Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Bias Invariant RNA-Seq Data Annotation and Liver Diseases Microbiome Analysis

Dissertation
zur Erlangung des Doktorgrades
an der Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Chemie
der Universität Hamburg

Hannes Wartmann

hanneswartmann@gmail.com
Matr.-Nr. 7194963

Erstgutachter: Professor Dr. Stefan Bonn
Zweitgutachter: Professor Dr. Andrew Torda

Abgabe: 03.2021
Verteidigung: 21.05.2021

Du sollst der werden, der du bist.
– *F. Nietzsche*

Die vorgelegte Arbeit wurde von Mai 2018 bis März 2021 am Institut für Medizinische Systembiologie am Zentrum für Molekulare Neurobiologie am Universitätsklinikum Hamburg-Eppendorf unter Anleitung von Herr Prof. Dr. Stefan Bonn angefertigt.

List of Publications

Wartmann, H., Heins, S., Kloiber, K. and Bonn, S. Bias invariant RNA-seq metadata annotation. *bioRxiv*. 2020

Contents

Acronyms

ANN artifical neural network.
ASV amplicon sequence variant.

BC Bray-Curtis distance.

CD14 cluster of differentiation 14.
cDNA complementary deoxyribonucleic acid.

DA domain adaptation.
DNA deoxyribonucleic acid.

GTEx Genotype-Tissue Expression Project.

HC healthy control.

IBD inflammatory bowel disease.

LFC log fold change.
LIN linear model.
LPB lipopolysaccharide-binding protein.
LPS lipopolysaccharides.

MCA mean class accuracy.
ML machine learning.
MLP multilayer perceptron.
mRNA messenger RNA.
MSA mean sample accuracy.
MSE mean squared error.

NTC no template control.

PBC primary biliary cholangitis.
PCA principal component analysis.
ppt percent point.
PSC primary sclerosing cholangitis.

RNA-seq ribosomal nucleic acid sequencing.

SRA Sequence Read Archive.
sRNA small RNA.

TCGA The Cancer Genome Atlas.

List of Figures

List of Tables

1. Abstract / Zusammenfasung

1.1. Abstract

Next-generation sequencing has become so cheap and ubiquitous that the amount of publicly available data is growing exponentially. Large public repositories have been created for RNA-seq data to facilitate data sharing and reusability. However, most of these datasets lack standardized metadata annotation, such as the sampled tissue and basic patient information such as sex. As public databases and the use of sequencing technologies grow, proper annotation will become increasingly important to make data truly searchable and findable for everyone.

Sequencing technology can be applied to measure a diverse range of signals. For example, targeted sequencing of the 16s ribosomal gene of prokaryotes can be used to measure the microbiome of a given environment. The human microbiome is well known to be linked with diseases and analysing it can lead to novel biomedical and diagnostic discovery.

This thesis presents two projects working with sequencing data. The first project is concerned with developing a machine learning algorithm. The aim is to annotate public RNA-seq samples with metadata automatically. The developed algorithm successfully predicts the tissue of origin, the source of the sample and the sex of the patient more accurately than a previously published linear model and a standard neural network. We were able to generate more than 10,000 novel metadata entries for 8,495 publicly available RNA-seq samples. The second project is concerned with analyzing microbial sequencing data (16s rRNA-seq) of blood serum samples. This project compares patients' blood microbiome between primary sclerosing cholangitis, primary biliary cholangitis, and healthy controls. It is well known that these two progressive liver diseases can be linked to a change in the microbial composition of the gut. We were able to confirm this finding in the blood. The liver disease patients showed an increased within-sample diversity compared to the healthy controls, thereby supporting a current hypothesis of heightened intestinal permeability and microbial translocation into the blood as a potential pathway of disease development.

1.2. Zusammenfassung

Die Erbgut-Sequenzierung ist so kostengünstig und allgegenwärtig geworden, dass die Menge der öffentlich verfügbaren Daten exponentiell ansteigt. Es wurden große öffentliche Datenbanken für RNA-Seq-Daten eingerichtet, um den Datenaustausch und die Wiederverwendbarkeit zu erleichtern. Den meisten dieser Datensätze fehlen jedoch standardisierte Metadaten-

Annotationen, wie z. B. das entnommene Gewebe und grundlegende Patienteninformationen wie das Geschlecht. Da öffentliche Datenbanken und der Einsatz von Sequenziertechnologien wachsen und zunehmen, wird eine korrekte Annotation immer wichtiger, um Daten wirklich suchbar und für jeden auffindbar zu machen.

Die Sequenzierungstechnologie kann zur Messung einer Vielzahl von Signalen eingesetzt werden. Zum Beispiel kann die gezielte Sequenzierung des 16s ribosomalen Gens von Prokaryoten verwendet werden, um das Mikrobiom einer bestimmten Umgebung zu messen. Es ist bekannt, dass das menschliche Mikrobiom mit Krankheiten in Verbindung steht und seine Analyse kann zu neuen biomedizinischen und diagnostischen Entdeckungen führen.

Dieser Arbeit stelle zwei Projekte vor, die mit Sequenzierungsdaten arbeiten. Das erste Projekt befasst sich mit der Entwicklung eines maschinellen Lernalgorithmus. Dieser soll öffentliche RNA-Seq-Proben automatisch mit Metadaten annotieren. Der entwickelte Algorithmus sagt erfolgreich das Herkunftsgewebe, die Quelle der Probe und das Geschlecht des Patienten genauer voraus als ein zuvor veröffentlichtes lineares Modell und ein standardmäßiges neuronales Netzwerk. Wir konnte mehr als 10.000 neue Metadateneinträge für 8.495 öffentlich verfügbare RNA-Seq-Proben generieren. Das zweite Projekt befasst sich mit der Analyse von mikrobiellen Sequenzierdaten (16s rRNA-seq) von Blutserumproben. Dieses Projekt vergleicht das Blutmikrobiom von Patienten mit primär sklerosierender Cholangitis, primär biliärer Cholangitis und gesunden Kontrollen. Es ist bekannt, dass diese beiden Lebererkrankungen mit einer Veränderung der mikrobiellen Zusammensetzung des Darms in Verbindung gebracht werden können. Diese Erkentniss konnte im Blut bestätigt werden. Die Patienten mit der Lebererkrankung zeigten eine erhöhte Diversität innerhalb der Probe im Vergleich zu den gesunden Kontrollen, wodurch eine aktuelle Hypothese der erhöhten intestinalen Permeabilität und der mikrobiellen Translokation ins Blut als möglicher Weg der Krankheitsentwicklung unterstützt wird.

2. Bias Invariant RNA-Seq Metadata Annotation

2.1. Statement of Contribution

This chapter is the result of a collaboration between Hannes Wartmann and Sven Heins. Sven kindly took the responsibility of reproducing the results published in [1], generated all results reported for the LIN model and wrote the method section *Linear Regression Model - LIN*.

2.2. Background

2.2.1. RNA-Sequencing

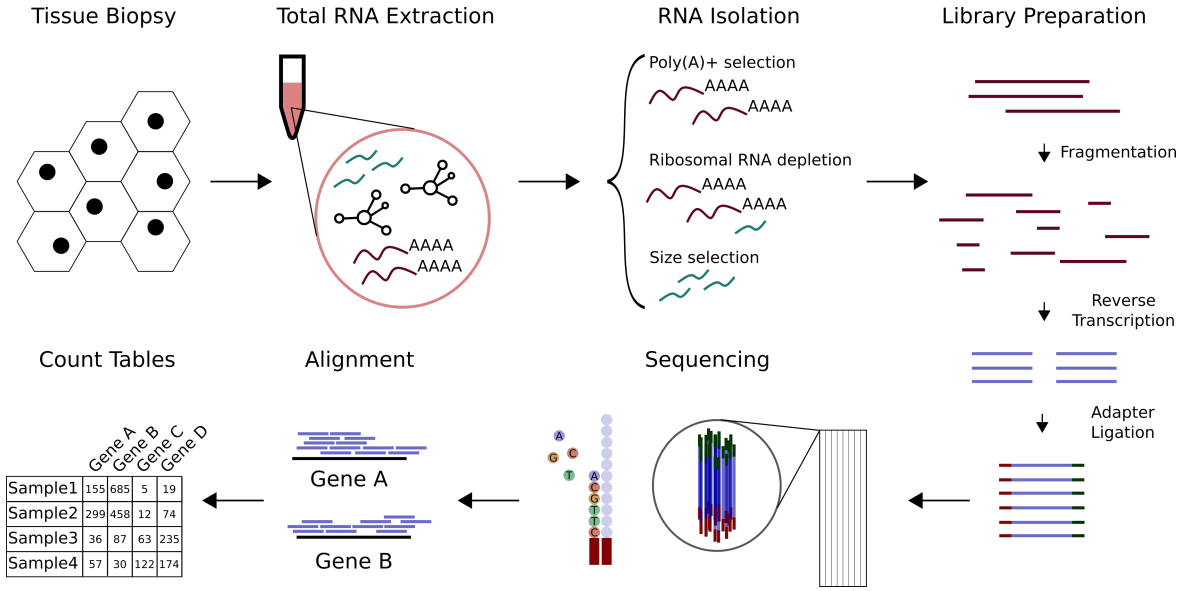

Figure 2.1.: RNA-sequencing workflow. The process of RNA-seq for gene expression measurement follows these steps: (i) cells are collected (e.g., from tissue biopsy), (ii) cells are lysed, total RNA is extracted, DNA, proteins, and other cell debris are removed, (iii) depending on the study design different RNA species are enriched or depleted, (iv) RNA library preparation; RNA is fragmented and reverse transcribed into the more stable cDNA, adapters are ligated, (v) next-generation sequencing; cDNA fragments are placed onto a flow cell, fragments are amplified to form clusters for signal amplification, one nucleotide at a time is added to each cluster emitting a light signal representing the fragment's sequence, (vi) sequenced fragments (reads) are aligned to a reference genome and (vii) gene count tables generated.

Ribosomal nucleic acid sequencing (RNA-seq) is a technology to measure the quantity of RNA in a biological sample at a given moment. It allows researchers to simultaneously quantify and compare the expression of tens of thousands of genomic transcripts. The workflow of RNA-seq and downstream analysis are briefly summarized in Figure 1 [2]. RNA-seq raw data can, for example, be used to discover novel exons, splicing events or information about non-protein-coding RNA species [3]. However, the most straightforward application is to measure gene expression (Figure 2.1). Gene expression count tables are generated from RNA-seq experiments focusing on messenger RNA (mRNA). This technique has helped, for example, establish the

human transcriptome and regulatory effects across healthy tissue and individuals [4, 5], to define a comprehensive transcriptional portrait of human cancer cell lines [6], to link primary sclerosing cholangitis to pro-inflammatory signaling [7], to identify the molecular etiology of Parkinson's disease [8], or to develop blood-based pan-cancer diagnostics [9].

A continuous drop in cost has made RNA-seq a widely available method of choice to uncover the molecular basis of biological development or disease [3, 10]. Between 2010 and 2015, the worldwide annual sequencing capacity has doubled every seven months [11]. As a result, recent years have seen substantial growth in publications and publicly accessible data linked to RNA-seq (Figure 2.2).

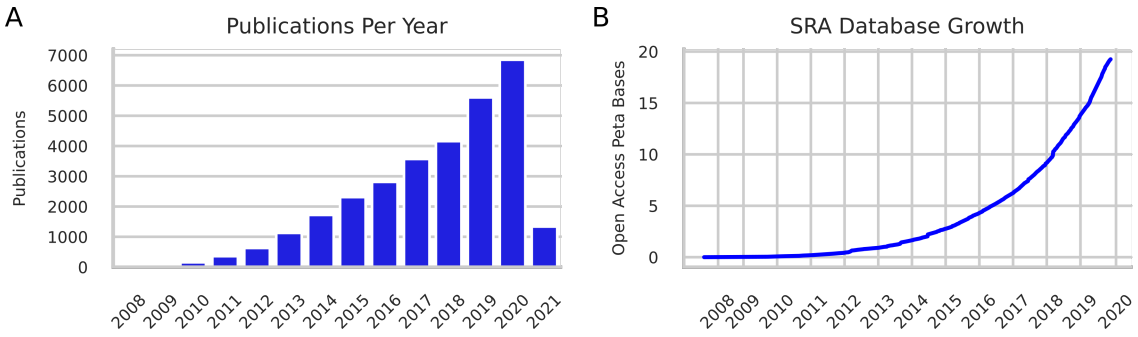

Figure 2.2.: Publications and publicly available data linked to RNA-seq have seen substantial growth in the past decade. A) The number of publications mentioning "RNA-Seq" archived in PubMed has been growing steadily to almost 7000 publications in the year 2020. B) The Sequence Read Archive (SRA) is the largest repository for raw sequencing data. Since its initiation in 2008, the number of data, measured in peta bases, has grown exponentially and is predicted to keep growing at a similar rate.

2.2.2. Meta-Analysis and Big Data

Meta-analysis is the synthesis of multiple research studies using statistical methods. The first medical meta-analysis was published in 1904 [12], and the number of publications has been growing exponentially in the past 30 years [13], reaching 29,317 publications on PubMed in 2020. Meta-analysis can be performed when multiple datasets are attempting to measure the same signal. Integrating datasets from different studies (and sources) can increase statistical power, analyze differences in results between studies or generate new hypotheses [14]. In genomics, meta-analysis has been a well-established method for many years [15, 16]. For example, publicly available RNA-seq datasets have been pooled to study gene expression across species and tissues [17], to find novel genes associated with dilated cardiomyopathy [18], or increasing statistical power to identify robust transcriptomic changes specific to Alzheimer's disease [19].

At least in biology, big data is essentially meta-analysis using a much larger number of studies and applying machine learning (ML) to infer new knowledge [20]. Genomics is predicted to produce 2-40 exabytes (1 exabyte = 10^6 terabytes) of data each year by 2025 [11], which

would make it the most extensive data-driven science by far [11]. Research done on hundreds of datasets is already commonplace today. For example, Taroni et al. [21] used a sizable public gene expression compendia to train ML models, which they then applied on smaller, rare disease datasets. Similarly, Tan et al. [22] extracted gene pathway information from a gene expression compendia of *P. aeruginosa* containing more than 125 datasets. Others have integrated hundreds of RNA-seq datasets to generate novel insights into mutually exclusive splicing [23] or to identify key splicing factors in Rett syndrome and cold-induced thermogenesis [24]. One challenge that all examples in this section have in common is the dependency on proper data management standards within the scientific community.

2.2.3. Sequence Data Reusability

Data reusability describes the ease of utilization of published data. Good data management will allow us to reproduce and verify results, minimize duplication effort, and build on others' work [25]. In 2015 Wilkson et al. [26] proposed the FAIR Principles for scientific data management and stewardship to standardize data handling across scientific fields. The authors postulated a code of conduct for making data findable, accessible, interoperable and reusable. In genomic science several projects have been started in the past decade to centralize data storage (accessibility), standardize data processing (interoperability) and homogenize and impute new metadata (findability).

The Sequence Read Archive (SRA) [27] is the National Center for Biotechnology Information's primary high-throughput sequencing repository. The SRA has grown exponentially and currently holds 20 peta bases of information (Figure 2.2B). As of the time of writing, about 10% of the data (95,000 samples belonging to 3,000 studies) is human RNA-seq data. A strict hierarchical order (projects SRP#, samples SRS#, experiment SRX# and sequencing runs SRR#) is maintained using unique identifiers for each identity. The SRA plays an integral part in data reusability by providing a single repository for all raw sequencing datasets and each sample's unique identifiability. However, another critical part is standardized metadata annotation to make data search- and findable. The repository is user-submission based, and currently, no standardized terms for metadata annotation are enforced. As a result, metadata, if available, includes many synonyms, spelling variants and different levels of granularity [28, 1] (Supplementary Table A.1).

MetaSRA [28] is a project aiming to homogenize the currently available SRA metadata. To this end, the metadata of 75,000 human SRA samples was downloaded. Metadata entries related to disease state and anatomy were mapped to ontologies to link and normalize manually entered entities. Sample source (e.g., tissue, cell line, stem cell) were predicted using ML with ontological information as input features. The generated metadata is freely available and continuously updated and expanded.

Recount2 [29] is a project established in 2011 intending to provide analysis-ready RNA-seq datasets. While the SRA collects raw sequencing files and makes them available for download, recount2 provides gene and exon count tables for more than 50'000 human SRA samples. Recount2 serves the community in two significant ways: (i) Raw SRA datasets are processed into analysis-ready gene and exon count tables that are easily accessible, and (ii) SRA datasets are processed using a single pipeline (Rail-RNA [30]) and therefore minimize the dataset bias originating from the use of different pipelines.

Phenopredict [1] is an R package for phenotype prediction of human RNA-seq samples. Publicly available gene expression data (i.e. GTEx, see section 2.2.4) with standardized metadata annotation (e.g., sex, tissue, sample type) was used to train a linear classifier. The classifier was then validated on TCGA and SRA samples downloaded from recount2 and can be used for automatic phenotype annotation. The algorithm is split into a predictor building and a prediction phase. First, expressed regions (ERs) [31] were determined resulting in more than 1 million potential input features. Next, a linear regression model (see section 2.2.6) was fit to determine coefficients (including an intercept term for base expression) relating categorical phenotype classes (independent variable) to the ERs (dependent variable). Next, the most discriminative ERs were selected as features for the model. A linear model (no intercept) relating phenotype classes to preselected ERs was fit to estimate the mean expression level for each ER across all classes. Given new ER expression values (outcome variable) and the estimated mean expression (coefficients) for each region, the model can be solved for the independent variable (class predictions).

2.2.4. Large Public Datasets

The **Genotype-Tissue Expression (GTEx)** [32] project is the largest public effort to study tissue-specific gene expression [4, 5]. GTEx strives to build a highly homogeneous dataset with strict guidelines on donor selection, biopsy and sequencing methodology. For example, dataset biases resulting from different laboratory methods, RNA extraction kits or sequencing technologies were minimized by stringent standards and centralized sequencing facilities. In the current version (v8), 17,382 samples from 948 healthy (albeit dead) donors covering 54 tissues are available.

The Cancer Genome Atlas (TCGA) is a project aiming to collect (among others) genomic sequencing information from a large variety of cancer tissues. TCGA depends on a submission model of biospecimen from all over the US. A high level of quality and standardization is ensured by centralized quality control and sequencing in few facilities. The TCGA provides raw sequencing files for 11,284 donors across 26 tissue types.

2.2.5. Dataset Bias

Dataset bias is a shift in the distribution between two datasets attempting to measure the same signal (Figure 2.3). For example, an algorithm trained on images (see section 2.2.6 for more details) of white cats might have trouble recognizing red cats in the test set. Both datasets attempted to measure the same signal (i.e., cat) but a shift in fur color led to a dataset bias in the test set. Similarly, two RNA-seq datasets measuring gene expression in the liver could be biased if one cohort contains only female, the other only male patients [33], or one group consists of cancer patients and the other not.

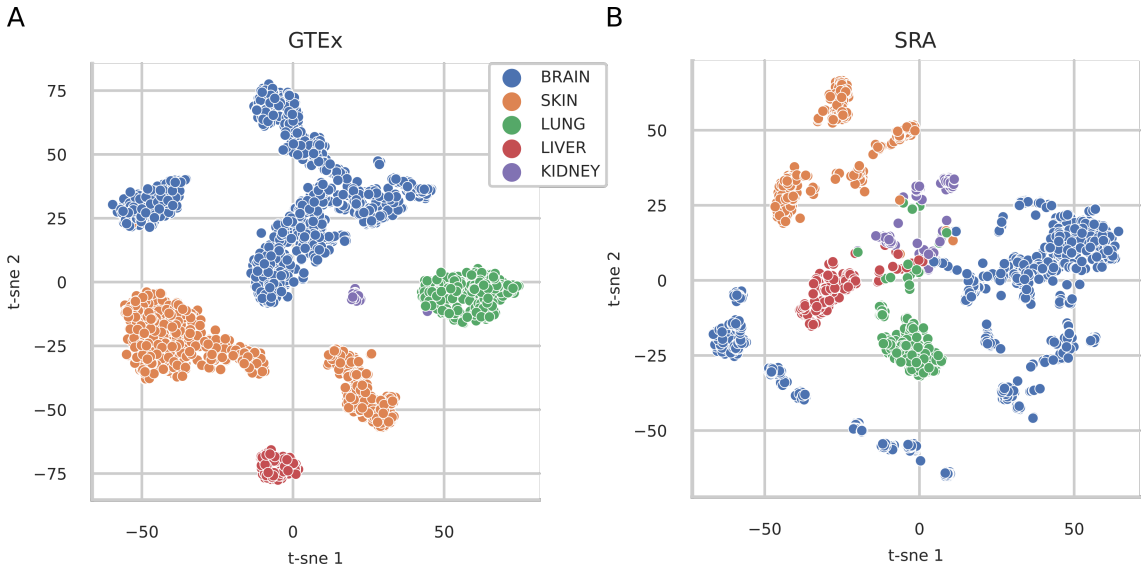

Figure 2.3.: Dataset bias in GTEx vs. SRA. A t-distributed stochastic neighbor embedding (t-SNE) [34] plot. T-SNE is a nonlinear dimensionality reduction method for the visualization of high-dimensional data. Plotted are the expression values of the five largest tissue classes in GTEx (A) and SRA (B). GTEx shows strong homogeneity within the tissues evident by the clear tissue specific clusters. SRA datasets show a diverse number of biological and technical biases evident by the fact that each tissue is split up into smaller clusters, some even mixed. The increased heterogeneity in the SRA data makes it hard to train classification models that generalize across all datasets.

In addition to these biological biases, many technical biases are known. For example, extraction and library preparation kits used during the sequencing workflow (Figure 2.1) have been shown to alter the starting concentration of DNA [35, 36]. T' Hoen et al. [37] showed that even using the same starting material and protocols, different laboratories can not perfectly replicate sequencing results. Arora et al. [38] showed that even the choice of bioinformatic pipelines for raw data processing (e.g., implementation choices made, statistical and algorithmic methods used, software version and run-time parameters) could lead to biases in the final gene count tables.

2.2.6. Machine Learning

Machine learning (ML) is the study of algorithms that improve automatically through experience [39]. In other words, an ML model defines a family of functions from which it chooses (i.e., learns) the best. The best function is determined based on the data provided and an optimization strategy. One of the most simple examples of this concept is linear regression.

Linear Regression

Linear regression aims to find a linear function f such that $y = f(\vec{x})$ where $\vec{x}$ is an input vector and y the output variable. For example, assuming a linear relationship between yearly salary (y) and years of education (x), the following function takes the form:

$$y = a + \beta x$$

In this example, a (the intercept) is the minimum wage, and β is the coefficient (or weight) relating yearly salary to years of education. The parameters a and β have to be estimated given the data x and the labels y.

Linear models perform well if a linear relationship between model input and output can be assumed. However, many problems offer more complex nonlinear relationships for which we need appropriate models (Figure 2.4). Artificial neural networks (ANNs) are such a family of models.

Perceptron

Neural networks belong to what is known as the deep learning methods [40]. The most basic form of ANNs, feedforward neural networks, are able to approximate some highly parameterized nonlinear function f [41]. In supervised learning, a data sample consists of a feature vector $\vec{x}$ and a label y. The function f is a mapping from the input $\vec{x}$ to an output y. The network learns weights $\boldsymbol{W}$ to define the mapping $\hat{y} = f_W(\vec{x})$ where $\hat{y} \simeq y$ is the model's approximation of the true output. Neural networks consist of neurons. A neuron is a computational unit connected to other neurons. The output z of a neuron is the weighted sum of its inputs x and a weight vector $\vec{w}$ plus a bias b.

$$z = \vec{x}^T \vec{w} + b$$

The most simple feedforward network is the perceptron [42] consisting of an input and output layer (Figure 2.5A). The perceptron is a binary classifier outputting 0 or 1. The output z is binarized by applying the activation function $\phi_{step}(x)$ to the output z.

$$\phi_{step}(z) = \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases}$$

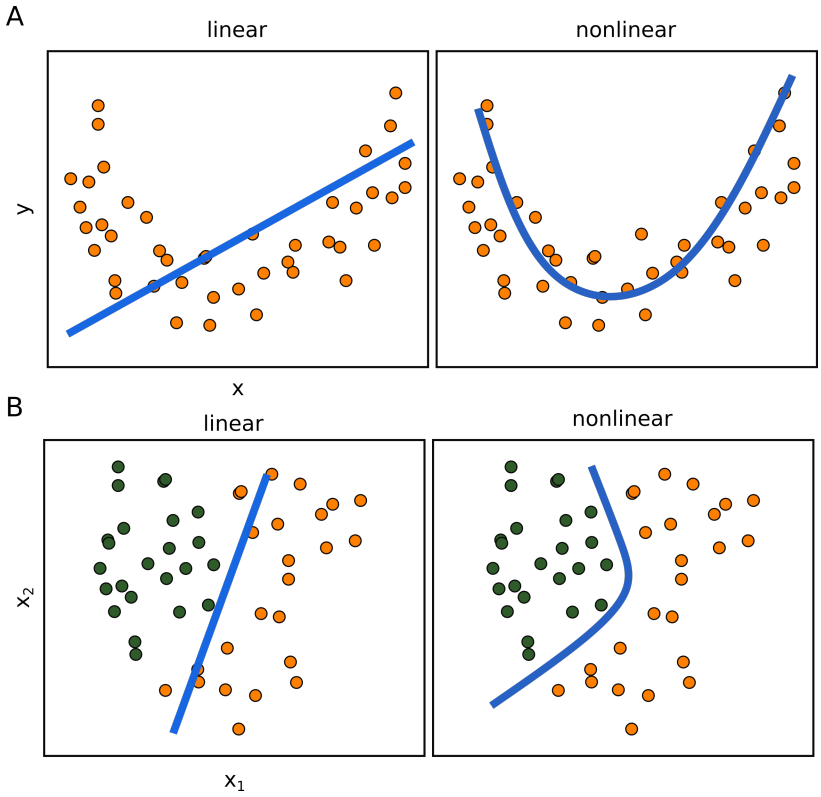

Figure 2.4.: Linear and nonlinear functions. A) Regression is the statistical process of estimating the relationship between input (x) and output (y) variables. In this example of a 2-dimensional plane, a nonlinear relationship between x and y is observed. Fitting a linear model (left panel) to this data results in a higher error than a nonlinear model (right panel). B) Classification is the task of differentiating between two or more groups (i.e., classes) of data points given input features (x_1, x_2). Similar to the regression example, these two classes (green and orange points) are not linearly separable (i.e., they can not be separated by a straight line). A classification model limited to linear combinations of the feature space will result in a higher misclassification rate compared to a nonlinear classifier (right panel).

The function f_W for the perceptron becomes thus

$$f_W(\vec{x}) = \phi_{step}(\vec{x}^T\vec{w} + b)$$

In a 2-dimensional space the equation $\vec{x}^T w + b = 0$ defines a decision boundary separating two classes (Figure 2.5B). Many such decision boundaries are possible. The weights of the perceptron have to be learned such that an optimal boundary can be found.
The learning process requires quantifying the error (or loss) between $\hat{y}$ and y. A very simple loss function l is the squared difference:

$$l_{mse}(f_W(\vec{x}), y) = (f_W(\vec{x}) - y)^2$$

For the total cost, given a training dataset $D_t = \{(\vec{x}_i, y_i), ..., (\vec{x}_N, y_N)\}$ with N samples and an activation function ϕ, a cost function c is defined which in this case is the mean squared error (MSE):

$$c_{mse}(D_t, \vec{w}) = \frac{1}{2N}\sum_{i=1}^{N}\phi(x_i^T\vec{w} + b) - y_i)^2$$

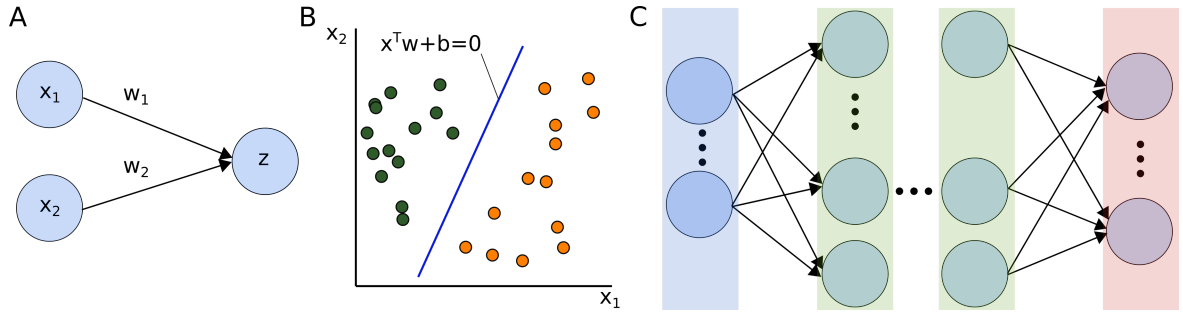

Figure 2.5.: Perceptron A) Symbolic representation of a 1-layer neural network, also known as a perceptron. The input variables x_1 and x_2 are connected to the output node (i.e., the neuron) z. Each input is multiplied by the weight (w_1, w_2) of the connection. B) A perceptron is used for binary classification using the step function. This linear classifier defines a decision boundary separating two classes. Inputs that produce an output of the perceptron ≥ 0 are classified as class 1, else 0. C) A multilayer perceptron extends the perceptron by connecting the input layer (blue) and the output layer (red) through multiple hidden layers (green).

The more accurately the classifier f_W predicts the true value y the lower the cost, thus $\vec{w}$ and b are changed such that the cost function is minimized.

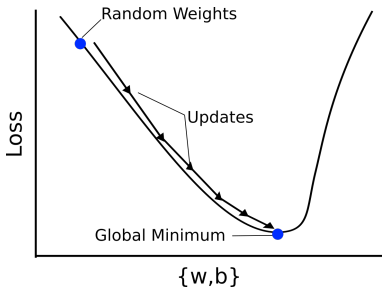

Figure 2.6.: Model optimization with gradient descent. The set of parameters $\{\vec{w}, b\}$ of a model are randomly initialized. Samples are fed forward through the network and the output is calculated. A cost function is used to calculate the error in the model's prediction $\hat{y}$ compared to the true target y. The aim is to find parameters such that the loss is minimal. Gradient descent calculates the partial derivative of the cost function with respect to each parameter and updates the parameters in the negative direction of the gradient until the model converges to a minimum of the cost function.

$c(D_t, \boldsymbol{W})$ is typically minimized using gradient descent (Figure 2.6). During training the parameters of the perceptron are adjusted in the negative direction of the gradient (i.e., the slope of the function at a given point) of the cost function.

$$w_{i+1} = w_i - \alpha \frac{\partial c(D_T)}{\partial w_i}$$

$$b_{i+1} = b_i - \alpha \frac{\partial c(D_T)}{\partial b_i}$$

Where w_i and b_i are the parameter values after the i^{th} iteration and α is the size step the algorithm takes known as the learning rate. Once the cost function c converges the model is optimized and the category of unseen data can be predicted.

Deep Learning

For more complex networks the idea of the perceptron is extended to multilayer perceptrons (MLP). The model can be visualized as a directed acyclic graph describing the flow of data (feed forward). A MLP is a chain of functions $f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$ where $f^{(1)}$,$f^{(2)}$ and $f^{(3)}$ are the layers of the network. In addition many choices for ϕ are available [43]. One of the most commonly used activation functions today is the Rectified Linear Unit (ReLU) [44]:

$$\phi_{ReLU}(x) = max(0, x)$$

Classification

MLPs that are trained on a set of class $Y = \{c_i, ..., c_C\}$ with $C > 2$ typically use the softmax activation function in their output layer $f_{out}(x)$ which has C nodes.

$$p_i = \phi_{softmax}(f_{out}(\vec{\boldsymbol{x}}))_i = \frac{e^{f_{out}(\vec{\boldsymbol{x}})_i}}{\sum_{j=1}^{C} e^{f_{out}(\vec{\boldsymbol{x}})_j}}$$

Softmax takes as input the output vector of the output layer of the network and returns the ratio of the exponential of the i^{th} output node and the sum of the exponential of all output nodes. f_{out_i} is the output value of the i^{th} node / class. The new output vector $\vec{\boldsymbol{p}}$ sums up to 1, and each output value can be interpreted as a probability to belong to the corresponding class such that

$$\hat{y} = argmax(p)$$

A typical loss function used for classification is the cross-entropy loss.

$$l_{ce}(\vec{\boldsymbol{p}}, \vec{\boldsymbol{y}}) = -\sum_{i=1}^{C} y_i log(p_i)$$

Where $\vec{\boldsymbol{y}}$ is a one-hot encoded vector of dimension C with $y_i = 1$ if $\vec{\boldsymbol{x}}$ is of class i else 0.

Siamese Networks

A siamese network [45] is an ANN architecture designed to learn similarities (Figure 2.7A). Siamese networks can be trained on triples, an anchor, a positive and a negative sample of equal and unequal class, respectively. Given a set of N triplets $D_{trip} = \{(\vec{\boldsymbol{x}}_i^a, \vec{\boldsymbol{x}}_i^p, \vec{\boldsymbol{x}}_i^n), ..., (\vec{\boldsymbol{x}}_N^a, \vec{\boldsymbol{x}}_N^p, \vec{\boldsymbol{x}}_N^n)\}$ and a function g_w parametrized by the weights matrix $\boldsymbol{W}$ the model is trained as follows: i) each sample of a tripled is passed through g_w, ii) the distances $d_{l_2}(\vec{\boldsymbol{x}}_i^a, \vec{\boldsymbol{x}}_i^p)$ and $d_{l2}(\vec{\boldsymbol{x}}_i^a, \vec{\boldsymbol{x}}_i^n)$ in the embedding space (i.e., output layer) are measured, iii) $\boldsymbol{W}$ is updated such that the model converges towards $d_{l_2}(\vec{\boldsymbol{x}}_i^a, \vec{\boldsymbol{x}}_i^p) < d_{l_2}(\vec{\boldsymbol{x}}_i^a, \vec{\boldsymbol{x}}_i^n)$ (Figure 2.7B) with

$$d_{l_2}(\vec{\boldsymbol{x}}_1, \vec{\boldsymbol{x}}_2) = \|g_w(\vec{\boldsymbol{x}}_1) - g_w(\vec{\boldsymbol{x}}_2)\|$$

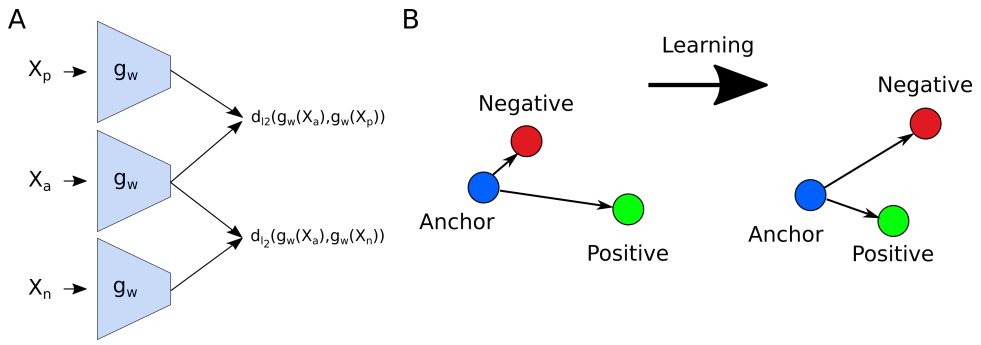

Figure 2.7.: Siamese neural network architecture. A) A siamese network is an ANN architecture learning similarities instead of decision boundaries. Siamese networks pass two or more samples in parallel through the same network, depending on the loss function. The triplet loss takes as input the distance between an anchor, a positive sample from the same class and a negative sample of a different class in the embedding space. B) The network is updated such that the distance between anchor and positive sample is minimized and between anchor and negative sample maximizes. (Adapted from [46])

Schroff et al. [46] introduced the triplet loss function. The triplet loss minimizes the distance from the anchor to the positive sample and maximizes the negative sample's distance.

$$l_t(\vec{x}^a, \vec{x}^p, \vec{x}^n) = max(d_{l_2}(\vec{x}^a, \vec{x}^p) - d_{l_2}(\vec{x}^a, \vec{x}^n) + m, 0)$$

Where m is a margin parameter.

Domain Adaptation

Domain adaptation (DA) is a branch of ML concerned with developing network architectures and training procedures that enable models trained on a source domain to perform well on a biased target domain. In other words, the aim is to develop models that extract bias invariant features from the training data. Many different types of algorithms have been developed in recent years attempting to solve this problem. Following is a brief overview of the literature that has been influential to this work. DLID [47] is an algorithm trained on various degrees of mixed source and target data. The model extracts features from each intermediate dataset and interpolates a feature path between the two domains. Another popular method is domain-adversarial training, as proposed by Ganin et al. [48]. Domain-adversarial training uses two loss functions, one to differentiate between classes and a second to differentiate between the domains. Samples of both domains are passed through the same network, forcing it to extract features relevant for class classification and suppress features relevant for domain classification.

Recently, Tzeng et al. [49] introduced adversarial-discriminative domain adaptation, which applies the GAN-loss [50]. Generative adversarial networks (GANs) train two networks, a generator and a discriminator. The discriminator takes as input a 'real' sample and a 'fake' generated sample. The networks are trained such that the generator outputs samples that can

not be discriminated from 'real' samples. In the DA approach proposed by Tzeng et al., a standard MLP is trained on the source domain. The MLP is then split into a source mapper (i.e., all layers before the output layer) and the classification layer (i.e., output layer). For DA, the output of the source mapper (with fixed weights) is used as 'real' samples, and the output of a new network (i.e., target mapper) is used as 'fake' input to a discriminator. The target mapper is trained such that the discriminator can not differentiate between samples from the source and the target domain. Once the target mapper is trained, it is used for classification connected to the pre-trained classification layer. Building on the previous approach, Motiian et al. [51] proposed a model combining adversarial discriminative learning and siamese network architecture.

2.3. Aim and Problem Statement

Next-generation RNA-seq has been a pillar of biomedical research for many years [52, 53]. A continuous drop in cost has made RNA-seq a widely available method of choice to uncover the molecular basis of biological development or disease [3, 10]. As a result of this, recent years have seen a strong growth in publicly accessible RNA-seq data.

The actual reuse and integration of this data, however, has been largely limited by the lack of consistent metadata annotation and individual dataset bias [37, 54]. The lack of metadata annotation for RNA-seq samples, such as tissue of origin, disease or sex phenotype, prohibits experimenters from finding data that is relevant to their research.

To allow for efficient data reuse, publicly available data has to be harmonized and well annotated with standardized metadata [11]. The primary database for next-generation sequencing projects, the SRA [27], provides a centralized repository for raw sequencing information. However, the SRA lacks rigorous standards of curation, which limits the reusability of its data. Efforts to predict missing or sparse metadata in public RNA-seq resources have shown promising results. Recently, a linear regression model fitted to GTEx data has been presented for the prediction of tissue, sex and other phenotypes of SRA and TCGA samples [1]. While this method performed well on homogenous TCGA data, results on SRA test data was less convincing. The authors identified the large number of different dataset biases in the SRA as a potential reason for the limited results on that data.

ANNs in their various forms and functions consistently outperform classical ML approaches in a large variety of biological tasks, including classification, data generation and segmentation [55, 56, 57, 58]. Given large training datasets these algorithms can learn complex representations of data by automatically weighting and combining features nonlinearly. This has led us to hypothesize that ANN-based models could increase the performance in metadata prediction beyond that of classical ML approaches such as linear regression. Of special interest in this context is domain adaptation [59], a subfield of ML which aims to specifically alleviate problems conferred by dataset bias [60].

Here we present a DA approach capable of leveraging a large number of dataset biases, boosting generalizability of phenotype prediction. We developed the model using three data sources (GTEx, TCGA and SRA) of different size and with a different degree of bias. To validate our approach we compare it to the previously suggested linear model (LIN) [1] as well as a standard multi-layer perceptron (MLP) on prediction of tissue of origin, sex and sample source. Importantly, we find that our DA network significantly outperforms the LIN model by up to 12.3% in prediction accuracy. We subsequently apply trained models to generate and make available new metadata for 8,495 unique SRA samples.

2.4. Experimental Setup

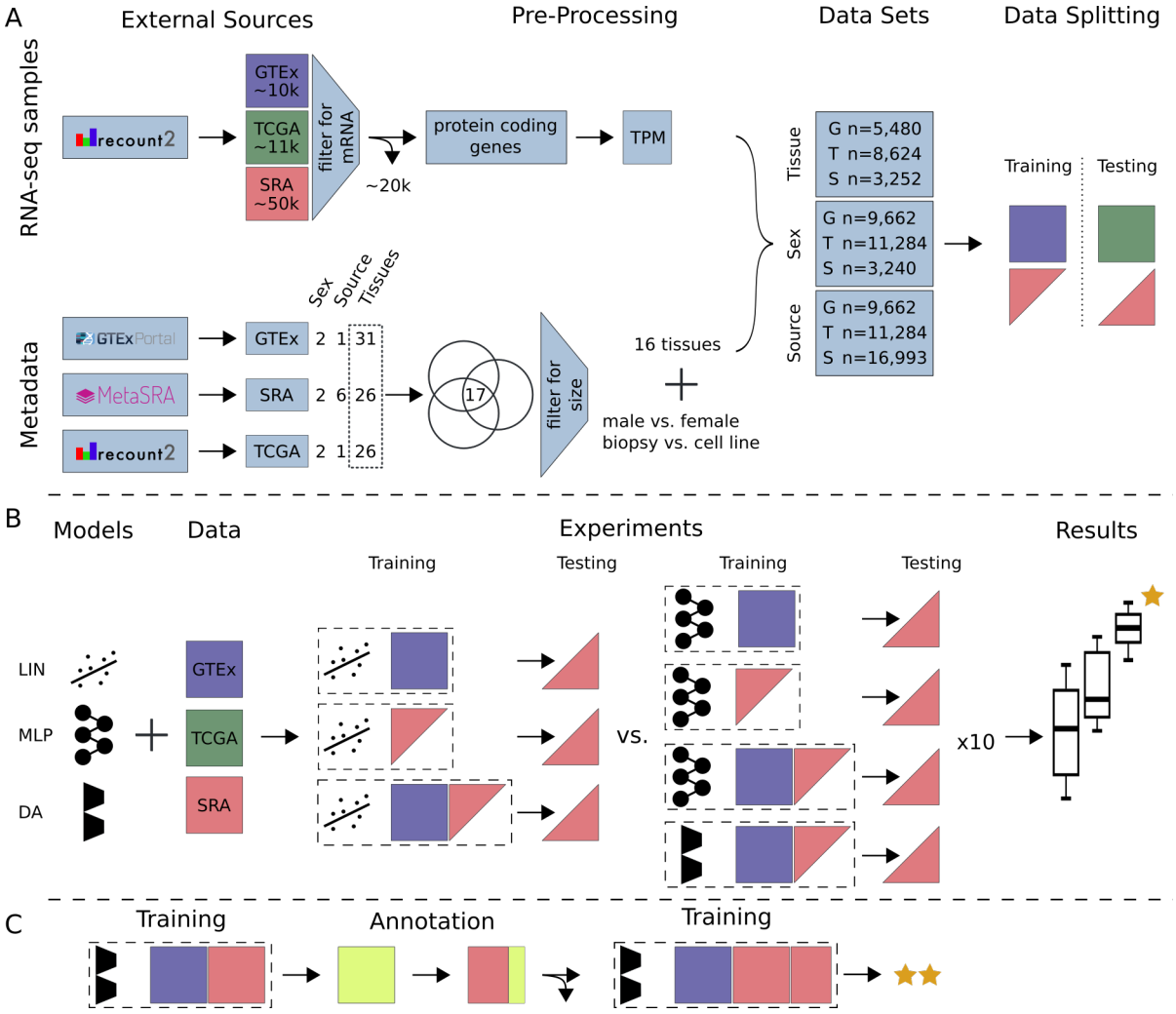

Figure 2.8.: Study overview. (A) Data was downloaded from recount2 and split into three data sources: (i) GTEx, (ii) TCGA and (iii) SRA. Non bulk mRNA data and technical replicates were removed. Protein coding genes were selected and TPM normalized. Metadata for tissue of origin (e.g. heart), source (e.g. biopsy) and sex phenotype was collected, if available. A subset of 16 tissues was selected. Samples were annotated, training and testing datasets were created. (B) Three models were compared: LIN (linear model), MLP (multilayer perceptron) and DA (novel domain adaptation algorithm). Each experiment (dashed box) is made up of a model and training data. The previously published LIN model served as a benchmark for our MLP and DA model. Each model experiment was repeated 10x with different seeds to give an estimation of uncertainty. The best model (orange star) was chosen by comparing average performance across all seeds. (C) All available data was used for training the best model. Previously unlabeled SRA data (yellow square) was automatically annotated with the appropriate metadata. Newly annotated metadata can be used to re-train existing models to further improve performance.

This study aims to find the best model for RNA-seq metadata annotation based on gene expression. Three different data sources were selected for which phenotype data was available (Figure 2.8A). Each of the three data sources comes with a different number of dataset biases. Briefly,

GTEx is a large homogeneous dataset containing healthy samples following a strict centralized standard protocol. TCGA contains pooled samples from different cancers, disease stages and sequencing centers. Our SRA data comprises hundreds of individual studies following no centralized standard, containing the largest number of biases of all three data sources. Bias in a test dataset that a model has not learned (domain shift) can severely compromise performance. We hypothesized that exposing classification models to a sufficient number of dataset biases will enable them to learn a generalized internal feature representation. Such a model would be able to classify data with previously unseen biases. To test and benchmark our models, we selected the classification tasks of (1) tissue of origin of a given RNA-seq sample, (2) biopsy vs. cell line origin of a sample (i.e., sample source), and (3) sample sex (Figure 2.8A).

Three different machine learning models were compared (Figure 2.8B). First, a fully connected ANN (MLP) was tested because of its capability to create novel latent features (see Methods for model details). Second, we developed a domain adaptation (DA) approach (Figure A.2), a subfield of machine learning dealing with dataset biases. Lastly, the LIN model trained on GTEx data, proposed in Ellis et al. [1], was used as the baseline for all tissue and sex classification experiments. Models were trained on either GTEx or a mix of GTEx and SRA data and tested on TCGA and SRA data. Uncertainties for MLP and DA models were estimated from 10 training runs with different random seeds (Figure 2.8B).

2.5. Methods

2.5.1. Data Acquisition and Processing

Data Source

To train and test models, we gathered data from three different sources, each with a different level of homogeneity, which we define as the number of unique dataset biases present within one data source. Datasets were defined as all the RNA-seq samples from one study based on the assumption that they were obtained and processed under identical conditions. To avoid additional biases by using different bioinformatic alignment pipelines [38], all data was downloaded from recount2. The RSE V2 files of all available RNA-seq projects (n=2,036) from recount2 (release 13.09.19[1]) were downloaded using the recount R package (v 1.11.13). The downloaded data was separated into three different data sources according to their origin.

GTEx - The Genotype-Tissue Expression Project [32] v6 (https://www.gtexportal.org/) comprises 9,662 samples from 554 healthy donors across 31 tissues. GTEx strives to build a highly homogeneous dataset with strict guidelines on donor selection, biopsy and sequencing methodology[2]. We considered the GTEx data source to have a single dataset bias.

SRA - A total of 2,034 SRA [27] studies containing 49,657 samples were downloaded from recount2 [29]. Every SRA study was potentially processed at a different site by a different technician following different standards. Besides, the underlying biological condition of the samples is often unclear. We assume each study to have a unique dataset bias which makes the SRA a highly heterogeneous data source. In addition, data annotation is not standardized, resulting in sparse metadata with low fidelity.

TCGA - RNA-seq data for The Cancer Genome Atlas was downloaded consisting of 11,284 samples spanning 26 tissues. While there are 740 samples of healthy donors across 20 tissues, more than 90% of the samples are tumor biopsies from various tissues and different stages of tumor progression. In contrast to GTEx, TCGA is a submission-based project leaving more room for potential dataset bias. Despite the high level of standardization and reliability of metadata information, heterogeneity is also inherent to the TCGA dataset due to the biological context (cancers, stages), albeit not as pronounced as in SRA.

Preprocessing of SRA Data Source

This study focuses on bulk mRNA-seq data, as it is by far the most frequent datatype in either of the three data sources used. The following approaches were used to exclude data from single-cell and small RNA (sRNA studies from further analysis: First, we identified sRNA-

[1] https://jhubiostatistics.shinyappts.io/recount/
[2] https://www.gtexportal.org/home/documentationPage

seq data based on the total fraction of sRNA counts and protein-coding RNAs. Specifically, we considered a subset of the Gencode gene types (i.e., protein_coding and processed_pseudogene vs. rRNA, miRNA, misc_RNA, snRNA and lincRNA). Every sample with its maximum total count fraction not allocated to either protein_coding or processed_pseudogene was removed from further analysis (Figure A.1). Second, we removed single-cell RNA-seq studies by scanning titles and abstracts for variations of the words 'single cell' and manually validated and excluded the identified samples. In addition to this semi-automatic validation step, we manually validated the 50 largest projects within the SRA data source and removed samples that did not qualify as bulk mRNA-seq data.

Most importantly, we noticed a large number of technical replicates in the remaining SRA data. Using technical replicates to train and test a classification model inflates the reported metrics. Therefore only samples with a unique experiment accession (SRX#) were retained. From the 49,657 SRA samples downloaded initially, 29,685 samples and 1,833 unique studies passed our preprocessing steps.

Metadata

Three different phenotypes for expression-based prediction were considered. Explicitly, we predicted the tissue of origin of a biopsy (e.g., heart, lung, kidney, ovary), the patients' sex, and sample source (denoting whether the sample was from a patient biopsy or a lab-grown cell line).

GTEx and TCGA - Tissue and sex annotation for GTEx were extracted from the official sample annotation table provided by GTEx (GTEx_Data_V6_Annotations_SampleAttributesDS.txt [3]). Recount2 provided an annotation file for TCGA from which we took columns gdc_cases.project.primary_site and gdc_cases.demographic.gender for tissue and sex annotation, respectively. Sample source was assumed to be of type biopsy for all GTEx and TCGA samples.

SRA - For the SRA samples, we relied on normalized metadata provided by MetaSRA. Available SRA identifiers were downloaded through the GUI [4] by searching for all 31 GTEx tissues (site accessed on 11.09.2019) (Supplementary Table A.2). Of the 31 tissues available for GTEx, we were able to identify samples for 26 in MetaSRA, resulting in 6,183 annotated SRA samples. Sample identifiers for sex were accessed through the same GUI by searching for male and female organisms + Homo sapiens cell line, which resulted in 3,240 annotated SRA samples. Sample source was determined using the SQLite file provided by MetaSRA (metasra.v1-5.sqlite [5]), resulting in 28,043 annotated samples across six sample source categories.

[3] https://storage.googleapis.com/gtex_analysis_v6/annotations
[4] http://metasra.biostat.wisc.edu
[5] http://metasra.biostat.wisc.edu/download.html, colum sample_type

Tissue Label Harmonization - GTEx, TCGA and SRA have 17 common tissue types (Figure A.3). Bladder was removed due to its small sample size (GTEX n=11). We kept tissues of comparable size in the SRA (adrenal gland n=14, testis n=14, pancreas n=17 in the SRA training data). The SRA training data was mainly used for bias injection and size was not considered an exclusion criterion. Filtering resulted in 5,480, 8,624, and 3,252 tissue annotated samples across 16 tissues for GTEx, TCGA and SRA, respectively (Supplementary Tables A.3 and A.4).

Dimensionality Reduction and Normalization

The downloaded gene count table provided counts for 58,037 genes (Gencode v25, GRCh38, 07.2016). First standard log2 Transcript per Million (TPM) normalization was applied to normalize for gene length and library size. Next, we reduced the number of input features (genes), aiming to keep features containing information and remove potentially uninformative features. All non-protein-coding genes were removed, reducing the gene set by 65.5% to 19,950 genes. For sex classification, only protein-coding genes on the X and Y chromosome (n=913) were selected. The Gini coefficient for each gene was computed. Only genes that showed significant dispersion between tissues [23, 61, 62] across all GTEx samples were retained. Housekeeping genes, for example, are known to be expressed similarly across tissues and would score a low Gini coefficient (i.e., high dispersion). Low and high cutoffs were determined during hyperparameter optimization. For tissue classification, genes with Gini coefficients g between 0.5 and 1 were retained, resulting in a features space of dimension d=6,974. For sex classification, genes with $0.4 < g < 0.7$ were used (d=190). Sample source classification included genes with $0.3 < g < 0.8$ (d=8,679) (Supplementary Table A.3).

Dataset Preparation

Phenotype Classification Experiments - **Tissue**: SRA data was always split on the study level (SRP#) into train and test sets. The two largest SRA studies per class were put in the training set for tissue prediction. This split ensured maximal bias variability in the remaining test data. Of the 178 SRA studies containing tissue annotated samples (SRR#), 30 studies were selected for the training set (n=1,721) and 148 studies for the test set (n=1,531) (Supplementary Tables A.3 and A.4). **Sex**: In total, 159 SRA studies contained samples annotated with male and or female by MetaSRA. These studies were combined into the training set (studies=78, n=2,317), and test set (studies=81, n=923) (Supplementary Tables A.3 and A.4). For model validation, GTEx was randomly split into training and test sets with an 80:20 ratio for both sex and tissue classification. **Sample Source**: A confidence cutoff of >= 0.7 on the predicted label was applied (provided by MetaSRA), reducing the total amount of annotated samples for SRA from 23,651 to 17,343. For each of the two selected SRA categories (i.e., biopsy and cell line), we sorted all available studies by the number of samples, placed the first third of studies into the training (studies=420, n=12,725), the second third into the test (studies=422, n=3,144) and the last third into the SRA validation set (studies=418, n=1,124) (Supplementary Tables A.3 and A.4).

Metadata Annotation - After determining the best model for each phenotype, we re-trained the models for automated metadata annotation (Figure 2.8C). The same datasets as defined above were used for the sex metadata annotation. Tissue: We followed the same pipeline as described above, the only difference being that no samples were discharged because of their tissue label. Samples from a tissue class other than the original 16 classes were pooled together into a 'catch-all' class, resulting in 17 classes. In total, 44 SRA studies were selected for the training set (n=3,370) and 203 studies for the test set (n=2,813). Sample Source: Contrary to before, we used all available classes in the SRA data source for metadata annotation. All classes that are not tissue (i.e., biopsy) were grouped into a single 'catch-all' class while the same cutoff as before was applied. The training set (n=16,463) comprises 974 SRA studies and the test set (n=3,707) of 492 studies.

2.5.2. Machine Learning Models

Multilayer Perceptron - MLP

All our ANN-based models were developed and trained on tf.keras (Tensorflow 2.1 [63]). The hyperparameters for each prediction task were determined using an exhaustive iterative random search (keras tuner 1.0.1) (Supplementary Table A.5). In the case of approximately equal accuracy on the validation set, the least complex model was chosen. A single hidden layer was selected for each problem with 128, 128 and 32 nodes for tissue, sample source, and sex prediction, respectively (Supplementary Table A.6, Figure A.4). Each network was trained for 10 epochs with a batch size of 64.

Domain Adaptation Model - DA

Many DA models correct bias between two domains, a source and a target domain. However, in biological research, one is often confronted with a large number of small datasets, each potentially with its unique dataset bias. Therefore, we specifically designed our DA model to learn from very few data by using a siamese network architecture. The siamese network learns bias from pairs or triplets of training samples by exposing each sample in multiple relationships to the model. We distinguished three different types of input data for our model. The source domain $D_{SD} = \{(\vec{x}_i, y_i), ..., (\vec{x}_n, y_n)\}$ with N samples is a sizable single-bias dataset used to learn the feature embedding for the classification task (in our case: GTEx). The bias domain $D_{BD} = \{(\vec{x}_i, y_i), ..., (\vec{x}_m, y_m)\}$ with M samples contains labeled samples from multiple smaller datasets (in our case: SRA), each with its own bias. The target domain $D_{TD} = \{\vec{x}_i, ..., \vec{x}_k\}$ with K samples refers to unlabeled and biased datasets we want to classify (unlabeled SRA or TCGA data).

Model Architecture - Our DA model is based on the siamese network architecture. It consists of three modules: A source mapper f_{SM}, bias mapper f_{BM}, as well as a classification layer f_{CL},

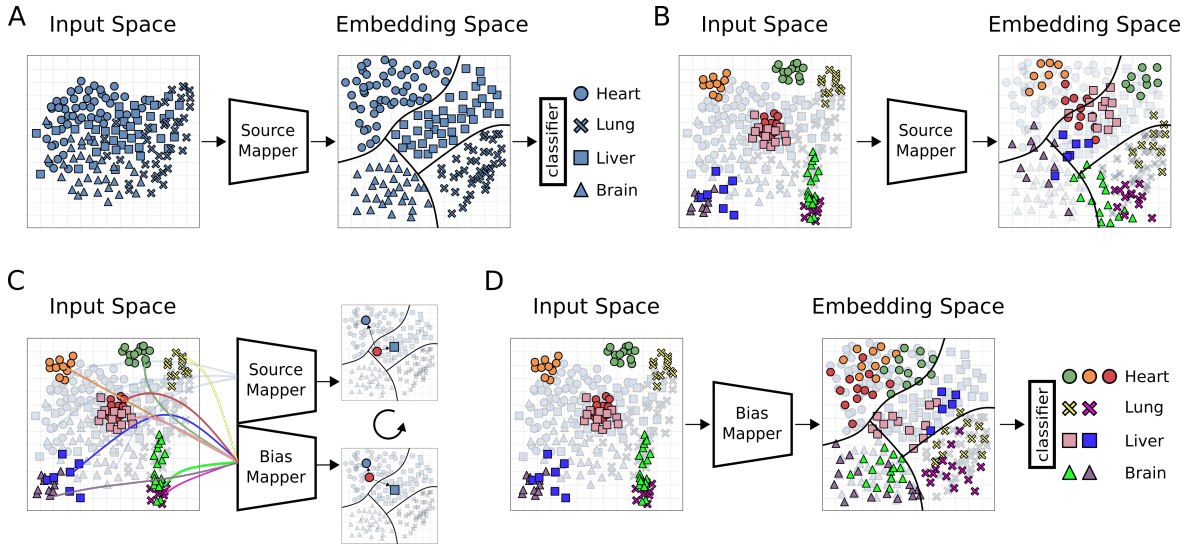

Figure 2.9.: Overview domain adaptation model. Illustration of our DA model architecture and training. Shapes of (hypothetical) data points represent classes, colors are datasets with unique biases. Source Mapper (SM), Bias Mapper (BM) and classifier layer (CL) are ANN modules. (A) First training cycle: The SM is trained on a single bias dataset, the source domain (SD). In this step, the SM learns a feature embedding. The CL learns how to partition this embedding space into classifiable regions and draws decision boundaries (black lines). (B) For biased test data (colored sample data points), same classes may occupy distinct regions in input space. In this case, the source mapper may not be able to map the samples to the correct region of embedding space, compromising classification performance of the CL. (C) In order to learn the mapping of different biases to the embedding learned in (A), a bias mapper (BM) is created by copying the SM, and trained weights of the SM are fixed. In this second training cycle, triplets of samples are passed through the SM-BM configuration, consisting of an anchor from the bias domain and two samples from the source domain, one of them with a matching label. The triplet loss function is defined to minimize distance of like labels in embedding space and to maximize distance of opposite labels. This process is repeated until the SM has learned to map all known biases into the previously learned embedding space. (D) The BM is now able to map data points from previously unseen datasets into the embedding space where the CL can classify them.

with weights $\boldsymbol{W_{SM}}$, $\boldsymbol{W_{BM}}$ and $\boldsymbol{W_{CL}}$, respectively. These modules give rise to three different configurations, i.e., two training and one prediction configuration (see Figure A.2 for a brief illustration). In the first training cycle, f_{SM} and f_{CL} are combined to form an MLP (Figure 2.9A).

$$f_{MLP}(D_{SM}) = f_{CL}(f_{SM}(D_{SM}))$$

The f_{SM}'s task is to learn a mapping $f_{SM} : D_{SD} \mapsto E$ from the input space D_{SD} to an embedding space E from which the f_{CL} can predict phenotype classes $f_{CL} : E \mapsto Y$. f_{MLP} is trained with a batch size of 64 for 10 epochs. Because the f_{MLP} is trained on a sizable single-bias dataset, it will likely overfit and thus not readily generalize to other datasets (Figure 2.9B). For a second training cycle, the bias mapper f_{BM} is created with the same architecture as the f_{SM}.

$$\boldsymbol{W_{BM}} \equiv \boldsymbol{W_{SM}}$$

f_{CL} is removed and the weight matrices $\boldsymbol{W_{SM}}$ and $\boldsymbol{W_{CL}}$ are frozen. The two networks f_{SM} and f_{BM} are now trained as a siamese network (Figure 2.7A). The network is trained using the triplet loss l_t (see Section 2.2.6) on triplets $\vec{\boldsymbol{x}}^a$ (anchor), $\vec{\boldsymbol{x}}^p$ (positive), $\vec{\boldsymbol{x}}^n$ (negative) with $\vec{\boldsymbol{x}}^a \in D_{BD}$ and $\vec{\boldsymbol{x}}^n, \vec{\boldsymbol{x}}^p \in D_{SD}$. For improved training time and robustness, our model is trained on semi-hard triplets [46] satisfying he following condition:

$$d_{l_2}(\vec{\boldsymbol{x}}^a, \vec{\boldsymbol{x}}^p) < d_{l_2}(\vec{\boldsymbol{x}}^a, \vec{\boldsymbol{x}}^n) < d_{l_2}(\vec{\boldsymbol{x}}^a, \vec{\boldsymbol{x}}^p) + m$$

The siamese network was trained for 10 epochs with a batch size of 64. Hyperparameters were determined as described above (Supplementary Table A.6, Figure A.4). As the second training cycle proceeds, the f_{BM} learns a mapping $f_{BM} : D_{BD} \mapsto E$. After training, the f_{BM} and f_{CL} are combined to form the final DA model f_{DA} and can be used to predict the target domain (Figure 2.9D).

$$f_{DA}(D_{TD}) = f_{CL}(f_{SM}(D_{TM}))$$

Linear Regression Model - LIN

We used the metadata prediction performance of the LIN model described in Ellis et al. [1] as a reference. The LIN model was optimized on the same data as all other models (see data section of methods). For each experimental setup, the following steps were conducted in R version 3.6.3 in order to build the corresponding phenotype predictor and evaluate its accuracy based on the test data:

1. calculating the coverage matrix for the training samples based on the regions reported in Ellis et al. [1] by employing the function 'coverage_matrix_bwtool' (R package recount.bwtool version 0.99.31).

2. building the model by running 'filter_regions' and 'build_predictor' (R package phenopredict version 0.99.0) with the same parameters used in Ellis et al. [1]

3. testing the model on the test samples with 'extract_data', 'predict_pheno', 'test_predictor' (R package phenopredict version 0.99.0)

Notably, our experiments differ from the original work [1] solely by applying additional preprocessing steps to the samples, which may be responsible for observed small differences in performance.

2.5.3. Nomenclature of Experiments

Each experiment was named after the model, the training and the test data used. The possible models are LIN, MLP and DA. The data sources are named G (GTEx), T (TCGA) and S (SRA). If only the SRA training data is used (i.e. if the model is evaluated on the SRA test data) we write S_{small}. If the SRA train and test sets are combined for training we write S_{large}. For instance,

an experiment using an MLP, trained on a mix of GTEx and SRA and evaluating on SRA data would be named MLP G+S_{small}-S.

2.5.4. Impact of Data Diversity and Quantity on Model Performance

To analyse the effect of training data diversity on prediction accuracy the following experiments were designed. First, MLP S-S models for sample source prediction were trained with an increasing number of unique SRA studies in the training data, systematically increasing bias diversity. Only SRA studies containing > 100 samples for either class were considered. In order to control for training set size, each SRA study was subsampled to 50 samples before training. Six iterations of this training process were conducted starting with one study (i.e. one bias) per class (biopsy vs. cell line). At each step one additional SRA study per class was subsampled ending with six SRA biases and 350 samples in the training set per class. As a control experiment we chose the largest SRA study available for each class to create a training set with a single bias per class. Starting with 50 samples per class in six iterations we subsampled an additional 50 samples ending with 350 samples, thereby assessing the effect on performance that can be attributed to the dataset size. Subsampling and random selection of SRA studies were repeated 10 times with different seeds and each configuration was trained on 10 different seeds, yielding an estimate of uncertainty.

2.5.5. Metrics

We report micro and macro accuracy which are equivalent to mean sample accuracy (MSA) and mean class accuracy (MCA) respectively. Sample accuracy is a measure of absolute performance on the test data. It reports the fraction of correctly classified samples over all classes:

$$\text{MSA} = \frac{\sum_i^N \mathbb{1}_{y_i}(\hat{y}_i)}{N}$$

Where N is the number of samples, y the true label and $\hat{y}$ the predicted label, and $\mathbb{1}$ is the indicator function. Given the large class imbalance in some of our experiments an increase in accuracy in a small class will not be captured by this metric. Average class accuracy, on the other hand, reports the average sample accuracy per class, weighing each class equally and thereby capturing local improvements of the models:

$$\text{MCA} = \frac{\sum_{j=1}^{C} \frac{1}{M_j} \sum_{i=1}^{M_j} \mathbb{1}_{y_{ij}}(\hat{y}_{ij})}{C}$$

Here, C is the number of classes, M_j is the number of samples for class j, y_{ij} and $\hat{y}_{ij}$ are the true and predicted values, and $\mathbb{1}$ is the indicator function.

2.5.6. Statistical Tests

Accuracy distributions for sex and tissue prediction were tested for statistically significant differences using a t-test (two distributions, scipy.stats.ttest_ind v 1.3.1) or ANOVA (more than two distributions, scipy.stats.f_oneway) with a significance threshold of 0.01.

2.6. Results

2.6.1. Domain Adaptation Outperforms Other Models on Tissue Classification

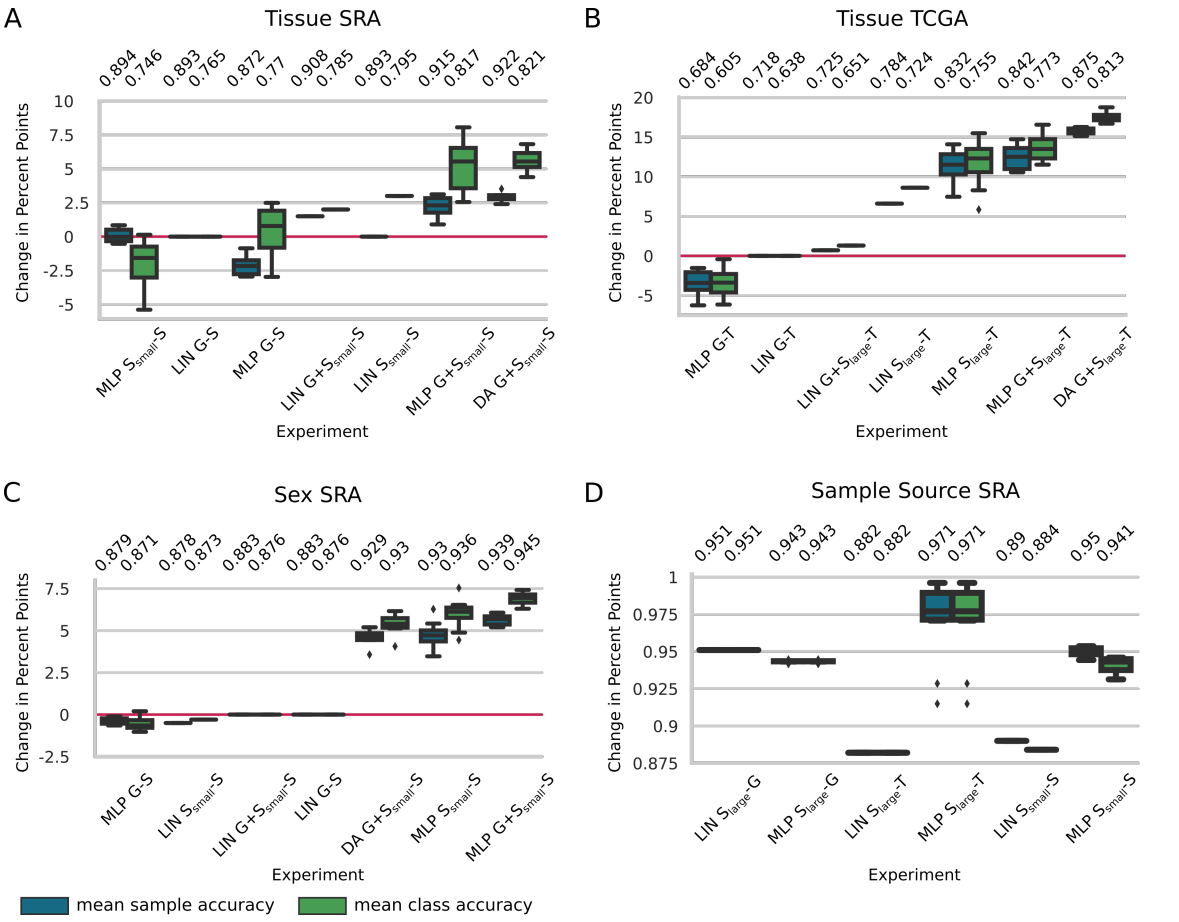

Figure 2.10.: Phenotype prediction results for A) prediction of tissue of origin on SRA, B) TCGA (16 classes), C) prediction of sex on SRA (2 classes) and D) sample source (2 classes) on SRA data. Indices 'small' and 'large' refer to the different size of SRA training data used due to splits of the dataset in SRA prediction. Box plots represent model uncertainty of ANN based models, estimated from training with different random seeds (n=10). Mean sample accuracy and mean class accuracy were calculated for each seed. For panel A-C) LIN G-X was chosen as the baseline model. Results for these panels are given in change in percentage points compared to the baseline (red line). Experiments are sorted by increasing mean class accuracy. LIn=linear regression, MLP=multilayer perceptron, DA=domain adaptation, G=GTEx, T=TCGA, S=SRA.

We first tested the performance of the LIN, MLP, and DA algorithms to predict the tissue of origin on GTEx (n=5,480), TCGA (n=8,624), and SRA (train n=1,721, test n=1,531) datasets. A subset of 16 tissue labels was chosen that is common to all three data sources (see Methods, Figure A.3, Supplementary Table A.4). First, we conducted a single-bias experiment, i.e., MLP G-G (see 2.5.3). The nearly perfect score of mean sample accuracy (MSA) of 0.996 and mean class accuracy (MCA) of 0.99 (data not shown) confirmed that the MLP yielded highly accurate results when trained and tested on a single-bias dataset.

Prediction of SRA Tissue - Metadata prediction on SRA was the most challenging and interesting task due to the numerous biases. The base model LIN G-S was re-trained and tested on our datasets and achieved a MSA of 0.893 and a MCA of 0.765 for the 16 tissues (Figure 2.10A). Of note is the significantly higher accuracy achieved with LIN G-S than the one reported by Ellis et al. [1] (0.519 MSA). MLP G-S (MSA: 0.872, MCA: 0.77) had a higher MCA but a lower MSA than the corresponding LIN model (Figure 2.10A). As a next step models were trained on multi-bias training data. Specifically, models were trained on S_{small}. MLP S_{small}-S (MSA: 0.894, MCA: 0.746) matched the base model's MSA score but performed slightly worse using the MCA metric. Similarly, the LIN S_{small}-S model matched the MSA of LIN G-S but showed an increased performance for MCA (MSA: 0.893, MCA: 0.795). Notably, by only using the small SRA training dataset, we lose the advantage of the large sample size of GTEx. Based on this, we hypothesized that by combining SRA and GTEx in the training data, we might leverage both sample size and diversity.

The LIN G+S_{small}-S model increased the MSA to 0.908 and MCA to 0.785, which is 1 percentage point (ppt) lower than the LIN S_{small}-S model. The two best-performing models were MLP G+S_{small}-S and DA G+S_{small}-S, outperforming LIN G-S on MSA by 2.5 ppts and MCA 5.5 ppts (MLP G+S_{small}-S MSA: 0.915, MCA: 0.817 and DA G+S_{small}-S MSA: 0.922, MCA: 0.821). No significant difference in the mean performance was detected between the best two models (MSA p-value > 0.01, MCA p-value > 0.01, t-test). Crucially, however, DA G+S_{small}-S exhibited the lowest standard deviation (st$d$=0.003 for MSA and st$d$=0.009 for MCA) of all models tested (Supplementary Table 6). For this reason, DA G+S_{small}-S was considered to be the best model for the prediction of tissue on the highly heterogeneous SRA test data. DA G+S_{small}-S increased the MSA score by 1.5% compared to LIN G+S_{small}-S and MCA by 3.3% compared to LIN S_{small}-S, the best performing linear models for the respective metrics.

Prediction of TCGA Tissue - Next, model performance on TCGA data was assessed (Figure 2.10B). The baseline model for this task, LIN G-T, achieved MSA 0.718 and MCA 0.638 (Figure 2.10B). Applying the MLP model on the same data resulted in a drop of MSA and MCA of 2.4 and 3.3 ppts, respectively (MLP G-T MSA: 0.684, MCA: 0.605). For TCGA tissue prediction, we used S_{large} for training, essentially doubling the SRA training data (SRA train + SRA test set: n=3,252). LIN S_{large}-T improved accuracy by 6.6 ppts for MSA and 8.6 ppts for MCA to 0.784 and 0.724, respectively. In comparison, MLP S_{large}-T increased model performance by 11.4 ppts to 0.832 (by 11.7 ppts to 0.755) for MSA (MCA) with respect to LIN G-T. Combining GTEx and SRA training data reduced LIN G+S_{large}-T performance to MSA 0.725 and MCA 0.651. The best accuracy was achieved by our MLP G+S_{large}-T (MSA: 0.842, MCA: 0.773) and DA G+S_{large}-T (MSA: 0.875, MCA: 0.813) models. The DA model had a 11.6% performance increase for MSA and a 12.3% increase for MCA compared to LIN S_{large}-T, the best linear model. In addition to being the top performer, DA G+S_{large}-T also was the most robust model for this task, having the lowest variation in its results (st$d$=0.004 for MSA and st$d$=0.006 for MCA) (Supplementary Table A.7). Prediction for TCGA was repeated with the models trained for SRA tissue pre-

diction (previous section), i.e., trained on S_{small}, which allowed us to assess the influence of bias injection on model performance. Whereas the addition of more SRA data to the training data had little influence on LIN models (except for a slight increase of ∼0.2 ppts for LIN G-S_{large}-T), both MLP and DA model accuracies improved significantly (5 to 9 ppts) upon addition of additional SRA data (Supplementary Table A.7).

Notably, adding 5,480 GTEx training samples to MLP S_{small} (MLP-S_{small} -> MLP G+S_{small}) increased MSA from 0.748 to 0.764 and MSA from 0.688 to 0.716 on the TCGA test set. On the other hand, adding 1,531 SRA samples (MLP-S_{small} -> MLP S_{large} increased MSA to 0.832 and MSA to 0.755, underlining our model's ability to incorporate multiple biases for better generalization (Supplementary Table A.7).

Multi-Bias Data Enhances Tissue Classification

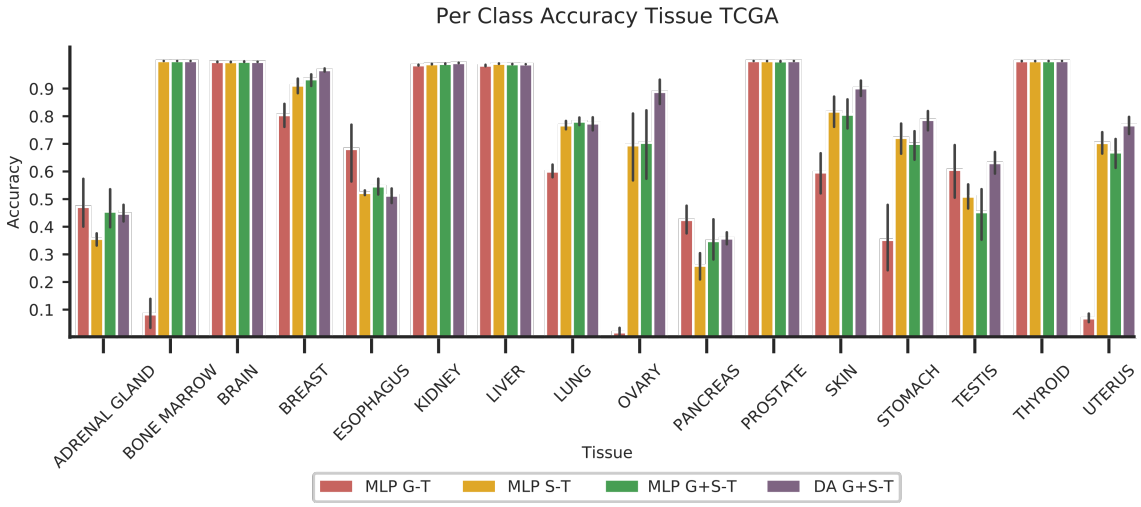

Figure 2.11.: Per class accuracy for TCGA tissue classification. Mean sample accuracy for each tissue and all ANN based models is shown. The error bar shows the standard deviation across 10 random seeds. The plot demonstrates the varied tissue classification performance of different tissues. For instance, it seems to be difficult to identify adrenal gland or pancreas with any of the models. In particular, the bad classification performance of MLP G-T for bone marrow, ovary and uterus is especially noticeable, along with the observation that performance can be salvaged by addition of (biased) SRA data to the training dataset. This highlights the strength of ANN based models in capturing bias from training data.

For tissue classification on TCGA, mean class accuracy increased by 16.8 ppts between MLP G-T and MLP G+S_{large}-T. This result confirmed our hypothesis that the GTEx data's homogeneity did not allow the MLP G-T model to generalize to TCGA data, while the addition of SRA training data in MLP G+S_{large}-T resulted in a model with significantly improved generalization. To further investigate this result, we took a closer look at the per-class accuracy for the TCGA tissue prediction (Figure 2.11). MLP G-T was unable to predict samples for three tissues, namely bone marrow (MSA: 0.08), ovary (MSA: 0.02) and uterus (MSA: 0.07), whereas all our

other models achieved accuracies between 0.7 and 1.0 on these tissues. Adding SRA data to the training set enabled the model to achieve per tissue sample accuracy of 1.00, 0.704 and 0.67 for bone marrow, ovary and uterus, respectively. We used principal component analysis (PCA) to visualize the dataset bias for these tissues. Interestingly, the GTEx-ovary and TCGA-ovary data points show little overlap in the PCA plot, while the SRA-ovary data overlaps with GTEx and TCGA-ovary data forming a 'bridge' (Figure 2.12A). Other tissues such as liver (MLP G-T MSA: 0.98), on the other hand, show an overlap between the GTEx and TCGA data which is reflected in the consistent accuracy across all models (Figures 2.11 and 2.12B).

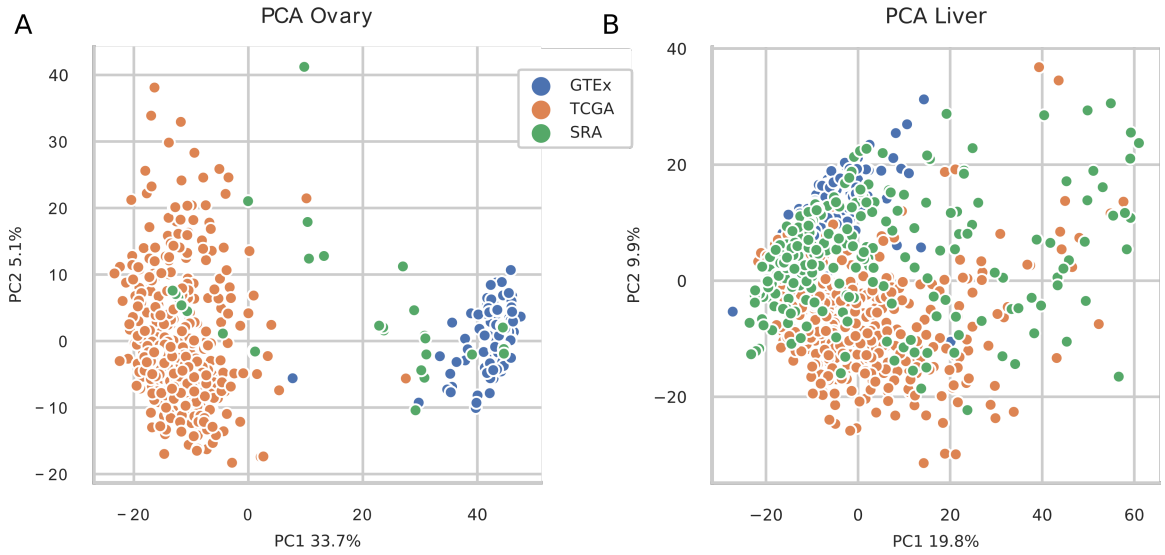

Figure 2.12.: Bias visualization. Principal Component Analysis on gene expression of available GTEx (blue), TCGA (orange) and SRA (green) samples for A) ovary and B) liver tissue. For ovary tissue samples GTEx and TCGA data do not overlap and SRA data is needed for proper model generalization.

Improved Sex Prediction with ANNs

For sex classification, only genes on the X and Y chromosome were used as input features (d=190). We first tested the trivial case MLP G-G by splitting GTEx into training and test sets, achieving sample and class accuracy of 0.995 (data not shown).

Prediction of TCGA SEX - Sex phenotype prediction on TCGA data was the only task where we could not perform significantly better than the linear model. The baseline LIN G-T and the other linear models LIN S_{large}-T and LIN G+S_{large}-T achieved almost perfect accuracy on the TCGA data (MSA/MCA 0.989 for LIN G-T and LIN G+S_{large}-T, MSA 0.988 and MCA 0.987 for LIN S_{large}-T). Based on the data annotation provided by MetaSRA, our best model was MLP G+S_{large}-T with MSA 0.947 and MCA 0.945 (Figure 2.13).

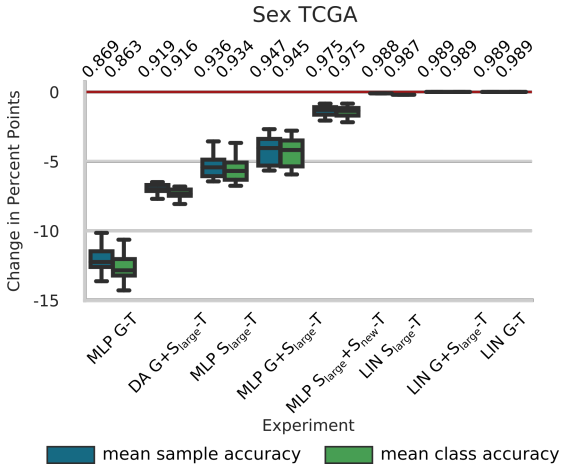

Figure 2.13.: TCGA sex classification results. Results are reported as change in percent points compared with the baseline model LIN G-T. Sample (blue) and class (green) accuracy are shown. LI*n*=linear model, MLP=multilayer perceptron, DA=domain adaptation, G=GTEx, S=SRA and T=TCGA. ANN based models yielded consistently worse results than the baseline model, until newly annotated data were incorporated into the training set.

Prediction of SRA Sex - All linear models for the prediction of sex for SRA data achieved an accuracy (MSA: 0.883 and MCA: 0.876 for LIN G-S and LIN G+S_{small}-S, MSA: 0.878 and MCA: 0.873 for LIN S_{small}-S) similar to what was previously reported (MSA: 0.863 [1]). The MLP G-S model (MSA: 0.879 and MCA: 0.871) did, on average, perform worse than all the linear models (Figure 2.10C). While adding SRA data to the training set did not improve the LIN model, it increased the MLP and DA models' performance. DA G+S_{small}-S (MSA: 0.929 and MCA: 0.93), MLP S_{small}-S (MSA: 0.93 and MCA: 0.936) and MLP G+S_{small}-S (MSA: 0.939 and MCA: 0.945) differ statistically (p-value <8e-5, ANOVA). A t-test corroborated that MLP G+S_{small}-S is statistically the best model (p-val=0.0066, t-test) with a performance increase of 6.3% for MSA and 7.9% for MCA compared to the best linear model LIN G-S. Results are shown in (Figure 2.10C).

According to MetaSRA, all our training and testing data for sex prediction on SRA stem from patient biopsies. However, at least two of the largest misclassified SRA studies in the test set are cultured cell lines. For example, SRP056612 is a study on the coronavirus's effect on cultured kidney and lung cells [64], and SRP045611 is a study involving HEK cells, which lack the Y chromosome but are annotated as male by MetaSRA [65]. These are two examples of mislabeled SRA data. Mislabeled data can compromise classifier accuracy, either by providing the wrong ground truth for training or by reporting the false label at the point of prediction.

Expression Based Prediction of Sample Source

SRA data stems from multiple different sources, from which we selected the two largest, namely biopsy and cell lines. As all GTEx and TCGA samples are exclusively from biopsies, models were only trained on SRA data. Of note, while we were able to approximately reproduce the

initial results for LIN S_{large}-G and LIN S_{small}-S, we could not do so for LIN S_{large}-T (MSA: 0.882 vs. 0.998 reported in [1]). LIN S_{large}-G (MSA/MCA 0.951) did slightly better than MLP S_{large}-G (MSA and MCA of 0.943). MLP S_{large}-T achieved MSA and MCA 0.971, outperforming LIN S_{large}-T with (MSA and MCA of 0.882). MLP S_{small}-S achieved MSA 0.95 and MCA 0.941, outperforming LIN S_{small}-S with MSA 0.89 and MCA of 0.884 (Figure 2.10D).

Training Data Diversity Outweighs Quantity

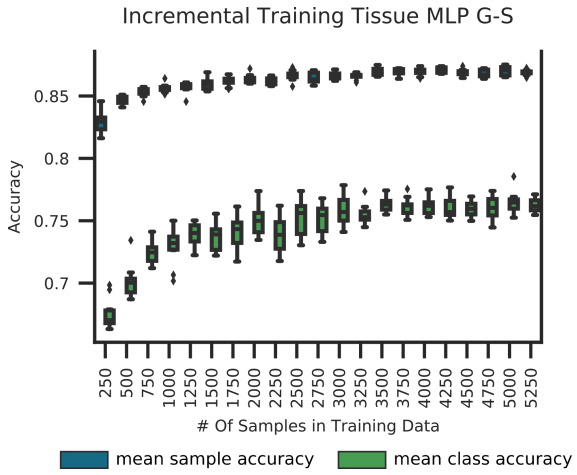

Figure 2.14.: Dependence of prediction performance on increasing training dataset sizes for MLP G-S. MLP models were trained on subsets of the GTEx data for SRA tissue classification on 10 seeds and averaged. At each step the subset was increased by 250 samples. Box Plots from 20 iterations for the MSA and MCA are shown in blue and green, respectively. Mean sample accuracy reaches its peak with only 25% of the training data while 50% of the data is sufficient for the mean class accuracy to saturate.

Our experiments on phenotype classification seem to indicate that increased training data diversity might enhance classification performance. MLP G-S was trained on an increasingly large subset of the GTEx training data for tissue classification to learn more about the relationship between the amount of training data and model performance. We observed a limited effect on model performance with increased training dataset size. The MSA reached its peak with one-third of the available training data, while the MCA saturated at about half of the available training data (Figure 2.14).

Training Data Diversity Outweighs Quantity

An increasing number of biases in the training set was used for training an MLP S_{small}-S for sample source classification to test the effect of bias in the training data. As a control experiment, an MLP was trained with the same amount of data but drawn from a single-bias source. We observed a positive correlation between MSA and the number of biases in the training set (Figure 2.15A). In contrast, increasing the number of training samples by the same amount but from a single-bias source did not lead to better model performance (Figure 2.15B), validating

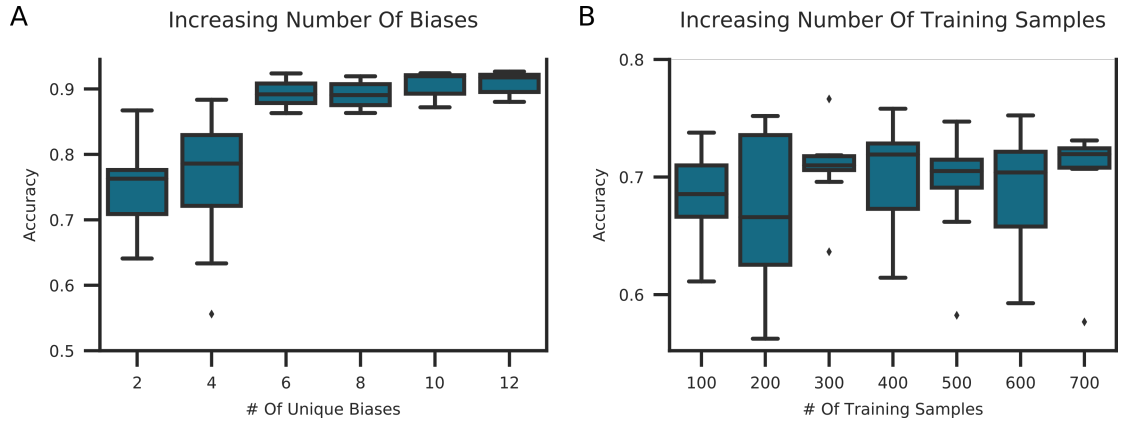

Figure 2.15.: Increasing bias vs. increasing sample size in training data. A) A MLP S_{small} for sample source prediction on SRA data was trained by randomly sampling an increasing number of SRA studies per class. Each study was subsampled to 50 samples. Studies were drawn from all SRA studies with $n > 100$ for either sample source tissue or cell line. B) To differentiate the effect of increased bias vs. increased sample size, the same model was trained by randomly subsampling the largest available SRA study per class. At each step an additional 50 samples were added to the training set per class. Models were run with 10 different seeds and the mean sample accuracy was computed. Box plots are produced by 10 random sampling iterations. We observe a positive correlation between training data diversity and accuracy.

our assumptions. Both experiments support our assumption that ANN-based models can integrate different biases in the training set and translate them into better model performance compared to other methods.

Prediction and Availability of Novel Metadata

We have used our best models to predict high-quality metadata for published SRA samples lacking information on tissue, sex, or sample source. Prediction of sex is straightforward because our models were trained on all possible biological categories. However, for tissue and sample source, our models were trained on a subset of all potential classes in the unlabeled data. If, for example, we try to label a sample of a tissue type unknown by the model, the model will force one of the learned classes onto that sample. To avoid this for sample source classification, we modified the classification task into 'one vs. all'. Specifically, a new MLP binary-class model was trained (biopsy vs. all) using all SRA data labeled by MetaSRA. This model (i.e., MLP S_{small}-S) achieved MSA 0.947 and MCA 0.93 on a test set (data not shown), and MLP S_{large} was subsequently trained and applied to identify all as of yet unannotated SRA samples of source type biopsy in our data. A total of 1,072 new SRA biopsy samples were identified.

Second, the tissue classification task was extended to 17 classes. A 'catch-all' class was added following the same reasoning as above. To this end, we extended the training data to all GTEx (n=9,366) and SRA (n=6,183) data with tissue labels and assigned the placeholder class for every sample that did not belong to the original set of 16 tissues. With this approach, the DA

$G+S_{small}$ model achieved MSA 0.912 and MCA 0.787 (data not shown). Training and test sets were subsequently combined to train DA $G+S_{large}$ for annotation prediction of unlabeled SRA samples. The tissue of origin was predicted for all SRA samples of source type biopsy for which no entry on MetaSRA was available (n=2,818). Third, 8,495 SRA biopsy samples with missing sex information were predicted using MLP $G+S_{large}$. Figure A.5 shows the true positive rate for each phenotype and each class on the test set.

Finally, we used the newly annotated data to improve our models (Figure 2.8C). To determine the additional tissue training set, we chose a probability cutoff of 0.9 and removed all brain samples (n=1,057) to avoid a further increase in class imbalance, adding a total of 530 new SRA samples to the training set for tissue prediction (i.e., S_{new}). While no increase in performance was observed for TCGA classification, MLP $S_{small}+S_{new}$-S for tissue prediction increased in performance compared to MLP S_{small}-S from 0.894 to 0.911 (MSA) and from 0.746 to 0.798 (MCA). DA $G+S_{large}+S_{new}$-S achieved the best accuracy of all tissue classification models with MSA 0.933 and MCA 0.854. All newly annotated SRA samples were added (S_{new}=8,495) to build a new SRA training dataset for sex. MLP $S_{small}+S_{new}$-S for sex classification improved only slightly upon the previous best model MLP $G+S_{small}$-S with MSA 0.945 and MCA 0.948 (compared to MSA 0.939 and MCA 0.945). However, for classification on TCGA, MLP $S_{large}+S_{new}$-T yielded sample and class accuracy 0.975, 4.1 ppts higher than the MLP S_{large}-T model trained on our default SRA training set (Figure 2.13). We thus successfully identified novel training data and used it in a positive feedback to enhance our models, validating the high-quality of the new annotations.

2.7. Discussion

We developed a novel deep-learning-based domain adaptation approach for automated bias invariant metadata annotation. To the best of our knowledge, this is the first time domain adaptation has been applied to this problem. We were able to outperform the current best model [1] on tissue prediction by 3.3% for SRA and 12.3% for TCGA data on mean class accuracy. As previously reported [66], we can confirm that ANNs trained on single-bias training data do not perform better than linear models. However, given multi-bias training data, we showed that MLPs, especially our DA algorithm, have an advantage over standard machine learning approaches.

Our current models help researchers to verify the sex, tissue and sample type of an RNA-seq sample in the presence of bias. This metadata information is currently rarely given for datasets downloaded from the SRA but can be crucial. Our method's main strength is its ability to incorporate dataset bias from datasets with only a few samples by applying a siamese network-like architecture. The model learns to ignore bias by repeated exposure to (few) samples in (many) different contexts, i.e., as triplets. Besides, it does not rely on feature selection but uses normalized gene count tables and lets the network learn which features carry essential information.

Different types of experiments showed the importance of training models on a multi-bias dataset. First, we showed for every phenotype classification that models with SRA samples included in the training data performed better than models trained only on GTEx data. For tissue classification, we further showed that the effect of adding SRA samples to the training data outweighs adding 3.2x as much GTEx data (MLP S_{small} -> MLP S_{large} vs. MLP S_{small} -> MLP G-S_{small}). Second, for SRA tissue classification, we showed a diminishing return on performance increase achieved with increasing training set size. Our experiment showed that peak accuracy is already reached by using 50% of the available data. Lastly, we directly compared the relationship between the number of biases in the training data, the number of samples, and the model performance for sample source classification. We found a positive correlation between the diversity of the training data and the accuracy achieved by that model.

Lastly, we generated novel metadata for SRA samples using our best performing models, adding over 10,000 novel metadata entries for 8,495 SRA samples. We established a positive feedback loop by re-training the existing models for phenotype prediction by adding the newly annotated data to the training set. Expanding the SRA training data worked exceptionally well for TCGA sex classification, where an additional 4.1 ppts in accuracy was achieved. The newly generated metadata is now publicly available and can be used for future research. We see this as a first and essential step in the general direction of making publicly available data more accessible and reusable in an automated way.

We observed some limitations to our DA approach. Our experiments showed that the DA model does not perform as well as the MLP for classification tasks with a low number of classes (e.g., sex). At least for the TCGA tissue classification, it seems that a minimum of about 8 classes

is needed for the DA model to unfold its full potential consistently. Our experiments indicate that the difference between DA and MLP performance will keep increasing, in favor of the DA model, the more classes we add (Figure A.6). Adding more tissue classes to our model is an important next step. Another limitation is posed by the need for labeled data to train the bias mapper. The limitations imposed by the need for labeled data could be avoided by using an unsupervised algorithm. For example, autoencoders [67], a type of neural network, use unlabeled data to learn a compact representation of input data and have successfully been applied to genomic expression data before [22, 66]. Briefly, autoencoders are split into two parts, an encoder and a decoder. The encoder learns a lower-dimensional embedding of the input (i.e., encoder output nodes « input nodes). The decoder takes as input the embedding space and aims to reconstruct the input data (i.e., the autoencoder's objective is input = output). Our method could be adapted in the following way. First, an autoencoder is trained on the source domain. Second, an MLP consisting of the source encoder (with fixed weights) and a classification layer is trained on the source domain using class labels. Third, a second bias autoencoder is created with the weights of the embedding layer equal to the fixed weights of the trained source embedding layer. Four, the bias autoencoder is trained on the bias domain. The bias encoder learns to map the biased data into the same embedding space as the source encoder. Next, the trained bias encoder could now be linked to the pre-trained classification layer and used as a classifier.

Whereas currently, the scope of our predictive models has been limited by the availability of data (e.g., intersecting tissue types between datasets, the limited size of datasets), the approach is ready to incorporate more data, biases, classes, and more phenotypes. There is reason to believe that this will confer increased ANN-based models' performance, in particular DA models. Simultaneously, automated annotation ensures that the vast amount of data currently lying idle in online repositories and institutional data centers can indeed be leveraged. We believe that this synergy can produce an extensive and comprehensive body of annotated biological data to boost knowledge discovery for biomedical research.

3. Analysis of Microbial 16S rRNA-Seq Data of the Blood

3.1. Background

3.1.1. Microbiome

The microbiome is defined as a microbial community occupying a habitat that has distinct physio-chemical properties [68]. More precisely, it refers to the abundance and richness of microbes involved and their biological activity (i.e. their collective genomes). Thus, the microbiome forms a dynamic and interactive identity prone to change in time and space, integrating into an ecosystem or host playing a crucial role in metabolic function [68].

The human microbiome is thought to match the number of somatic and germ cells in our bodies [69] and provide more than 100 times the number of genes in our genome [70]. By supplying genes that encode for functions that humans have not evolved, microorganisms take over critical metabolic processes in our bodies. For example, our distal intestine bacteria degrade otherwise indigestible polysaccharides [70].

3.1.2. 16S Ribosomal RNA Sequencing

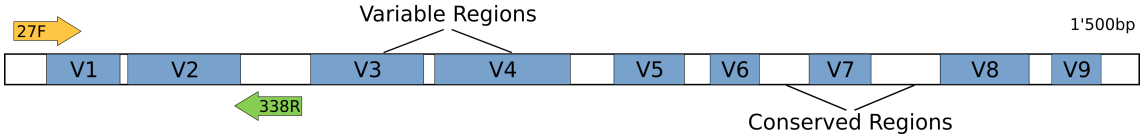

Figure 3.1.: 16S ribosomal RNA gene of prokaryotes. The 16S ribosomal gene can be divided into strictly conserved and 9 highly variable regions (blue). Conserved regions can be used for primer alignment to sequence variable regions. For this study, the 27F forward (orange) and the 338R (green) reverse primer were used to amplify the V1-V2 region.

The sequence of the 16S ribosomal RNA (rRNA) gene has been used for phylogenetic analysis since the late '70s [71]. Highly variable regions (V1-V9) (Figure 3.1) within the gene can be sequenced to establish the phylogenetic history and taxonomic classification [72].

16S rRNA sequencing has been applied for many decades [73] (Figure 3.2). The 16S rRNA gene is targeted by primers flanking a specific variable subregion. During sequencing errors can be introduced, making it challenging to identify the identical sequence amplicons. Previously this problem has been solved by creating so-called operational taxonomic units (OTU) (Figure 3.3). OTUs can be created *de novo* by clustering amplicons of high similarity (usually 97%) or by assigning amplicons to known reference OTUs. The results are dependent on the dataset or a reference frame subject to change [74]. In addition to these reproducibility problems, OTUs cannot resolve fine-scale variations, which are essential to determine population structure on the species level [75]. New methods have recently been proposed to recover true biological sequences from 16S rRNA sequencing data [76, 77, 78]. These methods infer parameters of an error model for each sequencing run and use it to determine the true biological sequence of an amplicon. The resulting amplicon sequence variants (ASVs) are independent of

the data and can resolve single nucleotide differences resulting in a finer taxonomic classification [74].

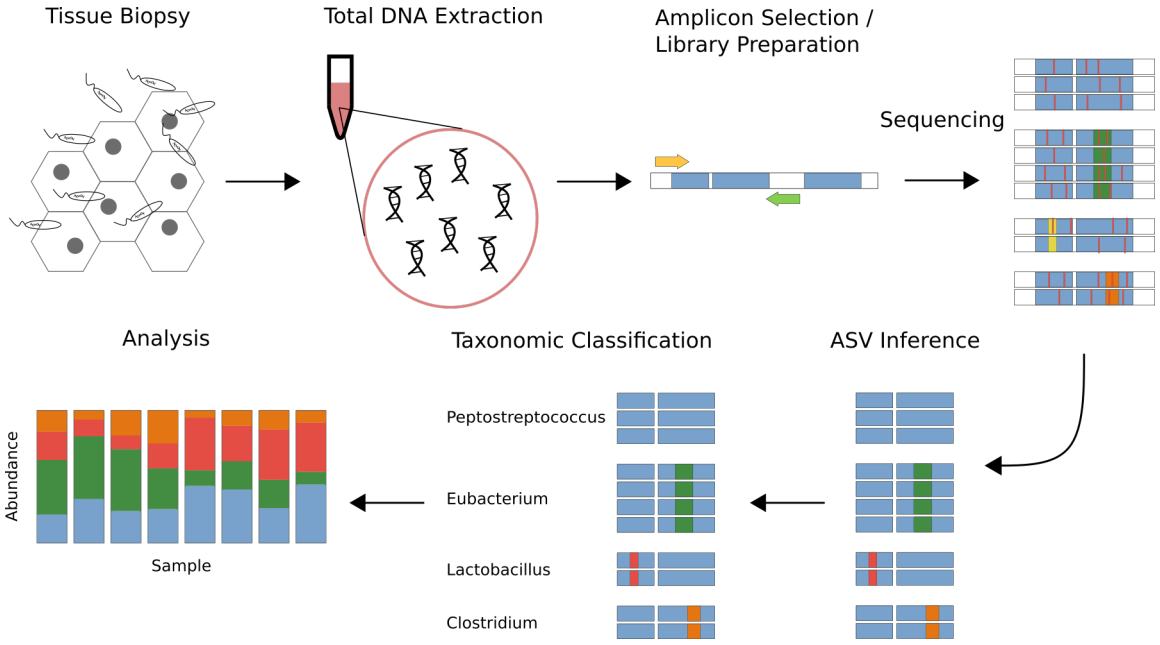

Figure 3.2.: 16S RNA sequencing workflow. After tissue biopsy (or sampling any other environment, e.g., soil), DNA is extracted, and the desired 16S ribosomal gene region is amplified. Sequencing errors lower the possible resolution of taxonomic classification. Amplicon sequence variants (ASVs) are inferred, taking error models into account to denoise the reads. After taxonomic classification, the data can be analysed.

3.1.3. Microbial Imbalance and Disease

Microbial imbalance in or on the body, known as dysbiosis, has been linked to many diseases [79]. One of the most studied microbial environments is the gut and the so-called gut-liver axis [80]. For example, differences in the gut microbiome of patients with chronic liver disease (e.g., cirrhosis) [80, 81] and inflammatory bowel disease (IBD) [82] are well established. For example, dysbiosis in cirrhosis can be caused by certain bacteria's overgrowth, releasing endotoxins such as bacterial lipopolysaccharides (LPS). Endotoxins can enter the blood circulation by bacterial translocation through the intestinal barrier. A positive correlation between LPS concentration in the blood and the severity of chronic liver disease has long been established [83]. It has been shown that endotoxins and whole bacteria can use the translocation route to enter circulation in cirrhosis patients [84].

Increased Intestinal Permeability

Translocation across the intestinal membrane can correlate with a condition known as 'leaky gut' [85]. An increased intestinal permeability was established as a primary defect in IBD 35

years ago [86]. More recently, Dhillon et al. [87] compared gut leakage markers in primary sclerosing cholangitis (n=166) and healthy controls (n=100). The authors were able to show that soluble CD14 (sCD14) and lipopolysaccharide-binding protein (LBP) concentration was significantly increased in the blood of primary sclerosing cholangitis patients. LPB is an acute phase protein (i.e., present during inflammation response), while sCD14 is a protein made mostly by macrophages to detect gram-negative bacteria. Furthermore, a decrease in zonulin concentration was reported, a known physiological regulator of intercellular tight junctions [87]. LPS entering the liver interacts with a specific receptor of the adaptive immune system (toll-like receptors) and can provoke an immune response [88]. An exaggerated immune response can lead to tissue damage in the liver, leading to fibrosis.

Primary Sclerosing Cholangitis

Primary sclerosing cholangitis (PSC) is a rare, progressive disease of the liver. It is characterized by inflammation and fibrotic strictures (sclerosing) in the bile ducts (cholangitis). In northern Europe, the incidence is estimated to be 1-3 per 100,000, and about 70-90% of all cases have comorbidity with ulcerative colitis which is a type of IBD [89]. The condition can eventually lead to cirrhosis, and more than 50% of patients need a liver transplant within 10-15 years of symptom development [90].

Genetic studies have accumulated many gene associations, but the combined impact of genetic susceptibility is less than 10% [91, 92] . The strongest associations are localized in the human leukocyte antigen and suggest an adaptive immune response [91].

The microbiome in PSC has recently received focused attention [92]. Cross-sectional studies comparing patients with PSC and healthy controls showed a decrease in alpha diversity (within-group diversity) in PSC patients [93, 94, 95, 96, 97]

Primary Biliary Cholangitis

Similar to PSC, primary biliary cholangitis (PBC) is a chronic and slowly progressive liver disease associated with autoimmune events. It is more prevalent than PSC, with an incidence of 1.91 to 49.2 per 100,000 inhabitants [98]. PBC primarily affects women with a ratio between 1.6 to 10 [98]. Unlike PSC, for PBC, administration of ursodeoxycholic acid, a microbial product, is a therapy to which most patients respond.

Similarly to PSC, environmental and genetic risk factors have been determined and linked to immune tolerance's breakdown [98]. An altered gut microbial profile has been reported for PBC patients [99, 100] both as treatment target and disease marker [99]. The difference between PBC and healthy controls does not seem to be as strong as with PSC [101]. However, both conditions show enrichment of specific species, e.g., *Streptococcus*, *Haemophilus* and *Veillonella*.

3.1.4. Microbiome in Blood

The blood, like other bodily fluids, is considered to be sterile. Preventing pathogenic microorganisms from invading the blood is, after all, the job of the immune system and the epithelial barriers. However, with the advent of RNA-sequencing technologies, the 16S ribosomal RNA gene has been first detected in human blood samples in 2001 [102]. The study of bacterial presence in the blood, both for healthy and diseased patients, has since received heightened attention [103].

The presence of a blood microbiome is heavily debated [103]. Some groups have explicitly given up on the idea [104]. It is well known that the signal of low-biomass environments might be drowned out by bacterial DNA contamination during sampling or sequencing (i.e., laboratory reagents) [105, 106, 107, 108, 109]. The low signal to noise ratio makes it especially hard to pin down the potential microbial composition in the blood.

Nevertheless, dysbiosis in the blood has been linked to schizophrenia [110], celiac disease [111], cardiovascular events [112], type 2 diabetes [113] and cirrhosis [114, 115]. All these studies show significant differences between groups to support their hypothesis. However, they exhibit a sizable between-study variance. For example, Santiago et al. [114] were able to detect microbial DNA in cirrhotic patients only, while the control samples did not yield enough biomass for analysis (less than 1,000 reads per sample). On the other hand, comparing patients with type 2 diabetes mellitus with healthy controls, Qiu et al. [113] reported an average read count of more than 60,000 per sample, including 100 healthy patients. Furthermore, while some report to have found 23 distinct phyla in the blood [110], others report six phyla [114] or one phylum (*Proteobacteria*) to dominate 99.58% of all reads [113]. However, the consensus across many studies is that the phyla *Proteobacteria*, *Actinobacteriota*, *Firmicutes* and *Bacteroidetes* [103] dominate the bacterial DNA found in the blood.

3.2. Aim and Problem Statement

PSC and PBC are two rare, progressive diseases of the liver. While a therapeutic intervention for PBC is available, more than 50% of all PSC patients will need a liver transplant [90]. The etiology of both diseases is unknown, making it challenging to develop new treatment avenues. Genetic risk factors have been established for both diseases, mainly pointing to the immune system [91, 92, 98]. PSC patients have comorbidity with IBD of 70-90% [89]. It is well known that IBD patients show an increased intestinal permeability [86] and dysbiosis in the intestine [82]. Similarly, it has recently been shown that PSC patients show increased concentrations of gut leakage markers associated with immune response [87]. A link between a microbial imbalance in the gut between PSC / PBC and healthy controls has been confirmed for several gut tissues [92, 93, 94, 95, 96, 97, 99, 100]. These findings support the theory of translocation of pro-inflammatory bacteria or their components from the gut into the portal circulation.

This study analyzed PSC and PBC patients' blood microbiome and compared it to healthy controls. While the blood microbiome has been studied for patients with cirrhosis [114, 115], to the best of our knowledge, this is the first microbial study of PSC / PBC patients of the blood. Analysing the provided that we were able to confirm an increase in microbial diversity in the blood of PSC and PBC patients compared to healthy controls thereby supporting the 'leaky gut' hypothesis.

3.3. Methods

3.3.1. Data

We received raw 16s rRNA sequencing reads from the Institute of Clinical Molecular Biology at the University of Kiel. Samples were collected at the University Medical Center Hamburg-Eppendorf. Serum samples were collected for 10 PSC patients, 10 PBC patients and 8 healthy controls (HC). For each patient, blood was drawn at three different time points: i) Before breakfast (time 0), ii) 5 min after breakfast and iii) 60 min after breakfast. Approximately 200 µl serum was used for DNA extraction, and the variable regions V1 and V2 of the 16S rRNA gene were amplified using the primer pair 27F-338R. Paired-end sequencing was done using Illumina MiSeq v3 2x300bp (Illumina Inc., San Diego, CA, USA). The following controls were provided in addition to the patient sample: MOCK (ZymoBIOMICS Microbial Community DNA Standard), no template control (NTC) and an empty DNA-extraction control.

3.3.2. ASV Inference With Divisive Amplicon Denoising Algorithm (DADA2)

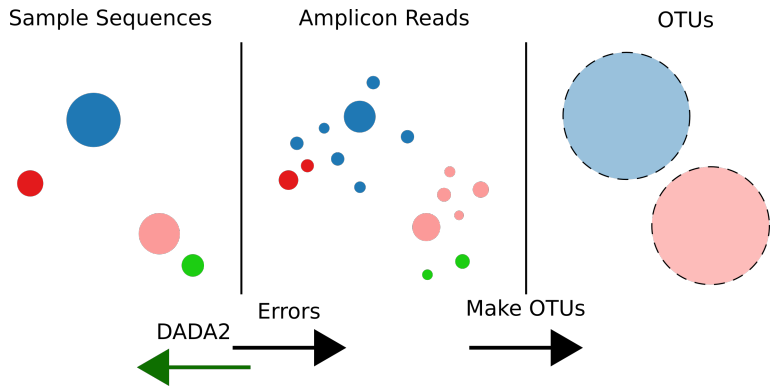

Figure 3.3.: Dealing with sequencing errors. After DNA extraction, samples of true biological sequences are present (colored circles left panel). Noise is introduced during the sequencing process, making it difficult to assign amplicon reads to their taxon of origin (scattered circles of equal color middle panel). Operational taxonomic units (OTUs) can be formed by clustering similar amplicons together for taxonomic assignment. One drawback of this method is the loss of taxonomic resolution (right panel). Algorithms such as the Divisive Amplicon Denoising Algorithm (DADA2) can be used to infer true biological sequences. Figure adapted from https://callahanlab.cvm.ncsu.edu/publications/.

DADA2 [76] is an algorithm implementing a complete workflow to produce error-corrected ASV tables from raw sequencing input. DADA2 uses a parametric error model to infer the true biological sequence. The model depends on input amplicon abundance and distance (in single base difference) between amplicons. It is assumed that true reads are likely more abundant and that minor differences between true clusters and reads are likely error-derived. A nucleotide substitution model is calculated based on the quality scores provided in the raw data. Based on read abundance, distance and the error model, p-values for cluster assignment are calculated.

For ASV inference, the *denoise-paired* function of the DADA2 plugin in QIIME2 [116] (version 2020.6) was used. DADA2 takes as input FASTQ files (raw sequencing files with quality information). Samples were processed for each Illumina run separately. Reads were quality filtered by truncating forward reads to 240bp and backward reads to 160bp and allowing only one expected error per read. Error rates were learned for each sequencing run, and reads iterative assigned to ASV clusters. Next, inferred forward and backward reads were merged, and the merged reads of both runs are combined into one ASV table. Merged reads smaller than 285 bps were removed.

3.3.3. Normalization

In any given environment, some species are more abundant than others. Low abundant taxa need to pass a sequencing threshold to be detected. If more reads are generated for a sample, more rare taxa pass the threshold. The number of reads generated is proportional to the starting concentration of DNA and other factors. Samples need to be normalized to be comparable. Rarefaction [117] has been the standard normalization method in microbial ecology for many years [118]. Samples are rarefied to a pre-selected size by random subsampling without replacement. Rarefaction curves, relating sampling depth to sample diversity, can determine the dataset's optimal sample size. The method has some drawbacks. For example, ASVs of low abundance could be lost during subsampling and statistical power reduced by throwing out data. Rarefaction has recently been critically reviewed from a theoretical perspective [119, 120] as well. Subsequently, other normalization methods have been developed [118] in the past years. Nevertheless, it has recently been shown that rarefaction is still one of the best normalization techniques available for this type of data [118, 121].

3.3.4. Alpha Diversity

Alpha diversity attempts to quantify the microbial within-sample diversity. Diversity indices are models that try to capture species richness (number of species) and abundance (proportion of species) simultaneously. Here we use the Shannon Index h [122] as suggested for situations where we do not emphasize rare or abundant taxa [123]. We use the implementation in scikit-bio (v 0.5.5), which is calculated as follows:

$$h(\vec{\boldsymbol{p}}) = -\sum_{i=1}^{S} p_i log_2(p_i)$$

Where p_i is the fraction of counts of the i^{th} ASV $\in S$. Willis [120] raised some concerns about rarefaction and alpha diversity estimation under the assumption that deeper sequencing leads to higher diversity. However, our rarefaction curves showed that alpha diversity saturated quickly for the three conditions.

3.3.5. Beta Diversity

Beta diversity attempts to quantify the between-sample diversity. It is a measurement of change in community composition between two environments. Here we applied the abundance-based Bray-Curtis (BC) distance [124], the most widely used and most robust beta diversity index [125]. We used the scikit-bio (v. 0.5.5) implementation, which is calculated as follows:

$$bc(\vec{\boldsymbol{x_j}}, \vec{\boldsymbol{x_k}}) = \frac{\sum_{i=1}^{S} |x_{ji} - x_{ki}|}{\sum_{i=1}^{S} (x_{ji} + x_{ki})}$$

Where x_{ji} and x_{ki} are the number of the i^{th} species $\in S$ of sample j and k, respectively.

3.3.6. Taxonomy Classification

A pre-trained Naive Bayes classifier was downloaded for the QIIME2's *feature-classifier* function (Silva version 138[1] [126]). ASVs that were not classified below the kingdom level (i.e., only as bacteria but no phylum) were removed (n=53), leaving 5,263 ASVs with a taxonomic assignment. A total of 5,071 ASV were assigned at family level, and 4,527 AVS were labeled at the genus level.

3.3.7. Statistical Tests

The applied tests are assumed to be known and only summarised. If not mentioned differently, statistical tests were corrected for the covariates sex, age and BMI (Supplementary Table B.1). Age was binned into intervals from $min(age)$ to 30, 30 to 50 and 50 $max(age)$ BMI was binned into the intervals $min(BMI)$ to 18, 18 to 25 and 25 to $max(BMI)$.

Analysis of Covariance (ANCOVA) is an extension of ANOVA and incorporates one or multiple covariates. To perform ANCOVA type III sums of squares is used. This type of sums of squares tests for the presence of the main effect conditional on the effect of the covariates.

Repeated measures analysis of variance (rANOVA) is the variant of ANOVA for paired data. In rANOVA, the total sum of squares is partitioned into within-group variance, within-subject variance and the between-group variance.

Permutational multivariate analysis of variance (PERMANOVA) [127] is used on distance matrices. It compares groups and tests the null hypothesis that the centroids and dispersion of the groups are different. The sum of squares is the squared distance between two samples. Equivalent to ANOVA, tight clusters of samples will have low dispersion. Significance is established by calculating a number (typically 999) of F-statistics on permuted data and comparing the F-statistics ratio larger and smaller than the actual data score. P-values for PERMANOVA depend thus on a random process and no exact value will be given.

[1] https://docs.qiime2.org/2020.6/data-resources/

3.3.8. Differential Abundance With DEseq2

Differential abundance (DA) analysis was applied to quantify differences in abundance between groups. DESeq2 [128] is the standard algorithm used to identify differential gene expression and is also recommended for microbial abundance analysis [118, 128]. It is well known that within-group variance increases with the mean expression. DEseq2 uses dispersion instead of variance to measure variation, which accounts for a gene's variance and mean expression level. A generalized linear model using the negative binomial distribution is fit and individual genes are tested for differential expression using Wald Test. P-values were automatically adjusted using the Benjamini-Hochberg method.

3.4. Results

3.4.1. Identification and Removal of Potential Decontamination Improves Clustering

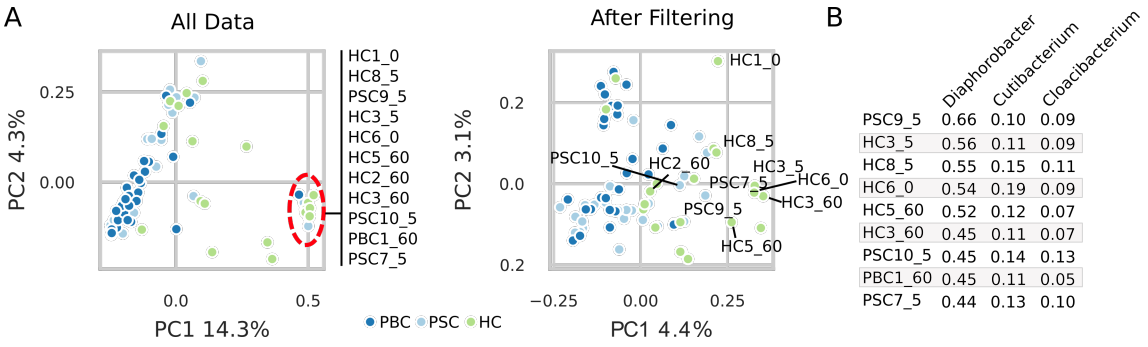

Figure 3.4.: Removal of major contaminants improved clustering. Between-sample diversity between all samples was calculated with the Bray-Curtis distance. A clear cluster of outliers formed containing samples from all conditions and time points (A, left panel). Members of that cluster showed a similar contamination pattern where more than 50% of reads belonged to known contaminants (B). After removing 37 genera previously reported to be contaminants, the outlier cluster dissolved (A, right panel), thereby validating the filtering.

We first confirmed the sequencing and downstream pipeline's validity by checking the positive control's MOCK community (Supplementary Figure B.1). Next, possible contaminants introduced during sequencing were detected using the R-package decontam (version 1.8.0). The NTC control had 1,703 counts across 5 ASVs and the reagent control 408 counts across 20 ASVs. In the patient samples, 8 ASVs were found to be equal with the ASVs found in the controls and subsequently removed. However, the mean count of 24,468 ($\pm$ 16,834) for low-biomass samples was still much higher than the 1,000-2,000 reads we would have expected [114, 108, 110] (5,316 AVS, 2,177,644 total reads). Between-sample diversity analysis using BC distance revealed an outlier cluster containing samples from all conditions and time points (Figure 3.4A, left panel). Compositional analysis for these samples showed a similar pattern of potential contaminants (Figure 3.4B).

Taxonomic classification identified 32 phyla (Supplementary Table B.2). First, we selected the 8 best-documented phyla in the gut [129] (*Proteobacteria*, *Actinobacteriota*, *Firmicutes*, *Bacteroidota*, *Deinococcota*, *Fusobacteriota*, *Spirochaetota* and *Vuryrrucomicrobiota*). Next, we inspected microbial content on the genera level. Some genera were obvious candidates for removal (i.e., *Cloacibacterium* 2.7% of all counts or *Legionella* with 1% of all counts) while others were less obvious. For example, *Diaphorobacter* present in 68 samples, making up 12.5% of all counts, is a close relative to a known contaminant [104], while it has been associated with IBD in cats and dogs [130]. As another example, *Cutibacterium*, present in 80 samples with 7.4% of all counts, is a well-characterized member of the skin microbiome. However, it has also been reported to be differentially abundant in PSC patients' bile fluid [97]. Of the 10 most abundant genera,

9 were listed as decontamination in the literature [104, 105] while accounting for 40% of all counts. With the help of literature, 37 genera were removed, accounting for 70% of all counts (Supplementary Table B.3). We decided on a stringent filter to reduce the false-positive rate. After filtering, 2,860 ASVs remained, and the mean library size dropped to 9,672 (± 7,279).

Removal of these contaminants resolved the cluster of highly contaminated samples (Figure 3.4A right panel). Given that 477 genera are left in the data, we can not be sure to have removed all contamination. However, we are confident that the biological signal has been significantly improved with our efforts.

3.4.2. High In-Patient and Between-Condition Variability Resulting from Undersampled Environment

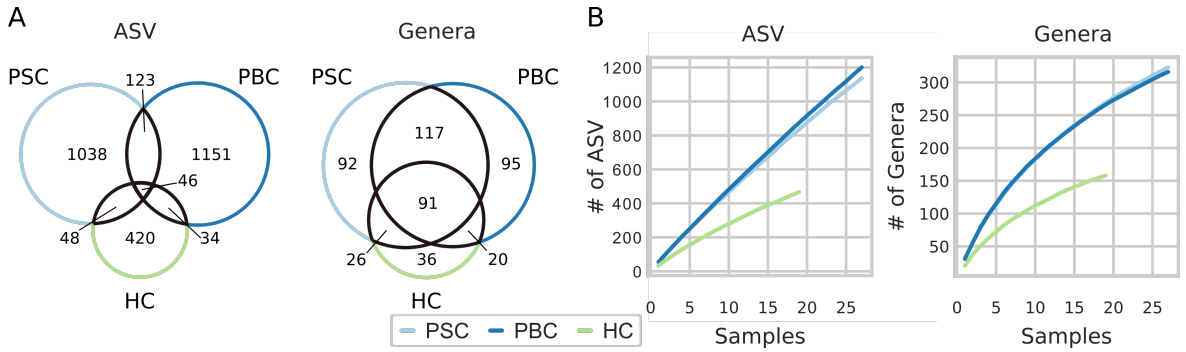

Figure 3.5.: Environments are not sampled to saturation. (A) Venn diagram of ASV and genera shared between PSC, PBC and HC showing a large between-condition variability. (B) Accumulation curves for ASV and genera for PSC (light blue), PBC (dark blue) and HC (green) showing an insufficient sampling of the environment.

Next we looked at the within-patient similarity to find out how similar replicates of patients are. The average of genera shared between replicates of patients was 3.4% (± 2.3%). On the family level, the average was 9.5% (±4.7%). While there is little agreement between all replicates, 21.1% (±10.2) of genera found in a patient are present in at least two out of three replicates. At the family level, this value improved to 35% (±7.6%). Next, we looked at the between-condition variability. Of the 2,860 ASVs, 1038 were unique to PSC, 1151 unique to PBC and 420 unique to HC (Figure 3.5A left). The overlap between groups was generally very low, with 1.2-4.3%. Of the total ASVs, only 1.6% are shared among all three conditions. On the genus level, 92 genera were unique to PSC, 95 unique to PBC and 36 unique to HC (Figure 3.5A right). The overlap between groups was lowest for PBC-HC with 4.2% and highest for PSC-PBC, with 24.5% of shared genera present in both groups. A total of 91 ASVs were shared across all three conditions.

For a qualitative assessment of how well each environment (PSC, PBC and HC) was sampled, accumulation curves were plotted (Figure 3.5B). Accumulation curves show the relationship

between diversity and the number of samples for an environment. For a sufficiently sampled environment, the accumulation curve is expected to saturate. Looking at the ASV and genera accumulations for each condition, we observed that the slope is almost linear for all three groups.

3.4.3. No Change in Microbial Diversity Observed After Food Intake

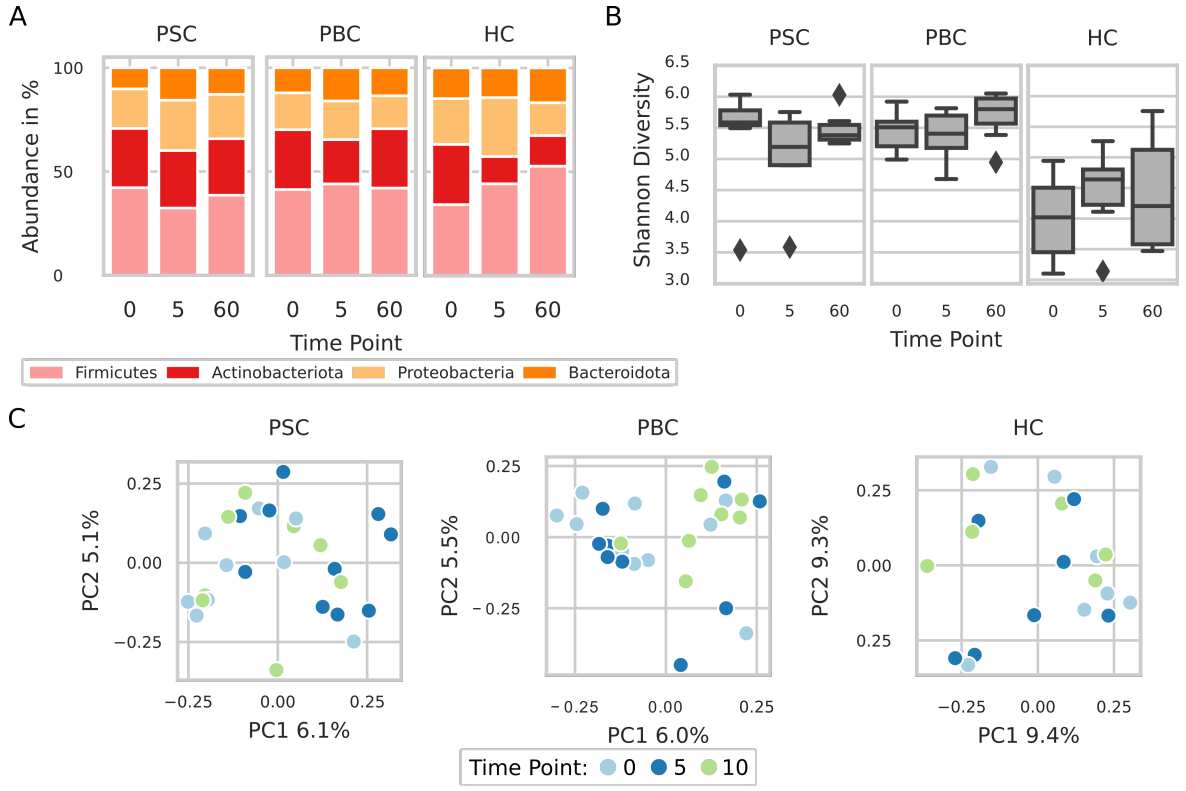

Figure 3.6.: Summary of time point analysis. A) Phylum abundance for samples rarefied to a depth of 1,500 of patients with replicates present in all time points (PSC n=8, PBC n=8, HC n=6). Only phyla with total abundance > 0.25% are shown. B) Alpha diversity box plots comparing within-sample diversity between time points within each condition. No significance was detected (p-values > 0.1). C) Between-sample diversity measured by Bray-Curtis distance showing no significant difference between samples (p-values > 0.1).

First, we wanted to investigate a potential within-group change in blood microbiome induced by food intake. For this, blood serum samples of patients with PSC (n=10), PBC (n=10) and HC (n=8) patients were obtained. Blood was drawn before (TP=0), 5 min after (TP=5) and 60 min (TP=60) after breakfast. For time point analysis, samples were rarified to a depth of 1,500 reads. Samples with less than 1,500 counts were removed, and only patients were selected that had a replicate at each time point (PBC n=8, PSC n=8 and HC n=6).

We first looked at the phyla abundance across the three patient groups and time points (Figure 3.6A). No differential abundance on the phyla level was detected between the four most

abundant phylum. Alpha diversity, or within-sample diversity, was determined for each sample using the Shannon diversity index (Figure 3.6B). Repeated measures ANOVA was performed, resulting in p-values > 0.1 for all three conditions. Beta diversity, or between-sample diversity, was determined using the Bray-Curtis distance (Figure 3.6C). PERMANOVA was applied to determine the statistical difference between the time points. Again no significant difference was found between the time points (p-value > 0.1 for all conditions).

3.4.4. PSC and PBC Patients Show Increased Within-Sample and Between-Sample Diversity

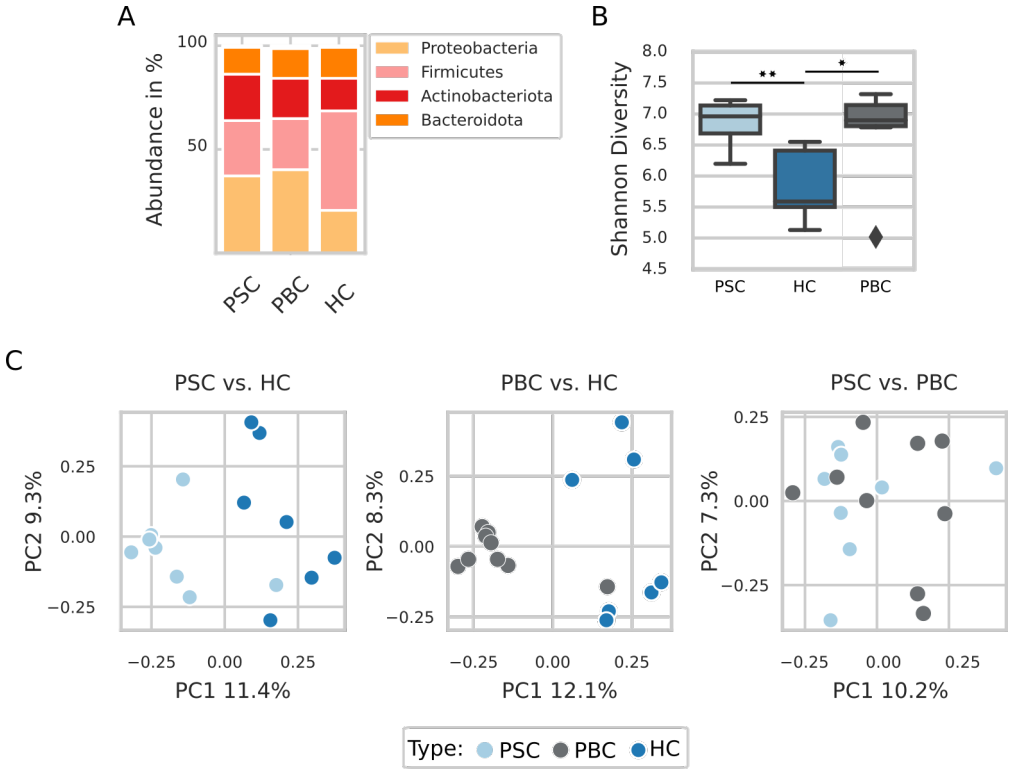

Figure 3.7.: Summary of between-condition analysis. Replicates were merged and rarified to 14,000 reads. (A) Abundance of most abundant phyla for each condition showing differential abundance for *Firmicus* and *Proteobacteria* between PSC / PBC and HC (p-value < 0.001). (B and C) showing alpha and beta diversity, respectively. PSC / PBC show significant higher within-sample diversity compared to HC (p-value < 0.05).

Next, we wanted to determine differences in microbial composition in the blood between PSC, PBC and HC, independent of time points. We were not able to find any systematic statistical difference between the patient's replicates. Therefore, we treated the different time points as technical replicates (i.e., samples sampled from the same environment under the same conditions). Technical replicates can be used to reduce noise inherent to the sampled environment or introduced during sequencing. We merged all available replicates per patient by summing

their ASV counts into one sample. Only merged samples with more than 14,000 ASV counts were retained. Filtering resulted in a reduced number of samples per group of 7, 8 and 9 for HC, PSC, and PBC, respectively.

Analyzing the phyla abundance (Figure 3.7A), we found *Proteobacteria* and *Firmicus* to be differentially abundant for PSC vs. HC with log fold change (LFC) LFC +0.77 and LFC -0.76, respectively. The same phyla were significant for PBC vs HC with LFC -1.13 for *Firmicus* and LFC +0.76 for *Proteobacteria* (all p-values > 0.001). Next, alpha diversity was calculated for each condition (Figure 3.7B). PSC was found to have a significantly higher within-sample diversity than HC (p-value = 0.011, ANCOVA). PBC also showed a higher alpha diversity, but only when not correcting for age (p-value = 0.028, ANCOVA). While all patients in the control group were under 40 years of age, all but one patient of the PBC group were above 40 years old (Supplementary Table B.1). No difference was detected between PSC and PBC. Beta diversity was again calculated using the Bray-Curtis dissimilarity (Figure 3.7C), and significance was determined using PERMANOVA. For the PSC vs. HC comparison, a significant difference in the between-sample diversity was detected (p-values < 0.01). Similarly to the alpha diversity, PBC vs. HC did not show a significant between-sample diversity (p-values > 0.1) when correcting for age. However, statistical testing showed a significant difference when age was left out at a p-value < 0.005. PSC and PBC showed a significant difference in between-sample diversity (p-values < 0.05).

3.4.5. Differential Abundance Analysis

We were interested in whether we could find differentially abundant taxa at the lower hierarchical level. Investigating taxa unique to one condition is trivial. Hence, we only focused on genera and families present in at least 50% of the patients for each condition. The family and genera uncultured were removed as these are independent of actual taxonomic classification. Given this restriction, we found 42 genera in PSC, 33 in PBC and 20 in HC (Figure 3.8A). A total of 62 genera were identified, 8 of which were shared across all three conditions. Similarly, 54 families present in at least half the samples were identified in PSC, 43 in PBC and 29 in HC (Figure 9A). Together they accounted for 63 unique families, with 18 of them shared between all three conditions.

We tested genera and families present in either PSC and HC, PBC and HC or PSC and PBC for differential abundance using DESeq2. For PBC vs. HC, we again excluded age as a covariable. The genus *Lachnospiraceae_ND3007_group* showed a LFC of -6.2 for PSC vs. PBC (p-value 0.037). The family taxon *Rhodocyclaceae* with a log fold change (LFC) -5.9 and *Sporomusaceae* (LFC -8.3) both showed a tendency to be lower in abundance in PSC compared to PBC (p-value 0.055 for both). Between PSC and HC, no differentially abundant genera were detected. On the family level, however, we identified *Microbacteriaceae* to have a LFC of +3.8 in PSC compared to HC (p-value < 0.001). The same result was found comparing PBC vs. HC on the family level, where *Microbacteriaceae* was found to have a LFC of +3.1 (p-value 0.005). In addition to this, we

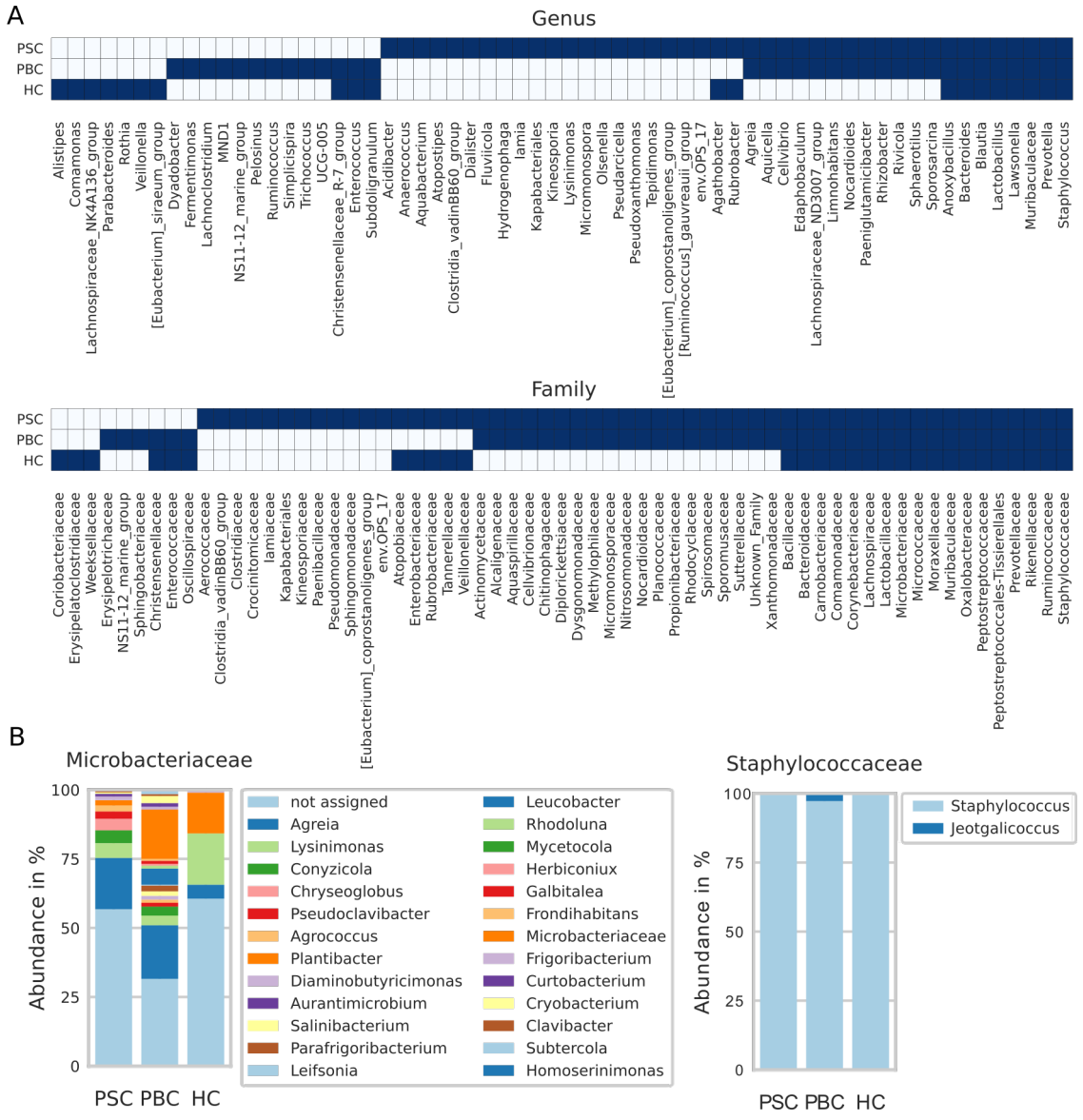

Figure 3.8.: Differential abundance. A) Genera and families that were present in at least 50% of samples for PSC, PBC and HC were selected (dark blue square) for differential abundance analysis. B) *Microbacteriaceae* was more abundant in PSC and PBC vs. HC (p-val < 0.01) and *Styphylococcaceae* was less abundant in PBC vs. HC (p-val > 0.01). *Microbacteriaceae* showed a more diverse set of genera in PSC and PBC vs. HC (left) while *Staphylococcus* was virtually the only genus found (right).

found the genus *Staphylococcus* to be less abundant in PBC compared to the controls (LFC -2.4, p-value <0.001) as well as the family it belongs to *Staphylococcaceae* (LFC -2.4, p-value < 0.001). Members of the family *Microbacteriaceae* were present in all three conditions, and 26 genera of the family have been detected (Figure 3.8B). Diversity within the family was different across the conditions, with PSC showing reads for 15 genera, PBC 22 and HC 5. Besides this, 43% of all reads assigned to *Microbacteriaceae* could not be assigned below the family level. Counts associated with *Staphylococcaceae* were almost exclusively assigned to the genus *Staphylococcus* (Figure 3.8B).

3.5. Discussion

PSC and PBC are two chronic liver disease associated with dysbiosis in the gut [92, 93, 94, 101, 95, 96, 97, 99, 100]. One theory currently investigated is that bacteria, bacterial components or bacterial endotoxins traverse the intestinal barrier into the portal circulation. This translocation would elicit an immune response, a disease pathway involved in IBD and other chronic diseases [85, 86, 87].

This study analyzed blood serum 16S Ribosomal RNA sequencing data from PSC, PBC, and control patients. Serum samples were taken from patients at three different time points, before, 5 min after and 60 min after breakfast. This experimental setup allowed us to investigate two questions. First, we analyzed if PSC / PBC patients show an altered blood microbiome possibly induced through digestion and increased intestinal wall permeability. Second, we investigated a difference in the blood microbiome between the groups independent of time points.

We first analyzed our samples for contaminants, a well-known issue with low-biomass 16S rRNA sequencing samples [105, 106, 107, 108, 109]. The negative controls provided from the sequencing center did not show significant contamination introduced during DNA-extraction and sequencing. However, unexpectedly large library sizes and a Bray-Curtis distance-based multi-condition outlier-cluster made us suspect contamination was still present. Others have reported a similar discrepancy between contamination found in negative controls and remaining contaminants in the data. Loohuis et al. [110], for example, reported 23 taxa in the blood while using religious positive and negative controls. This result was so unexpected that they replicated the whole experiment with a new set of patients and ended up with the same result. Schierwagen et al. [131] reported well-known reagent contaminants in their study, which prompted others to challenge their results [104]. However, it was later shown that negative controls were empty, and the claim that their samples were contaminated was refuted [132]. It is essential to know that blood samples in this study were not initially taken for 16S rRNA analysis, and sample contamination during that process is likely. With this in mind, we generously removed 70% of the data leading to the elimination of the contamination-based outlier-cluster (Figure 3.4A).

Undersampling of microbial environments is another issue that we had to consider [133]. Other than the apparent reason (i.e., not all species are equally distributed in the environment), a significant difference of species detected in samples of a given environment can depend on PCR primer selection, PCR template concentration and sequencing itself [134]. Considering PSC, PBC and HC as unique environments, they were vastly undersampled, as was shown by their respective accumulation curves (Figure 3.5B). Besides, the overlap between technical replicates (i.e., samples from the same patient) was 3.4% on the genus level. This overlap is far below the expected 13-20% [135, 134]. Due to the apparent undersampling of the environment, no statement about the true microbial diversity or abundance in PSC, PBC or healthy control patients is possible. However, a relative comparison in alpha and beta diversity can still generate important insight into disease mechanisms.

We first analyzed if food intake could trigger bacterial translocation. Within-condition samples were analyzed before food intake, 5 min after and 60 min after. No significant difference in phyla abundance, alpha, or beta diversity was detected. After 5 min, the patients' food has not had the time to pass through the stomach, so it is not surprising that no change due to intestinal activity could be detected. While the second time point might be more reasonable, even saccharides-based permeability measurements take between 0-2 hours to reflect small intestinal permeability [85].

Next, we looked at between-condition differences in the blood microbiome. We found the phyla *Firmicutes* and *Proteobacteria* to have a lower and higher abundance, respectively, in PSC / PBC compared to HC. An increase in abundance of *Proteobacteria* in PSC patients was recently reported in the bile fluid [97], with *Firmicus* remaining stable between the conditions. A reduction in *Firmicus* for PSC patients was reported by Amar et al. [112] for fecal samples, while *Proteobacteria* were not present in any significant amount. We identified a significant increase in alpha diversity between PSC / PBC and HC. Differences in alpha diversity for PSC and PBC patients have previously been reported for mucosa-associated bacteria [93], stool samples [95, 100] and bile fluid [97]. These studies showed a decrease in alpha diversity. However, working under the 'leaky gut' hypothesis, an increase in alpha diversity was expected. This finding suggests a similar result as in Dhillon et al. [87]. We were also able to show a significant difference in overall community structure between disease and healthy patients. Many differentially abundant genera and species have been reported for PSC / PBC [100, 101]. This study found the family *Microbacteriaceae* to be differentially abundant in PSC and PBC compared to the controls.

Microbacteriaceae are present in the human gut [129], and some species are known to be clinically relevant. However, the most abundant genus of the family, *Agreia*, is reportedly mostly found in plant material [136]. Most of the taxonomic family counts across all three groups were not defined (not assigned or generic assignment *Microbacteriaceae* bacterium), and none of the genera were differentially abundant. Comparing PBC and HC, we found the family *Staphylococcaceae* and the cirrhosis-associated pathobiont *Staphylococcus* [80] to be differentially under-represented in PBC. This genus has recently been reported to be over-represented in PSC patients' bile fluid [97]. Few studies have confirmed *Staphylococcus* to be over-represented in cirrhotic patients' serum [114]. On the other hand, *Staphylococcus* is of no significance for PSC patients in stool samples, while yet another study found *Staphylococcus* to be under-represented in stool samples of PSC patients. It is important to note that Staphylococcus has also been identified in multiple studies' negative controls [108].

This analysis would greatly benefit from much larger sample sizes, stricter procedures during drawing blood from patients and randomized patient selection. Nevertheless, based on the data analyzed, it is likely that PSC and PBC patients show differences in their blood microbiome. The observed increase in diversity between PSC / PBC patients and controls was expected. An increase in diversity indicates more active microbial activity in the blood or the increased translocation of microbes or microbial components across the intestinal membrane.

A. Bias Invariant RNA-Seq Metadata Annotation

A.1. Figures

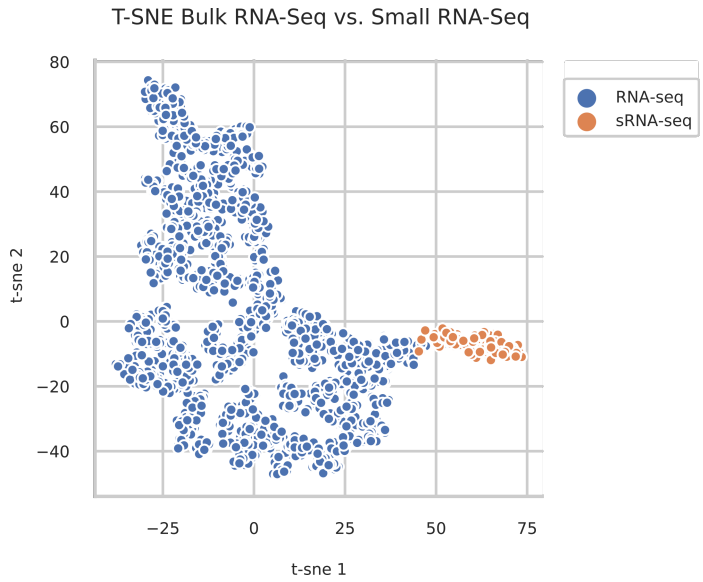

Figure A.1.: T-SNE on fraction of total gene count per gene type. The fraction of the total log TPM normalized counts per gene type was calculated for all types that can be associated with mRNA or small RNA. T-SNE was applied on the resulting vectors of fraction per gene type. Samples with their maximum fraction in a gene type belonging to a small RNA category were labeled orange, else blue. The scatter plot shows samples labeled as small RNA-seq all cluster together suggesting a valid approach.

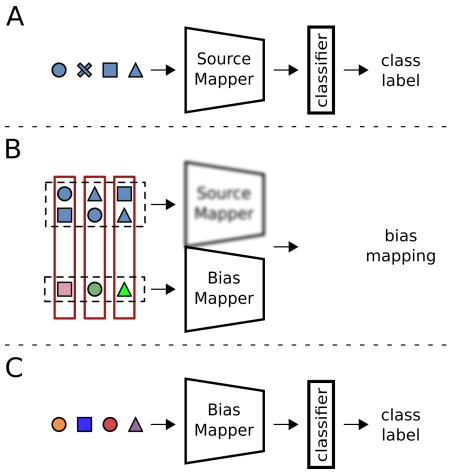

Figure A.2.: Overview of DA model short. Samples are indicated according to their classes (circles, squares, triangles) and their bias (blue: source domain, other colors: bias domain, target domain). The model is ready for prediction after two training steps: A) A source mapper is trained on single bias data together with a classification layer. B) A bias mapper is created as a duplicate of the source mapper, the weights of the source mapper are fixed. Triplets are passed through the source mapper and bias mapper configuration to learn a bias mapping. C) The bias mapper, equipped with a classification layer, can be used to predict data from previously unseen datasets.

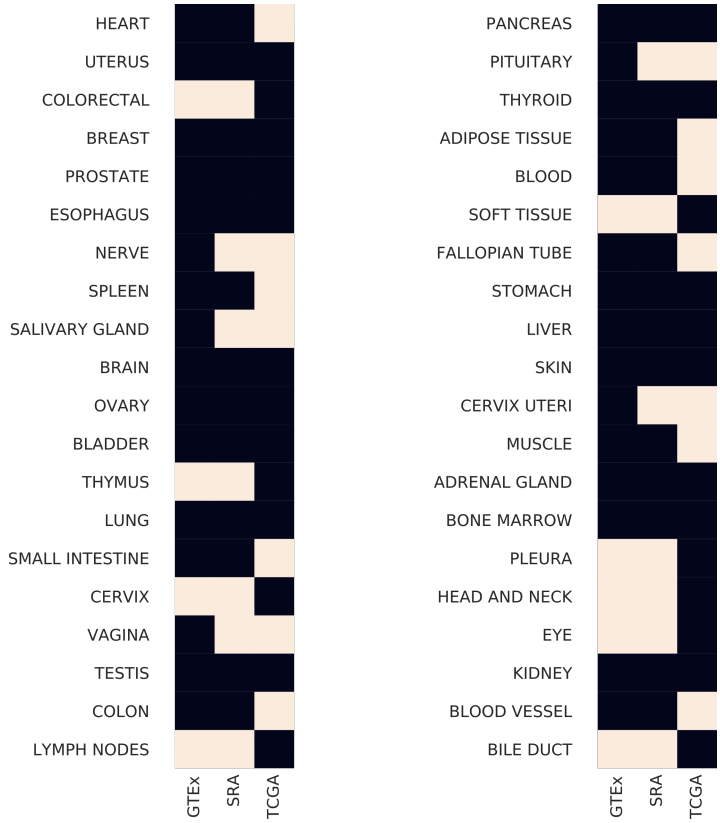

Figure A.3.: Tissue label overlap between GTEx, TCGA and SRA. GTEx v6 provides samples for 31 tissues and TCGA for 26. MetaSRA provided labels for 26 of the 31 GTEx tissues. This figure depicts the 40 tissues which form the union between the three data sources, a black square indicating that a tissue is present in the respective dataset. 17 Tissues are shared between GTEx, TCGA and SRA, 16 of which were used for tissue prediction.

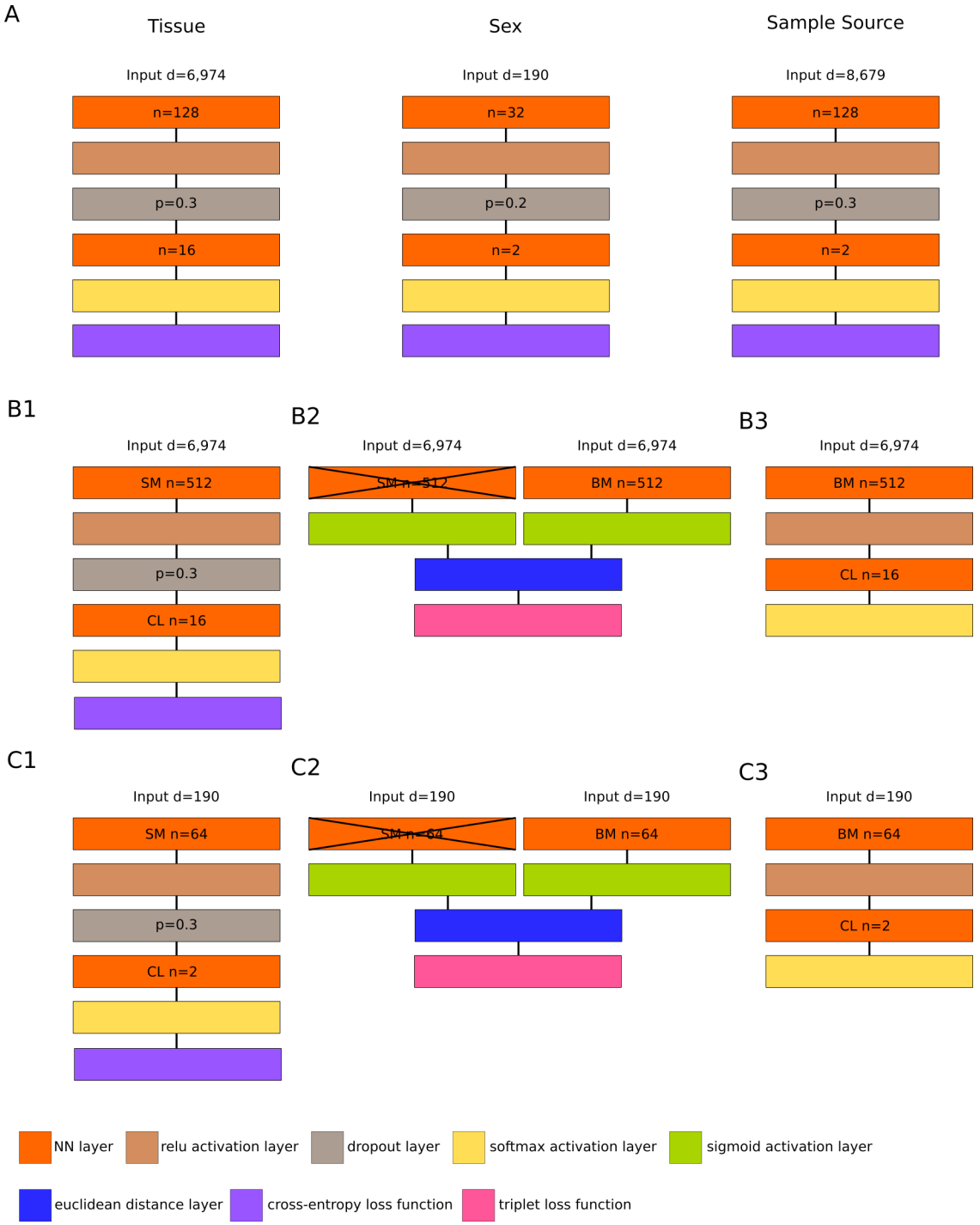

Figure A.4.: Architectures of all applied models. Graphical representation of architectures for ANN based models. A) MLP models for tissue, sex and sample source, B) are the (1) SM-CL MLP, (2) SM-BM Siamese Network and (3) BM-CL prediction model for tissue and C) sex. Each rectangle represents a layer in the neural network and is colored according to the type of layer that has been used. d = input dimension, n = number of nodes, p = drop out probability, SM = source mapper, BM = bias mapper, CL = classification layer. B2 and C2 show the SM to have frozen weights.

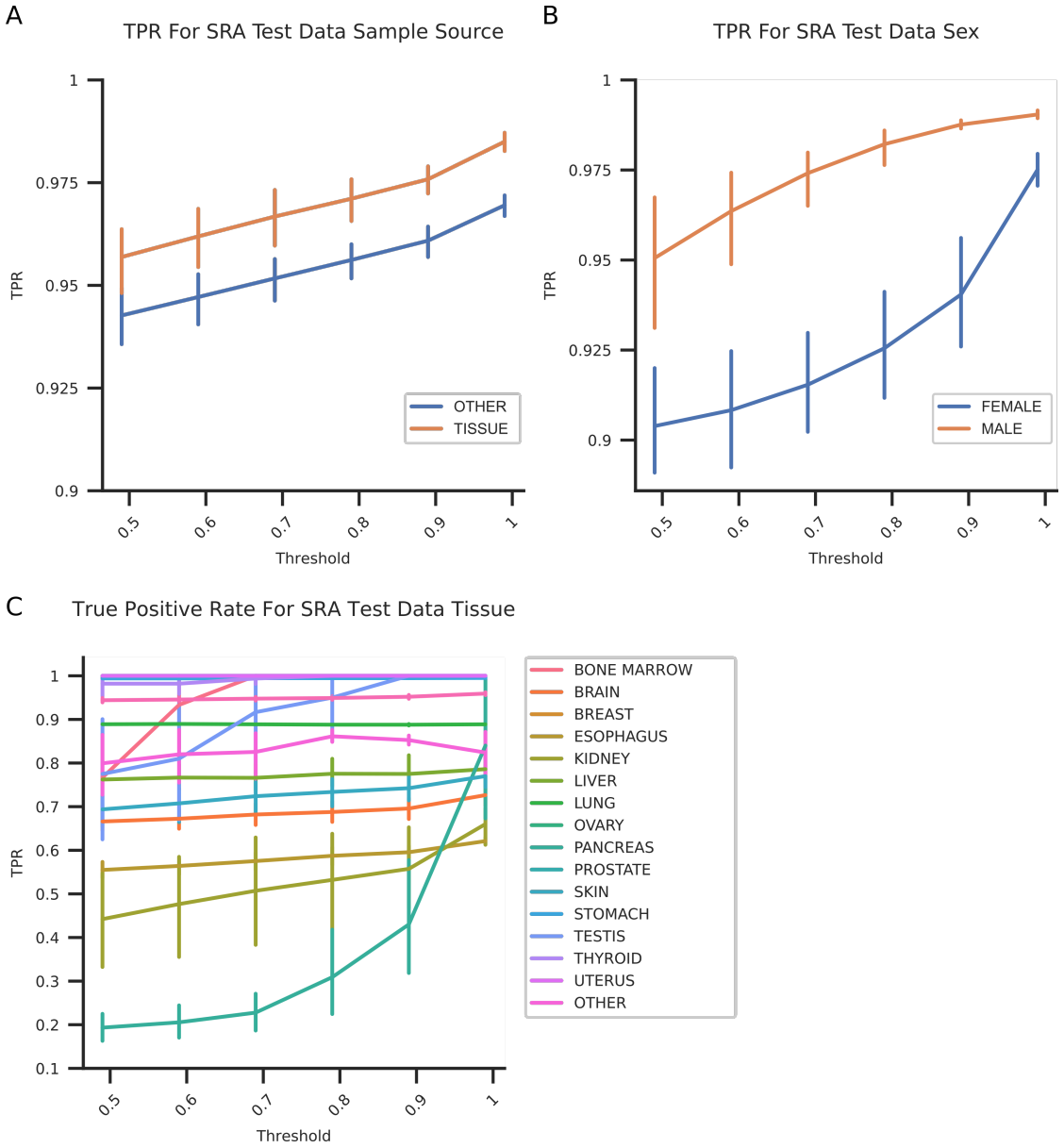

Figure A.5.: True positive rate for test data predicted with annotation models. A) Sample source, B) sex and C) tissue classification.

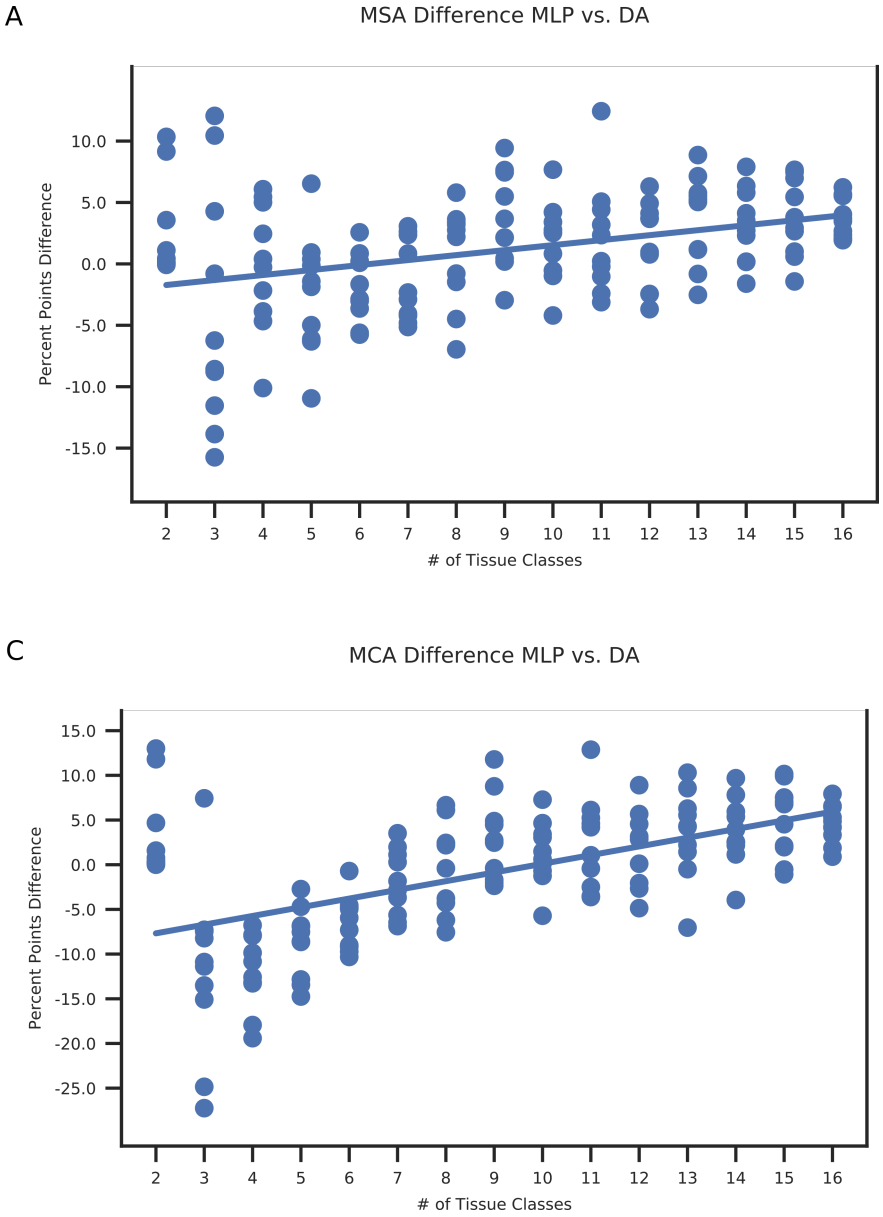

Figure A.6.: Relationship between number of classes and DA performance in DA G+S-T. The 16 tissues were sorted by sample size in GTEx, at each step one tissue was added to the classification problem, starting with the largest two. MLP and DA were trained as described above for 10 seeds each and tested on TCGA data. The mean sample accuracy for each seed (top panel) or mean class accuracy (bottom panel) are shown. Each dot shows the difference in accuracy (DA-MLP) at each step for each seed. Seaborn's regplot was used to a regression line. While, on average, MLP performs better for lower number of classes, the performance gain by the DA model with respect to MLP increases with the number of classes.

A.2. Tables

Table A.1.: Tissue annotation for brain tissue in SRA metadata

Brain
brain region BA22 (temporal cortex)
brain region BA41 (temporal cortex)
brain region BA09 (frontal cortex)
Peripheral brain tissue
Tumor brain tissue
brain
Normal human brain
brain (middle frontal gyrus)
brain (dorsal prefrontal cortex)
frozen postmortem brain from NICHD
Human brain
brain tumor tissue
Brain, Cerebellum
Brain, whole
Brain, fetal
brain (BA9 prefrontal cortex)
Dorsal Forebrain Equivalent
Human brain cortex (BA9)
Post morten brain

Table A.2.: Mapping from GTEx tissue names to MetaSRA tissue names.

GTEx	MetaSRA
ovary	female gonad
skin	anatomical skin
thyroid	thyroid gland
prostate	prostate gland
bladder	urinary bladder
cervix uteri	uterine cervix

Table A.3.: Summary of the datasets used for each phenotype after pre-processing.

Dataset	# Samples	# Classes	# Input genes	Gini cut off	
				Low	High
Tissue					
GTEx	5,480	16	6,974	0.5	1
TCGA	8,624				
SRA train	1,721				
SRA test	1,531				
SRA train annotation	3,370	17			
SRA test annotation	2,813				
Sex					
GTEx	9,662	2	190	0.4	0.7
TCGA	11,284				
SRA train	2,317				
SRA test	923				
Sample Source					
GTEx	9,662	1	8,679	0.3	0.8
TCGA	11,284				
SRA train	12,725				
SRA train	3,144	2			
SRA val	1,124				
SRA train annotation	16,463				
SRA test annotation	3,707				

Table A.4.: Number of samples per class for phenotype classification experiments.

	GTEx	TCGA	SRA train	SRA test
		Tissue		
Adrenal gland	159	266	14	5
Bone marrow	102	126	77	90
Brain	1,409	707	508	770
Breast	218	1246	123	30
Esophagus	790	198	35	5
Kidney	36	1030	94	88
Liver	136	424	111	134
Lung	374	1156	228	72
Ovary	108	430	23	12
Pancreas	197	183	17	5
Prostate	119	558	123	49
Skin	974	473	238	198
Stomach	204	453	25	11
Testis	203	156	14	18
Thyroid	361	572	51	32
Uterus	90	646	40	12
		Sex		
Male	6,036	5,395	1,246	575
Female	3,326	5,889	1,071	348
		Sample source		
Cell line	9,662	11,284	7,108	1,950
Biopsy	-	-	5,617	1,194

Table A.5.: Hyperparameters considered during model tuning and their initial range.

Hyperparameter	Range	Sampling mode
# Layers	[0,3]	linear
# Nodes per layer	[32,512]	linear
Batch size	[16,32,64]	step
Learning rate	[1e-4, 1e-2]	log
Optimizer	[Adam, SGD]	binary
Drop out	[0.1,0.2,0.3]	step
Gini cut off	manually	manually

Table A.6.: Summary of the hyperparameters used for each model.

Model	# Nodes	Dropout rate	Learning rate	Margin
MLP Tissue	128	0.3	0.0002	-
MLP Sex	32	0.2	0.0024	-
MLP Sample Source	128	0.3	0.0002	-
DA SM-CL Tissue	512 / 16	0.3	0.0001	-
DA SM-BM Tissue	512 / 512	-	0.0005	5
DA SM-CL Sex	64 / 2	0.3	0.0001	-
DA SM-BM Sex	64 / 64	-	0.0005	3

For every model 1 hidden layer was used, batch size was 64, trained epochs were 10 and the optimizer used Adam.

Table A.7.: Sample and class accuracy given are the mean over n=10 seeds

	msa	mca	msa std.	mca std.
Tissue				
SRA				
LIN G-S	0.893	0.765	NA	NA
LIN S_{small}-S	0.893	0.795	NA	NA
LIN G+S_{small}-S	0.908	0.785	NA	NA
MLP G-S	0.872	0.77	0.007	0.018
MLP S_{small}-S	0.894	0.746	0.005	0.017
MLP G+S_{small}-S	0.915	0.817	0.008	0.02
DA G+S_{small}-S	0.922	0.821	0.003	0.009
MLP S_{small}+S_{new}-S	0.911	0.798	0.007	0.022
DA G+S_{small}+S_{new}-S	0.933	0.854	0.002	0.009
TCGA				
LIN G-T	0.718	0.638	NA	NA
LIN S_{large}-T	0.784	0.724	NA	NA
LIN G+S_{large}-T	0.725	0.651	NA	NA
MLP G-T	0.684	0.605	0.015	0.017
MLP S_{large}-T	0.832	0.755	0.02	0.03
MLP G+S_{large}-T	0.842	0.773	0.015	0.017
DA G+S_{large}-T	0.875	0.813	0.004	0.006
LIN S_{small}-T	0.768	0.708	NA	NA
LIN G+S_{small}-T	0.729	0.658	NA	NA
MLP S_{small}-T	0.748	0.688	0.016	0.027
MLP G+S_{small}-T	0.764	0.716	0.033	0.028
DA G+S_{small}-T	0.81	0.763	0.014	0.024
MLP S_{large}+S_{new}-T	0.83	0.758	0.017	0.02
Sex				
SRA				
LIN G-S	0.883	0.876	NA	NA
LIN S_{small}-S	0.878	0.873	NA	NA
LIN G+S_{small}-S	0.883	0.876	NA	NA
MLP G-S	0.879	0.871	0.002	0.04
MLP S_{small}-S	0.93	0.936	0.008	0.009
MLP G+S_{small}-S	0.939	0.945	0.003	0.003
DA G+S_{small}-S	0.929	0.93	0.025	0.036
MLP S_{small}+S_{new}-S	0.945	0.948	0.003	0.004
TCGA				
LIN G-T	0.989	0.989	NA	NA
LIN S_{large}-T	0.988	0.987	NA	NA
LIN G+S_{large}-T	0.989	0.989	NA	NA
MLP G-T	0.869	0.863	0.011	0.011
MLP S_{large}-T	0.936	0.934	0.01	0.01
MLP G+S_{large}-T	0.947	0.945	0.011	0.011
DA G+S_{large}-T	0.919	0.916	0.004	0.004
MLP S_{large}+S_{new}-T	0.975	0.975	0.004	0.004
Sample source				
LIN S_{large}-G	0.951	0.951	NA	NA
LIN S_{large}-T	0.882	0.882	NA	NA
LIN S_{small}-S	0.89	0.884	NA	NA
MLP S_{large}-G	0.943	0.943	0.001	0.001
MLP S_{large}-T	0.971	0.971	0.028	0.028
MLP S_{small}-S	0.95	0.941	0.003	0.005

B. Analysis of Microbial 16S rRNA-Seq Data of the Blood

B.1. Figures

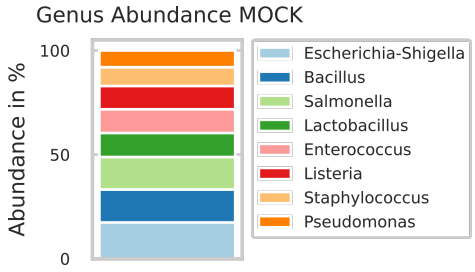

Figure B.1.: Genus Abundance of MOCK Control. A ZymoBIOMICS Microbial Community DNA Standard was sequenced as a positive control. This control has a known microbial composition. The expected composition was recovered at the genus level, validating the sequencing, ASV inference and taxonomic classification.

B.2. Tables

Table B.1.: Clinical patient characteristics. Median, range, count, or percentage are reported. PSC and HC were selected during the same time window by the same researchers. PBC samples were obtained one year later, involving a different researcher. HC samples were obtained from young employees of the UKE.

	PSC	PBC	HC
Patients, n	10	10	8
Female, n (Age, years (range)	34 (23-66)	50 (40-69)	33 (23-37)
BMI, (range)	23 (19.2-27.2)	25 (21-35.3)	22.6 (18.7-26.3)
Collected in	2011	2012	2011
Collected by	A+B	B+C	A+B

Table B.2.: List of all phyla found in serum samples ordered by total count.

Phylum	In Samples	Total Count
Proteobacteria	89	994629
Actinobacteriota	88	425930
Firmicutes	86	328938
Bacteroidota	88	265214
Patescibacteria	54	27983
Acidobacteriota	47	19042
Myxococcota	44	15125
Bdellovibrionota	36	14441
Campilobacterota	37	13311
Chloroflexi	41	10233
Cyanobacteria	29	8245
Verrucomicrobiota	52	8215
Fibrobacterota	18	5891
Desulfobacterota	21	4922
Gemmatimonadota	17	4721
Armatimonadota	19	3472
Fusobacteriota	12	2976
Deinococcota	8	2719
Spirochaetota	11	2666
Elusimicrobiota	12	2330
Nitrospirota	6	1806
Planctomycetota	23	1503
Synergistota	4	1279
Dependentiae	7	1205
Methylomirabilota	3	981
SAR324_clade(Marine_group_B)	4	697
WPS-2	3	598
MBNT15	2	371
Abditibacteriota	3	333
Cloacimonadota	1	247
Latescibacterota	1	148
Hydrogenedentes	2	124

Table B.3.: List of genera determined to be contamination according to literature and subsequently removed from the data.

Genus	In Samples	Total Count
Diaphorobacter	68	241047
Cutibacterium	80	142361
Pseudomonas	72	76359
Psychrobacter	57	59610
Flavobacterium	62	56349
Undibacterium	44	55695
Lamprocystis	59	55378
Cloacibacterium	53	52780
Enhydrobacter	53	45930
Corynebacterium	62	40821
Micrococcus	41	35417
Rhodoferax	50	35097
Streptococcus	29	28544
Polaromonas	48	25489
Acinetobacter	56	23869
Legionella	34	20470
Bacillus	37	19149
Pedobacter	43	16786
Clostridium_sensu_stricto_1	34	14004
Microbacterium	30	11792
Clostridium_sensu_stricto_13	35	10734
Massilia	9	10353
Escherichia-Shigella	26	9290
Mucilaginibacter	38	9258
Duganella	22	8695
Rhodococcus	32	8587
Chryseobacterium	38	7104
Kocuria	13	6450
Janthinobacterium	22	6429
67-14	20	6039
Acidovorax	28	6013
[Agitococcus]_lubricus_group	30	5779
Janibacter	15	5343
Faecalibacterium	21	5162
Variovorax	15	4409
CL500-29_marine_group	12	3701
Marinospirillum	12	1604

Bibliography

[1] S.E. Ellis, L. Collado-Torres, A. Jaffe, and J.T. Leek. Improving the value of public rna-seq expression data by phenotype prediction. *Nucleic Acids Res.*, 46(9):e54–e54, 2018.

[2] K.R. Kukurba and S.B. Montgomery. Rna sequencing and analysis. *Cold Spring Harb Protoc*, 11:pdb–top084970, 2015.

[3] V. Costa, M. Aprile, R. Esposito, and A. Ciccodicola. Rna-seq and human complex diseases: recent accomplishments and future perspectives. *Eur. J. Hum. Genet.*, 21(2):134–142, 2013.

[4] M. Melé, P.G. Ferreira, F. Reverter, D.S. DeLuca, J. Monlong, M. Sammeth, T.R. Young, and J.M. Goldmann. The human transcriptome across tissues and individuals. *Science*, 348(6235):660–665, 2015.

[5] GTEx Consortium. The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020.

[6] C. Klijn, S. Durinck, E.W. Stawiski, P.M. Haverty, Z. Jiang, H. Liu, J. Degenhardt, O. Mayba, F. Gnad, J. Liu, and G. Pau. A comprehensive transcriptional portrait of human cancer cell lines. *Nat Biotechnol*, 33(3):306–312, 2015.

[7] J.H. Tabibian, C.E. Trussoni, S.P. O'hara, P.L. Splinter, J.K. Heimbach, and N.F. LaRusso. Characterization of cultured cholangiocytes isolated from livers of patients with primary sclerosing cholangitis. *Lab Invest*, 94(10):1126–1133, 2014.

[8] A. Henderson-Smith, J.J. Corneveaux, M. De Both, L. Cuyugan, W.S. Liang, M. Huentelman, C. Adler, E. Driver-Dunckley, T.G. Beach, and T.L. Dunckley. Next-generation profiling to identify the molecular etiology of parkinson dementia. *Neurol Genet*, 2(3), 2016.

[9] M.G. Best, N. Sol, I. Kooi, J. Tannous, B.A. Westerman, F. Rustenburg, P. Schellen, H. Verschueren, E. Post, J. Koster, and B. Ylstra. Rna-seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. *Cancer Cell*, 28(5):66–676, 2015.

[10] Z. Wang, M. Gerstein, and M. Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63, 2009.

[11] Z.D. Stephens, S.Y. Lee, F. Faghri, R.H. Campbell, C. Zhai, M.J. Efron, R. Iyer, M.C. Schatz, S. Sinha, and G.E. Robinson. Big data: astronomical or genomical? *PLoS Biol.*, 13(7):p.e1002195, 2015.

[12] K. Pearson. Report on certain enteric fever inoculation statistics. *Br Med J*, 2(2288):1243–1246, 1904.

[13] J. Gurevitch, J. Koricheva, S. Nakagawa, and G. Stewart. Meta-analysis and the science of research synthesis. *Nature*, 555(7695):175–182, 2018.

[14] E. Walker, A.V. Hernandez, and M.W. Kattan. Meta-analysis: Its strengths and limitations. *Cleve Clin J Med*, 75(6):431, 2008.

[15] G.C. Tseng, D. Ghosh, and E. Feingold. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res*, 40(9):3785–3799, 2012.

[16] A. Rau, G. Marot, and F. Jaffrézic. Differential meta-analysis of rna-seq data from multiple studies. *BMC Bioinformatics*, 15(1):1–10, 2014.

[17] P.H. Sudmant, M.S. Alexis, and C.B. Burge. Meta-analysis of rna-seq expression data across species, tissues and studies. *Genome Biol*, 16(1):1–11, 2015.

[18] A. Alimadadi, P.B. Munroe, B. Joe, and X. Cheng. Meta-analysis of dilated cardiomyopathy using cardiac rna-seq transcriptomic datasets. *Genes*, 11(1):60, 2020.

[19] H. Patel, R.J. Dobson, and S.J. Newhouse. A meta-analysis of alzheimer's disease brain transcriptomic data. *J Alzheimers Dis*, 68(4):1635–1656, 2019.

[20] K. Dolinski and O.G. Troyanskaya. Implications of big data for cell biology. *Mol Biol Cell*, 26(14):2575–2578, 2015.

[21] J.N. Taroni, P.C. Grayson, Q. Hu, S. Eddy, M. Kretzler, P.A. Merkel, and C.S. Greene. Multiplier: a transfer learning framework for transcriptomics reveals systemic features of rare disease. *Cell Syst*, 8(5):380–394, 2019.

[22] J. Tan, G. Doing, K.A. Lewis, C.E. Price, K.M. Chen, K.C. Cady, B. Perchuk, M.T. Laub, D.A. Hogan, and C.S. Greene. Unsupervised extraction of stable expression signatures from public compendia with an ensemble of neural networks. *Cell Syst*, 5(1):63–71, 2017.

[23] K. Hatje, R.U. Rahman, R.O. Vidal, D. Simm, B. Hammesfahr, V. Bansal, A. Rajput, M.E. Mickael, T. Sun, S. Bonn, and M. Kollmar. The landscape of human mutually exclusive splicing. *Mol. Syst. Biol.*, 13(12):959, 2017.

[24] P. Yu, J. Li, S.P. Deng, F. Zhang, P.N. Grozdanov, E.W. Chin, S.D. Martin, L. Vergnes, M.S. Islam, D. Sun, and J.M. LaSalle. Integrated analysis of a compendium of rna-seq datasets for splicing factors. *Sci Data*, 7(1):1–16, 2020.

[25] C. Thanos. Research data reusability: Conceptual foundations, barriers and enabling technologies. *Publications*, 5(1):2, 2017.

[26] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.W. Boiten, L.B. da Silva Santos, P.E. Bourne, and J. Bouwman. The fair guiding principles for scientific data management and stewardship. *Sci Data*, 3(1):1–9, 2016.

[27] R. Leinonen, H. Sugawara, M. Shumway, and International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Res.*, 39(suppl1):D19–D21, 2010.

[28] M.N. Bernstein, A. Doan, and C.N. Dewey. Metasra: normalized human sample-specific metadata for the sequence read archive. *Bioinformatics*, 33(18):2914–2923, 2017.

[29] L. Collado-Torres, A. Nellore, K. Kammers, S.E. Ellis, M.A. Taub, K.D. Hansen, A.E. Jaffe, B. Langmead, and J.T. Leek. Reproducible rna-seq analysis using recount2. *Nat Biotechnol*, 35(4):319–321, 2017.

[30] A. Nellore, L. Collado-Torres, A.E. Jaffe, J. Alquicira-Hernández, C. Wilks, J. Pritt, J. Morton, J.T. Leek, and B. Langmead. Rail-rna: scalable analysis of rna-seq splicing and coverage. *Bioinformatics*, 33(24):4033–4040, 2017.

[31] L. Collado-Torres, A. Nellore, A.C. Frazee, C. Wilks, M.I. Love, B. Langmead, R.A. Irizarry, J.T. Leek, and A.E. Jaffe. Flexible expressed region analysis for rna-seq with derfinder. *Nucleic Acids Res*, 45(2):e9–e9, 2017.

[32] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, and B. Foster. The genotype-tissue expression (gtex) project. *Nat. Genet.*, 45(6):580–585, 2013.

[33] M. Oliva, M. Muñoz-Aguirre, S. Kim-Hellmuth, V. Wucher, A.D. Gewirtz, D.J. Cotter, P. Parsana, S. Kasela, B. Balliu, A. Viñuela, and S.E. Castel. The impact of sex on gene expression across human tissues. *Science*, 369(6509), 2020.

[34] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *J Mach Learn Res*, 9(11), 2008.

[35] M. Sultan, V. Amstislavskiy, T. Risch, M. Schuette, S. Dökel, M. Ralser, D. Balzereit, H. Lehrach, and M.L. Yaspo. Influence of rna extraction methods and library selection schemes on rna-seq data. *BMC Genomics*, 15(1):1–13, 2014.

[36] A.N. Scholes and J.A. Lewis. Comparison of rna isolation methods on rna-seq: implications for differential expression and meta-analyses. *BMC Genomics*, 21(1):1–9, 2020.

[37] P.A.C. 't Hoen, M.R. Friedländer, J. Almlöf, M. Sammeth, I. Pulyakhina, S.Y. Anvar, J.F. Laros, H.P. Buermans, O. Karlberg, M. Brännvall, and J.T den Dunnen. Reproducibility of high-throughput mrna and small rna sequencing across laboratories. *Nat. Biotechnol.*, 31(11):1015–1022, 2013.

[38] S. Arora, S.S. Pattwell, E.C. Holland, and H. Bolouri. Variability in estimated gene expression among commonly used rna-seq pipelines. *Sci. Rep.*, 10(1):1–9, 2020.

[39] T. Mitchell. *Machine Learning*. McGraw-Hill Education Ltd, 1997.

[40] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[41] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[42] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev*, 65(6):386, 1958.

[43] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall. Activation functions: Comparison of trends in practice and research for deep learning. *arXiv*, page 1811.03378, 2018.

[44] V. Nair and G.E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML 2010*, 2010.

[45] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a" siamese" time delay neural network. *Adv Neural Inf Process Syst*, pages 737–737, 1994.

[46] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[47] S. Chopra, S. Balakrishnan, and R. Gopalan. Dlid: Deep learning for domain adaptation by interpolating between domains. In *ICML workshop on challenges in representation learning*, volume 2, 2013.

[48] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *J Mach Learn Res*, 17(1):2096–2030, 2016.

[49] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[50] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farleyn, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[51] S. Motiian, Q. Jones, S.M. Iranmanesh, and G. Doretto. Few-shot adversarial domain adaptation. *arXiv*, page 1711.02536, 2017.

[52] R. Hrdlickova, M. Toloue, and B. Tian. Rna-seq methods for transcriptome analysis. *Wiley Interdiscip Rev: RNA*, 8(1):e1364, 2017.

[53] S. Goodwin, J.D. McPherson, and W.R. McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, 17(6):33, 2016.

[54] S. Li, P.P. Łabaj, P. Zumbo, P. Sykacek, W. Shi, L. Shi, J. Phan, P.Y. Wu, M. Wang, C. Wang, and D. Thierry-Mieg. Detecting and correcting systematic variation in large-scale rna sequencing data. *Nat. Biotechnol.*, 32(9):888–895, 2014.

[55] M. Marouf, P. Machart, V. Bansal, C. Kilian, D.S. Magruder, C.F. Krebs, and S. Bonn. Realistic in silico generation and augmentation of single-cell rna-seq data using generative adversarial networks. *Nat. Commun.*, 11(1):1–12, 2020.

[56] K. Menden, M. Marouf, S. Oller, A. Dalmia, D.S. Magruder, K. Kloiber, P. Heutink, and S. Bonn. Deep learning–based cell composition analysis from tissue expression profiles. *Sci. Adv.*, 6(30):eaba2619, 2020.

[57] P. Mamoshina, A. Vieira, E. Putin, and A. Zhavoronkov. Applications of deep learning in biomedicine. *Mol. Pharm.*, 13(5):1445–1454, 2016.

[58] M. Wainberg, D. Merico, A. Delong, and B.J. Frey. Deep learning in biomedicine. *Nat. Biotechnol.*, 36(9):829–838, 2018.

[59] G. Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv*, page 1702.05374, 2017.

[60] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars. A deeper look at dataset bias. *arXiv*, page 1505.01257, 2017.

[61] L. Ceriani and P. Verme. The origins of the gini index: extracts from variabilità e mutabilità (1912) by corrado gini. *J. Econ. Inequal.*, 10(2):421–443, 2012.

[62] J.D. Zhang, K. Hatje, G. Sturm, C. Broger, M. Ebeling, M. Burtin, F. Terzi, S.I. Pomposiello, and L. Badi. Detect tissue heterogeneity in gene expression data with bioqc. *BMC Genomics*, 18(1):1–9, 2017.

[63] Google Research. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[64] M.L. Yeung, Y. Yao, L. Jia, J.F. Chan, K.H. Chan, K.F. Cheung, H. Chen, V.K. Poon, A.K. Tsang, K.K. To, and M.K. Yiu. Mers coronavirus induces apoptosis in kidney and lung by upregulating smad7 and fgf2. *Nat. Microbiol.*, 1(3):1–8, 2016.

[65] Y. Kravtsova-Ivantsiv, I. Shomer, V. Cohen-Kaplan, B. Snijder, G. Superti-Furga, H. Gonen, T. Sommer, T. Ziv, A. Admon, I. Naroditsky, and M. Jbara. Kpc1-mediated ubiquitination and proteasomal processing of nf-kb1 p105 to p50 restricts tumor growth. *Cell*, 161(2):333–347, 2015.

[66] A.M. Smith, J.R. Walsh, J. Long, C.B. Davis, P. Henstock, M.R. Hodge, M. Maciejewski, X.J. Mu, S. Ra, S. Zhao, and D. Ziemek. Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. *BMC Bioinformatics*, 21(1):1–18, 2020.

[67] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[68] G. Berg, D. Rybakova, D. Fischer, T. Cernava, M.C.C. Vergès, T. Charles, X. Chen, L. Cocolin, G.H. Eversole, K. amd Corral, and M. Kazou. Microbiome definition re-visited: old concepts and new challenges. *Microbiome*, 8(1):1–22, 2020.

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

[70] F. Bäckhed, R.E. Ley, J.L. Sonnenburg, D.A. Peterson, and J.I. Gordon. Host-bacterial mutualism in the human intestine. *Science*, 307(5717):1915–1920, 2005.

[71] C.R. Woese and G.E. Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. In *Proceedings of the National Academy of Sciences*, volume 74, pages 5088–5090, 1977.

[72] J.S. Johnson, D.J. Spakowicz, B.Y. Hong, L.M. Petersen, P. Demkowicz, L. Chen, S.R. Leopold, B.M. Hanson, H.O. Agresta, M. Gerstein, and E. Sodergren. Evaluation of 16s rrna gene sequencing for species and strain-level microbiome analysis. *Nat Commun*, 10(1):1–11, 2019.

[73] D.M. Ward, R. Weller, and M.M. Bateson. 16s rrna sequences reveal numerous uncultured microorganisms in a natural community. *Nature*, 345(6270):63–65, 1990.

[74] B.J. Callahan, P.J. McMurdie, and S.P. Holmes. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J*, 11(12):2639–2643, 2017.

[75] M.J. Rosen, M. Davison, D. Bhaya, and D.S. Fisher. Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. *Science*, 348(6238):1019–1023, 2015.

[76] B.J. Callahan, P.J. McMurdie, M.J. Rosen, A.W. Han, A.J.A. Johnson, and S.P. Holmes. Dada2: high-resolution sample inference from illumina amplicon data. *Nat Methods*, 13(7):581–583, 2016.

[77] R.C. Edgar. Unoise2: improved error-correction for illumina 16s and its amplicon sequencing. *bioRxiv*, page 081257, 2016.

[78] A. Amir, D. McDonald, J.A. Navas-Molina, E. Kopylova, J.T. Morton, Z.Z. Xu, E.P. Kightley, L.R. Thompson, E.R. Hyde, A. Gonzalez, and R. Knight. Deblur rapidly resolves single-nucleotide community sequence patterns. *MSystems*, 2(2), 2017.

[79] B. Wang, M. Yao, L. Lv, Z. Ling, and L. Li. The human microbiota in health and disease. *Engineering*, 3(1):71–82, 2017.

[80] V. Giannelli, V. Di Gregorio, V. Iebba, M. Giusto, S. Schippa, M. Merli, and U. Thalheimer. Microbiota and the gut-liver axis: bacterial translocation, inflammation and infection in cirrhosis. *World J Gastroenterol*, 20(45):16795, 2014.

[81] C. Cesaro, A. Tiso, A. Del Prete, R. Cariello, C. Tuccillo, G. Cotticelli, C. del Vecchio Blanco, and C. Loguercio. Gut microbiota and probiotics in chronic liver diseases. *Dig Liver Dis*, 43(6):431–438, 2011.

[82] H. Takaishi, T. Matsuki, A. Nakazawa, T. Takada, S. Kado, T. Asahara, N. Kamada, A. Sakuraba, T. Yajima, H. Higuchi, and N. Inoue. Imbalance in intestinal microflora constitution could be involved in the pathogenesis of inflammatory bowel disease. *Int J Med Microbiol*, 298(5-6):463–472, 2008.

[83] R.S. Lin, F.Y. Lee, S.D. Lee, Y.T. Tsai, H.C. Lin, L. Rei-Hwa, H. Wan-Ching, H. Cheng-Chun, W. Sun-Sang, and L. Kwang-Juei. Endotoxemia in patients with chronic liver diseases: relationship to severity of liver diseases, presence of esophaegeal varices, and hyperdynamic circulation. *J Hepatol*, 22(2):165–172, 1995.

[84] J. Such, R. Francés, C. Muñoz, P. Zapater, J.A. Casellas, A. Cifuentes, Rodríǵguez-Valero F., S. Pascual, J. Sola-Vera, and F. Carnicer. Detection and identification of bacterial dna in patients with cirrhosis and culture-negative, nonneutrocytic ascites. *Hepatology*, 36(1):135–141, 2002.

[85] M. Camilleri. Camilleri, m., 2019. leaky gut: mechanisms, measurement and clinical implications in humans. *Gut*, 68(8):1516–1526, 2019.

[86] D. Hollander, C.M. Vadheim, E. Brettholz, G.M. Petersen, T. Delahunty, and J.I. Rotter. Increased intestinal permeability in patients with crohn's disease and their relatives: a possible etiologic factor. *Ann Intern Med*, 105(6):883–885, 1986.

[87] A.K. Dhillon, M. Kummen, M. Trøseid, S. Åkra, E. Liaskou, B. Moum, M. Vesterhus, I. Karlsen, T.H. amd Seljeflot, and J.R. Hov. Circulating markers of gut barrier function associated with disease severity in primary sclerosing cholangitis. *Liver Int*, 39(2):371–381, 2019.

[88] A. Mencin, J. Kluwe, and R.F. Schwabe. Toll-like receptors as targets in chronic liver diseases. *Gut*, 58(5):704–720, 2009.

[89] E. Langholz. Current trends in inflammatory bowel disease: the natural history. *Ther Adv Gastrointest Endosc*, 3(2):77–86, 2010.

[90] G.M. Hirschfield, T.H. Karlsen, K.D. Lindor, and D.H. Adams. Primary sclerosing cholangitis. *The Lancet*, 382(9904):1587–1599, 2013.

[91] T.H. Karlsen, T. Folseraas, D. Thorburn, and M. Vesterhus. Primary sclerosing cholangitis–a comprehensive review. *J Hepatol*, 67(6):1298–1323, 2017.

[92] J.E.R. Hov and T.H. Karlsen. The microbiome in primary sclerosing cholangitis: current evidence and potential concepts. *Semin Liver Dis*, 37(4):314–331, 2017.

[93] N.G. Rossen, S. Fuentes, K. Boonstra, G.R. D'Haens, H.G. Heilig, E.G. Zoetendal, W.M. de Vos, and C.Y. Ponsioen. The mucosa-associated microbiota of psc patients is characterized by low diversity and low abundance of uncultured clostridiales ii. *J Crohns Colitis*, 9(4):342–348, 2015.

[94] M. Kummen, K. Holm, J.A. Anmarkrud, S. Nygård, M. Vesterhus, M.L. Høivik, M. Trøseid, H.U. Marschall, E. Schrumpf, B. Moum, and H. Røsjø. The gut microbial profile in patients with primary sclerosing cholangitis is distinct from patients with ulcerative colitis without biliary disease and healthy controls. *Gut*, 66(4):611–619, 2016.

[95] L. Bajer, M. Kverka, M. Kostovcik, P. Macinga, J. Dvorak, Z. Stehlikova, J. Brezina, P. Wohl, J. Spicak, and P. Drastich. Distinct gut microbiota profiles in patients with primary sclerosing cholangitis and ulcerative colitis. *World J Gastroenterol*, 23(25):4548, 2017.

[96] M.C. Rühlemann, M.E.L. Solovjeva, R. Zenouzi, T. Liwinski, M. Kummen, W. Lieb, J.R. Hov, C. Schramm, A. Franke, and C. Bang. Gut mycobiome of primary sclerosing cholangitis patients is characterised by an increase of trichocladium griseum and candida species. *Gut*, 69(10):1890–1892, 2020.

[97] T. Liwinski, R. Zenouzi, C. John, H. Ehlken, M.C. Rühlemann, C. Bang, S. Groth, W. Lieb, M. Kantowski, N. Andersen, and G. Schachschal. Alterations of the bile microbiome in primary sclerosing cholangitis. *Gut*, 69(4):665–672, 2019.

[98] P.M. Rodrigues, M.J. Perugorria, A. Santos-Laso, L. Bujanda, U. Beuers, and J.M. Banales. Primary biliary cholangitis: A tale of epigenetically-induced secretory failure?. *J Hepatol*, 69(6):1371–1383, 2018.

[99] R. Tang, Y. Wei, Y. Li, W. Chen, H. Chen, Q. Wang, F. Yang, Q. Miao, X. Xiao, and H. Zhang. Gut microbial profile is altered in primary biliary cholangitis and partially restored after udca therapy. *Gut*, 67(3):534–541, 2018.

[100] L. Lv, D. Fang, D. Shi, D. Chen, R. Yan, Y. Zhu, Y. Chen, L. Shao, F. Guo, and W. Wu. Alterations and correlations of the gut microbiome, metabolism and immunity in patients with primary biliary cirrhosis. *Environ Microbiol*, 18(7):2272–2286, 2016.

[101] M. Kummen and J.R. Hov. The gut microbial influence on cholestatic liver disease. *Liver Int*, 39(7):1186–1196, 2019.

[102] S. Nikkari, I.J. McLaughlin, W. Bi, D.E. Dodge, and D.A. Relman. Does blood of healthy subjects contain bacterial ribosomal dna?. *J Clin Microbiol*, 39(5):1956–1959, 2001.

[103] D.J. Castillo, R.F. Rifkin, D.A. Cowan, and M. Potgieter. The healthy human blood microbiome: fact or fiction?. *Front Cell Infect Microbiol*, (9):148, 2019.

[104] B.V.H. Hornung, R.D. Zwittink, Q.R. Ducarmon, and Ed.J. Kuijper. Response to:'circulating microbiome in blood of different circulatory compartments' by schierwagen et al. *Gut*, 69(4):789–790, 2020.

[105] S.J. Salter, M.J. Cox, E.M. Turek, S.T. Calus, W.O. Cookson, M.F. Moffatt, P. Turner, J. Parkhill, N.J. Loman, and A.W. Walker. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol*, 12(1):1–12, 2014.

[106] M. Laurence, C. Hatzis, and D.E. Brash. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PloS One*, 9(5):e97876, 2014.

[107] A. Glassing, S.E. Dowd, S. Galandiuk, B. Davis, and R.J. Chiodini. Inherent bacterial dna contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathog*, 8(1):1–12, 2016.

[108] R. Eisenhofer, J.J. Minich, C. Marotz, A. Cooper, R. Knight, and L.S. Weyrich. Contamination in low microbial biomass microbiome studies: issues and recommendations. *Trends Microbiol*, 27(2):105–117, 2019.

[109] R. Gschwind, T. Fournier, S. Kennedy, V. Tsatsaris, A. Cordier, F. Barbut, M. Butel, and S. Wydau-Dematteis. Evidence for contamination as the origin for bacteria found in human placenta rather than a microbiota. *Plos One*, 15(8):e0237232, 2020.

[110] L.M.O. Loohuis, S. Mangul, A.P. Ori, G. Jospin, D. Koslicki, H.T. Yang, T. Wu, M.P. Boks, C. Lomen-Hoerth, M. Wiedau-Pazos, and R.M. Cantor. Transcriptome analysis in whole blood reveals increased microbial diversity in schizophrenia. *Transl Psychiatry*, 8(1):1–9, 2018.

[111] G. Serena, C. Davies, M. Cetinbas, R.I. Sadreyev, and A. Fasano. Analysis of blood and fecal microbiome profile in patients with celiac disease. *Hum Microb J*, 11:100049, 2019.

[112] J. Amar, C. Lange, G. Payros, C. Garret, C. Chabo, O. Lantieri, M. Courtney, M. Marre, M.A. Charles, B. Balkau, and R. Burcelin. Blood microbiota dysbiosis is associated with

the onset of cardiovascular events in a large general population: the desir study. *PloS One*, 8(1):e54461, 2013.

[113] J. Qiu, H. Zhou, Y. Jing, and C. Dong. Association between blood microbiome and type 2 diabetes mellitus: A nested case-control study. *J Clin Lab Anal*, 33(4):e22842, 2019.

[114] A. Santiago, M. Pozuelo, M. Poca, C. Gely, J.C. Nieto, X. Torras, E. Román, D. Campos, G. Sarrabayrouse, S. Vidal, and E. Alvarado-Tapias. Alteration of the serum microbiome composition in cirrhotic patients with ascites. *Sci Rep*, 6(1):1–9, 2016.

[115] D. Traykova, B. Schneider, M. Chojkier, and M. Buck. Blood microbiome quantity and the hyperdynamic circulation in decompensated cirrhotic patients. *PloS One*, 12(2):e0169310, 2017.

[116] E. Bolyen, J.R. Rideout, M.R. Dillon, N.A. Bokulich, C.C. Abnet, G.A. Al-Ghalith, H. Alexander, E.J. Alm, M. Arumugam, F. Asnicar, and Y. Bai. Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. *Nat Biotechnol*, 37(8):852–857, 2019.

[117] H.L. Sanders. Marine benthic diversity: a comparative study. *Am Nat*, 102(925):243–282, 1968.

[118] S. Weiss, Z.Z. Xu, S. Peddada, A. Amir, K. Bittinger, A. Gonzalez, C. Lozupone, J.R. Zaneveld, Y. Vázquez-Baeza, A. Birmingham, and E.R. Hyde. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1):1–18, 2017.

[119] P.J. McMurdie and S. Holmes. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol*, 10(4):e1003531, 2014.

[120] A.D Willis. Rarefaction, alpha diversity, and statistics. *Front Microbiol*, 10:2407, 2019.

[121] E.S. Cameron, P.J. Schmidt, B.J-M. Tremblay, M.B. Emelko, and K.M. Müller. To rarefy or not to rarefy: Enhancing microbial community analysis through next-generation sequencing. *bioRxiv*, 2020.

[122] C.E. Shannon and W. Weaver. The mathematical theory of communication. *Urbana: University of Illinois Press*, 1949.

[123] E.K. Morris, T. Caruso, F. Buscot, M. Fischer, C. Hancock, T.S. Maier, T. Meiners, C. Müller, E. Obermaier, D. Prati, and S.A. Socher. Choosing and using diversity indices: insights for ecological applications from the german biodiversity exploratories. *Ecol Evol*, 4(18):3514–3524, 2014.

[124] J.R. Bray and J.T. Curtis. An ordination of the upland forest communities of southern wisconsin. *Ecol Monogr*, 27(4):325–349, 1957.

[125] P.J. Schroeder and D.G. Jenkins. How robust are popular beta diversity indices to sampling error?. *Ecosphere*, 9(2):e02100, 2018.

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

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

[128] M.I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol*, 15(12):1–21, 2014.

[129] P. Hugon, J-C. Lagier, P. Colson, F. Bittar, and D. Raoult. Repertoire of human gut microbes. *Microb Pathog*, 106:103–112, 2017.

[130] P. Deng and K.S. Swanson. Gut microbiota of humans, dogs and cats: current knowledge and future opportunities and challenges. *Br J Nutr*, 113(S1):S6–S17, 2015.

[131] R. Schierwagen, C. Alvarez-Silva, M.S.A. Madsen, C.C. Kolbe, C. Meyer, D. Thomas, F.E. Uschner, F. Magdaleno, C. Jansen, A. Pohlmann, and M. Praktiknjo. Circulating microbiome in blood of different circulatory compartments. *Gut*, 68(3):578–580, 2019.

[132] R. Schierwagen, C. Alvarez-Silva, F. Servant, J. Trebicka, B. Lelouvier, and M. Arumugam. Trust is good, control is better: technical considerations in blood microbiome analysis. *Gut*, 69(7):1362–1363, 2020.

[133] J.B. Hughes, J.J. Hellmann, T.H. Ricketts, and B.J.M. Bohannan. Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl Environ Microbiol*, 67(10):4399–4406, 2001.

[134] C. Wen, L. Wu, Y. Qin, J.D. Van Nostrand, D. Ning, B. Sun, K. Xue, F. Liu, Y. Deng, Y. Liang, and J. Zhou. Evaluation of the reproducibility of amplicon sequencing with illumina miseq platform. *PloS One*, 12(4):e0176716, 2017.

[135] Jizhong Zhou, Liyou Wu, Ye Deng, Xiaoyang Zhi, Yi-Huei Jiang, Qichao Tu, Jianping Xie, Joy D Van Nostrand, Zhili He, and Yunfeng Yang. Reproducibility and quantitation of amplicon sequencing-based detection. *ISME J*, 5(8):1303–1313, 2011.

[136] L.I. Evtushenko and M. Takeuchi. The family microbacteriaceae. *The Prokaryotes*, 3:1020–1098, 2006.

Acknowledgements

Here I would like to take the opportunity to thank everyone who has supported me on my professional and personal journey through the past three years. I would like to especially thank the following persons.

First, I would like to thank Prof. Dr. Stefan Bonn for giving me the opportunity and freedom to pursue my research at the Institute of Medical Systems Biology. I would also like to thank Prof. Dr. Andrew Torda for taking on the role of co-supervisor and always having time to talk at his office.

Thanks to all the members, and past members, of the IMSB. Especially to Sabine Wehrmann for all her administrative help in dealing with the UKE and the UHH. Many thanks to Dr. Pierre Machart for taking the time to answer all my machine learning related questions and his patience. Thanks to Sven Heins for the collaboration on our project and the IT support. For IT support I would like to give special thanks to Dr. Sergio Oller as well, thanks to him I saved many frustrating hours. Also thanks to Dr. Karin Kloiber for taking the time to help me finalize my first manuscript. Finally, I would like to thank Pierre, Yu and Rotem for their friendship during the past years.

Lastly, I would like to thank my family and friends for their ongoing support. Special thanks to my godmother Dorothea for always believing in me.

List of the hazardous substances

No H&P-substances were used in this dissertation.