Department of Computer Science

Statistical and computational analysis of high-throughput ‘omics’ datasets for understanding the etiology and pathogenesis of autoimmune diseases

Juhi Somani

Aalto University

DOCTORAL
DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 90/2021

Statistical and computational analysis of high-throughput 'omics' datasets for understanding the etiology and pathogenesis of autoimmune diseases

Juhi Somani

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall 1017 TU1 of the school on 16 August 2021 at 12.00.

Aalto University
School of Science
Department of Computer Science
Computational Systems Biology Group

Supervising professor
Professor Harri Lähdesmäki, Aalto University, Finland

Thesis advisor
Professor Harri Lähdesmäki, Aalto University, Finland

Preliminary examiners
Professor Erik Bongcam-Rudloff, Swedish University of Agricultural Sciences, Sweden
Adjunct Professor Reija Anniina Autio, Tampere University, Finland

Opponent
Professor Erik Bongcam-Rudloff, Swedish University of Agricultural Sciences, Sweden

Aalto University publication series
DOCTORAL DISSERTATIONS 90/2021

ISBN 978-952-64-0435-6 (printed)
ISBN 978-952-64-0436-3 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-64-0436-3

Unigrafia Oy
Helsinki 2021

Finland

NORDIC SWAN ECOLABEL
Printed matter
4041-0619

A! Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Juhi Somani

Name of the doctoral dissertation
Statistical and computational analysis of high-throughput ‘omics’ datasets for understanding the etiology and pathogenesis of autoimmune diseases

Publisher School of Science

Unit Department of Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 90/2021

Field of research Information and Computer Science; Computational Systems Biology

Manuscript submitted 8 March 2021 **Date of the defence** 16 August 2021

Permission for public defence granted (date) 7 May 2021 **Language** English

□ **Monograph** ☒ **Article dissertation** □ **Essay dissertation**

Abstract

The primary function of the human immune system is to maintain our wellbeing by protecting ourselves from harmful substances and microbes (i.e. pathogens) that we continuously encounter through our surroundings. However, a variety of factors can lead to immune system dysfunction, which in turn can give rise to various incurable diseases, including autoimmune diseases (ADs), such as type 1 diabetes (T1D), immunoglobulin G4 related disease (IgG4-RD) and systemic sclerosis (SSc). In ADs, the immune system fails to distinguish between pathogens and body's own cells, and mistakenly attacks body's healthy tissues. Unfortunately, the factors that trigger ADs (i.e. etiology) and the molecular mechanisms by which ADs develop (i.e. pathogenesis) remain poorly understood. Genetics and environmental factors, such as gut microbiome, have been implicated in triggering or influencing the development of ADs, but the concerned mechanisms remain largely elusive. Therefore, the aim of this thesis is to further our understanding about the etiology and pathogenesis of ADs by performing robust statistical and computational analyses on high-throughput 'omics' datasets.

More specifically, one of the aims of this thesis was to study transcriptomics data from immune cells of T1D susceptible infants in order to identify gene expression markers that can aid in predicting the onset of autoimmunity and/or characterizing the disease progression. We found several genes to be associated with the pathogenesis of T1D, including *IL32* that has not been associated with T1D before. Another aim was to develop a personalised method that can robustly model longitudinal transcriptomics data from heterogeneous diseases and identify pathways associated with the pathogenesis of the disease. When applied to T1D data, this method was able to associate several key pathways to T1D pathogenesis that were missed by other methods. Additionally, this thesis aimed to study the gut microbial architecture of IgG4-RD and SSc patients (metagenomics data) and identify potential sources of microbial signals that may be contributing to the etiology of the two diseases. Among other interesting results, we found a specific strain of *Eggerthella lenta* that contains genes with the potential of influencing the immune system, to be significantly overabundant in both diseases. Finally, this thesis also aimed to identify the environmental and host-related factors that may be influencing the development of the highly dynamic early gut microbiome of T1D susceptible infants. In effect, we linked several new factors to the development of the early gut, such as household location at birth, maternal antibiotic treatments and average increase in height and weight per year, to name a few.

Keywords High-throughput data, statistical modelling, computational analysis, immune system, gut microbiome, genes, pathways, autoimmune diseases

ISBN (printed) 978-952-64-0435-6 **ISBN (pdf)** 978-952-64-0436-3

ISSN (printed) 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2021

Pages 208 **urn** http://urn.fi/URN:ISBN:978-952-64-0436-3

Preface

Most of the research work presented in this thesis was conducted in the Computational Systems Biology group led by Prof. Harri Lähdesmäki at the Department of Computer Science, Aalto University, Finland. A part of the thesis work, specifically pertaining to microbiome research and data analysis, was conducted in collaboration with the laboratory of Prof. Ramnik Xavier at the Broad Institute of MIT and Harvard, Greater Area of Boston, USA. Additionally, most of the data required to conduct the research work in this thesis was generously provided by Prof. Ramnik Xavier's laboratory and the Molecular Systems Immunology group led by Prof. Riitta Lahesmaa at Turku Bioscience Centre, University of Turku, Finland.

First and foremost, I would like to convey my heartfelt gratitude to my thesis supervisor, Prof. Harri Lähdesmäki, for his exceptional guidance, unwavering support and constant encouragement throughout the course of this research work. His thorough, far-reaching knowledge on a broad spectrum of different topics in Bioinformatics, Computational Systems Biology and Machine Learning, to name a few, has not only been beneficial for my thesis work, but also a constant source of inspiration for me. I have had the privilege of being mentored by Prof. Harri Lähdesmäki for several years even prior to this doctoral research and am extremely grateful to him for giving me the freedom of expressing ideas and encouraging me to explore cutting-edge methods for conducting research. I sincerely appreciate and admire him for his calm and graceful personality even in the face of adversity that has been a valuable guiding force during the compilation of this thesis. Additionally, I would like to thank him for enabling and unreservedly supporting my collaboration with Prof. Ramnik Xavier's laboratory, as well as facilitating my research visits to his laboratory at the Broad Institute.

I sincerely thank all my colleagues in the Computational Systems Biology group for their helpful companionship and fruitful scientific discussions. I am especially thankful to Dr. Sini Rautio for her unstinted peer support over the years and for our enjoyable discussions over lunch/coffee that I will always cherish. My sincere thanks to Dr. Tommi Vatanen for his able advice in the field of microbiome research and for always being available to clarify my queries regarding various topics. I warmly thank him for keenly checking the microbiome-related texts in this thesis as well as for helping me with the logistics during my research visit to Boston. I would like to thank Dr. Charles Gadd and Mr. Siddharth Ramchandran for sharing an office with me and for engaging in riveting conversations that often broke the monotony of long workdays.

I express my deep sense of gratitude to Prof. Ramnik Xavier for giving me with the unique opportunity of collaborating on many exciting projects of his laboratory. Especially, I am grateful to him for granting me with the opportunity to conduct research in his laboratory on several occasions. This provided me with ample scope of interacting with the leading experts in the field of microbiome research, thus expanding my horizon of knowledge in various related topics. I am thankful to all the members of Prof. Ramnik Xavier's laboratory for readily welcoming me into the lab and for helping me integrate into the Boston lifestyle and work culture. In particular, I would like to convey my sincere thanks to Dr. Hera Vlamakis for her valuable mentoring and guidance, and to Dr. Damian Plichta for successful collaboration and for promptly helping me understand specific details and logic behind some microbiome data analyses that previously eluded me.

I would like to express my earnest gratitude to Prof. Riitta Lahesmaa for giving me the opportunity to participate in several interesting projects from her group at Turku Bioscience Centre. I profusely thank the members of her group for the fruitful collaboration over the years. In particular, I would like to convey my sincere thanks to Dr. Henna Kallionpää and Dr. Soile Tuomela for productive exchange of ideas, lively discussions and constant interactions. In addition, they have been instrumental in my learning of the topics pertaining to type 1 diabetes, immunology and molecular biology. I greatly appreciate the support of Dr. Henna Kallionpää for lending me with her expertise in checking the immunology-related texts in this thesis.

I would like to extend my sincere appreciation to Dr. Mikael Knip for supporting me in my doctoral studies and constant encouragement throughout my research work.

I am deeply indebted to the pre-examiners of this thesis, Prof. Erik Bongcam-Rudloff and Docent, DSc. Reija Autio, for taking time out of their hectic schedules to carefully evaluate this thesis, and for providing constructive and insightful comments that undoubtedly improved the quality of the thesis.

The research work carried out in this thesis along with related logistics has been generously funded by the Academy of Finland, Emil Aaltosen Säätiö, Foundation for Aalto University Science and Technology, Juvenile Diabetes Research Foundation (JDRF), Instrumentariumin Tiedesäätiö, Diabetestutkimussäätiö (The Diabetes Research Foundation). I am extremely grateful to these funding organizations for supporting me financially, thus enabling me to conduct this research work smoothly.

I am extremely grateful to my friends and family members for constantly showering me with their immense love, support and motivation throughout this incredible journey. I would like to convey my deep gratitude to Dr. Tiina Murtola for her precious friendship, and for always being there for me whenever I needed her able guidance. My heartfelt appreciation also goes to Ms. Renee Aird for warmly welcoming me into her home in Boston during my research visits. Her amicable and caring personality made me feel at home whenever I visited Boston. I would like to thank my furry canine housemates in Boston—Philby, Darwin and Toby—for showing me immense love and affection that always brightened my days up. I would like to cordially thank my grandmother, Mrs. Manbhari Devi Somani, and grandparents-in-law, Mrs. Kamla Devi Maheshwari and Dr. Onkar Prasad Maheshwari, for their blessings and constant motivation to deliver my best. I wish to express my deepest gratitude to my parents-in-law, Mrs. Manju Maheshwari and Mr. Manoj Maheshwari, for always supporting, motivating and encouraging me to pursue my dreams. I am immensely thankful and indebted to my parents, Mrs. Aruna Somani and Prof. Mahesh Somani, for their undying love, affection and blessings that motivated me to thrive in my studies and career. I thank them for always believing in me, inspiring me to aim higher and knowing exactly what to say to boost my morale. Finally, words cannot express how grateful I am to my husband, Mr. Rohit Toshniwal, for being my best friend and standing by me through thick and thin for all these years. I thank him from the bottom of my heart for his endless love, moral support and unlimited supply of delicious Indian Chai (i.e. tea).

Espoo, July 5, 2021,

Juhi Somani

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals. Equal contributions are indicated by asterisks (*).

I Henna Kallionpää*, Juhi Somani*, Soile Tuomela*, Ubaid Ullah*, Rafael de Albuquerque, Tapio Lönnberg, Elina Komsi, Heli Siljander, Jarno Honkanen, Taina Härkönen, Aleksandr Peet, Vallo Tillmann, Vikash Chandra, Mahesh Kumar Anagandula, Gun Frisk, Timo Otonkoski, Omid Rasool, Riikka Lund, Harri Lähdesmäki, Mikael Knip, and Riitta Lahesmaa. Early Detection of Peripheral Blood Cell Signature in Children Developing β-Cell Autoimmunity at a Young Age. *Diabetes*, vol. 68, pp. 2024-2034, October 2019.

II Juhi Somani*, Siddharth Ramchandran*, Harri Lähdesmäki. A personalised approach for identifying disease-relevant pathways in heterogeneous diseases. *npj Systems Biology and Applications*, vol. 6, article number 17, June 2020.

III Damian R. Plichta, Juhi Somani, Matthieu Pichaud, Zachary S. Wallace, Ana D. Fernandes, Cory A. Perugino, Harri Lähdesmäki, John H. Stone, Hera Vlamakis, Daniel C. Chung, Dinesh Khanna, Shiv Pillai, Ramnik J. Xavier. Congruent microbiome signatures in fibrosis-prone autoimmune diseases: IgG4-related disease and systemic sclerosis. *Genome Medicine*, vol. 13, article no. 35, February 2021.

IV Tommi Vatanen, Damian R. Plichta, Juhi Somani, Philipp C. Münch, Timothy D. Arthur, Andrew Brantley Hall, Sabine Rudolf, Edward J. Oakeley, Xiaobo Ke, Rachek A. Young, Henry J. Haiser, Raivo Kolde, Moran Yassour, Kristiina Luopajärvi, Heli Siljander, Suvi M. Virtanen, Jorma Ilonen, Raivo Uibo, Vallo Tillmann, Sergei Mokurov, Natalya Dorshakova, Jeffrey A. Porter, Alice C. McHardy, Harri Lähdesmäki, Hera Vlamakis, Curtis Huttenhower, Mikeal Knip, and

Ramnik J. Xavier. Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. *Nature Microbiology*, vol. 4, pp. 470-479, March 2019.

Author Contributions

The contributions of all authors in each publication is defined below using the initials of each author's name. The doctoral candidate and author of this thesis, i.e. Juhi Somani, is indicated by the initials **J.S.** (in bold).

Publication I: "Early Detection of Peripheral Blood Cell Signature in Children Developing β-Cell Autoimmunity at a Young Age"

H.K., S.T., and U.U. were responsible for the interpretation of the results. **J.S.** conducted all bioinformatic analyses, including pre-processing of the high-throughout datasets, and performing all statistical and computational analyses. H.K., **J.S.**, S.T., and U.U. drafted the manuscript. H.K., **J.S.**, and U.U. prepared the figures. R.d.A., E.K., and O.R. were responsible for the isoform-specific IL32 RT-PCR assay and the intracellular IL32 staining in T cells and interpretation of the results. T.L. provided expertise in scRNA-seq study design, sample and data analysis, and interpretation of the results. H.S., J.H., T.H., A.P., and V.T. were responsible for sample collection, sample storage, and further clinical information of the children. V.C. and T.O. carried out the experiments and interpreted the results of the studies in pancreatic b-cells. M.K.A. and G.F. were responsible for experiments on virus-infected pancreatic islets. R.Lu. was responsible for study design, cell fractionation, sample analysis, and data production. H.L. was responsible for computational data analysis, interpretation of the results, editing the manuscript, and supervising **J.S.** M.K. was responsible for the DIABIMMUNE study design, sample collection, sample storage, clinical information for the children, directing of the clinical study, interpreting the results, and editing the manuscript. R.La. was responsible for study design, sample and data analysis, interpretation of the results, writing the manuscript, and supervision of the study. **All authors** contributed to the final version of the manuscript.

Publication II: “A personalised approach for identifying disease-relevant pathways in heterogeneous diseases”

J.S., S.R., and H.L. co-developed the method presented in this paper. S.R. implemented the method, and **J.S.** interpreted the results as well as supervised S.R. H.L. oversaw the whole project, and supervised **J.S.** and S.R. **J.S.** and S.R. prepared the figures. **All authors** contributed to the writing of this manuscript.

Publication III: “Congruent microbiome signatures in fibrosis-prone autoimmune diseases: IgG4-related disease and systemic sclerosis”

S.P. and R.J.X. coordinated and helped conceive the study. D.R.P., H.V., S.P. and R.J.X. designed the research. Z.S.W., A.D.F., C.A.P., J.H.S., D.C., D.K. and S.P. participated in data collection. H.V. and R.J.X. generated data. D.R.P. performed the pre-processing of the high-throughput data and **J.S.** performed the statistical and computational analyses. D.R.P. performed the accessory gene analysis. D.R.P. and **J.S.** interpreted the results (approximately 70% and 30%, respectively) and prepared the figures. M.P., H.L., D.K., H.V., S.P. and R.J.X. provided critical feedback. D.R.P., **J.S.**, H.V., and R.J.X. wrote the paper. **All authors** reviewed the paper.

Publication IV: “Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life”

T.V., D.R.P., **J.S.** and P.C.M. analysed the sequencing data and performed bioinformatic analyses. **J.S.** pre-processed the metadata and performed the association analyses between the metadata (i.e. external variables) and the microbiome community data. T.D.A., S.R., E.J.O., X.K., R.A.Y., H.J.H. and J.A.P. contributed to B. dorei isolate sequencing. A.B.H. and R.K. contributed to bioinformatic analysis. M.Y., K.L. and H.S. contributed to study design. J.I., S.M.V., R.U., V.T., S.M. and N.D. collected clinical samples. A.C.M., H.L., H.V., C.H., M.K. and R.J.X. served as principal investigators. T.V., D.R.P., **J.S.**, P.C.M., H.V., C.H., M.K. and R.J.X. drafted the manuscript. **All authors** discussed the results, contributed to critical revisions and approved the final manuscript.

Abbreviations

ADs	Autoimmune diseases
AIP	Autoimmune pancreatitis
AMP	Antimicrobial protein
APC	Antigen-presenting cell
APL	Adjusted profile likelihood
BCR	B cell receptor
BF	Bayes factor
CCD	Central composite design
cDNA	Complementary DNA
cgr	Cardiac glycoside reductase
CPM	Counts per million reads
CSS	Cumulative sum scaling
CTL	Cytotoxic T lymphocyte
DAA	Differential abundance analysis (or analyses)
DC	Dendritic cell
dcSSc	Diffuse cutaneous systemic sclerosis
DE	Differentially expressed (or differential expression)
DEA	Differential expression analysis (or analyses)
DEG	Differentially expressed gene
DPC	Density-peak clustering

E. lenta	*Eggertherlla lenta*
ERCC	External RNA Control Consortium
FACS	Flow-activated cell sorting
FC	Fold change
GADA	Glutamic acid decarboxylase
GC	Guanine-cytosine
GLMM	Generalized linear mixed model
GLM	Generalized linear model
GO	Gene Ontology
GP	Gaussian process
GWAS	Genome-wide association studies
HGP	Human Genome Project
HLA	Human leukocyte antigen
HT	High-throughput
HVG	Highly variable gene
IA-2A	Islet antigen-2
IAA	Insulin autoantibody
IAC	IgG4-associated cholangitis
IBD	Inflammatory bowel disease
IFN	Interferon
IgA	Immunoglobulin A
IgD	Immunoglobulin D
IgE	Immunoglobulin E
IgG	Immunoglobulin G
IgG4-RD	Immunoglobulin G4 related disease
IgM	Immunoglobulin M
IKZF1	Ikaros
IL32	Interleukin 32

IL	Interleukin
INS	Insulin
KEGG	Kyoto Encyclopedia of Genes and Genomes
KL	Kullback-Leibler
KO	KEGG Orthology
LCM	Laser capture microdissection
lcSSc	Limited cutaneous systemic sclerosis
LME	Linear mixed-effects
lncRNA	Long non-coding ribonucleic acid
LN	Lymph node
LRT	Likelihood ratio test
MANOVA	Multivariate analysis of variance
MDS	Multidimensional scaling
MHC	Major histocompatibility complex
miRNA	MicroRNA
ML-II	Type II maximum likelihood
MLE	Maximum likelihood estimation
MM	Mismatch
mRNA	Messenger ribonucleid acid
MSP	Metagenomic species pangenome
mtRNA	Mitochondrial ribonucleic acid
NB	Negative binomial
ncRNAs	Non-coding ribonucleic acid
NGS	Next generation sequencing
NK	Natural killer
ORF	Open reading frame
OTU	Operational taxonomic unit
PBMC	Peripheral blood mononuclear cell

PCA	Principal component analysis
PCoA	Principal coordinate analysis
PC	Principal component
PERMANOVA	Permutational multivariate analysis of variance
PM	Perfect-match
QC	Quality control
RA	Rheumatoid arthritis
RGE	Reverse graph embedding
RMA	Robust multi-array average
RNA	Ribonucleic acid
RNA-seq	RNA-sequencing
RPKM	Reads per kilobase per million reads
rRNA	Ribosomal ribonucleic acid
RTX	Rituximab
scRNA-seq	Single cell RNA-sequencing
SLE	Systemic lupus erythematosus
SNP	Single nucleotide polymorphism
SSc	Systemic sclerosis
t-SNE	T-distributed stochastic neighbour embedding
T1D	Type 1 diabetes
TC	Time-course
TCR	T cell receptor
TF	Transcription factor
TFBS	Transcription factor binding site
Tfh	T follicular helper
Th1	T helper 1
Th17	T helper 17
Th2	T helper 2

TMM	Trimmed mean of M-values
TNF	Tumor necrosis factors
TPM	Transcript-per-million
Treg	Regulatory T cells
tRNA	Transfer ribonucleic acid
TSS	Total sum scaling
UMI	Unique molecular identifier
UPGMA	Unweighted-pair group method using average linkages
WMS	Whole metagenome shotgun
WSC	Window before seroconversion
WT1D	Window before T1D diagnosis
ZnT8A	Zinc transporter 8

1. Introduction

We, humans, are an inseparable part of our environment and therefore, continuously encounter a variety of environmental components that may have the potential of influencing our well-being in either a beneficial or a detrimental manner. These environmental components that include microorganisms (i.e. microbes, such as bacteria, viruses, fungi, archaea, phages, and microbial eukaryotes), parasites, toxins, and allergens, to name a few, can come into contact with the human body through multiple routes, such as via the gastrointestinal tract (i.e. gut), the respiratory tract, and the skin [Janeway *et al.*, 2001]. Along with being subjected to the harmless components of the environment in our everyday lives, our bodies are also constantly exposed to various pathogens (i.e. harmful or infectious agents) that can lead to several diseases and thus jeopardise our well-being [Tortora and Derrickson, 2013].

Despite this continual exposure to pathogens, the occurrence of diseases is fortunately rare in an otherwise healthy human being. This is largely attributable to the defense mechanisms of the human body, i.e. the human immune system that protects the body from invading pathogens as well as helps the body to fight infections and other diseases by employing a complex network of organs, tissues, cells and molecules [Murphy and Weaver, 2017]. Undoubtedly, a healthy and properly functioning immune system is critical for maintaining good health and developing immunity towards harmful pathogens. An important characteristic of a healthy immune system is its ability to distinguish between the body's own healthy cells and molecules that are of no threat (i.e. self), and those that belong to the invading foreign bodies and are likely to be pathogenic (i.e. non-self). As the immune system develops and matures, especially during the early childhood years, it is trained to recognize self from non-self and thus becomes self-tolerant [Murphy and Weaver, 2017; Simon *et al.*, 2015].

However, a plethora of factors can influence the immune system; some of which can lead to its dysfunction (i.e. abnormality or impairment in function). A dysfunctional immune system often gives rise to various immune-related diseases, including autoimmune diseases (ADs) [Brodin and Davis, 2017; Wang *et al.*, 2015]. In ADs, the immune system fails to distinguish between self and non-self, and erroneously mounts an attack on the body's healthy tissues that leads to autoimmunity. There are more than 80 ADs in the world, including type 1 diabetes (T1D), systemic sclerosis (SSc), and immunoglobulin G4 related disease (IgG4-RD), that affect nearly 10% of the world's population [Gutierrez-Arcelus *et al.*, 2016; Theofilopoulos *et al.*, 2017]. In fact, T1D is among the most common of all chronic diseases in infants, especially in Finland that has the highest incidence of T1D in the world [Regnell and Lernmark, 2017; Atkinson *et al.*, 2014]. Although ADs are a diverse collection of diseases, most of them are believed to employ similar mechanisms for disease development. Even so, the etiology (i.e. the cause or trigger of a disease) and pathogenesis (i.e. development of the disease and/or the molecular mechanisms by which a disease develops) of most ADs remain poorly understood, and no cures exist for any of the ADs till date [Vojdani, 2014].

It has, however, been indicated through years of research that genetics and environmental factors play a significant role in the development of these diseases [Wang *et al.*, 2015]. Even though hundreds of genetic variants have been associated with ADs, their role in the breakdown of self-tolerance and development of autoimmunity remains unclear [Rosenblum *et al.*, 2015]. These genetic variants along with the possible presence of characteristic autoantibodies in the blood are some of the only markers that are currently available for predicting the onset of autoimmunity and assessing the progression of the disease; neither of which can do so in a reliable manner [Rose, 2016].

Moreover, while many environmental factors have been implicated in triggering ADs by inducing the breakdown of self-tolerance, the mechanisms by which they do so are still poorly understood [Murphy and Weaver, 2017]. The composition of the human gut microbiome[1] is one of the most prominent environmental factors (among many others) that has been implicated in triggering autoimmunity in genetically predisposed individuals and has already been associated with several ADs [Khan and Wang, 2020; Theofilopoulos *et al.*, 2017]. The gut commensals (i.e. the commensal microbes of the gut) are known to have a major impact on human health [Tibbs *et al.*, 2019; Belkaid and Hand, 2014]. One of the mechanisms by which they are believed to influence human health is by regulating the immune system [Belkaid and Harrison, 2017; Gianchecchi and Fierabracci, 2019]. From an early age, the gut commensals are believed to establish a cross-talk with the immune system that significantly influences the development and education of the immune system. The influence that the

[1] The human gut, like all surfaces of the human body, is colonized by trillions of microbial cells (collectively known as the gut microbiome) that are largely beneficial (i.e. commensal) and perform several important physiological processes that are essential for human health.

gut microbiome has on the immune system during the first 2-3 years of life is especially important as both the gut microbiome and the immune system are immature and in their developing stages at the time [Zhao and Elson, 2018; Gensollen *et al.*, 2016]. An infant's immune system is uniquely impressionable, and therefore the development and training it gets during these early years have life-long implications on host immune responses and health [Gianchecchi and Fierabracci, 2019; Belkaid and Hand, 2014]. On the other hand, the gut microbial composition during these early years is highly dynamic and largely modulated by several intrinsic and extrinsic factors, such as mode of delivery, use of antibiotics, host genetics, breastfeeding patterns, and diet, to name a few, which can then indirectly influence the immune system [Zhao and Elson, 2018; Liang *et al.*, 2018]. Generally, the establishment of a 'healthy'[2] gut microbiome (or a lack thereof) in early life is believed to encourage proper (or improper) development and education of the immune system and thus has long-term implications [Milani *et al.*, 2017]. Nevertheless, the exact mechanisms by which the gut affects or induces disease pathogenesis remain elusive and only a small fraction of the intrinsic and extrinsic factors that can significantly influence the early colonization of the gut has been identified, leaving many of them yet to be discovered [Liang *et al.*, 2018; Vieira *et al.*, 2014].

Therefore, there is an urgent need of improving our understanding of the etiology and pathogenesis of ADs. In other words, we need to identify reliable etiological signatures, such as presence of specific microbes, microbial genes, or other environmental factors, that may be involved in triggering the development of autoimmunity and thus ADs. Moreover, the pathogenesis of a disease can be understood, for instance, by identifying specific molecular alterations, such as changes in gene expression or disruptions in regulation of pathways, which take place before the onset of the disease and/or during disease progression. Identifying such molecular alterations in an autoimmune disease can lead to the discovery of better predictive biomarkers for that disease. Eventually, discovery of reliable etiological and predictive signatures can help tailor strategies for predicting, monitoring, treating and even preventing ADs. It is to be noted that the scope of this thesis is limited to studying T1D, IgG4-RD and SSc. However, the results from this thesis could perhaps aid in unraveling the pathogenic mechanisms that are shared by most ADs.

The recent advances in high-throughput (HT) technologies along with their ever decreasing costs have revolutionized biomedical research by making it more feasible to generate HT 'omics' datasets . In HT 'omics'

[2] The word 'healthy' is in quotes because an exact definition of a healthy gut microbiome that can be applied to all individuals does not exist [Huttenhower *et al.*, 2012]. In this context, a healthy gut microbiome generally refers to colonization by microbes that are functionally beneficial and carry out all the necessary processes [Lloyd-Price *et al.*, 2016].

datasets, several thousands to millions of biological molecules, such as deoxyribonucleic acid (DNA), ribonucleic acid (RNA), proteins, and metabolites, are interrogated in parallel in an unbiased manner [Hasin *et al.*, 2017; D'Argenio, 2018]. Needless to say, using 'omics' datasets in biomedical studies presents an unprecedented opportunity for elucidating the potential drivers and mechanisms that may be involved in the etiology and pathogenesis of diseases as well as discovering novel biomarkers [Lightbody *et al.*, 2019]. However, at the same time, analysing and interpreting such copious amounts of data can be challenging, especially when coupled with heterogeneity of the datasets and small sample sizes (common in human studies) [D'Argenio, 2018]. For instance, due to the heterogeneity in T1D disease, the data is often heterogeneous. Also, different 'omics' datasets often require specialized processing pipelines that account for the specific intricacies of the dataset [Hasin *et al.*, 2017]. Therefore, in order to ensure proper interpretation of the data and detection of truly significant results while ignoring the noise in the data, appropriate data analysis pipelines must be employed. This often includes quality control and other pre-processing steps as well as computational and statistical analyses of the data. For statistical analyses, several powerful methods already exist for modelling 'omics' datasets, but there is still a lot of scope for improvement and development especially when it comes to analyzing complex and heterogeneous datasets. For instance, barely any method exists that can appropriately model longitudinal data, while accounting for the heterogeneity of the disease.

Therefore, the aim of this thesis was to further our understanding about the etiology and pathogenesis of three autoimmune diseases namely, T1D, IgG4-RD and SSc, by analyzing HT 'omics' datasets using robust statistical and computational methods. Specifically, transcriptomics (i.e. study of gene expression) and microbiomics (i.e. study of the microbial communities) datasets have been analysed in this thesis. T1D was studied in Publications I, II and IV, whereas IgG4-RD and SSc were studied in Publication III. In Publication I, one of the main aims was to identify associated gene expression markers (from immune cells) that can aid in predicting the onset of autoimmunity in T1D susceptible infants and/or reflect upon the progression of the disease. Another aim of this study was to identify the specific types of immune cells that may be capable of expressing the gene expression markers. The aim in Publication II was to develop a new method that can: 1) identify differentially expressed genes (DEGs) by robustly modelling longitudinal gene expression data from heterogeneous diseases, such as T1D, in a personalised manner, and 2) summarize the DEGs on a pathway-level to identify disease-relevant pathways that can help predict the onset of consequential events in the pathogenesis of the disease and perhaps even be useful for biomarker identification. While the main aim of this study was to build the new personalised method, another aim was to apply the method on transcriptomics data from T1D susceptible infants to identify pathways that are enriched or disrupted in the 6 month window before the onset of autoimmunity and before clinical diagnosis of T1D, as well as those pathways that are perturbed over the whole time-course due to the disease phenotype. In Publication III, the main aim was to identify potential sources of microbial signals that may be contributing to the etiology of IgG4-RD and SSc. Finally, in Publication IV, one of the aims of the study was to investigate the influence of several intrinsic and extrinsic factors on the development of the early gut microbiome in T1D susceptible infants.

The structure of this thesis is as follows. After this introductory chapter, Chapter 2 proceeds by giving the basic biological background needed to understand the importance and relevance of the research conducted in this thesis. Altogether, the level of biological detail provided in Chapter 2 is designed to equip the reader with sufficient biological knowledge to understand and appreciate the study designs and results presented in this thesis. The chapter begins by introducing some of the major constituents of the human immune system; how the immune system develops self-tolerance; and how healthy immune responses are carried out in order to eliminate or inactivate the foreign pathogens. After establishing how a healthy immune system generally functions, the chapter dives into discussing how autoimmunity may arise and the factors that may lead up to it, such as genetic susceptibility and induction via environmental factors. Towards the end, the chapter also discusses about ADs that may develop due to autoimmunity. ADs, such as type 1 diabetes (T1D), immunoglobulin G4 related disease (IgG4-RD), and systemic sclerosis (SSc) are covered in more detail as these are the diseases that are primarily studied in this thesis.

Following the biological background, Chapter 3 begins by establishing the importance of 'omics' fields of studies and high-throughput (HT) technologies in biomedical research. Next, the chapter provides an overview of two prominent HT technologies availed in biomedical research, namely DNA microarrays and next generation sequencing (NGS). Thereafter, the remainder of the chapter focuses on discussing different types of transcriptomics and microbiomics datasets as well as their analyses. In this context, dataset analysis refers to the processing of the raw data and transforming it into a format suitable for further statistical and computational inference.

Chapter 4 presents several statistical and computational tools that can be used to analyse the processed 'omics' datasets, especially transcriptomics and microbiomics datasets, in order to address the research questions and fulfill the aims of the studies. The focus in this chapter will remain on covering specifically those analytical tools that have been employed in the publications of this thesis. In particular, this chapter presents computational techniques, such as dimension reduction, visualization, clustering and microbial diversity estimation techniques, that can be used to gain insights into the underlying structure and patterns of the data. Furthermore, it presents statistical modelling tools, such as linear models and Gaussian processes, that can be used for identifying covariates (such as disease status or intrinsic/extrinsic factors) that are associated with the variation in the observed data. The statistical models are commonly used for differential expression and differential abundance analyses in transcriptomics and microbiomics studies, respectively.

Finally, Chapters 5-8 present the main results of the four publications associated with this thesis and Chapter 9 provides the main conclusions and discussion of this thesis.

2. The human immune system and autoimmunity

The human body is continually exposed to various harmful pathogens, such as bacteria, viruses, fungi and parasites, as well as toxins and allergenic substances from the environment that can lead to a variety of diseases and threaten the normal homeostasis of the body. Yet, the incidence of disease is far less than what it could be due to the protection provided by the human immune system against the invading foreign bodies[Tortora and Derrickson, 2013]. To maintain homeostasis in the body, the immune system employs a complex network of lymphoid organs, tissues, cells, proteins and other molecules that recognize the presence of a variety of pathogens and aim at neutralizing or eliminating them. Central to these protective responses are its mechanisms of ensuring self-tolerance, i.e. its ability to distinguish between self from non-self, which prevents the immune system from mounting an attack on the body's own cells. There are two major constituents of the human immune system that are largely determined by the speed, specificity and memory of their responses: the innate immune system and the adaptive immune system [Parkin and Cohen, 2001; Murphy and Weaver, 2017; Chaplin, 2006]. Though initially thought to act independently, the two arms of the immune system interact with each other in a complementary and cooperative manner for the efficient recognition and eradication of pathogens [Chaplin, 2006; Parkin and Cohen, 2001; Clark and Kupper, 2005; Murphy and Weaver, 2017].

2.1 The innate immune system

The innate immune system represents a diverse collection of defense mechanisms that are present at birth and provide the initial host response towards invading pathogens and foreign substances [Tortora and Derrickson, 2013; Clark and Kupper, 2005]. Innate immunity is encoded in the germline genes of the host [Chaplin, 2006; Clark and Kupper, 2005] and provides an immediate response against invading pathogens (within a span of minutes or hours) in a non-specific manner, i.e. it does not target a specific pathogen and provides similar protection against all pathogens. Although non-specific, the innate immune cells are able to discriminate foreign molecules from self, for instance, via pattern-recognition receptors that recognize a broad repertoire of pathogen-associated molecular patterns present on microbes [Parkin and Cohen, 2001; Jain and Pasare, 2017]. However, after encountering a pathogen, it does not retain an immunological memory of the event for future reference [Tortora and Derrickson, 2013].

The first line of defense in innate immunity is provided by the physical and chemical barriers of the skin and mucous membranes of the body that prevent pathogens from penetrating and spreading throughout the body. In case a pathogen evades the first line of defense, it encounters the second line of defense that consists of natural killer (NK) cells, phagocytes, inflammation, fever and four main types of internal antimicrobial substances that prevent microbial growth: interferons (IFNs), complement proteins, iron-binding proteins and antimicrobial proteins (AMPs) [Tortora and Derrickson, 2013; Parkin and Cohen, 2001]. More specifically, subsequent to penetrating the first line of defense, the next non-specific defense consists of NK cells and phagocytes [Tortora and Derrickson, 2013]. NK cells are naturally occurring cytotoxic lymphocytes that represent 5-20% of the lymphocytes in human [Abel *et al.*, 2018]. They kill a wide variety of cells in the body that have either become cancerous or infected with a virus or other intracellular pathogens [Tortora and Derrickson, 2013; Abel *et al.*, 2018]. However, they do not kill the microbes inside the cells or released from the cells. The microbes are killed by phagocytes, such as neutrophils and macrophages, which are specialised cells that ingest and digest microbes and other particles, such as dying cells, in a process known as phagocytosis. When lymphocytes, phagocytes or epithelial cells are infected with viruses, they produce a class of cytokines (Section 2.2.2), known as interferons, that induce synthesis of antiviral proteins by healthy neighbouring cells [Tortora and Derrickson, 2013; Molnar and Gair, 2013]. In case of abrasions, chemical irritations, disturbances of cells, etc. caused by pathogens and toxins, an inflammatory response tries to dispose of the pathogen or toxin at the site of injury, prevent its further spread to neighboring tissues and initiate the healing process. Multiple types of proteins also play a prominent role in innate immunity, where complement proteins enhance certain immune responses, including phagocytosis; iron-binding proteins inhibit growth of certain microbes by reducing the availability of iron; and antimicrobial proteins kill a wide range of microbes and encourage immune response [Tortora and Derrickson, 2013].

2.2 The adaptive immune system

When innate immunity is insufficient to control the pathogenic activity of foreign bodies, the innate immune system activates the adaptive immune system through multiple highly orchestrated processes [Jain and Pasare, 2017; Clark and Kupper, 2005; Molnar and Gair, 2013]. The lymphatic system, made up of the thymus, bone marrow, lymph nodes, spleen, lymphatic nodules and lymph, is responsible for the adaptive immunity and some aspects of innate immunity. Adaptive immunity distinguishes from innate immunity in mainly three ways: (1) it takes much longer time to establish (days or even weeks), (2) it recognizes specific antigens released by the pathogens and provides an antigen-specific response, and (3) it preserves immunological memory for previously encountered antigens so that a subsequent encounter prompts a quicker and more intense immune response [Tortora and Derrickson, 2013; Molnar and Gair, 2013].

Antigens are large, complex molecules that are often proteins. They may originate from within the body as self-antigens or from pathogens and foreign substances as non-self antigens. They are immunogenic by nature, which means that they have the ability to provoke an immune response. In a healthy immune system, self-antigens are ignored by the immune system and only antigens recognized as non-self prompt an immune response. Entire microbes; certain parts of microbes, such as flagella, cell walls, capsules, and bacterial toxins; or non-microbial chemical components, such as pollen; may act as non-self antigens. Typically, immune response is triggered by small parts of a large antigen molecule, called epitopes. Remarkably, the human immune system can recognize at least a billion different types of epitopes [Tortora and Derrickson, 2013].

There are two types of adaptive immunity: (1) cell-mediated immunity, which involves lymphocytes called T cells and is particularly effective against intracellular pathogens and (2) humoral immunity (or antibody-mediated immunity), which involves lymphocytes called B cells and mainly combats extracellular pathogens in the body humors (i.e. fluids) outside cells. [Tortora and Derrickson, 2013; Molnar and Gair, 2013]. These two types of adaptive immune responses often work together to eradicate a large number of copies of antigens that circulate the body's humors as well as invade body's cells.

2.2.1 Major histocompatibility complex (MHC) molecules

Major histocompatibility complex (MHC) molecule, also known as human leukocyte antigen (HLA), is a glycoprotein that is displayed on the surface of each cell in the human body, except red blood cells [Tortora and Derrickson, 2013]. MHC molecules are encoded in a group of over 200 genes, commonly referred to as HLA genes, located on chromosome 6 in humans. These HLA genes are known to be the most polymorphic genes in the human genome. Due to the polygenic and polymorphic properties of MHC molecules, every individual possesses a unique set of thousand to several hundred thousand MHC molecules [Murphy and Weaver, 2017]. The function of MHC molecules is to bind to peptide fragments derived from self or non-self antigens and present them on the cell surface in order to help T cells assess the possible need to mount an adaptive immune response [Murphy and Weaver, 2017; Hewitt, 2003]. In a step called antigen processing, antigens are digested into peptide fragments and bound with MHC molecules to create antigen-MHC complexes. These antigen-MHC complexes are then inserted into the plasma membrane of the cell for presentation to lymphocytes in a consecutive step, called antigen presentation [Tortora and Derrickson, 2013]. In a healthy immune system, if the peptide fragment comes from a self-antigen, T cells ignore the antigen-MHC complex, whereas if the peptide is derived from a foreign protein, T cells initiate an immune response [Tortora and Derrickson, 2013; Hewitt, 2003].

There are mainly two types of MHC molecules: MHC class I and MHC class II. MHC class I molecules are present on all body cells, except red blood cells, where they process and display endogenous antigens that are present inside the body cells, such as self-proteins, viral proteins, bacterial toxins, and cancer-related abnormal proteins. MHC class II molecules appear on a special class of cells called antigen-presenting cells (APCs) that include dendritic cells, macrophages and B cells that are strategically located in those areas of the body where antigens are likely to penetrate innate defenses. APCs detect, engulf and process exogenous antigens, such as bacteria, bacterial toxins, parasitic worms, pollen and self-antigens that are present in the humors outside body cells and express the antigen-derived peptide fragments on the cell surface in an antigen-MHC complex [Tortora and Derrickson, 2013; Murphy and Weaver, 2017].

2.2.2 Cytokines

Cytokines are small proteins or peptides secreted by a broad range of cells throughout the body, including lymphocytes, APCs, endothelial cells, and fibroblasts, which convey regulatory signals from one cell to another cell that expresses the cytokine receptor [Tortora and Derrickson, 2013]. They are involved in regulating many routine cell functions, such as growth, differentiation and activation [Steinke and Borish, 2006; Murphy and Weaver, 2017; Parkin and Cohen, 2001]. Specifically, they play a central role in nearly every aspect of immunity and inflammation, including innate immunity, antigen presentation, and cellular recruitment and activity, as well as in determining the nature of the immune response [Steinke and Borish, 2006]. There are more than 60 different cytokines, including interleukins (ILs), interferons (IFNs), tumor necrosis factors (TNF) and chemokines, where many have redundant functionalities [Murphy and Weaver, 2017; Parkin and Cohen, 2001]. Some cytokines are produced by a variety of cell types, whereas others are specific to certain types of cells; and some cytokines influence a wide range of cell types, whereas others influence a certain few [Murphy and Weaver, 2017]. Moreover, a cytokine can be pro- or anti-inflammatory, or both. Among other factors, the cellular source, target, the phase at which a cytokine is released in an immune response, as well as the presence of other cytokines, govern the function of a cytokine, especially in the case of cytokines with both pro- and anti-inflammatory potential [Borish and Steinke, 2003].

2.2.3 T and B cells

T and B cells are two major types of lymphocytes, i.e. white blood cells, that are produced and made immunocompetent in the primary lymphatic organs, namely bone marrow and thymus. While the precursor cells of both populations are derived in the bone marrow from pluripotent hematopoietic stem cells, only the B cells complete their maturation there and become immunocompetent. Precursor T cells, on the other hand, migrate to the thymus where they proliferate, complete their maturation process and become immunocompetent [Tortora and Derrickson, 2013; Murphy and Weaver, 2017; Parkin and Cohen, 2001]. During their maturation process, each immature B or T cell produces a unique antigen-specific receptor through a process called somatic recombination that recognizes a single antigen. In this somatic recombination process, several interchangeable gene segments of receptor genes are randomly rearranged to form unique gene combinations that give rise to unique T cell and B cell receptors (i.e. TCRs and BCRs) per cell [Murphy and Weaver, 2017]. As a result, a remarkably diverse repertoire of millions of TCRs and BCRs are expressed on the surfaces of B and T cells that recognize a wide range of pathogenic antigens [Murphy and Weaver, 2017; Parkin and Cohen, 2001; Xing and Hogquist, 2012]. BCRs can directly recognize and bind to antigens, whereas TCRs require the antigens to be processed by other cells and presented to them in an antigen-MHC complex 2.2.1 [Murphy and Weaver, 2017].

Central tolerance

One drawback of producing an incredibly diverse set of receptors is the inevitable production of receptors that are non-functional or react to self-antigens, where the latter can result in autoimmune diseases. Therefore, as soon as the receptors are formed, the developing lymphocytes are subjected to central tolerance mechanisms in their respective sites of maturation, by which developing lymphocytes expressing non-functional or self-reactive receptors are eliminated [Murphy and Weaver, 2017; Xing and Hogquist, 2012]. In general, in a process of central tolerance known as positive selection, lymphocytes with receptors that weakly interact with self-antigens are allowed to survive; whereas in another process of central tolerance known as negative selection, lymphocytes with receptors that react strongly with self-antigens are eliminated (undergo apoptosis and die) or inactivated (via anergy, where the cell is alive but unresponsive to antigenic stimulations). Lymphocytes with receptors that have no affinity to self-antigens are usually deemed non-functional and are also eliminated or inactivated [Murphy and Weaver, 2017; Nemazee, 2017; Tortora and Derrickson, 2013]. These processes help ensure that the surviving repertoire of BCRs and TCRs is self-tolerant. In the thymus, dendritic cells and thymic epithelial cells present antigens on MHC molecules to the TCRs of developing T cells in order to assess their self-reactivity (self-affinity), thus determining their fate in the positive and negative selection processes. T cells that fail to bind to antigen-MHC complexes presented in the thymic environment, i.e. have no affinity to the self-antigens in that environment, are considered non-functional as they do not recognize the body's own MHC molecules, which is a crucial trait for a T cell to initiate an immune response [Tortora and Derrickson, 2013; Xing and Hogquist, 2012]. In fact, only 1-5% of the developing T cells survive elimination by central tolerance mechanisms and complete the maturation process [Tortora and Derrickson, 2013].

Peripheral tolerance

A limitation of central tolerance is that not all self-antigens, which the lymphocytes need to be tolerant of, are expressed at the primary sites of lymphocyte development. There are some self-antigens, such as food antigens and developmental antigens, that lymphocytes only encounter after leaving the thymus and bone marrow. Therefore, an additional layer of tolerance mechanisms exist in the immune periphery, known as peripheral tolerance. It manages and educates the self-reactive lymphocytes that have escaped central tolerance to circulate the lymph and colonize the secondary lymphatic organs, such as lymph nodes, spleen, and lymphatic nodules. Like central tolerance, peripheral tolerance also has several ways of dealing with self-reactive lymphocytes, including deletion by apoptosis, inactivation via anergy, and survival [Walker and Abbas, 2002; Xing and Hogquist, 2012; Nemazee, 2017; Murphy and Weaver, 2017]. In the periphery, mature lymphocytes are recognized as self-reactive usually when they react to self-antigens without receiving additional 'co-stimulatory' or 'danger' signals that are necessary for their activation (discussed further in Sections 2.2.4 and 2.2.5) [Gutierrez-Arcelus *et al.*, 2016; Tobón *et al.*, 2013; Murphy and Weaver, 2017].

Together, the central and peripheral tolerance mechanisms aim to prevent the development of autoimmune diseases.

CD4+ and CD8+ T cells

T cells that survive central tolerance mechanisms differentiate into two major types of T cells before exiting the thymus, namely T helper cells and cytotoxic T cells. T helper cells indirectly participate in eliminating foreign cells in the body by regulating the activity of other immune cells largely via cytokine signalling. Cytotoxic T cells, on the other hand, mount a direct attack on the foreign cells that destroys them, especially microbe-infected body cells and cancer cells. These two cell types are phenotypically distinguished from each other by the proteins expressed on their plasma membranes. T helper cells express a protein called CD4, which is why they are also known as CD4+ T cells; whereas, cytotoxic T cells express a protein called CD8, and hence are also know CD8+ T cells. CD4 and CD8 proteins are co-receptors that help maintain the interaction between the TCR of the cell and the antigen-MHC complex during antigen recognition process [Tortora and Derrickson, 2013; Murphy and Weaver, 2017; Parkin and Cohen, 2001; Molnar and Gair, 2013]. Approximately 1-5% of T cells that exit the thymus do not express either of the cell-surface proteins (CD4-CD8- T cells) and thus, do not recognize antigen-MHC complexes [Völkl, 2019; Parkin and Cohen, 2001].

2.2.4 Cell-mediated immunity

In cell-mediated immunity, T cells carry out an immune response that eventually leads to the elimination of the infected or foreign cells producing the pathogenic antigens. This immune response begins with the activation of a small number of T cells that express TCRs reactive to the pathogenic antigen. Subsequently, the activated T cells undergo a process called clonal selection, by which they proliferate and differentiate to form large populations of effector and memory cells with same antigen-specificity, that ultimately carry out the immune responses [Tortora and Derrickson, 2013; Murphy and Weaver, 2017].

T cell activation is a crucial process in the regulation of an adaptive immune response as well as maintenance of peripheral tolerance. Two simultaneous signals are needed to fully activate naive T cells. The first signal is provided by the antigen recognition process, in which TCRs recognize specific antigens presented to them by APCs; and the second signal is provided by co-stimulatory molecules, such as cytokines and plasma membrane molecules, in a process called co-stimulation, which enables adhesion of two cells for a prolonged period of time [Tortora and Derrickson, 2013; Sharpe, 2009], promotes cell survival, and increases cytokine production [Podojil and Miller, 2009]. Co-stimulation (or a lack thereof) is also one of the more prominent peripheral tolerance mechanisms that determines whether a T cell would become activated or anergic. If TCR-mediated antigen recognition process takes place in the absence of co-stimulation, the corresponding T cells are rendered unresponsive to subsequent antigenic stimuli [Tortora and Derrickson, 2013; Sharpe, 2009]. Lack of co-stimulation (or negative co-stimulation) often occurs to mediate peripheral tolerance and prevent autoimmunity when T cells encounter self-antigens [Xing and Hogquist, 2012].

Dendritic cells (DCs), which are a type of APCs, are the most potent activators of naive T cells as they excel in picking up virtually any type of antigen from the sites of infections, injury or vaccination in both lymphoid and non-lymphoid tissues (skin, intestine, lung, skeletal muscle and liver), and in presenting them on MHC class I and II molecules for antigen recognition by naive T cells [Murphy and Weaver, 2017; den Haan *et al.*, 2014; Dalod *et al.*, 2014]. Before encountering antigens, DCs are considered immature as they express low levels of MHC and co-stimulatory molecules. In response to activation by an antigen, DCs begin their maturation process, in which they increase the expression of MHC and co-stimulatory molecules at the cell surface while migrating to T-cell-rich zones in lymph nodes (LNs) [Dalod *et al.*, 2014]. Even though DCs primarily present antigens on MHC class II molecules, they can also present them on MHC class I molecules largely under two circumstances: 1) they are directly infected by a virus, or 2) they belong to a subset of DCs that are capable of presenting antigens from viral, bacterial and other sources, on MHC class I molecules without being infected, by a process called cross-presentation [Murphy and Weaver, 2017]. Since CD4+ and CD8+ T cells recognize antigens only when associated with MHC class II and class I molecules, respectively, the mature DCs in the LNs use antigens on MHC class II molecules to activate CD4+ T cells and MHC class I molecules to activate CD8+ T cells [Tortora and Derrickson, 2013; Dalod *et al.*, 2014]. Along with antigen presentation, DCs also provide relevant co-stimulation for full activation of the T cells through their cell-surface molecules and by secreting specific cytokines that communicate the nature of antigen and the type of immune response needed to the T cells [Dalod *et al.*, 2014].

The antigen recognition process and co-stimulation initiated by DCs (and in some cases, other APCs, such as B cells) are enough to activate CD4+ T cells. However, possibly due to the destructive actions of CD8+ T cells, they usually require additional co-stimulation from activated CD4+ T cells (by the same antigen) to become fully activated. Activated CD4+ T cells secrete various cytokines, including IL-2, that act as important co-stimulators for a variety of immune cells, such as CD8+ T cells, B cells, NK cells and themselves, and enhance their activity [Murphy and Weaver, 2017; Tortora and Derrickson, 2013]. Upon activation, naive CD8+ T cells differentiate into cytotoxic effector T cells that can recognize the foreign antigens associated with MHC class I molecules on infected cells outside the secondary lymphatic organs and kill the cells; whereas naive CD4+ T cells can differentiate into several subsets of effector T cells with different immunological functions, depending on the cytokines and other co-stimulatory signals they receive from the APCs during the activation process. The main CD4+ T cell subsets include T helper 1 (Th1), Th2, Th17, T follicular helper (Tfh) and regulatory T (Treg) cells [Murphy and Weaver, 2017]. They differ from each other in terms of the cytokines they produce and the type of pathogens they help in eradicating [Parkin and Cohen, 2001; Molnar and Gair, 2013].

2.2.5 Humoral immunity

In addition to T cells, the body also contains millions of different B cells, each capable of recognizing and responding to a specific antigen. These B cells form an integral part of humoral immunity, which protects the extracellular spaces of the body, i.e. body fluids or humors, from invading pathogens and toxins upon activation. Similar to T cells, activation of B cells takes place in the secondary lymphatic organs, but unlike TCRs, BCRs are capable of directly recognising and binding to their specific antigens. Nevertheless, most naive B cells require 'help' or co-stimulation from CD4+ T cells for optimal activation. Without the co-stimulation from CD4+ T cells, the B cells are usually rendered inactive or deleted, which commonly happens when a B cell recognizes self-antigens [Tobón *et al.*, 2013]. Notably, B cells can also act as APCs where they engulf certain antigens, process them into peptide fragments and present them on MHC class II molecules to activate or interact with antigen-specific CD4+ T cells, and in turn, receive co-stimulation for furthering their own activation. Upon activation, like T cells, B cells undergo clonal selection that produces a clone of plasma and memory cells. Plasma cells are the effector B cells that secrete hundreds of millions of copies of antibodies, i.e. immunoglobulins, specific to the antigen at hand [Tortora and Derrickson, 2013; Murphy and Weaver, 2017]. Antibodies are secreted forms of the receptors on the B cells [Murphy and Weaver, 2017] and are chemically similar to the antigen that triggered its production [Tortora and Derrickson, 2013]. They are the agents of the humoral immunity that circulate in the lymph and blood, and bind with its specific antigen whenever it is encountered in order to prevent its ability to infect cells. By binding to the antigens, antibodies act to disable them in many possible ways, such as neutralization, immobilization, agglutination, activation of complement system, and enhancement of phagocytosis [Molnar and Gair, 2013]. There are 5 types of antibodies, namely immunoglobulin G (IgG), IgA, IgM, IgD and IgE, that are characteristically and structurally different, but all aim to avert infection by pathogenic antigens or toxins in some way [Tortora and Derrickson, 2013].

2.3 Autoimmunity

Maintaining a healthy immune system is crucial for preserving homeostasis within the body and for efficiently protecting it against invading pathogens and potential diseases. However, a broad range of factors can perturb the human immune system and lead to its dysfunction [Brodin and Davis, 2017], which in turn gives rise to various chronic immune-related complications, including immune deficiencies, allergies and autoimmunity [Wang *et al.*, 2015].

Autoimmunity is the abnormal response of the adaptive immune system to self-antigens, where it fails to distinguish between pathogenic- and self-antigens, and erroneously mounts immune responses that damage body's healthy tissues [Murphy and Weaver, 2017; Gutierrez-Arcelus *et al.*, 2016]. Analogous to normal immune responses to pathogens, autoimmune responses are activated by specific antigens, called self-antigens, and give rise to effector lymphocytes as well as antibodies, called autoantibodies [Murphy and Weaver, 2017]. It creates an inflammatory environment involving multiple immune cells, cytokines, and other mediators that amplify the reaction [Rosenblum *et al.*, 2015].

As explained in Section 2.2.3, the immune system has installed a succession of tolerance mechanisms that act as synergistic checkpoints in order to eliminate self-reactive lymphocytes. However, the central tolerance mechanisms are not very stringent by design as they allow a small number of weakly self-reactive lymphocytes to escape elimination and exist in the periphery. This is because many weakly self-reactive lymphocytes also have the potential to respond to pathogenic antigens, and deleting them would dangerously narrow the repertoire of receptors available to respond to the foreign pathogens, which would inadvertently impair the immune system [Wang *et al.*, 2015; Walker and Abbas, 2002]. Therefore, self-reactive lymphocytes are a natural part of the immune repertoire that are not often activated by self-antigens, but if activated, they are usually suppressed by peripheral tolerance mechanisms.

Clearly, in order to induce tolerance mechanisms, the immune system should be able to efficiently distinguish between self and non-self. There are several clues that enable them to do so, but they are all imperfect and error-prone [Murphy and Weaver, 2017]. Isolated breakdowns in tolerance at one or more checkpoints are common in healthy individuals. However, it is the persistent breakdown of tolerance at multiple checkpoints and sustained activation of self-reactive lymphocytes that lead to loss of self-tolerance and autoimmunity, which in turn leads to various autoimmune diseases (ADs). Moreover, an imbalance between activated self-reactive T cells and Tregs (T cells that suppress immune response and help maintain homeostasis and self-tolerance [Xing and Hogquist, 2012]) is believed to play a role in the development of T cell-dependent ADs [Rosenblum *et al.*, 2015].

The mechanisms by which autoimmunity occurs are still incompletely understood [Murphy and Weaver, 2017; Gutierrez-Arcelus *et al.*, 2016], and thus the etiology and pathogenesis of most autoimmune diseases remain elusive [Vojdani, 2014; Wang *et al.*, 2015]. However, decades of research have indicated that most autoimmune diseases develop as a result of genetic susceptibility as well as induction via environmental factors, such as microbial exposure, dietary habits and other lifestyle choices [Wang *et al.*, 2015; Vojdani, 2014; Brodin and Davis, 2017].

2.3.1 Genetic susceptibility

In the past decade or so, hundreds of large-scale genome-wide association studies (GWAS) have been conducted to assess the genetic susceptibility of autoimmune diseases (ADs) in humans [Rosenblum *et al.*, 2015; Wang *et al.*, 2015]. These studies have identified hundreds of risk variants, typically single nucleotide polymorphisms (SNPs), in ~80 ADs [Gutierrez-Arcelus *et al.*, 2016; Brodin and Davis, 2017], where more than 80% of the risk variants fall in non-coding regions of the genome [Gutierrez-Arcelus *et al.*, 2016; Murphy and Weaver, 2017]. Also, most risk variants associated with ADs have small to moderate effect sizes (i.e. strength of association), making the contribution of each variant to a particular disease small [Gutierrez-Arcelus *et al.*, 2016; Rosenblum *et al.*, 2015]. In fact, in a majority of the ADs, the disease development is attributed to multiple genetic variants [Rosenblum *et al.*, 2015]. Furthermore, many of the risk variants, as well as the immune pathways that they are involved in, are common across different ADs, which indicates that many ADs have a common genetic etiology [Gutierrez-Arcelus *et al.*, 2016; Murphy and Weaver, 2017].

As established earlier, the MHC molecules (Section 2.2.1) play a critical role in distinguishing self from non-self as it is involved in the antigen presentation process. The HLA genes encoding these molecules are highly polymorphic and contain variants that have been known to be associated to ADs for over 50 years [Matzaraki *et al.*, 2017]. Interestingly, despite discovering hundreds of risk variants contributing to ADs via GWAS, till date, the polymorphisms in the HLA gene locus remain the most significantly and consistently associated genetic variants to ADs with large effect sizes [Wang *et al.*, 2015; Rosenblum *et al.*, 2015]. In fact, most associations are mediated by a handful of HLA genes [Gutierrez-Arcelus *et al.*, 2016]. It is hypothesised that certain variants (or genotypes) of MHC molecules may be more susceptible to presenting peptides from self-antigens to self-reactive T cells, or they may influence the shaping of the T cell receptor repertoire in the thymus during development by promoting positive selection of self-reactive T cells and avoid their negative selection

[Murphy and Weaver, 2017].

Unfortunately, regardless of the increased knowledge on the genetic variants associated with a number of ADs, the role of most of these genetic variants in the breakdown of self-tolerance and the development of autoimmunity is still poorly understood and remains a challenging task to resolve [Rosenblum *et al.*, 2015] perhaps because many of them occur in non-coding regions [Theofilopoulos *et al.*, 2017]. In fact, the way in which different HLA genotypes contribute to any ADs is still inadequately understood [Rosenblum *et al.*, 2015]. Moreover, most of these genetic variants, including HLA genotypes, fail to exhibit significant predictive strength to indicate onset of autoimmunity [Wang *et al.*, 2015]. Nevertheless, GWAS have been remarkable in identifying several genetic variants that would otherwise have not been possible [Wang *et al.*, 2015]. Notably, genetic variants are known to affect transcript levels or mRNA stability, which in turn could alter protein levels [Gutierrez-Arcelus *et al.*, 2016].

2.3.2 Environmental triggers

The alarming rate at which the prevalence and incidence of ADs have risen worldwide, especially in industrialized or urbanized countries, as well as the highly varying ranges of concordance rates in monozygotic twins, indicate non-genetic factors, i.e. environmental factors, to be involved in triggering autoimmunity in genetically susceptible individuals [Vojdani, 2014; Rose, 2016; Wang *et al.*, 2015]. In fact, environmental factors are considered to account for up to 70% of all ADs [Khan and Wang, 2020]. Some known environmental factors that have been linked to various ADs include—but are not limited to—infections [Kivity *et al.*, 2009; Danzer and Mattner, 2013], environmental toxins [Murphy and Weaver, 2017], dietary factors [Manzel *et al.*, 2014; Vieira *et al.*, 2014; Mackay, 2020], usage of drugs & vaccines [Wang *et al.*, 2015; Murphy and Weaver, 2017], vitamin D levels [Yang *et al.*, 2013; Murdaca *et al.*, 2019; Rebeca *et al.*, 2019], and gut microbial composition [Khan and Wang, 2020; Gianchecchi and Fierabracci, 2019; Opazo *et al.*, 2018].

The transition from an initial trigger to full-blown autoimmunity is not a well understood concept, but it is believed to be a cumulative process that is driven by a combination of factors, rather than a single one [Kivity *et al.*, 2009; Danzer and Mattner, 2013]. For instance, instead of a single infection, a 'burden of infections' from early childhood is considered to be responsible for the induction of autoimmunity in susceptible individuals [Kivity *et al.*, 2009].

Infections

Infections caused by infectious agents, such as bacteria, viruses and parasites, have long been suggested to play an important role in eliciting autoimmunity in susceptible individuals [Kivity *et al.*, 2009; Rosenblum *et al.*, 2015; Danzer and Mattner, 2013]. Indeed, almost every AD has been associated with one or more infectious agents [Vojdani, 2014]. Autoimmunity can be induced by infectious agents via multiple mechanisms, including molecular mimicry, epitope spreading, bystander activation, viral persistence and polyclonal activation, among others [Kivity *et al.*, 2009; Vojdani, 2014]. Molecular mimicry is believed to be the most likely mechanism by which infectious agents promote autoimmunity [Vojdani, 2014] and it has been implicated in the pathogenesis of several microbe-associated diseases [Danzer and Mattner, 2013]. In this mechanism, pathogenic antigens that bear structural similarity to self-antigens and are recognized by the same (auto)antibodies, may result in activation of T cells and production of (auto)antibodies that are cross-reactive with self-antigens [Murphy and Weaver, 2017; Danzer and Mattner, 2013; Vojdani, 2014]. Here, the structure of the pathogenic antigen need not be identical, but sufficiently similar, to the self-antigen [Murphy and Weaver, 2017].

Influence of the gut microbiome on host defenses

All areas of the human body, including the skin, gut (i.e. gastrointestinal tract), oral cavity, nasal cavity and urogenital tract, are colonized by trillions of commensal microbes (or microorganisms), such as bacteria, archaea, fungi, microbial eukaryotes, viruses and phages, that together constitute the human microbiome (or microbiota[1]). In fact, there are approximately as many microbial cells in or on the human body as human cells [Sender *et al.*, 2016; Allaband *et al.*, 2019] that humans have co-evolved with for millenia, establishing a symbiotic relationship [Liang *et al.*, 2018; Belkaid and Harrison, 2017; Vieira *et al.*, 2014]. Humans rely on the vast enzymatic properties and metabolic pathways of these microbes for regulating its various physiological processes [Belkaid and Hand, 2014; Liang *et al.*, 2018]. As a result, each body-site has evolved to harbour specific microbes essential for its physiological activities, resulting in strikingly different microbial communities between body-sites of an individual [Huttenhower *et al.*, 2012].

The gut—being the largest area of the body that is constantly exposed to environmental antigens and microbes [Gianchecchi and Fierabracci, 2019; Lynch and Pedersen, 2016]—houses the largest, most influential and a highly diverse reservoir of microbes (and antigens) in the human body

[1] The definitions of the terms 'microbiome' and 'microbiota' are inconsistent in literature and may be used interchangeably. Marchesi and Ravel [2015] have proposed a specific terminology where 'microbiota' refers to the collection of microbes in a defined environment, and 'microbiome' refers to the collection of microbes as well as their genomes (i.e. genes). However, the definition of 'microbiome' differs in Allaband *et al.* [2019] and clashes with the definition of 'metagenome' given in Marchesi and Ravel [2015]. Therefore, in this thesis, both of these terms refer to definition of 'microbiome' proposed by Marchesi and Ravel [2015].

[Tibbs *et al.*, 2019; Vieira *et al.*, 2014]. The gut microbiome consists tens of trillions of microbial cells that contain millions of unique genes [Zhu *et al.*, 2010; Lynch and Pedersen, 2016], which are approximately 150 times more genes than in the human genome [Zhu *et al.*, 2010]; and include about 2000 bacterial species [Opazo *et al.*, 2018]. In addition to microbes, the gut harbours the largest number of immune cells in the human body (up to 70% of the body's immune cells) [Takiishi *et al.*, 2017; West *et al.*, 2015; Mason *et al.*, 2008].

A rich and diverse gut microbiome plays an essential role in human health [Tibbs *et al.*, 2019; Belkaid and Hand, 2014]. In addition to playing a central role in nutrient and drug metabolism [Zhu *et al.*, 2010], from an early age, commensal microbes of the gut microbiome (i.e. gut commensals) calibrate nearly all aspects of the innate and adaptive immune systems, both local and systemic, [Belkaid and Harrison, 2017; Lazar *et al.*, 2018] in order to promote immune homeostasis [Gianchecchi and Fierabracci, 2019; Belkaid and Hand, 2014]. Here, immune homeostasis refers to the ability of protecting the human body from pathogenic microbes, while remaining tolerant to harmless food, commensals and self-antigens [Mason *et al.*, 2008]. By establishing a cross-talk with the immune system using a large repertoire of signalling mechanisms and pathways [Levy *et al.*, 2017; Lazar *et al.*, 2018; De Luca and Shoenfeld, 2019], the gut commensals significantly influence the development of the immune system, especially during early childhood [Zhao and Elson, 2018; Gensollen *et al.*, 2016; Tibbs *et al.*, 2019]. (In this context, the term development also refers to maturation and education of the immune system.) These actions train the immune system to differentiate between commensals and pathogenic microbes [Lazar *et al.*, 2018] as well as self and non-self antigens [Tibbs *et al.*, 2019], which in turn enables the immune system to shape and preserve the microbial ecology of the gut [Belkaid and Harrison, 2017; Levy *et al.*, 2017]. Moreover, gut commensals and components of the immune system compose the first two (of three) layers of the gut barrier [Assimakopoulos *et al.*, 2018], which contributes to the containment of the gut microbial cells [Assimakopoulos *et al.*, 2018; Opazo *et al.*, 2018]. This containment is especially crucial to human health as they prevent gut microbes (both commensal and pathogenic) from translocating to other parts of the body or into the systemic blood circulation, which can lead to systemic inflammatory response in various organs or sepsis [Assimakopoulos *et al.*, 2018; Belkaid and Harrison, 2017]. Furthermore, a thriving and symbiotic gut microbiome is also vital for inhibiting pathogens from invading the host and initiating infections [Belkaid and Hand, 2014; Opazo *et al.*, 2018] as well as in clearing existing infection [Pickard *et al.*, 2017]. This phenomenon is known as 'colonization resistance', which the gut commensals mediate via various mechanisms, such as direct killing, successfully competing for limited supply of nutrients, and promoting fast immune responses [Pickard *et al.*, 2017; Opazo *et al.*, 2018]. Thus, throughout life, the gut commensals plays an extremely significant role in protecting the host from various diseases, including intestinal, non-intestinal, autoimmune and inflammatory diseases [Brodin and Davis, 2017; Milani *et al.*, 2017; Khan and Wang, 2020].

The early microbial colonization of an infant's gut plays an instrumental role in the development of the immune system, and has long-term implications on host immune responses and health [Belkaid and Hand, 2014; Gensollen *et al.*, 2016; Rackaityte *et al.*, 2020]. The initial colonization of the gut happens *in utero* [Tanaka and Nakayama, 2017; Gianchecchi and Fierabracci, 2019; Rackaityte *et al.*, 2020]. However, the extensive colonization begins immediately after birth [Ferretti *et al.*, 2018; Gensollen *et al.*, 2016] and continues until 2-3 years (or ~1000 days [Lazar *et al.*, 2018]) of age. After these early years, the complex and highly dynamic gut microbial community of an infant stabilizes to resemble that of an adult [Zhao and Elson, 2018; Gensollen *et al.*, 2016] and remains henceforth relatively unchanged throughout life [Tibbs *et al.*, 2019; Gensollen *et al.*, 2016]. The establishment and development of the gut microbiome (or a lack thereof) during the first few years of life are driven and modulated by several extrinsic factors, such as maternal (gut) microbiome, [Ferretti *et al.*, 2018; Yassour *et al.*, 2018], mode of delivery (cesarean vs. vaginal delivery) [Gianchecchi and Fierabracci, 2019; Opazo *et al.*, 2018], breastfeeding patterns [Belkaid and Hand, 2014; Isolauri, 2012], use of antibiotics [Opazo *et al.*, 2018; Zhao and Elson, 2018] (by both mothers and infants [Ferretti *et al.*, 2018]), age at weaning [Tanaka and Nakayama, 2017], socio-economic status [Lazar *et al.*, 2018], geographical location [Arrieta *et al.*, 2014], exposure to farm environment [Brodin and Davis, 2017], infections [Wang *et al.*, 2015]; as well as intrinsic factors, such as host genetics [Zhuang *et al.*, 2019; Levy *et al.*, 2017] and gender [Mohammadkhah *et al.*, 2018]. While these factors can regulate the diversity, richness and composition of the microbes in the environment, the capability of accepting the microbes into the environment (i.e. allowing them to colonize) without an inflammatory response, can be explained by the unique nature of the infant immune system at the time [Belkaid and Hand, 2014]. During these early years of life, along with the gut microbiome, the infant immune system is also developing and is relatively immature [Gianchecchi and Fierabracci, 2019]. It is characterized by blunted inflammatory responses and a regulatory environment [Belkaid and Hand, 2014]. Simply put, the infant immune cells—unlike those of adults—favour regulatory responses, where they preferentially develop tolerance towards antigens and commensals introduced by the infant's new environment after birth, such as via food and gut microbes, ensuring establishment of a rich and stable gut microbiome without inflammatory responses [Lazar *et al.*, 2018]. Furthermore, the infant immune system is considered to be more durable and permissive to microbial instructions during infancy; providing a 'window of opportunity' for proper (or improper) immune development , and thus resilience (or susceptibility) towards diseases later in life [Zhao and Elson, 2018; Gensollen *et al.*, 2016]. A 'healthy' colonization of the gut by beneficial microbes during this critical window is believed to encourage proper development and training of the immune system, and thus promote immune homeostasis as well as long-term health. Even though a formal definition of what constitutes a 'healthy' infant gut microbiome has been difficult to ascertain, certain colonization trends of specific beneficial microbes for infant development have been inferred through various studies [Milani *et al.*, 2017].

However, owing to the high instability of the gut microbial composition during the early years of life, it is more vulnerable to environmental and host-related factors [Gensollen *et al.*, 2016; Milani *et al.*, 2017], such as those listed above. Certain factors can lead to reduced diversity or aberrant colonization of the infant gut, which in turn may result in significant defects or abnormalities in the development of the immune system and thus defective immunological tolerance [Gensollen *et al.*, 2016; Lazar *et al.*, 2018; Milani *et al.*, 2017]. In fact, many recent metagenomic studies have linked reduced diversity, aberrant colonizations or compositional shifts during infancy to illnesses that manifest during childhood or later in life, including type 1 diabetes (T1D), inflammatory bowel disease (IBD), asthma and metabolic disorders [Milani *et al.*, 2017; Kostic *et al.*, 2015]; although the mechanisms involved in disease pathogenesis remain largely elusive [Lazar *et al.*, 2018; Khan and Wang, 2020]. For instance, cesarean section (c-section)- delivery, which has shown to drive reduced gut microbial complexity in infants, has been associated with an increased risk of immune diseases, such as T1D [Milani *et al.*, 2017; Gianchecchi and Fierabracci, 2019]. Also, as proposed in the hygiene hypothesis (first suggested by Strachan in 1989 [Strachan, 1989]), the lack of infections (or insufficient microbial exposure) during childhood in westernised/urbanised countries (or cities) due to overuse of antibiotics, changes in diet, socioeconomic status, higher hygiene levels, etc., may result in underdeveloped gut microbiomes that lack the maturity and diversity required for establishing a stable and homeostatic immune system [Belkaid and Hand, 2014]. The hygiene hypothesis has been used to explain the recent worldwide increase in incidences of autoimmune diseases, especially T1D and IBD [Wang *et al.*, 2015]. Nonetheless, elucidating the exact impact of a specific microbe on human health or ascertaining the influences of certain extrinsic and intrinsic factors on the gut microbial composition, is still in its infancy and requires further research [Tibbs *et al.*, 2019].

After reaching an adult-like composition, the gut microbiome is considered mostly stable and symbiotic, but due to certain factors, such as use of antibiotics and drugs, diet, infections, host genetics, etc., dysbiosis may originate [Levy *et al.*, 2017]. Here, dysbiosis refers to the compositional and functional aberrations in the gut microbiome that are typically driven by an overgrowth of pathobionts (i.e. commensals that have the potential to be pathogenic under certain circumstances), loss of gut commensals, and/or loss of overall microbial diversity [Levy *et al.*, 2017; De Luca and Shoenfeld, 2019]. Gut microbial dysbiosis may increase local and systemic susceptibility to infections; induce chronic immune responses that may lead to inflammation and tissue damage [Lazar *et al.*, 2018]; as well as compromise gut barrier that may lead to increased microbial translocation [Assimakopoulos *et al.*, 2018] and gut permeability [Gianchecchi and Fierabracci, 2019]. Even though the precise mechanisms by which the gut microbiome affects disease pathogenesis are not well-known, many studies have associated gut microbiome dysbiosis to the pathogenesis of a plethora of diseases, including autoimmune diseases [Gianchecchi and Fierabracci, 2019; Theofilopoulos *et al.*, 2017].

2.4 Autoimmune diseases

Till date, there are more than 80 distinct autoimmune diseases (ADs) affecting 7.6-9.4% of the world's population [Gutierrez-Arcelus *et al.*, 2016] and have a significant effect on mortality and morbidity in populations [Wang *et al.*, 2015; Theofilopoulos *et al.*, 2017]. Being a diverse collection of diseases, there are various important demographic differences between different ADs [Cooper and Stroehla, 2003]. In particular, nearly all ADs disproportionately affect women more than men [Theofilopoulos *et al.*, 2017]; the age distribution among ADs is notably different [Wang *et al.*, 2015]; and specific ADs are more prevalent in certain countries or ethnic groups [Cooper and Stroehla, 2003; Gutierrez-Arcelus *et al.*, 2016].

Most ADs can be classified into two categories: organ-specific ADs, those affecting specific organs of the body, such as type 1 diabetes (T1D); and systemic ADs, those affecting various tissues of the body, such as systemic sclerosis (SSc), immunoglobulin G4 related disease (IgG4-RD), systemic lupus erythematosus (SLE) and rheumatoid arthritis (RA). Despite varying greatly in the organs or tissues they affect as well as in their clinical manifestations, many of the ADs employ similar mechanisms for immunopathogenesis [Vojdani, 2014].

2.4.1 Prediction and prevention

Unfortunately, no definitive cures exist for any of the ADs, probably because most ADs develop over a prolonged period of time in a clinically asymptomatic manner and become diagnostically detectable only after irreversible damage has taken place in the affected organs or tissues [Rose, 2016; Rosenblum *et al.*, 2015]. This has strengthened the need for etiological and predictive signatures (or factors or markers) that can enable early diagnosis of ADs and even provide an opportunity to prevent it altogether. Currently, genetic susceptibility conferred by HLA and non-HLA genes along with detection of characteristic autoantibodies in the serum are some of the most common and only indicators used to predict the onset of autoimmunity and to assess the prognosis of the disease. However, in most ADs, neither of these factors are able to do so in a reliable manner [Rose, 2016], especially since none can definitively assure a clinical illness in the future [Vojdani, 2008; Castro and Gourley, 2010], such as in T1D (Section 2.4.2) [Kallionpää *et al.*, 2014]. In fact, the presence of autoantibodies is indicative of an already active autoimmune response (i.e. loss of self-tolerance), which makes them poor predictive or prognostic biomarkers in most ADs, including T1D [Kallionpää *et al.*, 2014]. Furthermore, various environmental factors have also been implicated in the etiology of ADs, but the mechanisms by which they induce autoimmunity remain unclear [Theofilopoulos *et al.*, 2017; Vojdani, 2014].

Therefore, there is an urgent need of more reliable predictive signatures, such as gene expression profiles and pathways, as well as etiological signatures, such as microbes, microbial genes, and other environmental factors, that can help in monitoring and predicting the progression of ADs as well as establishing preventive treatments. These goals can efficiently be met by exploring high-throughput 'omics' datasets derived from host (i.e. humans) and gut microbial populations as well as by investigating the environmental factors that influence them.

2.4.2 Type 1 diabetes

Type 1 diabetes (T1D)—studied in Publications I, II and III—is a complex autoimmune disease that is characterised by the continuous infiltration of pancreatic islet cells by immune cells, particularly CD4+ and CD8+ T cells as well as NK cells and macrophages, that typically results in insulitis, i.e. islet inflammation [Bending *et al.*, 2012; Clark *et al.*, 2017; Regnell and Lernmark, 2017; DiMeglio *et al.*, 2018]. Over time, insulitis culminates in the selective destruction of insulin-producing β-cells that make up about 60% [Da Silva Xavier, 2018] of the pancreatic islet cells, and consequently leads to diminished insulin production [Bending *et al.*, 2012; Clark *et al.*, 2017].

T1D is among the most common chronic diseases in infants and adolescents; more common in males than in females [Atkinson *et al.*, 2014]. Globally, an annual increase of both the incidence and prevalence of T1D has been reported, with 2-3% increase in incidence per year. However, the disease incidence varies substantially between countries [DiMeglio *et al.*, 2018], including neighbouring countries (or regions) [Atkinson *et al.*, 2014]. For instance, Finland has the highest incidence of T1D in the world with more than 60 cases per 100,000 people each year [Regnell and Lernmark, 2017; Tuomilehto, 2013], which is about 6-times and 3-times higher than the incidence of T1D in the neighbouring countries (or regions) of Russian Karelia [Kondrashova *et al.*, 2013] and Estonia [Atkinson *et al.*, 2014; Tuomilehto, 2013], respectively.

Like most ADs, T1D can be caused by both genetic and environmental factors. An individual's genetic susceptibility to T1D can be conferred primarily using HLA genes, particularly two HLA class II haplotypes that are linked to 50% of the total genetic risk [Pociot and Lernmark, 2016; Atkinson, 2012]. These haplotypes are also used to identify individuals at high risk of T1D [Pociot and Lernmark, 2016; DiMeglio *et al.*, 2018] and classify all susceptible individuals into risk groups, termed as HLA risk classes [Kallionpää *et al.*, 2014]. Additionally, GWAS have identified over 60 non-HLA loci to be modestly associated with the risk of T1D, which has been shown to offer improved predictions of risk of T1D when combined with HLA loci screening [Pociot and Lernmark, 2016].

However, the rapidly increasing incidence of T1D globally and in genetically low-risk individuals; the large disparity in incidences between genetically similar populations that are separated by socioeconomic borders; concordance rates of only 30-60% in identical twins; <10% risk in children with a parent or sibling with T1D; and the increasing risk in second generation immigrants; cannot be explained by genetic factors and implicate a crucial role of environmental factors in the etiology as well as pathogenesis of T1D [DiMeglio *et al.*, 2018; Rewers and Ludvigsson, 2016; Dedrick *et al.*, 2020]. A plethora of environmental influences have been associated with T1D pathogenesis already, including dietary factors (e.g. breastfeeding), vitamin D insufficiency [Miettinen *et al.*, 2020], early-life viral infections (e.g. by enteroviruses) [DiMeglio *et al.*, 2018; Rewers and Ludvigsson, 2016], toxins [Rewers and Ludvigsson, 2016], gut microbial composition, and gut diversity [Vatanen *et al.*, 2018; Kostic *et al.*, 2015]. The hygiene hypothesis has also been strongly implicated in prevalence of T1D [Rewers and Ludvigsson, 2016; Atkinson, 2012; Kallionpää *et al.*, 2014]. However, the mechanisms by which environmental factors affect the disease process are poorly understood [Clark *et al.*, 2017] and largely debated [Regnell and Lernmark, 2017].

Currently, T1D-associated autoantibodies against islet antigens: insulin (IAA), glutamic acid decarboxylase (GADA), islet antigen-2 (IA-2A) and zinc transporter 8 (ZnT8A), are the first and only measurable biomarkers that can help predict the prognosis of T1D [Pociot and Lernmark, 2016; Kallionpää *et al.*, 2014]. Positive detection of one or more of these autoantibodies in at least 2 consecutive blood samples, called seroconversion [Ziegler *et al.*, 2013], can happen as early as 6 months of age [Atkinson *et al.*, 2014], with a median age of seroconversion for multiple autoantibodies at 2 years [Atkinson *et al.*, 2014; Ziegler *et al.*, 2013]. Notably,

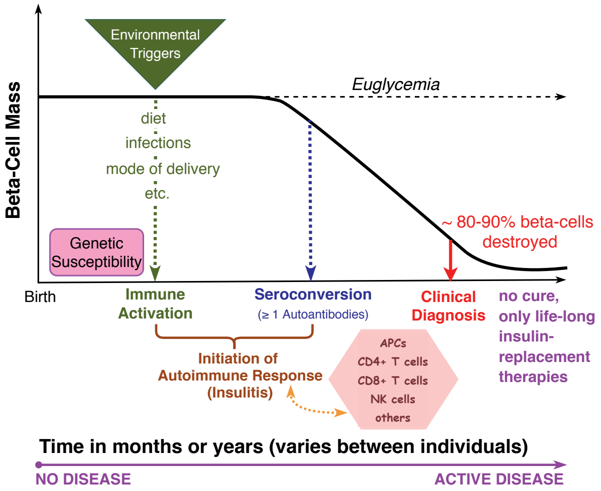


Figure 2.1. Disease progression of type 1 diabetes (T1D)

seroconversion is detected in >90% of the newly diagnosed individuals at the disease onset (~70% of diabetics present 3-4 autoantibodies and only 10% present 1 autoantibody [Regnell and Lernmark, 2017]) [Regnell and Lernmark, 2017; Atkinson *et al.*, 2014]. Generally, a larger number of circulating autoantibodies is indicative of a greater risk of rapid progression to clinical onset of T1D [Pociot and Lernmark, 2016; Kallionpää *et al.*, 2014]. However, seroconversion does not necessarily mean that autoantibodies are pathogenic [Pociot and Lernmark, 2016] since individuals can remain asymptomatic for months or years after seroconversion [Atkinson *et al.*, 2014], and some may not develop clinical T1D [Jasinski and Eisenbarth, 2005].

To date, there is no cure for T1D and it is managed with life-long insulin replacement therapies [DiMeglio *et al.*, 2018; Atkinson *et al.*, 2014]. The lack of cure is partly because the clinical diagnosis of T1D occurs at a very late stage of disease progression when 80-90% of the β-cells have already been destroyed [Kallionpää *et al.*, 2014]. Due to the heterogeneity of this disease [DiMeglio *et al.*, 2018; Pociot and Lernmark, 2016], predicting the onset of T1D remains a stiff challenge [DiMeglio *et al.*, 2018; Kallionpää *et al.*, 2014]. For instance, the age of seroconversion as well as clinical diagnosis of T1D varies extensively between individuals [Knip, 2017].

Figure 2.1 illustrates the loss of β-cell mass with age in a T1D susceptible individual, while highlighting the likely sequence of the events that take place in the pathogenesis of the disease as well as the involvement of the above-discussed factors.

2.4.3 Immunoglobulin G4 related disease and systemic sclerosis

Immunoglobulin G4 related disease (IgG4-RD) and systemic sclerosis (SSc)—studied in Publication III—are two complex autoimmune diseases that are characterized by chronic inflammation and generalised fibrosis in multiple organs as well as dysregulation of adaptive and innate immune responses [Stone *et al.*, 2012; Mahajan *et al.*, 2014; Brito-Zerón *et al.*, 2014]. IgG4-RD is a newly coined concept introduced in early 21st century [Mahajan *et al.*, 2014; Yamamoto *et al.*, 2014], which unifies a large number of single- or multi-organ fibroinflammatory conditions that were once regarded as entirely separate disorders, including autoimmune pancreatitis (AIP) and IgG4-associated cholangitis (IAC) [Stone *et al.*, 2012]. It has been reported in nearly every organ [Mahajan *et al.*, 2014] with similar serological and histopathological features regardless of the site of disease [Hubers *et al.*, 2018; Della-Torre *et al.*, 2015]. IgG4-RD has been reported mostly in Asian populations, but with increasing global awareness, incidences of the disease have been emerging in Europe and USA also [Yamamoto *et al.*, 2014; Kamisawa *et al.*, 2015]. Unfortunately, this disease has often been challenging to diagnose [Abraham and Khosroshahi, 2017] and is often overlooked as it may mimic other diseases [Celis *et al.*, 2017] and has neither sufficient genetic nor antibody biomarkers. Most of the genetic association studies in IgG4-RD are still in their infancy [Stone *et al.*, 2012] and have largely been conducted in Japanese and Korean populations with AIP, which have identified certain HLA haplotypes to be associated with AIP [Yamamoto *et al.*, 2014; Stone *et al.*, 2012]. High serum levels of IgG4, an anti-inflammatory antibody [Mattoo *et al.*, 2016; Mahajan *et al.*, 2014] with unique immunological properties [Hubers *et al.*, 2018], is a diagnostic hallmark of IgG4-RD. However, IgG4 is insufficient as a single diagnostic marker [Kamisawa *et al.*, 2015] since it is not elevated in all IgG4-RD patients [Stone *et al.*, 2012; Mahajan *et al.*, 2014] and the measurement of its concentration is error-prone [Kamisawa *et al.*, 2015; Mahajan *et al.*, 2014]. In fact, the role of IgG4 in the pathogenesis of IgG4-RD remains unclear [Yamamoto *et al.*, 2014; Mattoo *et al.*, 2016; Celis *et al.*, 2017] and it is usually not considered a driver of the pathogenesis [Kamisawa *et al.*, 2015].

SSc is a rare connective tissue disease [Steen, 2005] that may present patient-to-patient heterogeneity [Allanore *et al.*, 2015]. It is often classified into four major subgroups: limited cutaneous SSc (lcSSc), diffuse cutaneous SSc (dcSSc), sine scleroderma and overlap scleroderma, based on the extent of skin involvement, localization of the fibrosis, circulating autoantibodies and occurrence of other connective tissue disease [Allanore *et al.*, 2015; Denton and Khanna, 2017]. Moreover, SSc has a higher prevalence in Southern Europe, North America and Australia as well as in certain ethnic groups. For instance, a majority of African Americans develop the more severe form of the disease, i.e. dcSSc, that too at a younger mean age and have a higher mortality rate [Allanore *et al.*, 2015]. The genetic susceptibility of SSc has been strongly attributed to multiple HLA and non-HLA haplotypes [Burbelo *et al.*, 2019]. Additionally, distinct autoantibodies are commonly used to diagnose the disease, classify patients into subgroups [Allanore *et al.*, 2015], and predict the prognosis of the disease [Desbois and Cacoub, 2016], but they are not known to have a role in the pathogenesis of the disease [Steen, 2005]. Despite these biomarkers, early diagnosis of SSc is often challenging due to its similarity to some other ADs [McMahan and Hummers, 2013; Bellando-Randone, 2010].

While both diseases can lead to organ-failure, IgG4-RD is a relapsing-remitting disorder [Della-Torre *et al.*, 2015], whereas SSc has high mortality and morbidity with no cure and limited therapeutic options [Desbois and Cacoub, 2016; Steen, 2005; Patrone *et al.*, 2017]. Moreover, in line with most ADs, the etiology and pathogenesis of both IgG4-RD and SSc remain elusive [Allanore *et al.*, 2015; Kamisawa *et al.*, 2015; Yamamoto *et al.*, 2014]. However, it is known that the immunological characteristics of both the diseases are similar, where CD4+ T cells play a central role in the pathogenesis of the disease [Mattoo *et al.*, 2016; Allanore *et al.*, 2015] and are the primary immune cell type to infiltrate the fibrotic lesions [Mahajan *et al.*, 2014; Mattoo *et al.*, 2016; Laurent *et al.*, 2018]. Recently, an unusual subpopulation of CD4+ T cells that secrete IFN-γ, IL-1β and TGF-β, termed CD4+ cytotoxic T lymphocytes (CD4+ CTLs), were found to be clonally expanded in the blood and fibrotic lesions of both IgG4-RD and SSc [Mattoo *et al.*, 2016]. Also, B cells have been suggested to contribute to the pathogenesis of the diseases by acting as APCs to CD4+ CTLs and producing various autoantibodies [Sakkas and Bogdanos, 2016; Haldar and Hirschfield, 2018]. Furthermore, studies suggest these diseases to be driven by antigens (Section 2.2) [Haldar and Hirschfield, 2018; Kamisawa *et al.*, 2015], such as Annexin A11 [Hubers *et al.*, 2018] and galectin-3 [Perugino *et al.*, 2019], via mechanisms like molecular mimicry (Section 2.3.2) [Mahajan *et al.*, 2014]. These antigens are hypothesized to stem from environmental factors, such as long-term exposure to toxic industrial chemicals [Hubers *et al.*, 2018; Denton and Khanna, 2017] and microbial encounter.

3. High-throughput ‘omics’ datasets and their analyses

Over the last few decades, rapid advances in high-throughput (HT) technologies, such as DNA microarray and next generation sequencing (NGS), have revolutionized biomedical research by facilitating unprecedented development in the ‘omics’ fields of studies, such as genomics, epigenomics, transcriptomics, proteomics, metabolomics, and microbiomics, to name a few [D’Argenio, 2018; Hasin *et al.*, 2017; Manzoni *et al.*, 2018]. The ‘omics’ fields of study are the fast emerging disciplines in science and medicine that focus on obtaining global assessments of the concerned set of biological molecules in each field—such as DNA or genes in genomics and epigenomics, RNA in transcriptomics, proteins in proteomics, metabolites in metabolomics and microbes in microbiomics—instead of investigating single molecules at a time, as has been done previously [Yadav, 2007; Hasin *et al.*, 2017; Debnath *et al.*, 2010]. Indeed, pursuing these goals have been driven by the increased capacity, reliability, accuracy and availability of HT technologies. These technologies are capable of performing diverse simultaneous measurements on several thousands or millions of biological molecules at substantially reduced costs as well as time [D’Argenio, 2018; **?**; Hasin *et al.*, 2017]. Notably, the wealth of information that is supplied by the parallel interrogation of entire volumes of biological molecules in HT ‘omic’ datasets, has been highly beneficial in the biomedical domain. They have enabled the identification of large-scale disease-associated variations on genetic (or gene), protein, transcript, metabolite, microbial and other molecular levels that can aid in elucidating the drivers and mechanisms underlying disease development and progression as well as discovering novel biomarkers [D’Argenio, 2018; Lightbody *et al.*, 2019; Hasin *et al.*, 2017]. These insights into the etiology and pathogenesis of diseases will not only lead to improved disease diagnosis, monitoring and treatments, but can also provide an opportunity for early prediction and intervention of diseases in order to completely prevent the onset of disease or at least slow down its progression [D’Argenio, 2018]. However, the massive amounts of data that are generated by HT technologies coupled with small sample sizes (in human studies), heterogeneity and complexity of various datasets, as well as considerable differences between different 'omic' datasets, often pose great challenges in terms of analysis and interpretation [D'Argenio, 2018; Hasin *et al.*, 2017; Lightbody *et al.*, 2019]. While general pipelines usually exist for analysing each type of 'omics' dataset, many 'omics' fields of studies do not yet have a gold standard pipeline or method [Hasin *et al.*, 2017]. Moreover, novel HT strategies are continually being developed, for instance, single cell approaches [D'Argenio, 2018] where standard tools for analysis may not be applicable and may need special considerations [Lightbody *et al.*, 2019]. Therefore, there is a constant need of building robust pipelines and methods for analysing 'omics' datasets, which would ensure correct interpretation of the data and identification of truly significant results that may lead to promising medical discoveries [D'Argenio, 2018; Lightbody *et al.*, 2019].

This chapter focuses on outlining the general pipelines and methods for analysing transcriptomics and microbiomics datasets that are generated using HT technologies, such as DNA microarrays or next generation sequencing (NGS). The technologies and methods that have been used in the publications of this thesis have been indicated in the text whenever necessary. It should be noted that the computational methods that have been employed in the publications were chosen due to their wide-spread use at the time when the analyses were done and because they were considered state-of-the-art at the time by the experts in the field.

3.1 High-throughput technologies

3.1.1 DNA Microarrays

Microarrays have been one of the most widely-used HT technologies in the past few decades for performing simultaneous analysis on thousands of biomolecules, such as gene transcripts, [Sobek *et al.*, 2006; Zou *et al.*, 2008] in a single cost-effective experiment. After the development of the first DNA microarrays in 1990's [Fodor *et al.*, 1991; Schena *et al.*, 1995; Shalon *et al.*, 1996; Bumgarner, 2013], the microarray technology made rapid improvements in its efficiency, sensitivity and specificity, among other attributes; paving way for the development of other types of microarrays as well, such as protein, antibody [Miller and Tang, 2009; Sobek *et al.*, 2006], polysaccharide, lipid, and whole cell [Shiu and Borevitz, 2008] microarrays. However, this section will focus only on DNA microarrays as they are the type of microarray platforms that are used in this thesis (Publication II). Also, they are the most commonly used microarray platforms [Miller and Tang, 2009; Ventimiglia and Petralia, 2013] and have applications in multiple 'omics' fields of studies, such as transcriptomics, epigenetics, genomics, and microbiomics [Bumgarner, 2013]. While DNA microarrays have been predominantly used for investigating the gene expression profiles of cells, tissues [Bednar, 2000; Bumgarner, 2013; Stoughton, 2005] and microbes [Miller and Tang, 2009]; they have also been applied for SNP analyses (i.e. genotyping) [Heller, 2002; Bumgarner, 2013] and for identifying transcription factor binding sites [Bumgarner, 2013], to name a few of its applications.

Fundamentally, a typical DNA microarray consists of thousands of DNA sequences (i.e. probes) that are attached to a solid-surface support either by synthesis or immobilization techniques [Bumgarner, 2013; Miller and Tang, 2009]. These probes are usually either cDNA sequences or short oligonucleotides (25-60 bp) that are carefully selected to represent specific genes or genomic regions; to provide sufficient sensitivity and specificity; to have high coverage of the transcriptome or genome; and to avoid nonspecific or cross-hybridizations [Liu *et al.*, 2010]. The core aspect of DNA microarray analysis is the hybridization of a complex mixture of RNA or DNA fragments (i.e. targets) derived from cells, tissues or microbes, to their complementary probes on the microarray in order to identify (and/or quantify) the genes or genomic regions in the sample of interest [Stoughton, 2005; Ventimiglia and Petralia, 2013; Sobek *et al.*, 2006]. There are essentially two approaches for performing the hybridization step: (i) one-color approach, where targets from a single sample is hybridized on a microarray; (ii) two-color approach, where targets from two samples (e.g. case and control) are hybridized on the same microarray [Patterson *et al.*, 2006]. However, in this thesis, only the one-color approach will be discussed as it is the approach employed to collect the data analyzed in Publication II. Prior to hybridization, for some applications, the amount of target is first amplified [Stoughton, 2005; Ginsburg and Willard, 2009] to increase detectability [Ventimiglia and Petralia, 2013]. Subsequently, the target sequences are fluorescently labeled with one color and hybridized on the microarray [Miller and Tang, 2009]. After the hybridization step, the microarray is washed to remove unbound target sequences [Stoughton, 2005; Bumgarner, 2013]; and the target sequences that successfully hybridize to a probe are detected using fluorescent scanners [Miller and Tang, 2009; Stoughton, 2005] and the fluorescence intensity is measured using image processing software [Stoughton, 2005; Wu and Irizarry, 2004].

It is worth noting that there exists a variety of microarray platforms that differ from each other in many ways, such as in terms of fabrication techniques, nature of the probes, solid-surface support used, methods of labeling the targets for hybridization detection, and target detection methods. The basic types of DNA microarray platforms include spotted microarrays, in situ-synthesized microarrays, high-density bead arrays, and electronic and suspension bead microarrays, among others [Miller and Tang, 2009]. In ‘*in situ*-synthesized microarrays’, oligonucleotide probes (25-60 bp in length; depending on the manufacturer) are directly synthesized on the solid-surface support, which is usually quartz wafer, using photolithography techniques [Miller and Tang, 2009; Ginsburg and Willard, 2009]. These are extremely high-density microarrays that can be used to study tens of thousands of genes or genomic areas simultaneously [Miller and Tang, 2009]. Affymetrix GeneChips are the most widely-known and popular examples of this microarray that use 25-bp probes [Ginsburg and Willard, 2009]. Oligonucleotide probes are short and are generally considered to provide increased flexibility, sensitivity, specificity and accuracy [Ventimiglia and Petralia, 2013; Miller and Tang, 2009]. They are usually designed to capture regions of genes or the genome that have the lowest similarity with sequences from other genes or genomic regions in order to avoid cross-hybridization [Liu *et al.*, 2010; Bumgarner, 2013]. Also, in order to control for random hybridization events, Affymetrix GeneChips introduced the concept of using probe sets. Here, each probeset consists of pairs of probes, where one probe is an exact compliment to the target transcript, called perfect-match (PM) probe, whereas the other differs from the PM probe by 1-bp in the middle position, called mismatch (MM) probe [Liu *et al.*, 2010; Ginsburg and Willard, 2009]. Each PM maps to a different part of the target transcript [Jiang *et al.*, 2008]. Usually, there are 8 to 16 pairs of these PM and MM probes in each probe set [Liu *et al.*, 2010], where the MM probes act as a negative control for binding specificity [Miller and Tang, 2009; Ginsburg and Willard, 2009]. However, some of the advanced Affymetrix GeneChips, such as Human Genome U219 Array Strip, consist only of PM probes [Affymetrix, 2010].

While DNA microarray technology is still used in some applications, it has largely been superseded by next generation sequencing technologies in at least gene expression studies [Metzker, 2010; Lightbody *et al.*, 2019].

3.1.2 Next generation sequencing

The first sequencing technology (that had limited throughput [Reuter *et al.*, 2015]), i.e. Sanger sequencing, was introduced in 1977 [Sanger *et al.*, 1977; Metzker, 2010] and was used for completing the first draft of the human genome sequence between 1990 and 2001 as a part of the Human Genome Project (HGP) [Lander *et al.*, 2001] at the cost of about $2.7 billion [NHGRI, 2020]. Post 2001, tremendous progress was made in genome sequencing technologies, which brought about the first truly HT sequencing platform, i.e. next generation sequencing (NGS), in 2005 [Goodwin *et al.*, 2016; Mardis, 2017] with larger throughput than Sanger sequencing [Lightbody *et al.*, 2019]. Over the past decade or so, vast improvements have been made in nearly all aspects of the NGS pipeline in order to deal with the complexities of genomes and to increase performance as well as decrease the cost and processing time [Lightbody *et al.*, 2019;

Goodwin *et al.*, 2016]. Indeed, NGS technologies have gained 100-1000 times more capacity in the past years [Goodwin *et al.*, 2016], giving high sequence coverage per instrument [Lightbody *et al.*, 2019; Mardis, 2017]. They have decreased the cost of sequencing to about $1000 per human genome [Goodwin *et al.*, 2016; Mardis, 2017; Lightbody *et al.*, 2019], which will continue to decrease in the future as well [D’Argenio, 2018]. Also, some NGS platforms, such as the latest ones by Illumina, can process a human genome in an hour [Lightbody *et al.*, 2019]. Most importantly, unlike microarray technology, NGS-based approaches provide an unbiased perspective of complex biological systems without the requirement of any *a priori* information on the targets of interest. Unlike DNA microarrays, NGS-based approaches allow the assessment of both known and unknown molecules in a sample, thus facilitating the inference of novel biological insights [Lightbody *et al.*, 2019]. In the past few years, the avalanche of NGS-based approaches that have been developed to study different aspects of the biological systems, has substantially impacted practically every ‘omic’ field of study.

Nowadays, many commercial NGS technologies and platforms are available for studying different ‘omics’ disciplines. Most of their processing pipelines differ in terms of chemistry and certain details, but they tend to generally follow a similar paradigm [Reuter *et al.*, 2015], which involves sample collection, library preparation (includes fragmentation, addition of adaptor sequences, size selection, amplification), sequencing, base calling, and data analysis [Lightbody *et al.*, 2019]. Some technologies perform short-read sequencing (35-700 bp); whereas others focus on long-read sequencing, which is usually more expensive and has lower throughput [Goodwin *et al.*, 2016]. A technology or platform is usually chosen depending on the abilities, strengths and weaknesses, along with the consideration of sample type and the aim of the research. Currently, Illumina is the most widely-used sequencing technology [Reuter *et al.*, 2015], especially for short-read sequencing, in part due to its technological maturity as well as its broad range of platforms and high-level of compatibility across platforms [Goodwin *et al.*, 2016].

Illumina has produced a variety of sequencing platforms, where MiSeq and HiSeq series are the most established platforms that perform short-read sequencing, and each varies in terms of throughput and processing times. For instance, MiSeq platforms (used in Publication IV) are fast sequencers with low run times and are designed for sequencing small genomes as well as for targeted sequencing; whereas, HiSeq platforms (used in Publications I, III and IV) are geared towards high-throughput applications and have varying run times depending on model version [Reuter *et al.*, 2015]. Generally, Illumina short-read sequencers first generate clonal DNA template populations, where sample DNA is fragmented and common adaptors are ligated to either ends. Then, the DNA templates are amplified *in situ* on a solid support using bridge amplification, where each fragment creates thousands of copies in a cluster and ensures detection of the signal over background noise. Like this, millions of clusters are made, each with its own clonal DNA template population [Goodwin *et al.*, 2016]. Subsequently, the clonal DNA templates of all clusters are simultaneously subjected to sequencing by synthesis, which is one of the two broad approaches for sequencing short reads in a massively parallel fashion [Goodwin *et al.*, 2016; Mardis, 2017]. The sequencers can produce single-end or paired-end reads, whereby a DNA template is sequenced at only one end of the template or from both ends, respectively [Mardis, 2017].

3.2 Transcriptomics

Transcriptomics is the study of the transcriptome, which is the complete set of ribonucleic acid (RNA) transcripts in a single or a population of cells [Lowe *et al.*, 2017; Casamassimi *et al.*, 2017; Lightbody *et al.*, 2019]. RNA transcripts are transient molecules that are produced by transcribing (or copying) the information/instructions encoded in the stretches of DNA (i.e genes) of a cell. Notably, a large proportion (~75%) of the human genome can be transcribed [Djebali *et al.*, 2012] to produce several types of RNA transcripts that are generally classified into two groups: protein-coding and non-protein-coding RNA (in short, coding and non-coding RNA). Here, coding RNA, which generally refers to messenger RNA (mRNA), make up for a very small percentage (<5%) of the total RNA of a cell and serves as intermediary molecules that are translated to synthesise proteins. On the other hand, non-coding RNAs (ncRNAs), that include ribosomal RNA (rRNA), transfer RNA (tRNA), long non-coding RNA (lncRNA), microRNA (miRNA), and many other ncRNAs, make up for a majority of the total RNA and are directly involved in performing various structural as well as regulatory roles in the cell, such as gene regulation, protein translation, RNA splicing, among other cellular functions [Lowe *et al.*, 2017; Casamassimi *et al.*, 2017; Clancy, 2008].

Even though all cells of the human body contain the same genome, each cell expresses only a fraction of its genes depending on the cell's functions, type and developmental stage [Alberts, 2018]. Similar cells tend to express similar genes. Changes in gene expression of cells may be driven by varying pathological conditions of the body. By performing transcriptomics analyses and quantifying the abundances of the derived RNA transcripts, one can determine the gene expression patterns as well as gene expression levels of cell(s) under specific circumstances [Wang *et al.*, 2009]. This is called gene expression profiling and it can give a functional overview of the cell(s) of interest. In fact, one of the main purposes of many transcriptomics studies has been to quantify the (changes in) gene expression profiles of cells from different disease conditions, tissues, or time points, in order to identify differentially regulated genes, isoforms of genes, or enriched pathways that may contribute to the mechanisms underlying the development or progression of the disease [Casamassimi *et al.*, 2017; Lowe *et al.*, 2017].

Most studies perform transcriptomics analyses by enriching (or isolating) a specific type of RNA for targeted studies [Lowe *et al.*, 2017]. Messenger RNA has been the most frequently studied form of RNA [Lightbody *et al.*, 2019; Lowe *et al.*, 2017], but lately an increased interest has been seen in isolating specific ncRNAs for transcriptome level analyses as they have recently been implicated as important regulators in various disease processes [Hasin *et al.*, 2017], such as lncRNAs in cancer [Iyer *et al.*, 2015] and autoimmune diseases [Xu *et al.*, 2019; Wu *et al.*, 2015].

Both DNA microarrays and NGS, specifically RNA-sequencing (RNA-seq), technologies have been widely used for performing transcriptomic analyses, where the former has been largely supplanted by the latter in recent years due to its technological superiority and capability of analyzing all types of RNA [Lowe *et al.*, 2017]. In fact, recent technological advances in RNA-seq protocols have made it possible to perform analyses using much less starting material, which has enabled the transcriptomic profiling of single cells (i.e. single cell RNA-seq) in addition to its existing bulk RNA-seq capabilities [Blumenberg, 2019].

In this thesis, Publication I employs RNA-seq technologies (both bulk and single cell RNA-seq) for transcriptomic analyses, whereas Publication II studies DNA microarray datasets, specifically Affymetrix GeneChip microarrays.

3.2.1 Affymetrix GeneChip data normalization

For transcriptomics analyses using Affymetrix GeneChip, mRNA (or other RNA of interest) is first isolated from total RNA; processed using a suitable library kit to prepare fluorescent-labelled target libraries (outlined in Section 3.1.1); and hybridized to the probes on the microarray [Quackenbush, 2002]. Subsequently, the fluorescent intensities (i.e. hybridization data) that indicate the abundance of targets for each probe [Wu, 2009; Lowe *et al.*, 2017], are captured at each probe location using fluorescence scanners and presented in the form of an image.

Since the aim of transcriptomics analyses is to quantitate/quantify the gene (or transcript) expression levels in the sample of interest for further statistical analyses, the image data is usually converted to probe level intensity values using one of the many available image processing software [Quackenbush, 2002; Wu and Irizarry, 2004]. After image processing, these probe level intensity values usually contain noise from various sources, including non-specific hybridization events as well as the steps employed during sample preparation, hybridization, scanning, etc. that are of no biological interest and can produce misleading results [Wu, 2009; Quackenbush, 2002; Jiang *et al.*, 2008; Gregory Alvord *et al.*, 2007]. Therefore, these raw intensity values require appropriate normalization (i.e. preprocessing) in order to eliminate questionable measurements and remove variations from non-biological origins, which in turn would facilitate meaningful and relevant gene expression comparisons between different samples [Quackenbush, 2002; Wu and Irizarry, 2004].

Normalization of Affymetrix microarray data involves at least the following three steps [Jiang *et al.*, 2008; Gregory Alvord *et al.*, 2007]: (i) background correction, where the background noise arising primarily from non-specific hybridization events is adjusted from the observed intensities to estimate the accurate intensities at each probe, (ii) normalization, where systemic errors and biases are removed, and (iii) summarization, where intensities from multiple probes in a probe set are combined to yield a single intensity value depicting the expression level of the represented gene or transcript [Gregory Alvord *et al.*, 2007; Jiang *et al.*, 2008; Wu, 2009].

Several normalization procedures for Affymetrix data have been developed over the years that perform the above-mentioned steps using different methods and algorithms, such as robust multi-array average (RMA) [Irizarry *et al.*, 2003], MAS 5.0 [Affymetrix, 2002], GeneChip RMA (GCRMA) [Wu and Irizarry, 2004], and model based expression index (MBEI) [Li and Wong, 2001]. RMA normalization is one of the most popular approaches for normalizing Affymetrix data and it was used for normalizing the data in Publication II. It does not use MM probe intensities and assumes that the PM intensities are additive combinations of the true signal and background noise, which are modeled as exponentially and normally distributed, respectively, during background correction [Irizarry *et al.*, 2003]. Next, it performs quantile normalization and summarizes the (background corrected, normalized and log_2-transformed) probe level intensities by fitting a robust multi-array model using median polish algorithm [Irizarry *et al.*, 2003].

3.2.2 Bulk RNA-sequencing

RNA-sequencing is a revolutionary NGS technology that has taken transcriptomics studies to unprecedented heights [Wang *et al.*, 2009]. In the past decade, it has become an increasingly popular tool in transcriptomics studies as it enables proper and unbiased assessment of complex transcriptomes [Zhao *et al.*, 2016]. RNA-seq has various advantages over its predecessors, such as DNA microarrys, and circumvents some of their limitations. For instance, RNA-seq experiments are not limited to detecting only a known set of transcripts defined by the probes on the microarray and are in fact the first sequencing-based methods that have the potential of giving a comprehensive view of the entire transcriptome [Wang *et al.*, 2009; Zhao *et al.*, 2016; Garber *et al.*, 2011], including the lowly-expressed genes, small and long ncRNAs, as well as rare and novel transcripts [Wang *et al.*, 2009; Zhao *et al.*, 2016]. Moreover, RNA-seq has very low background noise and is extremely accurate in quantifying the expression levels of genes as well as their isoforms [Wang *et al.*, 2009]. However, the large amounts of data produced by RNA-seq, can be challenging to analyse and requires careful considerations [Garber *et al.*, 2011].

Data collection

In the bulk RNA-seq protocol (by Illumina), the RNA of interest (such as mRNA) is first isolated from 'bulk' samples containing thousands or millions of cells [Zhao *et al.*, 2016]. Subsequently, the captured RNA strands are fragmented, converted to complementary DNA (cDNA) fragments, size selected and subjected to the typical Illumina short-read DNA sequencing protocol as explained in Section 3.1.2 [Wang *et al.*, 2009; Zhao *et al.*, 2016]. In case of mRNA sequencing, 200-500 bp long fragments [Wang *et al.*, 2009] are usually selected for sequencing short reads (typically ~36-600 bases long [Garber *et al.*, 2011; Amarasinghe *et al.*, 2020]). Depending on the exact library preparations steps, mRNA libraries may also include certain long ncRNAs (>200 bp long).

Bulk mRNA RNA-seq data analysis

The large volumes of raw sequencing reads produced by Illumina sequencers for each sample are usually stored in FASTQ formatted files, which also contain the sequencing quality scores for each base [Ji and Sadreyev, 2018]. In case of paired-end sequencing, the two reads from either ends of each DNA template are saved in two separate FASTQ files. Before these raw reads can be analysed to draw any biological conclusions, they need to undergo a few crucial processing steps. Although there is no single optimal pipeline that can universally be applied to all RNA-seq datasets, there are, however, a few essential processing steps that are typically conducted in most (mRNA) RNA-seq analyses (as depicted in Figure 3.1) [Conesa *et al.*, 2016]. These steps include: (1) quality control, (2) alignment of reads to a reference transcriptome and/or genome (or *de novo* assembly of the transcripts), (3) quantification of gene (or transcript) expression levels, (4) normalization, and (5) differential expression analyses. Moreover, in order to understand the biological functions of the differentially expressed genes (DEGs), some pipelines also perform further downstream analyses that identify the pathways or gene sets the DEGs may be implicated in.

First, the raw reads in each FASTQ file are subjected to quality control

(QC) analysis in order to detect poor-quality reads, adaptor contamination, and biases in the data that may arise due to problems in library preparation or sequencing [Conesa *et al.*, 2016; Zhao *et al.*, 2016; Lowe *et al.*, 2017]. Minimizing such defects in the data is vital for downstream analyses as they may lead to inaccurate results [Zhao *et al.*, 2016]. FastQC [Andrews, 2010] is one of the most popular software for performing QC analyses on Illumina reads in bulk RNA-seq studies and was used in Publication I. It performs a series of analyses for evaluating the qualities of reads at each base position; duplication levels of reads, which may indicate PCR amplification bias; guanine-cytosine (GC) content; overrepresented sequences, which may indicate presence of adaptors or contaminants; and distribution of read lengths; to name a few [Andrews, 2010]. Generally, the quality scores of bases decrease towards the ends of long reads [Conesa *et al.*, 2016; Fonseca *et al.*, 2012]. If the QC analysis identifies any issues in the reads of a FASTQ file, tools such as Cutadapt [Martin, 2011] and Trimmomatic [Bolger *et al.*, 2014], can be used to remove poor-quality reads, trim low-quality bases, or trim adaptor sequences. Wrapper tools that integrate some of the above-mentioned tools to perform QC on sequencing data have also been developed, such as Trim Galore! [Krueger, 2012] (a wrapper tool around FastQC and Cutadapt).

Subsequently, the high-quality mRNA reads are computationally mapped (i.e. aligned) to a reference genome or transcriptome in order to identify the genes or genomic locations from which they originate [Zhao *et al.*, 2016; Conesa *et al.*, 2016]. If reads are aligned to a reference transcriptome, the analysis is easier and less computationally expensive [Fonseca *et al.*, 2012, 2014], but is limited to the identification of annotated (i.e. known) transcripts. Whereas, alignment to a reference genome allows the discovery of unannotated and novel transcripts as well [Conesa *et al.*, 2016; Yang and Kim, 2015]. In case a reference genome (and transcriptome) for the organism of study is unavailable, *de novo* assembly can be performed on the reads [Conesa *et al.*, 2016; Lowe *et al.*, 2017], where the reads are first assembled into longer contigs, creating an expressed transcriptome to which the reads can be mapped back for quantification [Conesa *et al.*, 2016]. The scope of this section, however, is limited to discussing only the alignment of reads to a reference.

The accurate alignment of mRNA reads to mammalian, such as human, reference genomes or transcriptomes is often complicated by specific challenges. Only a subset of the challenges are discussed here. First, since eukaryotic genes are generally composed of relatively multiple short coding regions (average length 235 bp [Kim *et al.*, 2013]), called exons, that are separated or intervened by non-coding regions (vary from 50 bp to >1 Mbp in length [Kim *et al.*, 2013, 2015]), called introns; and because these introns are spliced out from the mature RNA transcripts that are used to construct the short mRNA reads in RNA-seq data; many mRNA reads may span two exons (called exon-exon spanning reads or junction reads) and need to be appropriately split across potentially long stretches of intronic regions for accurate alignment [Kim *et al.*, 2013; Zhao *et al.*, 2016; Garber *et al.*, 2011; Lowe *et al.*, 2017]. Second, due to the short lengths of the reads (50-100 bp), a read may align uniquely to one location on the genome or align to multiple locations (i.e. multi-map reads or multi-reads) due to repetitive regions in the genome [Conesa *et al.*, 2016; Fonseca *et al.*, 2012], such as paralogous genes [Conesa *et al.*, 2016] and pseudogenes [Kim *et al.*, 2013]. Third, the reads may contain mismatches, insertions and deletions usually caused either due to genetic variation or sequencing errors [Zhao, 2014]. Also, with the increasing throughput of sequencing technologies and increasing read lengths, conducting alignment of all reads at a reasonable speed using limited computational resources is also a major concern [Lowe *et al.*, 2017; Kim *et al.*, 2015]. However, the increase in read length along with paired-end sequencing can reduce multi-mapping of reads [Wang *et al.*, 2009]. Moreover, since the size of a transcriptome is far smaller than that of a genome, by aligning the reads to a reference transcriptome the alignment speed can be substantially increased [Kim *et al.*, 2013]. The mapping accuracy of junction reads may also increase in this case [Zhao *et al.*, 2016].

Numerous read aligning algorithms have been proposed over the past decade that deal with the aforementioned challenges in different ways. The aligning approaches can largely be divided into two categories: 'unspliced read aligners' and 'spliced read aligners' [Garber *et al.*, 2011]. One of the major differences between the two is that the former does not allow any large gaps during alignment, whereas the latter allows large gaps for properly mapping the junction reads [Garber *et al.*, 2011]. Most aligners accommodate for short gaps and a few mismatches [Zhao *et al.*, 2016]. Unspliced read aligners, such as Bowtie [Langmead *et al.*, 2009], Bowtie2 [Langmead and Salzberg, 2012], and BWA [Li and Durbin, 2009], can be used for aligning to the reference transcriptome, whereas spliced read aligners, such as Tophat [Trapnell *et al.*, 2009], Tophat2 [Kim *et al.*, 2013] (used to align bulk RNA-seq data in Publication I), STAR [Dobin *et al.*, 2013] (used to align scRNA-seq data in Pulication I), and HISAT [Kim *et al.*, 2015], can be used to align to the reference genome [Conesa *et al.*, 2016; Garber *et al.*, 2011]. Some popular spliced read aligners, such as Tophat2 and STAR, make use of the annotations in the reference transcriptome as well in order to increase their alignment accuracy and speed [Kim *et al.*, 2013]. Engström *et al.* [2013] discovered that there are major performance differences between aligners on various benchmarks and each alignment tool usually exhibits distinct strengths and weaknesses.

After alignment, a common aim of most RNA-seq data analyses is to quantify the number of reads that align to a particular gene or transcript in order to estimate their expression levels [Fonseca *et al.*, 2014; Conesa *et al.*, 2016; Lowe *et al.*, 2017]. There are broadly two approaches to gene quantification: (1) union-exon-based approaches, where a read is counted towards the expression level of a gene if it aligns to any one of its exons, and (2) transcript-based approaches, where transcript-level (i.e. isoform-level) quantification is performed and used for gene-level quantification by summing over the expression levels all of the gene's isoforms [Zhao *et al.*, 2016, 2015]. The simplest and the most commonly used approach for quantification is the union-exon-based approach [Conesa *et al.*, 2016; Zhao *et al.*, 2015], which is implemented in tools such as HTSeq-count [Anders *et al.*, 2015] and featureCounts [Liao *et al.*, 2014]. HTSeq-count is one of the top ranking gene-level quantification tools [Fonseca *et al.*, 2014] and was used for quantifying genes in the bulk RNA-seq data from Publication I. It considers a gene as the union of its exons and assesses the read count of each gene to be equivalent to the number of reads aligning unambiguously to its exons [Anders *et al.*, 2015]. In contrast, transcript-level quantification, which is performed by tools such as Cufflinks [Trapnell *et al.*, 2010], RSEM [Li and Dewey, 2011], Sailfish [Patro *et al.*, 2014], Kallisto [Bray *et al.*, 2016], and Salmon [Patro *et al.*, 2017], is more complicated as related transcripts often share their reads due to common exons [Conesa *et al.*, 2016; Lowe *et al.*, 2017; Garber *et al.*, 2011; Zhao *et al.*, 2015]. Therefore, for gene-level expression analyses, a union-exon-based approach is more feasible, whereas for a transcript-level expression analysis, a transcript-based approach can be adopted.

As the raw read counts quantified for each gene (or transcript) are often riddled with technical and biological biases, such as those caused by varying sequencing depth of samples, lengths of genes, and composition of RNA population in each sample, they are often not directly comparable between or within samples [Robinson and Oshlack, 2010; Zhao *et al.*, 2016; Garber *et al.*, 2011]. Therefore, it is crucial to normalize these raw counts to ensure reliable estimation of gene expression in each sample and to infer accurate results from any subsequent analyses, such as differential expression analyses (DEA) [Robinson and Oshlack, 2010; Zhao *et al.*, 2016]. Several different normalization methods have been proposed over the years. Some of these methods involve re-scaling raw gene counts by calculating values such as counts per million mapped reads (CPM), which corrects only for the varying sequencing depths of samples; and reads per kilobase per million mapped reads (RPKM) [Mortazavi *et al.*, 2008] (used in Publication I), which combines between- and within-sample normalization by correcting for both the sequencing depths of samples as well as gene lengths. Certain other methods, such as trimmed mean of M-values (TMM) normalization [Robinson and Oshlack, 2010] (used in Publication I) and DESeq [Anders and Huber, 2010], estimate scaling factors that do not directly adjust the raw counts, but are built into the statistical model (as model offsets, Section 4.4.3) that are used for DEA. TMM and DESeq have been demonstrated to outperform other normalization methods [Dillies *et al.*, 2013]. Both of these methods normalize for the RNA composition biases, and estimate scaling factors on the assumption that most genes are not differentially expressed (DE) and the log-fold-changes of the non-DE genes should be close to 1 [Robinson and Oshlack, 2010; Anders and Huber, 2010; Dillies *et al.*, 2013]. The TMM approach estimates scaling factors by first removing (i.e. trimming) the most highly expressed (or repressed) genes and the genes with extreme log-fold changes, and then calculating the weighted mean of gene-wise log expression ratios (i.e. M values) on the remainder of the genes using one of the samples as a reference [Robinson and Oshlack, 2010]. DESeq, on the other hand, defines a pseudoreference sample, which is built with the geometric mean of gene counts across all samples, to estimate the scaling factor as the median of the ratios of observed counts over all genes [Anders and Huber, 2010].

Having performed all the aforementioned pre-processing steps, one of the fundamental steps in most RNA-seq studies is to understand how gene expression levels differ across distinct sample groups and to identify differentially expressed genes (DEGs). Statistical models, such as generalized linear models that approximate the count data to be distributed according to negative binomial distribution, are often used to reliably identify these DEGs in RNA-seq studies. Detailed discussion on this can be found in Section 4.4.3.

As differential expression analyses are usually performed on tens of thousands of genes at a time, they often result in long lists of differentially expressed genes (DEGs) that are challenging to interpret as such [Reimand *et al.*, 2019; Khatri *et al.*, 2012]. However, these lists of DEGs are often composed of genes that belong to the same pathways or gene sets [Zhao *et al.*, 2016; Yu *et al.*, 2017]. Here, a pathway refers to a group of genes that work cooperatively to carry out a biological process or function; and a gene set refers to groups of genes that have common biological function, regulation or chromosomal location [Reimand *et al.*, 2019; Subramanian *et al.*, 2005]. Details on the genes involved in each pathway or gene set are generally stored in various databases [Khatri *et al.*, 2012], such as Gene Ontology (GO) [Ashburner *et al.*, 2000], Kyoto Encyclopedia of Genes and Genomes (KEGG) [Kanehisa *et al.*, 2017] and Molecular Signatures Database (MSigDB) [Liberzon *et al.*, 2011, 2015]. For ease, the term 'pathway' will be used here as an overarching term to indicate both pathways and gene sets. Therefore, to make the list of DEGs more interpretable and to gain insights into the biological mechanisms that they are involved in (i.e. functionally annotate the DEGs), pathway enrichment analyses are commonly performed on the lists of DEGs [Reimand *et al.*, 2019; Zhao *et al.*, 2016; Khatri *et al.*, 2012; Yu *et al.*, 2017]. Essentially, these analyses identify the biological pathways that are enriched in a list of genes by performing statistical tests that check if the genes of a pathway are over-represented in the list of DEGs [Reimand *et al.*, 2019]. Various methods have been proposed for performing pathway enrichment analyses, such as Fisher’s exact test (as implemented in online tool, DAVID [Dennis *et al.*, 2003; Hosack *et al.*, 2003]) and GSEA [Subramanian *et al.*, 2005]. Gene regulatory analyses, such as transcription factor (TF) analyses using TRANSFAC [Matys *et al.*, 2006] database, can also be performed to identify the factors involved in regulation of the DEGs at hand.

3.2.3 Single Cell RNA-sequencing

While bulk RNA-sequencing (and even DNA microarray) techniques have enabled numerous valuable insights into the complex transcriptional profiles of different cell types, they are limited to providing gene expression measurements that are averaged across thousands of cells [Stegle *et al.*, 2015; Bacher and Kendziorski, 2016; Chen *et al.*, 2019; Poirion *et al.*, 2016]. Such global view of average gene expression profiles may be sufficient in many contexts, such as comparative transcriptomics, and have led to various biomarker discoveries in the field of biomedicine. However, they are insufficient in certain scenarios; for instance, when analyzing a small number of functionally distinct cells, such as embryonic cells; or complex tissues that are often composed of different cell types [Stegle *et al.*, 2015]. Moreover, populations of cells, even those from the same cell type, can exhibit substantial heterogeneity [Raj and Van Oudenaarden, 2008; Altschuler and Wu, 2010; Wagner *et al.*, 2016], owing to the presence of rare or novel subpopulations of cells [Kolodziejczyk *et al.*, 2015; Wagner *et al.*, 2016] as well as those cells transitioning between different states of biological processes, such as differentiation [Stegle *et al.*, 2015; Altschuler and Wu, 2010]. Major cell-to-cell heterogeneity stems from the inherent stochastic nature of cellular gene expression wherein cells, even from genetically homogeneous populations, undergo the phenomenon of ‘transcriptional bursts’ [Stegle *et al.*, 2015; Liu and Trapnell, 2016; Wagner *et al.*, 2016; Raj and Van Oudenaarden, 2008]. In other words, the genes of a cell are not transcribed continuously. Instead, they experience short bursts of transcription followed by silent intervals, which are regulated by nonsynchronous cellular processes [Liu and Trapnell, 2016; Kolodziejczyk *et al.*, 2015; Haque *et al.*, 2017]. Unfortunately, bulk RNA-seq experiments mask important cellular-level heterogeneity [Poirion *et al.*, 2016; Stegle *et al.*, 2015; Altschuler and Wu, 2010].

These limitations of bulk RNA-seq are largely overcome by single cell RNA-sequencing (scRNA-seq) technologies that investigate gene expression profiles of individuals cells in an unbiased and high-throughput manner [Poirion *et al.*, 2016; Kolodziejczyk *et al.*, 2015]. Ever since the generation of the first scRNA-seq dataset in 2009 [Kolodziejczyk *et al.*, 2015; Zhang *et al.*, 2019], single cell technologies have made tremendous experi-

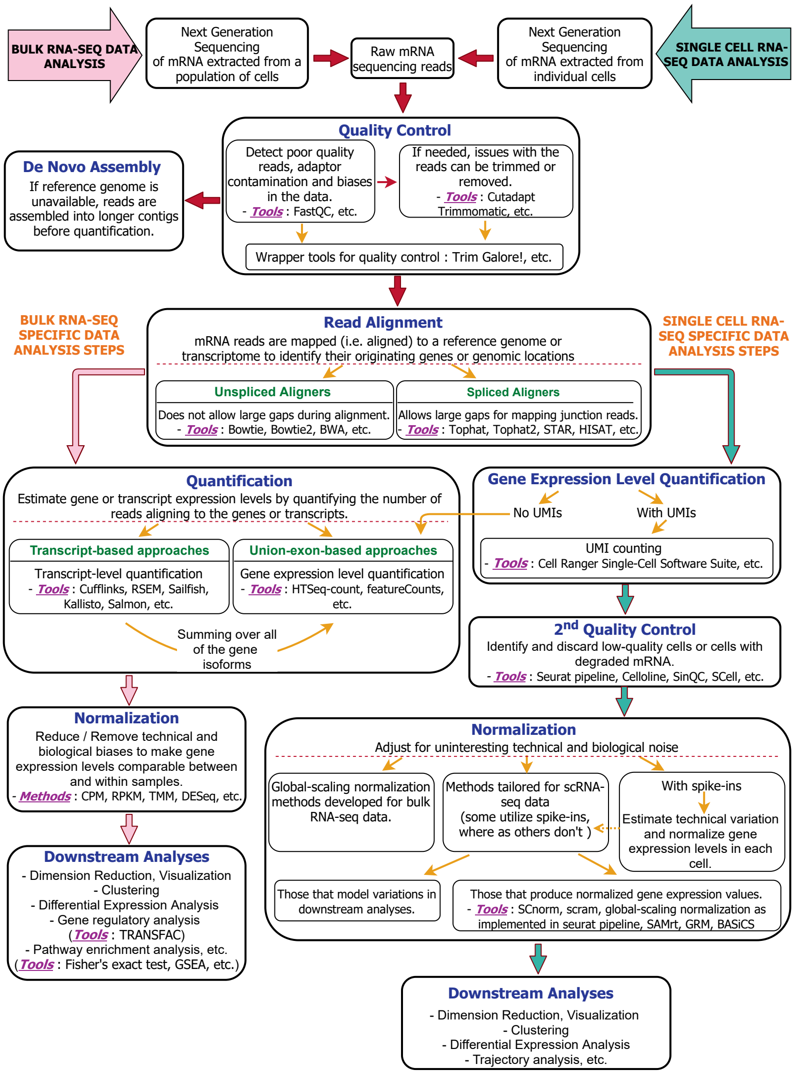

BULK RNA-SEQ DATA ANALYSIS
Next Generation Sequencing of mRNA extracted from a population of cells
Raw mRNA sequencing reads
Next Generation Sequencing of mRNA extracted from individual cells
SINGLE CELL RNA-SEQ DATA ANALYSIS
Quality Control
Detect poor quality reads, adaptor contamination and biases in the data.
- Tools : FastQC, etc.
If needed, issues with the reads can be trimmed or removed.
- Tools : Cutadapt Trimmomatic, etc.
Wrapper tools for quality control : Trim Galore!, etc.
De Novo Assembly
If reference genome is unavailable, reads are assembled into longer contigs before quantification.
BULK RNA-SEQ SPECIFIC DATA ANALYSIS STEPS
Read Alignment
mRNA reads are mapped (i.e. aligned) to a reference genome or transcriptome to identify their originating genes or genomic locations
Unspliced Aligners
Does not allow large gaps during alignment.
- Tools : Bowtie, Bowtie2, BWA, etc.
Spliced Aligners
Allows large gaps for mapping junction reads.
- Tools : Tophat, Tophat2, STAR, HISAT, etc.
SINGLE CELL RNA-SEQ SPECIFIC DATA ANALYSIS STEPS
Quantification
Estimate gene or transcript expression levels by quantifying the number of reads aligning to the genes or transcripts.
Transcript-based approaches
Transcript-level quantification
- Tools : Cufflinks, RSEM, Sailfish, Kallisto, Salmon, etc.
Union-exon-based approaches
Gene expression level quantification
- Tools : HTSeq-count, featureCounts, etc.
Summing over all of the gene isoforms
Gene Expression Level Quantification
No UMIs
With UMIs
UMI counting
- Tools : Cell Ranger Single-Cell Software Suite, etc.
2nd Quality Control
Identify and discard low-quality cells or cells with degraded mRNA.
- Tools : Seurat pipeline, Celloline, SinQC, SCell, etc.
Normalization
Reduce / Remove technical and biological biases to make gene expression levels comparable between and within samples.
- Methods : CPM, RPKM, TMM, DESeq, etc.
Downstream Analyses
- Dimension Reduction, Visualization
- Clustering
- Differential Expression Analysis
- Gene regulatory analysis
(Tools : TRANSFAC)
- Pathway enrichment analysis, etc.
(Tools : Fisher's exact test, GSEA, etc.)
Normalization
Adjust for uninteresting technical and biological noise
Global-scaling normalization methods developed for bulk RNA-seq data.
Methods tailored for scRNA-seq data
(some utilize spike-ins, where as others don't)
With spike-ins
Estimate technical variation and normalize gene expression levels in each cell.
Those that model variations in downstream analyses.
Those that produce normalized gene expression values.
- Tools : SCnorm, scram, global-scaling normalization as implemented in seurat pipeline, SAMrt, GRM, BASiCS
Downstream Analyses
- Dimension Reduction, Visualization
- Clustering
- Differential Expression Analysis
- Trajectory analysis, etc.

Figure 3.1. Depiction of the main steps in bulk and single cell RNA-sequencing (RNA-seq) data analysis pipelines. The red arrows indicate steps and tools that are common between the two RNA-seq technologies. The pink arrows indicate the steps specific to bulk RNA-seq data analysis. The green arrows indicate the steps specific to single cell RNA-seq data analysis. The red, pink and green arrows indicate transition from one step in the data analysis pipeline to another (multiple arrows from a box indicate multiple directions in which the analysis can transition). Yellow arrows are used to indicate the ways in which a step can be carried out.

mental and computational developments that have since led to profound new discoveries in biology [Stegle *et al.*, 2015]. In scRNA-seq data, each cell provides a snapshot of its transcriptional activity at a particular point in time, revealing the cell-to-cell heterogeneity that exists within a population of cells. This unprecedented level of information on individual cells provides us with the opportunity to capitalize upon several biological insights that were previously intangible. For instance, scRNA-seq data can be used to identify the cell types present in complex and heterogeneous cell populations, including rare and novel cell subpopulations; study the cellular transitions between different states of biological processes (i.e. temporal trajectories); determine the spatial organization of the cells or cell types; infer gene regulatory networks; and investigate the stochastic components of transcription; to name a few [Stegle *et al.*, 2015; Wagner *et al.*, 2016; Kolodziejczyk *et al.*, 2015; Poirion *et al.*, 2016; Liu and Trapnell, 2016].

Data collection

Most scRNA-seq library preparation protocols follow a similar basic strategy, which begins by capturing individual cells, lysing the cells and isolating the mRNA molecules (most scRNA-seq protocols thus far focus on isolating (poly-A tailed) mRNA [Liu and Trapnell, 2016]). The subsequent steps are similar to that of bulk RNA-seq wherein the mRNA molecules are reverse transcribed to obtain cDNA fragments, the cDNA is amplified and then sequenced using the NGS technology of choice [Kolodziejczyk *et al.*, 2015; Stegle *et al.*, 2015; Liu and Trapnell, 2016]. During reverse transcription, some protocols tag each cell’s transcriptome with unique oligonucleotide sequences, called barcodes, to preserve the information on each transcript’s cellular origin [Haque *et al.*, 2017]. Moreover, most protocols at the moment use Illumina platforms for sequencing [Kolodziejczyk *et al.*, 2015].

One of the most challenging tasks in sequencing mRNA from single cells is in capturing individual cells with high efficiency [Kolodziejczyk *et al.*, 2015]. Several approaches have been introduced over the years for capturing single cells. Most of the earlier methods that include micromanipulation, flow-activated cell sorting (FACS), and laser capture microdissection (LCM), have been limited to capturing hundreds or thousands of cells in a single experiment; whereas, recent methods based on microwell plate-based and droplet-based microfluidics strategies have enabled the capturing of tens of thousands of cells at a time in emulsion oil droplets [Kolodziejczyk *et al.*, 2015; Liu and Trapnell, 2016]. Of the three most widely used droplet-based methods, namely inDrop [Klein *et al.*, 2015], Drop-seq [Macosko *et al.*, 2015] and 10X Genomics Chromium [Zheng *et al.*, 2017], the 10X method (used in Publication I) was demonstrated by Zhang *et al.* [2019] to have higher sensitivity than the other two and is suitable for most applications.

Analytical challenges in scRNA-seq data

Despite the considerable progress that has been made by scRNA-seq technologies, scRNA-seq data contain very high levels of noise from both technical and biological sources that pose many analytical and computational challenges [Poirion *et al.*, 2016; Kolodziejczyk *et al.*, 2015; Bacher and Kendziorski, 2016; Stegle *et al.*, 2015]. Generally, the variations (or noise) present in scRNA-seq data are much higher than that of in bulk RNA-seq data [Bacher and Kendziorski, 2016; Haque *et al.*, 2017; Wagner *et al.*, 2016]. Therefore, while some bulk RNA-seq analysis tools are directly applicable in scRNA-seq data analysis, many new computational methods need to be adopted to accurately characterize the biological insights presented in this type of data [Stegle *et al.*, 2015].

One of the most prominent sources of biological variation emanates from the existence of intrinsic cell-to-cell heterogeneity within populations of cells (as detailed above) [Kolodziejczyk *et al.*, 2015; Bacher and Kendziorski, 2016; Wagner *et al.*, 2016]. Additionally, other biological factors that may add to the unwanted biological variations in the data include: cell sizes, where small cells typically contain less RNA and thus appear to be of inferior quality [Bacher and Kendziorski, 2016; Haque *et al.*, 2017; Wagner *et al.*, 2016]; varying RNA compositions of individual cells; varying rates of mRNA degradation; and difficulties in capturing or lysing certain cells [Wagner *et al.*, 2016]. Biological variations may even influence the extent of technical variation [Wagner *et al.*, 2016]; a few examples are highlighted below.

Most of the technical variation stems from the nature of single cell technologies. Due to the minute amounts of starting biological material (picograms of mRNA [Zhang *et al.*, 2019]) that are usually available in a single cell, substantial amount of amplification is needed, which can result in amplification bias (i.e. false positive detection) [Poirion *et al.*, 2016; Liu and Trapnell, 2016; Bacher and Kendziorski, 2016]. These starting quantities of RNA can further vary depending on biological factors, such as cell size and cell type [Wagner *et al.*, 2016]. Additionally, the capture efficiency of current scRNA-seq protocols is quite low [Liu and Trapnell, 2016; Bacher and Kendziorski, 2016; Stegle *et al.*, 2015; Chen *et al.*, 2019]; only about 10-20% of the transcripts of a cell are present in the final sequencing libraries [Kolodziejczyk *et al.*, 2015] and even moderately expressed genes are frequently undetected [Haque *et al.*, 2017; Stegle *et al.*, 2015]. This along with biological factors, such as the subpopulations of cells or transient states, when certain genes are not expressed, lead to high frequency of dropout events (i.e. false negatives; expressed but undetected transcripts), which in turn results in relatively sparse data [Wagner *et al.*, 2016; Haque *et al.*, 2017; Vallejos *et al.*, 2017]. Moreover, scRNA-seq protocols sometimes fail to dissociate cells and analyze two or more cells (often referred to as doublets) together. These doublets then manifest in the data as high-quality cells with relatively more complex libraries and more transcripts than other cells [Wagner *et al.*, 2016]. Finally, the technical sources of variations that are common to other NGS data are also prevalent in scRNA-seq data, such as batch effects and library preparation protocols.

UMIs and spike-ins

Most studies aim to account for the technical and biological variations during different steps of the data analysis pipeline, such as quality control, quantification, normalization and/or data modelling. For instance, to alleviate amplification bias and to improve quantification of scRNA-seq reads, some scRNA-seq protocols (such as 10X Genomics Chromium, inDrop and Drop-seq [Zhang *et al.*, 2019]) tag each individual mRNA molecule within a cell with short random sequences (4-20 bp), called unique molecular identifiers (UMIs), during the reverse transcription step of library preparation [Kolodziejczyk *et al.*, 2015; Wagner *et al.*, 2016; Stegle *et al.*, 2015; Poirion *et al.*, 2016]. After amplification, by counting the unique number of UMIs mapping to each gene, instead of the total number of mapped reads, the real gene expression levels that are reflected in the cell library can be estimated [Wagner *et al.*, 2016]. Another approach of accounting for some, but not all, technical variation in the data is to incorporate artificial spike-in molecules in each cell lysate [Kolodziejczyk *et al.*, 2015; Stegle *et al.*, 2015; Wagner *et al.*, 2016]. Spike-ins are exogeneous RNA sequences, such as those from External RNA Control Consortium (ERCC), that are added to the mRNA content of each cell in known and constant quantities, and are assumed to be unaffected by the biological covariates [Wagner *et al.*, 2016; Stegle *et al.*, 2015]. Therefore, they are considered to be good negative controls for normalizing the gene expression measurements of each cell and also for evaluating the library quality [Wagner *et al.*, 2016; Kolodziejczyk *et al.*, 2015]. Most droplet-based technologies, however, have been unable to accommodate spike-ins in their protocols [Bacher and Kendziorski, 2016; Lun *et al.*, 2016]. Therefore, as 10X Genomics Chromium technology was used to generate the scRNA-seq data in Publication I, the protocol and data included UMIs, but not spike-ins. Nevertheless, while there are a lot of benefits of using UMIs and spike-ins, they have their own challenges and limitations [Bacher and Kendziorski, 2016; Wagner *et al.*, 2016].

Single cell RNA-seq data analysis

Similar to bulk RNA-seq data analysis, scRNA-seq data analysis pipeline (as depicted in Figure 3.1) begins with performing quality control (QC) on the raw sequencing reads and subsequently aligning the good-quality reads from each cell to a reference genome or transcriptome. For carrying out both steps, the methods that are developed for bulk RNA-seq data analysis (as outlined in Section 3.2.2) can be applied to scRNA-seq data as well [Stegle *et al.*, 2015; Poirion *et al.*, 2016; Kolodziejczyk *et al.*, 2015]. After alignment, gene expression level quantification of the mapped reads is performed. In case the data is obtained from scRNA-seq protocols that do not incorporate UMIs, quantification can be performed using the same methods that are applied in bulk RNA-seq data analysis (see Section 3.2.2). However, in case UMIs are incorporated in the protocol, gene expression level quantification is done by counting the total number of unique UMIs associated with the reads mapping to a gene, called UMI counting (as done in Publication I) [Stegle *et al.*, 2015; Bacher and Kendziorski, 2016]. Sequencing errors can occur in UMIs that might result in the appearance of spurious UMIs with low copy numbers [Wagner *et al.*, 2016; Stegle *et al.*, 2015]. Therefore, in order to avoid over-counting, UMIs with low quality should be filtered out and those with low copy numbers should be removed or collapsed with other UMIs to which they have short edit distances using statistical models [Wagner *et al.*, 2016; Stegle *et al.*, 2015; Zheng *et al.*, 2017]. For instance, the Cell Ranger Single-Cell Software Suite [Zheng *et al.*, 2017], which was availed in Publication I for performing QC analyses and alignment of raw reads, filters out UMIs with <90% base call accuracy and corrects UMIs that are 1-Hamming-distance away from another UMI with more reads.

After quantifying the gene expression levels in each cell, it is extremely important to perform another set of QC analysis in an effort to identify and discard low quality cells or cells that may contain degraded mRNA as they can lead to misinterpretation of the data. Here, low quality cells refer to those cells that are broken or killed in the capturing process, or those capture sites that are empty or contained multiple cells [Ilicic *et al.*, 2016]. There are several metrics that can help identify such cells. These metrics include examining: the number of reads sequenced (i.e. sequencing depth) [Bacher and Kendziorski, 2016]; the proportion of uniquely mapping reads, where low mapping rates may indicate degraded mRNA, contamination or improper lysing [Stegle *et al.*, 2015; Bacher and Kendziorski, 2016; Kolodziejczyk *et al.*, 2015]; the number of genes expressed [Kolodziejczyk *et al.*, 2015]; the fraction of reads mapping to endogeneous genes [Kolodziejczyk *et al.*, 2015; Bacher and Kendziorski, 2016]; the proportion of reads mapping to mitochondrial genome, where high levels of mitochondrial RNA (mtRNA) may arise from broken cells that lose cytoplasmic RNA but retain mtRNA that is enclosed in the mitochondria [Ilicic *et al.*, 2016]; and the ratio of the number of reads mapped to the endogeneous RNA versus those that mapped to the extrinsic spike-ins (if spike-ins are incorporated), where high mapping to spike-ins would indicate low amount of RNA in the cell or broken cell [Stegle *et al.*, 2015; Kolodziejczyk *et al.*, 2015; Bacher and Kendziorski, 2016]. Additionally, dimension reduction and ordination of the data may help detect outlier cells [Stegle *et al.*, 2015]. However, identifying low quality cells requires setting arbitrary thresholds that differ according to the datasets [Ilicic *et al.*, 2016; Poirion *et al.*, 2016]. There are a variety of tools for performing these or a subset of these metrics on scRNA-seq data, such as Seurat [Satija *et al.*, 2015; Macosko *et al.*, 2015] pipeline, Celloline [Ilicic *et al.*, 2016], SinQC [Jiang *et al.*, 2016] and SCell [Diaz *et al.*, 2016].

One of the most critical and challenging steps in processing scRNA-seq data is normalization of the data to adjust for uninteresting technical and biological noise that may be masking the underlying signal of interest [Stegle *et al.*, 2015; Vallejos *et al.*, 2017; Vieth *et al.*, 2019]. The choice of normalization method can substantially impact the results of scRNA-seq data analyses, such as differential expression analysis [Vieth *et al.*, 2019; Bacher and Kendziorski, 2016]. Ideally, the data normalization strategies should capture the biases and variations specific to the type of data at hand [Vallejos *et al.*, 2017]. The most widely used normalization techniques for scRNA-seq data have been the global-scaling normalization methods developed for bulk RNA-seq data, such as library-based normalization, TMM, and DESeq [Vallejos *et al.*, 2017; Bacher and Kendziorski, 2016]. These methods attempt to remove cell-specific biases by calculating one scaling factor for all genes in each cell [Vallejos *et al.*, 2017]. However, normalization methods for bulk RNA-seq data make assumptions that do not always apply in the context of scRNA-seq data. For instance, they usually assume the total amount of RNA processed per sample (or cell) to be approximately same or vary only due to technical noise; and as explained in Section 3.2.2, methods like TMM and DESeq assume only a small fraction of genes to be differentially expressed between samples (or cells); which are not true when diverse cell sizes and types are considered [Stegle *et al.*, 2015; Vieth *et al.*, 2019]. Also, some bulk RNA-seq normalization methods, such as DESeq, perform poorly on zero-inflated data such as those generated by scRNA-seq [Vallejos *et al.*, 2017]. Even though a direct application of bulk RNA-seq normalization techniques has been found through various studies to be inappropriate and misinforming in the context of scRNA-seq data, they have been extensively used in scRNA-seq data analysis [Vallejos *et al.*, 2017; Lun *et al.*, 2016].

Some scRNA-seq protocols incorporate spike-ins to estimate technical variation and normalize gene expression levels in each cell (explained earlier). Essentially, spike-ins can be used to estimate the endogeneous mRNA content (i.e. total number of mRNA molecules) per cell by calculating a cell-specific scaling factor based on the differences between the observed and expected expressions of the spike-ins and using that scaling factor for normalizing the expression values of endogeneous mRNA [Stegle *et al.*, 2015; Vallejos *et al.*, 2017; Bacher and Kendziorski, 2016]. They can also be used to improve the estimation of global-scaling factors [Bacher and

Kendziorski, 2016]. However, spike-ins do not normalize for all sources of variation, such as the differences in mRNA content between cells [Vallejos *et al.*, 2017] and stochastic dropout of RNA molecules [Kolodziejczyk *et al.*, 2015]. Moreover, calibrating the amount of spike-ins added to each cell according to the endogeneous mRNA content of each cell is a crucial but non-trivial task, which, if done poorly, can invalidate the use of spike-ins as control sequences [Kolodziejczyk *et al.*, 2015; Vallejos *et al.*, 2017; Bacher and Kendziorski, 2016]. Also, as ERCC spike-in sequences are different from endogeneous mRNA molecules, their counts may be affected by the technical and biological factors differently [Vallejos *et al.*, 2017; Wagner *et al.*, 2016].

Several state-of-the-art normalization methods specifically tailored for scRNA-seq data have been introduced recently; some of which utilize the spike-in information, whereas others do not. These methods can be largely divided into two approaches: those that model the variations in downstream analyses and those that produce normalized gene expression values that can be used in subsequent analyses [Vallejos *et al.*, 2017]. Methods belonging to the latter category will be discussed here. One of the most recent methods is SCnorm [Bacher *et al.*, 2017], which addresses the issues that arise from estimating global scaling factors that assume the count-depth relationship to be same across all genes of the cell when that is not the case. They use one quantile regression to group genes according to the dependence of their expression on sequencing depth and a second quantile regression to estimate group-specific scaling factors. SCnorm does not require spike-in information, but it can be used to improve its performance [Bacher *et al.*, 2017]. The deconvolution approach used in scran [Lun *et al.*, 2016] does not employ spike-in information and performs robustly in zero-inflated data. This approach partitions cells into pools of comparable library sizes, normalizes across cells in each pool, and then uses the resulting system of linear equations to estimate cell-specific scaling factors [Lun *et al.*, 2016]. On the other hand, Seurat pipeline [Satija *et al.*, 2015; Macosko *et al.*, 2015] implements a simple global-scaling normalization approach where the unique UMI count per gene (i.e. gene count) is divided by the total number of unique UMIs in the cell (i.e. library size), and then multiplied by a scaling factor (10,000 by default). The result, x, is finally log-transformed by $ln(x+1)$ to account for zero counts [Macosko *et al.*, 2015]. Other normalization methods designed specifically for scRNA-seq data include Single-Cell Tagged Reverse Transcription (SAMstrt) [Katayama *et al.*, 2013], Gamma Regression Model (GRM) [Ding *et al.*, 2015] and Bayesian Analysis of Single-Cell sequencing Data (BASiCS) [Vallejos *et al.*, 2016], which utilize spike-in information.

A recent comparative analysis of seven scRNA-seq data normalization methods (that included the ones mentioned here) could not identify any one method that outperformed all others in every aspect and all datasets [Lytal *et al.*, 2020]. They also concluded that a simple normalization approach, such as the one implemented in Seurat, does not significantly differ in terms of performance as compared to the other more intensive methods [Lytal *et al.*, 2020]. In fact, in Publication I, the Seurat pipeline was used to perform the second set of QC analyses and normalization.

Finally, the normalized gene expression values of each cell can be used to perform downstream analyses, such as dimension reduction, clustering, differential expression analysis, and trajectory analysis, which are covered in Section 4.6.

3.2.4 Brief comparison between the transcriptomics technologies

As evident from Sections 3.2.1-3.2.3, DNA microarrays, bulk RNA-seq and scRNA-seq technologies differ from each other in several aspects and have their own advantages as well as disadvantages. In this section, some of the main differences between these transcriptomics technologies will be highlighted.

One of the most prominent differences between DNA microarrays and RNA-seq technologies is that DNA microarrays measure the expression levels of a limited number of genes or transcripts that are chosen *a priori*, whereas the RNA-seq technologies provide an unbiased view of the transcriptome without any *a priori* knowledge on the targets of interest. Thus, RNA-seq experiments have the potential to discover novel regions, which is useful especially while studying complex transcriptomes from higher eukaryotes [Wilhelm and Landry, 2009; Lightbody *et al.*, 2019]. Also, with sufficient sequencing depth, RNA-seq experiments are capable of detecting rare or lowly-expressed genes as well [Wilhelm and Landry, 2009]. On the other hand, if a study involves assessing the expression levels of specific known genes or transcripts, the use of custom-designed or commercialized DNA microarrays could be an easy and cost-effective option. However, it should be noted that DNA microarrays have certain technological shortcomings, such as high background levels due to cross-hybridization, and varying hybridization properties of the probes. These limit its ability of accurately measuring expression levels, especially in case of lowly abundant genes or transcripts [Marioni *et al.*, 2008]. Comparatively, RNA-sequencing is considered to have technical superiority over microarrays [Lightbody *et al.*, 2019; Marioni *et al.*, 2008], but they are also computationally more demanding (in terms of storage as well as analysis).

In order to resolve abundances of gene isoforms (i.e. alternatively spliced transcripts of genes), DNA microarrays (such as, custom-designed splice arrays) or bulk RNA-seq technologies can be employed. Unlike DNA microarrays, bulk RNA-sequencing has the potential of identifying novel isoforms, assuming that the sequencing depth of the cells is sufficiently deep

[Wang *et al.*, 2009; Wilhelm and Landry, 2009]. In principle, scRNA-seq data can also be used for quantifying isoform abundances, but it remains challenging due to higher technical variation in the data as compared to bulk RNA-seq data [Stegle *et al.*, 2015].

Finally, both DNA microarrays and bulk RNA-seq technologies result in gene expression profiles that are averaged across a large population of cells. Conversely, scRNA-seq technologies provide unbiased gene expression profiles of individual cells. Doing so, they capture cell-to-cell heterogeneity that population-based transcriptomics methods cannot do [Stegle *et al.*, 2010; Poirion *et al.*, 2016]. Single cell RNA-seq data enables us to gain several biological insights that are not possible using population-based technologies, such as identifying rare or novel cell types in samples [Wagner *et al.*, 2016]. However, compared to bulk RNA-seq, scRNA-seq technologies produce higher levels of noise and variation in the data [Mou *et al.*, 2020], which makes the computational analysis more challenging than that of bulk RNA-seq data.

3.3 Microbiomics - Microbial community analysis

As discussed in Section 2.3.2, the human gut microbiome plays a very crucial role in human health and its dysbiosis has been associated to a plethora of diseases. However, the mechanisms by which the gut microbiome affects the pathogenesis of most diseases remain largely elusive. Therefore, in order to elucidate these mechanisms, a major focus of human gut microbiome studies has been to identify and characterize the microbes inhabiting the human gut under specific conditions (or circumstances) as well as to decipher their involvement in biochemical pathways by which they may impact host health (i.e. the structural and functional properties of gut) [Malla *et al.*, 2019]. Such studies have unlocked a wealth of data and may offer potential biomarkers for early detection and targets for therapeutic interventions [Cullen *et al.*, 2020; Morgan and Huttenhower, 2012].

Prior to NGS technologies, the traditional culture-based approaches were largely used for studying microbial community profiles, where microbial cells were first isolated and grown in laboratory conditions (i.e. cultured). This method tends to generate a biased view of the microbiota as it generally provides information on only a small proportion of microbes that are capable of being cultured; and also because certain microbes grow faster than others in given culture conditions [Allaband *et al.*, 2019; Morgan and Huttenhower, 2012; Malla *et al.*, 2019]. However, with the advent of NGS technologies and the dramatic reduction in their cost, came about the era of culture-independent approaches that revolutionized microbiome studies by allowing the taxonomic and functional profiling of entire communities in an efficient and unbiased manner [Morgan and Huttenhower, 2012; Pereira *et al.*, 2018; Malla *et al.*, 2019]. Here, taxonomic profiling answers the question ‘who is there?’ and functional profiling answers ‘what are they doing?’. Essentially, NGS-based approaches directly analyse the DNA extracted from microbial cells of a sample without culturing them [Morgan and Huttenhower, 2012]. There are currently two main NGS-based methods for studying microbial communities: marker gene sequencing (also known as amplicon or targeted sequencing; used in Publication IV) and whole metagenome shotgun (WMS) sequencing (used in Publications III and IV) [Pérez-Cobas *et al.*, 2020; Hamady and Knight, 2009]. The marker gene sequencing approach targets specific genes for sequencing that can identify the genome that contains it without needing to sequence the entire genome, i.e. marker genes [Pérez-Cobas *et al.*, 2020; Morgan and Huttenhower, 2012]. Additionally, the chosen marker genes are required to be present in almost all bacteria (or other microbes of interest); and should be highly conserved, such that changes in the sequence would serve as an evolutionary clock and distance measure [Morgan and Huttenhower, 2012; Bharti and Grimm, 2019; Janda and Abbott, 2007]. The most commonly used marker gene for bacterial profiling of microbial samples is 16S ribosomal RNA (rRNA) [Janda and Abbott, 2007; Bharti and Grimm, 2019; Morgan and Huttenhower, 2012; Bik, 2016], which is part of the small subunit of the 70S ribosomes that are found in all prokaryotes (i.e. bacteria and archaea) [Alberts, 2018]. This approach is generally used to identify the members of a microbial community [Pérez-Cobas *et al.*, 2020; Bharti and Grimm, 2019; Bik, 2016]. The WMS sequencing approach, on the other hand, sequences all the genomic DNA in a sample, and is therefore also referred to as the metagenomic sequencing[1] [Hamady and Knight, 2009; Pérez-Cobas *et al.*, 2020; Morgan and Huttenhower, 2012]. WMS sequencing is capable of revealing the microbial composition of communities as well as their genetic content, where the latter can give insights into the functional potential of the microbes [Pérez-Cobas *et al.*, 2020; Bik, 2016]. Even though 16S rRNA sequencing is relatively cheaper and faster [Pérez-Cobas *et al.*, 2020], WMS is quickly displacing it due to its increased accuracy, greater microbial resolution (16S rRNA can assign taxonomy only till the genus-level, whereas WMS data can confidently provide species- and strain-level taxonomic classifications) and capability of detecting genes [Allaband *et al.*, 2019; Brumfield *et al.*, 2020]. However, both methods have their strengths and weaknesses, and can produce varying results [Knight *et al.*, 2018].

The material that is most commonly collected for gut microbial commu-

[1] The term ‘metagenomics’ is often inaccurately used in literature to refer to the entire body of high-throughput sequencing solutions for studying microbial communities, including marker gene sequencing [Morgan and Huttenhower, 2012; Bharti and Grimm, 2019; Hamady and Knight, 2009]

nity studies (including Publications III and IV) is the stool due to its easy accessibility [Allaband *et al.*, 2019]. Microbial cells of a target size are then isolated from the samples for DNA or RNA extraction and lysed using either mechanical or chemical lysing [Bharti and Grimm, 2019]. After library preparation (discussed in Sections 3.3.1 and 3.3.2), the extracted nucleic acid fragments are sequenced. The Illumina sequencing platforms are the most widely used sequencing platforms in microbiome studies at present (including Publications III and IV) as they yield higher output per run and have substantially lower error rates than their predecessor, 454 pyrosequencing, as well as the newer (3^{rd} generation) sequencing platforms, such as Pacific Biosicences, Oxford Nanopore MinION, and Ion Torrent [Bik, 2016; Malla *et al.*, 2019; Pérez-Cobas *et al.*, 2020; Escobar-Zepeda *et al.*, 2015].

3.3.1 16S rRNA sequencing

Data collection

Bacterial 16S rRNA gene is ~1500 bp in length and consists of highly conserved regions separated by nine hypervariable regions (V1-V9) that demonstrate sequence diversity between different bacterial taxa and can be used to identify microbial profiles [Bharti and Grimm, 2019; Allaband *et al.*, 2019]. However, each hypervariable region has different degrees of sequence diversity and no single region is capable of distinguishing between all bacteria [Chakravorty *et al.*, 2007]. Therefore, the choice of hypervariable region(s) for sequencing can influence, for instance, taxonomic coverage [Bik, 2016]. The most commonly sequenced regions include V3-V4, V5-V6 and V4, where V4 is a popular choice in combination with Illumina sequencing and was used in Publication IV [Bharti and Grimm, 2019].

Briefly, 16S rRNA library is prepared by amplifying the hypervariable region(s) of choice using barcoded PCR primer pairs that are complementary to the conserved regions flanking the region(s) of interest [Bharti and Grimm, 2019; Allaband *et al.*, 2019]. Before sequencing, the sequences are purified and constructed into DNA libraries [Bharti and Grimm, 2019]. The most common Illumina platform used for 16S rRNA sequencing is the MiSeq platform (introduced in Section 3.1.2; used in Publication IV) [Pérez-Cobas *et al.*, 2020; Bharti and Grimm, 2019].

16S rRNA data analysis

Similar to the above-mentioned sequencing datasets, microbial HTS data, such as 16S rRNA and WMS sequencing data, are first subjected to quality control (QC) analysis. If the samples have been multiplexed during sequencing, the QC step would be preceded by a demultiplexing step, wherein each sequence is assigned to its sample of origin based on the barcodes [Calle, 2019; Goodrich *et al.*, 2014]. The QC step is critical in microbiome data analysis, as the data can contain sequencing artifacts, such as low-quality reads and contaminating reads from host-genome that can mislead downstream results [Zhou *et al.*, 2014; Bharti and Grimm, 2019; Pérez-Cobas *et al.*, 2020]. While the 16S rRNA approach is generally undeterred by host-genome contamination in the sample [Knight *et al.*, 2018], it is sensitive to sequencing errors, as it can result in an overestimation of microbial diversity in a community and/or lead to incorrect taxonomic annotations [Pérez-Cobas *et al.*, 2020; Bharti and Grimm, 2019; Escobar-Zepeda *et al.*, 2015]. Generally, several QC tools that are applied to other NGS data, such as FastQC [Andrews, 2010], trimmomatic [Bolger *et al.*, 2014], cutadapt [Martin, 2011], etc. can be used to assess the quality of the data as well as trim or remove adapter sequences, short reads and low-quality reads, thus reducing sequencing error among other artifacts [Zhou *et al.*, 2014; Escobar-Zepeda *et al.*, 2015; Pérez-Cobas *et al.*, 2020]. Specific toolkits, such as ea-utils [Aronesty, 2013] (used in Publication IV), have also been developed for effortlessly performing multiple 16S rRNA data processing steps, such as barcode demultiplexing, adapter trimming, etc. in one go. Additionally, PCR errors can introduce chimeras in the 16S data, which can result in inflated diversity estimations [Goodrich *et al.*, 2014]. Therefore, chimera filtering can also be performed as part of the QC step using several tools, such as ChimeraSlayer [Haas *et al.*, 2011], UCHIME [Edgar *et al.*, 2011], and DECIPHER [Wright *et al.*, 2012].

After QC analysis, paired-end 16S rRNA reads are usually joined by overlapping to obtain single reads that are longer and of higher-quality [Aronesty, 2013; Pérez-Cobas *et al.*, 2020]. Several programs have been developed to perform this task, such as fastq-join (implemented in ea-utils) [Aronesty, 2013], PEAR [Zhang *et al.*, 2014], and SeqPrep [John, 2011].

Subsequently, to interpret the microbial structure of a community, the reads are clustered into operational taxonomic units (OTUs)—the lowest level of phylotypes detectable by 16S rRNA sequencing—based on a predefined sequence similarity threshold. The most commonly used threshold for sequence similarity within an OTU is >97%, which allows for some degree of sequence divergence possibly occurring due to sequencing errors [Morgan and Huttenhower, 2012; Bharti and Grimm, 2019; Knight *et al.*, 2018; Pérez-Cobas *et al.*, 2020]. Typically, >97% sequence similarity has been thought to reflect species-level classification [Morgan and Huttenhower, 2012; Edgar, 2013]. However, many studies nowadays consider it to reflect a genus-level classification because many species are identical along the full length of the 16S rRNA gene [Bik, 2016; Allaband *et al.*, 2019]. In fact, it is believed that it is not possible to distinguish taxonomic levels lower than the genus-level using marker gene regions because a wide range of species and strains distinguish from one another only on a

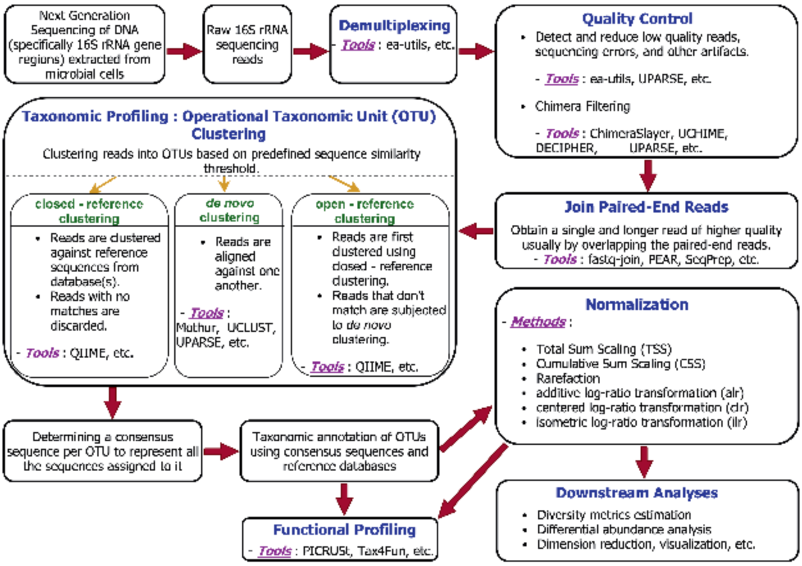

Next Generation Sequencing of DNA (specifically 16S rRNA gene regions) extracted from microbial cells
Raw 16S rRNA sequencing reads
Demultiplexing
- Tools : ea-utils, etc.
Quality Control
• Detect and reduce low quality reads, sequencing errors, and other artifacts.
- Tools : ea-utils, UPARSE, etc.
• Chimera Filtering
- Tools : ChimeraSlayer, UCHIME, DECIPHER, UPARSE, etc.
Taxonomic Profiling : Operational Taxonomic Unit (OTU) Clustering
Clustering reads into OTUs based on predefined sequence similarity threshold.
closed - reference clustering
• Reads are clustered against reference sequences from database(s).
• Reads with no matches are discarded.
- Tools : QIIME, etc.
de novo clustering
• Reads are aligned against one another.
- Tools : Mothur, UCLUST, UPARSE, etc.
open - reference clustering
• Reads are first clustered using closed - reference clustering.
• Reads that don't match are subjected to de novo clustering.
- Tools : QIIME, etc.
Join Paired-End Reads
Obtain a single and longer read of higher quality usually by overlapping the paired-end reads.
- Tools : fastq-join, PEAR, SeqPrep, etc.
Normalization
- Methods :
• Total Sum Scaling (TSS)
• Cumulative Sum Scaling (CSS)
• Rarefaction
• additive log-ratio transformation (alr)
• centered log-ratio transformation (clr)
• isometric log-ratio transformation (ilr)
Determining a consensus sequence per OTU to represent all the sequences assigned to it
Taxonomic annotation of OTUs using consensus sequences and reference databases
Functional Profiling
- Tools : PICRUSt, Tax4Fun, etc.
Downstream Analyses
• Diversity metrics estimation
• Differential abundance analysis
• Dimension reduction, visualization, etc.

Figure 3.2. Depiction of the main steps in 16S rRNA-sequencing data analysis pipeline. The red arrows indicate transition from one step in the data analysis pipeline to another (multiple arrows from a box indicate multiple directions in which the analysis can transition). Yellow arrows are used to indicate the ways in which a step can be carried out.

genome-level [Pérez-Cobas *et al.*, 2020; Allaband *et al.*, 2019].

OTU clustering techniques can be divided into largely three categories: closed-reference clustering, *de novo* clustering, and open-reference clustering. In closed-reference clustering, the reads are clustered against reference sequences from database(s) and those reads that do not match any reference sequence at the defined similarity threshold are discarded [Kopylova *et al.*, 2016; Rideout *et al.*, 2014]. Some databases that store annotated 16S rRNA sequences include GreenGenes [DeSantis *et al.*, 2006], Ribosomal Database Project (RDP) [Wang *et al.*, 2007], and SILVA [Quast *et al.*, 2012]. In *de novo* clustering, the reads are aligned against one another and similar reads (similarity higher than a given threshold) are clustered into the same OTU [Kopylova *et al.*, 2016; Rideout *et al.*, 2014]. Finally, in open-reference clustering, the previous two approaches are combined, such that the reads are first clustered using closed-reference clustering and any reads that fail to match the reference are clustered *de novo* instead of being discarded [Kopylova *et al.*, 2016; Rideout *et al.*, 2014]. Several methods belonging to each category have been proposed over the years. Mothur [Schloss *et al.*, 2009] is one of the most widely used tools that implements three agglomerative hierarchical clustering techniques (i.e. *de novo* clustering), namely nearest neighbour, furthest neighbour and unweighted-pair group method using average linkages (UPGMA) [Schloss and Handelsman, 2005]. Another popular clustering method is UCLUST [Edgar, 2010], which employs USEARCH to assess sequence similarities [Edgar, 2010] and implements a centroid-based greedy algorithm for assigning sequences to clusters [Rideout *et al.*, 2014; Navas-Molina *et al.*, 2013]. Both mothur and UCLUST are also implemented in QIIME [Caporaso *et al.*, 2010], which is an open-source bioinformatics software that integrates commonly used tools designed for 16S rRNA-based microbial community analysis [Navas-Molina *et al.*, 2013]. In fact, the implementations of UCLUST in QIIME can be used to perform all three types of clustering discussed above [Edgar, 2017]. The author of UCLUST also developed the UPARSE pipeline (used in Publication IV), which includes several quality control steps in addition to *de novo* OTU clustering using a novel greedy algorithm, where it performs chimera filtering and OTU clustering simultaneously, thus improving its accuracy [Edgar, 2013].

After (or during) clustering, a consensus sequence per OTU is determined to represent all the sequences assigned to it [Calle, 2019]. These consensus sequences are then usually used to retrieve taxonomic annotations for each OTU from the reference databases [Escobar-Zepeda *et al.*, 2015; Calle, 2019; Knight *et al.*, 2018; Pérez-Cobas *et al.*, 2020]. Some OTUs may remain unannotated or annotated only to a higher taxonomic level as the databases are usually incomplete and imperfect [Pérez-Cobas *et al.*, 2020; Calle, 2019; Knight *et al.*, 2018].

A few methods, such as PICRUSt [Langille *et al.*, 2013] and Tax4Fun [Aßhauer *et al.*, 2015], have been developed to understand the biological functions of the microbial community inferred from 16S rRNA data. However, 16S rRNA data is usually considered insufficient for functional analysis as it does not represent the genomic diversity of the microbial community very well [Brumfield *et al.*, 2020].

Finally, before any computational or statistical analyses (as discussed in Chapter 4) can be performed on the OTU abundance data, normalization steps (as explained in Section 3.3.3) need to be applied. The main steps in the data analysis pipeline for 16S rRNA sequencing data is also depicted in Figure 3.2.

3.3.2 Whole metagenome shotgun sequencing - Metagenomics

Data collection

After extracting the genomic DNA from microbial cells, they are fragmented, prepared for sequencing and then sequenced. Amplification is generally not required here, but can be included [Morgan and Huttenhower, 2012]. The most common Illumina platform used for WMS sequencing is the HiSeq platform (introduced in Section 3.1.2) [Escobar-Zepeda *et al.*, 2015].

Data Analysis
As discussed in Section 3.3.1, WMS sequencing data analysis begins with QC analysis. In addition to the usual read quality issues with NGS data, such as low-quality reads and presence of adapter, WMS data commonly includes contaminating sequences from the host-genome or other sources [Bharti and Grimm, 2019; Zhou *et al.*, 2014]. Therefore, in addition to performing the basic trimming and filtering of reads, contaminating reads are also identified by aligning the reads to the host genome (and/or a set of sequences) using unspliced aligners, such as Bowtie, Bowtie2 or BWA, and subsequently removed [Bharti and Grimm, 2019]. One of the easiest ways to perform QC analysis on WMS sequencing data is to use a wrapper tool, such as KneadData [Huttenhower, 2020] (used in Publication III), that performs all the QC steps one-by-one. KneadData integrates, for instance, FastQC, Trimmomatic, and Bowtie2, in order to perform quality assessment, quality trimming/filtering, and host sequence decontamination, respectively [Huttenhower, 2020; Pereira *et al.*, 2018].

After pre-processing, the WMS reads can be used for taxonomic and functional profiling of microbial communities using largely two different types of approaches, namely assembly-free (i.e. reference-based) and assembly-based approaches [Pérez-Cobas *et al.*, 2020; Knight *et al.*, 2018; Escobar-Zepeda *et al.*, 2015; Morgan and Huttenhower, 2012]. In assembly-free approaches, the short sequencing reads are directly compared against reference genomes and gene catalogues (from genome databases, such as RefSeq [O'Leary *et al.*, 2016] and IMG system [Markowitz *et al.*, 2012], and protein family databases, such as UniRef [Suzek *et al.*, 2015] and Pfam [El-Gebali *et al.*, 2019]) to determine the taxonomic composition and the functional potential of a given microbial community. Whereas, in assembly-based approaches, the short reads are assembled into longer sequences, called contigs, before they are used for taxonomic and functional profiling.

Several assembly-free methods for taxonomic profiling have been proposed over the year, including MetaPhlAn [Segata *et al.*, 2012], MetaPhlAn2 [Truong *et al.*, 2015], Kraken [Wood and Salzberg, 2014], and MEGAN [Huson *et al.*, 2016]. In MetaPhlAn and MetaPhlAn2 (used in Publication III), reads are aligned (using Bowtie2 [Langmead and Salzberg, 2012]) to a reduced catalog of marker genes that are computationally selected from publicly available reference genomes to explicitly identify specific microbial clades (i.e. a group of organisms that are phylogenetically linked) at species or higher taxonomic levels [Segata *et al.*, 2012]. These markers are chosen to be highly conserved within the clade and not present in other clades.

For functional profiling, the reads can be aligned against reference

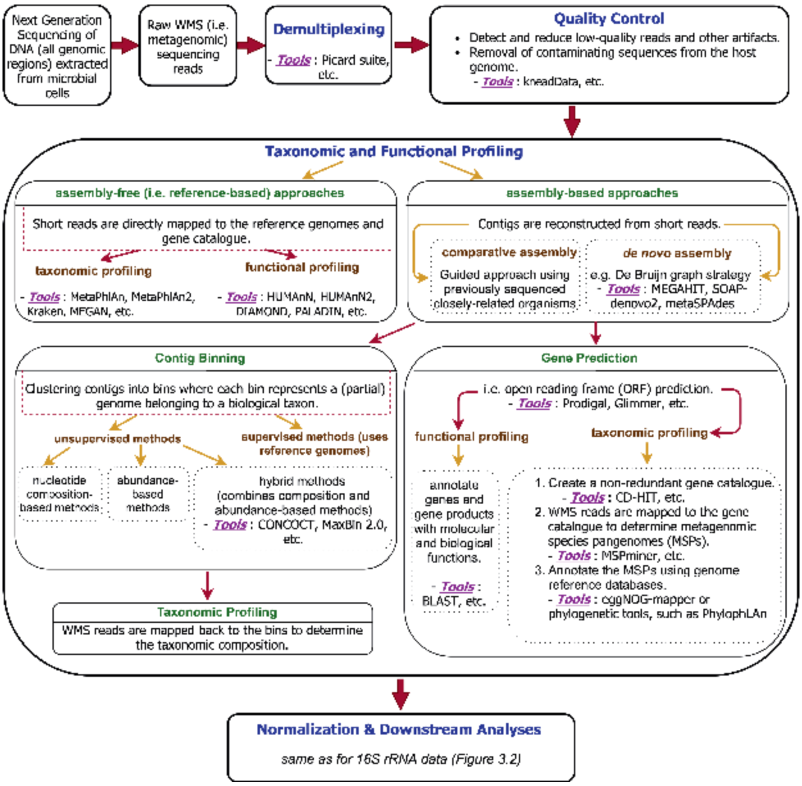

Next Generation Sequencing of DNA (all genomic regions) extracted from microbial cells
Raw WMS (i.e. metagenomic) sequencing reads
Demultiplexing
- Tools : Picard suite, etc.
Quality Control
• Detect and reduce low-quality reads and other artifacts.
• Removal of contaminating sequences from the host genome.
- Tools : kneadData, etc.
Taxonomic and Functional Profiling
assembly-free (i.e. reference-based) approaches
Short reads are directly mapped to the reference genomes and gene catalogue.
taxonomic profiling
functional profiling
- Tools : MetaPhlAn, MetaPhlAn2, Kraken, MEGAN, etc.
- Tools : HUMAnN, HUMAnN2, DIAMOND, PALADIN, etc.
assembly-based approaches
Contigs are reconstructed from short reads.
comparative assembly
Guided approach using previously sequenced closely-related organisms
de novo assembly
e.g. De Bruijn graph strategy
- Tools : MEGAHIT, SOAP-denovo2, metaSPAdes
Contig Binning
Clustering contigs into bins where each bin represents a (partial) genome belonging to a biological taxon.
unsupervised methods
supervised methods (uses reference genomes)
nucleotide composition-based methods
abundance-based methods
hybrid methods (combines composition and abundance-based methods)
- Tools : CONCOCT, MaxBin 2.0, etc.
Taxonomic Profiling
WMS reads are mapped back to the bins to determine the taxonomic composition.
Gene Prediction
i.e. open reading frame (ORF) prediction.
- Tools : Prodigal, Glimmer, etc.
functional profiling
taxonomic profiling
annotate genes and gene products with molecular and biological functions.
- Tools : BLAST, etc.
1. Create a non-redundant gene catalogue.
- Tools : CD-HIT, etc.
2. WMS reads are mapped to the gene catalogue to determine metagenomic species pangenomes (MSPs).
- Tools : MSPminer, etc.
3. Annotate the MSPs using genome reference databases.
- Tools : eggNOG-mapper or phylogenetic tools, such as PhylophLAn
Normalization & Downstream Analyses
same as for 16S rRNA data (Figure 3.2)

Figure 3.3. Depiction of the main steps in whole metagenome shotgun (WMS) sequencing data analysis pipeline. The red arrows indicate transition from one step in the data analysis pipeline to another (multiple arrows from a box indicate multiple directions in which the analysis can transition). Yellow arrows are used to indicate the ways in which a step can be carried out.

databases of protein sequences or KEGG using fast aligners, such as DIAMOND [Buchfink *et al.*, 2015] and PALADIN [Westbrook *et al.*, 2017]. More advanced assembly-free functional profiling tools, such as HUMAnN [Abubucker *et al.*, 2012] and HUMAnN2 [Franzosa *et al.*, 2018] (used in Publication III), have also been developed that allow the inference of the functional and metabolic potential in a microbial community.

Assembly-free methods can provide accurate profiling of microbial communities as well as scale efficiently to large and complex datasets without substantially increasing the computational costs [Knight *et al.*, 2018]. However, they are limited by the availability of reference genomes in databases; single species needs to be isolated and cultured for genome assessment, but some species are impossible to cultivate and culture. Also, the reference databases may represent species of public health interest more extensively than commensal bacteria [Plaza Oñate *et al.*, 2019]. Therefore, assembly-based methods have been developed for analysing metagenomic data, which are not restricted to identifying and quantifying just the known genomes in the community, but also novel or non-referenced genomes. An avalanche of tools have been proposed to carry out one or more steps in the assembly-based analysis pipeline of WMS data. Here, the key steps of the pipeline along with a few widely-used methods for conducting each step will be presented.

Assembly-based methods begin with the reconstruction of contigs from short reads in either a guided manner using previously sequenced closely-related organisms (i.e. 'comparative' assembly) or a *de novo* manner [Pérez-Cobas *et al.*, 2020; Morgan and Huttenhower, 2012]. Of the two, *de novo* methods, specifically those employing De Bruijn graph strategy, are most widely used for metagenomic data [Pérez-Cobas *et al.*, 2020; Bharti and Grimm, 2019]. Some of these methods include MEGAHIT [Li *et al.*, 2015] (used in Publication III), SOAP-denovo2 [Luo *et al.*, 2012], and metaSPAdes [Nurk *et al.*, 2017]. Regardless of the overwhelming number of methods developed for genome assembly, it remains a complicated and challenging task, especially in the context of metagenomes [Pérez-Cobas *et al.*, 2020]. Moreover, the choice of the assembly method is consequential in downstream analysis [Bharti and Grimm, 2019], but no such method exists as of now that is optimal for all datasets and research questions [Pérez-Cobas *et al.*, 2020].

Subsequent to assembly, the contigs can be binned (i.e. classified) into discrete clusters, wherein each cluster (i.e. bin) represents a (partial) genome belonging to a biological taxon [Wu *et al.*, 2016; Pérez-Cobas *et al.*, 2020]. Therefore, by mapping the WMS reads back to the bins, the taxonomic composition of a sample can be determined [Pérez-Cobas *et al.*, 2020]. Both supervised (using the reference genomes) and unsupervised methods have been proposed for contig binning, but the latter methods have become more popular as they do not rely on reference genomes [Pérez-Cobas *et al.*, 2020; Escobar-Zepeda *et al.*, 2015; Bharti and Grimm, 2019]. Unsupervised binning can be done using nucleotide composition-based methods, abundance-based methods or hybrid methods that combine composition- and abundance-based approaches [Pérez-Cobas *et al.*, 2020]. Hybrid methods, including CONCOCT [Alneberg *et al.*, 2014], MaxBin 2.0 [Wu *et al.*, 2016] and several others, tend to perform better than the other two types of methods [Pérez-Cobas *et al.*, 2020]. Additionally, the assembled contigs can also be used for gene prediction, i.e. open reading frame (ORF) prediction, [Escobar-Zepeda *et al.*, 2015; Pérez-Cobas *et al.*, 2020], using methods such as Prodigal [Hyatt *et al.*, 2010] (used in Publication III) and Glimmer [Kelley *et al.*, 2012], among many others. In addition to functional profiling of communities using these predicted genes, they can also be used for taxonomic profiling. One of the ways in which taxonomic profiling using predicted genes can be done is by first removing redundant gene sequences using tools, such as CD-HIT (used in Publication III), and creating a non-redundant gene catalogue. Next, the WMS reads can be mapped to the gene catalogue and metagenomic species pangenomes (MSPs)—core and accessory genes of each species—can be constructed using a tool called MSPminer [Plaza Oñate *et al.*, 2019] (used in Publication III), which bins co-abundant gene across metagenomic samples [Plaza Oñate *et al.*, 2019]. Genome reference databases can then be used to annotate each MSP at different taxonomic levels using tools such as eggNOG-mapper [Huerta-Cepas *et al.*, 2017] (used in Publication III). MSPs without matching taxa in the reference databases can be phylogenetically annotated, using tools such as PhyloPhlAn [Segata *et al.*, 2013] (used in Publication III).

Finally, to determine the functional potential of a microbial community, the predicted gene sequences or their translated gene products can be assigned molecular and biological functions [Morgan and Huttenhower, 2012; Escobar-Zepeda *et al.*, 2015] using tools such as BLAST [Ye *et al.*, 2006]. The genes and gene products can also be annotated with more informative functional categories, such as GO terms, KEGG terms and MetaCyc pathways; and orthologous families, such as COGs and KOs [Morgan and Huttenhower, 2012].

Furthermore, the functional capacity of human microbiomes can also be determined by the strain-level variants within microbial species [Nayfach and Pollard, 2016; Truong *et al.*, 2017]. Specific strains of many species can have a pathogenic potential and can be associated to the phenotype of the health condition [Truong *et al.*, 2017]. Strain-level differences in microbes can occur in the form of single nucleotide polymorphisms (SNPs) as well as addition/deletion of genomic elements, such as genes, plasmids or operons. With the refined resolution provided by WMS sequencing data, it is possible to characterize microbial communities at the strain-level by identifying genomic variations, such as SNPs (using methods like StrainPhlAn [Truong *et al.*, 2017], for instance) and accessory genes (as is done by MSPminer [Plaza Oñate *et al.*, 2019]). The latter method was used in Publication III.

Lastly, similar to the 16S rRNA data, the taxonomic and functional-level abundance data obtained from the above-mentioned processing pipelines have to be normalized before any computational or statistical analyses can be performed on them. Some of the normalization methods are discussed in Section 3.3.3. The main steps in the data analysis pipeline for WMS sequencing data is also depicted in Figure 3.3.

3.3.3 Normalization of abundance data

In order to perform meaningful comparisons between samples or other downstream analyses, the taxonomic- and functional-level raw abundance data (originating from 16S rRNA or WMS sequencing data [Weiss *et al.*, 2017; Pereira *et al.*, 2018]) need to be normalized. Some of the most simple and commonly used normalization methods for microbiome data include total sum scaling (TSS) (used in Publications III and IV), where individual raw counts are divided by the total number of counts per sample, resulting in relative abundances that sum to 1; and rarefaction, where the same number of reads are subsampled from each sample to ensure equal number of total counts in all samples [Pérez-Cobas *et al.*, 2020; Gloor *et al.*, 2017; Calle, 2019; Pereira *et al.*, 2018]. TMM and DESeq normalization methods from RNA-seq data analyses can also be applied to microbiome data [Knight *et al.*, 2018], but they are considered less suitable for highly asymmetrical and sparse datasets like those of microbiome data [Gloor *et al.*, 2017], and their assumptions are likely to be inappropriate for highly diverse microbial communities [Weiss *et al.*, 2017]. Many normalization methods have been proposed over the years for microbiome data, including cumulative sum scaling (CSS) [Paulson *et al.*, 2013]. Several comparative studies have been conducted to identify the best-fitting normalization methods for microbiome data [Paulson *et al.*, 2013; Pereira *et al.*, 2018; Weiss *et al.*, 2017], but the results tend to vary from one study to another. In fact, Paulson *et al.* [2013] claimed that CSS outperformed TSS, DESeq and TMM, but Costea *et al.* [2014] later refuted this and showed that the improved results were mere artifacts of preferential post-processing steps.

To account for the compositional nature of microbiome data, Aitchison proposed transforming the raw count data into compositional data using log-ratio transformation techniques, such as additive, centered and isometric log-ratio transformations (i.e. alr, clr, and ilr) [Gloor *et al.*, 2017; Calle, 2019]. Essentially, TSS-normalized relative abundance data is also considered compositional data, but the data is in the Simplex space where Euclidean metrics are not valid [Gloor *et al.*, 2017; Calle, 2019]. However, the log-ratio transformed data are in the Euclidean space. In fact, TMM and DESeq are similar to log-ratio transformations [Gloor *et al.*, 2017].

4. Statistical and Computational Analyses

As established in Chapter 3, high-throughput ‘omics’ data analysis starts with a collection of raw and noisy data that often mandates ‘cleaning’ (i.e. quality checking) and has to undergo several data processing steps. This essentially transforms the raw information into a more meaningful and analysable format, such as a molecular-level characterization or quantification. Subsequently, various statistical and computational techniques can be employed to analyse the processed data and thereby address the research questions pertaining to the study. For instance, computational techniques, such as dimension reduction, clustering, visualization and microbial diversity estimations, can be performed to gain insights into the underlying structure of the data and to extract specific patterns that may exist. Statistical modelling provides a powerful mathematical framework for explaining the data in terms of underlying covariates (i.e. explanatory variables) that may be associated with the variation in the data. These models can also be used to account for confounding factors while testing the association of more interesting covariates. Statistical modelling tools, such as linear models and Gaussian processes, are often used for performing differential expression and differential abundance analyses in transcriptomics and microbiomics studies, respectively. Moreover, computational analyses can be performed prior to statistical modelling in order to gain perspective on the covariates that should be included in the model.

This chapter will cover some of the most prominent computational and statistical analyses that can be performed on the transcriptomics and microbiomics datasets.

4.1 Microbiome Diversity Analysis

One of the most widely studied attributes of a microbial community is its diversity, as it can be an important indicator of a symbiotic or dysbiotic microbiome (as mentioned in Section 2.3.2). Two types of measures are commonly used to describe the microbial diversity: alpha and beta diversity.

Alpha diversity quantifies the variety of taxa within a microbial community or sample, whereas beta diversity measures the (dis)similarity between two communities or samples. Several estimators have been proposed to quantify the alpha and beta diversities of microbial communities derived from either 16S rRNA or WMS sequencing data.

4.1.1 Measuring alpha diversity

The alpha diversity of a community can be estimated by measuring its richness (i.e. the number of distinct microbial species/OTUs in the community) or its evenness (i.e. the homogeneity in the abundances of the species/OTUs).

Richness of a microbial community can be estimated, for instance, by merely counting the number of observed species/OTUs in the community or by computing the Chao1 index [Chao, 1984] (used in Publication IV), S_{chao1}, according to Equation 4.1, which adjusts for the lowly abundant or rare species in the community that are likely to remain undetected due to sequencing depth limitations. This is performed by taking into account the frequencies of species that are observed exactly once and twice in the community (one or two reads with specific sequence), i.e. singletons and doubletons, respectively.

$$S_{chao1} = S_{obs} + \frac{f_1^2}{2f_2}, \tag{4.1}$$

where S_{chao1} refers to the number of species/OTUs observed in the community, f_1 refers to the number of singleton species, and f_2 refers to the number of doubleton species.

Since communities that are dominated by a few species are generally considered less diverse than those with several species with similar abundances, alpha diversity estimators that measure the evenness within communities are also widely used. One of the most commonly used measures for evenness is the Shannon index [Shannon, 1948] (used in Publications III and IV), H, which increases with the richness of the sample and gives more weight to lowly abundant species:

$$H = -\sum_{i=1}^{k} p_i \ln(p_i) \tag{4.2}$$

where p_i is the relative abundance of taxon i in the sample, such that $\sum_{i=1}^{k} p_i = 1$; and k is the total number of taxa in the sample.

4.1.2 Measuring beta diversity

One of the most widely used measures for estimating beta diversities between samples is Bray-Curtis dissimilarity [Bray and Curtis, 1957] (used in Publications III and IV):

$$\mathrm{BC}_{ij} = \frac{\sum_{t=1}^{k} \left| x_{ti} - x_{tj} \right|}{\sum_{t=1}^{k} \left| x_{ti} + x_{tj} \right|}, \tag{4.3}$$

where x_{ti} and x_{tj} are the relative abundances (for instance, TSS-normalized data) or counts of taxon t in samples i and j, respectively; and k is the total number of taxa observed in the samples altogether. Due to the proportional nature and highly skewed distributions of the relative abundances, Euclidean distances would not be suitable to estimate beta diversities between microbial communities (Section 3.3.3).

4.2 Dimension reduction and visualization

An informative analysis that is often performed on high-dimensional datasets is dimensionality reduction (or ordination), where the data is projected from a high-dimensional space onto a lower dimensional (D) space, such as 2- or 3-D space, while preserving as much of the significant trends, structure and information of the original data as possible in the low-dimensional space. This essentially enables us to visualize the high-dimensional data in a 2- or 3-D space, which in turn allows us to obtain a compact global view of the data and visually identify possible structures, trends or outliers in the data.

Principal component analysis (PCA) [Hotelling, 1933] (performed in Publication I) is one of the most widely-used dimensionality reduction methods that identifies a set of orthogonal variables (i.e. basis vectors), called principal components (PCs), as a linear combination of the original set of variables (such as, genes) by maximizing the amount of variance preserved from the original dataset in the lower dimension projection. Fundamentally, PCA performs eigen decomposition on the covariance matrix of the original set of variables. The resulting eigenvectors correspond to the PCs, whereas the respective eigenvalues inform about the variance in the data explained by each PC. Here, the first PC corresponds to the eigenvector with the largest eigenvalue and subsequent PCs correspond to eigenvectors in the order of decreasing eigenvalues [Alpaydin, 2010].

However, since PCA functions in the Euclidean space, it is not suitable for data in non-Euclidean space, such as microbiome relative abundance data. In such cases, methods that can obtain a low-dimensional Euclidean representation of data points from non-Euclidean space by measuring their relationships using any distance or (dis)similarity metric are often preferred. One such method is the principal coordinate analysis (PCoA), also known as classical multidimensional scaling (MDS) [Torgerson, 1952] (performed in Publication III), which tries to preserve as much of the original relationships (i.e. distances or (dis)similarities) between the data points from the high-dimensional space as possible in the low-dimensional projections. In the case of microbiome data, any beta diversity measure, including Bray-Curtis dissimilarity measure (as done in Publication III), can be used to compute the dissimilarities between samples for performing PCoA ordination [Legendre *et al.*, 1998].

Another method that has become widely popular for visualizing high-dimensional datasets on 2- or 3-D space is t-distributed stochastic neighbour embedding (t-SNE) [Maaten and Hinton, 2008] (availed in Publication I). Briefly, this state-of-the-art method aims to preserve as much of the high-dimensional structure of the data as possible in the low-dimensional representation. It does so by minimizing the Kullback-Leibler (KL) divergence [Kullback and Leibler, 1951] between the probability distributions of the pairwise similarities between the data points in the high-dimensional space, P, and low-dimensional space, Q:

$$\mathrm{KL}(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} , \tag{4.4}$$

where p_{ij} and q_{ij} are similarity measures between data points i and j in the high- and low-dimensional spaces, respectively, defined by Gaussian and Student-t distributions. As KL divergence is a way of comparing two probability distributions, minimizing Equation 4.4 would ensure that the pairwise similarities between data points in the original data is preserved in its low-dimensional representation. Moreover, t-SNE visualization can be applied to data in the Euclidean space [Maaten and Hinton, 2008] and has been extended to visualize non-Euclidean (dis)similarity data as well [Van der Maaten and Hinton, 2012].

4.3 Clustering

Clustering is a popular machine learning technique that aims to classify features (or elements) of a data[1] into groups (or clusters) based on their similarities in an unsupervised manner. Cluster analysis often provides valuable insights into the structure of the data as well as reveal meaningful patterns that may exist in the data. For instance, clustering can be used to identify co-regulated genes in transcriptomics analyses; and in single cell RNA-seq analyses, clustering techniques are often used to form clusters of single cells, such that the cells belonging to each cluster share similar gene expression profiles and are thus likely to represent a specific cell type.

Numerous clustering algorithms have been proposed over the past decades, where a few methods are used more regularly than others, including k-means clustering, hierarchical agglomerative clustering, and density-based clustering. In this thesis, particularly in Publication I, k-means and hi-

[1] Referred to as data points while explaining the algorithms

erarchical clustering were performed on bulk RNA-seq data to identify co-regulating genes; whereas density-based clustering was performed on scRNA-seq data to cluster the single cells. k-means is one of the simplest and most well-known clustering techniques that iteratively partitions the data points into a predefined number of clusters, k, by minimizing the distance between the data points and the nearest cluster centers [Alpaydin, 2010]. Here, k needs to be decided beforehand, which can be done based on prior knowledge of the data or by using techniques, such as silhouette scoring. A silhouette score evaluates how similar a data point is to the members of its own cluster as compared to those of other clusters [De Amorim and Hennig, 2015]. However, k-means cannot detect non-spherical clusters [Rodriguez and Laio, 2014].

Hierarchical agglomerative clustering, on the other hand, does not require the number of clusters to be specified beforehand and can be inferred after completing the clustering process. The method starts by treating each data point as a single cluster and then successively merges (or agglomerates) pairs of clusters with the smallest distance until a stopping criterion is met or all clusters are merged into a single cluster containing all the data points. The hierarchy of clustering is often represented as a dendogram (i.e. tree) [Alpaydin, 2010].

Another widely-used clustering method is the density-based clustering, which can easily identify non-spherical (or arbitrarily-shaped) clusters. The notion behind such clustering is that the regions in the data space that contain the clusters have high density of data points, whereas regions that contain noise/outliers or no clusters have low density of data points. Density-based spatial clustering of applications with noise (DBSCAN) [Ester *et al.*, 1996] is one of the most popular density-based clustering methods. It identifies clusters such that each data point, p, in a cluster has either a minimum number of data points, $MinPts$, in its neighborhood (i.e. in a given distance, ϵ, of p) or is in the neighborhood of such high-density data point(s); and the data points that do not meet this criteria are considered as noise. Simply put, in DBSCAN, data points that are densely packed and contain many neighbors are clustered together, whereas those that are alone with far away or insufficient number of neighbours are classified as noise [Ester *et al.*, 1996]. Recently, Rodriguez and Laio [2014] proposed a new density-peak clustering (DPC) method that performs the clustering process by first identifying the density peaks (or cluster centers) of clusters and subsequently assigning data points to the nearest density peaks. Briefly, two quantities are computed for each data point i: local density, ρ_i and minimum distance, δ_i, to the nearest data point with higher local density. The density peaks are then identified as the data points with relatively high ρ and δ values; and outliers are the data points with relatively low ρ, but high δ values. Subsequently, the remaining data points are assigned to the nearest density peaks [Rodriguez and Laio, 2014]. Both DBSCAN and DPC intuitively define the number of clusters in the data without prior input [Ester *et al.*, 1996; Rodriguez and Laio, 2014].

4.4 Linear modelling

Linear models are some of the most powerful and prevalent statistical modelling tools that are employed in 'omics' studies, including transcriptomics and microbiomics. They can be used to efficiently describe or summarize the observed data (such as gene expression and microbial abundance data) as a linear combination of a set of explanatory variables (such as disease status, treatment information) that may explain some of the variation in the observed data. In statistics, the observed data are referred to as response or dependent variable(s), whereas the explanatory variables are referred to as predictors or independent variables. Essentially, linear models can capture the relationship or association between the response variable(s) and the predictors while accounting for any uninteresting sources of variation from confounding factors (such as subject ID). In fact, linear models are typically used for statistically inferring the significance of the association between the predictors and the response variable(s). This inference can be made upon estimating the relevant parameters (i.e. fitting the linear model to the data) and subsequently performing hypothesis testing on the parameters of interest, such as regression coefficients.

There are several types of linear models, some of which will be covered in this section.

4.4.1 Linear regression models

The most basic type of linear models that are used to understand data are simple and multiple linear regression models [Milton and Arnold, 2003; Rencher and Schaalje, 2008], hereon collectively referred to as linear regression models. These models express the value of the response variable, y_i, in sample i, as a function of p predictors, $x_{i1}, x_{i2}, ..., x_{ip}$:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i \ , \ i = 1, ..., n \tag{4.5}$$

Here, β_0 and β_j ($j = 1, ..., p$) that denote the intercept and the regression coefficients (or slopes) of x_{ij}, respectively, are estimated by fitting the linear model to data. The ϵ_i is the error term (or residual) that follows a Gaussian distribution, $\mathcal{N}(0, \sigma^2)$ [Milton and Arnold, 2003]. In this model, the predictors can also be referred to as fixed-effects, which indicate that they have a predictable and constant influence across all samples in the data.

After estimating the regression coefficients, hypothesis testing can be performed to infer the significance of the corresponding fixed-effects in terms of their association to the response variable. The hypothesis testing can be performed either individually or simultaneously on the regression coefficients of the fixed-effects. For instance, t-tests can be used to assess the significance of individual coefficients, whereas F-tests can be used for testing a combination of coefficients [Pinheiro and Bates, 2006; Allen, 1997]. Notably, when assessing the significance of a single fixed-effect, the F-test statistic is equivalent to the t-test statistic [Galecki and Burzykowski, 2013]. In the t-test, the null hypothesis states that the regression coefficient is equal to zero, and the t-statistic can be computed as such:

$$\text{t-statistic} = \frac{\beta_{\text{estimated}} - \beta_{\text{null}}}{\text{se}(\beta_{\text{estimated}})}\ , \tag{4.6}$$

where $\beta_{\text{estimated}}$ is the coefficient estimate of the fixed-effect of interest, β_{null} is the null hypothesis (i.e. equal to zero), and $se(\beta_{\text{estimated}})$ is the estimated standard error of the coefficient estimate [Allen, 1997; Galecki and Burzykowski, 2013]. Finally, a p-value can be determined using the t-distribution with the respective degrees of freedom [Allen, 1997; Galecki and Burzykowski, 2013].

Linear regression models, such as the ones implemented in a tool called Multivariate Association with Linear Models (MaAsLin) [Morgan *et al.*, 2012] and the *lm()* function of the *stats* R package [R Core Team, 2019], are commonly used to model microbial community data. Both of these methods use t-tests to assess the significance of the regression coefficients [R Core Team, 2019; Pinheiro *et al.*, 2018]. Since the distribution of microbial data is highly skewed, the relative abundances are often transformed before linear modelling to make the data more suitable for linear model with Gaussian error residual and homoscedastic noise. MaAsLin, which is a tool developed specifically for microbial data analysis, implements an arcsine square root transformation of the data, $\arcsin\sqrt{y_i}$, that stabilizes the variance (i.e. ensures homoscedasticity); but generally logarithmic transformation, $\log(y_i)$ can also be applied. In Publication III, the microbiome data was $\log_{10}$-transformed and analysed using the *lm()* function [R Core Team, 2019] in R, whereas in Publication IV, data transformation and linear modelling were performed using MaAsLin.

Alternatively, data from non-normal observation model can be modelled directly using methods such as generalized linear models (discussed further in Section 4.4.3) and they are likely to be more robust for modelling non-normally distributed data.

4.4.2 Linear mixed-effects models

One of the main assumptions that is made by a linear regression model (Equation 4.5) is that the observations, y_i, are independent of each other

[Rencher and Schaalje, 2008]. Violation of the independence assumptions may lead to spurious and/or misleading results. Typically, when a single observation is collected per individual or group, the response data is considered to be independent and uncorrelated. However, when multiple samples are collected from an individual or a group, the repeated observations cannot be regarded as independent and are typically correlated [Rencher and Schaalje, 2008; Agresti, 2015].

Therefore, to model such non-independent response data, linear mixed-effects (LME) model can be employed. In an LME model, two types of predictors are used to model the data: fixed-effects (introduced in Section 4.4.1) and random-effects [Agresti, 2015]. As opposed to fixed-effects, random-effects are predictors that have an unpredictable, idiosyncratic or 'random' influence on the data. A general form of linear mixed-effects model can be represented as:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i \ , \quad i = 1,...,m \tag{4.7}$$

where, for individual or group i, $\mathbf{y}_i$ is a vector of n_i observations, $\mathbf{X}_i$ is the design matrix for the fixed-effects; $\boldsymbol{\beta}$ is a vector corresponding to the regression coefficients of the fixed-effects; $\mathbf{Z}_i$ is the design matrix of the random-effects; $\boldsymbol{\gamma}_i$ is a vector corresponding to the regression coefficients of the random-effects that has a Gaussian prior, $\mathcal{N}(0,D)$; and $\boldsymbol{\epsilon}_i$ is a vector of error terms (or residuals) corresponding to each observation in i and follows a Gaussian distribution, $\mathcal{N}(0,\Sigma_i)$.

In longitudinal data, when multiple samples are collected from an individual over a period of time, a random effect predictor can be used to model within-individual correlations. For instance, the random effect predictor can model random intercepts, such that the samples from each individual will have their own intercept. An LME model with a random intercept predictor can be presented as:

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \gamma_i + \epsilon_{ij} \ , \quad j = 1,...,n_i \tag{4.8}$$

where, y_{ij} is the j^{th} observation from individual i; $\mathbf{x}_{ij}$ is a vector of predictors for observation j corresponding to the fixed-effects; γ_i is the random intercept of individual i; and ϵ_{ij} is the error term (or residual) that follows a Gaussian distribution, $\mathcal{N}(0,\sigma^2)$.

MaAsLin [Morgan *et al.*, 2012] can also fit linear mixed-effects model (as implemented in the *MASS* R package [Venables and Ripley, 2002]) to each microbial taxon at a time and was employed in Publication IV to analyze longitudinal data. It tests the significance of each fixed-effect individually by performing linear regression t-tests (as explained in Section 4.4.1 and depicted in Equation 4.6) on the regression coefficients [Pinheiro and Bates, 2006].

4.4.3 Generalized linear models

While linear models discussed in Sections 4.4.1 and 4.4.2 are powerful statistical modelling tools, they make certain assumptions about the structure of the data that may not be applicable to all datasets. In particular, these linear models assume a linear relationship between the mean of the response variable and the predictors, and assume that the variation of the response variable (i.e. error distribution) is homoscedastic and follows a normal distribution [Agresti, 2015]. Therefore, they are not appropriate for modelling datasets with non-linear and/or non-normal structure. One way of addressing this issue is to suitably transform the response variable such that it complies with the linear model assumptions, as mentioned in Section 4.4.1. However, for most data, it can be challenging to find an appropriate transformation [Agresti, 2015; Gelman *et al.*, 2004]. Therefore, it may be more robust and plausible to conform the linear model to the characteristics of the data instead, as is done by generalized linear models (GLMs) [Nelder and Wedderburn, 1972].

GLMs are an extension of the linear models that can be applied to response variables that are non-linearly related to the predictors and/or are non-normally distributed [Gelman *et al.*, 2004; Agresti, 2015; Crosbie and Hinch, 1985]. They consist of three components, namely a random component that specifies the probability distribution of the response variable; a linear predictor that specifies the linear combination of the explanatory variables; and a link function, $g(\cdot)$, that specifies how the mean of the response variable is related to the linear predictor,

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} \quad \Longleftrightarrow \quad \boldsymbol{\mu} = g^{-1}(\mathbf{X}\boldsymbol{\beta}) \tag{4.9}$$

where $\mathbf{X}$ is a matrix that contains the predictors (column-wise) and $\boldsymbol{\beta}$ is the vector of regression coefficients.

GLMs are widely-used in the statistical analysis of RNA-seq data, especially in differential expression analyses of genes (or transcripts) between different experimental conditions, as they can accommodate the count-based nature of RNA-seq data and thus give higher statistical power than approximate normal models [McCarthy *et al.*, 2012]. Even though Poisson distribution models were initially introduced for modelling RNA-seq count data [Marioni *et al.*, 2008; Garber *et al.*, 2011], they were soon found to be unsuitable for such data since they can account only for the technical variations and not the biological variations across samples [Anders *et al.*, 2013; Garber *et al.*, 2011; Conesa *et al.*, 2016]. Instead, a generalization of the Poisson distribution, called negative binomial (NB) distribution (i.e. gamma-Poisson), became a widely-adopted and befitting approximation for modelling RNA-seq data [McCarthy *et al.*, 2012; Anders *et al.*, 2013; Agresti, 2015; Conesa *et al.*, 2016], as it has an additional parameter for estimating overdispersion (i.e. greater variation) in the data that generally exists due to the biological variation between samples [Agresti, 2015; Robinson *et al.*, 2010]. Basically, an NB model makes an assumption that an observation, y_{gi} (i.e. the read count of gene g in sample i), has a mean μ_{gi} and a variance that is a function of μ_{gi} and a dispersion parameter, ϕ_g:

$$\mathrm{var}(y_{gi}) = \mu_{gi} + \phi_g \mu_{gi}^2 \tag{4.10}$$

Several methods have been developed for performing differential expression analyses (DEA) on RNA-seq count data [Hardcastle and Kelly, 2010; Zhou *et al.*, 2011; Tarazona *et al.*, 2011; Love *et al.*, 2014], including edgeR [Robinson *et al.*, 2010] and DESeq [Anders and Huber, 2010]. Comparative studies have found large differences between different DEA algorithms and have been unable to determine a single method that performs optimally in all datasets under all circumstances [Zhao *et al.*, 2016; Soneson and Delorenzi, 2013]. However, edgeR [Robinson *et al.*, 2010] is among the top performers [Anders *et al.*, 2013; Soneson and Delorenzi, 2013] and is one of the most widely used tools for DEA of RNA-seq data [Anders *et al.*, 2013].

edgeR is a powerful and flexible framework that implements an NB model using GLMs. It is suitable for modeling count data even from complex experimental designs, such as paired samples. Essentially, it fits a log-linear model for each gene [McCarthy *et al.*, 2012]:

$$\log \mu_{gi} = \mathbf{x}_i^T \beta_g + \log N_i \ , \tag{4.11}$$

where $\mathbf{x}_i^T$ is a vector of predictors and β_g are regression coefficients of the predictors for gene g, that are estimated using maximum likelihood estimation (MLE). Here, N_i is either the original library size (i.e. the total number of reads mapped in sample i) or the effective library size (i.e. the original library size multiplied/divided by the square root of the estimated TMM scaling factor [Robinson and Oshlack, 2010]). The N_i parameter is an offset that is built into the model to account for the library size and thus normalize the data (Section 3.2.2) [McCarthy *et al.*, 2012].

edgeR estimates the dispersion, ϕ_g, of each gene g individually by maximizing the Cox-Reid adjusted profile likelihood (APL) of the gene [McCarthy *et al.*, 2012; Cox and Reid, 1987]. However, due to small sample sizes in most biological studies, McCarthy *et al.* [2012] (the authors of edgeR) advise against using these gene-wise estimates of dispersion for modelling. They state that unless the data contains a large number of samples, a reliable estimation of the dispersion requires some sort of sharing of information between the genes. EdgeR provides three options for dispersion calculations that shares information across genes in some way: 1) common dispersion, which maximizes a shared profile likelihood function, 2) trended dispersion, where dispersion-per-gene is modeled as a smooth function of the average read count of each gene, and 3) gene-wise dispersions, where a weighted shared likelihood component is added to the individual genewise dispersions for maximizing [McCarthy *et al.*, 2012].

Finally, with reliable estimates of all the parameters, edgeR performs gene-wise likelihood ratio tests (LRT) to assess the significance of the regression coefficient(s) of interest, where the null hypothesis is that the coefficient is equal to zero [McCarthy *et al.*, 2012].

Lastly, the predictors in GLMs are usually fixed effects, but GLMs can be extended to include random effects in the linear predictor (as shown in Section 4.4.2). This extension of the generalized linear model is known as the generalized linear mixed model (GLMM) [Agresti, 2015; Schall, 1991], which can be useful in modelling, for example, longitudinal non-linear or non-normal data.

4.4.4 Non-parametric multivariate analysis of variance

As microbial composition can be influenced by a variety of external variables, such as age, disease group, treatment type, geographical location, etc., a prominent aspect of microbial community analysis is to explore whether the compositional differences between samples could be attributable to these variables. Such explorations can be performed using univariate models, such as linear models, that can identify individual taxa (or functional categories, such as genes or pathways), whose relative abundances are associated with the external variables; or using multivariate models that can identify specific external variables that may be associated with the whole community-level compositional variations.

Traditional multivariate models, such as multivariate analysis of variance (MANOVA) and other ANOVA-based methods, assume the data to be normally distributed. However, owing to the highly skewed abundance distributions of taxa in microbial datasets, such traditional multivariate models would not be suitable for analysing them. Instead, several non-parametric methods have been developed for such microbial community analyses [Mantel, 1967; Clarke, 1993; Anderson, 2001]. The non-parametric multivariate analysis of variance (PERMANOVA) method proposed by Anderson [2001] (implemented in the *adonis()* function of the *vegan* R package [Oksanen *et al.*, 2019]) is one of the most popular and suitable methods for performing multivariate microbial association analyses with both categorical and continuous external variables. Thus, it was availed in Publications III and IV for performing multivariate association analyses.

Briefly, PERMANOVA tests the microbial compositional variations between different groups by comparing the within-groups variability with the between-groups variability of samples (i.e. data points) using any distance or dissimilarity measure, including Bray-Curtis dissimilarity measure (Section 4.1.2), and a pseudo F-statistic. Traditionally, ANOVA-based methods compare within-groups and between-groups variability by calculating sum of squared Euclidean distances between the data points and their group mean (i.e. centroid of the data points). However, the centroid of data points in the Simplex space (such as microbiome data) does not usually correspond to the mean of the data points and can be problematic to determine [Anderson, 2001]. Therefore, Anderson [2001] circumvents this problem by identifying that the sum of squared distances from the data points to their centroid was equal to the sum of squared inter-point distances (or dissimilarities) divided by the number of data points. Thus, PERMANOVA computes the total sum of squares (SS_T), within-group SS (SS_W) and between-groups SS (SS_B) as such:

$$\begin{aligned} SS_T &= \frac{1}{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}^2 \ , \\ SS_W &= \sum_{k=1}^{g} W_k \ , \text{ where } \ W_k = \frac{1}{n_k} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}^2 \epsilon_{ij}^{[k]} \ , \ \text{ and} \\ SS_B &= SS_T - SS_W \end{aligned} \tag{4.12}$$

where N is the total number of data points; g is the total number of groups; n_k is the number of data points in the k^{th} group; d_{ij} is the distance or dissimilarity between data points i and j; and $\epsilon_{ij}^{[k]}$ is 1 if data points i and j are in group k, and 0 otherwise [Anderson, 2014, 2001]. The null hypothesis is then tested using the following pseudo F-statistic:

$$F = \frac{SS_B/(k-1)}{SS_W/(N-k)} \tag{4.13}$$

Finally, a permutation-based p-value is computed to assess the statistical significance of the pseudo F-statistic.

4.5 Gaussian Processes

With the decreasing costs of high-throughput ‘omics’ technologies (Section 3.1), longitudinal (i.e. time-series) studies have become commonplace in biomedical research. For instance, longitudinal gene expression data are routinely analysed to understand the transcriptional dynamics involved in various diseases. Traditionally, parametric models, such as LMM and GLMM (Sections 4.4.2 and 4.4.3), that ‘absorb’ the training data into a finite set of parameters while training the model and make predictions on unobserved input data independent of the training data [Rasmussen, 2003; Roberts *et al.*, 2013], have been used for modelling longitudinal data [Rasmussen, 2003]. Recently, non-parametric models, such as Gaussian processes (GPs), have become a popular choice for modelling longitudinal datasets (such as in Publication II), as they make less assumptions about the underlying structure of the data and are generally more flexible than parametric models [Alpaydin, 2010; Cheng *et al.*, 2019]. Contrary to parametric models, non-parametric models do not assume that the training data can be defined in terms of a finite set of parameters and they take into account the training data while making predictions [Roberts *et al.*, 2013; Alpaydin, 2010].

A Gaussian process (GP) is a powerful non-parametric modelling tool that can be used for both regression and classification tasks. It is a flexible class of models that can capture the true underlying signal of the data and give a reliable estimate of the model uncertainty [Rasmussen and Williams, 2006]. This section will focus on discussing Gaussian process regression that has become a popular method for modelling of longitudinal data, such as gene expression data, in a non-linear manner.

The main objective of a regression task is to fit a function to the training data observed at specific input variables and thus enable prediction of data at any input variable. In case of longitudinal data, the input variables could be the time points as well as other covariates. Given a set of training data, there are generally infinitely many functions that can fit the data. A Gaussian process, which is defined as a collection of any finite number of random variables[2] that have a joint Gaussian distribution (i.e. multivariate Gaussian distribution) [Rasmussen, 2003], offers a probabilistic solution to the problem by assigning a probability to each of the possible functions that can fit the data. Thus, a GP can be seen as defining a probability distribution over functions, where the mean of the distribution represents the most probable characterization of the data. For input variables $\{t_i, t_j\} \in \mathbf{T}$, where $\mathbf{T} = (t_1, t_2, \ldots, t_N)$, a GP can be fully specified by its mean function, $\mu(t)$, and covariance function, $k(t_i, t_j)$, as:

$$f(t) \sim \mathrm{GP}(\mu(t),\ k(t_i, t_j)) \tag{4.14}$$

Here, the covariance function, also called the kernel of GP, generates a symmetric and positive semi-definite covariance matrix that describes the similarity between all pairwise combinations of the random variables and also controls the shape of the fitted function. A wide variety of kernels are available for inferring GPs [Roberts *et al.*, 2013; Cheng *et al.*, 2019]. Among these, the squared exponential function is one of the most widely-used kernels [Rasmussen, 2003; Roberts *et al.*, 2013] and is defined as:

$$k(t_i, t_j) = \sigma_{\text{se}}^2 \exp\left(-\frac{(t_i - t_j)^2}{2\ell_{\text{se}}^2}\right) \tag{4.15}$$

where σ_{se}^2 is the signal variance of the covariance function that determines distance of the function from the mean, and ℓ_{se} is the length-scale parameter that controls the smoothness. This kernel assigns high correlation to

[2] Here, random variables refer to the training data and/or the test data

the input variables that are close to each other.

A GP regression has a Gaussian prior, $f(\mathbf{T}) \sim \mathcal{N}(0, \mathbf{K}_{\mathbf{T},\mathbf{T}})$, where $f(\mathbf{T}) = (f(t_1), f(t_2), \ldots, f(t_N))^{\mathrm{T}}$, the mean is set to zero for convenience, and $\mathbf{K}_{\mathbf{T},\mathbf{T}}$ is the covariance matrix that contain $k(t_i, t_j)$ elements [Rasmussen, 2003; Roberts *et al.*, 2013]. Moreover, given a set of training data (such as gene expression measurements) consisting of N noisy observations, $\mathbf{X} = (x_{t_1}, x_{t_2}, \ldots, x_{t_N})^{\mathrm{T}} \in \mathbb{R}^N$, that are measured at input variables (such as time points), $\mathbf{T}$, each observation, x_t, can be modelled as

$$x_t = f(t) + \epsilon_t \ , \tag{4.16}$$

where $\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is an additive Gaussian noise term; and $f(t)$ is the true underlying model with a Gaussian process prior. Therefore, the likelihood can be represented as $\mathbf{X} \sim \mathcal{N}(f(\mathbf{T}), \sigma_\epsilon^2 \mathbf{I})$.

Given the GP prior and the training data, function values, $\mathbf{f}_*$, can be evaluated at new input variables, $\mathbf{T}_*$. In this context, the function values, $\mathbf{f}_* = f(\mathbf{T}_*)$, are referred to as the test data. In GP, the model that generates the training and test data is assumed to be a joint Gaussian distribution, which can be written as:

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{f}_* \end{pmatrix} = \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \mathbf{K}_{\mathbf{T},\mathbf{T}} + \sigma_\epsilon^2 \mathbf{I} & \mathbf{K}_{\mathbf{T},\mathbf{T}_*} \\ \mathbf{K}_{\mathbf{T}_*,\mathbf{T}} & \mathbf{K}_{\mathbf{T}_*,\mathbf{T}_*} \end{pmatrix} \right) \tag{4.17}$$

where $\mathbf{K}_{\mathbf{T},\mathbf{T}_*} = \mathbf{K}_{\mathbf{T}_*,\mathbf{T}}{}^{\mathrm{T}}$. Therefore, the distribution of the test data is modeled by conditioning the joint distribution on the training data and deriving the conditional distribution, $\mathbf{f}_* | \mathbf{X}$. It essentially forces the functions to go close to each training point. This conditional distribution is known as the predictive posterior distribution of the GP regression model and is formally derived as:

$$\mathbf{f}_* | \mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*), \tag{4.18}$$

where

$$\begin{aligned} \boldsymbol{\mu}_* &= \mathbf{K}_{\mathbf{T}_*,\mathbf{T}} \, (\mathbf{K}_{\mathbf{T},\mathbf{T}} + \sigma_\epsilon^2 \mathbf{I})^{-1} \, \mathbf{X} \\ \boldsymbol{\Sigma}_* &= \mathbf{K}_{\mathbf{T}_*,\mathbf{T}_*} - \mathbf{K}_{\mathbf{T}_*,\mathbf{T}} \, (\mathbf{K}_{\mathbf{T},\mathbf{T}} + \sigma_\epsilon^2 \mathbf{I})^{-1} \, \mathbf{K}_{\mathbf{T},\mathbf{T}_*} \end{aligned} \tag{4.19}$$

Finally, though GPs are considered non-parametric models, they usually include hyperparameters from the covariance function, such as $\{\sigma_{\mathrm{se}}^2, \ \ell_{\mathrm{se}}\}$ (Equation 4.15), and parameter(s) from the likelihood, such as σ_ϵ^2 (Equation 4.16), which need to be estimated. One way is to obtain type II maximum likelihood (ML-II) estimates (i.e. point estimates) by maximizing the analytically tractable marginal likelihood that marginalizes over the function values $\mathbf{f}$:

$$p(\mathbf{X} | \mathbf{T}, \boldsymbol{\theta}) = \int p(\mathbf{X} | \mathbf{f}, \mathbf{T}, \boldsymbol{\theta}) p(\mathbf{f} | \mathbf{T}, \boldsymbol{\theta}) d\mathbf{f} \, , \tag{4.20}$$

where $p(\mathbf{X}|\mathbf{f},\mathbf{T},\boldsymbol{\theta})$ is the likelihood and $p(\mathbf{f}|\mathbf{T},\boldsymbol{\theta})$ is the GP prior stated earlier, and $\boldsymbol{\theta} = \{\sigma^2_{\text{se}},\ \ell_{\text{se}},\ \sigma^2_\epsilon\}$. Alternatively, instead of finding point estimates for $\boldsymbol{\theta}$, one can also characterize the posterior distribution of $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\mathbf{X})$, which requires defining the priors for $\boldsymbol{\theta}$ also. By characterizing the posterior of $\boldsymbol{\theta}$, the posterior predictive distribution can be computed by marginalizing over $\boldsymbol{\theta}$:

$$p(\mathbf{f}_*|\mathbf{X}) = \int p(\mathbf{f}_*|\mathbf{X},\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{X})d\boldsymbol{\theta}\ , \tag{4.21}$$

However, the integral in Equation 4.21 does not have a closed-form solution and estimating the posterior distribution of $\boldsymbol{\theta}$ is challenging. Therefore, the posterior and predictive distributions in Equation 4.21 need to be approximated using methods such as central composite design (CCD) [Rue *et al.*, 2009; Vanhatalo *et al.*, 2010], or Markov chain Monte Carlo (MCMC) methods can be used to sample from the posterior distribution Timonen *et al.* [2021]. CCD estimates the posterior distribution of $\boldsymbol{\theta}$ and performs numerical integration approximation.

4.6 Cell type identification and trajectory inference of single-cells

Single-cell RNA-sequencing (scRNA-seq) data (discussed in Section 3.2.3) consists of high-throughput gene expression profiles from thousands of individual cells that are commonly used to identify the cell types present in a sample and to understand the dynamic cellular states associated with biological processes, such as differentiation [Stegle *et al.*, 2015; Wagner *et al.*, 2016; Bacher and Kendziorski, 2016].

Traditionally, cells have been classified into specific cell types based on their morphology, physiology, and marker gene expression [Wagner *et al.*, 2016]. However, at present, the concept of a ‘cell type’ is poorly defined (at least in humans) [Kolodziejczyk *et al.*, 2015], as the current catalog of human cell types include those types that can be further sub-categorized by functional differences as well as unique gene expression profiles, such as muscle cells [Trapnell, 2015]. Moreover, many cell types, especially rare and novel cell types, may have insufficient (if any) number of reliable marker genes that can be used for cell-typing [Wagner *et al.*, 2016; Trapnell, 2015]. In this scenario, scRNA-seq data provides a crucial advantage, as it can be used to identify cell types in an unbiased and systematic manner, without relying on an incomplete catalogue of marker genes [Wagner *et al.*, 2016; Kolodziejczyk *et al.*, 2015]. Typically, unsupervised clustering (Section 4.3) and/or dimensionality reduction methods (Section 4.2) are applied on the scRNA-seq gene expression data to group the cells by transcriptomic similarity [Wagner *et al.*, 2016; Kolodziejczyk *et al.*, 2015; Liu and Trapnell, 2016; Poirion *et al.*, 2016]. Several methods, including Seurat [Macosko *et al.*, 2015] (Section 4.6.1), perform linear or non-linear dimensionality reduction on the data prior to clustering in order to avoid the challenge of clustering in high-dimensional spaces. On the other hand, some methods either refrain from or perform solely dimensionality reduction to identify the cell types [Wagner *et al.*, 2016; Kolodziejczyk *et al.*, 2015]. Moreover, either all of the expressed genes or a subset of the genes, such as highly variable genes (HVGs), are used for dimension reduction and/or clustering [Stegle *et al.*, 2015; Kolodziejczyk *et al.*, 2015; Bacher and Kendziorski, 2016]. Finally, after grouping the cells, one main objective is to characterize the clusters of cells (i.e. cell types) by, for instance, identifying marker genes that best discriminate the different clusters [Stegle *et al.*, 2015; Kolodziejczyk *et al.*, 2015; Poirion *et al.*, 2016]. One of the most commonly used ways to do this is to perform differential expression analysis (DEA) between pairs of clusters [Stegle *et al.*, 2015; Poirion *et al.*, 2016]. While, the approaches for bulk RNA-seq DEA, such as edgeR [Robinson *et al.*, 2010] (Section 4.4.3), can be employed here [Bacher and Kendziorski, 2016; Stegle *et al.*, 2015], scRNA-seq data is generally more noisy and contains many zeros. Therefore, methods that account for the bimodality in the data have been developed for scRNA-seq DEA, such as zero-inflated models (i.e. mixture-model-based approaches) [Bacher and Kendziorski, 2016; Poirion *et al.*, 2016; Finak *et al.*, 2015; Kharchenko *et al.*, 2014].

Another important aim of scRNA-seq studies is to study the dynamic transitions that cells undergo as a response to biological processes, such as cell differentiation. Due to the unsynchronized nature of single-cells, a population of cells studied at any given timepoint is likely to contain cells at different stages of the biological process [Wagner *et al.*, 2016; Bacher and Kendziorski, 2016]. Therefore, most single-cell datasets provide a snapshot of the entire biological process under study [Trapnell *et al.*, 2014; Wagner *et al.*, 2016]. Also, since each biological process is typically reflected in the cell's molecular (i.e. RNA or protein) profile, scRNA-seq data can be used to position or order the cells along a (pseudo)temporal trajectory of the corresponding biological process [Wagner *et al.*, 2016]. More than 70 single-cell trajectory inference tools have been proposed over the years [Saelens *et al.*, 2019], including Monocle [Trapnell *et al.*, 2014], Monocle 2 [Qiu *et al.*, 2017], Wanderlust [Bendall *et al.*, 2014], Wishbone [Setty *et al.*, 2016], and SCUBA [Marco *et al.*, 2014]. Most of these tools vary substantially in terms of their algorithms [Saelens *et al.*, 2019].

Monocle was the first method that introduced the strategy of computationally inferring the trajectories of single-cells from scRNA-seq data, and ordering the single-cell expression profiles in 'pseudotime' [Trapnell *et al.*, 2014]. Here, 'pseudotime' is an abstract unit of the progress of an individual cell along the trajectory of a biological process; it is simply the shortest distance between a cell and the start of the trajectory [Trapnell *et al.*, 2014].

This section will also briefly discuss the Seurat pipeline (version 2) and Monocle 2 as they have been employed for scRNA-seq data analysis in Publication I (discussed in Chapter 5).

4.6.1 Seurat (version 2) Pipeline

Seurat [Macosko *et al.*, 2015] is an R package that has been developed for performing several single cell data analysis steps, including quality control (QC), normalization, dimension reduction, clustering and marker gene identification. As an input, Seurat takes a count matrix, for instance, the UMI count matrix generated by the Cell Ranger Single-Cell Software Suite pipeline that has undergone the first tier of QC (as explained in Section 3.2.3), read alignment and gene level quantification [Zheng *et al.*, 2017]. Subsequently, Seurat performs the second tier of QC analysis, where specific QC metrics are used to identify and discard the cells that are empty, duplets/multiplets, of low-quality, and/or contain degraded mRNA. These QC metrics include the total number of genes expressed in a cell (i.e. gene count), total number of unique UMIs detected in a cell (i.e. UMI count), and the proportion of reads mapping to mitochondrial genome. The filtered data is then normalized using a simple global-scaling normalization (as explained in Section 3.2.3) and variations from uninteresting sources, such as the percentage of mitochondrial genes expressed per cell and the UMI count per cell, are also regressed out.

Prior to performing dimension reduction, Seurat identifies a set of genes that are highly variable across all cells (HVGs) in order to capture the heterogeneity of the single cell data. The HVGs are determined by computing the average expression and dispersion for each gene, placing the genes into bins based on their average expressions, and computing a z-score for the dispersion values of all genes within each bin. Genes with an average expression and z-normalized dispersion value above certain threshold are then identified as the HVGs. Then, PCA (Section 4.2) is performed on the HVGs and DBSCAN clustering (Section 4.3) is performed on the top PCs. Finally, to determine the cell type represented by each cluster of cells, the defining marker genes are identified using differential expression analysis (as implemented in McDavid *et al.* [2013]), and by comparing cells of a single cluster to the cells of all other clusters combined. A gene is considered a marker of a cluster if it is expressed in at least a certain proportion of the cells of the cluster and with a minimum log fold change.

Additionally, Seurat visualizes the single-cell data by performing t-SNE (Section 4.2) on the top PCs that were used for clustering.

4.6.2 Monocle 2

Monocle 2 (implemented in an R package) [Qiu *et al.*, 2017] includes algorithms for reconstructing the cellular trajectories that take place during dynamic biological processes, such as cell differentiation, among other algorithms. In this thesis (specifically in Publication I), Monocle 2 was used for only the trajectory inference of single-cells. An advantage of Monocle 2 is that it does not require any *a priori* knowledge, for instance, about the marker genes that characterize the biological processes or the number of branch points in the trajectory.

The single-cell trajectory analysis using Monocle 2 has three main steps. The first step involves feature selection, where a set of genes that capture the variation between the cells are chosen for understanding the shape of the trajectory. This step can be performed in a completely unsupervised manner or using a set of known genes that define the biological process that Monocle augments with related genes (i.e. 'semi-supervised'). The unsupervised method, called 'dpFeature' in Monocle, starts by performing PCA on the genes that are expressed in a minimum percentage of cells. Then it applies t-SNE dimension reduction on the top PCs and performs density-peak clustering (Section 4.3) to identify the clusters on the 2-dimensional t-SNE space. Finally, a differential expression analysis is performed to identify the genes that differ between the clusters and the top 1000 significant genes are selected for trajectory reconstruction.

In the second step, Monocle 2 uses a machine learning technique, called reverse graph embedding (RGE), to simultaneously perform dimensionality reduction on the high-dimensional data and learn a principal graph (or tree) on the population of cells that describes how cells transition from one state to another (i.e. trajectory). Monocle 2 uses DDRTree [Mao *et al.*, 2015], which is a scalable implementation of the RGE framework. Finally, in the third step, the cells are ordered along the trajectory by performing manifold learning on the tree from step 2.

5. Publication I: *IL32* gene expression - a novel signature for early detection of β-cell autoimmunity

As discussed in Sections 2.4.1 and 2.4.2, predicting the progression of an autoimmune disease, such as T1D, remains elusive. Despite the ability of conferring genetic susceptibility to T1D using HLA genes, when (or whether) the onset of autoimmunity (i.e. appearance of autoantibodies or seroconversion) or clinical diagnosis of T1D happens is unpredictable as of now. The only predictive biomarkers that exist in T1D are autoantibodies, which are usually identified only after self-tolerance is already broken. Therefore, new biomarkers are needed that can predict seroconversion or indicate progressive β-cell destruction, providing a window of opportunity for therapeutic interventions aimed at preventing the disease progression. On that premise, the aim of this study was to analyse the gene expression profiles of specific immune cells collected from T1D susceptible children during the first three years of life in order to: 1) identify early gene expression markers that may help predict the onset of autoimmunity and/or reflect upon progression of the disease, and 2) determine the specific cell types that may be expressing the gene expression markers. This chapter will present some of the main findings of this study.

5.1 Study Design

Participants for this study were selected from the DIABIMMUNE project's birth cohort [Peet *et al.*, 2012]. Briefly, the DIABIMMUNE project is an international collaboration that was initiated to test the hygiene hypothesis (explained in Section 2.3.2) in the development of T1D. They have collected longitudinal blood and stool samples as well as extensive metadata from hundreds of T1D susceptible individuals from Finland, Estonia and Karelian Republic of Russia (i.e. Russian Karelia). These three countries (or regions) represent a unique 'living laboratory' for such testing as there exists one of the steepest welfare gradients worldwide between Finland and Russian Karelia, (where Estonia represents a country in rapid transition); and the incidence rate of T1D is much higher in Finland than the other two countries (discussed in Section 2.4.2) despite the fact that the populations share similar genetics. In the birth cohort, the participants were sampled from birth till 3 years of age.

From this cohort, seven cases and seven healthy ‘matched’ controls were selected to be part of this study (totalling fourteen participants; except one case who had two control individuals due to insufficient number of samples). All the case individuals had developed T1D-associated autoantibodies (i.e. seroconverted, details in 2.4.2) before turning 2 years of age; and a non-seroconverted (i.e. healthy) individual from the cohort was paired with each case sample as a matched control based on the same date and place of birth, gender and HLA-conferred genetic risk category. Blood samples from each individual were collected (as part of the DIABIMMUNE project) at 3, 6, 12, 18, 24 and 36 months of age.

First, the peripheral blood mononuclear cells (PBMCs) from all the blood samples of all individuals were fractionated (i.e. separating cells) into CD4-enriched (CD4+), CD8-enriched (CD8+), and CD4- and CD8-depleted (CD4-CD8-) cells. Fractionation enables identification of cell type-specific gene expression profiles, which may be masked when analysing PBMCs due to the varying compositions of different cell types in the blood. Moreover, as established in Sections 2.2.3 - 2.3, CD4+ and CD8+ cells play a crucial role in mediating the adaptive immune response and in facilitating autoimmunity. Therefore, studying their gene expression profiles may provide novel insights into the cellular mechanisms that may lead to T1D pathogenesis. Altogether, 306 samples were included in this study.

Subsequently, bulk RNA-sequencing (using Illumina HiSeq 2500 instrument and general pipelines explained in Sections 3.1.2 and 3.2.2) was performed on the mRNA extracted from the cells of each fraction of each sample, including the PBMC fractions. Additionally, single-cell RNA-sequencing (using 10X Genomics Chromium method, introduced in 3.2.3) was performed on eight out of 78 PBMC samples (four case samples and their four closest control samples) that had high (or low) expressions of the interleukin 32 (*IL32*) or insulin (*INS*) genes.

5.2 Bulk RNA-seq results

The raw sequencing reads from bulk RNA-seq were processed according to the steps mentioned in Section 3.2.2 using: FastQC for quality checking, Tophat2 for alignment, and HTSeq-count for quantification of gene expression levels. Subsequently, the genes were divided into coding and non-coding categories, and the genes from each category were filtered based on selected RPKM value thresholds (>3 and >0.5 for coding and non-coding genes, respectively) to remove lowly expressed genes. TMM normalization and DE analysis (DEA) were conducted separately on the

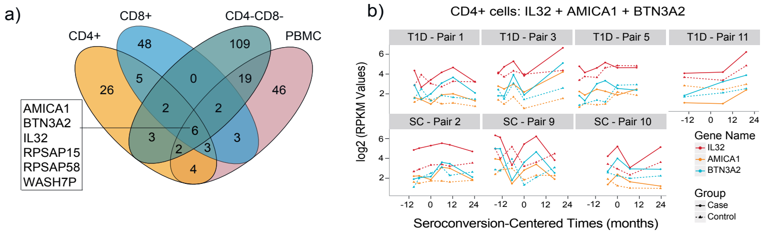


Figure 5.1. a) Number of DEGs identified between cases and matched controls over all time points in each fraction as well as the overlaps between fractions. b) Concerted gene expression profiles (presented as $\log_2$ RPKM values) of *IL32*, *AMICA1* and *BTN3A2* in CD4+ samples.

raw read counts of the filtered coding and non-coding genes. Finally, differentially expressed genes (DEGs) were identified using a set of post-filtering criteria, including false discovery rate (FDR) and median $\log_2$ fold change (FC) thresholds. Details of each step can be found in the Supplementary Data of Publication I.

In one of the main analyses of this study, all case samples (over all timepoints) were compared to their age-matched control samples using GLMs (with trended dispersions) as implemented in edgeR (Section 4.4.3), in order to identify genes that are DE between cases and controls regardless of the sampling ages. This analysis identified 51, 69, 143, and 85 (coding and non-coding) genes to be DE in CD4+, CD8+, CD4-CD8- and PBMC fractions, respectively (Figure 5.1a). While most of the DEGs were unique to specific fractions, six genes were found to be differentially upregulated in the case samples of all four fractions. Three of the genes were pseudogenes with unknown functions, whereas the other three genes, namely *IL32*, *BTN3A2* and *AMICA1*, have previously been associated with autoimmune diseases (ADs); *BTN3A2* has been associated with T1D. *IL32* is a gene that encodes a proinflammatory cytokine (Section 2.2.2) and whose overexpression has been observed in ADs, such as rheumatoid arthritis (RA) and inflammatory bowel disease (IBD). However, it has not been associated with T1D until now. In fact, these three genes (along with a few other DEGs that are highlighted in red in Figure 2B of Publication I) were found to be co-regulated in most of the fractions upon clustering using the k-means method with silhouette scoring for determining the value of k (Section 4.3) as well as an Euclidean distance-based co-clustering selection criteria (Figure 5.1b, details of cluster analysis can be found in Supplementary Data of Publication I). Additionally, transcription factor binding site (TFBS) analysis using TRANSFAC database identified the TFBS of Ikaros (IKZF1), a T1D-associated TF, to be enriched on the promoters of *IL32* and its co-regulated genes in CD4+ and PBMC fractions.

IL32 was also found to be differentially upregulated in the case samples collected in the 12 months window before seroconversion, suggesting its increased expression to be a critical immunological signature in infants progressing towards β-cell autoimmunity. This upregulation was identified in all four fractions using differential expression analysis and was validated using qRT-PCR (in PBMC fraction).

In order to identify the specific cell sub-populations from which *IL32* gene expression signature originated, scRNA-seq analysis was conducted.

5.3 Single-cell RNA-seq results

The eight scRNA-seq samples were pre-processed individually using the Cell Ranger Single-Cell Software Suite. Specifically, this tool was used to perform QC analysis of the raw sequencing reads, alignment of the reads to human reference genome using STAR, barcode pre-processing and UMI counting (as explained in Section 3.2.3). To identify rare cell types, the cells from the eight samples were pooled together using Cell Ranger's multi-library aggregation algorithm where the samples were normalized using subsampling normalization, retaining ~31,000 confidently mapped reads per cell that mapped to a median of 801 genes per cell. Altogether, expression of ~33,000 genes from ~20,000 cells was obtained after pooling.

Subsequently, Seurat (version 2) pipeline was used to perform further QC filtering steps (retaining ~18k cells expressing ~20k genes for downstream analyses), normalization, clustering and cluster-specific marker gene identification, as explained in Section 4.6.1 (details about parameter choices can be found in Supplementary Data of Publication I). Clustering of the cells identified thirteen clusters, where two clusters—both representing cells from naive T cells with no perceivable biological differences—were later merged into a single cluster, reducing the number of clusters to twelve. As shown in Figure 5.2a, these clusters represented various sub-populations of immune cells, including sub-populations of CD4+ and CD8+ T cells, NK cells (Section 2.1), B cells (Section 2.2.5), monocytes/dendritic cells (DCs) (Section 2.2.4), naive and developing T cells. Cells from the control samples dominated the naive T cell cluster; whereas monocytes/DCs cluster was dominated by cells from case samples (Supplementary Figure S10B in Publication I). Congruent with the bulk RNA-seq results, the *IL32* gene was found to be overexpressed in the cells from case samples (Supplementary Figure S11 in Publication I). As seen in Figure 5.2b, *IL32* was highly expressed by T cells, specifically activated (and proliferating) GZMA+ CD8+ T cells, as well as NK cells. This result is in concordance with previous observations of *IL32* expression in immune cells [Steinke and Borish, 2006].

After clustering, the QC filtered cells from the Seurat analysis were

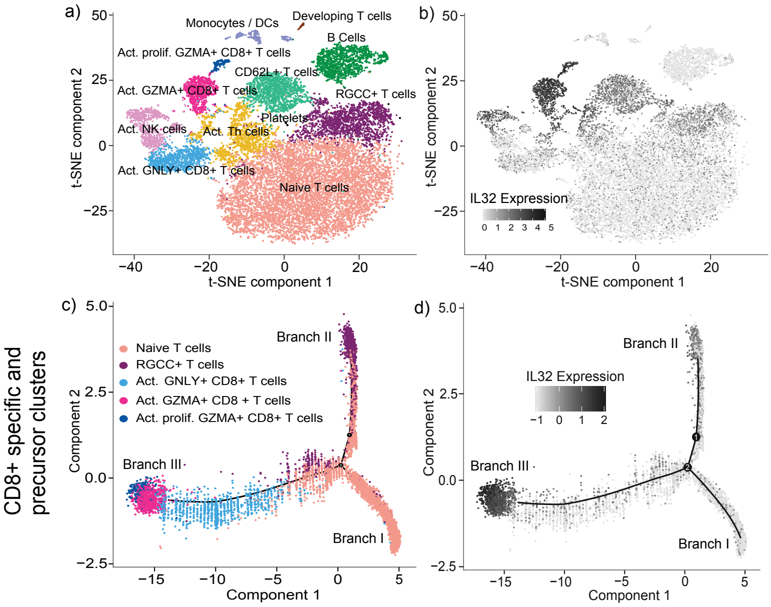


Figure 5.2. a) t-SNE visualization of the single cell clusters (~18k cells from eight samples) identified using the Seurat pipeline, where the colours correspond to specific sub-populations of immune cells determined using cluster-specific markers. b) and d) *IL32* expression levels of each cell. c) Pseudotemporal ordering of CD8+ T cells and precursor cells (naive and RGCC+ T cells), where the colours correspond to specific cluster classifications from Seurat results.

ordered in pseudotime using Monocle 2 (pipeline explained in Section 4.6.2; details about parameter or specific method choices can be found in Supplementary Data of Publication I). Specifically, separate pseudotemporal trajectory analyses were performed on the cells from CD4+ and CD8+ T cell sub-populations (Figure 3D and 3G in Publication I), where naive and RGCC+ T cell clusters (denoted ‘precursor cells’ in Figure 5.2c) were also included in each analysis to represent the less activated or differentiated immune cells. This analysis revealed the CD8+ T cells expressing the highest levels of *IL32* to be in the more advanced stages of the differentiation process (Figures 5.2c and 5.2d).

Details on the parameters chosen for each step of the Seurat and Monocle 2 analyses can be found in the Supplementary Data of Publication I.

6. Publication II: A personalised approach for identifying disease-relevant pathways

With the decreasing costs of high-throughput technologies (as discussed in Section 3.1), numerous time-course gene expression datasets are being routinely generated for studying the molecular mechanisms underlying the pathogenesis of various complex diseases. However, hardly any methods have been developed that can appropriately model longitudinal data as well as account for the heterogeneity that entails most complex diseases. Therefore, the aim of this publication is to introduce a new method that can robustly model time-course gene expression data from heterogeneous diseases in a personalised manner and identify disease-relevant pathways that can aid in predicting critical events in the disease progression and perhaps even in identifying biomarkers.

6.1 The personalised approach

This section provides a broad overview of the personalised approach, as illustrated in Figure 6.1. The full description can be found in the 'Methods' of Publication II.

6.1.1 Step 1: Identifying DEGs in a personalised manner using Gaussian processes

One of the most prominent goals of gene expression studies (i.e. transcriptomics studies) is to identify differentially expressed genes (DEGs) between a case (e.g. disease) and a control (e.g. healthy) group. Several methods have been developed for modelling time-course data for identifying DEGs. Some of these methods account for the dynamic nature of the data, whereas others disregard it (as elucidated in the 'Introduction' of Publication II). Recently, Gaussian processes (GPs) (discussed in Section 4.5) have gained popularity for modelling time-course data due to their capabilities of capturing the true underlying signal and embedded noise in a non-linear manner [Rasmussen and Williams, 2006]. They have been

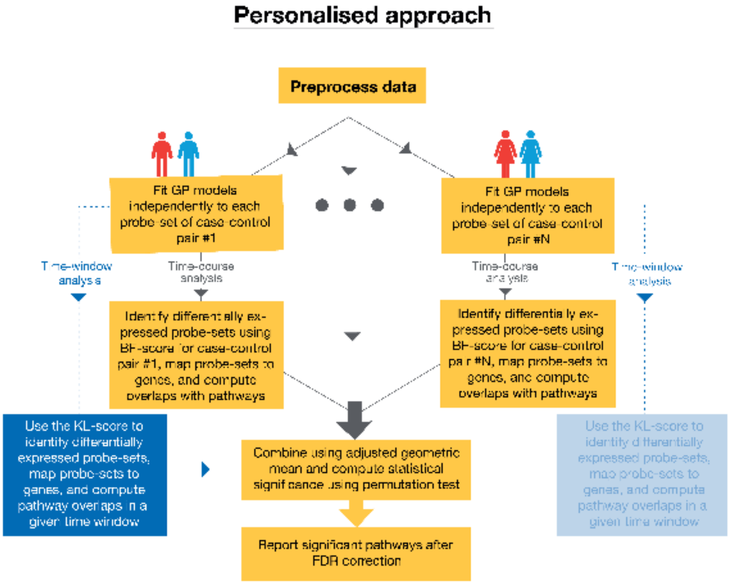

Personalised approach
Preprocess data
Fit GP models independently to each probe-set of case-control pair #1
Fit GP models independently to each probe-set of case-control pair #N
Time-window analysis
Time-course analysis
Time-course analysis
Time-window analysis
Identify differentially expressed probe-sets using BF-score for case-control pair #1, map probe-sets to genes, and compute overlaps with pathways
Identify differentially expressed probe-sets using BF-score for case-control pair #N, map probe-sets to genes, and compute overlaps with pathways
Use the KL-score to identify differentially expressed probe-sets, map probe-sets to genes, and compute pathway overlaps in a given time window
Combine using adjusted geometric mean and compute statistical significance using permutation test
Use the KL-score to identify differentially expressed probe-sets, map probe-sets to genes, and compute pathway overlaps in a given time window
Report significant pathways after FDR correction

Figure 6.1. A schematic outline of the personalised approach for identifying DEGs and significant pathways

applied for identifying DEGs across the whole time-course [Äijö *et al.*, 2012; Cheng *et al.*, 2019] as well as between specific time-windows containing few or no observations [Stegle *et al.*, 2010; Heinonen *et al.*, 2015; Yang *et al.*, 2016]. However, most of these methods model all samples in a group (such as cases and controls) together to detect genes that exhibit differential expression across all or most case individuals in the study population (referred to as 'the combined method' in Publication II). This is an unrealistic proposition in the case of heterogeneous diseases, where different genes with similar functionalities may be perturbed across different case individuals. Therefore, by assuming a study design, where each case individual is matched with a control individual, the personalised approach presented in Publication II identifies DEGs for each case-control pair separately (i.e. in a personalised manner) in a direction-agnostic manner using a robust and efficient method involving Gaussian processes to model the time-course data. This method can be used to detect DEGs over the whole time-course as well as in specific time-windows, as explained below and in Figure 1a of Publication II.

Time-course analysis

For identifying the features (e.g. probe-sets or genes) that are differentially expressed (DE) in each case as compared to its matched control across the whole time-course, the personalised approach models the expression data from each feature individually. Essentially, two models are fit per feature, namely the *joint* and *separate* models. In the *joint* model, a single Gaussian process (GP) regression (Section 4.5), is fit to all the time-course data points of a particular feature from the case-control pair; whereas in the *separate* model, two GP regressions are fit to the data points from the cases and controls separately. Subsequent to model fitting, a model selection step is conducted to quantify the fit of each model and thus assess whether the case and control expressions come from the same process (i.e. feature is not DE, null hypothesis) or different processes (i.e. feature is DE, alternative hypothesis). The model selection step for the time-course analysis is done by calculating the log ratio of the marginal likelihoods of the *joint* and *separate* models, which corresponds to a Bayes factor score (BF-score) [Kass and Raftery, 1995]

$$\mathrm{BF-score} = \log \frac{p(\mathbf{x^A}|M^A)p(\mathbf{x^B}|M^B)}{p(\mathbf{x^S}|M^S)}, \tag{6.1}$$

where $\mathbf{x^A}$ and $\mathbf{x^B}$ correspond to the time-course data points from cases and controls, respectively; $\mathbf{x^S}$ is the pooled data points from $\mathbf{x^A}$ and $\mathbf{x^B}$; M^A and M^B are the models fit to $\mathbf{x^A}$ and $\mathbf{x^B}$ separately in the *separate* model; and M^S is the *joint* model fit to $\mathbf{x^S}$. In this study, if a feature had a BF-score > 4, it was considered to be differentially expressed in the case.

Time-window analysis

As mentioned earlier, the personalised approach can also be used to identify DE features within specific time-windows of any chosen size. Similar to the time-course analysis explained above, the time-window analysis also begins with fitting a *joint* and *separate* model for each feature from each case-control pair. Subsequently, predictions at specific intervals in the time-window are made from the predictive posterior distributions of the fit GP regressions. In order to compare the predictions from the *separate* and *joint* models, the predictions from each model are assumed to follow multivariate Gaussian distributions (one for each model). Then the two distributions are compared using the Kullback-Leibler (KL) divergence (similar to Equation 4.4):

$$\mathrm{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx, \tag{6.2}$$

where $p(x)$ and $q(x)$ are the corresponding distributions. Specifically, a continuous KL-score is computed using the symmetric KL divergence:

$$\frac{1}{2}\mathrm{KL}(P||Q) + \frac{1}{2}\mathrm{KL}(Q||P). \tag{6.3}$$

In this study, a feature with KL-score > 250 in a specific time-window was considered to be differentially expressed in that window.

6.1.2 Step 2: Summarising DEG lists on a pathway-level

Since gene-level results in similar studies of the heterogeneous diseases show alarmingly little overlap and are often inconsistent [Menche *et al.*, 2017; Segal *et al.*, 2004; Drier *et al.*, 2013; Jin *et al.*, 2014; Subramanian *et al.*, 2005; Chen *et al.*, 2013], several methods have been developed to summarize the gene-level results on a pathway-level [Subramanian *et al.*, 2005; Segal *et al.*, 2004; Drier *et al.*, 2013; Lee *et al.*, 2008; Vaske *et al.*, 2010; Chen *et al.*, 2008]. Therefore, in the personalised approach, the DEG lists obtained from Step 1 (Section 6.1.1) are summarized on a pathway-level using a permutation-based empirical hypothesis testing that is customized for personalised DE analysis. The definition of a pathway given in Section 3.2.2 applies here.

Specifically, an overall enrichment score was defined for each pathway from the MSigDb [Subramanian *et al.*, 2005] by first computing a scaled pathway overlap statistic, $f_{i,j}$, for each pathway i, and each case-control pair j as

$$f_{i,j} = \frac{\text{overlap}_{i,j}}{\text{diff. exp. genes}_j} + \alpha, \tag{6.4}$$

where $\text{overlap}_{i,j}$ denotes the number of DEGs from the j^{th} case-control pair that overlaps with the genes in the i^{th} pathway; $\text{diff. exp. genes}_j$ refers to the total number of genes detected as DE in the j^{th} case-control pair; and α is a small constant (10^{-6} by default). Then, the enrichment score for each pathway i was defined as the geometric mean of the scaled pathway overlaps (Equation (6.4)) across all case-control pairs (Equation (15) in Publication II). Finally, the statistically enriched pathways were identified by performing a permutation test and empirically computing the p-values for each pathway.

6.2 Data

The personalised approach developed in this study was applied to three type 1 diabetes (T1D) time-course gene expression microarray datasets (microarrays introduced in Section 3.1.1). Two of these datasets (*Datasets 1* and *2*) were published by Kallionpää *et al.* [2014] (generated using Affymetrix U219 arrays), whereas the third dataset (*Dataset* 3) was published by Ferreira *et al.* [2014] (generated using Affymetrix Human Gene 1.1 ST arrays). *Datasets 1* and *2* comprised of six and 15 case-control pairs, respectively, where each case individual was matched with a healthy (but T1D susceptible) control individual based on date and place of birth, gender and HLA risk class. As *Dataset 1* was sampled before and after seroconversion of the case, it was used to identify disrupted pathways during the early progression of T1D (time-course (TC) analysis) as well as in the

6 months window prior to seroconversion (window before seroconversion (WSC) analysis). The sampling of individuals in *Dataset 2* started after seroconversion and continued till T1D diagnosis of the case, so it was used to understand pathway-level disruptions in the 6 months window prior to clinical diagnosis of T1D (window before T1D diagnosis (WT1D) analysis). Meanwhile, *Dataset 3* comprised of 9 case-control pairs[1] that were sampled before and after seroconversion of the case. TC analysis was performed on *Dataset 3* to assess the generalisability of the results obtained using the personalised approach.

All raw microarray samples from the three datasets were pre-processed using RMA normalization technique (discussed in Section 3.2.1) prior to any analysis. More details on the data can be found in the 'Data' section and Supplementary Notes of Publication II.

6.3 Results

For comparative purposes, TC, WSC and WT1D analyses using *Datasets 1* and *2* were performed using the personalised approach (briefly explained in Section 6.1) and the combined method (introduced in Section 6.1.1). Additionally, the pathway-level results from the personalised and combined methods were also compared with the equivalent results from Kallionpää *et al.* [2014] to establish the biological and disease-specific relevance of the achieved results.

In general, the personalised approach was able to identify several immunological and disease-relevant pathways in the time-course as well as window-analyses than its non-personalised counterparts (the combined method and the rank-product method used by Kallionpää *et al.* [2014]), thus revealing more insight into the intrinsic mechanisms involved in the progression of disease.

At gene-level, the personalised approach generally identified many more DEGs per case-control pair than the combined method. On an average, only 14% of the DEGs overlapped between different case-control pairs, demonstrating the heterogeneity among case-control pairs. A closer study of the 'T1D pathway' revealed that only a subset of the pathway's genes are found DE in each case-control pair and this subset varied between pairs unpredictably (Figure 4 of Publication II). However, when the genes of the T1D pathway were divided into sub-processes, it was observed that usually at least one gene per sub-process was detected as DE in each pair. Other pathways might also follow a similar phenomenon of regulation.

[1] Pairing was performed as part of this study based on time of birth, gender and sampling ages

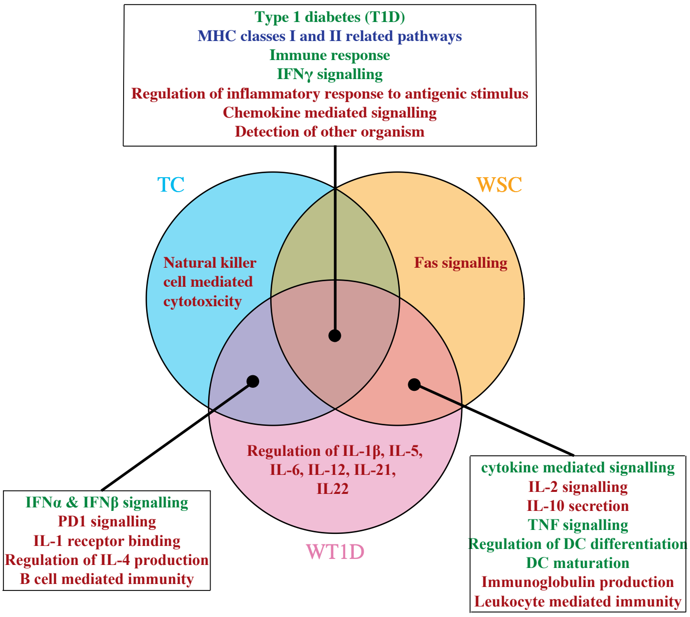

Type 1 diabetes (T1D)
MHC classes I and II related pathways
Immune response
IFNγ signalling
Regulation of inflammatory response to antigenic stimulus
Chemokine mediated signalling
Detection of other organism
TC
WSC
Natural killer cell mediated cytotoxicity
Fas signalling
Regulation of IL-1β, IL-5, IL-6, IL-12, IL-21, IL22
WT1D
IFNα & IFNβ signalling
PD1 signalling
IL-1 receptor binding
Regulation of IL-4 production
B cell mediated immunity
cytokine mediated signalling
IL-2 signalling
IL-10 secretion
TNF signalling
Regulation of DC differentiation
DC maturation
Immunoglobulin production
Leukocyte mediated immunity

Figure 6.2. A Venn diagram summarizing the most disease-relevant pathways identified by the personalised approach that are specific to certain analyses or overlapping between analyses. Blue text: found enriched by personalised and both non-personalised approaches; green text: found enriched by personalised approach and Kallionpää *et al.* [2014]; and red text: found enriched only by the personalised approach

6.3.1 Enriched pathways

The most disease- and immunologically-relevant pathways identified by the personalised approach in the three analyses are summarized in Figure 6.2, where the colour of the text indicates whether the pathway was also found to be enriched in at least one of the analyses using the non-personalised method(s).

Due to the conceptual stringency by which the combined method identifies DEGs, it detected far fewer number of enriched pathways than the other methods as seen in Table 1 of Publication II. In fact, a majority of the disease-relevant pathways that were identified as enriched by the personalised approach, were not identified by the combined method in any of its analyses, including the basic pathways related to immune response and 'T1D pathway'. Among the most disease-relevant pathways, it identified only a few pathways involved in the initiation of an immune response, i.e. those related to MHC class I and II molecules and its functions 2.2.1, in its

TC analysis.

Comparatively, Kallionpää *et al.* [2014] identified many of the significant pathways identified by the personalised approach. However, they detected mostly the overarching pathways as enriched, whereas the personalised approach identified more specialised pathways as well. For instance, while both of these methods identify the 'cytokine mediated signalling' pathway as enriched in at least one of the analyses, the personalised approach also identified pathways related specific interleukins. Moreover, Kallionpää *et al.* [2014] detected significance of certain pathways in different analyses as compared to the personalised approach. A comprehensive comparison of the results from the personalised approach and Kallionpää *et al.* [2014] can be found in the 'Results' of Publication II; and the relevance of the pathway-level results (in terms of immunology and T1D pathogenesis) obtained from all three analyses using the personalised approach is discussed in detail in 'Discussion' of the Publication II.

Briefly, the personalised approach identified the significance of the 'T1D pathway' in all three analyses along with several crucial pathways involving many of the key players of an active immune response, such as cytokines (Section 2.2.2), dendritic cells (DCs) (Section 2.2.4), and B cells (Section 2.2.5), to name a few. Among the influential cytokine pathways that were found to be enriched in the case individuals were the pathways related to signalling of: IFN-γ, which is produced by self-reactive CD4+ and CD8+ T cells and plays a vital role in driving the pathogenesis of T1D; IFN-α, which is a known initiator of T1D pathogenesis; and IL-2, which is secreted by CD4+ T cells to co-stimulate a variety of immune cells, including CD8+ T cells (explained in Section 2.2.4). Additionally, the PD-1 signalling pathway, which is known to promote self-tolerance (Section 2.2.3) and has been proposed as a target for a novel therapy for preventing autoimmunity, was found to be disrupted in the early stages of the disease. The Fas signalling pathway, which is one of the pathways by which CD8+ T cells kill target cells, was found to be uniquely enriched before seroconversion, indicating that β-cell destruction may be observed much before clinical onset of T1D. In fact, even the immunoglobulin (i.e. antibody) production pathway is found to be enriched in cases before they present autoantibodies. Many more interesting and disease-relevant pathways revealed to be disrupted by the personalised approach at different stages of disease progression can be found in Publication II.

6.4 Generalizability of the results & Robustness of the method

The generalizability of the pathway-level results obtained using the personalised approach was assessed using the Spearman's rank correlation tests on the FDR values of the pathways from the TC analyses of two independent T1D datasets, namely *Datasets 1* and *3*. The correlation tests were performed using all the pathways as well as a subset of 32 disease-relevant pathways. Both tests revealed the results from the two TC analyses to be highly correlated, thus demonstrating the generalizability of the pathway-level results using the personalised approach. Further details can be found in the ‘Methods’ section of Publication II.

To demonstrate the robustness of the personalised approach in terms of efficiently modelling time-course data and estimating unobserved values, a leave-one-out cross-validation analysis was performed. Furthermore, to demonstrate the robustness of the approach to noise in the data, additional noise was added to *Dataset 1* and the results were compared to the original results using correlation tests. These analyses confirmed the robustness of the personalised approach. A detailed account of these analyses can be found in the Supplementary Methods of Publication II.

7. Publication III: Differences in the gut microbial architectures of IgG4-RD and SSc patients compared to healthy controls

As introduced in Section 2.4.3, immunoglobulin G4-related disease (IgG4-RD) and systemic sclerosis (SSc) are two rare fibroinflammatory systemic autoimmune diseases with no established etiology or pathogenesis. Both diseases have been associated with genetic variations in the HLA genes and are characterized by similar immunological characteristics, such as the presence of an unusual subset of cytotoxic CD4+ T cells (CD4+ CTLs). These characteristics of the diseases together with the suggestion by recent studies that they may be driven by environmentally-sourced antigens, such as microbial antigens, indicate that these diseases may stem from the dysfunctional immune recognition of microbial signal. Analogous to other autoimmune diseases, the disease-triggering or -sustaining microbial signals could emerge from a dysbiotic gut microbiome (as explained in Section 2.3.2). Therefore, the aim of this study was to characterize the compositions and functional capabilities of the gut microbiomes of IgG4-RD and SSc patients along with healthy individuals in order to identify potential sources of microbial signals that might be contributing to the etiology of the diseases.

7.1 Study design and data processing

Stool samples were collected (one sample per individual) from 58 IgG4-RD and 90 SSc patients as well as 165 healthy (i.e. control) individuals. The SSc cohort included patients from all four major subgroups of the diseases outlined in 2.4.3, and the IgG4-RD cohort included samples from patients in remission (i.e. inactive disease) as well as with active disease. Moreover, approximately half of the individuals from both disease cohorts were being treated with immunosuppressive agents, such as rituximab (RTX) and prednisone, or another form of treatment. Some individuals were receiving a combination of treatments. Therefore, to investigate the effects of individual as well as a combination of drugs on the gut microbiome, the medication metadata (later referred to as 'treatment information') was classified into six treatment categories: no treatment, RTX, prednisone, other medication, RTX with prednisone, and prednisone with other medication. Additional metadata, such as age and gender, were also available for each individual in the study. More cohort-specific metadata can be found in Table 1 of Publication III.

All stool samples were subjected to whole metagenome shotgun (WMS) sequencing using Illumina HiSeq 2500 instrument and demultiplexed using the Picard suite [Picard-toolkit, 2019]. The raw reads were then processed according to the steps explained in Section 3.3.2. For QC analysis, a wrapper tool around FastQC and Cutadapt, called Trim Galore! [Krueger, 2012], was used for adapter trimming; and KneadData [Huttenhower, 2020] was used for quality trimming and removing contaminating reads from the human DNA. Subsequently, the reads were taxonomically profiled using both assembly-free and assembly-based methods. Since the latter method detected more taxa, known and unknown, in all cohorts (i.e. higher alpha diversities), it was used for all taxonomic analyses in the study. Meanwhile, the assembly-free profiling, which was done using MetaPhlAn2, was used for the functional analysis using HUMAnN2.

For assembly-based profiling, MEGAHIT was used for contig reconstruction, Prodigal for gene prediction, CD-HIT for creating a non-redundant gene catalogue, BWA for mapping the reads to the gene catalogue, and MSPminer for binning the genes into metagenomic species pangenomes (MSPs) (Section 3.3.2). The abundance of each MSP in a sample was determined as the median TPM (transcript-per-million) of top 30 core genes per MSP. Finally, the MSPs were annotated using eggNOG-mapper and PhloPhlAn.

7.2 Microbiome community analyses and results

The taxonomic classification of the WMS reads using assembly-based approaches identified 504 metagenomic species pangenomes (MSPs). Several comparative analyses were performed to identify diversity- and taxonomic abundance-level differences between samples from different cohorts, treatment groups, gender, age, IgG4-RD disease status (i.e. active or inactive disease), and/or SSc subgroups.

Alpha and beta diversities of samples were quantified based on the relative abundances at MSP levels (i.e. species-level) using Shannon indices (Section 4.1.1) and Bray-Curtis dissimilarity measures (Section 4.1.2), respectively. The Shannon diversity indices were compared between disease and control cohorts using linear regression modelling with age, gender, cohort information, and treatment information as fixed-effect covariates (i.e. predictors) (Section 4.4.1). This analysis revealed decreased alpha diversities in IgG4-RD patients as compared to the controls (FDR = 0.06), which is one of the hallmarks of a dysbiotic gut microbiome (Section 2.3.2). In the SSc cohort, decreased alpha diversities were associated only with those patients, who were being treated with prednisone in combination with other drugs (FDR = 0.004, Table S2 in Publication III).

Multivariate association analysis using PERMANOVA (Section 4.4.4) with Bray-Curtis dissimilarity measures was performed in order to identify covariates (age, gender, cohort information and treatment information) that may be correlated with the differences in the gut microbial compositions of samples as a whole. Here, the cohort classification of the samples (IgG4-RD, SSc and healthy control) was identified to have the strongest association with the variations in the gut microbiomes of patients (FDR < 0.05, Table S3 of Publication III), indicating that there are compositional-level differences in the gut microbiomes of healthy and diseased individuals. These differences were also revealed by principle coordinate analysis (PCoA, Section 4.2), which is visualized in Figure 1b of Publication III.

Additionally, univariate association analyses using linear fixed effects modelling were performed to identify specific taxonomic groups (phyla and MSPs) that were differentially abundant in different cohorts, SSc subgroups, and patients with active and inactive IgG4-RD disease. These analyses are henceforth referred to as differential abundance analysis (DAA). Here, the metadata available for each sample (explained in Section 7.1) was added as fixed effect covariates to each linear model (depending on its relevance in the analysis). The phylum-level relative abundances were obtained by summing up the relative abundances of the MSPs belonging to the respective phylum. Also, the taxa (i.e. taxonomic groups) that were present in only a few samples were excluded from these analyses due to loss in statistical power (detailed explanation in 'Methods' of Publication III). Below, some of the most interesting results from the main analyses are presented. More detailed discussion of the results can be found in Publication III.

At the phylum-level, the DAA identified alterations in the abundances of Firmicutes and Bacteroidetes in one or both of the disease cohorts, which are common phenomenon seen in several other diseases [Liang *et al.*, 2018; Opazo *et al.*, 2018]. Specifically, a consistent depletion of Bacteroidetes was seen in both IgG4-RD and SSc cohorts (FDR < 0.001), with an overabundance of Firmicutes in SSc (FDR = 0.06) and Actinobacteria in IgG4-RD (FDR = 0.15), relative to the healthy controls (Figure 2b in Publication III).

At the species-level, 38 known MSPs were found to be concordantly overabundant or depleted in both diseases as compared to the controls with FDR < 0.05. Additionally, 67 known MSPs and 36 MSPs with unknown species-level annotation were found to be differentially abundant (FDR < 0.05) in either of the two diseases compared to the controls (full list in Table S4 of Publication III). From Figure 7.1, where the beta coefficients

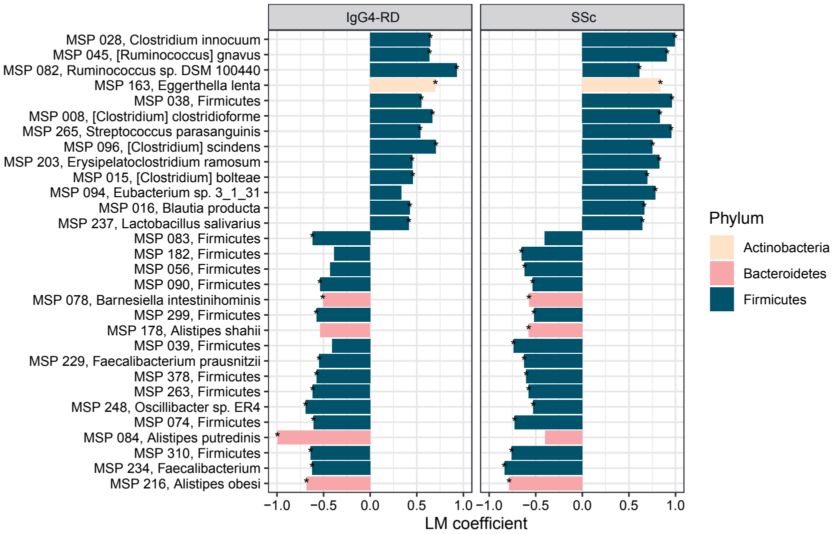


Figure 7.1. The top 30 differentially abundant species in IgG4-RD and/or SSc patients when compared to healthy controls. Here, FDR < 0.05 is indicated with '*' on the horizontal bars and the colors of the bars reflect the phylum classification of the species.

of the top 30 differentially abundant MSPs (FDR < 0.05) are plotted, and Figure S3 from Publication III, it can be seen that the MSPs that were differentially abundant in either of the diseases, were generally observed to follow a consistent trend of overabundance or depletion in the other disease as well. This indicates that similar gut microbiome differences exist in both diseases. Additionally, a species-level DAA analysis between the two disease cohorts found no MSPs (known or unknown) to be differentially abundant, which reinforces the notion that these two fibrosis-prone diseases share a common microbiome signature and architecture.

Overall, several opportunistic and pathogenic species from the *Clostridium* genus were observed to be significantly overabundant in the two diseases; all significant MSPs from the Bacteroidetes phylum were depleted in one or both of the diseases (Figure 2d in Publication III); multiple commensals typically found in the oral microbiome, such as *Streptococcus*, were overabundant in the disease cohorts (a phenomenon also seen in other autoimmune diseases, such as RA and IBD); and butyrate-producing species, such as *Faecalibaterium prausnitzii*, that promote good colonic health by inhibiting pro-inflammatory cytokines and up-regulating anti-inflammatory cytokines, were depleted in both diseases.

Additionally, *Eggertherlla lenta* (*E. lenta*), which is a species from the Actinobacteria phylum and one of the top differentially abundant species, was found to be significantly overabundant in both diseases. *E. lenta* is also

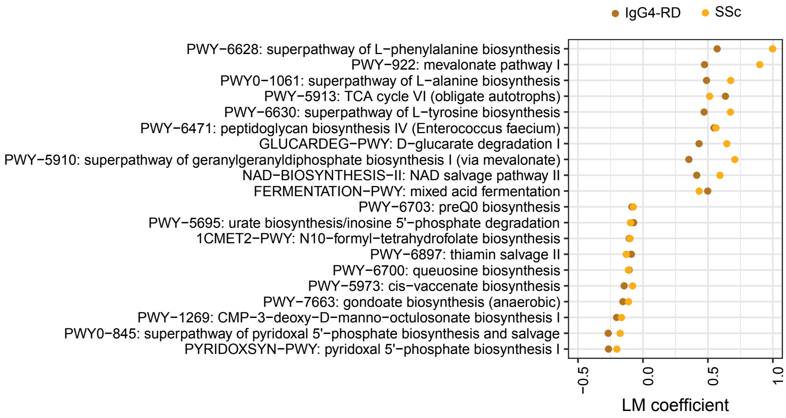


Figure 7.2. The top 20 differentially enriched pathways in both disease cohorts as compared to healthy controls with FDR < 0.05.

found to be overabundant in the gut microbiomes of patients with other ADs, such as MS and RA. In fact, a specific strain of *E. lenta* that contains the cardiac glycoside reductase (*cgr*) operon was found to be enriched in the disease cohorts using accessory gene content analysis (details of the analysis can be found in 'Methods' of Publication III). In literature, *E. lenta* with the *cgr* locus (i.e. *cgr+ E. lenta* strains) are known to have the potential to drive the activation of Th17 cells and modulate production of pro-inflammatory cytokines that may lead to the breakdown of immune homeostasis, whereas *cgr- E. lenta* strains do not.

Differences in the functional potential of the gut microbiomes of IgG4-RD and SSc patients were also assessed in comparison to healthy controls by performing linear fixed effects modelling on the relative abundances of various functional categories, such as pathways, enzymes, genes and GO categories, obtained using HUMAnN2. Several pathways were differentially enriched in the disease cohorts with FDR < 0.05 (top 20 illustrated in Figure 7.2), including the classical mevalonate pathway that leads to the synthesis of specific metabolites that play an important role in signalling the immune system and possibly altering immune responses in IgG4-RD and SSc patients. A full list can be found in Table S7 of Publication III. Additionally, assembled genes that were annotated with KEGG Orthology (KO) genes using eggNOG-mapper, were also analysed for enrichment in the disease cohorts. One of the striking results included the overabundance of a group of 12 genes belonging to the ethanolamine utilization compartment, in both IgG4-RD and SSc patients (Figure 4b in Publication III). Ethanolamine is a chemical compound that is prevalent in the gut, especially during inflammation, and can be metabolized only by specific microbes, which give them a growth advantage with limited competition for nutrition sources. A few ethanolamine metabolizers were found to be overabundant in this study also.

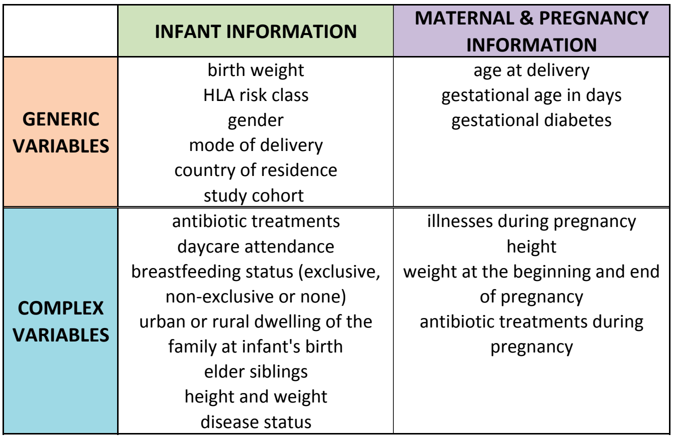
8. Publication IV: Influence of extrinsic and intrinsic factors on the early gut microbial development of T1D susceptible infants

As discussed in Section 2.3.2, the maturation of the human gut microbiome towards an adult composition takes place during the first 2-3 years (or ~1000 days) of life. The microbial exposures and colonization during these early years of life also play a crucial role in the development of the immune system and have long-term health implications. Aberrations in the early gut microbiome colonization have also been linked to several childhood diseases, including T1D. Recently, an increasing number of studies have linked the influence of the host microbiome on human health to be the consequence of the presence or absence of certain individual strains of specific microbes. Therefore, the main objective of this study was to explore the strain diversity in the gut microbiomes of T1D susceptible infants during the first three years of their lives. Additionally, the early gut microbial compositions are highly unstable and vulnerable to alterations by several environmental and host-related factors, such as diet, mode of delivery, breastfeeding patterns, antibiotic usage, etc. Therefore, another aim of this study was to investigate the impact of numerous intrinsic and extrinsic factors in the shaping of early gut microbial compositions of T1D susceptible infants. This chapter will focus on presenting results from the latter objective of this study.

8.1 Study design

This study was conducted using longitudinally collected stool samples from nearly 300 T1D susceptible infants that belong to the DIABIMMUNE project's birth cohort (introduced in Section 5.1). In the DIABIMMUNE project, monthly stool samples were collected from these infants for the first three years of their lives. The microbiome data, such as 16S rRNA and whole metagenome shotgun (WMS) sequencing, for these samples were generated and published as part of multiple different studies in the past (labeled as 'study cohort' in Table 8.1) [Kostic *et al.*, 2015; Vatanen *et al.*, 2016; Yassour *et al.*, 2018]. In all the studies, the 16S rRNA dataset was

Table 8.1. A list of some of the additional information (i.e. metadata) that was collected in the DIABIMMUNE project about each of the study participants and their mothers (arranged in separate columns) and were investigated in this study. Here, 'generic variables' refer to the information that was available for all participants, whereas the complex variables refer to those that contained missing values and often required pre-processing.



	INFANT INFORMATION	MATERNAL & PREGNANCY INFORMATION
GENERIC VARIABLES	birth weight HLA risk class gender mode of delivery country of residence study cohort	age at delivery gestational age in days gestational diabetes
COMPLEX VARIABLES	antibiotic treatments daycare attendance breastfeeding status (exclusive, non-exclusive or none) urban or rural dwelling of the family at infant's birth elder siblings height and weight disease status	illnesses during pregnancy height weight at the beginning and end of pregnancy antibiotic treatments during pregnancy

generated by sequencing the V4 hypervariable region of the 16S rRNA gene [Gevers *et al.*, 2014]; and sequencing was performed on Illumina MiSeq and Illumina HiSeq 2500 platforms.

In this publication, the 16S rRNA sequencing data was used to investigate the development of the early gut microbiome in association with the several extrinsic and intrinsic factors (i.e. external variables) that may influence it; whereas, the WMS data was used to characterize the strain-specific variations in the early gut microbiome of T1D susceptible infants.

In addition to blood and stool samples, the DIABIMMUNE project has also collected a comprehensive amount of metadata (i.e. information) related to each of the study participants (or infants) as well as their mothers and events during the pregnancy. Table 8.1 lists some of the infant and maternal information that has been collected (in the form of questionnaires) during the participants' study visits at 3, 6, 12, 18, 24 and 36 months of age. For the statistical association analyses conducted in this study, the metadata variables and corresponding covariates were divided into two categories: generic and complex variables (as shown in Table 8.1). Information from the 'generic variables' were available for all the study participants (i.e. no missing data), whereas the 'complex variables' contained missing values and often required pre-processing (discussed in Section 8.2).

8.2 Pre-processing of the data and metadata for statistical association analyses

Prior to any statistical analyses, the raw paired-end 16S rRNA sequencing reads were processed using ea-utils for demultiplexing the data, and UPARSE for performing quality control as well as OTU clustering (methods introduced in Section 3.3.1). Subsequently, a filtering step was performed on the obtained OTU table, where the samples that contained too few OTUs and/or the OTUs that were present in too few samples, were removed. Finally, a total of 3,204 samples from 289 individuals and 920 OTUs remained for further analyses.

A majority of the metadata variables also required pre-processing. Most of the pre-processing involved transforming or summarizing the raw information into a format (or multiple formats) that could be meaningfully incorporated into the statistical models. This applied to at least all the 'complex variables' in Table 8.1. Sometimes, there were multiple ways of transforming or summarizing the raw information of a variable, which resulted in multiple modelling covariates that were statistically tested one by one. For instance, two modelling covariates were derived from the infant antibiotic treatment information; five covariates were computed from the height and weight of the infants; mother's illnesses during pregnancy were classified into two groups: serious illnesses and any illness, thus establishing two covariates; and so on. Full list of covariates can be found in the Supplementary Tables 2-4 of Publication IV. Moreover, as the information was gathered using questionnaires over a span of 3 years, some answers varied over time in a contradictory manner. For instance, for some participants, the information about the 'family's dwelling at infant's birth' varied from urban to rural (or vice versa) over the years. Therefore, such situations had to either be filtered from the data or fixed using the best judgement.

8.3 Associations between the early gut microbial development and metadata variables

In this study, both multivariate and univariate statistical association analyses were performed to understand the development of the early gut microbial communities relative to the various intrinsic and extrinsic factors that may influence it. Specifically, permutational multivariate analysis of variance (PERMANOVA, explained in Section 4.4.4) was used to identify the metadata variables associated with the compositional-level variations in the gut microbiomes of infants at 2, 6 and 18 months of age; whereas linear mixed-effects (LME) modelling, as implemented in MaAsLin (Section 4.4.2), was used to identify associations between individual bacterial genera and the metadata variables in a cross-sectional (at 2, 6, and 18 months of age) as well as longitudinal manner. Associations of gut microbial diversities (Chao1 richness and Shannon diversity index) with the metadata variables were also investigated in a longitudinal manner.

In each cross-sectional analysis, the associations between the generic variables and the gut microbial communities were determined by modelling all the related covariates together in a single analysis. However, associations of the covariates related to the complex variables were determined by modelling them one at a time with the covariates of all the generic variables. Similar modelling idea was adopted in the longitudinal analyses using LME modelling, with the exception that the breastfeeding information was also considered as a generic variable. In the LME statistical analyses, the generic variables, complex variables and 'age at sample collection' were used as fixed effects, and the subject IDs were included in the models as random effects.

Overall, the statistical association analyses performed in this study uncovered several interesting environmental and host-related factors to be associated with infant gut microbial development in the DIABIMMUNE cohort. For instance, the cross-sectional PERMANOVA analysis found factors such as mode of delivery, maternal antibiotic treatments during pregnancy, breastfeeding status and country of residence to be significantly associated ($FDR < 0.05$) with the gut microbial composition variations at 2 months of age. The country of residence was linked to the differences in microbial compositions at 6 and 18 months of age as well.

Consistent with current knowledge on gut microbiome colonization and development, the microbiome diversity association analyses in this study found the alpha diversities of infants to significantly increase with age and decrease when given antibiotic treatments. Alpha diversities were also associated with breastfeeding status, country of residence, and maternal illnesses (especially serious illnesses) during pregnancy. Moreover, infants from rural households were observed to harbor richer microbiomes than those from urban households throughout the first 3 years of life. The gut microbial diversity was also positively correlated ($FDR \leq 0.10$) with linear growth of an infant (i.e. the average increase in height per year and height at 3 years of age), indicating that taller and faster growing infants had more diverse gut microbiomes in the early years of life. The taxonomic level association analysis identified that an infant's average increase in height and weight during the first three years as well as the height and weight of infants at 3 years of age were positively correlated to the relative abundance of genus *Dialister*.

More results and discussion on the association analyses can be found in the Supplementary Tables 2-4 and Supplementary Note 1 of Publication IV.

9. Concluding Remarks

To conclude, the aim of this thesis was to perform computational and statistical analyses on HT 'omics' datasets in order to improve our understanding about the etiology and pathogenesis of autoimmune diseases, specifically T1D, IgG4-RD and SSc. Overall, the objectives of the studies in this thesis were to identify predictive signatures, such as gene expression markers and pathways, which can help predict the onset of autoimmunity (i.e. loss of self-tolerance) and/or understand the molecular changes that occur during disease progression; as well as to determine etiological signatures, such as presence of microbes, microbial genes or other environmental factors, that could be influencing the pathogenesis of the disease and/or contributing to its etiology.

The aim of Publication I was to identify early gene expression markers from PBMCs,CD4-enriched, CD8-enriched, and CD4- and CD8-depleted fractions of immune cells that are predictive of the onset of autoimmunity and/or capable of characterizing T1D disease progression. Indeed, this study revealed several T1D-associated genes that have not been reported before, including *IL32* that has previously been associated with other ADs, but not T1D. *IL32* is a gene that encodes a proinflammatory cytokine that is expressed by many immune and epithelial cells. It was found to be differentially expressed in the case individuals across the time-course as well as in the window before seroconversion, in all four fractions of immune cells. This indicates that *IL32* could be one of the important immunological signatures that is involved in development of β-cell autoimmunity as well as in the progression of the disease. The connection between *IL32* and autoimmunity was further strengthened by the fact that it was identified to be co-regulated in all four cell fractions with two genes that have been previously linked with autoimmunity; one of which has been associated with T1D. Moreover, several cytokines have already been implicated in the pathogenesis of T1D and are considered to have the potential to be immunotherapeutic targets for T1D [Lu *et al.*, 2020]. Given that cytokines are important signalling molecules in immunity (Section 2.2.2), identifying cytokine genes to be differentially expressed in T1D case individuals indicates that immune cell signaling undergoes dynamic changes during the development of the disease. Additionally, since the second aim of the study was to identify the specific immune cell types that likely express *IL32*, it was shown that the high levels of *IL32* in case individuals were mostly expressed by T cells, especially highly differentiated CD8+ T cells, and NK cells. However, it should be noted here that the results presented in this publication largely apply to infants with prediabetes and may not apply to adolescents and adults. Also, the results of this study need to be validated on a larger cohort of prediabetic children.

In Publication II, the goal was to develop a method (called personalized approach) that first models longitudinal gene expression data in a personalised manner in order to identify DEGs, and then summarizes the DEGs on a pathway-level in order to identify disease-associated pathways. Essentially, several statistical methods can be used to model the longitudinal data, but in this method, Gaussian processes (GPs; Section 4.5) were used. This is because GPs can capture the real underlying structure of longitudinal data along with embedded noise in a non-linear manner and without imposing strong modelling assumptions. Also, they can robustly estimate unobserved data (demonstrated in the robustness analysis of this publication), which is an advantageous feature when modelling small sample sizes. The personalised approach assumes an experimental design, where each case individual is matched with a control individual. DEGs are then identified for each case-control pair individually using GPs. Finally, the list of DEGs from each pair is combined on a pathway-level using a permutation-based empirical hypothesis testing. Summarizing the results on a pathway-level often overcomes the gene-level variability and inconsistencies that exist in complex heterogeneous diseases, such as T1D. The generalizability of the pathway-level results across datasets has also been shown in this publication by comparing the results obtained from analysing two independent T1D datasets using the personalised approach. This method can be used to identify pathways that are perturbed in case individuals across the course of disease progression or within specific time-windows. By comparing the results from the personalised approach to those of non-personalised approaches, it was demonstrated that the personalised approach was capable of providing more insights into the progression of heterogeneous diseases. It identified several critical pathways involving many of the key immunological players in the time-window as well as time-course analyses that were missed by the other methods.

The aim of publication III was to study the gut microbial compostions of IgG4-RD and SSc patients and identify potential sources of microbial signals that may be contributing to the etiology of the diseases. Firstly, this study showed that IgG4-RD patients and SSc patients that were being treated by prednisone in combination with other drugs, had lower alpha diversities and this is a hallmark of a dysbiotic gut. Additionally, compositional-level differences were seen between diseased and healthy individuals. Specifically, Firmicutes and Bacteroidetes were either depleted or overabundant in at least on one of the disease cohorts. Alteration in the abundances of Firmicutes to Bacteroidetes is a common phenomenon seen in many other diseases [Liang *et al.*, 2018; Opazo *et al.*, 2018]. Overall, many pathogenic and opportunistic species were found to be significantly overabundant in the diseased patients, whereas several commensal bacteria that promote good colonic health were found to be significantly depleted. Among the most significantly overabundance species in both diseases was *Eggerthella lenta (E.lenta)*, which is also found to be overabundant in other autoimmune diseases. Interestingly, a specific strain of this species that contains a cluster of genes with the potential of activating and modulating certain components of the immune system was found to be enriched in the two disease cohorts. This study also uncovered several disease-relevant functional-level differences in the gut microbiomes of the patients as compared to the healthy controls. For instance, a pathway that is known to play a crucial role in signalling the immune system and possibly influencing the immune responses, was found to be enriched in disease cohorts. Several functional signatures also link the microbiomes of IgG4-RD and SSc patients to inflammation. Overall, the results of this study indicate that the gut microbiome composition likely plays a significant role in the etiology and pathogenesis of IgG4-RD and SSc. The results of this study warrant further studies into the identification of specific microbiome-derived antigens or other molecules that can drive the autoimmune pathogenesis.

One of the objectives of Publication IV was to identify the environmental and host-related factors that may be influencing the development of the early gut microbiome in T1D susceptible infants and in the process contributing the development and education of the infant immune system. This study identified the association of several interesting maternal- and infant-related factors with the development of the gut microbiome in infants predisposed to T1D. Some factors, such as age, country of birth, mode of delivery, breastfeeding and use of antibiotics, that have previously been linked to the early gut microbial composition in T1D [Vatanen *et al.*, 2016; Yassour *et al.*, 2016], were also found in this study. Additionally, this study also linked several new factors to the early gut microbial development in the context of T1D, including maternal antibiotic treatments and maternal illnesses during pregnancy; household location at the time of infant's birth (i.e. urban or rural dwelling); average increase in height and weight per year; as well as height and weight at 3 years of age. Overall, the results support the hygiene hypothesis and suggest that increased microbial exposure during childhood may encourage the development of a rich and diverse early gut microbiome. Additionally, the results also indicate that maternal health and antibiotic usage during pregnancy as well as antibiotic usage by the infant in the early years of life, could influence the early colonization of the infant gut. Interestingly, taller and faster growing children harbor a diverse gut microbiome; and the rate at which a child grows in both height and weight during the first 3 years of life can influence the presence of specific bacteria. It should be noted that malnutrition and stunted growth during infancy has been linked to immature or different gut microbial development during childhood [Subramanian *et al.*, 2014; Dinh *et al.*, 2016].

Altogether, the results in this thesis have advanced our knowledge about the environmental and host-related factors that may be contributing to the etiology of specific autoimmune diseases as well as about the markers and pathways that may be involved in the disease pathogenesis. Of course, there is a lot that is still unknown about the development of autoimmune diseases, including T1D, IgG4-RD and SSc. However, with more interdisciplinary studies similar to the ones in this thesis, the biomedical community will soon be able to unravel the underlying mechanisms that drive autoimmune diseases and establish the treatments for preventing or curing them. With the help of well-designed studies, HT 'omics' datasets, and robust statistical and computational tools, we can definitely cross this finish line.

References

Abel, A., Yang, C., Thakar, M., and Malarkannan, S. (2018). Natural killer cells: development, maturation, and clinical utilization. front immunol 9: 1869.

Abraham, M. and Khosroshahi, A. (2017). Diagnostic and treatment workup for igg4-related disease. *Expert review of clinical immunology*, **13**(9), 867–875.

Abubucker, S., Segata, N., Goll, J., Schubert, A. M., Izard, J., Cantarel, B. L., Rodriguez-Mueller, B., Zucker, J., Thiagarajan, M., Henrissat, B., *et al.* (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol*, **8**(6), e1002358.

Affymetrix (2002). Statistical algorithms description document `http://tools.thermofisher.com/content/sfs/brochures/sadd_whitepaper.pdf` (accessed: 23 july 2020).

Affymetrix (2010). Affymetrix human genome u219 array plate `https://www.affymetrix.com/support/technical/datasheets/hgu219_ap_datasheet.pdf` (accessed: 15 july 2020).

Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.

Äijö, T., Edelman, S. M., Lönnberg, T., Larjo, A., Kallionpää, H., Tuomela, S., Engström, E., Lahesmaa, R., and Lähdesmäki, H. (2012). An integrative computational systems biology approach identifies differentially regulated dynamic transcriptome signatures which drive the initiation of human t helper cell differentiation. *BMC genomics*, **13**(1), 572.

Alberts, B. (2018). Molecular biology of the cell.

Allaband, C., McDonald, D., Vázquez-Baeza, Y., Minich, J. J., Tripathi, A., Brenner, D. A., Loomba, R., Smarr, L., Sandborn, W. J., Schnabl, B., *et al.* (2019). Microbiome 101: studying, analyzing, and interpreting gut microbiome data for clinicians. *Clinical Gastroenterology and Hepatology*, **17**(2), 218–230.

Allanore, Y., Simms, R., Distler, O., Trojanowska, M., Pope, J., Denton, C. P., and Varga, J. (2015). Systemic sclerosis. *Nature reviews Disease primers*, **1**(1), 1–21.

Allen, M. P. (1997). The t test for the simple regression coefficient. *Understanding Regression Analysis*, pages 66–70.

Alneberg, J., Bjarnason, B. S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F., and Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature methods*, **11**(11), 1144–1146.

Alpaydin, E. (2010). *Introduction to Machine learning Second Edition*. The MIT Press.

Altschuler, S. J. and Wu, L. F. (2010). Cellular heterogeneity: do differences make a difference? *Cell*, **141**(4), 559–563.

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome biology*, **21**(1), 1–16.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, **11**(1), 0.

Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., Smyth, G. K., Huber, W., and Robinson, M. D. (2013). Count-based differential expression analysis of rna sequencing data using r and bioconductor. *Nature protocols*, **8**(9), 1765.

Anders, S., Pyl, P. T., and Huber, W. (2015). Htseq—a python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**(2), 166–169.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral ecology*, **26**(1), 32–46.

Anderson, M. J. (2014). Permutational multivariate analysis of variance (permanova). *Wiley statsref: statistics reference online*, pages 1–15.

Andrews, S. (2010). Fastqc: A quality control for high throughput sequence data `https://www.bioinformatics.babraham.ac.uk/projects/fastqc/` (accessed: 31 july 2020).

Aronesty, E. (2013). Comparison of sequencing utility programs. *The open bioinformatics journal*, **7**(1).

Arrieta, M.-C., Stiemsma, L. T., Amenyogbe, N., Brown, E. M., and Finlay, B. (2014). The intestinal microbiome in early life: health and disease. *Frontiers in immunology*, **5**, 427.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.* (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, **25**(1), 25–29.

Aßhauer, K. P., Wemheuer, B., Daniel, R., and Meinicke, P. (2015). Tax4fun: predicting functional profiles from metagenomic 16s rrna data. *Bioinformatics*, **31**(17), 2882–2884.

Assimakopoulos, S. F., Triantos, C., Maroulis, I., and Gogos, C. (2018). The role of the gut barrier function in health and disease. *Gastroenterology research*, **11**(4), 261.

Atkinson, M. A. (2012). The pathogenesis and natural history of type 1 diabetes. *Cold Spring Harbor perspectives in medicine*, **2**(11), a007641.

Atkinson, M. A., Eisenbarth, G. S., and Michels, A. W. (2014). Type 1 diabetes. *The Lancet*, **383**(9911), 69–82.

Bacher, R. and Kendziorski, C. (2016). Design and computational analysis of single-cell rna-sequencing experiments. *Genome biology*, **17**(1), 63.

Bacher, R., Chu, L.-F., Leng, N., Gasch, A. P., Thomson, J. A., Stewart, R. M., Newton, M., and Kendziorski, C. (2017). Scnorm: robust normalization of single-cell rna-seq data. *Nature methods*, **14**(6), 584.

Bednar, M. (2000). Dna microarray technology and application. *Medical Science Monitor*, **6**(4), MT796–MT800.

Belkaid, Y. and Hand, T. W. (2014). Role of the microbiota in immunity and inflammation. *Cell*, **157**(1), 121–141.

Belkaid, Y. and Harrison, O. J. (2017). Homeostatic immunity and the microbiota. *Immunity*, **46**(4), 562–576.

Bellando-Randone, S. (2010). Patient subgroups and potential risk factors in systemic sclerosis: is there a possibility of an early diagnosis? *International Journal of Clinical Rheumatology*, **5**(5), 555.

Bendall, S. C., Davis, K. L., Amir, E.-a. D., Tadmor, M. D., Simonds, E. F., Chen, T. J., Shenfeld, D. K., Nolan, G. P., and Pe'er, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell*, **157**(3), 714–725.

Bending, D., Zaccone, P., and Cooke, A. (2012). Inflammation and type one diabetes. *International immunology*, **24**(6), 339–346.

Bharti, R. and Grimm, D. G. (2019). Current challenges and best-practice protocols for microbiome analysis. *Briefings in Bioinformatics*, **00**(00), 1–16.

Bik, E. M. (2016). Focus: microbiome: the hoops, hopes, and hypes of human microbiome research. *The Yale journal of biology and medicine*, **89**(3), 363.

Blumenberg, M. (2019). Introductory chapter: Transcriptome analysis. In *Transcriptome Analysis*. IntechOpen.

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

Borish, L. C. and Steinke, J. W. (2003). 2. cytokines and chemokines. *Journal of Allergy and Clinical Immunology*, **111**(2), S460–S475.

Bray, J. R. and Curtis, J. T. (1957). An ordination of the upland forest communities of southern wisconsin. *Ecological monographs*, **27**(4), 326–349.

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic rna-seq quantification. *Nature biotechnology*, **34**(5), 525–527.

Brito-Zerón, P., Ramos-Casals, M., Bosch, X., and Stone, J. H. (2014). The clinical spectrum of igg4-related disease. *Autoimmunity reviews*, **13**(12), 1203–1210.

Brodin, P. and Davis, M. M. (2017). Human immune system variation. *Nature reviews immunology*, **17**(1), 21.

Brumfield, K. D., Huq, A., Colwell, R. R., Olds, J. L., and Leddy, M. B. (2020). Microbial resolution of whole genome shotgun and 16s amplicon metagenomic sequencing using publicly available neon data. *Plos one*, **15**(2), e0228899.

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

Bumgarner, R. (2013). Overview of dna microarrays: types, applications, and their future. *Current protocols in molecular biology*, **101**(1), 22–1.

Burbelo, P. D., Gordon, S. M., Waldman, M., Edison, J. D., Little, D. J., Stitt, R. S., Bailey, W. T., Hughes, J. B., and Olson, S. W. (2019). Autoantibodies are present before the clinical diagnosis of systemic sclerosis. *PloS one*, **14**(3).

Calle, M. L. (2019). Statistical analysis of metagenomics data. *Genomics & informatics*, **17**(1).

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Pena, A. G., Goodrich, J. K., Gordon, J. I., *et al.* (2010). Qiime allows analysis of high-throughput community sequencing data. *Nature methods*, **7**(5), 335–336.

Casamassimi, A., Federico, A., Rienzo, M., Esposito, S., and Ciccodicola, A. (2017). Transcriptome profiling in human diseases: new advances and perspectives. *International journal of molecular sciences*, **18**(8), 1652.

Castro, C. and Gourley, M. (2010). Diagnostic testing and interpretation of tests for autoimmunity. *Journal of Allergy and Clinical Immunology*, **125**(2), S238–S247.

Celis, I., Kriekaart, R., Aliredjo, R., van Lochem, E., van der Vorst, M., and Hassing, R. (2017). Igg4-related disease: a disease we probably often overlook. *Neth J Med*, **75**(1), 27–31.

Chakravorty, S., Helb, D., Burday, M., Connell, N., and Alland, D. (2007). A detailed analysis of 16s ribosomal rna gene segments for the diagnosis of pathogenic bacteria. *Journal of microbiological methods*, **69**(2), 330–339.

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of statistics*, pages 265–270.

Chaplin, D. D. (2006). 1. overview of the human immune response. *Journal of allergy and clinical immunology*, **117**(2), S430–S435.

Chen, G., Ning, B., and Shi, T. (2019). Single-cell rna-seq technologies and related computational data analysis. *Frontiers in genetics*, **10**, 317.

Chen, J., Wang, Y., Shen, B., and Zhang, D. (2013). Molecular signature of cancer at gene level or pathway level? case studies of colorectal cancer and prostate cancer microarray data. *Computational and mathematical methods in medicine*, **2013**.

Chen, X., Wang, L., Smith, J. D., and Zhang, B. (2008). Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes. *Bioinformatics*, **24**(21), 2474–2481.

Cheng, L., Ramchandran, S., Vatanen, T., Lietzén, N., Lahesmaa, R., Vehtari, A., and Lähdesmäki, H. (2019). An additive gaussian process regression model for interpretable non-parametric analysis of longitudinal data. *Nature communications*, **10**(1), 1–11.

Clancy, S. (2008). Dna transcription. *Nature education*, **1**(1), 41.

Clark, M., Kroger, C. J., and Tisch, R. M. (2017). Type 1 diabetes: a chronic anti-self-inflammatory response. *Frontiers in immunology*, **8**, 1898.

Clark, R. and Kupper, T. (2005). Old meets new: the interaction between innate and adaptive immunity. *Journal of Investigative Dermatology*, **125**(4), 629–637.

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian journal of ecology*, **18**(1), 117–143.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., *et al.* (2016). A survey of best practices for rna-seq data analysis. *Genome biology*, **17**(1), 13.

Cooper, G. S. and Stroehla, B. C. (2003). The epidemiology of autoimmune diseases. *Autoimmunity reviews*, **2**(3), 119–125.

Costea, P. I., Zeller, G., Sunagawa, S., and Bork, P. (2014). A fair comparison. *Nature methods*, **11**(4), 359–359.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, **49**(1), 1–18.

Crosbie, S. and Hinch, G. (1985). An intuitive explanation of generalised linear models. *New Zealand journal of agricultural research*, **28**(1), 19–29.

Cullen, C. M., Aneja, K. K., Beyhan, S., Cho, C. E., Woloszynek, S., Convertino, M., McCoy, S. J., Zhang, Y., Anderson, M. Z., Alvarez-Ponce, D., *et al.* (2020). Emerging priorities for microbiome research. *Frontiers in Microbiology*, **11**, 136.

Da Silva Xavier, G. (2018). The cells of the islets of langerhans. *Journal of clinical medicine*, **7**(3), 54.

Dalod, M., Chelbi, R., Malissen, B., and Lawrence, T. (2014). Dendritic cell maturation: functional specialization through signaling specificity and transcriptional programming. *The EMBO journal*, **33**(10), 1104–1116.

Danzer, C. and Mattner, J. (2013). Impact of microbes on autoimmune diseases. *Archivum immunologiae et therapiae experimentalis*, **61**(3), 175–186.

De Amorim, R. C. and Hennig, C. (2015). Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences*, **324**, 126–145.

De Luca, F. and Shoenfeld, Y. (2019). The microbiome in autoimmune diseases. *Clinical & Experimental Immunology*, **195**(1), 74–85.

Debnath, M., Prasad, G. B., and Bisen, P. S. (2010). *Molecular diagnostics: promises and possibilities*. Springer Science & Business Media.

Dedrick, S., Sundaresh, B., Huang, Q., Brady, C., Yoo, T., Cronin, C., Rudnicki, C., Flood, M., Momeni, B., Ludvigsson, J., *et al.* (2020). The role of gut microbiota and environmental factors in type 1 diabetes pathogenesis. *Frontiers in Endocrinology*, **11**, 78.

Della-Torre, E., Lanzillotta, M., and Doglioni, C. (2015). Immunology of igg4-related disease. *Clinical & Experimental Immunology*, **181**(2), 191–206.

den Haan, J. M., Arens, R., and van Zelm, M. C. (2014). The activation of the adaptive immune system: cross-talk between antigen-presenting cells, t cells and b cells. *Immunology letters*, **162**(2), 103–112.

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., and Lempicki, R. A. (2003). David: database for annotation, visualization, and integrated discovery. *Genome biology*, **4**(9), 1–11.

Denton, C. P. and Khanna, D. (2017). Systemic sclerosis. *The Lancet*, **390**(10103), 1685–1699.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006). Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Applied and environmental microbiology*, **72**(7), 5069–5072.

Desbois, A. C. and Cacoub, P. (2016). Systemic sclerosis: an update in 2016. *Autoimmunity reviews*, **15**(5), 417–426.

Diaz, A., Liu, S. J., Sandoval, C., Pollen, A., Nowakowski, T. J., Lim, D. A., and Kriegstein, A. (2016). Scell: integrated analysis of single-cell rna-seq data. *Bioinformatics*, **32**(14), 2219–2220.

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., *et al.* (2013). A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in bioinformatics*, **14**(6), 671–683.

DiMeglio, L. A., Evans-Molina, C., and Oram, R. A. (2018). Type 1 diabetes. *The Lancet*, **391**(10138), 2449–2462.

Ding, B., Zheng, L., Zhu, Y., Li, N., Jia, H., Ai, R., Wildberg, A., and Wang, W. (2015). Normalization and noise reduction for single cell rna-seq experiments. *Bioinformatics*, **31**(13), 2225–2227.

Dinh, D. M., Ramadass, B., Kattula, D., Sarkar, R., Braunstein, P., Tai, A., Wanke, C. A., Hassoun, S., Kane, A. V., Naumova, E. N., *et al.* (2016). Longitudinal analysis of the intestinal microbiota in persistently stunted young children in south india. *PLoS One*, **11**(5), e0155405.

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., *et al.* (2012). Landscape of transcription in human cells. *Nature*, **489**(7414), 101–108.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). Star: ultrafast universal rna-seq aligner. *Bioinformatics*, **29**(1), 15–21.

Drier, Y., Sheffer, M., and Domany, E. (2013). Pathway-based personalized analysis of cancer. *Proceedings of the National Academy of Sciences*, **110**(16), 6388–6393.

D'Argenio, V. (2018). The high-throughput analyses era: are we ready for the data struggle? *High-throughput*, **7**(1), 8.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than blast. *Bioinformatics*, **26**(19), 2460–2461.

Edgar, R. C. (2013). Uparse: highly accurate otu sequences from microbial amplicon reads. *Nature methods*, **10**(10), 996–998.

Edgar, R. C. (2017). Accuracy of microbial community diversity estimated by closed-and open-reference otus. *PeerJ*, **5**, e3889.

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. (2011). Uchime improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**(16), 2194–2200.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., *et al.* (2019). The pfam protein families database in 2019. *Nucleic acids research*, **47**(D1), D427–D432.

Engström, P. G., Steijger, T., Sipos, B., Grant, G. R., Kahles, A., Alioto, T., Behr, J., Bertone, P., Bohnert, R., Campagna, D., *et al.* (2013). Systematic evaluation of spliced alignment programs for rna-seq data. *Nature methods*, **10**(12), 1185–1191.

Escobar-Zepeda, A., Vera-Ponce de Leon, A., and Sanchez-Flores, A. (2015). The road to metagenomics: from microbiology to dna sequencing technologies and bioinformatics. *Frontiers in genetics*, **6**, 348.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., *et al.* (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

Ferreira, R. C., Guo, H., Coulson, R. M., Smyth, D. J., Pekalski, M. L., Burren, O. S., Cutler, A. J., Doecke, J. D., Flint, S., McKinney, E. F., *et al.* (2014). A type i interferon transcriptional signature precedes autoimmunity in children genetically at risk for type 1 diabetes. *Diabetes*, **63**(7), 2538–2550.

Ferretti, P., Pasolli, E., Tett, A., Asnicar, F., Gorfer, V., Fedi, S., Armanini, F., Truong, D. T., Manara, S., Zolfo, M., *et al.* (2018). Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. *Cell host & microbe*, **24**(1), 133–145.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., *et al.* (2015). Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome biology*, **16**(1), 1–13.

Fodor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T., and Solas, D. (1991). Light-directed, spatially addressable parallel chemical synthesis. *science*, **251**(4995), 767–773.

Fonseca, N. A., Rung, J., Brazma, A., and Marioni, J. C. (2012). Tools for mapping high-throughput sequencing data. *Bioinformatics*, **28**(24), 3169–3177.

Fonseca, N. A., Marioni, J., and Brazma, A. (2014). Rna-seq gene profiling-a systematic empirical comparison. *Plos one*, **9**(9), e107026.

Franzosa, E. A., McIver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., Lipson, K. S., Knight, R., Caporaso, J. G., Segata, N., *et al.* (2018). Species-level functional profiling of metagenomes and metatranscriptomes. *Nature methods*, **15**(11), 962–968.

Galecki, A. and Burzykowski, T. (2013). *Linear Mixed-Effects Models Using R: A Step-by-Step Approach*. Springer Science + Business Media New York.

Garber, M., Grabherr, M. G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using rna-seq. *Nature methods*, **8**(6), 469–477.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). Bayesian data analysis chapman & hall. *CRC Texts in Statistical Science*.

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

Gevers, D., Kugathasan, S., Denson, L. A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., Schwager, E., Knights, D., Song, S. J., Yassour, M., *et al.* (2014). The treatment-naive microbiome in new-onset crohn's disease. *Cell host & microbe*, **15**(3), 382–392.

Gianchecchi, E. and Fierabracci, A. (2019). Recent advances on microbiota involvement in the pathogenesis of autoimmunity. *International journal of molecular sciences*, **20**(2), 283.

Ginsburg, G. S. and Willard, H. F. (2009). *Essentials of genomic and personalized medicine*. Academic Press.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology*, **8**, 2224.

Goodrich, J. K., Di Rienzi, S. C., Poole, A. C., Koren, O., Walters, W. A., Caporaso, J. G., Knight, R., and Ley, R. E. (2014). Conducting a microbiome study. *Cell*, **158**(2), 250–262.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, **17**(6), 333.

Gregory Alvord, W., Roayaei, J. A., Quinones, O. A., and Schneider, K. T. (2007). A microarray analysis for differential gene expression in the soybean genome using bioconductor and r. *Briefings in Bioinformatics*, **8**(6), 415–431.

Gutierrez-Arcelus, M., Rich, S. S., and Raychaudhuri, S. (2016). Autoimmune diseases—connecting risk alleles with molecular traits of the immune system. *Nature Reviews Genetics*, **17**(3), 160.

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S. K., Sodergren, E., *et al.* (2011). Chimeric 16s rrna sequence formation and detection in sanger and 454-pyrosequenced pcr amplicons. *Genome research*, **21**(3), 494–504.

Haldar, D. and Hirschfield, G. M. (2018). Deciphering the biology of igg4-related disease: specific antigens and disease? *Gut*, **67**(4), 602–605.

Hamady, M. and Knight, R. (2009). Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome research*, **19**(7), 1141–1152.

Haque, A., Engel, J., Teichmann, S. A., and Lönnberg, T. (2017). A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. *Genome medicine*, **9**(1), 1–12.

Hardcastle, T. J. and Kelly, K. A. (2010). bayseq: empirical bayesian methods for identifying differential expression in sequence count data. *BMC bioinformatics*, **11**(1), 1–14.

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome biology*, **18**(1), 1–15.

Heinonen, M., Guipaud, O., Milliat, F., Buard, V., Micheau, B., Tarlet, G., Benderitter, M., Zehraoui, F., and d'Alche Buc, F. (2015). Detecting time periods of differential gene expression using gaussian processes: an application to endothelial cells exposed to radiotherapy dose fraction. *Bioinformatics*, **31**(5), 728–735.

Heller, M. J. (2002). Dna microarray technology: devices, systems, and applications. *Annual review of biomedical engineering*, **4**(1), 129–153.

Hewitt, E. W. (2003). The mhc class i antigen presentation pathway: strategies for viral immune evasion. *Immunology*, **110**(2), 163–169.

Hosack, D. A., Dennis, G., Sherman, B. T., Lane, H. C., and Lempicki, R. A. (2003). Identifying biological themes within lists of genes with ease. *Genome biology*, **4**(10), R70.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, **24**(6), 417.

Hubers, L. M., Vos, H., Schuurman, A. R., Erken, R., Elferink, R. P. O., Burgering, B., Van De Graaf, S. F., and Beuers, U. (2018). Annexin a11 is targeted by igg4 and igg1 autoantibodies in igg4-related disease. *Gut*, **67**(4), 728–735.

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., Von Mering, C., and Bork, P. (2017). Fast genome-wide functional annotation through orthology assignment by eggnog-mapper. *Molecular biology and evolution*, **34**(8), 2115–2122.

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.-J., and Tappu, R. (2016). Megan community edition-interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS computational biology*, **12**(6), e1004957.

Huttenhower (2020). Kneaddata: Tool designed to perform quality control on metagenomic sequencing data `https://huttenhower.sph.harvard.edu/kneaddata/` (accessed: 20 september 2020).

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., FitzGerald, M. G., Fulton, R. S., *et al.* (2012). Structure, function and diversity of the healthy human microbiome. *nature*, **486**(7402), 207.

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, **11**(1), 119.

Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., and Teichmann, S. A. (2016). Classification of low quality cells from single-cell rna-seq data. *Genome biology*, **17**(1), 1–15.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**(2), 249–264.

Isolauri, E. (2012). Development of healthy gut microbiota early in life. *Journal of paediatrics and child health*, **48**, 1–6.

Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T. R., Prensner, J. R., Evans, J. R., Zhao, S., *et al.* (2015). The landscape of long noncoding rnas in the human transcriptome. *Nature genetics*, **47**(3), 199–208.

Jain, A. and Pasare, C. (2017). Innate control of adaptive immunity: beyond the three-signal paradigm. *The Journal of Immunology*, **198**(10), 3791–3800.

Janda, J. M. and Abbott, S. L. (2007). 16s rrna gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, **45**(9), 2761–2764.

Janeway, C. A., Travers, P., Walport, M., and Shlomchik, M. J. (2001). *Immunobiology, 5th edition: The immune system in health and disease*. Garland Science.

Jasinski, J. and Eisenbarth, G. (2005). Insulin as a primary autoantigen for type 1a diabetes. *Journal of Immunology Research*, **12**(3), 181–186.

Ji, F. and Sadreyev, R. I. (2018). Rna-seq: Basic bioinformatics analysis. *Current protocols in molecular biology*, **124**(1), e68.

Jiang, N., Leach, L. J., Hu, X., Potokina, E., Jia, T., Druka, A., Waugh, R., Kearsey, M. J., and Luo, Z. W. (2008). Methods for evaluating gene expression from affymetrix microarray datasets. *BMC bioinformatics*, **9**(1), 284.

Jiang, P., Thomson, J. A., and Stewart, R. (2016). Quality control of single-cell rna-seq by sinqc. *Bioinformatics*, **32**(16), 2514–2516.

Jin, L., Zuo, X.-Y., Su, W.-Y., Zhao, X.-L., Yuan, M.-Q., Han, L.-Z., Zhao, X., Chen, Y.-D., and Rao, S.-Q. (2014). Pathway-based analysis tools for complex diseases: a review. *Genomics, proteomics & bioinformatics*, **12**(5), 210–220.

John, J. S. (2011). Seqprep: Tool for stripping adaptors and/or merging paired reads with overlap into single reads. *URL: https://githubcom/jstjohn/SeqPrep*.

Kallionpää, H., Elo, L. L., Laajala, E., Mykkänen, J., Ricaño-Ponce, I., Vaarma, M., Laajala, T. D., Hyöty, H., Ilonen, J., Veijola, R., *et al.* (2014). Innate immune activity is detected prior to seroconversion in children with hla-conferred type 1 diabetes susceptibility. *Diabetes*, **63**(7), 2402–2414.

Kamisawa, T., Zen, Y., Pillai, S., and Stone, J. H. (2015). Igg4-related disease. *The Lancet*, **385**(9976), 1460–1471.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, **45**(D1), D353–D361.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**(430), 773–795.

Katayama, S., Töhönen, V., Linnarsson, S., and Kere, J. (2013). Samstrt: statistical test for differential expression in single-cell transcriptome with spike-in normalization. *Bioinformatics*, **29**(22), 2943–2945.

Kelley, D. R., Liu, B., Delcher, A. L., Pop, M., and Salzberg, S. L. (2012). Gene prediction with glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic acids research*, **40**(1), e9–e9.

Khan, M. F. and Wang, H. (2020). Environmental exposures and autoimmune diseases: Contribution of gut microbiome. *Frontiers in immunology*, **10**, 3094.

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature methods*, **11**(7), 740–742.

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol*, **8**(2), e1002375.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, **14**(4), R36.

Kim, D., Langmead, B., and Salzberg, S. L. (2015). Hisat: a fast spliced aligner with low memory requirements. *Nature methods*, **12**(4), 357–360.

Kivity, S., Agmon-Levin, N., Blank, M., and Shoenfeld, Y. (2009). Infections and autoimmunity–friends or foes? *Trends in immunology*, **30**(8), 409–414.

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., and Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, **161**(5), 1187–1201.

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolek, T., McCall, L.-I., McDonald, D., *et al.* (2018). Best practices for analysing microbiomes. *Nature Reviews Microbiology*, **16**(7), 410–422.

Knip, M. (2017). Type 1 diabetes mellitus is a heterogeneous disease. *Nature Reviews Endocrinology*, **13**(9), 1.

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015). The technology and biology of single-cell rna sequencing. *Molecular cell*, **58**(4), 610–620.

Kondrashova, A., Seiskari, T., Ilonen, J., Knip, M., and Hyöty, H. (2013). The 'hygiene hypothesis' and the sharp gradient in the incidence of autoimmune and allergic diseases between russian karelia and finland. *Apmis*, **121**(6), 478–493.

Kopylova, E., Navas-Molina, J. A., Mercier, C., Xu, Z. Z., Mahé, F., He, Y., Zhou, H.-W., Rognes, T., Caporaso, J. G., and Knight, R. (2016). Open-source sequence clustering methods improve the state of the art. *MSystems*, **1**(1).

Kostic, A. D., Gevers, D., Siljander, H., Vatanen, T., Hyötyläinen, T., Hämäläinen, A.-M., Peet, A., Tillmann, V., Pöhö, P., Mattila, I., *et al.* (2015). The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell host & microbe*, **17**(2), 260–273.

Krueger, F. (2012). Trim galore! a wrapper tool around cutadapt and fastqc to consistently apply quality and adapter trimming to fastq files `https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/` (accessed: 22 october 2020).

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, **22**(1), 79–86.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome.

Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., Clemente, J. C., Burkepile, D. E., Thurber, R. L. V., Knight, R., *et al.* (2013). Predictive functional profiling of microbial communities using 16s rrna marker gene sequences. *Nature biotechnology*, **31**(9), 814–821.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, **9**(4), 357.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biology*, **10**(3), R25.

Laurent, P., Sisirak, V., Lazaro, E., Richez, C., Duffau, P., Blanco, P., Truchetet, M.-E., and Contin-Bordes, C. (2018). Innate immunity in systemic sclerosis fibrosis: recent advances. *Frontiers in immunology*, **9**, 1702.

Lazar, V., Ditu, L.-M., Pircalabioru, G. G., Gheorghe, I., Curutiu, C., Holban, A. M., Picu, A., Petcu, L., and Chifiriuc, M. C. (2018). Aspects of gut microbiota and immune system interactions in infectious diseases, immunopathology, and cancer. *Frontiers in immunology*, **9**, 1830.

Lee, E., Chuang, H.-Y., Kim, J.-W., Ideker, T., and Lee, D. (2008). Inferring pathway activity toward precise disease classification. *PLoS comput biol*, **4**(11), e1000217.

Legendre, P., Legendre, L., *et al.* (1998). Numerical ecology: developments in environmental modelling. *Developments in Environmental Modelling*, **20**.

Levy, M., Kolodziejczyk, A. A., Thaiss, C. A., and Elinav, E. (2017). Dysbiosis and the immune system. *Nature Reviews Immunology*, **17**(4), 219.

Li, B. and Dewey, C. N. (2011). Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, **12**(1), 323.

Li, C. and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*, **98**(1), 31–36.

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

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, **25**(14), 1754–1760.

Liang, D., Leung, R. K.-K., Guan, W., and Au, W. W. (2018). Involvement of gut microbiome in human health and disease: brief overview, knowledge gaps and research opportunities. *Gut pathogens*, **10**(1), 3.

Liao, Y., Smyth, G. K., and Shi, W. (2014). featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**(7), 923–930.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (msigdb) 3.0. *Bioinformatics*, **27**(12), 1739–1740.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *Cell systems*, **1**(6), 417–425.

Lightbody, G., Haberland, V., Browne, F., Taggart, L., Zheng, H., Parkes, E., and Blayney, J. K. (2019). Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. *Briefings in bioinformatics*, **20**(5), 1795–1811.

Liu, H., Bebu, I., and Li, X. (2010). Microarray probes and probe sets. *Frontiers in bioscience (Elite edition)*, **2**, 325.

Liu, S. and Trapnell, C. (2016). Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Research*, **5**.

Lloyd-Price, J., Abu-Ali, G., and Huttenhower, C. (2016). The healthy human microbiome. *Genome medicine*, **8**(1), 1–11.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, **15**(12), 550.

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., and Shafee, T. (2017). Transcriptomics technologies. *PLoS computational biology*, **13**(5), e1005457.

Lu, J., Liu, J., Li, L., Lan, Y., and Liang, Y. (2020). Cytokines in type 1 diabetes: mechanisms of action and immunotherapeutic targets. *Clinical & Translational Immunology*, **9**(3), e1122.

Lun, A. T., Bach, K., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell rna sequencing data with many zero counts. *Genome biology*, **17**(1), 75.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., *et al.* (2012). Soapdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, **1**(1), 2047–217X.

Lynch, S. V. and Pedersen, O. (2016). The human intestinal microbiome in health and disease. *New England Journal of Medicine*, **375**(24), 2369–2379.

Lytal, N., Ran, D., and An, L. (2020). Normalization methods on single-cell rna-seq data: An empirical survey. *Frontiers in genetics*, **11**, 41.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, **9**(Nov), 2579–2605.

Mackay, C. R. (2020). Diet, the gut microbiome, and autoimmune diseases. In *The Autoimmune Diseases*, pages 331–342. Elsevier.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., *et al.* (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**(5), 1202–1214.

Mahajan, V. S., Mattoo, H., Deshpande, V., Pillai, S. S., and Stone, J. H. (2014). Igg4-related disease. *Annual Review of Pathology: Mechanisms of Disease*, **9**, 315–347.

Malla, M. A., Dubey, A., Kumar, A., Yadav, S., Hashem, A., and Abd_Allah, E. F. (2019). Exploring the human microbiome: The potential future role of next-generation sequencing in disease diagnosis and treatment. *Frontiers in Immunology*, **9**, 2868.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research*, **27**(2 Part 1), 209–220.

Manzel, A., Muller, D. N., Hafler, D. A., Erdman, S. E., Linker, R. A., and Kleinewietfeld, M. (2014). Role of "western diet" in inflammatory autoimmune diseases. *Current allergy and asthma reports*, **14**(1), 404.

Manzoni, C., Kia, D. A., Vandrovcova, J., Hardy, J., Wood, N. W., Lewis, P. A., and Ferrari, R. (2018). Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings in bioinformatics*, **19**(2), 286–302.

Mao, Q., Wang, L., Goodison, S., and Sun, Y. (2015). Dimensionality reduction via graph structure learning. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 765–774.

Marchesi, J. R. and Ravel, J. (2015). The vocabulary of microbiome research: a proposal. *Microbiome*, **3**(31), 1–3.

Marco, E., Karp, R. L., Guo, G., Robson, P., Hart, A. H., Trippa, L., and Yuan, G.-C. (2014). Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proceedings of the National Academy of Sciences*, **111**(52), E5643–E5650.

Mardis, E. R. (2017). Dna sequencing technologies: 2006–2016. *Nature protocols*, **12**(2), 213.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, **18**(9), 1509–1517.

Markowitz, V. M., Chen, I.-M. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J., Williams, P., *et al.* (2012). Img: the integrated microbial genomes database and comparative analysis system. *Nucleic acids research*, **40**(D1), D115–D122.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, **17**(1), 10–12.

Mason, K. L., Huffnagle, G. B., Noverr, M. C., and Kao, J. Y. (2008). Overview of gut immunology. In *GI microbiota and regulation of the immune system*, pages 1–14. Springer.

Mattoo, H., Mahajan, V. S., Maehara, T., Deshpande, V., Della-Torre, E., Wallace, Z. S., Kulikova, M., Drijvers, J. M., Daccache, J., Carruthers, M. N., *et al.* (2016). Clonal expansion of cd4+ cytotoxic t lymphocytes in patients with igg4-related disease. *Journal of Allergy and Clinical Immunology*, **138**(3), 825–838.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., *et al.* (2006). Transfac® and its module transcompel®: transcriptional gene regulation in eukaryotes. *Nucleic acids research*, **34**(suppl_1), D108–D110.

Matzaraki, V., Kumar, V., Wijmenga, C., and Zhernakova, A. (2017). The mhc locus and genetic susceptibility to autoimmune and infectious diseases. *Genome biology*, **18**(1), 76.

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic acids research*, **40**(10), 4288–4297.

McDavid, A., Finak, G., Chattopadyay, P. K., Dominguez, M., Lamoreaux, L., Ma, S. S., Roederer, M., and Gottardo, R. (2013). Data exploration, quality control and testing in single-cell qpcr-based gene expression experiments. *Bioinformatics*, **29**(4), 461–467.

McMahan, Z. H. and Hummers, L. K. (2013). Systemic sclerosis—challenges for clinical practice. *Nature Reviews Rheumatology*, **9**(2), 90.

Menche, J., Guney, E., Sharma, A., Branigan, P. J., Loza, M. J., Baribaud, F., Dobrin, R., and Barabási, A.-L. (2017). Integrating personalized gene expression profiles into predictive disease-associated gene pools. *NPJ systems biology and applications*, **3**(1), 1–10.

Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature reviews genetics*, **11**(1), 31–46.

Miettinen, M. E., Niinistö, S., Erlund, I., Cuthbertson, D., Nucci, A. M., Honkanen, J., Vaarala, O., Hyöty, H., Krischer, J. P., Knip, M., *et al.* (2020). Serum 25-hydroxyvitamin d concentration in childhood and risk of islet autoimmunity and type 1 diabetes: the trigr nested case–control ancillary study. *Diabetologia*, pages 1–8.

Milani, C., Duranti, S., Bottacini, F., Casey, E., Turroni, F., Mahony, J., Belzer, C., Palacio, S. D., Montes, S. A., Mancabelli, L., *et al.* (2017). The first microbial colonizers of the human gut: composition, activities, and health implications of the infant gut microbiota. *Microbiology and Molecular Biology Reviews*, **81**(4).

Miller, M. B. and Tang, Y.-W. (2009). Basic concepts of microarrays and potential applications in clinical microbiology. *Clinical microbiology reviews*, **22**(4), 611–633.

Milton, J. S. and Arnold, J. C. (2003). *Introduction to Probability and Statistics*. McGraw-Hill.

Mohammadkhah, A. I., Simpson, E. B., Patterson, S. G., and Ferguson, J. F. (2018). Development of the gut microbiome in children, and lifetime implications for obesity and cardiometabolic disease. *Children*, **5**(12), 160.

Molnar, C. and Gair, J. (2013). *Concepts of Biology: 1st Canadian Edition*. Rice University.

Morgan, X. C. and Huttenhower, C. (2012). Human microbiome analysis. *PLoS Comput Biol*, **8**(12), e1002808.

Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., Reyes, J. A., Shah, S. A., LeLeiko, N., Snapper, S. B., *et al.* (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome biology*, **13**(9), R79.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, **5**(7), 621–628.

Mou, T., Deng, W., Gu, F., Pawitan, Y., and Vu, T. N. (2020). Reproducibility of methods to detect differentially expressed genes from single-cell rna sequencing. *Frontiers in genetics*, **10**, 1331.

Murdaca, G., Tonacci, A., Negrini, S., Greco, M., Borro, M., Puppo, F., and Gangemi, S. (2019). Emerging role of vitamin d in autoimmune diseases: an update on evidence and therapeutic implications. *Autoimmunity reviews*, page 102350.

Murphy, K. M. and Weaver, C. (2017). *Janeway's immunobiology 9th Edition*. Garland Science, Taylor & Francis Group.

Navas-Molina, J. A., Peralta-Sánchez, J. M., González, A., McMurdie, P. J., Vázquez-Baeza, Y., Xu, Z., Ursell, L. K., Lauber, C., Zhou, H., Song, S. J., *et al.* (2013). Advancing our understanding of the human microbiome using qiime. In *Methods in enzymology*, volume 531, pages 371–444. Elsevier.

Nayfach, S. and Pollard, K. S. (2016). Toward accurate and quantitative comparative metagenomics. *Cell*, **166**(5), 1103–1116.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, **135**(3), 370–384.

Nemazee, D. (2017). Mechanisms of central tolerance for b cells. *Nature Reviews Immunology*, **17**(5), 281.

NHGRI (2020). Human genome project faq `https://www.genome.gov/human-genome-project/Completion-FAQ` (accessed: 18 july 2020).

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017). metaspades: a new versatile metagenomic assembler. *Genome research*, **27**(5), 824–834.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., and Wagner, H. (2019). *vegan: Community Ecology Package*. R package version 2.5-4.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., *et al.* (2016). Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, **44**(D1), D733–D745.

Opazo, M., Ortega-Rocha, E., Coronado-Arrázola, I., Bonifaz, L., Boudin, H., Neunlist, M., Bueno, S., *et al.* (2018). Intestinal microbiota influences non-intestinal related autoimmune diseases. front microbiol. 2018; 9: 432.

Parkin, J. and Cohen, B. (2001). An overview of the immune system. *The Lancet*, **357**(9270), 1777–1789.

Patro, R., Mount, S. M., and Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from rna-seq reads using lightweight algorithms. *Nature biotechnology*, **32**(5), 462–464.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, **14**(4), 417–419.

Patrone, V., Puglisi, E., Cardinali, M., Schnitzler, T. S., Svegliati, S., Festa, A., Gabrielli, A., and Morelli, L. (2017). Gut microbiota profile in systemic sclerosis patients with and without clinical evidence of gastrointestinal involvement. *Scientific reports*, **7**(1), 1–11.

Patterson, T. A., Lobenhofer, E. K., Fulmer-Smentek, S. B., Collins, P. J., Chu, T.-M., Bao, W., Fang, H., Kawasaki, E. S., Hager, J., Tikhonova, I. R., *et al.* (2006). Performance comparison of one-color and two-color platforms within the microarray quality control (maqc) project. *Nature biotechnology*, **24**(9), 1140–1150.

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature methods*, **10**(12), 1200–1202.

Peet, A., Kool, P., Ilonen, J., Knip, M., Tillmann, V., and Group, D. S. (2012). Birth weight in newborn infants with different diabetes-associated hla genotypes in three neighbouring countries: Finland, estonia and russian karelia. *Diabetes/metabolism research and reviews*, **28**(5), 455–461.

Pereira, M. B., Wallroth, M., Jonsson, V., and Kristiansson, E. (2018). Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC genomics*, **19**(1), 274.

Pérez-Cobas, A. E., Gomez-Valero, L., and Buchrieser, C. (2020). Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses. *Microbial Genomics*, **6**(8), e000409.

Perugino, C. A., AlSalem, S. B., Mattoo, H., Della-Torre, E., Mahajan, V., Ganesh, G., Allard-Chamard, H., Wallace, Z., Montesi, S. B., Kreuzer, J., *et al.* (2019). Identification of galectin-3 as an autoantigen in patients with igg4-related disease. *Journal of Allergy and Clinical Immunology*, **143**(2), 736–745.

Picard-toolkit (2019). Picard toolkit (broad institute, github repository). `http://broadinstitute.github.io/picard/`.

Pickard, J. M., Zeng, M. Y., Caruso, R., and Núñez, G. (2017). Gut microbiota: Role in pathogen colonization, immune responses, and inflammatory disease. *Immunological reviews*, **279**(1), 70–89.

Pinheiro, J. and Bates, D. (2006). *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2018). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-137.

Plaza Oñate, F., Le Chatelier, E., Almeida, M., Cervino, A. C., Gauthier, F., Magoulès, F., Ehrlich, S. D., and Pichaud, M. (2019). Mspminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. *Bioinformatics*, **35**(9), 1544–1552.

Pociot, F. and Lernmark, Å. (2016). Genetic risk factors for type 1 diabetes. *The Lancet*, **387**(10035), 2331–2339.

Podojil, J. R. and Miller, S. D. (2009). Molecular mechanisms of t-cell receptor and costimulatory molecule ligation/blockade in autoimmune disease therapy. *Immunological reviews*, **229**(1), 337–355.

Poirion, O. B., Zhu, X., Ching, T., and Garmire, L. (2016). Single-cell transcriptomics bioinformatics and computational challenges. *Frontiers in genetics*, **7**, 163.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nature methods*, **14**(10), 979.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature genetics*, **32**(4), 496–501.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F. O. (2012). The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic acids research*, **41**(D1), D590–D596.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rackaityte, E., Halkias, J., Fukui, E., Mendoza, V., Hayzelden, C., Crawford, E., Fujimura, K., Burt, T., and Lynch, S. (2020). Viable bacterial colonization is highly limited in the human intestine in utero. *Nature medicine*, **26**(4), 599–607.

Raj, A. and Van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, **135**(2), 216–226.

Rasmussen, C. and Williams, C. (2006). *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA.

Rasmussen, C. E. (2003). Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer.

Rebeca, I.-M., Lucia, M.-R., Concepción, R., and Víctor, J. C.-R. (2019). Vitamin d and autoimmune diseases. *Life sciences*, page 116744.

Regnell, S. E. and Lernmark, Å. (2017). Early prediction of autoimmune (type 1) diabetes. *Diabetologia*, **60**(8), 1370–1381.

Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C., *et al.* (2019). Pathway enrichment analysis and visualization of omics data using g: Profiler, gsea, cytoscape and enrichmentmap. *Nature protocols*, **14**(2), 482–517.

Rencher, A. C. and Schaalje, G. B. (2008). *Linear models in statistics*. John Wiley & Sons.

Reuter, J. A., Spacek, D. V., and Snyder, M. P. (2015). High-throughput sequencing technologies. *Molecular cell*, **58**(4), 586–597.

Rewers, M. and Ludvigsson, J. (2016). Environmental risk factors for type 1 diabetes. *The Lancet*, **387**(10035), 2340–2348.

Rideout, J. R., He, Y., Navas-Molina, J. A., Walters, W. A., Ursell, L. K., Gibbons, S. M., Chase, J., McDonald, D., Gonzalez, A., Robbins-Pianka, A., *et al.* (2014). Subsampled open-reference clustering creates consistent, comprehensive otu definitions and scales to billions of sequences. *PeerJ*, **2**, e545.

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N., and Aigrain, S. (2013). Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **371**(1984), 20110550.

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, **11**(3), 1–9.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139–140.

Rodriguez, A. and Laio, A. (2014). Clustering by fast search and find of density peaks. *Science*, **344**(6191), 1492–1496.

Rose, N. R. (2016). Prediction and prevention of autoimmune disease in the 21st century: a review and preview. *American journal of epidemiology*, **183**(5), 403–406.

Rosenblum, M. D., Remedios, K. A., and Abbas, A. K. (2015). Mechanisms of human autoimmunity. *The Journal of clinical investigation*, **125**(6), 2228–2233.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series B (Statistical Methodology)*, **71**(2), 319–392.

Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nature biotechnology*, **37**(5), 547–554.

Sakkas, L. I. and Bogdanos, D. P. (2016). Systemic sclerosis: new evidence re-enforces the role of b cells. *Autoimmunity reviews*, **15**(2), 155–161.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, **74**(12), 5463–5467.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, **33**(5), 495–502.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, **78**(4), 719–727.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, **270**(5235), 467–470.

Schloss, P. D. and Handelsman, J. (2005). Introducing dotur, a computer program for defining operational taxonomic units and estimating species richness. *Applied and environmental microbiology*, **71**(3), 1501–1506.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., *et al.* (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, **75**(23), 7537–7541.

Segal, E., Friedman, N., Koller, D., and Regev, A. (2004). A module map showing conditional activity of expression modules in cancer. *Nature genetics*, **36**(10), 1090–1098.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, **9**(8), 811–814.

Segata, N., Börnigen, D., Morgan, X. C., and Huttenhower, C. (2013). Phylophlan is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature communications*, **4**(1), 1–11.

Sender, R., Fuchs, S., and Milo, R. (2016). Are we really vastly outnumbered? revisiting the ratio of bacterial to host cells in humans. *Cell*, **164**(3), 337–340.

Setty, M., Tadmor, M. D., Reich-Zeliger, S., Angel, O., Salame, T. M., Kathail, P., Choi, K., Bendall, S., Friedman, N., and Pe'er, D. (2016). Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature biotechnology*, **34**(6), 637–645.

Shalon, D., Smith, S. J., and Brown, P. O. (1996). A dna microarray system for analyzing complex dna samples using two-color fluorescent probe hybridization. *Genome research*, **6**(7), 639–645.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, **27**(3), 379–423.

Sharpe, A. H. (2009). Mechanisms of costimulation. *Immunological reviews*, **229**(1), 5–11.

Shiu, S. and Borevitz, J. (2008). The next generation of microarray research: applications in evolutionary and ecological genomics. *Heredity*, **100**(2), 141–149.

Simon, A. K., Hollander, G. A., and McMichael, A. (2015). Evolution of the immune system in humans from infancy to old age. *Proceedings of the Royal Society B: Biological Sciences*, **282**(1821), 20143085.

Sobek, J., Bartscherer, K., Jacob, A., Hoheisel, J. D., and Angenendt, P. (2006). Microarray technology as a universal tool for high-throughput analysis of biological systems. *Combinatorial chemistry & high throughput screening*, **9**(5), 365–380.

Soneson, C. and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of rna-seq data. *BMC bioinformatics*, **14**(1), 91.

Steen, V. D. (2005). Autoantibodies in systemic sclerosis. In *Seminars in arthritis and rheumatism*, volume 35, pages 35–42. Elsevier.

Stegle, O., Denby, K. J., Cooke, E. J., Wild, D. L., Ghahramani, Z., and Borgwardt, K. M. (2010). A robust bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. *Journal of Computational Biology*, **17**(3), 355–367.

Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, **16**(3), 133–145.

Steinke, J. W. and Borish, L. (2006). 3. cytokines and chemokines. *Journal of Allergy and Clinical Immunology*, **117**(2), S441–S445.

Stone, J. H., Zen, Y., and Deshpande, V. (2012). Igg4-related disease. *New England Journal of Medicine*, **366**(6), 539–551.

Stoughton, R. B. (2005). Applications of dna microarrays in biology. *Annu. Rev. Biochem.*, **74**, 53–82.

Strachan, D. P. (1989). Hay fever, hygiene, and household size. *BMJ: British Medical Journal*, **299**(6710), 1259.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, **102**(43), 15545–15550.

Subramanian, S., Huq, S., Yatsunenko, T., Haque, R., Mahfuz, M., Alam, M. A., Benezra, A., DeStefano, J., Meier, M. F., Muegge, B. D., *et al.* (2014). Persistent gut microbiota immaturity in malnourished bangladeshi children. *Nature*, **510**(7505), 417–421.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and Consortium, U. (2015). Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**(6), 926–932.

Takiishi, T., Fenero, C. I. M., and Câmara, N. O. S. (2017). Intestinal barrier and gut microbiota: shaping our immune responses throughout life. *Tissue Barriers*, **5**(4), e1373208.

Tanaka, M. and Nakayama, J. (2017). Development of the gut microbiota in infancy and its impact on health in later life. *Allergology International*, **66**(4), 515–522.

Tarazona, S., García, F., Ferrer, A., Dopazo, J., and Conesa, A. (2011). Noiseq: a rna-seq differential expression method robust for sequencing depth biases. *EMBnet. journal*, **17**(B), 18–19.

Theofilopoulos, A. N., Kono, D. H., and Baccala, R. (2017). The multiple pathways to autoimmunity. *Nature immunology*, **18**(7), 716.

Tibbs, T. N., Lopez, L. R., and Arthur, J. C. (2019). The influence of the microbiota on immune development, chronic inflammation, and cancer in the context of aging. *Microbial Cell*, **6**(8), 324.

Timonen, J., Mannerström, H., Vehtari, A., and Lähdesmäki, H. (2021). lgpr: an interpretable non-parametric method for inferring covariate effects from longitudinal data. *Bioinformatics*.

Tobón, G. J., Izquierdo, J. H., and Cañas, C. A. (2013). B lymphocytes: development, tolerance, and their role in autoimmunity—focus on systemic lupus erythematosus. *Autoimmune diseases*, **2013**.

Torgerson, W. S. (1952). Multidimensional scaling: I. theory and method. *Psychometrika*, **17**(4), 401–419.

Tortora, G. J. and Derrickson, B. (2013). *Principles of Anatomy and Physiology 14th Edition*. John Wiley & Sons.

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. *Genome research*, **25**(10), 1491–1498.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, **25**(9), 1105–1111.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, **28**(5), 511–515.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, **32**(4), 381.

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., and Segata, N. (2015). Metaphlan2 for enhanced metagenomic taxonomic profiling. *Nature methods*, **12**(10), 902–903.

Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C., and Segata, N. (2017). Microbial strain-level population structure and genetic diversity from metagenomes. *Genome research*, **27**(4), 626–638.

Tuomilehto, J. (2013). The emerging global epidemic of type 1 diabetes. *Current diabetes reports*, **13**(6), 795–804.

Vallejos, C. A., Richardson, S., and Marioni, J. C. (2016). Beyond comparisons of means: understanding changes in gene expression at the single-cell level. *Genome biology*, **17**(1), 1–14.

Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J. C. (2017). Normalizing single-cell rna sequencing data: challenges and opportunities. *Nature methods*, **14**(6), 565.

Van der Maaten, L. and Hinton, G. (2012). Visualizing non-metric similarities in multiple maps. *Machine learning*, **87**(1), 33–55.

Vanhatalo, J., Pietiläinen, V., and Vehtari, A. (2010). Approximate inference for disease mapping with sparse gaussian processes. *Statistics in medicine*, **29**(15), 1580–1607.

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., and Stuart, J. M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics*, **26**(12), i237–i245.

Vatanen, T., Kostic, A. D., d'Hennezel, E., Siljander, H., Franzosa, E. A., Yassour, M., Kolde, R., Vlamakis, H., Arthur, T. D., Hämäläinen, A.-M., *et al.* (2016). Variation in microbiome lps immunogenicity contributes to autoimmunity in humans. *Cell*, **165**(4), 842–853.

Vatanen, T., Franzosa, E. A., Schwager, R., Tripathi, S., Arthur, T. D., Vehik, K., Lernmark, Å., Hagopian, W. A., Rewers, M. J., She, J.-X., *et al.* (2018). The human gut microbiome in early-onset type 1 diabetes from the teddy study. *Nature*, **562**(7728), 589–594.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Ventimiglia, G. and Petralia, S. (2013). Recent advances in dna microarray technology: an overview on production strategies and detection methods. *BioNanoScience*, **3**(4), 428–450.

Vieira, S. M., Pagovich, O. E., and Kriegel, M. A. (2014). Diet, microbiota and autoimmune diseases. *Lupus*, **23**(6), 518–526.

Vieth, B., Parekh, S., Ziegenhain, C., Enard, W., and Hellmann, I. (2019). A systematic evaluation of single cell rna-seq analysis pipelines. *Nature communications*, **10**(1), 1–11.

Vojdani, A. (2008). Antibodies as predictors of complex autoimmune diseases. *International journal of immunopathology and pharmacology*, **21**(2), 267–278.

Vojdani, A. (2014). A potential link between environmental triggers and autoimmunity. *Autoimmune diseases*, **2014**.

Völkl, S. (2019). Human double-negative regulatory t-cells induce a metabolic and functional switch in effector t-cells by suppressing mtor activity. *Frontiers in immunology*, **10**, 883.

Wagner, A., Regev, A., and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. *Nature biotechnology*, **34**(11), 1145–1160.

Walker, L. S. and Abbas, A. K. (2002). The enemy within: keeping self-reactive t cells at bay in the periphery. *Nature Reviews Immunology*, **2**(1), 11–19.

Wang, L., Wang, F.-S., and Gershwin, M. E. (2015). Human autoimmune diseases: a comprehensive update. *Journal of internal medicine*, **278**(4), 369–395.

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, **73**(16), 5261–5267.

Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, **10**(1), 57–63.

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., *et al.* (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, **5**(1), 27.

West, C. E., Renz, H., Jenmalm, M. C., Kozyrskyj, A. L., Allen, K. J., Vuillermin, P., Prescott, S. L., MacKay, C., Salminen, S., Wong, G., *et al.* (2015). The gut microbiota and inflammatory noncommunicable diseases: associations and potentials for gut microbiota therapies. *Journal of Allergy and Clinical Immunology*, **135**(1), 3–13.

Westbrook, A., Ramsdell, J., Schuelke, T., Normington, L., Bergeron, R. D., Thomas, W. K., and MacManes, M. D. (2017). Paladin: protein alignment for functional profiling whole metagenome shotgun data. *Bioinformatics*, **33**(10), 1473–1478.

Wilhelm, B. T. and Landry, J.-R. (2009). Rna-seq—quantitative measurement of expression through massively parallel rna-sequencing. *Methods*, **48**(3), 249–257.

Wood, D. E. and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, **15**(3), 1–12.

Wright, E. S., Yilmaz, L. S., and Noguera, D. R. (2012). Decipher, a search-based approach to chimera identification for 16s rrna sequences. *Applied and environmental microbiology*, **78**(3), 717–725.

Wu, G.-C., Pan, H.-F., Leng, R.-X., Wang, D.-G., Li, X.-P., Li, X.-M., and Ye, D.-Q. (2015). Emerging role of long noncoding rnas in autoimmune diseases. *Autoimmunity reviews*, **14**(9), 798–805.

Wu, Y.-W., Simmons, B. A., and Singer, S. W. (2016). Maxbin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, **32**(4), 605–607.

Wu, Z. (2009). A review of statistical methods for preprocessing oligonucleotide microarrays. *Statistical methods in medical research*, **18**(6), 533–541.

Wu, Z. and Irizarry, R. A. (2004). Preprocessing of oligonucleotide array data. *Nature biotechnology*, **22**(6), 656–658.

Xing, Y. and Hogquist, K. A. (2012). T-cell tolerance: central and peripheral. *Cold Spring Harbor perspectives in biology*, **4**(6), a006957.

Xu, F., Jin, L., Jin, Y., Nie, Z., and Zheng, H. (2019). Long noncoding rnas in autoimmune diseases. *Journal of biomedical materials research Part A*, **107**(2), 468–475.

Yadav, S. P. (2007). The wholeness in suffix-omics,-omes, and the word om. *Journal of biomolecular techniques: JBT*, **18**(5), 277.

Yamamoto, M., Takahashi, H., and Shinomura, Y. (2014). Mechanisms and assessment of igg4-related disease: lessons for the rheumatologist. *Nature Reviews Rheumatology*, **10**(3), 148.

Yang, C.-Y., Leung, P. S., Adamopoulos, I. E., and Gershwin, M. E. (2013). The implication of vitamin d and autoimmunity: a comprehensive review. *Clinical reviews in allergy & immunology*, **45**(2), 217–226.

Yang, I. S. and Kim, S. (2015). Analysis of whole transcriptome sequencing data: workflow and software. *Genomics & informatics*, **13**(4), 119.

Yang, J., Penfold, C. A., Grant, M. R., and Rattray, M. (2016). Inferring the perturbation time from biological time course data. *Bioinformatics*, **32**(19), 2956–2964.

Yassour, M., Vatanen, T., Siljander, H., Hämäläinen, A.-M., Härkönen, T., Ryhänen, S. J., Franzosa, E. A., Vlamakis, H., Huttenhower, C., Gevers, D., *et al.* (2016). Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Science translational medicine*, **8**(343), 343ra81–343ra81.

Yassour, M., Jason, E., Hogstrom, L. J., Arthur, T. D., Tripathi, S., Siljander, H., Selvenius, J., Oikarinen, S., Hyöty, H., Virtanen, S. M., *et al.* (2018). Strain-level analysis of mother-to-child bacterial transmission during the first few months of life. *Cell host & microbe*, **24**(1), 146–154.

Ye, J., McGinnis, S., and Madden, T. L. (2006). Blast: improvements for better sequence analysis. *Nucleic acids research*, **34**(suppl_2), W6–W9.

Yu, C., Woo, H. J., Yu, X., Oyama, T., Wallqvist, A., and Reifman, J. (2017). A strategy for evaluating pathway analysis methods. *BMC bioinformatics*, **18**(1), 1–11.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). Pear: a fast and accurate illumina paired-end read merger. *Bioinformatics*, **30**(5), 614–620.

Zhang, X., Li, T., Liu, F., Chen, Y., Yao, J., Li, Z., Huang, Y., and Wang, J. (2019). Comparative analysis of droplet-based ultra-high-throughput single-cell rna-seq systems. *Molecular cell*, **73**(1), 130–142.

Zhao, Q. and Elson, C. O. (2018). Adaptive immune education by gut microbiota antigens. *Immunology*, **154**(1), 28–37.

Zhao, S. (2014). Assessment of the impact of using a reference transcriptome in mapping short rna-seq reads. *PLoS One*, **9**(7), e101374.

Zhao, S., Xi, L., and Zhang, B. (2015). Union exon based approach for rna-seq gene quantification: To be or not to be? *PLoS One*, **10**(11), e0141910.

Zhao, S., Zhang, B., Zhang, Y., Gordon, W., Du, S., Paradis, T., Vincent, M., and von Schack, D. (2016). Bioinformatics for rna-seq data analysis. *Bioinformatics—Updated Features and Applications: InTech*, pages 125–49.

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., *et al.* (2017). Massively parallel digital transcriptional profiling of single cells. *Nature communications*, **8**(1), 1–12.

Zhou, Q., Su, X., and Ning, K. (2014). Assessment of quality control approaches for metagenomic data analysis. *Scientific reports*, **4**(1), 1–11.

Zhou, Y.-H., Xia, K., and Wright, F. A. (2011). A powerful and flexible approach to the analysis of rna sequence count data. *Bioinformatics*, **27**(19), 2672–2678.

Zhu, B., Wang, X., and Li, L. (2010). Human gut microbiome: the second genome of human body. *Protein & cell*, **1**(8), 718–725.

Zhuang, L., Chen, H., Zhang, S., Zhuang, J., Li, Q., and Feng, Z. (2019). Intestinal microbiota in early life and its implications on childhood health. *Genomics, proteomics & bioinformatics*, **17**(1), 13–25.

Ziegler, A. G., Rewers, M., Simell, O., Simell, T., Lempainen, J., Steck, A., Winkler, C., Ilonen, J., Veijola, R., Knip, M., *et al.* (2013). Seroconversion to multiple islet autoantibodies and risk of progression to diabetes in children. *Jama*, **309**(23), 2473–2479.

Zou, S., He, H.-J., Zong, Y., Shi, L., and Wang, L. (2008). Dna microarrays: applications, future trends, and the need for standardization. In *Standardization and quality assurance in fluorescence measurements II*, pages 215–237. Springer.

Aalto-DD 90/2021

ISBN 978-952-64-0435-6 (printed)
ISBN 978-952-64-0436-3 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Computer Science
www.aalto.fi