immunity and inflammation. *J. Immunol.* **204**, 1075–1083 (2020).
47. Wang, X. *et al.* Ulcerative colitis is characterized by a decrease in regulatory B cells. *J Crohns Colitis* **10**, 1212–1223 (2016).

48. Ma, H., Tao, W. & Zhu, S. T lymphocytes in the intestinal mucosa: defense and tolerance. *Cell. Mol. Immunol.* **16**, 216–224 (2019).
49. Rudd-Schmidt, J. A. *et al.* Lipid order and charge protect killer T cells from accidental death. *Nat. Commun.* **10**, 5396 (2019).
50. Giuffrida, P. & Di Sabatino, A. Targeting T cells in inflammatory bowel disease. *Pharmacol. Res.* **159**, 105040 (2020).
51. Schreiber, S. *et al.* Increased activation of isolated intestinal lamina propria mononuclear cells in inflammatory bowel disease. *Gastroenterology* **101**, 1020–1030 (1991).
52. Luci, C., Vieira, E., Perchet, T., Gual, P. & Golub, R. Natural Killer Cells and Type 1 Innate Lymphoid Cells Are New Actors in Non-alcoholic Fatty Liver Disease. *Front. Immunol.* **10**, 1192 (2019).
53. Panda, S. K. & Colonna, M. Innate lymphoid cells in mucosal immunity. *Front. Immunol.* **10**, 861 (2019).
54. Abel, A. M., Yang, C., Thakar, M. S. & Malarkannan, S. Natural killer cells: development, maturation, and clinical utilization. *Front. Immunol.* **9**, 1869 (2018).
55. Poggi, A. *et al.* Human Gut-Associated Natural Killer Cells in Health and Disease. *Front. Immunol.* **10**, 961 (2019).
56. Geremia, A. & Arancibia-Cárcamo, C. V. Innate lymphoid cells in intestinal inflammation. *Front. Immunol.* **8**, 1296 (2017).
57. Owens, B. M. J. & Simmons, A. Intestinal stromal cells in mucosal immunity and homeostasis. *Mucosal Immunol.* **6**, 224–234 (2013).
58. Bajaj-Elliott, M., Breese, E., Poulsom, R., Fairclough, P. D. & MacDonald, T. T. Keratinocyte growth factor in inflammatory bowel disease. Increased mRNA transcripts in ulcerative colitis compared with Crohn's disease in biopsies and isolated mucosal myofibroblasts. *Am. J. Pathol.* **151**, 1469–1476 (1997).
59. Roulis, M. & Flavell, R. A. Fibroblasts and myofibroblasts of the intestinal lamina propria in physiology and disease. *Differentiation.* **92**, 116–131 (2016).
60. Barnhoorn, M. C. *et al.* Stromal cells in the pathogenesis of inflammatory bowel disease. *J Crohns Colitis* **14**, 995–1009 (2020).
61. Mifflin, R. C., Pinchuk, I. V., Saada, J. I. & Powell, D. W. Intestinal myofibroblasts: targets for stem cell therapy. *Am. J. Physiol. Gastrointest. Liver Physiol.* **300**, G684-96 (2011).
62. Li, C. & Kuemmerle, J. F. The fate of myofibroblasts during the development of fibrosis in Crohn's disease. *J. Dig. Dis.* **21**, 326–331 (2020).

63. Cheng, L. *et al.* gutMGene: a comprehensive database for target genes of gut microbes and microbial metabolites. *Nucleic Acids Res.* **50**, D795–D800 (2022).
64. Turnbaugh, P. J. *et al.* The human microbiome project. *Nature* **449**, 804–810 (2007).
65. Zhu, B., Wang, X. & Li, L. Human gut microbiome: the second genome of human body. *Protein Cell* **1**, 718–725 (2010).
66. Gupta, V. K., Paul, S. & Dutta, C. Geography, Ethnicity or Subsistence-Specific Variations in Human Microbiome Composition and Diversity. *Front. Microbiol.* **8**, 1162 (2017).
67. Idris, A., Hasnain, S. Z., Huat, L. Z. & Koh, D. Human diseases, immunity and the oral microbiota—Insights gained from metagenomic studies. *Oral Science International* **14**, 27–32 (2017).
68. Deo, P. N. & Deshmukh, R. Oral microbiome: Unveiling the fundamentals. *J. Oral Maxillofac. Pathol.* **23**, 122–128 (2019).
69. Byrd, A. L., Belkaid, Y. & Segre, J. A. The human skin microbiome. *Nat. Rev. Microbiol.* **16**, 143–155 (2018).
70. Eisenstein, M. The skin microbiome. *Nature* **588**, S209 (2020).
71. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
72. Saxena, R. *et al.* Comparison of healthy and dandruff scalp microbiome reveals the role of commensals in scalp health. *Front. Cell. Infect. Microbiol.* **8**, 346 (2018).
73. Saxena, R. *et al.* Longitudinal study of the scalp microbiome suggests coconut oil to enrich healthy scalp commensals. *Sci. Rep.* **11**, 7220 (2021).
74. Quigley, E. M. M. Gut bacteria in health and disease. *Gastroenterol Hepatol (N Y)* **9**, 560–569 (2013).
75. Poyet, M. *et al.* A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nat. Med.* **25**, 1442–1452 (2019).
76. Davenport, E. R. *et al.* Seasonal variation in human gut microbiome composition. *PLoS ONE* **9**, e90731 (2014).
77. Chen, J. *et al.* Altered gut microbiota correlated with systemic inflammation in children with Kawasaki disease. *Sci. Rep.* **10**, 14525 (2020).
78. Chen, T. *et al.* The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database (Oxford)* **2010**, baq013 (2010).

79. Rupf, S. *et al.* Comparison of initial oral microbiomes of young adults with and without cavitated dentin caries lesions using an in situ biofilm model. *Sci. Rep.* **8**, 14010 (2018).
80. Dewhirst, F. E. *et al.* The human oral microbiome. *J. Bacteriol.* **192**, 5002–5017 (2010).
81. Alexeev, E. E. *et al.* Microbiota-Derived Indole Metabolites Promote Human and Murine Intestinal Homeostasis through Regulation of Interleukin-10 Receptor. *Am. J. Pathol.* **188**, 1183–1194 (2018).
82. Struzycka, I. The oral microbiome in dental caries. *Pol. J. Microbiol.* **63**, 127–135 (2014).
83. DeGruttola, A. K., Low, D., Mizoguchi, A. & Mizoguchi, E. Current understanding of dysbiosis in disease in human and animal models. *Inflamm. Bowel Dis.* **22**, 1137–1150 (2016).
84. Zeng, M. Y., Inohara, N. & Nuñez, G. Mechanisms of inflammation-driven bacterial dysbiosis in the gut. *Mucosal Immunol.* **10**, 18–26 (2017).
85. Hemarajata, P. & Versalovic, J. Effects of probiotics on gut microbiota: mechanisms of intestinal immunomodulation and neuromodulation. *Therap. Adv. Gastroenterol.* **6**, 39–51 (2013).
86. Gong, D., Gong, X., Wang, L., Yu, X. & Dong, Q. Involvement of reduced microbial diversity in inflammatory bowel disease. *Gastroenterol. Res. Pract.* **2016**, 6951091 (2016).
87. Vandeputte, D. *et al.* Temporal variability in quantitative human gut microbiome profiles and implications for clinical research. *Nat. Commun.* **12**, 6740 (2021).
88. Ma, Z. S. Testing the Anna Karenina Principle in Human Microbiome-Associated Diseases. *iScience* **23**, 101007 (2020).
89. Holmes, I., Harris, K. & Quince, C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS ONE* **7**, e30126 (2012).
90. Rigottier-Gois, L. Dysbiosis in inflammatory bowel diseases: the oxygen hypothesis. *ISME J.* **7**, 1256–1261 (2013).
91. Kho, Z. Y. & Lal, S. K. The Human Gut Microbiome - A Potential Controller of Wellness and Disease. *Front. Microbiol.* **9**, 1835 (2018).
92. Gagliardi, A. *et al.* Rebuilding the gut microbiota ecosystem. *Int. J. Environ. Res. Public Health* **15**, (2018).
93. Magne, F. *et al.* The firmicutes/bacteroidetes ratio: A relevant marker of gut dysbiosis in obese patients? *Nutrients* **12**, (2020).

94. Lobionda, S., Sittipo, P., Kwon, H. Y. & Lee, Y. K. The Role of Gut Microbiota in Intestinal Inflammation with Respect to Diet and Extrinsic Stressors. *Microorganisms* **7**, (2019).
95. Kashef, N., Huang, Y.-Y. & Hamblin, M. R. Advances in antimicrobial photodynamic inactivation at the nanoscale. *Nanophotonics* **6**, 853–879 (2017).
96. Beveridge, T. J. Structures of gram-negative cell walls and their derived membrane vesicles. *J. Bacteriol.* **181**, 4725–4733 (1999).
97. Brown, L., Wolf, J. M., Prados-Rosales, R. & Casadevall, A. Through the wall: extracellular vesicles in Gram-positive bacteria, mycobacteria and fungi. *Nat. Rev. Microbiol.* **13**, 620–630 (2015).
98. Park, S.-Y. *et al.* Oral-Gut Microbiome Axis in Gastrointestinal Disease and Cancer. *Cancers (Basel)* **13**, (2021).
99. Bhat, A. A. *et al.* Tight junction proteins and signaling pathways in cancer and inflammation: A functional crosstalk. *Front. Physiol.* **9**, 1942 (2018).
100. Groschwitz, K. R. & Hogan, S. P. Intestinal barrier function: molecular regulation and disease pathogenesis. *J. Allergy Clin. Immunol.* **124**, 3–20; quiz 21 (2009).
101. Yang, B. A., Westerhof, T. M., Sabin, K., Merajver, S. D. & Aguilar, C. A. Engineered tools to study intercellular communication. *Adv Sci (Weinh)* **8**, 2002825 (2021).
102. Singer, S. J. Intercellular communication and cell-cell adhesion. *Science* **255**, 1671–1677 (1992).
103. Zhou, H., Fan, E. K. & Fan, J. Cell-Cell Interaction Mechanisms in Acute Lung Injury. *Shock* **55**, 167–176 (2021).
104. Mukherjee, S. *et al.* Abstract 3119: LRRC15, a candidate immunotherapy target, regulates cell-cell interaction, migration and spheroid formation in osteosarcoma cells. in *Tumor Biology* 3119–3119 (American Association for Cancer Research, 2021). doi:10.1158/1538-7445.AM2021-3119.
105. Andrews, N. *et al.* An unsupervised method for physical cell interaction profiling of complex tissues. *Nat. Methods* **18**, 912–920 (2021).
106. Hui, K. P.-Y. *et al.* Role of epithelial-endothelial cell interaction in the pathogenesis of SARS-CoV-2 infection. *Clin. Infect. Dis.* (2021) doi:10.1093/cid/ciab406.
107. Türei, D. *et al.* Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Mol. Syst. Biol.* **17**, (2021).
108. Hou, R., Denisenko, E., Ong, H. T., Ramiłowski, J. A. & Forrest, A. R. R. Predicting cell-to-cell communication networks using NATMI. *Nat. Commun.* **11**, 5011 (2020).

109. Dimitrov, D. *et al.* Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. *Nat. Commun.* **13**, 3224 (2022).
110. Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods* **17**, 159–162 (2020).
111. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).
112. Cheng, J., Zhang, J., Wu, Z. & Sun, X. Inferring microenvironmental regulation of gene expression from single-cell RNA sequencing data using scMLnet with an application to COVID-19. *Brief. Bioinformatics* **22**, 988–1005 (2021).
113. Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* **28**, 573–580 (2012).
114. Armingol, E., Officer, A., Harismendy, O. & Lewis, N. E. Deciphering cell–cell interactions and communication from gene expression. *Nat. Rev. Genet.* (2020) doi:10.1038/s41576-020-00292-x.
115. Harris, T. J. C. & Tepass, U. Adherens junctions: from molecules to morphogenesis. *Nat. Rev. Mol. Cell Biol.* **11**, 502–514 (2010).
116. Mehta, S., Nijhuis, A., Kumagai, T., Lindsay, J. & Silver, A. Defects in the adherens junction complex (E-cadherin/ β -catenin) in inflammatory bowel disease. *Cell Tissue Res.* **360**, 749–760 (2015).
117. Al-Ghadban, S., Kaissi, S., Homaidan, F. R., Naim, H. Y. & El-Sabban, M. E. Cross-talk between intestinal epithelial cells and immune cells in inflammatory bowel disease. *Sci. Rep.* **6**, 29783 (2016).
118. Delva, E., Tucker, D. K. & Kowalczyk, A. P. The desmosome. *Cold Spring Harb. Perspect. Biol.* **1**, a002543 (2009).
119. Schlegel, N., Boerner, K. & Waschke, J. Targeting desmosomal adhesion and signalling for intestinal barrier stabilization in inflammatory bowel diseases-Lessons from experimental models and patients. *Acta Physiol (Oxf)* **231**, e13492 (2021).
120. Guryanov, I., Fiorucci, S. & Tenukova, T. Receptor-ligand interactions: Advanced biomedical applications. *Mater. Sci. Eng. C Mater. Biol. Appl.* **68**, 890–903 (2016).
121. Sklar, L. A., Sayre, J., McNeil, V. M. & Finney, D. A. Competitive binding kinetics in ligand-receptor-competitor systems. Rate parameters for unlabeled ligands for the formyl peptide receptor. *Mol. Pharmacol.* **28**, 323–330 (1985).
122. Annis, D. A., Nazef, N., Chuang, C.-C., Scott, M. P. & Nash, H. M. A general technique to rank protein-ligand binding affinities and determine allosteric versus direct binding site competition in compound mixtures. *J. Am. Chem. Soc.* **126**, 15495–15503 (2004).

123. Ramilowski, J. A. *et al.* A draft network of ligand-receptor-mediated multicellular signalling in human. *Nat. Commun.* **6**, 7866 (2015).
124. Coudert, J. D. *et al.* Altered NKG2D function in NK cells induced by chronic exposure to NKG2D ligand-expressing tumor cells. *Blood* **106**, 1711–1717 (2005).
125. Song, Z., Yu, X. & Zhang, Y. Altered expression of programmed death-ligand 1 after neo-adjuvant chemotherapy in patients with lung squamous cell carcinoma. *Lung Cancer* **99**, 166–171 (2016).
126. Mikolaj, C. *et al.* Early host-microbe interaction in a peri-implant oral mucosa-biofilm model. *Cell. Microbiol.* **22**, e13209 (2020).
127. Guven-Maiorov, E. *et al.* HMI-PRED: A Web Server for Structural Prediction of Host-Microbe Interactions Based on Interface Mimicry. *J. Mol. Biol.* **432**, 3395–3403 (2020).
128. Franzosa, E. A. & Xia, Y. Structural principles within the human-virus protein-protein interaction network. *Proc Natl Acad Sci USA* **108**, 10538–10543 (2011).
129. Guven-Maiorov, E., Tsai, C.-J. & Nussinov, R. Pathogen mimicry of host protein-protein interfaces modulates immunity. *Semin. Cell Dev. Biol.* **58**, 136–145 (2016).
130. Plüddemann, A., Mukhopadhyay, S. & Gordon, S. The interaction of macrophage receptors with bacterial ligands. *Expert Rev. Mol. Med.* **8**, 1–25 (2006).
131. Encarnação, J. C., Schulte, T., Achour, A., Björkelund, H. & Andersson, K. Detecting ligand interactions in real time on living bacterial cells. *Appl. Microbiol. Biotechnol.* **102**, 4193–4201 (2018).
132. Raffa, R. B. *et al.* Bacterial communication (“quorum sensing”) via ligands and receptors: a novel pharmacologic target for the design of antibiotic drugs. *J. Pharmacol. Exp. Ther.* **312**, 417–423 (2005).
133. Han, Z. *et al.* Structure-based drug design and optimization of mannoside bacterial FimH antagonists. *J. Med. Chem.* **53**, 4779–4792 (2010).
134. Tribble, G. D. & Lamont, R. J. Bacterial invasion of epithelial cells and spreading in periodontal tissue. *Periodontol. 2000* **52**, 68–83 (2010).
135. Brown, E. M., Sadarangani, M. & Finlay, B. B. The role of the immune system in governing host-microbe interactions in the intestine. *Nat. Immunol.* **14**, 660–667 (2013).
136. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
137. Kline, K. A., Fälker, S., Dahlberg, S., Normark, S. & Henriques-Normark, B. Bacterial adhesins in host-microbe interactions. *Cell Host Microbe* **5**, 580–592 (2009).

138. Gaffen, S. L. & Moutsopoulos, N. M. Regulation of host-microbe interactions at oral mucosal barriers by type 17 immunity. *Sci. Immunol.* **5**, (2020).
139. Negrini, T. de C., Koo, H. & Arthur, R. A. Candida-Bacterial Biofilms and Host-Microbe Interactions in Oral Diseases. *Adv. Exp. Med. Biol.* **1197**, 119–141 (2019).
140. Chen, L. *et al.* Inflammatory responses and inflammation-associated diseases in organs. *Oncotarget* **9**, 7204–7218 (2018).
141. Kany, S., Vollrath, J. T. & Relja, B. Cytokines in inflammatory disease. *Int. J. Mol. Sci.* **20**, (2019).
142. Olbei, M. *et al.* SARS-CoV-2 Causes a Different Cytokine Response Compared to Other Cytokine Storm-Causing Respiratory Viruses in Severely Ill Patients. *Front. Immunol.* **12**, 629193 (2021).
143. Zhang, J.-M. & An, J. Cytokines, inflammation, and pain. *Int. Anesthesiol. Clin.* **45**, 27–37 (2007).
144. Scheller, J., Chalaris, A., Schmidt-Arras, D. & Rose-John, S. The pro- and anti-inflammatory properties of the cytokine interleukin-6. *Biochim. Biophys. Acta* **1813**, 878–888 (2011).
145. Altan-Bonnet, G. & Mukherjee, R. Cytokine-mediated communication: a quantitative appraisal of immune complexity. *Nat. Rev. Immunol.* **19**, 205–217 (2019).
146. Diesch, T., Filippi, C., Fritschi, N., Filippi, A. & Ritz, N. Cytokines in saliva as biomarkers of oral and systemic oncological or infectious diseases: A systematic review. *Cytokine* **143**, 155506 (2021).
147. Suárez, L. J., Garzón, H., Arboleda, S. & Rodríguez, A. Oral dysbiosis and autoimmunity: from local periodontal responses to an imbalanced systemic immunity. A review. *Front. Immunol.* **11**, 591255 (2020).
148. Ben-Sasson, S. Z. *et al.* IL-1 acts directly on CD4 T cells to enhance their antigen-driven expansion and differentiation. *Proc Natl Acad Sci USA* **106**, 7119–7124 (2009).
149. Lapérine, O. *et al.* Interleukin-33 and RANK-L Interplay in the Alveolar Bone Loss Associated to Periodontitis. *PLoS ONE* **11**, e0168080 (2016).
150. Santos-Lima, E. K. N. *et al.* Production of interferon-gamma, interleukin-6, and interleukin-1 β by human peripheral blood mononuclear cells stimulated with novel lysingipain synthetic peptides. *J. Periodontol.* **90**, 993–1001 (2019).
151. Shaker, O. G. & Ghallab, N. A. IL-17 and IL-11 GCF levels in aggressive and chronic periodontitis patients: relation to PCR bacterial detection. *Mediators Inflamm.* **2012**, 174764 (2012).
152. Mogensen, T. H. Pathogen recognition and inflammatory signaling in innate immune defenses. *Clin. Microbiol. Rev.* **22**, 240–73, Table of Contents (2009).

153. Walsh, D., McCarthy, J., O'Driscoll, C. & Melgar, S. Pattern recognition receptors--molecular orchestrators of inflammation in inflammatory bowel disease. *Cytokine Growth Factor Rev.* **24**, 91–104 (2013).
154. Yamamoto, M. & Takeda, K. Current views of toll-like receptor signaling pathways. *Gastroenterol. Res. Pract.* **2010**, 240365 (2010).
155. Kim, M. *et al.* Bacterial interactions with the host epithelium. *Cell Host Microbe* **8**, 20–35 (2010).
156. Konradt, C. & Hunter, C. A. Pathogen interactions with endothelial cells and the induction of innate and adaptive immunity. *Eur. J. Immunol.* **48**, 1607–1620 (2018).
157. Moser, M. & Leo, O. Key concepts in immunology. *Vaccine* **28 Suppl 3**, C2-13 (2010).
158. Andersen, M. H., Schrama, D., Thor Straten, P. & Becker, J. C. Cytotoxic T cells. *J. Invest. Dermatol.* **126**, 32–41 (2006).
159. Fukata, M. & Abreu, M. T. Pathogen recognition receptors, cancer and inflammation in the gut. *Curr. Opin. Pharmacol.* **9**, 680–687 (2009).
160. Fucikova, J., Palova-Jelinkova, L., Bartunkova, J. & Spisek, R. Induction of tolerance and immunity by dendritic cells: mechanisms and clinical applications. *Front. Immunol.* **10**, 2393 (2019).
161. Kayisoglu, Ö., Schlegel, N. & Bartfeld, S. Gastrointestinal epithelial innate immunity-regionalization and organoids as new model. *J. Mol. Med.* **99**, 517–530 (2021).
162. Souza-Moreira, L., Campos-Salinas, J., Caro, M. & Gonzalez-Rey, E. Neuropeptides as pleiotropic modulators of the immune response. *Neuroendocrinology* **94**, 89–100 (2011).
163. Gareb, B., Otten, A. T., Frijlink, H. W., Dijkstra, G. & Kosterink, J. G. W. Review: Local Tumor Necrosis Factor- α Inhibition in Inflammatory Bowel Disease. *Pharmaceutics* **12**, (2020).
164. Mahapatro, M., Erkert, L. & Becker, C. Cytokine-Mediated Crosstalk between Immune Cells and Epithelial Cells in the Gut. *Cells* **10**, (2021).
165. Csabai, L. *et al.* Signalink3: a multi-layered resource to uncover tissue-specific signaling networks. *Nucleic Acids Res.* **50**, D701–D709 (2022).
166. O'Neill, L. A. J. & Bowie, A. G. The family of five: TIR-domain-containing adaptors in Toll-like receptor signalling. *Nat. Rev. Immunol.* **7**, 353–364 (2007).
167. Kawasaki, T. & Kawai, T. Toll-like receptor signaling pathways. *Front. Immunol.* **5**, 461 (2014).

168. Török, H. P. *et al.* Functional Toll-Like Receptor (TLR)2 polymorphisms in the susceptibility to inflammatory bowel disease. *PLoS ONE* **12**, e0175180 (2017).
169. Kordjazy, N. *et al.* Role of toll-like receptors in inflammatory bowel disease. *Pharmacol. Res.* **129**, 204–215 (2018).
170. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **18**, 83 (2017).
171. Conesa, A. & Beck, S. Making multi-omics data accessible to researchers. *Sci. Data* **6**, 251 (2019).
172. Huang, S. *et al.* Longitudinal Multi-omics and Microbiome Meta-analysis Identify an Asymptomatic Gingival State That Links Gingivitis, Periodontitis, and Aging. *MBio* **12**, (2021).
173. Imhann, F. *et al.* The 1000IBD project: multi-omics data of 1000 inflammatory bowel disease patients; data release 1. *BMC Gastroenterol.* **19**, 5 (2019).
174. Vich Vila, A. *et al.* Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. *Sci. Transl. Med.* **10**, (2018).
175. Franzosa, E. A. *et al.* Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.* **4**, 293–305 (2019).
176. Uniken Venema, W. T. *et al.* Single-Cell RNA Sequencing of Blood and Ileal T Cells From Patients With Crohn's Disease Reveals Tissue-Specific Characteristics and Drug Targets. *Gastroenterology* **156**, 812-815.e22 (2019).
177. Abu-Ali, G. S. *et al.* Metatranscriptome of human faecal microbial communities in a cohort of adult men. *Nat. Microbiol.* **3**, 356–366 (2018).
178. Sink, R., Gobec, S., Pečar, S. & Zega, A. False positives in the early stages of drug discovery. *Curr. Med. Chem.* **17**, 4231–4255 (2010).
179. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).
180. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
181. Zhou, G., Ewald, J. & Xia, J. OmicsAnalyst: a comprehensive web-based platform for visual analytics of multi-omics data. *Nucleic Acids Res.* **49**, W476–W482 (2021).
182. Bah, S. Y., Morang'a, C. M., Kengne-Ouafo, J. A., Amenga-Etego, L. & Awandare, G. A. Highlights on the application of genomics and bioinformatics in the fight against infectious diseases: challenges and opportunities in africa. *Front. Genet.* **9**, 575 (2018).

183. Salzberg, S. L. Open questions: How many genes do we have? *BMC Biol.* **16**, 94 (2018).
184. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
185. Church, D. M. A next-generation human genome sequence. *Science* **376**, 34–35 (2022).
186. Rao, M. S. *et al.* Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies. *Front. Genet.* **9**, 636 (2018).
187. 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).
188. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
189. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2018).
190. Melit Devassy, B. & George, S. Dimensionality reduction and visualisation of hyperspectral ink data using t-SNE. *Forensic Sci. Int.* **311**, 110194 (2020).
191. Tenenbaum, J. B., de Silva, V. & Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–2323 (2000).
192. Coifman, R. R. *et al.* Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc Natl Acad Sci USA* **102**, 7426–7431 (2005).
193. Caetano, A. J. *et al.* Defining human mesenchymal and epithelial heterogeneity in response to oral inflammatory disease. *eLife* **10**, (2021).
194. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* **9**, 75 (2017).
195. Li, Y., Ma, L., Wu, D. & Chen, G. Advances in bulk and single-cell multi-omics approaches for systems biology and precision medicine. *Brief. Bioinformatics* **22**, (2021).
196. Graves, P. R. & Haystead, T. A. J. Molecular biologist's guide to proteomics. *Microbiol. Mol. Biol. Rev.* **66**, 39–63; table of contents (2002).
197. Al-Amrani, S., Al-Jabri, Z., Al-Zaabi, A., Alshekaili, J. & Al-Khabori, M. Proteomics: Concepts and applications in human medicine. *World J. Biol. Chem.* **12**, 57–69 (2021).

198. Dupree, E. J. *et al.* A Critical Review of Bottom-Up Proteomics: The Good, the Bad, and the Future of this Field. *Proteomes* **8**, (2020).
199. Xing, S., Wallmeroth, N., Berendzen, K. W. & Grefen, C. Techniques for the Analysis of Protein-Protein Interactions in Vivo. *Plant Physiol.* **171**, 727–758 (2016).
200. Sarkar, D. & Saha, S. Machine-learning techniques for the prediction of protein-protein interactions. *J. Biosci.* **44**, (2019).
201. Schiebenhoefer, H. *et al.* Challenges and promise at the interface of metaproteomics and genomics: an overview of recent progress in metaproteogenomic data analysis. *Expert Rev. Proteomics* **16**, 375–390 (2019).
202. Siggins, A., Gunnigle, E. & Abram, F. Exploring mixed microbial community functioning: recent advances in metaproteomics. *FEMS Microbiol. Ecol.* **80**, 265–280 (2012).
203. Toh, M. C. & Allen-Vercoe, E. The human gut microbiota with reference to autism spectrum disorder: considering the whole as more than a sum of its parts. *Microb. Ecol. Health Dis.* **26**, 26309 (2015).
204. Bonnet, M., Lagier, J. C., Raoult, D. & Khelaifia, S. Bacterial culture through selective and non-selective conditions: the evolution of culture media in clinical microbiology. *New Microbes New Infect.* **34**, 100622 (2020).
205. Wade, W. Unculturable bacteria--the uncharacterized organisms that cause oral infections. *J. R. Soc. Med.* **95**, 81–83 (2002).
206. Cho, I. & Blaser, M. J. The human microbiome: at the interface of health and disease. *Nat. Rev. Genet.* **13**, 260–270 (2012).
207. Nadkarni, M. A., Martin, F. E., Hunter, N. & Jacques, N. A. Methods for optimizing DNA extraction before quantifying oral bacterial numbers by real-time PCR. *FEMS Microbiol. Lett.* **296**, 45–51 (2009).
208. Xu, H. S. *et al.* Survival and viability of nonculturable *Escherichia coli* and *Vibrio cholerae* in the estuarine and marine environment. *Microb. Ecol.* **8**, 313–323 (1982).
209. Pace, N. R., Stahl, D. A., Lane, D. J. & Olsen, G. J. The analysis of natural microbial populations by ribosomal RNA sequences. in *Advances in microbial ecology* (ed. Marshall, K. C.) vol. 9 1–55 (Springer US, 1986).
210. Schmidt, T. M., DeLong, E. F. & Pace, N. R. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J. Bacteriol.* **173**, 4371–4378 (1991).
211. Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* **5**, R245-9 (1998).

212. Marchesi, J. R. & Ravel, J. The vocabulary of microbiome research: a proposal. *Microbiome* **3**, 31 (2015).
213. Zhang, X., Li, L., Butcher, J., Stintzi, A. & Figeys, D. Advancing functional and translational microbiome research using meta-omics approaches. *Microbiome* **7**, 154 (2019).
214. Lim, J. W., Park, T., Tong, Y. W. & Yu, Z. The microbiome driving anaerobic digestion and microbial analysis. in vol. 5 1–61 (Elsevier, 2020).
215. Xu, P. & Gunsolley, J. Application of metagenomics in understanding oral health and disease. *Virulence* **5**, 424–432 (2014).
216. Zhang, R.-Y. *et al.* Design of targeted primers based on 16S rRNA sequences in meta-transcriptomic datasets and identification of a novel taxonomic group in the Asgard archaea. *BMC Microbiol.* **20**, 25 (2020).
217. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).
218. Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**, 7537–7541 (2009).
219. Hilton, S. K. *et al.* Metataxonomic and Metagenomic Approaches vs. Culture-Based Techniques for Clinical Pathology. *Front. Microbiol.* **7**, 484 (2016).
220. Yilmaz, P. *et al.* The SILVA and “All-species Living Tree Project (LTP)” taxonomic frameworks. *Nucleic Acids Res.* **42**, D643-8 (2014).
221. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069–5072 (2006).
222. Jeske, J. T. & Gallert, C. Microbiome Analysis via OTU and ASV-Based Pipelines-A Comparative Interpretation of Ecological Data in WWTP Systems. *Bioengineering (Basel)* **9**, (2022).
223. Douglas, G. M. *et al.* PICRUSt2: An improved and extensible approach for metagenome inference. *BioRxiv* (2019) doi:10.1101/672295.
224. Aßhauer, K. P., Wemheuer, B., Daniel, R. & Meinicke, P. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics* **31**, 2882–2884 (2015).
225. Breitwieser, F. P., Lu, J. & Salzberg, S. L. A review of methods and databases for metagenomic classification and assembly. *Brief. Bioinformatics* **20**, 1125–1136 (2019).
226. Ladoukakis, E., Kolisis, F. N. & Chatziioannou, A. A. Integrative workflows for metagenomic analysis. *Front. Cell Dev. Biol.* **2**, 70 (2014).

227. Hsiao, W. W. *et al.* Microbial transformation from normal oral microbiota to acute endodontic infections. *BMC Genomics* **13**, 345 (2012).
228. McLean, J. S. *et al.* Identifying low pH active and lactate-utilizing taxa within oral microbiome communities from healthy children using stable isotope probing techniques. *PLoS ONE* **7**, e32219 (2012).
229. Ge, X., Rodriguez, R., Trinh, M., Gunsolley, J. & Xu, P. Oral microbiome of deep and shallow dental pockets in chronic periodontitis. *PLoS ONE* **8**, e65520 (2013).
230. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).
231. Shakya, M., Lo, C.-C. & Chain, P. S. G. Advances and challenges in metatranscriptomic analysis. *Front. Genet.* **10**, 904 (2019).
232. Marcelino, V. R. *et al.* Metatranscriptomics as a tool to identify fungal species and subspecies in mixed communities - a proof of concept under laboratory conditions. *IMA Fungus* **10**, 12 (2019).
233. Rodríguez-Valera, F. Environmental genomics, the big picture? *FEMS Microbiol. Lett.* **231**, 153–158 (2004).
234. Heyer, R. *et al.* Challenges and perspectives of metaproteomic data analysis. *J. Biotechnol.* **261**, 24–36 (2017).
235. Muth, T., Renard, B. Y. & Martens, L. Metaproteomic data analysis at a glance: advances in computational microbial community proteomics. *Expert Rev. Proteomics* **13**, 757–769 (2016).
236. Kint, G., Fierro, C., Marchal, K., Vanderleyden, J. & De Keersmaecker, S. C. J. Integration of “omics” data: does it lead to new insights into host-microbe interactions? *Future Microbiol.* **5**, 313–328 (2010).
237. Eisenstein, M. Microbial matters: modelling the complex web of host-microbiome interactions. *Nature* **581**, 489–490 (2020).
238. Karahalil, B. Overview of systems biology and omics technologies. *Curr. Med. Chem.* **23**, 4221–4230 (2016).
239. Bhan, A. & Mjolsness, E. Static and dynamic models of biological networks. *Complexity* **11**, 57–63 (2006).
240. Iñiguez, G., Battiston, F. & Karsai, M. Bridging the gap between graphs and networks. *Commun. Phys.* **3**, 88 (2020).
241. Zhang, B., Tian, Y. & Zhang, Z. Network biology in medicine and beyond. *Circ. Cardiovasc. Genet.* **7**, 536–547 (2014).

242. Withall, M., Phillips, I. & Parish, D. IET Digital Library: Network visualisation: a review. *IET Communications* (2007).
243. Koutrouli, M., Karatzas, E., Paez-Espino, D. & Pavlopoulos, G. A. A guide to conquer the biological network era using graph theory. *Front. Bioeng. Biotechnol.* **8**, 34 (2020).
244. Csermely, P., Korcsmáros, T., Kiss, H. J. M., London, G. & Nussinov, R. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol. Ther.* **138**, 333–408 (2013).
245. Vu, K. V. *et al.* Systematic deletion of the ER lectin chaperone genes reveals their roles in vegetative growth and male gametophyte development in Arabidopsis. *Plant J.* **89**, 972–983 (2017).
246. Calloni, G. *et al.* DnaK functions as a central hub in the E. coli chaperone network. *Cell Rep.* **1**, 251–264 (2012).
247. Ren, Y., Ay, A. & Kahveci, T. Shortest path counting in probabilistic biological networks. *BMC Bioinformatics* **19**, 465 (2018).
248. Di Nanni, N., Bersanelli, M., Milanesi, L. & Mosca, E. Network diffusion promotes the integrative analysis of multiple omics. *Front. Genet.* **11**, 106 (2020).
249. Tong, H., Faloutsos, C. & Pan, J.-Y. Random walk with restart: fast solutions and applications. *Knowl. Inf. Syst.* **14**, 327–346 (2008).
250. Paull, E. O. *et al.* Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics* **29**, 2757–2764 (2013).
251. Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat. Methods* **10**, 1108–1115 (2013).
252. Zhang, W., Ma, J. & Ideker, T. Classifying tumors by supervised network propagation. *Bioinformatics* **34**, i484–i493 (2018).
253. Di Nanni, N., Gnocchi, M., Moscatelli, M., Milanesi, L. & Mosca, E. Gene relevance based on multiple evidences in complex networks. *Bioinformatics* **36**, 865–871 (2020).
254. Blomsma, N. *et al.* Minimum spanning tree analysis of brain networks: A systematic review of network size effects, sensitivity for neuropsychiatric pathology, and disorder specificity. *Netw. Neurosci.* **6**, 301–319 (2022).
255. Ribeiro-Gonçalves, B., Francisco, A. P., Vaz, C., Ramirez, M. & Carriço, J. A. PHYLOViZ Online: web-based tool for visualization, phylogenetic inference, analysis and sharing of minimum spanning trees. *Nucleic Acids Res.* **44**, W246-51 (2016).
256. Shaik, F., Bezawada, S. & Goveas, N. *CySpanningTree*: Minimal Spanning Tree computation in Cytoscape [version 1; peer review: 1 approved, 1 approved with reservations]. *F1000Res.* **4**, 476 (2015).

257. Liu, C. *et al.* Computational network biology: Data, models, and applications. *Physics Reports* (2019) doi:10.1016/j.physrep.2019.12.004.
258. Liu, A. *et al.* From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL. *BioRxiv* (2019) doi:10.1101/541888.
259. Scardoni, G., Tosadori, G., Pratap, S., Spoto, F. & Laudanna, C. Finding the shortest path with PesCa: a tool for network reconstruction. [version 2; peer review: 2 approved, 2 approved with reservations]. *F1000Res.* **4**, 484 (2015).
260. Sambaturu, N., Pusadkar, V., Hannenhalli, S. & Chandra, N. PathExt: a general framework for path-based mining of omics-integrated biological networks. *Bioinformatics* **37**, 1254–1262 (2021).
261. Akhmedov, M. *et al.* PCSF: An R-package for network-based interpretation of high-throughput data. *PLoS Comput. Biol.* **13**, e1005694 (2017).
262. Tuncbag, N. *et al.* Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. *PLoS Comput. Biol.* **12**, e1004879 (2016).
263. Azeloglu, E. U. & Iyengar, R. Signaling networks: information flow, computation, and decision making. *Cold Spring Harb. Perspect. Biol.* **7**, a005934 (2015).
264. Emmert-Streib, F., Dehmer, M. & Haibe-Kains, B. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Front. Cell Dev. Biol.* **2**, 38 (2014).
265. Lambert, S. A. *et al.* The human transcription factors. *Cell* **172**, 650–665 (2018).
266. Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237–1251 (2013).
267. Rao, V. S., Srinivas, K., Sujini, G. N. & Kumar, G. N. S. Protein-protein interaction detection: methods and analysis. *Int. J. Proteomics* **2014**, 147648 (2014).
268. Bossi, A. & Lehner, B. Tissue specificity and the human protein interaction network. *Mol. Syst. Biol.* **5**, 260 (2009).
269. Baxevanis, A. D. The Molecular Biology Database Collection: an updated compilation of biological database resources. *Nucleic Acids Res.* **29**, 1–10 (2001).
270. Bry, F. & Kröger, P. A Computational Biology Database Digest: Data, Data Analysis, and Data Management. *Distributed and Parallel Databases* (2003).
271. Imker, H. J. 25 years of molecular biology databases: A study of proliferation, impact, and maintenance. *Front. Res. Metr. Anal.* **3**, (2018).
272. Chen, C., Huang, H. & Wu, C. H. Protein bioinformatics databases and resources. *Methods Mol. Biol.* **1558**, 3–39 (2017).

273. Merali, Z. & Giles, J. Databases in peril. *Nature* **435**, 1010–1011 (2005).
274. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
275. Vénien-Bryan, C., Li, Z., Vuillard, L. & Boutin, J. A. Cryo-electron microscopy and X-ray crystallography: complementary approaches to structural biology and drug discovery. *Acta Crystallogr. F Struct. Biol. Commun.* **73**, 174–183 (2017).
276. Lee, L., Leopold, J. L. & Frank, R. L. Protein secondary structure prediction using BLAST and Relaxed Threshold Rule Induction from Coverings. in *2011 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* 1–8 (IEEE, 2011). doi:10.1109/CIBCB.2011.5948462.
277. Davey, N. E., Cyert, M. S. & Moses, A. M. Short linear motifs - ex nihilo evolution of protein regulation. *Cell Commun. Signal.* **13**, 43 (2015).
278. Wigington, C. P. *et al.* Systematic discovery of short linear motifs decodes calcineurin phosphatase signaling. *Mol. Cell* **79**, 342-358.e12 (2020).
279. Kumar, M. *et al.* ELM-the eukaryotic linear motif resource in 2020. *Nucleic Acids Res.* **48**, D296–D306 (2020).
280. Pawson, T. Protein modules and signalling networks. *Nature* **373**, 573–580 (1995).
281. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
282. Berman, H. M. *et al.* The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
283. Pieper, U. *et al.* ModBase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.* **39**, D465-74 (2011).
284. Lo Conte, L. *et al.* SCOP: a structural classification of proteins database. *Nucleic Acids Res.* **28**, 257–259 (2000).
285. Kiefer, F., Arnold, K., Künzli, M., Bordoli, L. & Schwede, T. The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res.* **37**, D387-92 (2009).
286. Lehne, B. & Schlitt, T. Protein-protein interaction databases: keeping up with growing interactomes. *Hum Genomics* **3**, 291–297 (2009).
287. Kerrien, S. *et al.* Broadening the horizon--level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.* **5**, 44 (2007).
288. Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612 (2021).
289. Hermjakob, H. *et al.* IntAct: an open source molecular interaction database. *Nucleic Acids Res.* **32**, D452-5 (2004).

290. Oughtred, R. *et al.* The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.* **30**, 187–200 (2021).
291. Bajpai, A. K. *et al.* Systematic comparison of the protein-protein interaction databases from a user's perspective. *J. Biomed. Inform.* **103**, 103380 (2020).
292. Huang, H. *et al.* A comprehensive protein-centric ID mapping service for molecular data integration. *Bioinformatics* **27**, 1190–1191 (2011).
293. Berriz, G. F. & Roth, F. P. The Synergizer service for translating gene, protein and other biological identifiers. *Bioinformatics* **24**, 2272–2273 (2008).
294. Bussey, K. J. *et al.* MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome Biol.* **4**, R27 (2003).
295. Diehn, M. *et al.* SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res.* **31**, 219–223 (2003).
296. Mudunuri, U., Che, A., Yi, M. & Stephens, R. M. bioDBnet: the biological database network. *Bioinformatics* **25**, 555–556 (2009).
297. Jassal, B. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).
298. Griss, J. *et al.* ReactomeGSA - Efficient Multi-Omics Comparative Pathway Analysis. *Mol. Cell. Proteomics* **19**, 2115–2125 (2020).
299. 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).
300. Masseroli, M. Biological and medical ontologies: introduction. in *Encyclopedia of bioinformatics and computational biology* 813–822 (Elsevier, 2019). doi:10.1016/B978-0-12-809633-8.20395-6.
301. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267-70 (2004).
302. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514-7 (2005).
303. Schoch, C. L. *et al.* NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)* **2020**, (2020).
304. The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOLD mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).
305. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).

306. Ceccarelli, F., Turei, D., Gabor, A. & Saez-Rodriguez, J. Bringing data from curated pathway resources to Cytoscape with OmniPath. *Bioinformatics* **36**, 2632–2633 (2020).
307. GBD 2017 Inflammatory Bowel Disease Collaborators. The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Gastroenterol. Hepatol.* **5**, 17–30 (2020).
308. Mak, W. Y., Zhao, M., Ng, S. C. & Burisch, J. The epidemiology of inflammatory bowel disease: East meets west. *J. Gastroenterol. Hepatol.* **35**, 380–389 (2020).
309. Glasser, A.-L. & Darfeuille-Michaud, A. Abnormalities in the handling of intracellular bacteria in Crohn's disease: a link between infectious etiology and host genetic susceptibility. *Arch Immunol Ther Exp (Warsz)* **56**, 237–244 (2008).
310. Baumgart, D. C. & Sandborn, W. J. Crohn's disease. *Lancet* **380**, 1590–1605 (2012).
311. Head, K. A. & Jurenka, J. S. Inflammatory bowel disease Part 1: ulcerative colitis--pathophysiology and conventional and alternative treatment options. *Altern. Med. Rev.* **8**, 247–283 (2003).
312. Smillie, C. S. *et al.* Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. *Cell* **178**, 714-730.e22 (2019).
313. Hanai, H. *et al.* A new paradigm in ulcerative colitis: regulatory T cells are key factor which induces/exacerbates UC through an immune imbalance. *Mol. Immunol.* **54**, 173–180 (2013).
314. Thul, P. J. & Lindskog, C. The human protein atlas: A spatial map of the human proteome. *Protein Sci.* **27**, 233–244 (2018).
315. Sprenger, J. *et al.* LOCATE: a mammalian protein subcellular localization database. *Nucleic Acids Res.* **36**, D230-3 (2008).
316. Veres, D. V. *et al.* CompPPI: a cellular compartment-specific database for protein-protein interaction network analysis. *Nucleic Acids Res.* **43**, D485-93 (2015).
317. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
318. Cheadle, C., Vawter, M. P., Freed, W. J. & Becker, K. G. Analysis of Microarray Data Using Z Score Transformation. *J. Mol. Diagn.* **5**, 73–81 (2003).
319. Hart, T., Komori, H. K., LaMere, S., Podshivalova, K. & Salomon, D. R. Finding the active genes in deep RNA-seq gene expression studies. *BMC Genomics* **14**, 778 (2013).
320. Valatas, V., Filidou, E., Drygiannakis, I. & Kolios, G. Stromal and immune cells in gut fibrosis: the myofibroblast and the scarface. *Ann. Gastroenterol.* **30**, 393–404 (2017).

321. Dorofeyev, A. E., Vasilenko, I. V., Rassokhina, O. A. & Kondratiuk, R. B. Mucosal barrier in ulcerative colitis and Crohn's disease. *Gastroenterol. Res. Pract.* **2013**, 431231 (2013).
322. Bates, J. & Diehl, L. Dendritic cells in IBD pathogenesis: an area of therapeutic opportunity? *J. Pathol.* **232**, 112–120 (2014).
323. Yamada, A. *et al.* Role of regulatory T cell in the pathogenesis of inflammatory bowel disease. *World J. Gastroenterol.* **22**, 2195–2205 (2016).
324. Dharmasiri, S. *et al.* Human intestinal macrophages are involved in the pathology of both ulcerative colitis and crohn disease. *Inflamm. Bowel Dis.* **27**, 1641–1652 (2021).
325. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize Implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
326. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
327. Krutzik, P. O., Crane, J. M., Clutter, M. R. & Nolan, G. P. High-content single-cell drug screening with phosphospecific flow cytometry. *Nat. Chem. Biol.* **4**, 132–142 (2008).
328. Gerritsen, J. S. & White, F. M. Phosphoproteomics: a valuable tool for uncovering molecular signaling in cancer cells. *Expert Rev. Proteomics* **18**, 661–674 (2021).
329. Wei, W. *et al.* Single-Cell Phosphoproteomics Resolves Adaptive Signaling Dynamics and Informs Targeted Combination Therapy in Glioblastoma. *Cancer Cell* **29**, 563–573 (2016).
330. Ben-Shlomo, I., Yu Hsu, S., Rauch, R., Kowalski, H. W. & Hsueh, A. J. W. Signaling receptome: a genomic and evolutionary perspective of plasma membrane receptors involved in signal transduction. *Sci. STKE* **2003**, RE9 (2003).
331. Kirouac, D. C. *et al.* Dynamic interaction networks in a hierarchically organized tissue. *Mol. Syst. Biol.* **6**, 417 (2010).
332. Andoh, A., Fujino, S., Okuno, T., Fujiyama, Y. & Bamba, T. Intestinal subepithelial myofibroblasts in inflammatory bowel diseases. *J. Gastroenterol.* **37 Suppl 14**, 33–37 (2002).
333. Armingol, E., Officer, A., Harismendy, O. & Lewis, N. E. Deciphering cell-cell interactions and communication from gene expression. *Nat. Rev. Genet.* **22**, 71–88 (2021).
334. Mueller, M., Rasoulinejad, S., Garg, S. & Wegner, S. V. The Importance of Cell-Cell Interaction Dynamics in Bottom-Up Tissue Engineering: Concepts of Colloidal Self-Assembly in the Fabrication of Multicellular Architectures. *Nano Lett.* **20**, 2257–2263 (2020).

335. Azadian, S., Zahiri, J., Shahriar Arab, S. & Hassan Sajedi, R. Reconstruction of Intercellular Signaling Network by Cytokine-Receptor Interactions. *Iran. J. Biotech.* **19**, e2560 (2021).
336. Stagg, A. J., Hart, A. L., Knight, S. C. & Kamm, M. A. The dendritic cell: its role in intestinal inflammation and relationship with gut bacteria. *Gut* **52**, 1522–1529 (2003).
337. He, T. *et al.* The p38 MAPK Inhibitor SB203580 Abrogates Tumor Necrosis Factor-Induced Proliferative Expansion of Mouse CD4⁺Foxp3⁺ Regulatory T Cells. *Front. Immunol.* **9**, 1556 (2018).
338. Forward, N. A., Furlong, S. J., Yang, Y., Lin, T.-J. & Hoskin, D. W. Signaling through TLR7 enhances the immunosuppressive activity of murine CD4⁺CD25⁺ T regulatory cells. *J. Leukoc. Biol.* **87**, 117–125 (2010).
339. Nyirenda, M. H. *et al.* TLR2 stimulation regulates the balance between regulatory T cell and Th17 function: a novel mechanism of reduced regulatory T cell function in multiple sclerosis. *J. Immunol.* **194**, 5761–5774 (2015).
340. Cao, A. T. *et al.* TLR4 regulates IFN- γ and IL-17 production by both thymic and induced Foxp3⁺ Tregs during intestinal inflammation. *J. Leukoc. Biol.* **96**, 895–905 (2014).
341. Sepehri, Z. *et al.* TLR3 and its roles in the pathogenesis of type 2 diabetes. *Cell Mol Biol (Noisy-le-grand)* **61**, 46–50 (2015).
342. Xiao, X. *et al.* Inflammatory regulation by TLR3 in acute hepatitis. *J. Immunol.* **183**, 3712–3719 (2009).
343. Drexler, M. & Institute of Medicine (US). *How Infection Works*. (2010).
344. Dani, A. Colonization and infection. *Cent. European J. Urol.* **67**, 86–87 (2014).
345. *Medical Microbiology*. (University of Texas Medical Branch at Galveston, 1996).
346. Casadevall, A. & Pirofski, L. A. Host-pathogen interactions: basic concepts of microbial commensalism, colonization, infection, and disease. *Infect. Immun.* **68**, 6511–6518 (2000).
347. Ribet, D. & Cossart, P. How bacterial pathogens colonize their hosts and invade deeper tissues. *Microbes Infect.* **17**, 173–183 (2015).
348. Young, V. B. Old and new models for studying host-microbe interactions in health and disease: *C. difficile* as an example. *Am. J. Physiol. Gastrointest. Liver Physiol.* **312**, G623–G627 (2017).
349. Lee, S.-A. *et al.* Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. *BMC Bioinformatics* **9 Suppl 12**, S11 (2008).

350. Rahmati, S. *et al.* pathDIP 4: an extended pathway annotations and enrichment analysis resource for human, model organisms and domesticated species. *Nucleic Acids Res.* **48**, D479–D488 (2020).
351. Park, D., Singh, R., Baym, M., Liao, C.-S. & Berger, B. IsoBase: a database of functionally related proteins across PPI networks. *Nucleic Acids Res.* **39**, D295–300 (2011).
352. Sonnhammer, E. L. L. & Östlund, G. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.* **43**, D234–9 (2015).
353. Kotlyar, M., Pastrello, C., Sheahan, N. & Jurisica, I. Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic Acids Res.* **44**, D536–41 (2016).
354. Singh, R., Xu, J. & Berger, B. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc Natl Acad Sci USA* **105**, 12763–12768 (2008).
355. Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86–90 (1999).
356. Galperin, M. Y. *et al.* COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.* **49**, D274–D281 (2021).
357. Wojcik, J. & Schächter, V. Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics* **17 Suppl 1**, S296–305 (2001).
358. Alborzi, S. Z., Ahmed Nacer, A., Najjar, H., Ritchie, D. W. & Devignes, M.-D. PPIDomainMiner: Inferring domain-domain interactions from multiple sources of protein-protein interactions. *PLoS Comput. Biol.* **17**, e1008844 (2021).
359. Sarkar, D., Jana, T. & Saha, S. LMPID: a manually curated database of linear motifs mediating protein-protein interactions. *Database (Oxford)* **2015**, (2015).
360. Encinar, J. A. *et al.* ADAN: a database for prediction of protein-protein interaction of modular domains mediated by linear motifs. *Bioinformatics* **25**, 2418–2424 (2009).
361. Lian, X., Yang, S., Li, H., Fu, C. & Zhang, Z. Machine-Learning-Based Predictor of Human-Bacteria Protein-Protein Interactions by Incorporating Comprehensive Host-Network Properties. *J. Proteome Res.* **18**, 2195–2205 (2019).
362. Fritz, J. V., Desai, M. S., Shah, P., Schneider, J. G. & Wilmes, P. From meta-omics to causality: experimental models for human microbiome research. *Microbiome* **1**, 14 (2013).
363. Bandyopadhyay, S., Ray, S., Mukhopadhyay, A. & Maulik, U. A review of in silico approaches for analysis and prediction of HIV-1-human protein-protein interactions. *Brief. Bioinformatics* **16**, 830–851 (2015).

364. Zhang, Q. C. *et al.* Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **490**, 556–560 (2012).
365. Singhal, M. & Resat, H. A domain-based approach to predict protein-protein interactions. *BMC Bioinformatics* **8**, 199 (2007).
366. Wootton, J. C. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.* **18**, 269–285 (1994).
367. Walsh, I., Martin, A. J. M., Di Domenico, T. & Tosatto, S. C. E. ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics* **28**, 503–509 (2012).
368. Ghalwash, M. F., Dunker, A. K. & Obradović, Z. Uncertainty analysis in protein disorder prediction. *Mol. Biosyst.* **8**, 381–391 (2012).
369. Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F. & Jones, D. T. The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20**, 2138–2139 (2004).
370. Ishida, T. & Kinoshita, K. PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res.* **35**, W460-4 (2007).
371. Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K. & Obradovic, Z. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* **7**, 208 (2006).
372. Romero, P. *et al.* Sequence complexity of disordered protein. *Proteins: Structure, Function, and Bioinformatics* (2001).
373. Mészáros, B., Erdos, G. & Dosztányi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **46**, W329–W337 (2018).
374. Hugo, W., Sung, W.-K. & Ng, S.-K. Discovering interacting domains and motifs in protein-protein interactions. *Methods Mol. Biol.* **939**, 9–20 (2013).
375. Gouw, M. *et al.* The eukaryotic linear motif resource - 2018 update. *Nucleic Acids Res.* **46**, D428–D434 (2018).
376. Mosca, R., Céol, A., Stein, A., Olivella, R. & Aloy, P. 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.* **42**, D374-9 (2014).
377. Zhang, A., He, L. & Wang, Y. Prediction of GCRV virus-host protein interactome based on structural motif-domain interactions. *BMC Bioinformatics* **18**, 145 (2017).
378. Halehalli, R. R. & Nagarajaram, H. A. Molecular principles of human virus protein-protein interactions. *Bioinformatics* **31**, 1025–1033 (2015).
379. Sudhakar, P. *et al.* Targeted interplay between bacterial pathogens and host autophagy. *Autophagy* **15**, 1620–1633 (2019).

380. Andrichetti, T., Bohar, B., Lemke, N., Sudhakar, P. & Korcsmaros, T. MicrobioLink: An Integrated Computational Pipeline to Infer Functional Effects of Microbiome-Host Interactions. *Cells* **9**, (2020).
381. Ng, S.-K., Zhang, Z., Tan, S.-H. & Lin, K. InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res.* **31**, 251–254 (2003).
382. Luo, Q., Pagel, P., Vilne, B. & Frishman, D. DIMA 3.0: domain interaction map. *Nucleic Acids Res.* **39**, D724-9 (2011).
383. Kim, Y., Min, B. & Yi, G.-S. IDDI: integrated domain-domain interaction and protein interaction analysis system. *Proteome Sci.* **10 Suppl 1**, S9 (2012).
384. Yellaboina, S., Tasneem, A., Zaykin, D. V., Raghavachari, B. & Jothi, R. DOMINE: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Res.* **39**, D730-5 (2011).
385. Davis, F. P. & Sali, A. PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics* **21**, 1901–1907 (2005).
386. Weimer, B. C., Chen, P., Desai, P. T., Chen, D. & Shah, J. Whole Cell Cross-Linking to Discover Host-Microbe Protein Cognate Receptor/Ligand Pairs. *Front. Microbiol.* **9**, 1585 (2018).
387. Mix, A.-K., Goob, G., Sontowski, E. & Hauck, C. R. Microscale communication between bacterial pathogens and the host epithelium. *Genes Immun.* **22**, 247–254 (2021).
388. Maffei, B., Francetic, O. & Subtil, A. Tracking proteins secreted by bacteria: what's in the toolbox? *Front. Cell. Infect. Microbiol.* **7**, 221 (2017).
389. Merkel, D. Docker: lightweight linux containers for consistent development and deployment. *Linux j* (2014).
390. Peabody, M. A., Laird, M. R., Vlasschaert, C., Lo, R. & Brinkman, F. S. L. PSORTdb: expanding the bacteria and archaea protein subcellular localization database to better reflect diversity in cell envelope structures. *Nucleic Acids Res.* **44**, D663-8 (2016).
391. Nair, R. & Rost, B. Sequence conserved for subcellular localization. *Protein Sci.* **11**, 2836–2847 (2002).
392. Goodacre, N. F., Gerloff, D. L. & Uetz, P. Protein domains of unknown function are essential in bacteria. *MBio* **5**, e00744-13 (2013).
393. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
394. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**, D344–D354 (2021).

395. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
396. Dinkel, H. *et al.* ELM--the database of eukaryotic linear motifs. *Nucleic Acids Res.* **40**, D242-51 (2012).
397. Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* **13**, 966–967 (2016).
398. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.* **29**, 1363–1375 (2019).
399. Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* **48**, 838–847 (2016).
400. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
401. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).
402. Våremo, L., Nielsen, J. & Nookaew, I. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res.* **41**, 4378–4391 (2013).
403. Gansner, E. R. & North, S. C. An open graph visualization system and its applications to software engineering. *Softw: Pract. Exper.* **30**, 1203–1233 (2000).
404. Zanzoni, A., Spinelli, L., Braham, S. & Brun, C. Perturbed human sub-networks by *Fusobacterium nucleatum* candidate virulence proteins. *Microbiome* **5**, 89 (2017).
405. Levy, R., Carr, R., Kreimer, A., Freilich, S. & Borenstein, E. NetCooperate: a network-based tool for inferring host-microbe and microbe-microbe cooperation. *BMC Bioinformatics* **16**, 164 (2015).
406. Brunk, E. *et al.* Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nat. Biotechnol.* **36**, 272–281 (2018).
407. Nomura, K. *et al.* Bacteroidetes Species Are Correlated with Disease Activity in Ulcerative Colitis. *J. Clin. Med.* **10**, (2021).
408. Hansen, J., Gulati, A. & Sartor, R. B. The role of mucosal immunity and host genetics in defining intestinal commensal bacteria. *Curr. Opin. Gastroenterol.* **26**, 564–571 (2010).
409. Zakerska-Banaszak, O. *et al.* Dysbiosis of gut microbiota in Polish patients with ulcerative colitis: a pilot study. *Sci. Rep.* **11**, 2166 (2021).

410. Zhou, Y. & Zhi, F. Lower Level of Bacteroides in the Gut Microbiota Is Associated with Inflammatory Bowel Disease: A Meta-Analysis. *Biomed Res. Int.* **2016**, 5828959 (2016).
411. Salyers, A. A. Bacteroides of the human lower intestinal tract. *Annu. Rev. Microbiol.* **38**, 293–313 (1984).
412. Townsend, G. E. *et al.* A Master Regulator of Bacteroides thetaiotaomicron Gut Colonization Controls Carbohydrate Utilization and an Alternative Protein Synthesis Factor. *MBio* **11**, (2020).
413. Murphy, E. C., Mörgelin, M., Cooney, J. C. & Frick, I.-M. Interaction of Bacteroides fragilis and Bacteroides thetaiotaomicron with the kallikrein-kinin system. *Microbiology (Reading, Engl)* **157**, 2094–2105 (2011).
414. Gilmore, M. S. & Ferretti, J. J. Microbiology. The thin line between gut commensal and pathogen. *Science* **299**, 1999–2002 (2003).
415. Schwechheimer, C. & Kuehn, M. J. Outer-membrane vesicles from Gram-negative bacteria: biogenesis and functions. *Nat. Rev. Microbiol.* **13**, 605–619 (2015).
416. Gill, S., Catchpole, R. & Forterre, P. Extracellular membrane vesicles in the three domains of life and beyond. *FEMS Microbiol. Rev.* **43**, 273–303 (2019).
417. Chang, X. *et al.* Extracellular Vesicles with Possible Roles in Gut Intestinal Tract Homeostasis and IBD. *Mediators Inflamm.* **2020**, 1945832 (2020).
418. Berezow, A. B. *et al.* The structurally similar, penta-acylated lipopolysaccharides of Porphyromonas gingivalis and Bacteroides elicit strikingly different innate immune responses. *Microb. Pathog.* **47**, 68–77 (2009).
419. Jacobson, A. N., Choudhury, B. P. & Fischbach, M. A. The Biosynthesis of Lipooligosaccharide from Bacteroides thetaiotaomicron. *MBio* **9**, (2018).
420. Steimle, A. *et al.* Weak agonistic LPS restores intestinal immune homeostasis. *Mol. Ther.* **27**, 1974–1991 (2019).
421. Jones, E. J. *et al.* The Uptake, Trafficking, and Biodistribution of Bacteroides thetaiotaomicron Generated Outer Membrane Vesicles. *Front. Microbiol.* **11**, 57 (2020).
422. Hickey, C. A. *et al.* Colitogenic Bacteroides thetaiotaomicron Antigens Access Host Immune Cells in a Sulfatase-Dependent Manner via Outer Membrane Vesicles. *Cell Host Microbe* **17**, 672–680 (2015).
423. Kaparakis-Liaskos, M. & Ferrero, R. L. Immune modulation by bacterial outer membrane vesicles. *Nat. Rev. Immunol.* **15**, 375–387 (2015).
424. Cecil, J. D., Sirisaengtaksin, N., O'Brien-Simpson, N. M. & Krachler, A. M. Outer Membrane Vesicle-Host Cell Interactions. *Microbiol. Spectr.* **7**, (2019).

425. Durant, L. *et al.* Bacteroides thetaiotaomicron-derived outer membrane vesicles promote regulatory dendritic cell responses in health but not in inflammatory bowel disease. *Microbiome* **8**, 88 (2020).
426. Stentz, R., Carvalho, A. L., Jones, E. J. & Carding, S. R. Fantastic voyage: the journey of intestinal microbiota-derived microvesicles through the body. *Biochem. Soc. Trans.* **46**, 1021–1027 (2018).
427. Chang, C.-J. *et al.* Next generation probiotics in disease amelioration. *J. Food Drug Anal.* **27**, 615–622 (2019).
428. Shen, Z.-H. *et al.* Relationship between intestinal microbiota and ulcerative colitis: Mechanisms and clinical application of probiotics and fecal microbiota transplantation. *World J. Gastroenterol.* **24**, 5–14 (2018).
429. Peng, L., Zhong, Y., Wang, A. & Jiang, Z. Probiotics combined with aminosalicic acid affiliates remission of ulcerative colitis: a meta-analysis of randomized controlled trial. *Biosci. Rep.* **39**, (2019).
430. Ihezor-Ejiofor, Z. *et al.* Probiotics for maintenance of remission in ulcerative colitis. *Cochrane Database Syst. Rev.* **3**, CD007443 (2020).
431. Fedorak, R. N. Probiotics in the management of ulcerative colitis. *Gastroenterol Hepatol (N Y)* **6**, 688–690 (2010).
432. Sartor, R. B. Mechanisms of disease: pathogenesis of Crohn's disease and ulcerative colitis. *Nat. Clin. Pract. Gastroenterol. Hepatol.* **3**, 390–407 (2006).
433. Kelly, D. *et al.* Commensal anaerobic gut bacteria attenuate inflammation by regulating nuclear-cytoplasmic shuttling of PPAR-gamma and RelA. *Nat. Immunol.* **5**, 104–112 (2004).
434. Wrzosek, L. *et al.* Bacteroides thetaiotaomicron and Faecalibacterium prausnitzii influence the production of mucus glycans and the development of goblet cells in the colonic epithelium of a gnotobiotic model rodent. *BMC Biol.* **11**, 61 (2013).
435. Hooper, L. V. *et al.* Molecular analysis of commensal host-microbial relationships in the intestine. *Science* **291**, 881–884 (2001).
436. Hoffmann, T. W. *et al.* Microorganisms linked to inflammatory bowel disease-associated dysbiosis differentially impact host physiology in gnotobiotic mice. *ISME J.* **10**, 460–477 (2016).
437. Gul, L. *et al.* Extracellular vesicles produced by the human commensal gut bacterium Bacteroides thetaiotaomicron affect host immune pathways in a cell-type specific manner that are altered in inflammatory bowel disease. *J. Extracell. Vesicles* **11**, e12189 (2022).
438. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

439. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **6**, e21800 (2011).
440. Heberle, H., Meirelles, G. V., da Silva, F. R., Telles, G. P. & Minghim, R. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinformatics* **16**, 169 (2015).
441. Tsukamoto, H. *et al.* Lipopolysaccharide (LPS)-binding protein stimulates CD14-dependent Toll-like receptor 4 internalization and LPS-induced TBK1-IKK ϵ -IRF3 axis activation. *J. Biol. Chem.* **293**, 10186–10201 (2018).
442. Carroll, R. K. *et al.* The lone S41 family C-terminal processing protease in *Staphylococcus aureus* is localized to the cell wall and contributes to virulence. *Microbiology (Reading, Engl)* **160**, 1737–1748 (2014).
443. Rajpoot, S., Kumar, A., Zhang, K. Y. J., Gan, S. H. & Baig, M. S. TIRAP-mediated activation of p38 MAPK in inflammatory signaling. *Sci. Rep.* **12**, 5601 (2022).
444. Charlet, R., Bortolus, C., Sendid, B. & Jawhara, S. *Bacteroides thetaiotaomicron* and *Lactobacillus johnsonii* modulate intestinal inflammation and eliminate fungi via enzymatic hydrolysis of the fungal cell wall. *Sci. Rep.* **10**, 11510 (2020).
445. Coats, S. R. *et al.* The lipid A phosphate position determines differential host Toll-like receptor 4 responses to phylogenetically related symbiotic and pathogenic bacteria. *Infect. Immun.* **79**, 203–210 (2011).
446. Kuehn, M. J. & Kesty, N. C. Bacterial outer membrane vesicles and the host-pathogen interaction. *Genes Dev.* **19**, 2645–2655 (2005).
447. Delday, M., Mulder, I., Logan, E. T. & Grant, G. *Bacteroides thetaiotaomicron* Ameliorates Colon Inflammation in Preclinical Models of Crohn's Disease. *Inflamm. Bowel Dis.* **25**, 85–96 (2019).
448. Swirski, F. K., Hilgendorf, I. & Robbins, C. S. From proliferation to proliferation: monocyte lineage comes full circle. *Semin. Immunopathol.* **36**, 137–148 (2014).
449. Lih-Brody, L. *et al.* Increased oxidative stress and decreased antioxidant defenses in mucosa of inflammatory bowel disease. *Dig. Dis. Sci.* **41**, 2078–2086 (1996).
450. D'Inca, R. *et al.* Oxidative DNA damage in the mucosa of ulcerative colitis increases with disease duration and dysplasia. *Inflamm. Bowel Dis.* **10**, 23–27 (2004).
451. Dincer, Y. *et al.* Oxidative DNA damage and antioxidant activity in patients with inflammatory bowel disease. *Dig. Dis. Sci.* **52**, 1636–1641 (2007).
452. Aslan, M. *et al.* Peripheral lymphocyte DNA damage and oxidative stress in patients with ulcerative colitis. *Pol. Arch. Med. Wewn.* **121**, 223–229 (2011).

453. Beltrán, B. *et al.* Mitochondrial dysfunction, persistent oxidative damage, and catalase inhibition in immune cells of naïve and treated Crohn's disease. *Inflamm. Bowel Dis.* **16**, 76–86 (2010).
454. Pasquier, L. D. Germline and somatic diversification of immune recognition elements in Metazoa. *Immunol. Lett.* **104**, 2–17 (2006).
455. Collins, L. E., DeCoursey, J., Soledad di Luca, M., Rochfort, K. D. & Loscher, C. E. An emerging role for SNARE proteins in dendritic cell function. *Front. Immunol.* **6**, 133 (2015).
456. Hardbower, D. M. *et al.* EGFR regulates macrophage activation and function in bacterial infection. *J. Clin. Invest.* **126**, 3296–3312 (2016).
457. Thabet, N. A., El-Guendy, N., Mohamed, M. M. & Shouman, S. A. Suppression of macrophages- Induced inflammation via targeting RAS and PAR-4 signaling in breast cancer cell lines. *Toxicol. Appl. Pharmacol.* **385**, 114773 (2019).
458. Levin, A. & Shibolet, O. Toll-like receptors in inflammatory bowel disease-stepping into uncharted territory. *World J. Gastroenterol.* **14**, 5149–5153 (2008).
459. Horng, T., Barton, G. M., Flavell, R. A. & Medzhitov, R. The adaptor molecule TIRAP provides signalling specificity for Toll-like receptors. *Nature* **420**, 329–333 (2002).
460. Coats, S. R. *et al.* Cardiolipins Act as a Selective Barrier to Toll-Like Receptor 4 Activation in the Intestine. *Appl. Environ. Microbiol.* **82**, 4264–4278 (2016).
461. Lee, H.-J. & Zheng, J. J. PDZ domains and their binding partners: structure, specificity, and modification. *Cell Commun. Signal.* **8**, 8 (2010).
462. Harris, B. Z. & Lim, W. A. Mechanism and role of PDZ domains in signaling complex assembly. *J. Cell Sci.* **114**, 3219–3231 (2001).
463. Rajpoot, S. *et al.* TIRAP in the mechanism of inflammation. *Front. Immunol.* **12**, 697588 (2021).
464. Griffen, A. L. *et al.* CORE: a phylogenetically-curated 16S rDNA database of the core oral microbiome. *PLoS ONE* **6**, e19051 (2011).
465. Socransky, S. S., Haffajee, A. D., Cugini, M. A., Smith, C. & Kent, R. L. Microbial complexes in subgingival plaque. *J. Clin. Periodontol.* **25**, 134–144 (1998).
466. Le Bars, P. *et al.* The oral cavity microbiota: between health, oral disease, and cancers of the aerodigestive tract. *Can. J. Microbiol.* **63**, 475–492 (2017).
467. Bik, E. M. *et al.* Bacterial diversity in the oral cavity of 10 healthy individuals. *ISME J.* **4**, 962–974 (2010).

468. Colombo, A. P. V. *et al.* Comparisons of subgingival microbial profiles of refractory periodontitis, severe periodontitis, and periodontal health using the human oral microbe identification microarray. *J. Periodontol.* **80**, 1421–1432 (2009).
469. Nayak, A. *et al.* Detection of red complex organisms in chronic periodontitis by multiplex polymerase chain reaction. *JCRI* **5**, 139–144 (2018).
470. Sterzenbach, T., Helbig, R., Hannig, C. & Hannig, M. Bioadhesion in the oral cavity and approaches for biofilm management by surface modifications. *Clin. Oral Investig.* **24**, 4237–4260 (2020).
471. Bamashmous, S. *et al.* Human variation in gingival inflammation. *Proc Natl Acad Sci USA* **118**, (2021).
472. Wade, W. G. The oral microbiome in health and disease. *Pharmacol. Res.* **69**, 137–143 (2013).
473. Genco, R. J. & Van Dyke, T. E. Prevention: Reducing the risk of CVD in patients with periodontitis. *Nat. Rev. Cardiol.* **7**, 479–480 (2010).
474. Lalla, E. & Papapanou, P. N. Diabetes mellitus and periodontitis: a tale of two common interrelated diseases. *Nat. Rev. Endocrinol.* **7**, 738–748 (2011).
475. Irani, S., Barati, I. & Badiiei, M. Periodontitis and oral cancer - current concepts of the etiopathogenesis. *Oncol. Rev.* **14**, 465 (2020).
476. Lenartova, M. *et al.* The oral microbiome in periodontal health. *Front. Cell. Infect. Microbiol.* **11**, 629723 (2021).
477. Abusleme, L. *et al.* The subgingival microbiome in health and periodontitis and its relationship with community biomass and inflammation. *ISME J.* **7**, 1016–1025 (2013).
478. Griffen, A. L. *et al.* Distinct and complex bacterial profiles in human periodontitis and health revealed by 16S pyrosequencing. *ISME J.* **6**, 1176–1185 (2012).
479. Lovegrove, J. M. Dental plaque revisited: bacteria associated with periodontal disease. *J. N. Z. Soc. Periodontol.* 7–21 (2004).
480. Coelho, E. D. *et al.* Computational prediction of the human-microbial oral interactome. *BMC Syst. Biol.* **8**, 24 (2014).
481. Bao, K., Belibasakis, G. N., Selevsek, N., Grossmann, J. & Bostanci, N. Proteomic profiling of host-biofilm interactions in an oral infection model resembling the periodontal pocket. *Sci. Rep.* **5**, 15999 (2015).
482. Stathopoulou, P. G., Benakanakere, M. R., Galicia, J. C. & Kinane, D. F. Epithelial cell pro-inflammatory cytokine response differs across dental plaque bacterial species. *J. Clin. Periodontol.* **37**, 24–29 (2010).

483. Rouabhia, M. Interactions between host and oral commensal microorganisms are key events in health and disease status. *Can. J. Infect. Dis.* **13**, 47–51 (2002).
484. Radaic, A. & Kapila, Y. L. The oralome and its dysbiosis: New insights into oral microbiome-host interactions. *Comput. Struct. Biotechnol. J.* **19**, 1335–1360 (2021).
485. Zaura, E., Keijser, B. J. F., Huse, S. M. & Crielaard, W. Defining the healthy “core microbiome” of oral microbial communities. *BMC Microbiol.* **9**, 259 (2009).
486. Zhu, B., Macleod, L. C., Kitten, T. & Xu, P. Streptococcus sanguinis biofilm formation & interaction with oral pathogens. *Future Microbiol.* **13**, 915–932 (2018).
487. Perera, D. *et al.* Mechanisms underlying proximity between oral commensal bacteria. *BioRxiv* (2020) doi:10.1101/2020.09.29.318816.
488. Lim, Y. K. *et al.* Lautropia dentalis sp. nov., Isolated from Human Dental Plaque of a Gingivitis Lesion. *Curr. Microbiol.* **76**, 1369–1373 (2019).
489. Yee, M., Kim, S., Sethi, P., Düzgüneş, N. & Konopka, K. Porphyromonas gingivalis stimulates IL-6 and IL-8 secretion in GMSM-K, HSC-3 and H413 oral epithelial cells. *Anaerobe* **28**, 62–67 (2014).
490. Bloch, S., Thurnheer, T., Murakami, Y., Belibasakis, G. N. & Schäffer, C. Behavior of two Tannerella forsythia strains and their cell surface mutants in multispecies oral biofilms. *Mol. Oral Microbiol.* **32**, 404–418 (2017).
491. Wang, Q., Wright, C. J., Dingming, H., Uriarte, S. M. & Lamont, R. J. Oral community interactions of Filifactor alocis in vitro. *PLoS ONE* **8**, e76271 (2013).
492. Yu, N. Y. *et al.* PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* **26**, 1608–1615 (2010).
493. Stadnyk, A. W. Cytokine production by epithelial cells. *FASEB J.* **8**, 1041–1047 (1994).
494. Gao, N., Yin, J., Yoon, G. S., Mi, Q.-S. & Yu, F.-S. X. Dendritic cell-epithelium interplay is a determinant factor for corneal epithelial wound repair. *Am. J. Pathol.* **179**, 2243–2253 (2011).
495. Kirkwood, K. L. & Rossa, C. The potential of p38 MAPK inhibitors to modulate periodontal infections. *Curr. Drug Metab.* **10**, 55–67 (2009).
496. Li, Q., Valerio, M. S. & Kirkwood, K. L. MAPK usage in periodontal disease progression. *J. Signal Transduct.* **2012**, 308943 (2012).
497. Kajita, K. *et al.* Quantitative messenger RNA expression of Toll-like receptors and interferon-alpha1 in gingivitis and periodontitis. *Oral Microbiol. Immunol.* **22**, 398–402 (2007).

498. Ren, L., Leung, W. K., Darveau, R. P. & Jin, L. The expression profile of lipopolysaccharide-binding protein, membrane-bound CD14, and toll-like receptors 2 and 4 in chronic periodontitis. *J. Periodontol.* **76**, 1950–1959 (2005).
499. Patel, S. Pathogenicity-associated protein domains: The fiercely-conserved evolutionary signatures. *Gene Rep.* **7**, 127–141 (2017).
500. Djinic Krasavcevic, A. *et al.* Impact of Notch signalling molecules and bone resorption regulators on clinical parameters in periodontitis. *J. Periodont. Res.* **56**, 131–138 (2021).
501. Mijailovic, I. *et al.* The down-regulation of Notch 1 signaling contributes to the severity of bone loss in aggressive periodontitis. *J. Periodontol.* **91**, 554–561 (2020).
502. Figueredo, C. M., Lira-Junior, R. & Love, R. M. T and B cells in periodontal disease: new functions in A complex scenario. *Int. J. Mol. Sci.* **20**, (2019).
503. Kim, Y. C. *et al.* Presence of Porphyromonas gingivalis and plasma cell dominance in gingival tissues with periodontitis. *Oral Dis.* **16**, 375–381 (2010).
504. Dutzan, N., Abusleme, L., Konkell, J. E. & Moutsopoulos, N. M. Isolation, characterization and functional examination of the gingival immune cell network. *J. Vis. Exp.* 53736 (2016) doi:10.3791/53736.
505. Artese, L. *et al.* Immunohistochemical analysis of inflammatory infiltrate in aggressive and chronic periodontitis: a comparative study. *Clin. Oral Investig.* **15**, 233–240 (2011).
506. Dutzan, N., Konkell, J. E., Greenwell-Wild, T. & Moutsopoulos, N. M. Characterization of the human immune cell network at the gingival barrier. *Mucosal Immunol.* **9**, 1163–1172 (2016).
507. Mahanonda, R. *et al.* Human memory B cells in healthy gingiva, gingivitis, and periodontitis. *J. Immunol.* **197**, 715–725 (2016).
508. Ma, N. *et al.* Involvement of interleukin-23 induced by Porphyromonas endodontalis lipopolysaccharide in osteoclastogenesis. *Mol. Med. Report.* **15**, 559–566 (2017).
509. Liu, J., Zhang, X., Cheng, Y. & Cao, X. Dendritic cell migration in inflammation and immunity. *Cell. Mol. Immunol.* **18**, 2461–2471 (2021).
510. Hansen, A. S. *et al.* CD46 activation induces distinct CXCL-10 response in monocytes and monocyte-derived dendritic cells. *Cytokine* **113**, 466–469 (2019).
511. Sivori, S. *et al.* TLR/NCR/KIR: which one to use and when? *Front. Immunol.* **5**, 105 (2014).
512. Larsen, S. B., Cowley, C. J. & Fuchs, E. Epithelial cells: liaisons of immunity. *Curr. Opin. Immunol.* **62**, 45–53 (2020).

513. Davidson, S. *et al.* Fibroblasts as immune regulators in infection, inflammation and cancer. *Nat. Rev. Immunol.* **21**, 704–717 (2021).
514. Davoodi, S. & Foley, E. Host-Microbe-Pathogen Interactions: A Review of *Vibrio cholerae* Pathogenesis in *Drosophila*. *Front. Immunol.* **10**, 3128 (2019).
515. Rosman-Urbach, M., Niv, Y., Birk, Y., Morgenstern, S. & Schwartz, B. Relationship between nutritional habits adopted by ulcerative colitis relevant to cancer development patients at clinical remission stages and molecular-genetic parameters. *Br. J. Nutr.* **95**, 188–195 (2006).
516. Elson, C. O. & Cong, Y. Host-microbiota interactions in inflammatory bowel disease. *Gut Microbes* **3**, 332–344 (2012).
517. Bartold, P. M. & Van Dyke, T. E. Periodontitis: a host-mediated disruption of microbial homeostasis. Unlearning learned concepts. *Periodontol. 2000* **62**, 203–217 (2013).
518. Onyiah, J. C. & Colgan, S. P. Cytokine responses and epithelial function in the intestinal mucosa. *Cell. Mol. Life Sci.* **73**, 4203–4212 (2016).
519. Ji, S., Choi, Y. S. & Choi, Y. Bacterial invasion and persistence: critical events in the pathogenesis of periodontitis? *J. Periodont. Res.* **50**, 570–585 (2015).
520. Flemmig, T. F., Shanahan, F. & Miyasaki, K. T. Prevalence and severity of periodontal disease in patients with inflammatory bowel disease. *J. Clin. Periodontol.* **18**, 690–697 (1991).
521. Habashneh, R. A., Khader, Y. S., Alhumouz, M. K., Jadallah, K. & Ajlouni, Y. The association between inflammatory bowel disease and periodontitis among Jordanians: a case-control study. *J. Periodont. Res.* **47**, 293–298 (2012).
522. Van Dyke, T. E., Dowell, V. R., Offenbacher, S., Snyder, W. & Hersh, T. Potential role of microorganisms isolated from periodontal lesions in the pathogenesis of inflammatory bowel disease. *Infect. Immun.* **53**, 671–677 (1986).
523. Stein, J. M. *et al.* Clinical periodontal and microbiologic parameters in patients with Crohn's disease with consideration of the CARD15 genotype. *J. Periodontol.* **81**, 535–545 (2010).
524. Read, E., Curtis, M. A. & Neves, J. F. The role of oral bacteria in inflammatory bowel disease. *Nat. Rev. Gastroenterol. Hepatol.* **18**, 731–742 (2021).
525. Brito, F. *et al.* Prevalence of periodontitis and DMFT index in patients with Crohn's disease and ulcerative colitis. *J. Clin. Periodontol.* **35**, 555–560 (2008).
526. Grössner-Schreiber, B. *et al.* Prevalence of dental caries and periodontal disease in patients with inflammatory bowel disease: a case-control study. *J. Clin. Periodontol.* **33**, 478–484 (2006).

527. Figueredo, C. M. *et al.* Expression of cytokines in the gingival crevicular fluid and serum from patients with inflammatory bowel disease and untreated chronic periodontitis. *J. Periodont. Res.* **46**, 141–146 (2011).
528. Vavricka, S. R. *et al.* Periodontitis and gingivitis in inflammatory bowel disease: a case-control study. *Inflamm. Bowel Dis.* **19**, 2768–2777 (2013).
529. Koutsochristou, V. *et al.* Dental Caries and Periodontal Disease in Children and Adolescents with Inflammatory Bowel Disease: A Case-Control Study. *Inflamm. Bowel Dis.* **21**, 1839–1846 (2015).
530. Pan, J. *et al.* DWPPi: A Deep Learning Approach for Predicting Protein-Protein Interactions in Plants Based on Multi-Source Information With a Large-Scale Biological Network. *Front. Bioeng. Biotechnol.* **10**, 807522 (2022).
531. Wong, F. *et al.* Benchmarking AlphaFold-enabled molecular docking predictions for antibiotic discovery. *Mol. Syst. Biol.* **18**, e11081 (2022).
532. Akdel, M. *et al.* A structural biology community assessment of AlphaFold2 applications. *Nat. Struct. Mol. Biol.* **29**, 1056–1067 (2022).
533. Lee, C., Su, B.-H. & Tseng, Y. J. Comparative studies of AlphaFold, RoseTTAFold and Modeller: a case study involving the use of G-protein-coupled receptors. *Brief. Bioinformatics* **23**, (2022).
534. Straight, P. D. & Kolter, R. Interspecies chemical communication in bacterial development. *Annu. Rev. Microbiol.* **63**, 99–118 (2009).
535. Dogra, S. K., Doré, J. & Damak, S. Gut microbiota resilience: definition, link to health and strategies for intervention. *Front. Microbiol.* **11**, 572921 (2020).
536. Wang, L. *et al.* Facial Skin Microbiota-Mediated Host Response to Pollution Stress Revealed by Microbiome Networks of Individual. *mSystems* **6**, e0031921 (2021).
537. Kwon, Y.-K. & Cho, K.-H. Quantitative analysis of robustness and fragility in biological networks based on feedback dynamics. *Bioinformatics* **24**, 987–994 (2008).
538. Visconti, A. *et al.* Interplay between the human gut microbiome and host metabolism. *Nat. Commun.* **10**, 4505 (2019).
539. Zhang, H. *et al.* A volatile from the skin microbiota of flavivirus-infected hosts promotes mosquito attractiveness. *Cell* (2022) doi:10.1016/j.cell.2022.05.016.
540. Gul, L., Korcsmaros, T. & Hall, N. Flaviviruses hijack the host microbiota to facilitate their transmission. *Cell* **185**, 2395–2397 (2022).
541. Treveil, A. *et al.* ViralLink: An integrated workflow to investigate the effect of SARS-CoV-2 on intracellular signalling and regulatory pathways. *PLoS Comput. Biol.* **17**, e1008685 (2021).

542. Poletti, M. *et al.* Mapping the epithelial-immune cell interactome upon infection in the gut and the upper airways. *NPJ Syst. Biol. Appl.* **8**, 15 (2022).
543. Olbei, M. *et al.* CytokineLink: A Cytokine Communication Map to Analyse Immune Responses-Case Studies in Inflammatory Bowel Disease and COVID-19. *Cells* **10**, (2021).
544. Jacomin, A.-C., Gul, L., Sudhakar, P., Korcsmaros, T. & Nezis, I. P. What we learned from big data for autophagy research. *Front. Cell Dev. Biol.* **6**, 92 (2018).
545. Jones, E. J. *et al.* Integrative analysis of Paneth cell proteomic and transcriptomic data from intestinal organoids reveals functional processes dependent on autophagy. *Dis. Model. Mech.* **12**, (2019).

Appendix 1: Supplementary material

Supplementary Table 2.1: Number of condition-specific intercellular PPIs				
Source cell	Target cell	Healthy condition	UC condition	Difference
Goblet cell	Myofibroblast	343	316	27
Goblet cell	Dendritic cell	654	166	488
Goblet cell	Regulatory T cell	450	486	36
Goblet cell	Macrophage	416	266	150
Myofibroblast	Goblet cell	458	206	252
Myofibroblast	Dendritic cell	653	164	489
Myofibroblast	Regulatory T cell	472	478	6
Myofibroblast	Macrophage	428	254	174
Dendritic cell	Goblet cell	372	158	214
Dendritic cell	Myofibroblast	253	250	3
Dendritic cell	Regulatory T cell	355	343	12
Dendritic cell	Macrophage	299	182	117
Regulatory T cell	Goblet cell	427	183	244
Regulatory T cell	Myofibroblast	304	286	18
Regulatory T cell	Dendritic cell	610	146	464
Regulatory T cell	Macrophage	365	217	148
Macrophage	Goblet cell	622	272	350
Macrophage	Myofibroblast	456	428	28
Macrophage	Dendritic cell	865	218	647
Macrophage	Regulatory T cell	450	486	36









Supplementary Table 2.2: Top ten overrepresented pathways in upstream Treg signalling network

Healthy condition			Non-inflamed UC		
Pathway	Entities found	Entities total	Pathway	Entities found	Entities total
Signaling by RAS mutants	30	54	MyD88-independent TLR4 cascade	62	97
Signaling by moderate kinase activity BRAF mutants	30	54	TRIF(TICAM1)-mediated TLR4 signaling	62	97
Signaling downstream of RAS mutants	30	54	Toll Like Receptor 3 (TLR3) Cascade	61	93
Paradoxical activation of RAF signaling by kinase inactive BRAF	30	54	VEGFA-VEGFR2 Pathway	62	98
Oncogenic MAPK signaling	43	93	Transcriptional Regulation by TP53	163	367
Toll Like Receptor TLR6:TLR2 Cascade	52	118	Cell Cycle	228	651
MyD88:MAL(TIRAP) cascade initiated on plasma membrane	52	118	Diseases of signal transduction by growth factor receptors and second messengers	168	393
Toll Like Receptor 2 (TLR2) Cascade	52	121	Signaling by NTRK1 (TRKA)	69	117
Toll Like Receptor 7/8 (TLR7/8) Cascade	47	103	RNA Polymerase II Transcription	390	1379
Toll Like Receptor TLR1:TLR2 Cascade	52	121	Generic Transcription Pathway	359	1257

Supplementary table 5.1: Number of bacterial proteins derived from UniProt Proteome		
Condition	Strain	Number of proteins
Healthy gum	<i>Streptococcus sanguinis</i> SK36	2269
	<i>Haemophilus parainfluenzae</i> ATCC 33392	2010
	<i>Lautropia mirabilis</i> ATCC 51599	2665
	<i>Veillonella parvula</i> ATCC 10790	1843
Periodontitis	<i>Treponema denticola</i> ATCC 35405	2753
	<i>Porphyromonas gingivalis</i> ATCC BAA-308	1863
	<i>Tannerella forsythia</i> ATCC 43037	2978
	<i>Filifactor alocis</i> ATCC 35896	1616

Appendix 2: Peer-reviewed publications

Integrated intra- and intercellular signaling knowledge for multicellular omics analysis

Dénes Türei¹ , Alberto Valdeolivas¹ , Lejla Gul², Nicolás Palacio-Escat^{1,3,4} , Michal Klein⁵ ,
 Olga Ivanova¹ , Márton Ölbei^{2,6} , Attila Gábor¹ , Fabian Theis^{5,7} , Dezső Módos^{2,6} ,
 Tamás Korcsmáros^{2,6}  & Julio Saez-Rodriguez^{1,3,*} 

Abstract

Molecular knowledge of biological processes is a cornerstone in omics data analysis. Applied to single-cell data, such analyses provide mechanistic insights into individual cells and their interactions. However, knowledge of intercellular communication is scarce, scattered across resources, and not linked to intracellular processes. To address this gap, we combined over 100 resources covering interactions and roles of proteins in inter- and intracellular signaling, as well as transcriptional and post-transcriptional regulation. We added protein complex information and annotations on function, localization, and role in diseases for each protein. The resource is available for human, and via homology translation for mouse and rat. The data are accessible via *OmniPath's* web service (<https://omnipathdb.org/>), a Cytoscape plug-in, and packages in R/Bioconductor and Python, providing access options for computational and experimental scientists. We created workflows with tutorials to facilitate the analysis of cell–cell interactions and affected downstream intracellular signaling processes. *OmniPath* provides a single access point to knowledge spanning intra- and intercellular processes for data analysis, as we demonstrate in applications studying SARS-CoV-2 infection and ulcerative colitis.

Keywords intercellular signaling; ligand–receptor interactions; omics integration; pathways; signaling network

Subject Categories Computational Biology; Methods & Resources; Signal Transduction

DOI 10.15252/msb.20209923 | Received 14 August 2020 | Revised 11 February 2021 | Accepted 15 February 2021

Mol Syst Biol. (2021) 17: e9923

Introduction

Cells process information by physical interactions of molecules. These interactions are organized into an ensemble of signaling

pathways that are often represented as a network. This network determines the response of the cell under different physiological and disease conditions. In multicellular organisms, the behavior of each cell is regulated by higher levels of organization: the tissue and, ultimately, the organism. In tissues, multiple cells communicate to coordinate their behavior to maintain homeostasis. For example, cells may produce and sense extracellular matrix (ECM), and release enzymes acting on the ECM as well as ligands. These ligands are detected by receptors in the same or different cells, that in turn trigger intracellular pathways that control other processes, including the production of ligands and the physical binding to other cells. The totality of these processes mediates the intercellular communication in tissues. Thus, to understand physiological and pathological processes at the tissue level, we need to consider both the signaling pathways within each cell type as well as the communication between them.

Since the end of the nineties, databases have been collecting information about signaling pathways (Xenarios *et al.*, 2000). These databases provide a unified source of information in formats that users can browse, retrieve, and process. Signaling databases have become essential tools in systems biology and to analyze omics data. A few resources provide ligand–receptor interactions (Kirouac *et al.*, 2010; Fazekas *et al.*, 2013; Ramiłowski *et al.*, 2015; Armstrong *et al.*, 2019; Efremova *et al.*, 2020). However, their coverage is limited, they do not include some key players of intercellular communication such as matrix proteins or extracellular enzymes, and they are not integrated with intracellular processes. This is increasingly important as new techniques allow us to measure data from single cells, enabling the analysis of inter- and intracellular signaling. For example, the recent *CellPhoneDB* (Efremova *et al.*, 2020) and *ICELNET* (Noël *et al.*, 2021) tools provide computational methods to prioritize the most likely intercellular connections from single-cell transcriptomics data, and *NicheNet* (Browaeys *et al.*, 2019) expands this to intracellular gene regulation. A comprehensive resource of inter- and intracellular signaling knowledge would enhance and expedite these analyses.

1 Faculty of Medicine and Heidelberg University Hospital, Institute of Computational Biomedicine, Heidelberg University, Heidelberg, Germany

2 Earlham Institute, Norwich, UK

3 Faculty of Medicine, Joint Research Centre for Computational Biomedicine (JRC-COMBINE), RWTH Aachen University, Aachen, Germany

4 Faculty of Biosciences, Heidelberg University, Heidelberg, Germany

5 Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany

6 Quadram Institute Bioscience, Norwich, UK

7 Department of Mathematics, Technical University of Munich, Garching, Germany

*Corresponding author. Tel: +49 6221 5451334; E-mail: julio.saez@uni-heidelberg.de

To effectively study multicellular communication, a resource should (i) classify proteins by their roles in intercellular communication, (ii) connect them by interactions from the widest possible range of resources, and (iii) integrate all this information in a transparent and customizable way, where the users can select the resources to evaluate their quality and features, and adapt them to their context and analyses. Prompted by the lack of comprehensive efforts addressing principle (i), we built a database on top of *OmniPath* (Túrei et al, 2016), a resource which has already shown the benefits of principles (ii) and (iii). The first version of *OmniPath* focused on literature curated intracellular signaling pathways. It has been used in many computational projects and omics studies. For example, to model cell senescence from phosphoarray data (An et al, 2020), or as part of a computational pipeline to predict the effect of microbial proteins on human genes (Andrighetti et al, 2020), and a community effort to integrate knowledge about the COVID-19 disease mechanism (Ostaszewski et al, 2020). The new *OmniPath* extends its scope to intercellular communication and its integration with intracellular signaling, providing prior knowledge for modeling and analysis methods. It combines 103 resources to build an integrated database of molecular interactions, enzyme-PTM (*post-translational modification*) relationships, protein complexes and annotations about intercellular communication, and other functional attributes of proteins.

We demonstrate with two case studies that we provide a versatile resource for the analysis of single-cell and bulk omics data. Leveraging the intercellular communication knowledge in *OmniPath*, we present two examples where autocrine and paracrine signaling are key parts of pathomechanism. First, we studied the potential influence of ligands secreted in severe acute respiratory syndrome coronavirus 2 (*SARS-CoV-2*) infection on the inflammatory response through autocrine signaling. We identified signaling mechanisms that may lead to the dysregulated inflammatory and immune response shown in severe cases. Second, we examined the rewiring of cellular communication in *ulcerative colitis* (UC) based on single-cell data from the colon. By analyzing downstream signaling from the intercellular interactions, we found pathways associated with the regulatory T cells targeted by myofibroblasts in UC.

Results

We used four major types of resources: (i) molecular interactions, (ii) enzyme-PTM relationships, (iii) protein complexes, and (iv) molecule annotations about function, localization, and other attributes (Fig 1A). The *ppath* Python package combined the resources from those four types to build four corresponding integrated databases. Using the *annotations*, *ppath* compiled a fifth database about the roles in intercellular communication (*intercell*; Fig 1B). The ensemble of these five databases is what we call *OmniPath*, combining data from 103 resources (Fig 1A and Dataset EV1).

A focus on intercellular signaling

To create a database of intercellular communication, we defined the roles that proteins play in this process. Ligands and receptors are main players of intercellular communication. Many other kinds of

molecules have a great impact on the behavior of the cells, such as matrix proteins and transporters (Fig 2A). We defined eight major (Fig 2) and 17 minor generic functional categories of intercellular signaling roles (Datasets EV6 and EV10). We also defined ten locational categories (e.g., *plasma membrane peripheral*), using in addition structural resources and prediction methods to annotate the transmembrane, secreted and peripheral membrane proteins. Furthermore, we provide 994 specific categories (e.g., *neurotrophin receptors*). Each generic category can be accessed by resource (e.g., *ligands from HGNC*) or as the combination of all contributing resources (Fig EV4). To provide highly curated annotations, we checked every entry in each category manually against UniProt datasets to exclude wrong annotations. Overall we defined 1,170 categories and provided 54,330 functional annotations about intercellular communication roles of 5,781 proteins.

We collected the proteins for each intercellular communication functional category using data from 27 resources (Fig 2B, Dataset EV6). Combining them with molecular interaction networks from 48 resources (Dataset EV2), we created a corpus of putative intercellular communication pathways (Fig 2C). To have a high coverage on the intercellular molecular interactions, we also included ten resources focusing on ligand-receptor interactions (Figs 3 and EV1).

Many of the proteins in intercellular communication work as parts of complexes. We therefore built a comprehensive database of protein complexes and inferred their intercellular communication roles: a complex belongs to a category if and only if all members of the complex belong to it. We obtained 14,348 unique, directed transmitter-receiver (e.g., ligand-receptor) connections, around seven times more than the largest of the resources providing such kind of data. We also mapped a textbook table (Cameron & Kelvin, 2013) of 131 cytokine-receptor interactions to the ligand-receptor resources. As the textbook contains well-known interactions, many of the resources cover more than 90% of them (Fig 2D). This large coverage is achieved by not only integrating ten ligand-receptor resources, but also complementing these with data from annotation and interaction resources.

An essential feature of this novel resource is that it combines knowledge about intercellular and intracellular signaling (Table 1). Thus, using *OmniPath* one can, for example, easily analyze the intracellular pathways triggered by a given ligand or check the transcription factors (TFs) and microRNAs (miRNAs) regulating the expression of such ligands.

OmniPath: an ensemble of five databases

The abovementioned intercellular database exists in *OmniPath* together with four further databases (Fig 1B), supporting an integrated analysis of inter- and intracellular signaling.

The network of molecular interactions

The *network* database part covers four major domains of molecular signaling: (i) protein-protein interactions (PPI), (ii) transcriptional regulation of protein-coding genes, (iii) miRNA-mRNA interactions, and (iv) transcriptional regulation of miRNA genes (TF-miRNA). We further differentiated the PPI data into four subsets based on the interaction mechanisms and the types of supporting evidence: (i) literature curated activity flow (directed and signed; corresponds to the original release of *OmniPath*; Túrei et al, 2016), (ii) activity flow

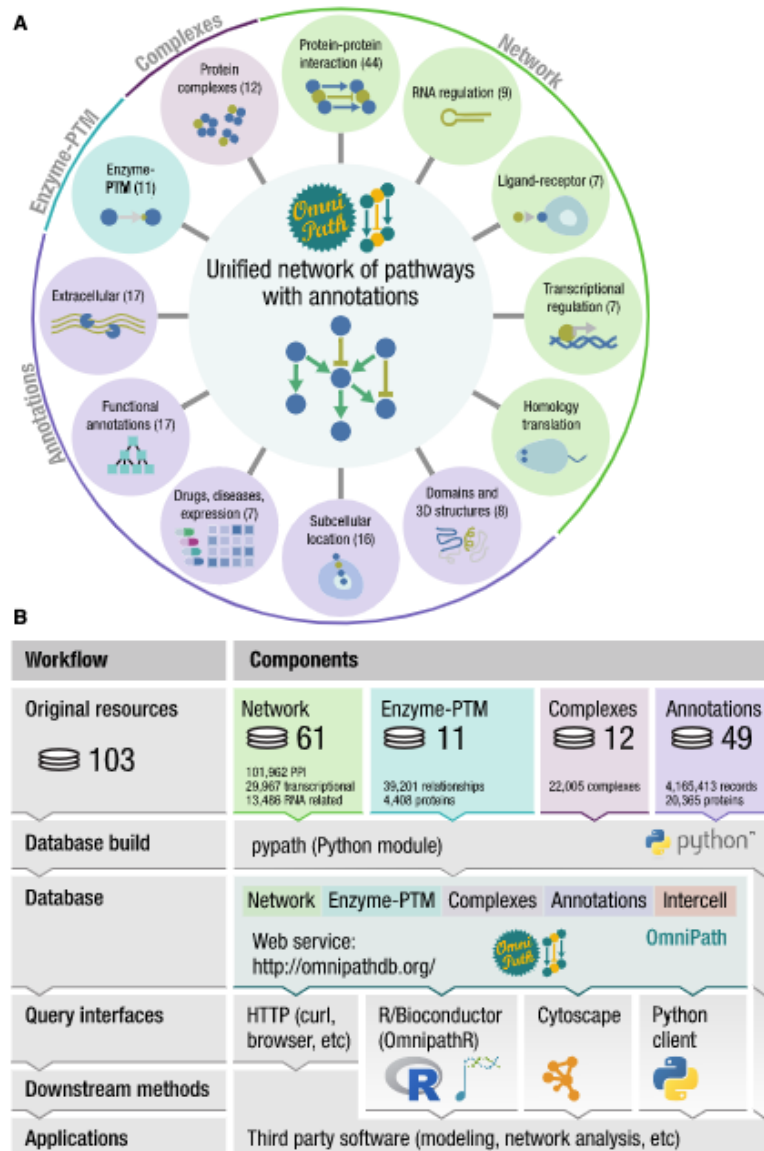


Figure 1. The composition and workflow of OmniPath.

A Database contents with the respective number of resources in parentheses.

B Workflow and design: OmniPath is based on four major types of resources, and the pypath Python package combines the resources to build five databases. The databases are available by the database builder software pypath, the web resource at <https://omnipathdb.org/>, the R package OmnipathR, the Python client omnipath, the Cytoscape plug-in and can be exported to formats such as Biological Expression Language (BEL).

with no literature references, (iii) enzyme-PTM, and (iv) ligand-receptor interactions (Fig 3A-C). Interaction data are extensively used for a variety of purposes: for building mechanistic models, deriving pathway and TF activities from transcriptomics data and graph-based analysis methods. In total, the resource contained

103,396 PPI interactions between 12,469 proteins from 38 original resources (Dataset EV2). The large number of unique interactions added by each resource underscores the importance of their integration (Figs EV1 and EV2, Appendix Fig S1). The interactions with effect signs, essential for mechanistic modeling, are provided by the

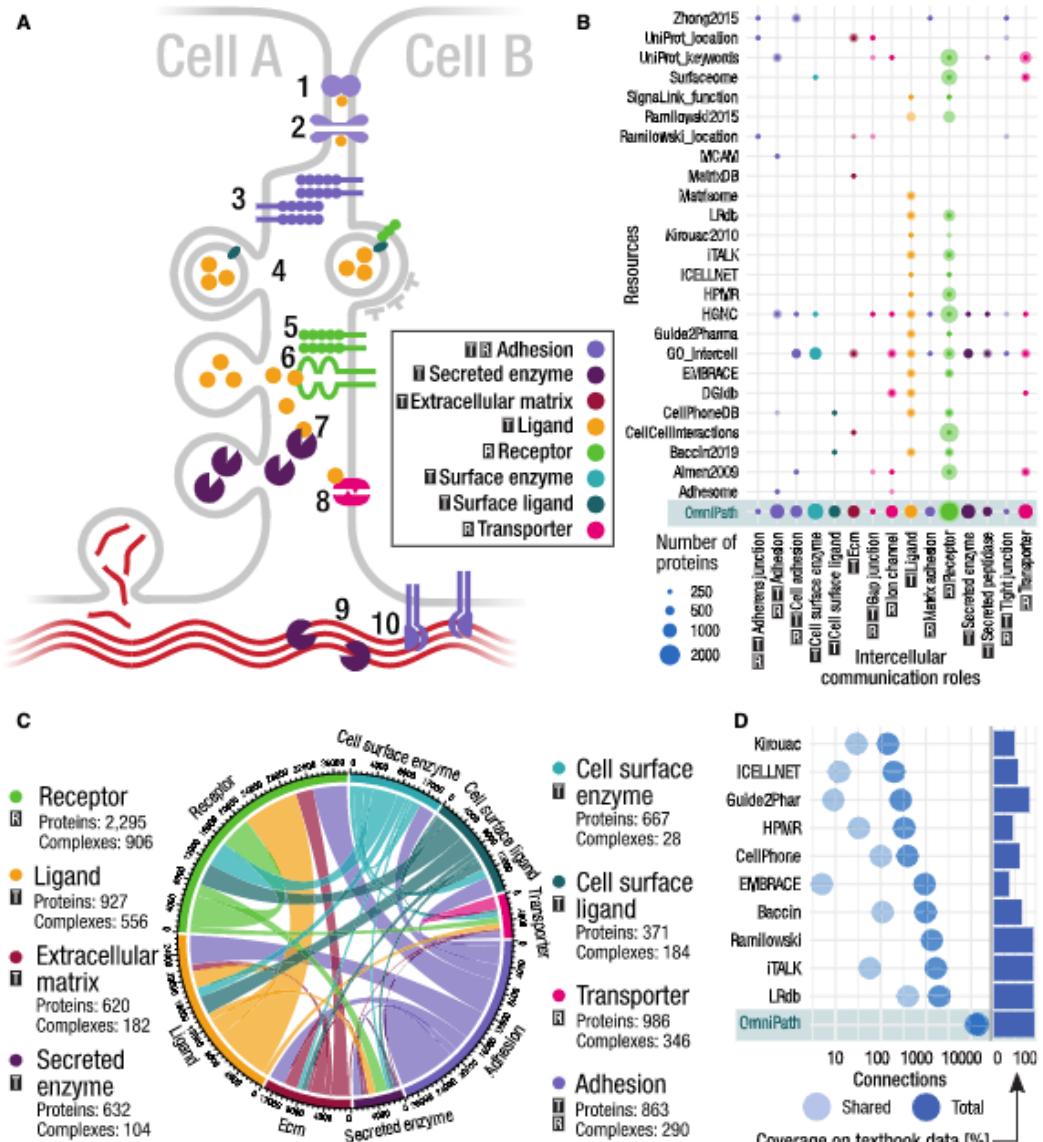


Figure 2.

Figure 2. The composition and representation of the intercellular signaling network.
 We assigned intercellular communication roles to proteins based on evidence from multiple resources. In all panels **A**—transmitter; **B**—receiver.
A Schematic illustration of the intercellular communication roles and their possible connections. Cells are physically connected by proteins forming tight junctions (1), gap junctions (2), and other adhesion proteins (3); they release vesicles which can be taken up by other cells (4), some receptors form complexes (5) to detect secreted ligands (6), transporters might also be affected by factors released by other cells (8); enzymes released into the extracellular space act on ligands and the extracellular matrix (7, 9); cells release the components of the extracellular matrix and bind to the matrix by adhesion proteins (10).
B The main intercellular communication roles (x axis) and the major contributing resources (y axis). Size of the dots represents the number of proteins annotated to have a certain role in a given resource. The darker areas represent the overlaps (proteins annotated in more than one resource for the same role) while the lighter color denotes those unique to that resource.
C The intercellular communication network. The circle segments represent the eight main intercellular communication roles. The edges are proportional to the number of interactions in the OmniPath PPI network connecting proteins of one role to the other.
D Number of unique, directed transmitter-receiver [e.g., ligand-receptor] connections by resources. Bars on the right show the coverage of each resource on a textbook dataset of 131 well-known ligand-receptor interactions.

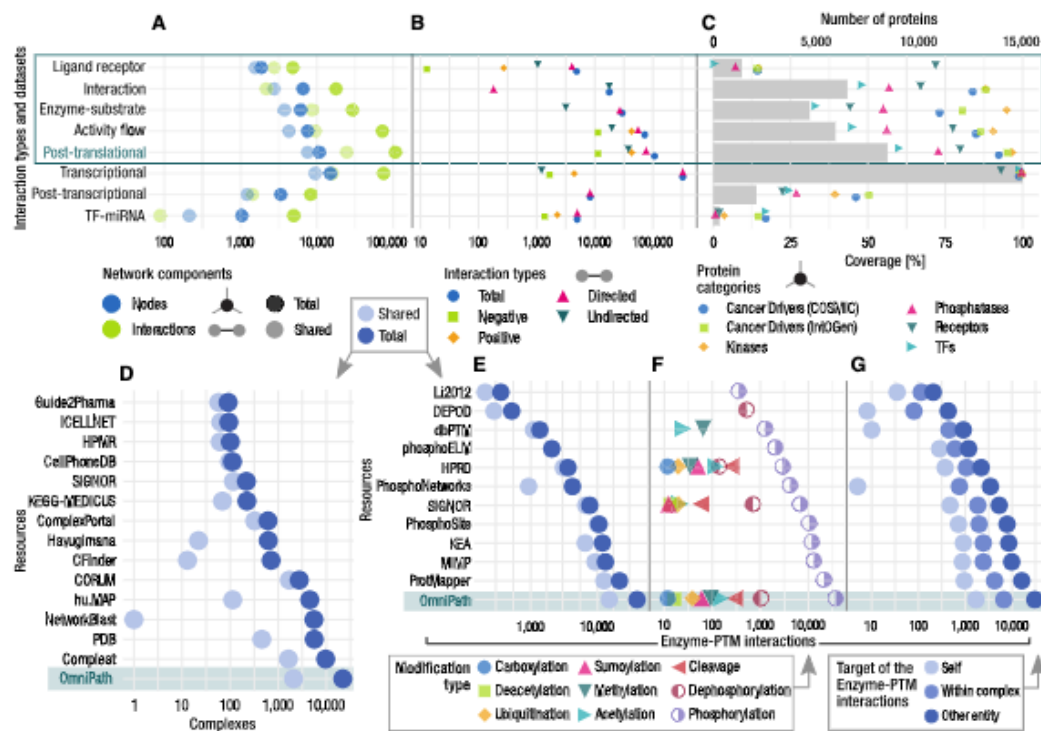


Figure 3. Quantitative description of the network, complex, and enzyme-PTM databases.
A–C Networks by interaction types and the network datasets within the PPI network. (A) Number of nodes and interactions. The light dots represent the shared nodes and edges (in more than one resource), while the dark ones show their total numbers. (B) Causality: number of connections by direction and effect sign. (C) Coverage of the networks on various groups of proteins. Dots show the percentage of proteins covered by network resources for the following groups: cancer driver genes from COSMIC and IntOGen, kinases from kinase.com, phosphatases from Phosphatome.net, receptors from the Human Plasma Membrane Receptorome (HIPMR) and transcription factors from the TF census. Gray bars show the number of proteins in the networks. The information for individual resources is in Figs EV1 and EV2, Appendix Fig S1.
D–G On each panel, the bottom rows represent the combined complex and enzyme-PTM databases contained in OmniPath (D, E). Number of complexes (D) and enzyme-PTM (E) interactions by resource. (F) Enzyme-PTM relationships by PTM type. (G) Enzyme-PTM interactions by their target. Light, medium, and dark dots represent the number of enzyme-PTM relationships targeting the enzyme itself, another protein within the same molecular complex or an independent protein, respectively.

Table 1. Qualitative comparison of ligand-receptor and integrative databases.

Resource	Interactions	Directed interactions	Signs (positive/negative)	Transcriptional regulation	Intracellular pathways	Intercellular communication roles	Protein complexes	Integrative resource	Literature curated
Baccin2019 (e)	yes	yes (a)	no	no	no	yes (f)	yes	yes	yes (g)
CellCellInteractions	yes	yes (a)	no	no	no	yes (f)	no	yes	no
CellPhoneDB	yes	yes (a)	no	no	no	yes (d)	yes	yes	yes
ConsensusPathDB	yes	no	no	yes	yes	no	no	yes	yes (g)
EMBRACE (e)	yes	yes (a)	no	no	no	yes	no	yes (k)	yes (g)
HPMR	yes	yes (a)	no	no	no	yes	no	no	yes
ICELNET	yes	yes (a)	no	no	no	yes	yes	no	yes
ITALK (h)	yes	yes (a)	no	no	no	yes	no	yes	yes (g)
Kirouac2010	yes	yes (a)	no	no	no	yes	no	no	yes
LRdb	yes	yes (a)	no	no	no	yes	no	yes	yes (g)
PathwayCommons	yes	yes (m)	no	yes	yes	no	yes	yes	yes (g)
Ramilowski2015	yes	yes (a)	no	no	no	yes	no	yes	yes (g)
SignalLink	yes	yes	yes	yes (i)	yes	yes	no	yes (j)	yes (g)
OmniPath	yes	yes (b)	yes	yes	yes	yes (c)	yes	yes	yes (g)

OmniPath combines resources to build a network with directions and effect signs, including intra- and intercellular signaling, transcriptional regulation, and annotates proteins as ligands or receptors. Here, we show which of these features are covered by other databases: those specialized in ligand-receptor interactions and two large integrative network databases (ConsensusPathDB and Pathway Commons). (a) Implicit: if we assume always the ligand affects the receptor; (b) As in some of the constituent resources the directions are implicit, certain directions in the combined network are implicit; (c) Provides not only ligand and receptor annotation but further categories, for example adhesion, transporter, ECM, etc; (d) Apart from secreted (mostly ligand) and receptor provides a few further categories: integrin, collagen, transmembrane, peripheral, etc; (e) Data are for mouse, homology translation is necessary to derive human data; (f) For ligands, provides certain classification, e.g., cytokine, ECM, secreted, etc; (g) Only in part is literature curated; (h) Ligand-receptor interactions are classified as growth factor, cytokine, checkpoint, or other; (i) Contains transcriptional regulation but that part is not integrated by OmniPath; (j) OmniPath only integrates its original literature curation, not the secondary resources; (k) Only builds on Ramilowski et al; (l) Besides ligand and receptor only ECM; (m) Directionality information might be extracted from BioPAX.

activity flow resources (Appendix 1; Fig 3B). The combined PPI network covered 53% of the human proteome (SwissProt), with an enrichment of kinases and cancer driver genes (Fig 3C). The transcriptional regulation data in *OmniPath* were obtained from *DoRothEA* (Garcia-Alonso et al, 2019), a comprehensive resource of TF regulons integrating data from 18 sources. In addition, six literature curated resources were directly integrated into *OmniPath* (Dataset EV8). The miRNA-mRNA and TF-miRNA interactions were integrated from five and two literature curated resources, with 6,213 and 1,803 interactions, respectively. Combining multiple resources not only increases the coverage, but also improves quality. It makes it possible to select higher confidence records based on the number of resources and references. Cross-checking the interaction directions and effect signs between resources reveal contradictory information which is either a sign of mistakes or reflects on limitations of our data representation (Appendix 1; Appendix Figs S4). Overall, we included 61 network resources in *OmniPath* (Dataset EV2). Furthermore, *pypath* provides access to additional resources, including the *Human Reference Interactome* (Luck et al, 2020), *ConsensusPathDB* (Kamburov et al, 2013), *Reactome* (Jassal et al, 2020), *ACSN* (Kuperstein et al, 2015), and *WikiPathways* (Slenter et al, 2018).

Enzyme-PTM relationships

In enzyme-PTM relationships, enzymes (e.g., kinases) alter specific residues of their substrates, producing so-called post-translational modifications (PTM). Enzyme-PTM relationships are essential for

deriving networks from phosphoproteomics data or estimating kinase activities. We combined 11 resources of enzyme-PTM relationships mostly covering phosphorylation (94% of all) and dephosphorylations (3%) (Fig 3F). Overall, we included 39,201 enzyme-PTM relationships, 1,821 enzymes targeting 16,467 PTM sites (Fig 3E-G). Besides phosphorylation and dephosphorylation, only proteolytic cleavage and acetylation account for more than one hundred interactions. Most of the databases curated only phosphorylation, and *DEPOD* (Damle & Köhn, 2019) exclusively dephosphorylation. Only *SIGNOR* (Licata et al, 2020) and *HPRD* (Keshava Prasad et al, 2009) contained a large number of other modifications (Fig 3F). 60% of the interactions were described by only one resource, and 92% of them by only one literature reference (Fig 3E). Self-modifications, e.g., autophosphorylation and modifications between members of the same complex comprised 4 and 18% of the interactions, respectively (Fig 3G).

Protein complexes

Many proteins operate in complexes, for example, receptors often detect ligands in complexes. To facilitate analyses taking into consideration complexes, we added to *OmniPath* a comprehensive collection of 22,005 protein complexes described by 12 resources from 4,077 articles (Fig 3D). A complex is defined by its combination of unique members. 14% of them were homomultimers, 54% had four or less unique components while 20% of them had 18 or more. 71% of the complexes had stoichiometry information.

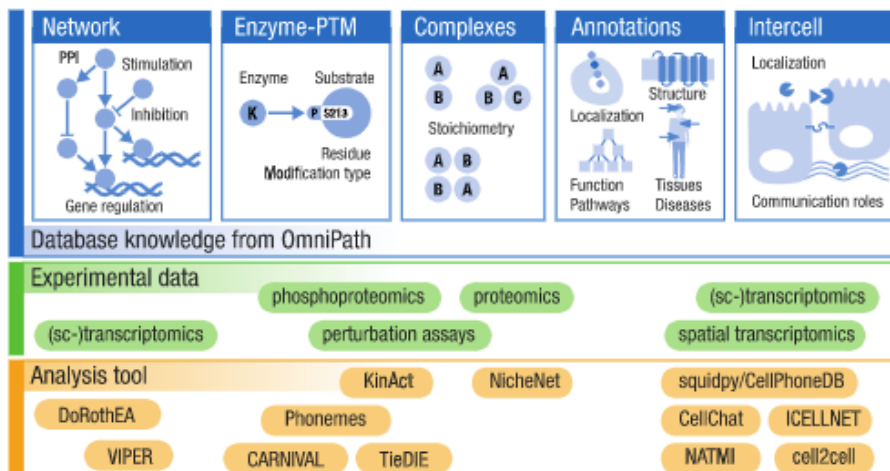


Figure 4. Examples of tools for omics data analysis that can be applied with the prior knowledge available in OmniPath

Annotations: function, structure and localization

Annotations provide information about the function, structure, localization, classification, and other properties of molecules. We compiled the *annotations* database from 49 resources. The format of the records from each of these resources is different. The simplest ones only define a category of proteins, like *Cell Surface Protein Atlas (CSPA)* (Bausch-Fluck et al, 2015) that collects the proteins localized on the cell surface. More complex annotation records express a combination of multiple attributes. For example, each of the annotations from the *Cancer Pathway Association Database (CPAD)* (Li et al, 2020) contain seven attributes to describe a relationship between a protein or miRNA, a pathway, and their effect on a specific cancer type (Fig EV3). The pathway and gene sets are also part of the annotation database, as these are useful for functional characterization of omics data and enrichment analysis.

Overall, the *annotations* database included 5,475,532 records about 20,365 proteins, virtually the whole protein-coding genome, 19,566 complexes, and 182 miRNAs. The majority of the annotations for complexes were the result of our *in silico* inference: If all members of a complex share a certain annotation, we assign this annotation to the complex itself.

The *annotations* database can be used in different ways: Selecting one resource, its data can be reconstituted into a conventional data frame with attributes as columns and annotations as rows. Alternatively, specific sets of proteins can be queried, e.g., "the members of the *Notch pathway* according to *SIGNOR*" (Licata et al, 2020) or "the *hypoxia upregulated genes* according to *MSigDB*" (Subramanian et al, 2005). The annotations are helpful in omics data analysis; for example, they can be used for contextualization or enrichment analysis.

Homology translation to rodents

OmniPath comprises human resources. We translated the network and the enzyme-PTM relationships to mouse and rat by protein

homology using *NCBI HomoloGene*, covering 81 and 31% of the interactions for mouse and rat, respectively (Dataset EV9). In addition, *pypath* is able to translate to other organisms.

Close connection to the analysis of omics data

The *OmniPath* databases are built by the *pypath* Python module and are distributed by the web service at <https://omnipathdb.org/>. We provide web service clients in R, Python, and Cytoscape (Ceccarelli et al, 2019). The clients not only query the *OmniPath* data but also offer convenient post-processing methods and integration with other software (Figs 1B and 4). The *OmniPathR* R client implements a full integration with *NicheNet*, a method for prioritizing ligands affecting cells based on transcriptomics data (Browaeys et al, 2019): A single *OmniPathR* function can be used to generate all inter- and intracellular knowledge required for *NicheNet*. The *omnipath* Python module, together with the single-cell data processing *scanpy* module (Wolf et al, 2018) and the *squidpy* reimplementation of the *CellPhoneDB* algorithm to infer ligand-receptor interactions between cell types (Efremova et al, 2020), provides an easy and efficient way to analyze intercellular communication. These applications and further examples are available as detailed tutorials at <https://workflows.omnipathdb.org/>. Here, a number of guides are available demonstrating various features of *OmniPath*, presenting the query parameters of the databases and showcasing downstream workflows.

Case studies

OmniPath provides a single-access point to resources covering diverse types of knowledge. Thus, it can be used as an input to many analysis tools and is particularly useful for applications that span over molecular processes typically considered separately (Fig 4). To illustrate this, we used two examples where we extracted

from *OmniPath* different types of intra- and intercellular knowledge for computational analysis of bulk and single-cell RNA-Seq data.

Analysis of intra- and intercellular processes in

SARS-CoV-2-infected lung epithelial cancer cells

NicheNet is a recently developed method to prioritize ligand–target relationships between interacting cells by combining their expression data with prior knowledge on interaction networks (Browaeys et al, 2019). For this purpose, *NicheNet* explores the most consistent inter- and intracellular protein interactions in accordance with a given gene expression dataset. In the *NicheNet* publication, the authors collected different types of interactions from more than 20 databases to build a ligand–receptor network, a signaling network, and a gene regulatory network. Here, we built a network for analysis with *NicheNet* using exclusively *OmniPath*.

We used the *OmniPath* built network to investigate the mechanistic processes leading to the excessive inflammatory response and dysregulated adaptive host immune defense that may occur in severe *COVID-19* cases (Catanzaro et al, 2020). We studied the autocrine regulatory effect of ligands secreted in *SARS-CoV-2* infection of epithelial lung cancer cells (*Calu3*; Methods and Appendix 2; data from Blanco-Melo et al, 2020). Out of 117 ligands over-expressed in *SARS-CoV-2* infection, we selected for subsequent analysis the 12 best predictors of inflammatory response genes according to the distribution of correlation values (Fig EV5B) and *nichenetr* guidelines (Methods and Appendix 2).

Among them, we found various cytokines: interleukins (*IL23A* and *IL1A*), tumor necrosis factors (*TNF* and *TNFSF13B*), and chemokines (*CXCL5*, *CXCL9*, and *CXCL10*), known to be involved in the inflammatory response. *NicheNet* scores describing the potential influence of the 12 selected ligands on the set of inflammatory genes are significantly higher than on sets of randomly selected genes (average *P*-value = 3.25e-08 from Fisher's exact tests after 10 cross-validation rounds). Then, we explored the signaling events linking these ligands to their target genes (Fig 5A, Methods and Appendix 2). We identified several key proteins of the *JAK-STAT* pathway (*JAK2*, *STAT1*, *STAT3*, and *STAT4*), a main regulator of the inflammatory response, that has been suggested as a potential target to treat *COVID-19* (Bagca & Avci, 2020). We also found ligands that potentially trigger the *MAPK* pathway that has also been reported to be promoted by *SARS-CoV-2* infection (Bouhaddou et al, 2020; Treveil et al, 2021). To further characterize the potential medical relevance of these results, we investigated the drugs targeting the genes shown in Fig 5A (Dataset EV14). Among the most interesting results, we identified minocycline, an antibiotic, and anti-inflammatory drug targeting *CASP3* and *TNF*. Minocycline has been very recently proposed to alleviate the effects of *SARS-CoV-2* severe infection in the central nervous system (Oliveira et al, 2020) (see extended results in Appendix 2).

In summary, we found mechanistic insights about inflammatory-related signaling cascades triggered by *SARS-CoV-2* infection. The underlying interactions spanned different curated (and thus supported by literature) individual inter- and intracellular resources that we could leverage as they are all integrated in *OmniPath* (Fig 5A in Dataset EV13).

Alteration of intercellular communication in ulcerative colitis

As a second case study, we used single-cell RNA-Seq data (Smillie et al, 2019) from *ulcerative colitis* (UC) to investigate paracrine

signaling using *OmniPath*'s intra- and intercellular knowledge. UC is an inflammatory bowel disease (IBD) driven by an interplay of epithelial cells and resident mucosal immune cells. Hence, it would be desirable to investigate it with considering both cell type-specific intracellular signaling and cell–cell communication.

We limited our analysis to five cell types relevant in UC: dendritic cell (DC), macrophage, regulatory T cell (Treg), myofibroblast, and Goblet cell. We combined the cell type and condition-specific expression data with *OmniPath* to build intracellular and intercellular signaling networks (Appendix Fig S5). The total number of cell–cell connections was similar (Table EV1), while their identity and distribution were different between healthy and UC conditions. In healthy condition, all cell types were tightly connected to DCs while in UC to Treg cells (Fig 5B).

Using the *intercell* annotation database of *OmniPath*, we examined the type of intercellular interactions between these cell types. We found that in both healthy state and UC the ligand–receptor and adhesion connections were dominant and the cell junction type connections were less abundant in UC—which was expected due to the pathophysiology of the disease. Also in UC, we found a higher amount of ligand–receptor and adhesion connections between Treg cells and the other four cell types, supporting previously described alteration of Treg signaling in UC (Yamada et al, 2016).

To analyze the changes in Treg signaling more in detail, we combined the intercellular and intracellular databases from *OmniPath* and focused on the connection between myofibroblasts and Treg cells. The total number of intercellular connections are nearly the same in healthy and in UC conditions 472 and 478, respectively. However, the actual interacting proteins and their downstream effects are remarkably different (Fig 5C). This is mainly due to ligands from myofibroblasts or receptors on Treg cells expressed uniquely in one of the conditions. For example TGF-beta signaling is a known regulatory input of Treg cells (Wan & Flavell, 2008), and we found *BMPRI1A* and *ACVRL1*, two receptors for the TGF-beta pathway, to be specific for healthy and UC conditions, respectively. Although there is no evidence for the role of *ACVRL1* in Treg cells, the knockout of *Bmpr1a* contributes to gut inflammation (Shroyer & Wong, 2007). The changes in intercellular connections lead to major downstream signaling difference in Treg cells. To map the downstream effect, we built an intracellular network of Treg cells including two steps downstream of all recipient proteins targeted by myofibroblast effectors (Fig 5B). There were roughly two times more affected downstream proteins in Treg cells in UC than in healthy condition (835 versus 1,971), suggesting a wider regulatory impact of myofibroblasts on Treg cells. Using Reactome (Jassal et al, 2020) pathway enrichment analysis (Dataset EV11), we identified the main pathways in Treg cells affected differently by myofibroblasts in the two conditions. In healthy state, the *MAPK*, *Toll-like receptor (TLR) 2/6*, and *TLR7/8* pathways were enriched that are known as key processes regulating immunosuppressive functions and suppressing the proinflammatory Th17 cells (Forward et al, 2010; Nyirenda et al, 2015; He et al, 2018). Meanwhile in UC, *TLR4* and *TLR3* pathways were affected by myofibroblasts, and these pathways are relevant in UC as they regulate inflammatory cytokine expression and decrease the abundance of Treg cells (Xiao et al, 2009; Cao et al, 2014).

Our analysis supports the fact that the normally anti-inflammatory effect of Treg cells in UC is deteriorated partially by myofibroblasts

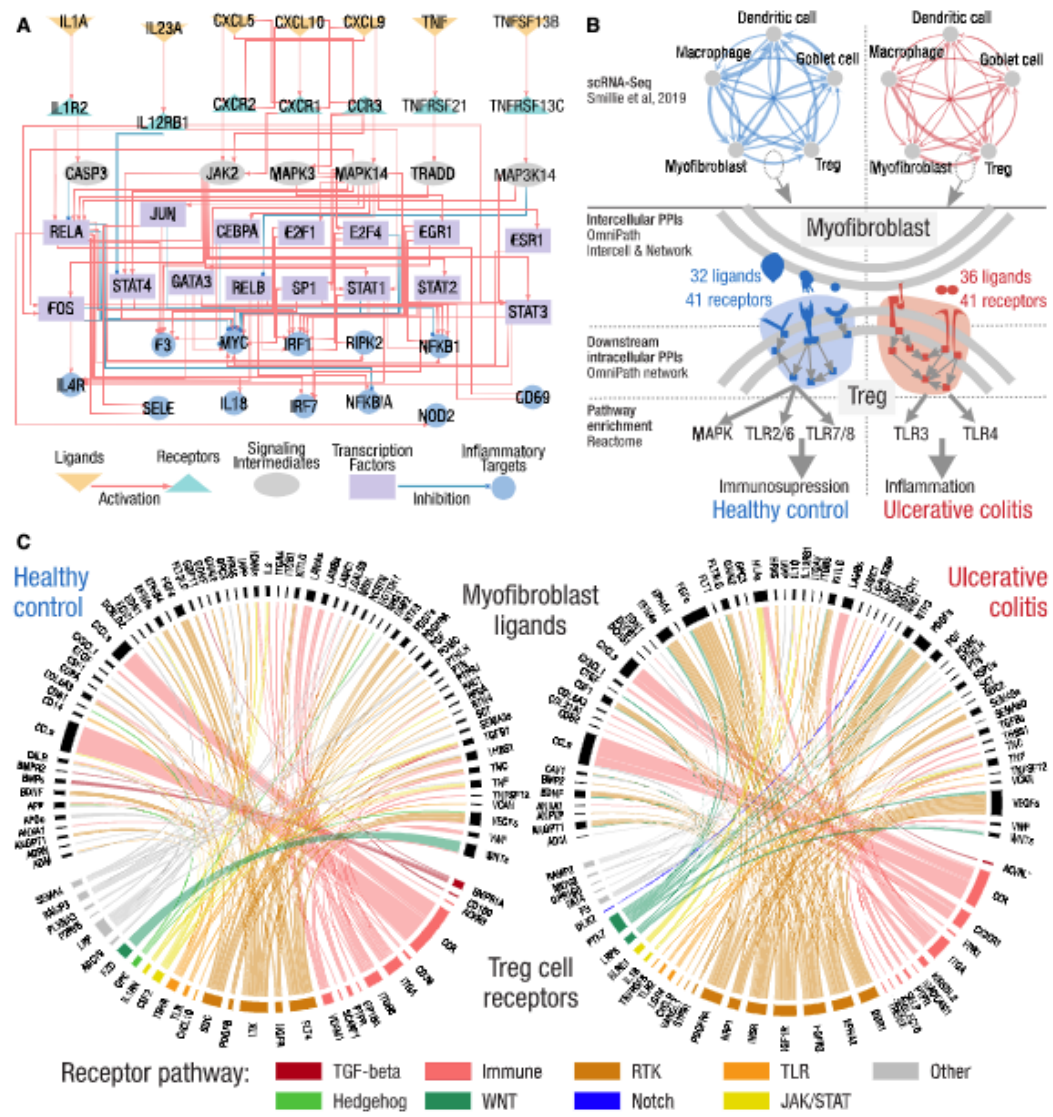


Figure 5. Illustrations of the integrated analysis of inter- and intracellular signaling.

A. Inter- and intracellular signaling interactions linking the top predicted ligands over-expressed after SARS-CoV-2 infection to their potential immune response targets in the Calu3 cell line. Top ranked ligands (orange) connect to their potential receptors (turquoise) that trigger an intracellular cascade until reaching TFs (purple), that in turn regulate the expression of the target genes (blue). Signaling intermediates (gray) connect receptors to TFs across their shortest path.

B. Intercellular connections and their downstream effect in UC compared with healthy control. Top: communication network of five cell types reconstructed from scRNA-Seq; the thickness of the edges is proportional to the number of intercellular connections. Bottom: condition-specific ligand-receptor connections between myfibroblasts and regulatory T cells trigger an immunosuppressive versus an inflammatory signaling in T cells, in healthy and UC, respectively.

C. Condition-specific connections between myfibroblast ligands (upper semicircles, black) and Treg cell receptors (lower semicircles, colored by pathways) in ulcerative colitis (right) and healthy control (left). Pathway annotations from SignalLink. Immune—innate immune response, RTK—receptor tyrosine kinase, TLR—Toll-like receptor.

Table 2. Number of unique receptors and their first two neighbors in each comparison.

Conditions	Pathway Commons network				OmniPath network	
	Ramilowski annotations		OmniPath annotations		Unique receptors	First two neighbors
	Unique receptors	First two neighbors	Unique receptors	First two neighbors		
Healthy	2 receptors	7553 proteins	13 receptors	9371 proteins	36 receptors	2476 proteins
Uninflamed UC	2 receptors	10138 proteins	6 receptors	11441 proteins	41 receptors	2879 proteins

(Takahashi et al, 2006; West, 2019). We found key intercellular mechanisms leading to well-defined differential pathway activation profiles. This was achieved via our novel approach to: (i) determine cell–cell interactions both in a healthy and in disease states and (ii) map affected downstream intracellular signaling processes based on the proteins interacting between cells.

Comparing Omnipath to other resources for cell–cell interaction analysis

The UC use case provided an opportunity to compare *OmniPath* against alternative sources of prior knowledge. We chose two widely used resources, *Pathway Commons (PC)* (Cerami et al, 2011) for network and *Ramilowski et al* (Ramilowski et al, 2015) for ligand–receptor annotations. Using the same workflow and expression data (Table 2), we investigated the myofibroblast–Treg cell interaction in three different network–annotation combinations (Appendix Fig S3): (i) *PC-Ramilowski*; (ii) *PC-OmniPath*; and (iii) *OmniPath-OmniPath* (i.e., as we presented in the use case).

Using the undirected *PC* network with the *Ramilowski* and the *OmniPath* annotation resulted in 523 and 3,136 ligand–receptor connections, respectively. The *OmniPath* PPI network with *OmniPath* annotations revealed 4,473 ligand–receptor connections indicating that this combination provides the largest coverage and more detailed data with directionality. In the intracellular network of Treg cells, using the *Ramilowski* annotation with *PC* network we found around 20 times less condition-specific receptors, compared with using *OmniPath* for both network and annotation, leading to a subsequent loss of downstream pathways in the former case. At the same time, the *PC* network with *OmniPath* annotations provided a large Treg cell downstream network containing ~ 50% of the *PC* network, while using the *OmniPath* network resulted in a three times smaller network, covering ~ 30% of the total *OmniPath* network. This is mainly due to the fact that *PC* provides a denser network than *OmniPath*, but undirected. Overall, *OmniPath* provides a high number of ligand–receptor interactions and directed interactions for downstream intracellular pathway analysis, complementing other meta-resources.

Discussion

In the first version of *OmniPath* (Türei et al, 2016), we built a comprehensive knowledge of intracellular signaling pathways with the aim of providing prior knowledge for modeling methods. Here, we present a major redesign and extension of this resource, offering a single-access point to over 100 resources containing prior knowledge of not only intra- but also intercellular processes. To achieve this, we developed versatile annotations of intercellular communication roles, combined with a network covering intra- and intercellular

signaling as well as gene regulation. By defining the *transmitter*, *receiver*, and *mediator* roles, we laid out a new conceptual framework to describe intercellular communication and generalized the terms of *ligand* and *receptor* (Dataset EV10). This framework allows *OmniPath* to combine diverse resources in a uniform way. In *OmniPath*, the intercellular annotations and the network connections are independent from each other, achieving together a great flexibility. As intercellular communication becomes increasingly popular thanks to single-cell technologies, we believe that supporting it with database knowledge deserves a dedicated effort instead of doing *ad hoc* data integration within each study.

While integrative resources such as *STRING* (Szklarczyk et al, 2019), *PathwayCommons* (Cerami et al, 2011), *ConsensusPathDB* (Kamburov et al, 2013), *PathMe*, and *ComPath* (Domingo-Fernández et al, 2019) use mostly the major process description resources (e.g., *Reactome* (Jassal et al, 2020) and *ACSN* (Kuperstein et al, 2015)) and resources with undirected interactions (e.g., *IntAct* (Orchard et al, 2014) and *BioGRID* (Oughtred et al, 2019)), the *network* database of *OmniPath* focuses on activity flow representation, providing a convenient input for multiple analysis techniques (Touré et al, 2020; Appendix 1). *OmniPath* is not limited to literature curated interactions and it also includes activity flow, kinase–substrate, and ligand–receptor interactions without references as separate datasets, so that the users can decide which ones to use according to their purposes (Dataset EV2). The rich annotations allow users to dive into specific knowledge and extract information across resources. The knowledge in *OmniPath* is general in terms of cell type or physiological condition. In the process of data analysis and modeling, omics data help to make the database knowledge more context specific. As an alternative, one can use for example *Human Protein Atlas* (Uhlén et al, 2015) in the *OmniPath* annotations database to build tissue specific networks (<https://workflows.omnipathdb.org/>).

As we demonstrated here, *OmniPath* is able to deliver the input knowledge for different data analysis tools, such as *CellPhoneDB* (Efremova et al, 2020), *NicheNet* (Browaeys et al, 2019), *CellChat* (Jin et al, 2021), *ICELNET* (Noël et al, 2021), *NATMI* (Hou et al, 2020), *cell2cell* (preprint: Armingol et al, 2020a), and *CARNIVAL* (Liu et al, 2019) to infer communication between (Armingol et al, 2020b) and within cell types. For some of the analysis tools, we provide dedicated software integration and workflows (<https://workflows.omnipathdb.org/>).

As our case studies illustrate, *OmniPath* can replace the tedious collection of information from many different databases. The first case study pointed to potential signaling mechanisms of autocrine origin in SARS-CoV-2 infection which can contribute to the dysregulated inflammatory and immune response characteristic of severe COVID cases. Our study is limited to the relationship of autocrine signaling and inflammatory response and hence it does not cover the complete process of viral infection. In the second study, we illustrated

how conveniently *OmniPath* supports a combined analysis of inter- and intracellular signaling from single-cell transcriptomics data. While multiple studies mapped intracellular signaling pathways to intestinal tissue, only a few of them were able to do it in a cell type-specific manner using single-cell transcriptomics data (Smillie et al, 2019). Due to the lack of integrated resources, combined intra- and intercellular studies have been so far challenging and not standardized. This is currently a major bottleneck to understand better conditions such as gut inflammation, which is modulated by the interplay of epithelial cells and resident mucosal immune cells. The results of the case studies can guide designing co-culture experiments by prioritizing the most relevant cell types and pointing out the key cell–cell interaction types. For example, testing the role of CASP3 in the autocrine signaling we pointed out in the first study, and the specific ligand–receptor connections that altered the intestinal paracrine signaling in diseased condition in the second case study. In general, the outcome of *OmniPath*-based analyses can define key candidates for more in depth investigations.

Over the past 4 years, we have kept developing *OmniPath*, adding new features and resources regularly. One of our main objectives for the future is to add more context information, e.g., cell type and physiological condition to the signaling network, and use scores to prioritize interactions and paths which contribute stronger to indirect causal relationships. Toward these aims, we plan to leverage text mining methods (Gyori et al, 2017; Kveler et al, 2018). We are also working on benchmarking the intercellular communication knowledge by deriving ground truth from experimental data (Armingol et al, 2020b). Furthermore, we envision to extend *OmniPath* with pathogen–host interactions (Treveil et al, 2021) and microbiome–host interactions (Andrighetti et al, 2020) in the near future.

In summary, we provide a new integrated resource of biological knowledge particularly valuable for network analysis and modeling of bulk and single-cell omics data. We anticipate that this knowledge will also be valuable to analyze the emergent spatially resolved omics data (Asp et al, 2020). To understand tissue architecture and function, it is crucial to study the spatial arrangement of the different cell types. Spatial transcriptomics technologies provide this information and hence help to prioritize the most likely ligand–receptor interactions. Fundamental questions about cell communication in tissues, such as how secreted ligands act on neighboring cells, can be addressed by analyzing spatially resolved data, combining data-driven (Sun et al, 2020; preprint: Tanevski et al, 2020) with prior knowledge-based (Browaays et al, 2019; Liu et al, 2019; Efremova et al, 2020) approaches. *OmniPath* provides a framework to support these endeavors.

Materials and Methods

Terminology

In the manuscript, we use consistently the following three definitions to describe the structure of *OmniPath*:

- **Database:** collection of similar records in a uniform format integrated from multiple resources (network, enzyme-PTM, complexes, annotations, intercell).

- **Dataset:** a subset or variant of a database, e.g., the transcriptional interaction network is a dataset of the network database.
- **Resource:** any data source we use for building the databases.

Database build

To build *OmniPath*, we developed a free software, the *pypath* Python module (<https://github.com/saezlab/pypath>, version 0.11.39). We built each segment of the database by the corresponding submodules and classes in *pypath*. In addition to the database building process, all modules rely on common utility modules from *pypath* such as the identifier translator or the downloading and caching service. *Pypath* downloads all data from the original sources. Many resources integrate data from other resources, we call these secondary resources and their relationships are listed in Dataset EV7.

Network

For the *OmniPath* network, we converted the identifiers of the different molecules and merged their pairwise connections, preserving the literature references, the information about the direction, and effect sign (activation or inhibition).

In *OmniPath*, we included nine network datasets built from 61 resources (Dataset EV2). The first four datasets provide PPI ("post_translational" in the web service) while the others transcriptional and post-transcriptional regulation. At each point below, we highlight the label of the dataset in the web service.

- 1 We compiled the "omnipath" network as described in Túrei et al (Túrei et al, 2016). Briefly, we combined all resources we could get access to, that are literature curated and are either activity flow, enzyme-PTM, or undirected interaction resources. We also added network databases with high-throughput data. Then, we added further directions and effect signs from resources without literature references.
- 2 The "kinaseextra" network contains additional kinase–substrate interactions without literature references. The direction of these interactions points from the enzyme to the substrate.
- 3 In the "pathwayextra" network, we combined further activity flow resources without literature references. However, they are manually curated and many have effect signs.
- 4 In the "ligreextra" network, we provide additional ligand–receptor interactions from large, comprehensive collections.
- 5 The "dorothea" network comes from DoRothEA database, a comprehensive resource of transcription factor–gene promoter interactions from literature curated databases, high-throughput experiments, binding motif and gene expression-based *in silico* inference, overall 18 resources (Garcia-Alonso et al, 2019). We included the interactions from DoRothEA subclassified by confidence levels from A to D, excluding the lowest confidence level E. In *OmniPath*, users are able to filter the TF–target interactions by confidence level.
- 6 Transcriptional regulation ("tf_target") directly from 6 literature curated resources. We show the size of the TF–target network at different settings in Dataset EV8.
- 7 In the "post_transcriptional" network, we combined 5 literature curated miRNA–mRNA interactions.

- 8 Transcriptional regulation of miRNA ("tf_mirna") from 2 literature curated resources.
- 9 lncRNA-mRNA interactions from 3 literature curated resources ("lncrna_mrna").

Enzyme-PTM interactions

After translating the identifiers, we merged enzyme-PTM interactions from 11 databases (Dataset EV3) based on the identity of the enzyme, the substrate and the modified residue and its position. In addition, we discarded the records where the residue could not be found in any of the isoform sequences from UniProt (UniProt Consortium, 2019). For each enzyme-PTM interaction, we included the original sources and the literature references. We also kept the records without literature support, e.g., from high-throughput screenings or in silico prediction.

Complexes

We combined 12 databases to build a comprehensive set of protein complexes (Dataset EV4). Seven of these databases provide information about the stoichiometry of the complexes while three contain only the lists of components. We translated the names of the components to UniProtKB accession numbers. We defined the complexes by their unique combination of members regardless of how the original resource processed the underlying experimental data. We merged the complexes based on their identical sets of components and preserved the stoichiometry if available. We represent each complex by the UniProt IDs of their components sorted alphabetically, separated by underscores and prefixed with "COMPLEX:". From the original sources, we kept the literature references, the human readable names (synonyms) and the PDB structure identifiers if available.

Annotations

Annotation resources provide diverse information about the localization, function, or other characteristics of the molecules. We obtained annotations from 49 databases (Dataset EV5). For these databases, we translated IDs and extracted the fields with relevant information. Due to the heterogeneous nature of the data, in the annotation database, the content of the resources is not merged, but rather all entries are provided independently.

Each annotation record assigns one or more attributes to a molecule. One protein might have more than one annotation record from the same database. For example, *Vesiclepedia* (Pathan et al, 2019) provides two attributes: the vesicle type and the tissue where the protein has been detected. We combined the annotation resources into a uniform table where one column is the name of the attribute and the other is the value. As one record might have multiple attributes, the records are identified by unique numbers (Fig EV3). Providing the data in this format in our web service, it can be easily reconstituted to conventional tables with standard tools like *tidyr* (<https://tidyr.tidyverse.org>) in R or *pandas* (<https://pandas.pydata.org>) in Python.

Complex annotations

Only four resources curate annotations of protein complexes; from these, we processed the complex annotations as we did for proteins. Furthermore, we inferred annotations for complexes based on the annotations of their components. We assigned the annotations to

the complex if all components agreed in all attributes that were considered relevant, e.g., if all members of a complex belong to the WNT pathway then the complex is also annotated as a member of the WNT pathway.

Intercellular signaling roles

From the resources used in *annotations*, we selected 26 with function, location, or structure information relevant in intercellular signaling. The relevant attributes we processed and combined to account for main roles in intercellular communication (e.g., ligand, receptor, ECM proteins) as well as the locational and topological properties (e.g., secreted, transmembrane). In addition, we built Boolean expressions from *Gene Ontology* terms to define the same categories. Overall we created 25 functional and 10 locational categories (Dataset EV6). Each category carries the attributes described in Dataset EV10 (Fig EV4). We manually checked the members of all the annotation groups, relying on literature knowledge and *UniProt* datasheets (UniProt Consortium, 2019), discarding the wrong annotations. We provide the classification of proteins and complexes by these categories in the *intercell* query of the web service.

Identifier translation

For each type of molecule, we chose a reference database: for proteins the *UniProtKB* ACs while for miRNAs the *miRBase* (Kozomara et al, 2019) mature Acs. From these databases, we obtained a reference set of identifiers for each type of molecular entity and organism. We then used translation tables provided by them to map other kinds of identifiers to the reference set. For *UniProt*, we corrected for deprecated or secondary Acs by translating to primary gene symbol and then to primary UniProt AC. We applied corrections to handle non-standard notations (e.g., extra dashes, Greek letters). We also accounted for the ambiguity in the mapping, i.e., if one foreign identifier may correspond to multiple reference identifiers we keep all target identifiers in *OmniPath*.

Translation by homology to rodent species

The homology translation in *pypath* uses the *NCBI HomoloGene* database (NCBI Resource Coordinators, 2018). Because *HomoloGene* uses *RefSeq* IDs, the translation takes three steps: from *UniProt* to *RefSeq*, then to the homologous *RefSeq* and finally from *RefSeq* to *UniProt*. The success rate of this translation is around 80% for mouse and roughly 30% for rat (Dataset EV9). We translated the network data and the enzyme-PTM interactions from human to mouse and rat, the two most popular mammalian model organisms. In addition, we looked up PTMs in *PhosphoSite* (Hornbeck et al, 2015) which provides homology data for PTM sites. Then, we checked the residues in the *UniProt* sequences (UniProt Consortium, 2019) and discarded the ones that did not match. The homology-translated data are included also in the *OmniPath* web service.

Data download and caching

At the database build, we download all input data directly from the original sources (Dataset EV1). Certain databases either temporarily or ultimately went offline; we deposited their data in the *OmniPath Rescued Data Repository* (<https://rescued.omnipathdb.org/>). *Pypath* contains the URLs for all resources used including the identifier translation tables. It automatically downloads, extracts, and preprocesses the data for each operation. Then, it stores the downloaded

data in a local cache directory which belongs to the user account on the computer. Once cache is created, *pyPath* reads from it and performs the download only if requested by the user.

Joint analysis of intra- and intercellular processes in SARS-CoV-2 infection

The *NicheNet* method (Browaeys et al, 2019) was built, trained, and applied to a case study using interactions and annotations from *OmniPath* resources. This information was downloaded via our R package, *OmniPathR*.

Network construction

NicheNet generates a model based on prior knowledge to describe potential regulatory effects of ligands on target genes. To reproduce their procedure, we first built three networks accounting for protein interactions of different categories retrieved from *OmniPath*:

- 1 Ligand–receptor network: We downloaded the “ligreextra” network which specifically contains known interactions between ligands and receptors. In addition, we selected proteins annotated as *ligands* or *receptors* as their main “intercellular signaling role”. Then, we extended this network with PPI whose source is a ligand and its target a receptor.
- 2 Signaling network: we retrieved PPI from the original *OmniPath* network (Túrei et al, 2016), the “kinaseextra” network and the “pathwayextra” network.
- 3 Gene regulatory network: We selected the most reliable TF–target interactions from the *DoRothEA* dataset (confidence levels A, B, and C) and the literature curated “tf_target” dataset of the “transcriptional” network of *OmniPath* to be in line with the curation level of the ligand–receptor and signaling networks.

Then, we computed ligand–target regulatory potential scores based on the topology of our aforementioned networks, following the protocols described in the *NicheNet* original study and using its associated *nichenetr* package (Browaeys et al, 2019). Briefly, *Personalized PageRank* was applied on the union of the ligand–receptor and signaling networks considering every individual ligand as starting node. To estimate the impact of every ligand in the expression of target genes, a matrix containing the *PageRank* scores is multiplied by the weighted adjacency matrix of the gene regulatory network.

Analysis of altered ligands and pathways

We applied our *OmniPath*-based version of *NicheNet* analysis on RNA-Seq data of a human lung cell line, *Calu3* (GSE147507) (Blanco-Melo et al, 2020). In this study, differential expression analysis at the gene level between controls and SARS-CoV-2-infected cells was carried out using the *DESeq2* package (Love et al, 2014). We selected over-expressed ligands (adjusted *P*-value < 0.1 and Log₂ fold-change > 1) after SARS-CoV-2 infection for further analysis. Then, we executed *Gene Set Enrichment Analysis* (GSEA) taking the Wald statistic and the hallmark gene sets from *MSigDB* (Liberzon et al, 2011) as inputs using the *fgsea* package (preprint: Korotkevich et al, 2016). Inflammatory response appeared as one of the top enriched sets. We therefore selected the leading edge genes of inflammatory response, i.e., genes contributing the most to the enrichment of this particular set, as

potential targets of the over-expressed ligands. We chose the inflammatory response genes, similarly to the original *NicheNet* study investigating the epithelial–mesenchymal transition-related genes (Browaeys et al, 2019), because these processes are likely to be regulated by extrinsic signals.

Ligand activity analysis on the aforementioned samples was conducted using the *nichenetr* package (Browaeys et al, 2019). We then selected the shortest paths between receptors (the ones interacting with the top predicted ligands) and transcription factors (the ones regulating the expression of the inflammatory target genes). These paths were exported to *Cytoscape* (Shannon et al, 2003) to generate Fig 5A.

Intercellular communication in ulcerative colitis

Intercellular interactions from OmniPath

We downloaded intercellular interactions using the “*import_intercell_network()*” method in *OmniPathR* and filtered for direct cell–cell connections: We discarded extracellular matrix proteins, extracellular matrix regulators, ligand regulators, receptor regulators, and matrix adhesion regulators and kept only membrane-bound (transmembrane or peripheral site of the membrane) proteins on the receiver side. This resulted in connections involving ligands, receptors, junction, adhesion, ion channel, transporter, and cell surface or secreted enzyme proteins.

Single-cell RNA-Seq data processing

We downloaded the raw scRNA-Seq data and processed it according to Smillie et al (Smillie et al, 2019). 51 cell types have been characterized by average gene expressions in healthy (*n* = 12) state and non-inflamed UC (*n* = 18). A gene was considered expressed if its log₂ expression value was above the mean minus 2 standard deviations of the expressed genes within the cell type.

Specific interactions between cell types

We examined all possible connections among the selected 5 cell types. We considered interactions condition specific if they appeared either only in healthy or in UC, i.e., at least one member was expressed only in the given condition. We counted the unique PPIs between each cell pair in the two conditions separately (Fig 5B). We visualized the condition-specific connections from myofibroblasts to T cells on circos plots using the *circlize* R package (Gu et al, 2014). On these figures, we grouped similar ligands (e.g., CCR2 and CCR5) and merged the connections within groups. Then, we grouped the receptors by pathways defined in *SignalLink* (Fazekas et al, 2013) to improve biological insight and visual clarity (Fig 5C).

Cell type-specific network of regulatory T cell and downstream pathway analysis

To highlight the downstream effect connections from myofibroblasts to regulatory T cells, we created a cell-specific signaling network and we carried out a pathway enrichment analysis. We used the *OmniPath Cytoscape* application (Ceccarelli et al, 2019) to combine the gene expression data with the *OmniPath* network. We limited the network to genes expressed in regulatory T cells. We selected the receptors targeted by condition-specific ligand–receptor connections in regulatory T cells. Finally, we pruned the network to the two steps neighborhood of the T cell-specific receptors. We

performed a pathway enrichment analysis on the network described above, using the online interface of the *Reactome* database with its default settings (hypergeometric test, Benjamini–Hochberg FDR correction, the human genome as the universe).

Comparing OmniPath to other resources for cell–cell interaction analysis

For the protein interaction network, we downloaded *Pathway Commons* (Cerami et al, 2011), which is an integrated resource containing undirected protein–protein connections from public pathways and interaction databases. *Pathway Commons* was downloaded from the version <https://www.pathwaycommons.org/archives/PC2/v12/PathwayCommons12.All hgnc.sif.gz>. Ligand, receptor annotations were derived from Ramiłowski et al and were downloaded from https://fantom.gsc.riken.jp/5/suppl/Ramiłowski_et_al_2015/data/PairsLigRec.txt. We run our pipeline for three different network–annotation combinations: (i) *Pathway Commons* network with Ramiłowski annotations; (ii) *Pathway Commons* network with *OmniPath* ligand, receptor annotations; and (iii) *OmniPath* network with *OmniPath* ligand, receptor annotations.

Data availability

OmniPath is available via the Python package *pypath* (<https://github.com/saezlab/pypath>), the web resource (<https://omnipathdb.org>), the R/Bioconductor package *OmniPathR* (<https://saezlab.github.io/OmniPathR>), the *omnipath* Python client (<https://github.com/saezlab/omnipath>), and the *OmniPath* Cytoscape plug-in (Ceccarelli et al, 2019). In addition, *pypath* is able to export the network and the enzyme–PTM databases in *BEL* (*Biological Expression Language*) format (Hoyt et al, 2018b), as well as to generate input files for *CellPhoneDB*. The *BEL* format databases are available in *BEL Commons* (Hoyt et al, 2018a). Code is licensed open source (GPLv3 or MIT). *Pypath* builds the *OmniPath* databases directly from the original resources, hence it gives the highest flexibility for customization and the richest API for queries and manipulation among all access options. Furthermore, it is possible to convert each database to a plain data frame and export in a tabular format. *Pypath* also generates the web resource's contents which is accessible for any HTTP client at <https://omnipathdb.org>. Information about the resources is available at <https://omnipathdb.org/info>. *OmniPathR* and the *OmniPath* Cytoscape plug-in work from the web resource data with convenient post-processing features. All data in *OmniPath* carry the licenses of the original resources (Dataset EV12), for profit users can easily limit their queries to fit the legal requirements. We maintain a directory of workflows and tutorials at <https://workflows.omnipathdb.org/>.

Apart from the figures presented in this paper, further regularly updated statistics and visualizations are available at <https://insights.omnipathdb.org>.

A Python and R package for producing the figures and tables of this paper is available at https://github.com/saezlab/omnipath_analysis. The code to build and train the *NicheNet* method (Browaeys et al, 2019) exclusively using *OmniPath* resources as well as to reproduce the SARS-CoV-2 case study is freely available at https://github.com/saezlab/NicheNet_OmniPath. The code for building the cell type-specific inter- and intracellular networks is available at https://github.com/korcsmarosgroup/uc_intercell.

Expanded View for this article is available online.

Acknowledgements

This work was partially supported by the JRC-COMBINE, partially funded by Bayer, the European Union Innovative Medicines Initiative TransQST (agreement No. 116030), the Federal Ministry of Education and Research (BMBF, Computational Life Sciences grant no. 031L0181B), the DFG (Deutsche Forschungsgemeinschaft / German Research Council; Funding code: SA 3554/1-2), to J.S.R. T.K. and D.M. were supported by the Earlham Institute (Norwich, UK) in partnership with the Quadram Institute (Norwich, UK) and strategically supported by the UKRI Biotechnological and Biosciences Research Council (BBSRC) UK grants (BB/J004529/1, BB/P016774/1, and BB/CSP17270/1) and by a BBSRC ISP grant for Gut Microbes and Health BB/R012490/1 and its constituent projects, BBS/E/F/000PR10353 and BBS/E/F/000PR10355. L.G. and M.O. were supported by the BBSRC Norwich Research Park Biosciences Doctoral Training Partnership grant number BB/M011216/1. Thanks to John P Thomas, Robin Broeways, Yvan Saey, Mirjana Eftemova, Daniel Domingo-Fernandez, Charles Tapley Hoyt, Lu Li and Paul D Thomas for their helpful feedback and discussions. Open Access funding enabled and organized by Projekt DEAL.

Author contributions

Design and development of *pypath* and *OmniPath* and creating descriptive figures and tables: DT; *OmniPathR* package: AV, DT, and AG; Designing and performing case study on SARS-CoV-2 infection data: AV; Designing and performing case study on ulcerative colitis: LG and DM; Development of *pypath* and visualization of the database contents: NP-E, OI, and LG; *omnipath* Python module: MK; Supervision of *omnipath* Python module: FT; Tutorials and analysis of coverage of ligand–receptor databases on textbook data: MO; Supervision of the project: JS-R and TK; Manuscript writing: All authors.

Conflict of interest

JSR receives funding from GSK and Sanofi and consultant fees from Travers Therapeutics.

References

- An S, Cho S-Y, Kang J, Lee S, Kim H-S, Min D-J, Son E, Cho K-H (2020) Inhibition of 3-phosphoinositide-dependent protein kinase 1 (PKD1) can revert cellular senescence in human dermal fibroblasts. *Proc Natl Acad Sci USA* 3: 201920338
- Andrighetti T, Bohar B, Lemke N, Sudhakar P, Korcsmaros T (2020) Microbiolink: an integrated computational pipeline to infer functional effects of microbiome-host interactions. *Cells* 9: 1278
- Armingol E, Joshi C, Baghdassarian H, Shamie I, Ghaddar A, Chan J, Her H-L, O'Rourke E, Lewis NE (2020a) Inferring the spatial code of cell–cell interactions and communication across a whole animal body. *bioRxiv* <https://doi.org/10.1101/2020.11.22.392217> [PREPRINT]
- Armingol E, Officer A, Harismendy O, Lewis NE (2020b) Deciphering cell–cell interactions and communication from gene expression. *Nat Rev Genet* 21: 71–88
- Armstrong JF, Faccenda E, Harding SD, Pawson AJ, Southan C, Sharman JL, Campo B, Cavanagh DR, Alexander SPH, Davenport AP et al (2019) The IUPHAR/BPS guide to PHARMACOLOGY in 2020: extending immunopharmacology content and introducing the IUPHAR/MMV guide to MALARIA PHARMACOLOGY. *Nucleic Acids Res* 48: D1006–D1021

- Asp M, Bergenstråhle J, Lundberg J (2020) Spatially resolved transcriptomes—next generation tools for tissue exploration. *BioEssays* 42: 1900221
- Bagca BG, Avci CB (2020) The potential of JAK/STAT pathway inhibition by ruxolitinib in the treatment of COVID-19. *Cytokine Growth Factor Rev* 54: 51–61
- Bausch-Fluck D, Hofmann A, Bock T, Frei AP, Cerciello F, Jacobs A, Maest H, Omasits U, Gundry RL, Yoon C et al (2015) A mass spectrometric-derived cell surface protein atlas. *PLoS One* 10: e0121314
- Bianco-Melo D, Nilsson-Payant BE, Liu W-C, Uhl S, Hoagland D, Møller R, Jordan TX, Oishi K, Panis M, Sachs D et al (2020) Imbalanced host response to SARS-CoV-2 drives development of COVID-19. *Cell* 181: 1036–1045.e9
- Bouhaddou M, Memon D, Meyer B, White KM, Rezelj VV, Correa Marrero M, Polacco BJ, Melnyk JE, Ulferts S, Kaake RM et al (2020) The global phosphorylation landscape of SARS-CoV-2 infection. *Cell* 182: 685–712
- Broweys R, Saelens W, Saeyns Y (2019) NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat Methods* 17: 159–162
- Cameron MJ, Kelvin DJ (2013) Cytokines, chemokines and their receptors. In *Madame Curie Bioscience Database Landes Bioscience*
- Cao AT, Yao S, Stefka AT, Liu Z, Qin H, Liu H, Evans-Marin HL, Elson CO, Nagler CR, Cong Y (2014) TLR4 regulates IFN- γ and IL-17 production by both thymic and induced Foxp3+ Tregs during intestinal inflammation. *J Leukoc Biol* 96: 895–905
- Catanzaro M, Fagiani F, Racchi M, Corsini E, Govoni S, Lanni C (2020) Immune response in COVID-19: addressing a pharmacological challenge by targeting pathways triggered by SARS-CoV-2. *Signal Transduct Target Ther* 5: 1–10
- Ceccarelli F, Túrei D, Gabor A, Saez-Rodriguez J (2019) Bringing data from curated pathway resources to Cytoscape with OmniPath. *Bioinformatics* 36: 2632–2633
- Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, Schultz N, Bader GD, Sander C (2011) Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res* 39: D685–D690
- Damle NP, Köhn M (2019) The human DEPhosphorylation Database DEPOD: 2019 update. *Database* 2019
- Domingo-Fernández D, Mubeen S, Marín-Llaoá J, Hoyt CT, Hofmann-Apitius M (2019) PathMe: merging and exploring mechanistic pathway knowledge. *BMC Bioinformatics* 20: 243
- Efremova M, Vento-Torres M, Teichmann SA, Vento-Torres R (2020) Cell PhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat Protoc* 15: 1484–1506
- Fazekas D, Koltai M, Túrei D, Módos D, Pálfi M, Dúl Z, Zsákai L, Szalay-Bekó M, Lenti K, Farkas IJ et al (2013) Signalink 2 - a signaling pathway resource with multi-layered regulatory networks. *BMC Syst Biol* 7: 7
- Forward NA, Furlong SJ, Yang Y, Lin T-J, Hoskin DW (2010) Signaling through TLR7 enhances the immunosuppressive activity of murine CD4 CD25 T regulatory cells. *J Leukoc Biol* 87: 117–125
- García-Alonso L, Holland CH, Ibrahim MM, Túrei D, Saez-Rodriguez J (2019) Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res* 29: 1363–1375
- Gu Z, Gu L, Ellis R, Schlesner M, Brors B (2014) circlize Implements and enhances circular visualization in R. *Bioinformatics* 30: 2811–2812
- Gyori BM, Bachman JA, Subramanian K, Muhlich JL, Galescu L, Sorger PK (2017) From word models to executable models of signaling networks using automated assembly. *Mol Syst Biol* 13: 954
- He T, Liu S, Chen S, Ye J, Wu X, Bian Z, Chen X (2018) The p38 MAPK inhibitor SB203580 abrogates tumor necrosis factor-induced proliferative expansion of mouse CD4Foxp3 regulatory T cells. *Front Immunol* 9: 1556
- Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* 43: D512–D520
- Hou R, Denisenko E, Ong HT, Ramilowski JA, Forrest ARR (2020) Predicting cell-to-cell communication networks using NATMI. *Nat Commun* 11: 5011
- Hoyt CT, Domingo-Fernández D, Hofmann-Apitius M (2018a) BEL Commons: an environment for exploration and analysis of networks encoded in biological expression language. *Database* 2018
- Hoyt CT, Konotopez A, Ebeling C, Wren J (2018b) PyBEL: a computational framework for biological expression language. *Bioinformatics* 34: 703–704
- Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, Sidiropoulos K, Cook J, Gillespie M, Haw R et al (2020) The reactome pathway knowledgebase. *Nucleic Acids Res* 48: D498–D503
- Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan C-H, Myung P, Pliusk MV, Nie Q (2021) Inference and analysis of cell-cell communication using CellChat. *Nat Commun* 12: 1088
- Kamburov A, Stelzl U, Lehrach H, Herwig R (2013) The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res* 41: D793–800
- Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A et al (2009) Human protein reference database-2009 update. *Nucleic Acids Res* 37: D767–D772
- Kirouac DC, Ito C, Csaszar E, Roch A, Yu M, Sykes EA, Bader GD, Zandstra PW (2010) Dynamic interaction networks in a hierarchically organized tissue. *Mol Syst Biol* 6: 417
- Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A (2016) Fast gene set enrichment analysis. *bioRxiv* <https://doi.org/10.1101/060012> [PREPRINT]
- Kozomara A, Birgaoanu M, Griffiths-Jones S (2019) miRBase: from microRNA sequences to annotation. *Nucleic Acids Res* 47: D155–D162
- Kuperstein I, Bonnet E, Nguyen H-A, Cohen D, Viara E, Grieco L, Fourquet S, Calzone L, Russo C, Kondratova M et al (2015) Atlas of Cancer signalling network: a systems biology resource for integrative analysis of cancer data with google maps. *Oncogenesis* 4: e160
- Kweler K, Starostvetsky E, Ziv-Kenet A, Kalugny Y, Garelík Y, Shalev-Malul G, Aizenbud-Reshef N, Dubovik T, Brilller M, Campbell J et al (2018) Immune-centric network of cytokines and cells in disease context identified by computational mining of PubMed. *Nat Biotechnol* 36: 651–659
- Li F, Wu T, Xu Y, Dong Q, Xiao J, Xu Y, Li Q, Zhang C, Gao J, Liu L et al (2020) A comprehensive overview of oncogenic pathways in human cancer. *Brief Bioinform* 21: 957–969
- Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27: 1739–1740
- Licata L, Lo Surdo P, Iannuccelli M, Palma A, Micarelli E, Perfetto L, Peluso D, Calderone A, Castagnoli L, Cesareni G (2020) SIGNOR 2.0, the signaling network open resource 2.0: 2019 update. *Nucleic Acids Res* 48: D504–D510
- Liu A, Trairatphisan P, Gjerga E, Didangelos A, Barratt J, Saez-Rodriguez J (2019) From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL. *Npj Syst Biol Appl* 5: 40
- Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15: 550

- Luck K, Kim D-K, Lambourne L, Spirohn K, Begg BE, Bian W, Brignall R, Cafarelli T, Campos-Laborie FJ, Charleatoux B et al (2020) A reference map of the human binary protein interactome. *Nature* 580: 402–408
- NCBI Resource Coordinators (2018) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 46: D8–D13
- Noël F, Massenet-Regad L, Carmi-Levy I, Cappuccio A, Grandclaudon M, Trichot C, Kieffer Y, Mehta-Grigoriou F, Saumelis V (2021) Dissection of intercellular communication using the transcriptome-based framework ICELLNET. *Nat Commun* 12: 1089
- Nyirenda MH, Morandi E, Vinkemeier U, Constantin-Teodosiu D, Drinkwater S, Mee M, King L, Poddá G, Zhang G-X, Ghaemmaghani A et al (2015) TLR2 stimulation regulates the balance between regulatory T cell and Th17 function: a novel mechanism of reduced regulatory T cell function in multiple sclerosis. *J Immunol* 194: 5761–5774
- Oliveira AC, Richards EM, Karas MM, Pepine CJ, Raizada MK (2020) Would repurposing minocycline alleviate neurologic manifestations of COVID-19? *Front Neurosci* 14: 577780
- Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, Deltoro N et al (2014) The MINTAct project—intAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* 42: D358–D363
- Ostaszewski M, Niarakis A, Mazein A, Kuperstein I, Phair R, Orta-Resendiz A, Singh V, Aghamiri SS, Acencio ML, Glaab E et al (2020) (2020) COVID-19 disease map, a computational knowledge repository of SARS-CoV-2 virus-host interaction mechanisms. *Cold Spring Harbor Lab* 10: 356014
- Oughtred R, Stark C, Breitkreutz B-J, Rust J, Boucher L, Chang C, Kolas N, O'Donnell L, Leung C, McAdam R et al (2019) The BioGRID interaction database: 2019 update. *Nucleic Acids Res* 47: D529–D541
- Pathan M, Fonseka P, Chitti SV, Kang T, Sanwlani R, Van Deun J, Hendrix A, Mathivanan S (2019) Vesiclepedia 2019: a compendium of RNA, proteins, lipids and metabolites in extracellular vesicles. *Nucleic Acids Res* 47: D516–D519
- Ramilowski JA, Goldberg T, Harshbarger J, Kloppmann E, Lizio M, Satagopam VP, Itoh M, Kawaji H, Carninci P, Rost B et al (2015) A draft network of ligand-receptor-mediated multicellular signalling in human. *Nat Commun* 6: 7866
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13: 2498–2504
- Shroyer NF, Wong MH (2007) BMP signaling in the intestine: cross-talk is key. *Gastroenterology* 133: 1035–1038
- Sleater DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, Mélius J, Cirillo E, Coort SL, Digles D et al (2018) WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res* 46: D661–D667
- Smillie CS, Biton M, Ordovas-Montanes J, Sullivan KM, Burgin G, Graham DB, Herbst RH, Rogel N, Slyper M, Waldman J et al (2019) Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. *Cell* 178: 714–730.e22
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102: 15545–15550
- Sun S, Zhu J, Zhou X (2020) Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nat Methods* 17: 193–200
- Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P et al (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 47: D607–D613
- Takahashi M, Nakamura K, Honda K, Kitamura Y, Mizutani T, Araki Y, Kabemura T, Chijiwa Y, Harada N, Nawata H (2006) An inverse correlation of human peripheral blood regulatory T cell frequency with the disease activity of ulcerative colitis. *Dig Dis Sci* 51: 677–686
- Tanevski J, Attila G, Ramirez-Flores RO, Schapiro D, Saez-Rodriguez J (2020) Explainable multi-view framework for dissecting inter-cellular signaling from highly multiplexed spatial data. *bioRxiv* <https://doi.org/10.1101/2020.05.08.084145> [PREPRINT]
- Touré V, Flobak Å, Niarakis A, Verduyck S, Kuiper M (2020) The status of causality in biological databases: data resources and data retrieval possibilities to support logical modeling. *Brief Bioinform* <https://doi.org/10.1093/bib/bbaa390>
- Treveil A, Bohar B, Sudhakar P, Gul L, Csabai L, Olbei M, Poletti M, Madgwick M, Andrighetti T, Hautefort I et al (2021) ViralLink: An integrated workflow to investigate the effect of SARS-CoV-2 on intracellular signalling and regulatory pathways. *PLoS Comput Biol* 17: e1008685
- Túrei D, Korcsmáros T, Saez-Rodriguez J (2016) OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat Methods* 13: 966–967
- Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjödstedt E, Asplund A et al (2015) Proteomics. Tissue-based map of the human proteome. *Science* 347: 1260419
- UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 47: D506–D515
- Wan YY, Flavell RA (2008) TGF-beta and regulatory T cell in immunity and autoimmunity. *J Clin Immunol* 28: 647–659
- West NR (2019) Coordination of immune-stroma crosstalk by IL-6 family cytokines. *Front Immunol* 10: 1093
- Wolf FA, Angerer P, Theis FJ (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 19: 15
- Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D (2000) DIP: the database of interacting proteins. *Nucleic Acids Res* 28: 289–291
- Xiao X, Zhao P, Rodriguez-Pinto D, Qi D, Henegariu O, Alexopoulou L, Flavell RA, Wong FS, Wen L (2009) Inflammatory regulation by TLR3 in acute hepatitis. *J Immunol* 183: 3712–3719
- Yamada A, Arakaki R, Saito M, Tsunematsu T, Kudo Y, Ishimaru N (2016) Role of regulatory T cell in the pathogenesis of inflammatory bowel disease. *World J Gastroenterol* 22: 2195–2205



License: This is an open access article under the terms of the Creative Commons Attribution 4.0 License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

RESEARCH ARTICLE

Extracellular vesicles produced by the human commensal gut bacterium *Bacteroides thetaiotaomicron* affect host immune pathways in a cell-type specific manner that are altered in inflammatory bowel disease

Lejla Gul¹ | Dezso Modos^{1,2} | Sonia Fonseca² | Matthew Madgwick^{1,2} |
John P. Thomas^{1,3} | Padhmanand Sudhakar^{1,2,4} | Catherine Booth⁵ | Régis Stentz² |
Simon R. Carding^{2,6} | Tamas Korcsmaros^{1,2}

¹ Earlham Institute, Norwich, Norwich, UK

² Gut Microbes and Health Research Programme, Quadram Institute Bioscience, Norwich, UK

³ Department of Gastroenterology, Norfolk and Norwich University Hospital, Norwich, UK

⁴ KU Leuven Department of Chronic Diseases, Metabolism and Ageing, Translational Research Centre for Gastrointestinal Disorders (TARGID), Leuven, Belgium

⁵ Core Science Resources, Quadram Institute Bioscience, Norwich, UK

⁶ Norwich Medical School, University of East Anglia, Norwich, UK

Correspondence

Simon R. Carding and Tamas Korcsmaros, Gut Microbes and Health Research Programme, Quadram Institute Bioscience, Norwich, UK. Email: Simon.Carding@quadram.ac.uk and Tamas.Korcsmaros@earlham.ac.uk

Funding information

UKRI Biotechnological and Biosciences Research Council (BBSRC), Grant/Award Numbers: BB/J004529/1, BB/P016774/1, BB/CSPI7270/1; Quadram Institute's Gut Microbes and Health Institute Strategic Programme, Grant/Award Numbers: BB/R012490/1, BBS/E/F/000PR10353, BBS/E/F/000PR10355; BBSRC Norwich Research Park Biosciences Doctoral Training Partnership, Grant/Award Number: BB/M01216/1; the National Institute of Health Research (NIHR)

Abstract

The gastrointestinal (GI) tract harbours a complex microbial community, which contributes to its homeostasis. A disrupted microbiome can cause GI-related diseases, including inflammatory bowel disease (IBD), therefore identifying host-microbe interactions is crucial for better understanding gut health. Bacterial extracellular vesicles (BEVs), released into the gut lumen, can cross the mucus layer and access underlying immune cells. To study BEV-host interactions, we examined the influence of BEVs generated by the gut commensal bacterium, *Bacteroides thetaiotaomicron*, on host immune cells. Single-cell RNA sequencing data and host-microbe protein-protein interaction networks were used to predict the effect of BEVs on dendritic cells, macrophages and monocytes focusing on the Toll-like receptor (TLR) pathway. We identified biological processes affected in each immune cell type and cell-type specific processes including myeloid cell differentiation. TLR pathway analysis highlighted that BEV targets differ among cells and between the same cells in healthy versus disease (ulcerative colitis) conditions. The *in silico* findings were validated in BEV-monocyte co-cultures demonstrating the requirement for TLR4 and Toll-interleukin-1 receptor domain-containing adaptor protein (TIRAP) in BEV-elicited NF- κ B activation. This study demonstrates that both cell-type and health status influence BEV-host communication. The results and the pipeline could facilitate BEV-based therapies for the treatment of IBD.

KEYWORDS

extracellular vesicles, host-microbe interactions, single-cell data analysis, toll-like receptor pathway, ulcerative colitis

1 | INTRODUCTION

The human gastrointestinal (GI) tract microbiota consisting of bacteria, viruses, archaea, and eukaryotic microbes, contributes to intestinal homeostasis by communicating with various host cells in the intestinal mucosa. Structural, compositional, and

This is an open access article under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Journal of Extracellular Vesicles* published by Wiley Periodicals, LLC on behalf of the International Society for Extracellular Vesicles

functional alterations of the microbiota (“dysbiosis”) are associated with various GI-related diseases, including Crohn’s disease (CD) and ulcerative colitis (UC), two major forms of inflammatory bowel disease (IBD) (Delday et al., 2019). Dysbiosis in IBD is characterised by a reduction in bacterial diversity (UC) or altered composition (CD) that involves *Bacteroides* and *Firmicutes* species (Kabeerdoss et al., 2015). Despite recent advances in our understanding of IBD pathogenesis, the complex interactions between the dysbiotic gut microbiota and the host mucosa that result in aberrant immune activation and inflammation in the gut, are yet to be defined in detail.

Bacteroides thetaioamicron (Bt) is a Gram-negative anaerobe that is a major constituent of the human caecal and colonic microbiota (ScienceDirect Topics, 2021). The administration of Bt in murine models of IBD ameliorates inflammation (Chang et al., 2020; Fábrega et al., 2017) with the anti-inflammatory effects being at least in part mediated by its production of bacterial extracellular vesicles (BEVs). BEVs are released by both commensal Gram-negative and Gram-positive bacteria and have the potential to mediate cross-kingdom interactions with host cells via the delivery of their contents and cargo to affect host cell physiology and function (Chang et al., 2020). BEVs produced by Gram-negative bacteria, such as Bt, are small, spherical bilayered structures (20–400 nm) composed of phospholipids, lipopolysaccharides, peptidoglycan, outer membrane proteins, periplasmic contents including proteins, and some inner membrane and cytoplasmic fractions (Chronopoulos & Kalluri, 2020; Schwechheimer & Kuehn, 2015). BEVs can permeate through the sterile mucus layer of the colon to access and transmigrate boundary intestinal epithelial cells through different routes (Jones et al., 2020) enabling them to interact with underlying mucosal immune cells (Cecil et al., 2019; Durant et al., 2020; Hickey et al., 2015; Kaparakis-Liaskos & Ferrero, 2015; Shen et al., 2012) and the intestinal vasculature which facilitates their wider, systemic dissemination (Durant et al., 2020; Jones et al., 2020; Stentz et al., 2018).

For Gram-negative bacteria, a defined pathway of interaction with the host immune system is via membrane-associated molecules, including lipopolysaccharide (LPS) (Matsuura, 2013). Immune cells interact with LPS via their pattern recognition molecules such as Toll-like receptors (TLRs). LPS consists of three main structural components of diverse functions: lipid anchor (lipid A), core oligosaccharide region, and O-antigen. Lipid A is the most conserved part of LPS. The core region connects the anchor and antigen units, the O-antigen is the immunogenic portion of LPS consisting of long polysaccharide chains (Arenas, 2012). The structure of LPS is diverse among bacterial taxa resulting in taxon-specific immune responses in the host. Bt contains lipooligosaccharides (LOS) which are structurally distinct from the prototypical LPS of *Escherichia coli* (*E. coli*) (Jacobson et al., 2018). For example, while lipid A is both hexa-acylated and diphosphorylated in *E. coli*, Bt has penta-acylated and monophosphorylated lipid A that does not promote proinflammatory responses in immune cells (Jacobson et al., 2018; Steimle et al., 2019).

Host cells acquire and degrade BEVs by several pathways including dynamin-dependent endocytosis, macropinocytosis, and caveolin-mediated endocytosis (Jones et al., 2020). BEVs and their protein cargo can trigger intracellular signalling cascades in various immune cells such as dendritic cells (DCs). In the healthy gut, this interaction leads to the production of anti-inflammatory cytokines (such as IL-10), whereas in the inflamed gut of IBD patients, this anti-inflammatory response is lost (Durant et al., 2020). Another recent study showed that Bt BEVs enhance regulatory T cell and helper T cell 1 (Th1) responses, while decreasing the activation of Th2 and Th17 cell (Li et al., 2021). These anti-inflammatory properties of BEVs have led to their incorporation into probiotic-based therapeutics in murine models of IBD (Chang et al., 2020; Fábrega et al., 2017). There is a major need for such novel therapeutic strategies as despite the advent of biologic therapies in IBD, ~25% of patients with UC and up to 75% with CD eventually require surgical intervention. One such strategy being explored is the ability to modulate the host immune system through microbiota-based therapies (Zhang et al., 2017). Given the ability of Bt BEVs to influence host immune cell signalling they may have untapped therapeutic potential.

However, the effects of Bt BEVs on different host immune cells are poorly understood. Single-cell transcriptomics (scRNAseq) provides an opportunity to understand how Bt BEVs influence gut mucosal immune cell populations with cell-type specific resolution. Of particular interest are monocytes, macrophages and DCs, which play key roles in initiating and determining the outcome of local and systemic immune responses to non-harmful and harmful stimuli (Scott & Mann, 2020), and shaping the immune response in IBD (Steinbach & Plevy, 2014).

Here, we have utilised single-cell RNAseq datasets in combination with Bt BEV proteomes to develop a computational workflow of the predicted effect of BEVs on immune cells at different stages of their development, in healthy and disease (UC) states. In a proof-of-concept study, we experimentally confirm the predicted interaction of BEVs with human monocytes via TLR4.

2 | MATERIAL & METHODS

2.1 | Characterisation of Bt BEV proteins

The bacterium Bt VPI-5482 was grown anaerobically at 37°C with agitation using a magnetic stirrer in Brain Heart Infusion (BHI) medium (Oxoid/Thermo Fisher, Basingstoke, UK) supplemented with 0.5 mg/L haemin. BHI (three independent cultures) was inoculated with an overnight culture of Bt at an initial OD₆₀₀ of 0.05. After 5 h of growth (OD approximately 3.0, early stationary phase), the cells were centrifuged at 5500 g for 45 min at 4°C. The supernatants were filtered through polyethersulfone

(PES) membranes (0.22 μm pore-size) (Sartorius) to remove debris and cells. The sterility of the vesicle-containing filtrates was confirmed by plating onto BHI–haemin agar. BEVs in the 500 ml filtrates were concentrated by crossflow ultrafiltration (100 kDa MWCO, Vivaflow 50R, Sartorius) to 0.5 ml, diluted by addition of 500 ml of ice-cold phosphate buffered saline (PBS), pH 7.4, and the suspensions were concentrated again by crossflow filtration to 0.5 ml and filter-sterilised through a 0.22 μm PES membrane (Sartorius). Following crossflow ultrafiltration, further purification of BEVs was performed by fractionation of the suspension (Durant et al., 2020) by size-exclusion chromatography using a CL2-B Sepharose (Sigma-Aldrich) (120 cm \times 1 cm column) in PBS buffer. The absorbance of the fractions was measured at 280 nm and the first fractions corresponding to the first absorbance peak were pooled and concentrated to 1 ml with a Vivaspin 20 centrifugal concentrator (100 kDa molecular weight cut-off, Sartorius) and filtered through a 0.22 μm PES membrane (Sartorius). Vesicle concentration was determined by Nanoparticle Tracking Analysis (NTA). The BEV suspension was centrifuged (150,000 g at 4°C or 2 h in a Ti70 rotor (Beckman Instruments)), the supernatant removed using a vacuum pump and the vesicle pellets were snap frozen in liquid nitrogen and stored at -80°C prior to extraction.

2.2 | Proteomic analysis

Samples for proteomics analysis consisted of 100 μg of BEV or cell protein extract prepared and labelled at the Bristol University proteomics facility using TMT reagents (10-Plex format, Isobaric Mass Tagging kit, Thermo Scientific). Labelled samples were pooled and then fractionated using High pH Reverse Phase Liquid Chromatography. The resulting fractions were subjected to nano-LC MS/MS using an Orbitrap Fusion Tribrid mass spectrometer with an SPS-MS3 acquisition method. Fragmentation of the isobaric tag released the low molecular mass reporter ions which were used to quantify the peptides. Protein quantitation was based on the median values of multiple peptides identified from the same protein, resulting in highly accurate protein quantitation between samples. The data sets were analysed using the Proteome Discoverer v2.1 software and run against the Bt VPI-5482 and filtered with a 1% and 5% FDR cut-off.

2.3 | Transmission electron microscopy

Samples were visualized using negative staining with TEM. Briefly, 4 μl BEV suspension was adsorbed to plasma-pretreated carbon-coated copper EM grids (EM Solutions) for 1 min before wicking off with filter paper and negatively staining with 1% Uranyl Acetate solution (BDH 10288) for 1 min. Grids were air-dried before analysis using a FEI Talos F200C electron microscope at 36,000 \times –92,000 \times magnification with a Gatan Oneview digital camera.

2.4 | Isolation and characterisation of Bt BEVs for the experimental validation

Bt (strain VPI-5482) was grown with agitation under anaerobic conditions at 37°C in 50 ml (three replicates) of brain heart infusion (BHI) broth medium (Oxoid/Thermo Fisher, Basingstoke, UK) supplemented with 0.5 mg/L haemin (Sigma-Aldrich, St Louis, MO, USA) (BHI–haemin) at 37°C to early stationary phase (OD approximately 2.5). 20 ml of each culture was centrifuged at 5,500 g for 20 min at 4°C and the supernatants vacuum-filtered through polyethersulfone (PES) membranes (0.22 μm pore-size) (Sartorius) to remove debris and cells. Supernatants were concentrated by ultrafiltration using Amicon ultra-15 centrifugal filter units (100 kDa molecular weight cut-off), the retentate was rinsed twice with 15 ml of PBS (pH 7.4) and concentrated to 150 μl . To separate out BEVs from remaining proteins and lipids, qEVsingle/35 nm columns (Izon) were used to perform SEC according to manufacturer instructions. Fractions containing BEVs were combined and the suspensions were stored at 4°C. The size and concentration of the isolated BEVs was determined using a ZetaView PMX-220 TWIN instrument according to manufacturer instructions (Particle Metrix GmbH). Aliquots of BEVs suspension were diluted 1000- to 20,000-fold in particle-free PBS for analysis. Size distribution video data was acquired using the following settings: temperature: 25°C; frames: 60; duration: 2 s; cycles: 2; positions: 11; camera sensitivity: 80 and shutter value: 100. The ZetaView NTA software (version 8.05.12) was used with the following post acquisition settings: minimum brightness: 20; max area: 2000; min area: 5 and trace length: 30.

2.5 | Single-cell transcriptomic datasets analysis

A publicly available scRNAseq dataset describing gene expressions in 51 cell-types from the colon in three conditions (healthy, non-inflamed UC, and inflamed UC) was analysed by using the average expression of genes (Smillie et al., 2019). From the 51 cell-type datasets, cycling monocytes, inflammatory monocytes, macrophages, DC1 (healthy mucosa-related subset) and DC2 (inflammation-related subset) populations appearing in healthy and non-inflamed UC conditions were selected for further

analysis. Raw data is available on the Single Cell Portal (<https://singlecell.broadinstitute.org/>) under SCP259 study ID. While the original dataset contains inflamed samples, in order to avoid inflammation-related bias in cell communication we focused our analysis on non-inflamed cells from the same UC patients.

Raw scRNAseq data was processed using scripts and parameters by Smillie et al (Smillie et al., 2019) (http://www.github.com/cssmillie/ulcerative_colitis). To discard genes expressed at extremely low levels, we applied a z-score test based on the method of Hart et al (Hart et al., 2013). A gene was considered not to be expressed if its log₂ expression value was less than three standard deviations of the mean expressed genes in that cell.

2.6 | THP-1 monocyte transcriptomic analysis

Two publicly available bulk RNAseq datasets of the human monocytic cell line THP-1 were used for experimental validation. Raw counts from GSE132408 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE132408>) and GSE157052 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE157052>) datasets were normalized using the DESeq2 package in R. Due to the different gene symbols and gene IDs in the datasets, we unified them to gene symbols using Uniprot and used only genes detected in both experiments. We applied the same protocol as for the single cell RNA-seq datasets: first we log₂ transformed the count number and then we used a Z normalisation. We considered a gene expressed if its z-score was above -3 (mean -3 standard deviation). We used these Z transformed values for the analysis.

2.7 | Constructing a host cell-BEV interactome

We predicted the effect of BEV proteins on different cell-types based on host-microbe protein-protein interaction (PPI) networks using our MicrobioLink pipeline (Andrighetti et al., 2020). The connections were based on experimentally verified domain-motif interactions from the Eukaryotic Linear Motif (ELM) database (Kumar et al., 2020). It was assumed that a BEV protein containing a domain can bind to a human protein having the corresponding interacting motif within its sequence. First, we downloaded the sequence of BEV and human proteins from the Uniprot database (Consortium, 2019). Then Pfam domains of BEV proteins were predicted by InterProScan and human motifs identified by the ELM database. To avoid large numbers of false-positive PPIs, a quality filter was applied using IUPred tool (Mészáros et al., 2018) which uses scores based on two methods (IUPred and ANCHOR2) to measure residue-level energy terms. The energy terms correlate how intrinsically disordered the protein region is. Higher disordered regions are more accessible for the bacterial domain. Two cut-off values (IUPred > 0.5 and ANCHOR2 > 0.4) were set up to select human motifs which are presented out of globular domains and at an intrinsic disordered protein region (Mészáros et al., 2018).

2.8 | Functional analysis of BEV target proteins

Functional analysis was performed using the Gene Ontology (GO) database. GO database orders the annotations in a tree-like structure where parent and child categories are represented in a hierarchical way. GOrilla was used to highlight the enriched biological processes of the BEV targets in different cell-types (Eden et al., 2009). As a background dataset, all expressed genes were examined in cells facilitating the identification of cell-type specific functions. An annotation was significantly overrepresented among the Bt targets if the *P*-value was less than 10⁻³ and the FDR *q*-value calculated by Benjamini and Hochberg method was less than 0.05. We used REVIGO to reduce the dimensionality of the annotations, thereby avoiding the overlapping processes that belong to the same function and identify significant differences among functions (Supek et al., 2011). simRel scores were applied to measure the GO semantic similarity. To visualise the functional overlap among cell-types, InteractiVenn was used (Heberle et al., 2015). Although this analysis is suitable for depicting processes that are specific to a cell-type or condition due to the large number of BEV interacting proteins in each cell-type, the output of this analysis focuses mainly on common processes. A more fine-grained analysis can be achieved by involving gene expression values, and not only the presence or absence of a gene's expression when establishing condition specific differences.

2.9 | Cell-type and condition specific TLR pathway modelling

Members of the TLR pathway were derived from the Reactome database due to its high and reliable coverage of associating proteins to pathways. (Jassal et al., 2020). The OmniPath database was used to collect the interactions due to slightly larger coverage of interaction data compared to Reactome (Türei et al., 2016). Signalling in different cell-types was interpreted by adding the expression values from scRNAseq datasets (monocytes, dendritic cells, macrophage) and bulk RNAseq (THP-1 cells).

To compare the signal flow under different conditions (healthy and non-inflamed UC), expression values were added to the genes/proteins. We created one network for each cell type to represent both conditions. We avoided using differentially expressed genes because it focuses only on the differences at the gene level and not the pathway of the spreading signal. Therefore, the healthy log₂ gene expression was subtracted from the diseased expression value to indicate differences in signal flow in the TLR pathway.

2.10 | TLR-signalling in THP1-Blue cells

THP1-Blue NF- κ B reporter cell line (Invivogen) was derived from the human THP-1 monocytic cell line by stable integration of an NF- κ B-inducible secreted alkaline phosphatase (SEAP) reporter construct. THP1-Blue cells were cultivated in RPMI-1640 (Sigma-Aldrich) supplemented with 10% heat-inactivated FBS (Biosera), 1% Pen/Strep (Sigma-Aldrich) and 100 μ g/ml Normocin (Invivogen) at 37°C and 5% CO₂ in a humidified incubator. To maintain selection pressure during cell subculturing, 10 μ g/ml blasticidin (Invivogen) was added to the growth medium at every other passage. To identify TLR4 and TIRAP mediated activation THP-1 cells were seeded in flat-bottomed 96-well plates at a density of 5×10^5 cells/ml and incubated with *E. coli* derived LPS (10 ng/ml, Sigma-Aldrich) 1 h at 37°C. Control cultures were incubated with PBS. In some cases, cells were pre-treated with the TLR4 inhibitor CLI-095 (2 μ g/ml) (Invivogen) or peptide-based TIRAP inhibitor (50 μ g/ml) (Merck) and incubated for 1.5 h at 37°C and 5% CO₂ in a humidified incubator. For BEV-THP-1 co-culture cells were incubated for 24 h with different concentrations of BEVs (3×10^9 , 3×10^8 , and 3×10^7 /ml) after which 20 μ l of the cell suspension was added to flat-bottomed 96-well plates, mixed with 180 μ l of Quanti-Blue (Invivogen) colorimetric assay reagent and incubated for 1 h at 37°C. Secreted alkaline phosphatase (SEAP) levels were quantified by absorbance reading at 620 nm. All incubations were performed in triplicate.

2.11 | Statistical analysis

Data were subjected to one-way or two-way ANOVA followed by Bonferroni's multiple comparison post hoc test using GraphPad Prism 5 software. Statistically significant differences between two mean values were established by adjusted *P*-value < 0.05. Data are presented as the mean \pm standard deviation.

2.12 | Data availability

Raw scRNAseq data was extracted from Smillie et al. (2019). Bulk transcriptomics for THP-1 cell line analysis can be found in GEO [GSE132408, GSE157052]. The workflow containing Python and R scripts, input files and results is accessible on GitHub (https://github.com/korcsmarosgroup/BT_BEV_project/).

3 | RESULTS

3.1 | The BEV-Immune cell protein interactome

To analyse the effect of BEV proteins on human cell-type specific signalling pathways we developed a computational workflow to process single-cell data, combine information from network resources, and incorporate bioinformatics prediction tools (Figure 1).

Using this workflow, we identified potential candidates from the proteome of BEVs obtained from a culture of Bt grown in the complex medium BHI, which totalled 2068 proteins. The same proteins were identified in BEVs extracted from the caecum of germ-free mice monocolonized with Bt (Stentz et al., 2020). TEM was used to determine the purity of BEV preparations (Figure S1). For host cells, scRNAseq data identifying genes expressed in each of five immune cell-types was used (Figure 2). For the purpose of developing the protein-protein interaction (PPI) network, we assumed that all of the expressed genes were translated into functional proteins.

BEVs can interact with the host via cell surface receptors and after internalisation, with cytoplasmic receptors. We did not therefore filter host proteins based on their cellular location. Despite the large number of BEV-human PPIs (Figure 2) the majority of bacterial proteins were hubs indicating they can potentially interact with thousands of host proteins. In total, 48 BEV proteins interact with the host immune cells (Table S1), the majority of which are hydrolases, proteases, and other catabolic enzymes without a specific cleavage site. In terms of individual interactions, five BEV helicase proteins (BT_0831, BT_1154, BT_3303, BT_3844 and BT_3938) were predicted to target the same host protein PAPD5, a non-canonical poly(A) polymerase whose function is impaired in IBD (Boele et al., 2014; Rammelt et al., 2011).

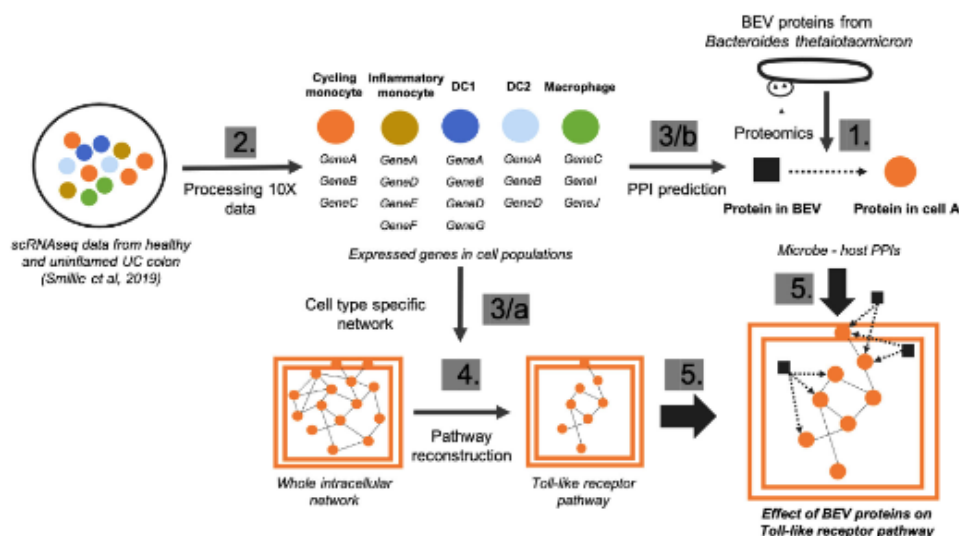


FIGURE 1 Computational workflow to analyse cell-type specific effects of BEVs. Numbers indicate the sequence of the main steps: 1, Extraction of BEV proteins from the proteomic dataset 2, Processing the raw single-cell transcriptomics from human colon 3/a, Creating cell-type specific network using protein-protein interactions from OmniPath (Türei et al., 2016) 3/b, Predicting protein-protein interactions (PPIs) between BEV and host proteins in each cell-type separately 4, Reconstruction of Toll-like receptor pathway using Reactome database (Jassal et al., 2020) 5, Combining cell-specific signalling with BEV targeted human proteins

3.2 | Functional analysis

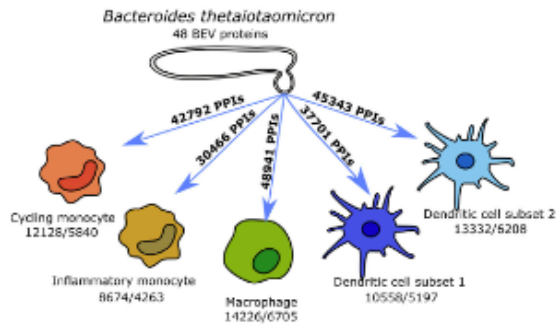
Cell-type specific BEV-host interactomes are complex due to the large number of proteins and interactions involved. Therefore, a functional analysis based on the GO database was initially carried out to identify the biological processes affected by microbial proteins in healthy (non-inflamed) and inflamed UC conditions. Most of the over-represented functions were overlapping among the different cell-types. However, comparing the cells under different conditions enabled us to identify specific effects of BEV proteins with the unique functions (Table S2 and S3, Figure S2–S4).

In the healthy state 209 functions were shared among the five cell-types containing basic cellular functions, such as chromatin organisation and macromolecule synthesis. Most of the unique processes (59) were found in inflammatory monocytes and were related to the endoplasmic reticulum (ER), apoptosis and myeloid cell differentiation. Counter to these results, in cycling monocytes—in terms of unique functions (16)—cell cycle-related processes were uncovered. Interestingly, among BEV targets in DC1 cells (20) somatic diversification of immune receptors and B cell apoptosis were uniquely over-represented. In contrast, negative regulation of myeloid leukocyte mediated immunity and cell differentiation were prominent in DC2 cells (11). Among BEV-targeted human proteins, the signalling pathways of both the epidermal growth factor (EGF) receptor and the regulation of transforming growth factor beta (TGF-beta) receptor were affected specifically in macrophages, based on 27 individual processes (Figure 3a).

BEV targets in the non-inflamed UC state included 174 overlapping processes that play vital roles in cell function. Uniquely over-represented functions were observed in inflammatory monocytes (30) that were similar in non-inflamed UC and healthy conditions and included positive regulation of the endoplasmic-reticulum-associated protein degradation (ERAD) pathway and intrinsic apoptotic signalling pathways. Among the 28 cycling monocyte-related annotations, similarly to the healthy condition, the cell cycle associated proteins were overrepresented. Here, we also found the negative regulation of G1/S phase transition overrepresented. Other targeted human proteins identified in this study are involved in the regulation of DNA repair and cyclin-dependent protein kinase activity, positive regulation of protein ubiquitination, and signal transduction by p53 class mediator. Whereas BEV proteins affected cell-cycle processes in DC2 (35), target proteins in DC1 (28) related to vesicle fusion, negative regulation of apoptotic signalling pathways, and the intracellular steroid hormone receptor signalling pathway. Among the 22 unique processes in macrophages, regulation of RAS protein signal transduction, base-excision repair, and diverse histone modification steps were identified (Figure 3b).

FIGURE 2 Interactions of 48 BEV proteins with monocytes, macrophages and dendritic cells in healthy (a) and UC (b) conditions. Number of expressed genes/number of interacting proteins are highlighted for each cell-type

a, Healthy condition



b, Ulcerative colitis

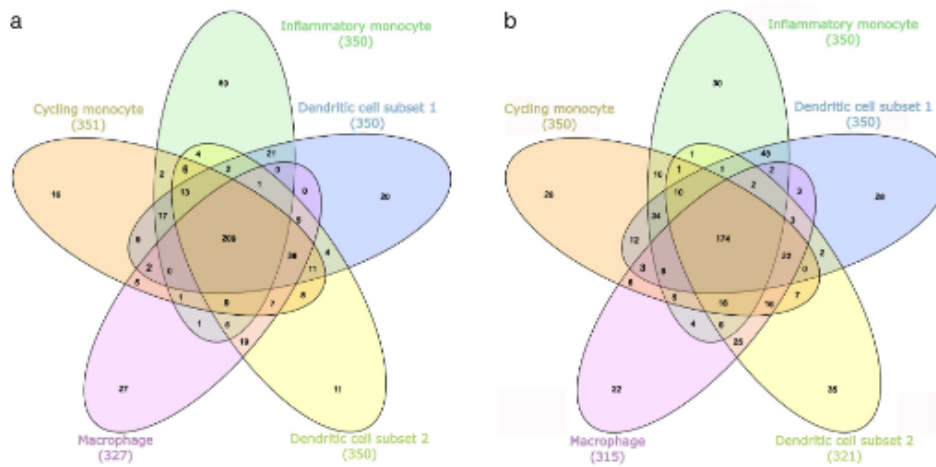
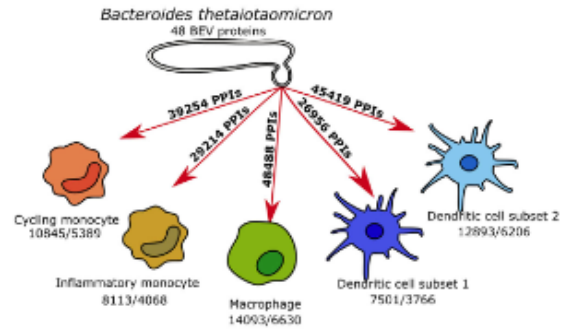


FIGURE 3 Overlap of biological processes over-represented in the BEV-host interactomes corresponding to cell-types in healthy (a) and uninfamed UC (b) conditions

In both conditions, macromolecule metabolism, DNA-related processes, and RNA-related processes were affected in all five cell-types by BEVs. Additionally, endoplasmic reticulum (ER)-stress response related processes and vesicle organisation and transport were influenced by BEVs in most cell-types.

3.3 | Effect of BEV proteins on TLR pathway in dendritic cells, monocytes and macrophages in healthy and UC conditions

As previous results established that Bt may alter immune pathways, we focused on the potential interactions between BEVs and TLR pathways. To do this, we created cell-type and condition specific signalling networks for BEVs and TLR pathways based on the scRNAseq data. These networks revealed that whilst the expression of TLR pathway-related transcription factors remained the same in both healthy and non-inflamed UC conditions in all examined cell-types, the level of TLR receptor expression was different amongst different immune cell-types. Due to the cell-type specific expression of different pathway members, BEV proteins established diverse interactions with immune cells (Figure 4–7).

Analysis of TLR pathways demonstrated cell specificity, especially in monocytes and DC1 cells, with differences occurring mostly at the level of receptor proteins. The BEV-targeted genes/proteins which show cell or condition specific expression may relate to the activation of different signalling pathways in healthy or UC. The network predictions also indicate that bacterial proteins can have intracellular immunomodulatory effects by binding to downstream elements of the TLR pathway (Figure 4–7). The expression of genes encoding transcription factors (TFs) did not show divergence between healthy and diseased conditions.

Dendritic cell subsets (DC1–DC2) show diverse characteristics regarding expression of TLR pathway members with fewer pathway members being expressed in DC1s. Also, in DC1 cells under healthy conditions, a large number of TLR pathway members were expressed in a condition-specific manner, including TLR1, 2, 3 and 7. In DC2 cells, three proteins were uniquely found in healthy (TLR3, MAPK7, and PP2R1B) and three in non-inflamed UC (TAB3, DNMI, and PELI3) conditions. In addition, more TLR receptors (TLR1–8, TLR10) were represented in DC2 cells compared to DC1 cells. However, a smaller number of differences were detected in the expression of TLR pathway members in DC2 cells compared to DC1 cells. While no receptor was targeted in DC1, TLR4 was identified as a potential BEV target in DC2 cells (Figure 4). These results raise an interesting issue regarding the DC subpopulation-specific LOS mediated activation via TLR2/4 mediated signalling: DC1s are likely to not bind LOS in diseased condition due to the lack of *TLR4* expression and health-related TLR2 expression. In contrast, *TLR2* and *TLR4* expressed in inflammation-related DC2 in both healthy and UC conditions, enabling LOS mediated activation in both health and disease states.

In monocytes, the majority of TLR pathway members were expressed with signals being spread through diverse paths due to a few key signalling proteins being represented only in the healthy or diseased network. In terms of cycling monocytes, *TLR1*, 2, 5, 6, 7, 8 were expressed at equivalent levels in both conditions, with *TLR4* expression strongly related to the diseased condition.

Amongst downstream signalling components, nine proteins were represented in the healthy state and two proteins in non-inflamed UC with BEV proteins being able to bind most of them. In inflammatory monocytes several condition-specific pathways were identified including TLR4 and TLR5 in non-inflamed UC, and TLR7 and TLR10 pathways in the healthy state. The network shows a high number of condition-specific proteins downstream (17 healthy and 12 UC specific proteins) (Figure 5). These results show that BEV proteins bind one TLR receptor (TLR4) which is expressed in both cell-types but only in inflammation-related monocytes in non-inflamed UC.

We analysed bulk RNAseq datasets to verify the role of BEVs on the TLR4 pathway in THP-1 monocytic cell line derived from human leukaemia (Tsuchiya et al., 1980). Results showed a more similar network to the output of the cycling monocyte scRNAseq data analysis. However, we found some differences in TLR pathways, revealing more potential for BEV-interacting proteins (PELI2-3, IRAK2, DNMI, RPS6K2, MAPK11) (Figure 6).

Based on the pipeline, macrophages depict no significant alteration in UC compared to the healthy state. While 9/10 receptors are potentially represented, TLR4 was the only candidate interacting with BEV proteins. MAPK10-11 helped spread the signal in healthy cells, while PELI2 was expressed only in diseased macrophages (Figure 7).

3.4 | Inhibition of TLR4 and TIRAP signalling abrogates BEV-driven monocyte activation

Our pipeline identified TLR4 as the only receptor associated with BEVs in cycling monocytes, DC1 and macrophage cells. We therefore investigated the effects of BEVs on TLR4-mediated activation of monocytes in BEV-monocyte co-cultures. Serial dilutions of Bt BEVs (3×10^9 – 3×10^7 /ml) were cultured with THP1 monocytes expressing an NF-kB reporter gene (THP1-Blue). These experiments were carried out in the presence or absence of the TLR4 inhibitor, CLI-095 (Ii et al., 2006; Kawamoto et al., 2008), which in pre-optimisation experiments using *E. coli* derived LPs was shown to selectively inhibit TLR4 mediated activation of NF-kB (data not shown). CLI-095 achieved significant levels of inhibition of BEV-mediated NF-kB activation with the highest level of inhibition (~37%) seen at the lower dose of BEVs (3×10^7). By comparison, THP1-Blue cells exposed to CLI-095

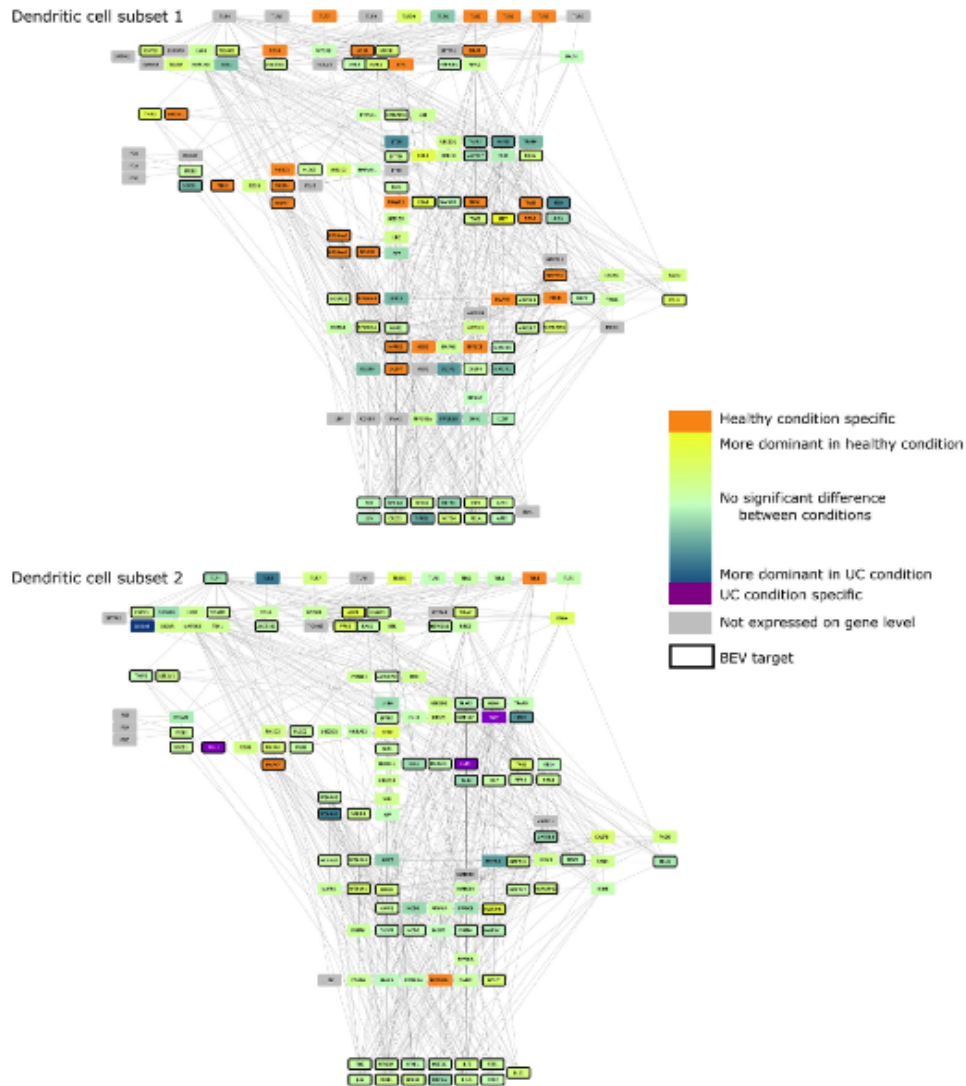


FIGURE 4 TLR pathway in DCs. Edges between nodes represent protein-protein interactions. Figures have been created with Cytoscape (Shannon et al., 2003)

in the absence of BEVs showed no significant inhibition ($P > 0.05$) of NF- κ B activation (Figure 8a). The inability to completely inhibit BEV-induced THP-1 activation by CLI-095 suggests TLR4-independent effects and pathways of BEVs induced NF- κ B-activation. This potential is revealed in the TLR signalling network that identifies the BEV interacting downstream pathway components.

To substantiate and confirm the BEV-TLR4 interaction in NF- κ B activation, we repeated the BEV-THP-1 co-culture experiments using an inhibitor of TIRAP, which is an intracellular adaptor protein and component of the TLR4 and TLR2 signalling pathways. Pre-incubation of THP1-Blue cells with the TIRAP inhibitor prior to incubation with 3×10^8 BEVs/ml demonstrated

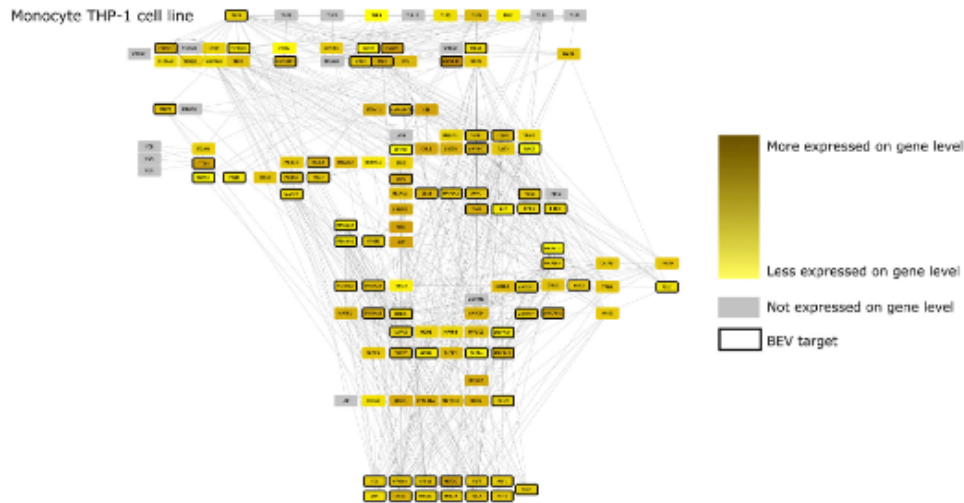


FIGURE 6 TLR pathway in THP-1 monocytes (based on bulk transcriptomic datasets). Edges between nodes represent protein-protein interactions. Figures have been created with Cytoscape (Shannon et al., 2003)

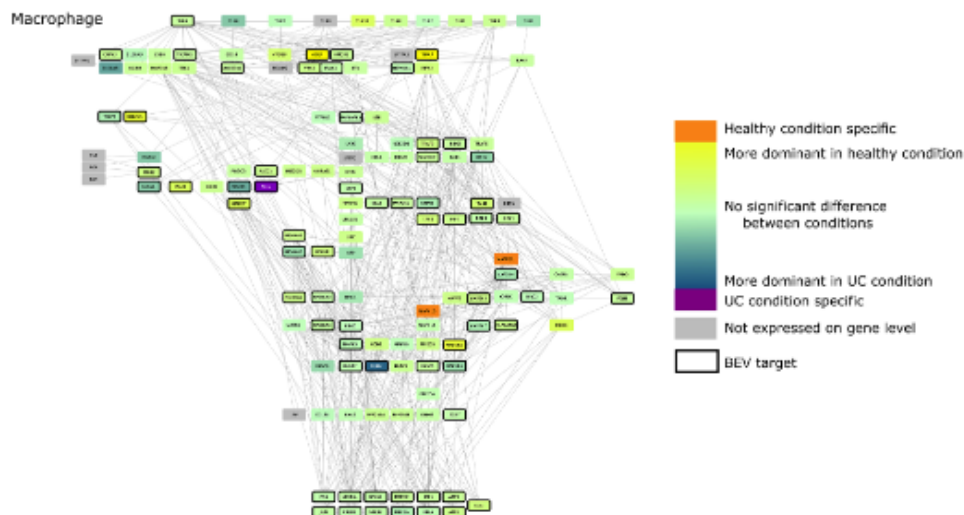


FIGURE 7 TLR pathway in macrophages. Edges between nodes represent protein-protein interactions. Figures have been created with Cytoscape (Shannon et al., 2003)

of biological pathways at cell-type specific resolution, which we utilised here to develop a computational workflow to identify the differential effects of BEV exposure on different populations of host immune cells.

Specifically, we examined proteins in BEVs generated by the major human commensal gut bacterium, Bt, which is a potential therapeutic agent in IBD (Delday et al., 2019). Hence, it is important to understand which, and how, specific cell-types are affected by Bt BEVs. Considering gene expression profiles are different not only among cells but also in the same cells under different conditions, the possible protein-protein interactions will vary between microbes and its host.

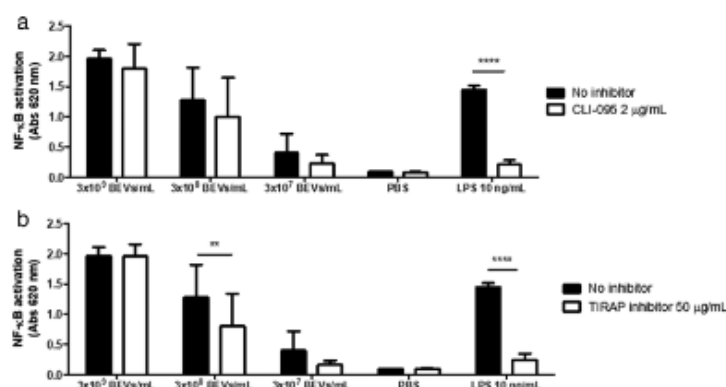


FIGURE 8 Inhibition of TLR4 and TIRAP signalling pathway abrogates THP1-Blue cells activation by Bt BEVs. NF- κ B activation was assessed using different doses of BEVs in 5×10^5 THP1-Blue cells/ml in the presence or absence of the TLR4 inhibitor CLI-095 (a) or TIRAP inhibitor (b) and by measuring absorbance at 620 nm after incubation with the colorimetric assay reagent Quanti-Blue. LPS from *E. coli* was used as a positive control and PBS as a negative control. Data are presented as mean \pm SD ($n = 9$). Significant differences were determined by using two-way ANOVA followed by Bonferroni's multiple comparison post hoc test. ** ($P < 0.01$), **** ($P < 0.0001$)

The computational pipeline combines single-cell transcriptomics with network biology approaches to reconstruct the interactomes and model the effect of Bt BEVs on different immune cells. In particular, we used a publicly available human scRNA-seq dataset to examine how Bt BEVs could potentially impact cycling monocytes, inflammatory monocytes, DC1s, DC2s, and macrophages in both the healthy, disease-free colon and the non-inflamed UC, diseased colon (Smillie et al., 2019). The output of the workflow highlighted that Bt BEVs have a large number of interactions with these immune cells. The majority of candidate interacting BEV proteins are catabolic enzymes with numerous non-specific connections with our workflow highlighting bacterial proteins carrying PDZ domains. PDZ domains can assemble signalling complexes recognising a C-terminal motif on the interacting protein and can change non-specific PPIs to more specific interactions. The two main functions of PDZ domains are related to protein location determination and signalling, including cell-cell communication (Harris & Lim, 2001; Lee & Zheng, 2010). Beside catabolic and PDZ domain-containing BEV proteins, we identified microbial helicases targeting specifically the human polymerase protein PAPD5. Binding of helicases to polymerase proteins is critical to initiate leading-strand DNA synthesis (Zhang et al., 2011). PAPD5 is also a well-known negative regulator of miR-21. Among the targets of this miRNA are genes involved in the immune responses and pathogenesis of autoimmune diseases, including IBD (Boele et al., 2014; Wang et al., 2016).

Despite the large overlap of connections, we identified in five types of immune cells unique functions triggered by Bt BEVs in the healthy and UC colon. For example, cell division is significantly enriched in cycling monocytes in the healthy state. In healthy conditions, bone marrow-derived monocytes circulate in the blood and differentiate to macrophages in various tissues. Therefore, the proliferation of monocytes is required to maintain a pool of tissue-specific monocytes and macrophages (Swirski et al., 2014). We also inferred that in UC, DNA repair activity might be influenced by BEV proteins interacting with cycling monocytes. Prior work has demonstrated that patients with UC have higher levels of mucosal oxidative DNA damage, even under non-inflamed conditions, which increases with the duration and severity of disease (Aslan et al., 2011; Beltrán et al., 2010; D'incà et al., 2004; Dincer et al., 2007; Lih-Brody et al., 1996). This is a potential explanation for the higher incidence of colorectal cancer in UC patients. Indeed, mice with chronic colitis that are deficient in a key DNA repair enzyme have increased susceptibility to developing colorectal carcinoma in response to oxidative stress (Liao et al., 2008). Our findings suggest Bt BEV proteins may play an important role in promoting DNA repair activity against oxidative DNA damage in cycling monocytes in patients with UC.

In inflammatory monocytes, BEV proteins upregulate apoptosis and the ERAD pathway in both healthy and UC states. Both these cellular processes are critical components of the unfolded protein response (UPR), which is important for resolving ER stress. Interestingly, our analysis also showed that BEVs influence ER-stress response related processes in most immune cell types we studied. In UC, several risk variants affect genes involved in these pathways and together with environmental factors (such as intestinal microbial dysbiosis, metabolites and/or inflammatory cytokines), disrupt the UPR in intestinal epithelial cells. The resultant unabated ER stress has been shown to precipitate intestinal inflammation. However, in monocytes and macrophages higher levels of UPR transcripts have been found in DSS-colitic mice compared to control mice, suggesting that the UPR may permit these cells to survive in the inflamed mucosal milieu of colitis (Jones et al., 2018). Thus, BEV proteins may help

promote resolution of ER stress and maintain the survival of inflammatory monocytes, macrophages, and other immune cells by upregulating key components of the UPR.

Dendritic cells are key antigen presenting cells and play important roles in innate and adaptive immunity including responses to microbial pathogens. Interactions between DCs and BEVs can direct inflammation in the gut (Durant et al., 2020). The microbiome can promote the differentiation of immature DCs into diverse subpopulations therefore maintaining immune homeostasis (Stagg, 2003).

We focused on the effects of BEVs on TLR pathways examining different activities and outcomes of the pathways. A prior study in mice indicates that Bt is capable of binding TLR4 (Coats et al., 2016). Here we discovered that Bt BEVs interact with TLR4 in a cell-type and condition specific manner. Of note, TLR4 expression is upregulated in the inflamed colonic mucosa of UC patients at both mRNA and protein levels (Hug et al., 2018; Levin & Shibolet, 2008).

Our pipeline was used to investigate if Bt BEV proteins might trigger immune response not only extracellularly via surface receptor interactions, but also by interacting with intracellular proteins. Based on our Bt BEV-human PPI network, a bacterial carboxyl-terminal protease (BT_2239) is predicted to bind TLR4. There is however no evidence as to how this enzyme affects TLR4 activation, although in chickens TLR15 can be triggered by microbial proteases (De Zoete et al., 2011). The domain-motif prediction approach of our pipeline provides more structural details about host-microbe interactions: BT_2239 interacts with TLR4 by a PDZ domain which catch a short motif —between 833 and 839 amino acid positions — at the end of the host protein's intracellular TIR domain. This suggests a possible intracellular BEV-TLR4 interaction separate or in addition to the extracellular LOS-TLR4 interactions. Evidence in support of this proposal was obtained using the TLR4 inhibitor CLI-095 which binds to and inhibits interactions with the intracellular domains of TLR4 and abrogated BEV-mediated NF- κ B activation of THP-1 monocytes. Further confirmation of the nature of BEV-TLR4 interactions was obtained by blocking the TLR2/4 adaptor protein TIRAP that similarly inhibited BEV-mediated NF- κ B activation of THP-1 monocytes. Of note, *in silico* analysis revealed cell and condition specific expression of TIRAP. It is expressed in both healthy and disease states in cycling monocytes, DC2s and macrophages. Nevertheless, the adaptor protein does not appear to be expressed by inflammatory monocytes and only under healthy conditions in DCs. PPI prediction revealed 19 BEV proteins which may interact with the TIRAP protein through diverse domain-motif interactions. The differential expression and the high number of interacting bacterial proteins highlights a potentially important role of TIRAP in BEV-related regulation of inflammation that could be explored further as a potential therapeutic target in IBD. The co-localization of Bt BEVs with various intracellular compartments and in particular, the nucleus, of intestinal epithelial cells that have acquired BEVs (Jones et al., 2020) demonstrates the feasibility of Bt BEVs interactions with various cytoplasmic constituents of host cells.

Whilst providing new and potentially important insights into BEV-host immune cell interactomes our pipeline is limited to one available scRNAseq dataset that describes gene expression in healthy and non-inflamed UC colonic mucosal cells. Some expressed genes could be missed with the 10X single-cell transcriptomics approach, and we also do not have corresponding protein levels (or their activities) in the cells of interest. In inferring a microbe-host PPI network, we assumed that all genes were translated to functional proteins regardless of post-transcriptional modifications that could affect protein abundance. Regarding the PPI predictions for the microbe-host interactions we used a limited list of domain-motif interactions from the ELM database and also only eukaryotic Pfam domains are represented in the analysis which means prokaryotic-specific domains (e.g., S41 proteases) are missing from the network analysis. Finally, our workflow cannot predict the activation or inhibitory effects of BEV proteins, but only whether they act on a particular receptor and pathway. Further investigations are needed to establish the binding mechanism and impact of for example, the BEV carboxy-peptidase on host TLR4 receptors. Despite these limitations, our pipeline provides a deeper insight into the effect of BEV proteins on host immunity at the protein level and shows the importance of condition and cell specificity. In addition to predicting the affected host processes supported by the literature our computational pipeline also identified new targets for experimental validation.

5 | CONCLUSION

We have developed a computational pipeline that predicts both the cell and condition specific effects of Bt BEV proteins on key host immune cell populations. Focusing on the inflammation-related TLR pathway, which plays a role in IBD pathogenesis, our workflow highlighted the importance of single-cell based analysis identifying differences in TLR4 receptor expression in diverse DC subpopulations. The current pipeline offers potentially interesting connection points and detailed mechanistic insight — using structural information about proteins — into the relationship between Bt and host immune cells that will aid in understanding how BEVs and their protein cargo may be of therapeutic value in IBD.

ACKNOWLEDGEMENT

This work was supported by the UKRI Biotechnological and Biosciences Research Council (BBSRC) UK grant awarded to the Earlham Institute (BB/J004529/1, BB/P016774/1, and BB/CSP17270/1) and to the Quadram Institute's Gut Microbes and Health Institute Strategic Programme (BB/R012490/1, BBS/E/F/000PR10353 and BBS/E/F/000PR10355). LG was supported by

the BBSRC Norwich Research Park Biosciences Doctoral Training Partnership grant number BB/M011216/1. JPT is an Academic Clinical Fellow funded by the National Institute of Health Research (NIHR).

CONFLICT OF INTEREST

The authors report no conflict of interest.

REFERENCES

- Andrighetti, T., Bohar, B., Lemke, N., Sudhakar, P., & Korcsmaros, T. (2020). MicrobioLink: An integrated computational pipeline to infer functional effects of microbiome-host interactions. *Cells*, *9*, 1278.
- Arenas, J. (2012). The role of bacterial lipopolysaccharides as immune modulator in vaccine and drug development. *Endocrine, Metabolic & Immune Disorders Drug Targets*, *12*, 221–235.
- Aslan, M., Nazligul, Y., Bolukbas, C., Bolukbas, F. F., Horoz, M., Dulger, A. C., Erdur, F. M., Celik, H., & Kocycigit, A. (2011). Peripheral lymphocyte DNA damage and oxidative stress in patients with ulcerative colitis. *Polskie Archiwum Medycyny Wewnętrznej*, *121*, 223–229. <doi:10.1111/1365-3113.11751111>
- Beltrán, B., Nos, P., Dasí, F., Iborra, M., Bastida, G., Martínez, M., O'connor, J.-E., Sáez, G., Moret, I., & Ponce, J. (2010). Mitochondrial dysfunction, persistent oxidative damage, and catalase inhibition in immune cells of naive and treated Crohn's disease. *Inflammatory Bowel Diseases*, *16*, 76–86.
- Boele, J., Persson H., Shin J. W., Ishizu Y., Newie I. S., Sokilde R., Hawkins S. M., Coarfa C., Ikeda K., Takayama K.-i., Horie-Inoue K., Ando Y., Burroughs A. M., Sasaki C., Suzuki C., Sakai M., Aoki S., Ogawa A., Hasegawa A., Lizio M., Kaida K., Teusink B., Carninci P., Suzuki H., Inoue S., Gunaratne P. H., Rovira C., Hayashizaki Y., del Hoon M. J. L. (2014). PAPD5-mediated 3' adenylation and subsequent degradation of miR-21 is disrupted in proliferative disease. *Proceedings of the National Academy of Sciences*, *111*(31), 11467–11472. <https://doi.org/10.1073/pnas.1317751111>
- Cecil, J. D., Sirisangtaksin, N., O'brien-Simpson, N. M., & Krachler, A. M. (2019). Outer Membrane Vesicle-Host Cell Interactions. *Microbiology Spectrum*, *7*.
- Chang, X., Wang, S.-L., Zhao, S.-B., Shi, Y.-H., Pan, P., Gu, L., Yao, J., Li, Z. S., & Bai, Y. (2020). Extracellular vesicles with possible roles in gut intestinal tract homeostasis and IBD. *Mediators of Inflammation*, *2020*, 1945832.
- Chronopoulos, A., & Kalluri, R. (2020). Emerging role of bacterial extracellular vesicles in cancer. *Oncogene*, *39*, 6951–6960.
- Coats, S. R., Hashim, A., Paramonov, N. A., To, T. T., Curtis, M. A., & Darveau, R. P. (2016). Cardiolipins act as a selective barrier to toll-like receptor 4 activation in the intestine. *Applied and Environmental Microbiology*, *82*, 4264–4278.
- Consortium, U. P. (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, *47*, D506–D515.
- De Zoete, M. R., Bouwman, L. I., Keestra, A. M., & Van Putten, J. P. M. (2011). Cleavage and activation of a Toll-like receptor by microbial proteases. *Proceedings National Academy of Science USA*, *108*, 4968–4973.
- Delday, M., Mulder, I., Logan, E. T., & Grant, G. (2019). Bacteroides thetaiotaomicron ameliorates colon inflammation in preclinical models of Crohn's disease. *Inflammatory Bowel Diseases*, *25*, 85–96.
- D'incà, R., Cardin, R., Benazzato, L., Angriman, I., Martinez, D., & Sturniolo, G. C. (2004). Oxidative DNA damage in the mucosa of ulcerative colitis increases with disease duration and dysplasia. *Inflammatory Bowel Diseases*, *10*, 23–27.
- Dincer, Y., Erzincan, Y., Himmetoglu, S., Gunes, K. N., Bal, K., & Akcay, T. (2007). Oxidative DNA damage and antioxidant activity in patients with inflammatory bowel disease. *Digestive Diseases and Sciences*, *52*, 1636–1641.
- Durant, L., Stentz, R., Noble, A., Brooks, J., Gicheva, N., Reddi, D., O'connor, M. J., Hoyles, L., Mccartney, A. L., Man, R., Pring, E. T., Dilke, S., Hendy, P., Segal, J. P., Lim, D. N. F., Misra, R., Hart, A. L., Arebi, N., Carding, S. R., & Knight, S. C. (2020). Bacteroides thetaiotaomicron-derived outer membrane vesicles promote regulatory dendritic cell responses in health but not in inflammatory bowel disease. *Microbiome*, *8*, 88.
- Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *Bmc Bioinformatics [Electronic Resource]*, *10*, 48.
- Fábrega, M.-J., Rodríguez-Nogales, A., Garrido-Mesa, J., Algieri, F., Badía, J., Giménez, R., Gálvez, J., & Baldomà, L. (2017). Intestinal anti-inflammatory effects of outer membrane vesicles from Escherichia coli Nissle 1917 in DSS-Experimental Colitis in Mice. *Frontiers in Microbiology*, *8*, 1274.
- Harris, B. Z., & Lim, W. A. (2001). Mechanism and role of PDZ domains in signaling complex assembly. *Journal of Cell Science*, *114 Pt 18*, 3219–3231.
- Hart, T., Komori, H., Lamere, S., Podshivalova, K., & Salomon, D. R. (2013). Finding the active genes in deep RNA-seq gene expression studies. *BMC Genomics [Electronic Resource]*, *14*, 778.
- Heberle, H., Meirelles, G. V., Da Silva, F. R., Telles, G. P., & Minghim, R. (2015). InteractiVenn: A web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinformatics [Electronic Resource]*, *16*, 169.
- Hickey, C. A., Kuhn, K. A., Donermeyer, D. L., Porter, N. T., Jin, C., Cameron, E. A., Jung, H., Kaiko, G. E., Wegorzewska, M., Malvin, N. P., Glowacki, R. W. P., Hansson, G. C., Allen, P. M., Martens, E. C., & Stappenbeck, T. S. (2015). Colitogenic bacteroides thetaiotaomicron antigens access host immune cells in a sulfatase-dependent manner via outer membrane vesicles. *Cell Host & Microbe*, *17*, 672–680.
- Hug, H., Mohajeri, M. H., & La Fata, G. (2018). Toll-like receptors: Regulators of the immune response in the human gut. *Nutrients*, *10*.
- Ii, M., Matsunaga, N., Hazeki, K., Nakamura, K., Takashima, K., Seya, T., Hazeki, O., Kitazaki, T., & Iizawa, Y. (2006). A novel cyclohexene derivative, ethyl (6R)-6-[N-(2-Chloro-4-fluorophenyl)sulfamoyl]cyclohex-1-ene-1-carboxylate (TAK-242), selectively inhibits toll-like receptor 4-mediated cytokine production through suppression of intracellular signaling. *Molecular Pharmacology*, *69*, 1288–1295.
- Jacobson, A. N., Choudhury, B. P., & Fischbach, M. A. (2018). The biosynthesis of lipooligosaccharide from bacteroides thetaiotaomicron. *MBio*, *9*.
- Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F., May, B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Weiser, J., ...D'Eustachio, P. (2020). The reactome pathway knowledgebase. *Nucleic Acids Research*, *48*, D498–D503.
- Jones, E. J., Booth, C., Fonseca, S., Parker, A., Cross, K., Miquel-Clopès, A., Hautefort, I., Mayer, U., Wileman, T., Stentz, R., & Carding, S. R. (2020). The uptake, trafficking, and biodistribution of bacteroides thetaiotaomicron generated outer membrane vesicles. *Frontiers in Microbiology*, *11*, 57.
- Jones, G.-R., Bain, C. C., Fenton, T. M., Kelly, A., Brown, S. L., Ivens, A. C., Travis, M. A., Cook, P. C., & Macdonald, A. S. (2018). Dynamics of colon monocyte and macrophage activation during colitis. *Frontiers in Immunology*, *9*, 2764.
- Kabeerdoss, J., Jayakanthan, P., Pugazhendhi, S., & Ramakrishna, B. S. (2015). Alterations of mucosal microbiota in the colon of patients with inflammatory bowel disease revealed by real time polymerase chain reaction amplification of 16S ribosomal ribonucleic acid. *Indian Journal of Medical Research*, *142*, 23–32.
- Kaparakis-Liaskos, M., & Ferrero, R. L. (2015). Immune modulation by bacterial outer membrane vesicles. *Nature Reviews Immunology*, *15*, 375–387.
- Kawamoto, T., Ii, M., Kitazaki, T., Iizawa, Y., & Kimura, H. (2008). TAK-242 selectively suppresses Toll-like receptor 4-signaling mediated by the intracellular domain. *European Journal of Pharmacology*, *584*, 40–48.

- Kuehn, M. J., & Kesty, N. C. (2005). Bacterial outer membrane vesicles and the host-pathogen interaction. *Genes & Development*, 19, 2645–2655.
- Kumar, M., Gouw, M., Michael, S., Sámano-Sánchez, H., Pancsa, R., Glavina, J., Diakogianni, A., Valverde, J. A., Bukirova, D., Čalyševa, J., Palopoli, N., Davey, N. E., Chernes, L. B., & Gibson, T. J. (2020). ELM-the eukaryotic linear motif resource in 2020. *Nucleic Acids Research*, 48, D296–306.
- Lee, H.-J., & Zheng, J. J. (2010). PDZ domains and their binding partners: Structure, specificity, and modification. *Cell Communication and Signaling*, 8, 8.
- Levin, A., & Shibolet, O. (2008). Toll-like receptors in inflammatory bowel disease—stepping into uncharted territory. *World Journal of Gastroenterology*, 14, 5149–5153.
- Li, K., Hao, Z., Du, J., Gao, Y., Yang, S., & Zhou, Y. (2021). Bacteroides thetaiootaomicron relieves colon inflammation by activating aryl hydrocarbon receptor and modulating CD4+T cell homeostasis. *International Immunopharmacology*, 90, 107183.
- Liao, J., Seril, D. N., Lu, G. G., Zhang, M., Toyokuni, S., Yang, A. L., & Yang, G.-Y. (2008). Increased susceptibility of chronic ulcerative colitis-induced carcinoma development in DNA repair enzyme Ogg1 deficient mice. *Molecular Carcinogenesis*, 47, 638–646.
- Lih-Brody, L., Powell, S. R., Collier, K. P., Reddy, G. M., Cerchia, R., Kahn, E., Weissman, G. S., Katz, S., Floyd, R. A., Mckinley, M. J., Fisher, S. E., & Mullin, G. E. (1996). Increased oxidative stress and decreased antioxidant defenses in mucosa of inflammatory bowel disease. *Digestive Diseases and Sciences*, 41, 2078–2086.
- Matsuura, M. (2013). Structural modifications of bacterial lipopolysaccharide that facilitate gram-negative bacteria evasion of host innate immunity. *Frontiers in Immunology*, 4, 109.
- Mészáros, B., Erdős, G., & Dosztányi, Z. (2018). IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Research*, 46, W329–W337.
- Rammelt, C., Bilen, B., Zavolan, M., & Keller, W. (2011). PAPD5, a noncanonical poly(A) polymerase with an unusual RNA-binding motif. *RNA*, 17, 1737–1746.
- Schwechheimer, C., & Kuehn, M. J. (2015). Outer-membrane vesicles from Gram-negative bacteria: Biogenesis and functions. *Nature Reviews Microbiology*, 13, 605–619.
- Bacteroides Thetaiootaomicron - an overview | ScienceDirect Topics. <https://www.sciencedirect.com/topics/immunology-and-microbiology/bacteroides-thetaiootaomicron>.
- Scott, N. A., & Mann, E. R. (2020). Regulation of mononuclear phagocyte function by the microbiota at mucosal sites. *Immunology*, 159, 26–38.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13, 2498–2504.
- Shen, Y., Torchia, M. L. G., Lawson, G. W., Karp, C. L., Ashwell, J. D., & Mazmanian, S. K. (2012). Outer membrane vesicles of a human commensal mediate immune regulation and disease protection. *Cell Host & Microbe*, 12, 509–520.
- Smillie, C. S., Biton, M., Ordovas-Montanes, J., Sullivan, K. M., Burgin, G., Graham, D. B., Herbst, R. H., Rogel, N., Slyper, M., Waldman, J., Sud, M., Andrews, E., Velomias, G., Haber, A. L., Jagadeesh, K., Vickovic, S., Yao, J., Stevens, C., Dionne, D., ... Regev, A. (2019). Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. *Cell*, 178, 714–730.e22.
- Stagg, A. J. (2003). The dendritic cell: Its role in intestinal inflammation and relationship with gut bacteria. *Gut*, 52, 1522–1529.
- Steimle, A., Michaelis, L., Di Lorenzo, F., Kliem, T., Münzner, T., Maerz, J. K., Schäfer, A., Lange, A., Parusel, R., Gronbach, K., Fuchs, K., Sälipo, A., Öz, H. H., Pichler, B. J., Autenrieth, I. B., Molinaro, A., & Frick, J.-S. (2019). Weak agonistic LPS restores intestinal immune homeostasis. *Molecular Therapy*, 27, 1974–1991.
- Steinbach, E. C., & Plevy, S. E. (2014). The role of macrophages and dendritic cells in the initiation of inflammation in IBD. *Inflammatory Bowel Diseases*, 20, 166–175.
- Stentz, R., Carvalho, A. L., Jones, E. J., & Carding, S. R. (2018). Fantastic voyage: The journey of intestinal microbiota-derived microvesicles through the body. *Biochemical Society Transactions*, 46, 1021–1027.
- Stentz, R., Wegmann, U., Guirro, M., Bryant, W., Ranjit, A., & Goldson, A. J. et al. (2020). Extracellular vesicles released by the human gut symbiont Bacteroides thetaiootaomicron in the mouse intestine are enriched in a selected range of proteins that influence host cell physiology and metabolism. <https://doi.org/10.21203/rs.3.rs-124947/v1>
- Supek, E., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *Plos One*, 6, e21800.
- Swirski, F. K., Hilgendorf, I., & Robbins, C. S. (2014). From proliferation to proliferation: Monocyte lineage comes full circle. *Seminars in Immunopathology*, 36, 137–148.
- Tsuchiya, S., Yamabe, M., Yamaguchi, Y., Kobayashi, Y., Konno, T., & Tada, K. (1980). Establishment and characterization of a human acute monocytic leukemia cell line (THP-1). *International Journal of Cancer*, 26, 171–176.
- Türei, D., Korcsmáros, T., & Saez-Rodriguez, J. (2016). OmniPath: Guidelines and gateway for literature-curated signaling pathway resources. *Nature Methods*, 13, 966–967.
- Wang, S., Wan, X., & Ruan, Q. (2016). The MicroRNA-21 in autoimmune diseases. *International Journal of Molecular Sciences*, 17(6), 864. <https://doi.org/10.3390/ijms17060864>
- Zhang, H., Lee, S.-J., Zhu, B., Tran, N. Q., Tabor, S., & Richardson, C. C. (2011). Helicase-DNA polymerase interaction is critical to initiate leading-strand DNA synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 108, 9372–9377.
- Zhang, M., Sun, K., Wu, Y., Yang, Y., Tso, P., & Wu, Z. (2017). Interactions between intestinal microbiota and host immune response in inflammatory bowel disease. *Frontiers in Immunology*, 8, 942.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Gul, L., Modos, D., Fonseca, S., Madgwick, M., Thomas, J. P., Sudhakar, P., Booth, C., Stentz, R., Carding, S. R., & Korcsmáros, T. (2022). Extracellular vesicles produced by the human commensal gut bacterium *Bacteroides thetaiootaomicron* affect host immune pathways in a cell-type specific manner that are altered in inflammatory bowel disease. *Journal of Extracellular Vesicles*, 11, e12189. <https://doi.org/10.1002/jev2.12189>