Identifying potential microbial aetiologies of Kawasaki Disease with metagenomics and metaproteomics

Andrew James McArdle

Department of Infectious Disease, Imperial College London

Thesis submitted for the award of DPhil

Abstract

Kawasaki disease (KD) is an acute inflammatory disorder of early childhood. It causes coronary artery aneurysms, due to the vasculitis which underlies much pathology.

Despite epidemiological evidence supporting an infectious trigger in predisposed individuals, decades of research have found no strong evidence to implicate an individual organism. A diverse range of organisms have been suggested as the cause, frequently with limited or contradictory evidence. Studies are typically small and consider a narrow range of organisms, mostly pathogens.

Two promising sites to search for microbial material are the oropharynx and immune complexes (IC). The former is the main microbial entry portal, and the latter form in the subacute phase of illness and may include microbial antigen.

I present bioinformatic analyses aiming to identify microbes associated with KD from these two sources. Data comprises oropharyngeal metagenomics (116 cases; 101 controls) and metaproteomics of ICs (112 cases; 128 controls).

Extensive method development was required, with a focus on mitigating against laboratory and reference database contamination for metagenomics and finding database search methods which could cope with a vast metaproteomic search space.

Metaproteomic analyses yielded no strong evidence of microbial associations, though the data allowed novel, exploratory analysis of immunoglobulin peptides and germline immunoglobulin locus usage. Since immunoglobulin is strongly implicated both genetically and pathologically in KD, the methods developed here show potential for further development and application.

Metagenomic analyses found organisms associated with KD, including *Abiotrophia defectiva* and *Lautropia mirabilis* (six-fold higher abundance, Q values 0.06 and 0.05 respectively). The strengthening of relationships when age-profiles are modelled, and previous implication of *A. defectiva* in infective endocarditis add to the potential significance of the findings.

I interpret the results in the light of the epidemiology of KD and consider the limitations of this study, and challenges inherent in the search for causes of KD.

Acknowledgments

In this thesis I present the major outputs of three years of research to identify potential microbial causes of KD from metagenomic and proteomic data. My analyses sit at the end of many years of work by superb and dedicated scientists, and I can truly say that I "stand on the shoulders of giants."

I am grateful to Mike Levin, Myrsini Kaforou and the whole Department of Paediatric Infectious Disease for welcoming me into this extraordinary and collaborative group, and express my gratitude for the opportunities that this period has afforded for intellectual and personal development. This would not have been possible without the generous funding of the Lee Family Foundation to the Imperial 4i programme. Mauricio Barahona was generous with time and advice.

Stephanie Menikou deserves special thanks for allowing me to join this exciting project and sharing the data and her considerable expertise and learning developed over several years preceding my fellowship.

The entire project would not have been possible without the generous collaboration of Jane Burns and her excellent team at University of California San Diego, since the majority of samples came from their team. It has been a privilege to join the weekly Kawasaki Zooms, initially in a pre-pandemic period where such an online meeting seemed quite novel!

I'm extremely grateful to Lesley Hoyles who kindly agreed to provide additional advice for the metagenomic analyses, and has done so patiently and supremely well through both e-mails and online conversations. Much of the work presented here would not have been possible without her support.

My work depends critically on myriad pieces of software produced by academics around the world and supported (often without specific funding) amazingly well. I raised exactly 100 issues on GitHub through the course of this fellowship, and 81 of these have been closed. Significant support was also provided by several individuals through e-mail. Special recognition goes to Matthias The and Thys Potgeiter who provided endless help with their cutting-edge software Quandenser and Metanovo, respectively.

The high-performance computing infrastructure at Imperial has been vital for many aspects of this work, and at multiple stages I have needed the support of the team to get specific pieces of software working as required.

I am grateful to my early and late stage reviewers, Tim Ebbels and Mike Cox, for encouragement and helpful direction.

Jane Webb did an excellent job of welcoming me into the group and helping orient me to systems, people, spaces and processes, and I know that a lot of what we take for granted in the group wouldn't happen without her quiet support.

I thank my fellow traveller scientists of room 224 for the collaborative, helpful and fun environment, which made both days of progress and stasis more enjoyable.

Contents

Figures

Tables

Abbreviations

ANCOVA	Analysis of covariance
AP	Affinity purification
BCG	Bacille Calmette-Guérin
BCR	B-cell receptor
BLAST	Basic local alignment and search tool
Bp	Base pair
CAA	Coronary artery aneurysm
cDNA	Complementary deoxyribonucleic acid
CDR	Complementarity determining region
CMV	Cytomegalovirus
COG	Cluster of orthologous genes
CRP	C-reactive protein
DDA	Data-dependent acquisition
DNA	Deoxyribonucleic acid
EBV	Epstein-Barr Virus
ED	Emergency department
EM	Electron microscopy
FDR	False discovery rate
Gb	Gigabyte
Gbp	Giga base pair
GC	Guanine-Cytosine
GTDB	Genome Taxonomy Database
HCoV	Human coronavirus
HHV	Human Herpesvirus
HIV	Human immunodeficiency virus
HLA	Human leukocyte antigen
HMP	Human microbiome project
HMW	High molecular weight
HPC	High performance computing
iBAQ	Intensity based absolute quantification
IC	Immune complex
IgA	Immunglobulin A
IgE	Immunoglobulin E

IGF	Imperial Genomics Facility
IgG	Immunoglobulin G
IgM	Immunoglobulin M
IL	Interleukin
IMGT	Immunogenetics
IQR	Interquartile range
IVIG	Intravenous Immunoglobulin
KD	Kawasaki disease
kDa	Kilodalton
LC	Liquid chromatography
LMW	Low molecular weight
MAG	Metagenome assesmbled genome
MBR	Match-by-run
MDS	Multidimensional scaling
MIS-C	Multisystem inflammatory syndrome in children
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
MS1	First round mass spectrum (precursors)
MS2	Second round mass spectrum (fragments)
MW	Molecular weight
NCBI	National Center for Biotechnology Information
ORF	Open reading frame
OUT	Operational taxonomic unit
PCA	Principal components analysis
PCoA	Principal co-ordinates analysis
PCR	Polymerase chain reaction
PEG	Polyethylene glycol
PIMS-TS	Paediatric Inflammatory Syndrome Temporally Associated with SARS-CoV-2
PMN	Polymorphonuclear cells
PSM	Peptide spectrum match
RA	Relative abundance
RAM	Random access memory
REOF	Rotated empirical orthogonal function

RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid
RT	Room temperature
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SEC	Size exclusion chromatography
SNP	Single nucleotide polymorphism
TB	Tuberculosis
TNF	Tumour necrosis factor
UCSD	University of California San Diego
UV	Ultraviolet
VZV	Varicella Zoster Virus
WBC	White blood cells

Declaration of originality

Unless otherwise stated in the text the work presented herein is my own, produced under supervision of Professor Mike Levin, Dr Myrsini Kaforou and Professor Mauricio Barahona. The data analysed has been largely produced by Dr Stephanie Menikou over a number of years of painstaking laboratory work.

Dr Andrew McArdle

15 August 2023

Copyright declaration

1 | Introduction

Kawasaki disease (KD) is an acute auto-inflammatory disorder predominantly affecting preschool children. KD has a worldwide distribution, with incidence estimates (per 100 000 children under 5 years old) ranging from 8 in England and New Zealand, to 18 and 22 in the USA and Canada, and 83, 134 and 264 in Taiwan, Korea and Japan respectively (Singh, Vignesh & Burgner, 2015). While the incidence in Western countries appears stable, numbers of cases in Japan and Korea continue to rise, with most recent estimates in Japan of 300 per 100 000 (Makino et al., 2018).

The disease presents with prolonged fevers, misery and a range of clinical features (McCrindle et al., 2017). Many organ systems are typically involved, predominantly the skin, mucosa, lymphatic system and blood vessels.

The inflammation of blood vessels (vasculitis) is responsible for the major complication of KD: coronary artery aneurysms (CAA). These dilatations of the vessels supplying the myocardium occur in up to 30% of untreated patients in historical cohorts, and can have lifelong consequences (see Figure 1). 50% of those with giant coronary artery aneurysms will require vascular intervention within 30 years of follow-up (Newburger, Takahashi & Burns, 2016). Around 1% of untreated patients die, usually from

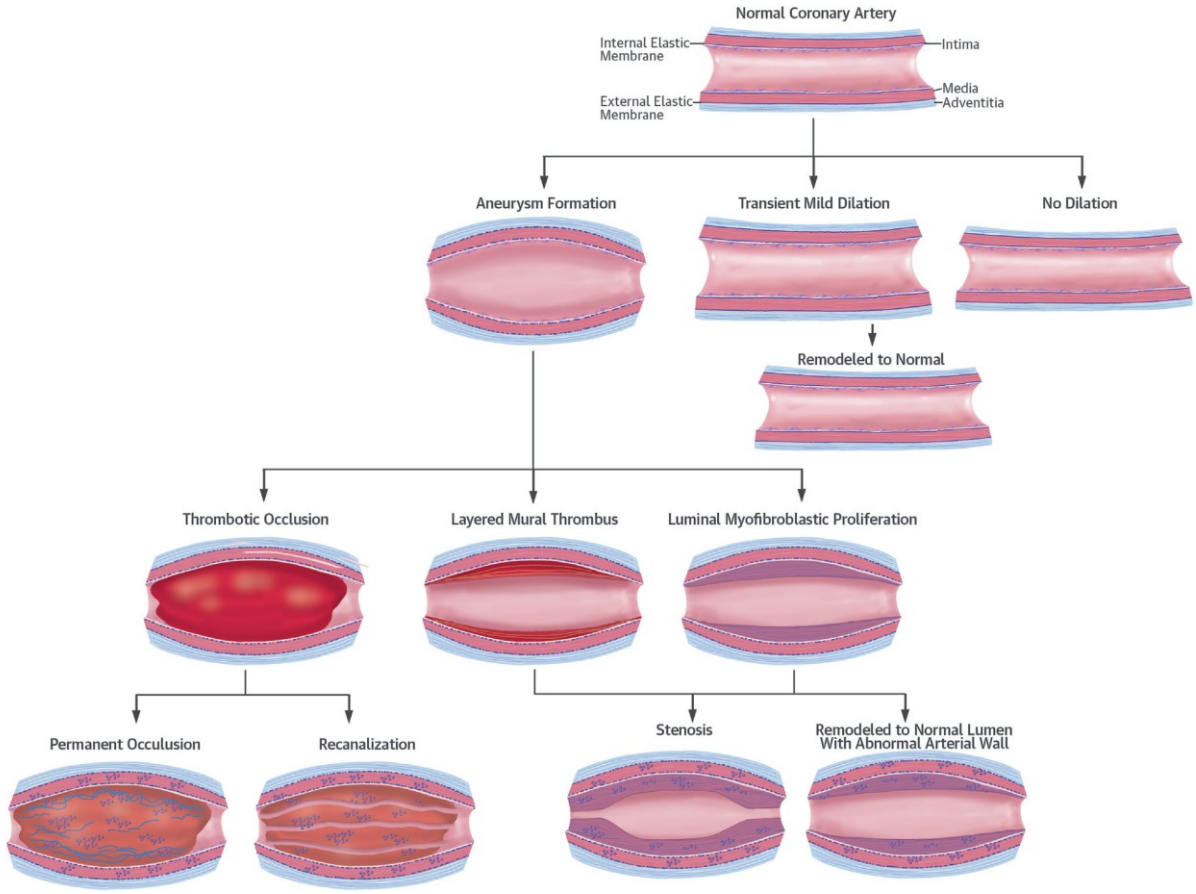

Figure 1 Aneurysm formation and complications in Kawasaki Disease. Reproduced from Newburger, Takahashi and Burns (2018).

cardiac complications. Outside of low resource settings, KD represents the commonest cause of acquired heart disease in children, and children with CAA require lifelong follow-up (McCrindle et al., 2017).

It is not well understood why the vasculitis in KD has a predilection for the coronary arteries. Vasculitides more generally are often classified according to the size of vessels they target, but Hoffman & Calabrese (2014) highlight the nuances of anatomical specificity in vasculitis. They describe the diversity of vessel phenotypes, with varied protein expression (including Toll-like receptors and adhesion molecules), matrix composition and interactions with circulating cells. They posit that anatomical specificity may be driven by autoantibodies to anatomically-restricted antigens, or restricted deposition of circulating antigen or immune complexes. Mice can be driven to exhibit disease which recapitulates similar coronary artery inflammation and dilatation as in KD following intraperitoneal injection of a *Lactobacillus casei* cell wall extract or *Candida albicans* water-soluble fraction (Noval Rivas & Arditi, 2020). Despite this, it is not known why these triggers drive inflammation specifically in coronary arteries.

Clinical features and diagnosis

KD has remained a syndromic diagnosis (*i.e.* based on a combination of symptoms, signs and investigations) from its first published description (Kawasaki, 1967), and there remains no diagnostic test. Current diagnostic criteria from the American Heart Association (McCrindle et al., 2017) rest upon prolonged fevers and the principal clinical features: inflammatory changes of the oral mucosa; bilateral non-purulent conjunctivitis; generalised erythematous rash; swelling or erythema of hands or feet with later desquamation, and cervical lymphadenopathy. Four or more of these features with fever for 5 or more days are the main way to meet the criteria, though diagnoses can be made with fewer criteria, especially in the presence of CAA ("incomplete" or "partial" KD). A range of other clinical features are recognised, while not forming part of the criteria – perhaps most interestingly, inflammation of the Bacille Calmette-Guérin (BCG) scar. This feature, though rare in Western practice, is highly specific for KD (Uehara et al., 2010).

The prevalent clinical features of KD are shared with other disorders, mostly of an infectious or inflammatory nature, making diagnosis difficult. Further, many children do not present with the full set of clinical features, and the diagnosis is only made by the detection of CAA, often at a late stage (Pilania, Bhattarai & Singh, 2018). Scarlet fever (a disease caused by group A streptococcus), measles or adenovirus infection and Stevens-Johnson syndrome are common differential diagnoses in children with suspected KD.

In Japan, the diagnosis of KD is often made early, even before the fifth day of fever. However, in the UK, the diagnosis is made a median of 7 days after symptom onset (Tulloh et al., 2019).

Gene-expression based diagnostics have been developed, but await clinical validation and implementation. These include a 13-transcript based classifier to distinguish from children with fever of diverse causes developed within our group (Wright et al., 2018). Other groups have presented classifiers that discriminate KD from adenovirus infection and group A streptococcus infection (Jaggi et al., 2018) and from viral and bacterial infections (Hu et al., 2020).

Treatment

Effective treatments for KD include a range of anti-inflammatory medicines, foremost intravenous immunoglobulin (IVIG). This has been shown to reduce the incidence of CAA if given within the first 10 days of illness (Oates-Whitehead et al., 2003). There is good evidence for adjunctive treatment with corticosteroids in patients whose response to IVIG is inadequate, and it is increasingly being considered as part of first-line therapy (Wardle et al., 2017). Other treatments include immunomodulators like anti-tumour necrosis factor (TNF) (Tremoulet et al., 2014) and anti-interleukin 1 (IL-1) biologics (Koné-Paut et al., 2021), though there is limited evidence especially for primary treatment (Yamaji et al., 2019).

Current outcomes

Despite a range of available treatments, children still experience adverse outcomes. A recent UK surveillance study found that 19% of children had CAA on initial echocardiogram (Z-score ≥ 2.5) with 8% having persistent CAA beyond 30 days and 1.5% with giant CAA (Tulloh et al., 2019). The mortality rate was 0.4% (contributed by cases diagnosed at post-mortem).

Younger patients are at higher risk for both delayed diagnosis and poor outcomes. This is related in part to the greater proportion of children under 1 year who present with atypical disease (fewer than four of the major clinical features while still having CAA): 21% vs 6% in older children.

Importance of understanding aetiology

From its recognition, researchers have sought to understand the aetiology of KD, from genetic predisposition through to precipitating triggers or causative agents. A microbial triggering aetiology has long been hypothesised, most likely prompted by the resemblance of KD to scarlet fever and viral exanthemous illnesses, though there has been limited evidence of any response to antimicrobial agents.

Identification of precipitating triggers of KD would be a major advance in scientific understanding of this modern disease. Such knowledge would aid further understanding of the pathogenesis of KD and

the explanation of its epidemiology. More clinically, routes to prevention, better treatment and diagnostics could be opened. Transferable insights could be developed into other inflammatory disorders.

Current state of knowledge of KD aetiology

History

The first patient in the series which prompted Dr Tomisako Kawasaki (1925-2020) to recognise KD presented in 1961, but it took until 1970 to publish a series of 50 patients. In retrospect, cases have been identified as early as the 1950s. It was only around 1970 that cardiac involvement, including vasculitis of the coronary arteries, was recognised as part of the syndrome. This still preceded the English language publication of Kawasaki's case series in 1974. Interestingly, KD was independently being recognised by clinicians in Hawaii around the same time, although it was on seeing photographs of cases in Japan that the link was established (Burns et al., 2000).

To this day, it is not certain whether KD represents a genuinely modern disease, or a longstanding condition, whose recognition was prompted by changing epidemiology and the lack of response to antibiotic treatment seen in clinically similar conditions like scarlet fever. One hypothesis, based on the identification of cases in the West as far back as the 1870s, but the absence of similar cases in Japan, is that KD has been endemic in the West for a long time but was new in Japan (Kushner et al., 2008). The premise is that in the West, the condition was frequently classified as unusual forms of other illnesses (including scarlet fever and Stevens-Johnson syndrome) or labelled as infantile polyarteritis nodosa. The earliest identifiable cases in Japan in the 1950s were classified on the basis of similar reports in the western literature, due to the lack of precedent in Japan. Oral history reveals a consensus among contemporary Japanese paediatricians that children with the full symptom cluster of KD were a novelty to them.

Genetics

The genetics of KD has been widely studied from the 1980s onwards. Early studies were predominantly family based, using sibling cases of KD (Onouchi, 2018). Findings included increasing sharing of human leukocyte antigen (HLA) haplotypes, a single nucleotide polymorphism (SNP) in the IL-4 gene, and 10 chromosome regions identified by genome-wide linkage. Subsequent case-control population studies were hypothesis driven and frequently failed to replicate findings; however valid risk alleles of ITPKC, CASP3 and ORAI1 were found.

The most confident findings awaited the application of microarray-based genome-wide association studies. These have included replicated identification of risk alleles of FCGR2A, BLK and CD40.

Identical, or co-located, SNPs in BLK and CD40 have also been associated with rheumatoid arthritis and systemic lupus erythematosus. In contrast, ITPKC and CASP3 SNPs have not been associated with other diseases and may therefore give specific insight into KD pathogenesis.

Recently, researchers sought to increase power by pooling studies of children with both IgA vasculitis and KD, in the hope of finding shared risk alleles. A SNP in the NAGPA intron was found to be positively associated with KD and IgA vasculitis in the discovery cohort, but could only be validated in the latter condition (Carmona et al., 2021).

No monogenic causes of KD have been identified. There are rare monogenic disorders (Jain et al., 2018) which give rise to vasculitis, including STING-associated vasculopathy of infancy (SAVI) and deficiency of adenosine deaminase 2 (DADA2), but coronary artery vasculitis is not specifically reported.

Less studied is the genetics of outcomes in KD. In addition to potentially guiding therapy, such findings could also help elucidate aetiopathogenesis. Patients with the risk alleles of ITPKC and CASP3 are at higher risk for resistance to IVIG (Onouchi et al., 2013). More recently, among European populations, a SNP in an intergenic region of chromosome 20, with potential long-range interactions with PLCB1, has been identified to confer higher risk of CAA (Hoggart et al., 2021).

Unsurprisingly, findings are not completely consistent between different regions and ethnicities. Although rendering the pooling and meta-analysis of results challenging, greater insight may be added by the heterogeneity. CASP3 and CD40 risk alleles are more common, for example, in the lower risk European population than the high risk Japanese, and ITPKC is not found as an association in Korean populations (Onouchi, 2018). In the previous study, significant associations with CD40 could not be found in European patients (Hoggart et al., 2021). This same study estimated the heritability of KD in Europeans to be between 5 and 20%, of which only a small fraction is explained. Although heritability in high risk populations may be higher, it is clear that much remains to be discovered about the genetics of KD, and environmental factors may be of greater importance.

Aetiological studies

Over decades, there have been myriad studies looking for factors involved in the aetiology of KD, and more specifically, the proximal triggers of KD. The methodologies are diverse and results frequently conflicting or difficult to integrate. In 2005, Burgner & Harnden reviewed proposed causes of Kawasaki Disease, listing 16 infectious organisms or groups of organisms and four environmental causes (see Figure 2).

Search terms: "kawasaki disease"[title] AND (aetiology[title] OR etiology[title] OR causative[title] OR trigger*[title] OR origin[title] OR agent[title] OR association[title] OR infection)

Filters: Case reports, classical article, clinical study, comparative study, corrected and republished article, journal article, letter, meta-analysis, multicentre study, observational study, review, systematic review

Date: 2 August 2021

Results: 1089

A broad search of PubMed (see above right) produces 164 results relevant to KD aetiology, including case reports, case series, case-control studies and epidemiological studies. 63 of these have been published from 2006.

Thirty-seven studies are comparative cohort studies, four are case-control studies and 21 are epidemiological or registry studies, which have the potential to detect associations. The remainder comprise case reports, case series, single cohorts, post-mortem studies and reviews. Culture, serology and polymerase chain reaction (PCR) methods predominate.

Proposed causative organisms span both RNA and DNA viruses and a diverse range of bacteria. DNA viruses comprise adenovirus (Jaggi et al., 2013), cytomegalovirus (CMV; Catalano-Pons et al., 2005), Epstein-Barr virus (EBV; most frequently as a negative association; Fuse *et al.*, 2010), Human Herpesvirus (HHV) 6 (Okano et al., 1989) and HHV7 (Burns et al., 1994:p.7), varicella zoster virus (VZV; Lee & Huang, 2004) and parvovirus B19 (Nigro et al., 1994). RNA viruses comprise coronaviruses (NL63, 229E, SARS-CoV-2) (Shirato et al., 2014; Cazzaniga et al., 2020), bocavirus (Bajolle et al., 2014), enteroviruses (Weng et al., 2018), dengue (Guleria et al., 2018), influenza (Joshi et al., 2011), measles (Kuijpers et al., 2000), parainfluenza 3 (Moreira et al., 2010) and retroviruses (Lin et al., 1992). Arboviruses as a class have also been epidemiologically linked (Paniz-Mondolfi et al., 2020), and a series of studies implicate an as-yet unidentified virus (Rowley et al., 2004, 2005, 2008, 2011). Bacteria comprise group A

Putative etiological agent/ environmental association
Adenovirus
Herpesvirus
Mycoplasma species
Toxigenic streptococci
Viridans streptococci
Toxigenic staphylococci
Propionibacterium acnes
Ehrlichia chaffeensis
Rickettsia species
Epstein Barr virus
Retrovirus
Human coronavirus New Haven
Measles virus
Chlamydia pneumoniae
Bartonella henselae
Coxiella burnetii
House dust mite
Mercury
Carpet shampoo
Residence near body of water

Figure 2 Agents and exposures implicated in the aetiology of Kawasaki Disease. Reproduced from Burgner and Harnden (2005).

Streptococci and *Staphylococcus aureus* (primarily considered through toxin or superantigen production) (Matsubara & Fukaya, 2007; Shen et al., 1990), *Streptococcus sanguis* (Shinomiya et al., 1987), *Mycoplasma pneumoniae* (Lee et al., 2011), *Cutibacterium* (formerly Propionibacterium) *acnes* (Kato et al., 1983), *Yersinia pseudotuberculosis* (Horinouchi et al., 2015), *Chlamydia pneumoniae* (Strigl et al., 2000), *Coxiella burnetii* (Swaby et al., 1980), *Ehrlichia canis* (Edlinger, Benichou & Labrune, 1980) and Rickettsiae (Shishido, 1979). Overall, it appears that the focus has shifted from bacterial to viral causes over the last four decades.

Importantly, the evidence for any of these organisms having a causative role in KD is generally weak and initial reports are frequently contradicted by further studies. To give an example of a relatively well-studied association, a novel coronavirus was first reported as a potential trigger in 2005 (Belay et al., 2005) in parallel with its identification as a cause of respiratory illness (Esper et al., 2005). These authors used archived respiratory tract secretions from 11 children with KD and 22 matched controls (only secretions from those negative by direct immunofluorescence for other respiratory viruses were included in the archive). They found the "New Haven" coronavirus (a strain of the HCoV-NL63 lineage) in 8 of 11 KD patients and 1 of 22 controls, giving an odds ratio of 16.0 (95% confidence interval 3.4-74.4).

Ebihara *et al.* (2005) subsequently screened 19 children with KD in Japan and 208 children with respiratory illnesses, finding only 5 positive samples among the controls and none among the children with KD. Shimizu *et al.* (2005) studied 57 respiratory samples from 48 KD patients at centres in the USA and Netherlands, finding only one positive sample.

Another group studied 21 children with complete and incomplete KD and 33 healthy controls in Germany (Lehmann et al., 2009). HCoV-NL63 was found in one healthy child and no cases, with similar rates of IgG positivity. Kim *et al.* (2012) conducted a larger study in South Korea of 55 children with KD and 78 age-matched healthy controls without recent respiratory symptoms. The overall detection rate of respiratory viruses was similar, and HCoV-NL63 was identified in only one control patient.

In contrast, Chang *et al.* (2014) found a strong association of respiratory virus detection with KD (50.4% vs 16.4% in healthy controls), with 12 of 226 KD (5%) positive for HCoV-NL63 and none of 226 in the control group. However, rhinovirus, followed by enterovirus, was responsible for the bulk of the excess virus detections among KD cases. Shirato *et al.* (2014) studied 15 KD cases and 52 controls serologically, and found no excess of seropositivity to HCoV-NL63 among cases before treatment with intravenous immunoglobulin and in recovery. However, they noted an excess of seropositivity to HCoV-229E.

At an epidemiological level in South Korea no association was be found between monthly rates of coronavirus detection in sentinel hospitals and monthly incidence of KD in the subsequent month (Choe, An & Choe, 2021). In a study published after the literature search described, Patra et al. (2022) conducted a systematic review comprising eight case-control and two prospective studies (including most cited above) and concluded with low certainty that there is an "increased risk of KD in children infected with HCoV."

The recent discovery of an inflammatory disorder precipitated by exposure to SARS-CoV-2 is clearly relevant, and will be considered extensively in the discussion (page 142).

A smaller number of studies have applied less targeted molecular approaches to the problem. The earliest used DNA hybridisation to detect parvovirus and herpesvirus DNA in serum, and 16S rRNA PCR to detect bacterial sequences in buffy-coat, synovial fluid and post-mortem tissue samples (Rowley et al., 1994).

In the last decade, there have been 7 published studies applying next-generation sequencing techniques. No metaproteomic studies have been identified. Four studies have investigated the gut microbiome in KD. Chen *et al.* (2020) assessed the faecal microbiome of 30 children with KD (acutely and in convalescence) and 30 age- and sex- matched controls using 16S ribosomal ribonucleic acid (rRNA) sequencing. They reported lower biodiversity in acute KD than healthy controls, but no differences in richness. The greater difference in microbiome composition was noted in convalescence, and the authors noted the majority of patients had received antibiotics. Several genera were found to be overabundant in acute KD cases, including Staphylococcus.

Shen *et al.* (2020) studied fecal microbiota of 48 children with KD and 46 healthy controls, finding reduced richness and diversity. Bacteriodetes and Dorea were found to be less abundant in KD patients, and no organisms were reported to be over-abundant. Kinumaki *et al.* (2015) applied metagenomic sequencing to 28 KD patients alone acutely and 4-6 months later. At genus level, abundance of Rothia and Staphylococci were associated with the acute phase, and at species level, several streptococcus species, including Streptococcus pneumoniae. Khan *et al.* (2020) similarly studied 5 KD patients and 3 healthy controls, though the small numbers and lack of measures of precision or hypothesis tests renders the results uninformative.

Thissen *et al.* (2018) conducted a pilot study of 11 patients with KD and 22 controls. For each patient, DNA was extracted from whole blood. Other sample types were pooled across sample groups and are therefore uninformative. A torque tenovirus was identified in two KD patients.

Hamada *et al.* (2016) made mixed DNA and cDNA libraries from serum and pharyngeal swabs during two episodes of KD in a single patient, obtaining two acute serum libraries, one acute pharyngeal library and one convalescent serum. Streptococcal reads were identified in many samples, but further reporting is limited.

Anne Rowley and colleagues' painstaking work investigating a possible novel viral aetiology was triggered by the finding of a KD-associated antigen in the cytoplasm of bronchial epithelium from children with KD (10/13) and not control subjects (n=9). A synthetic antibody derived from IgA-secreting plasma cells in the vascular tissue of a KD patient was used as a probe (Rowley et al., 2004). Further work identified that the detected material resides in cytoplasmic inclusion bodies with staining consistent with the presence of protein and nucleic acid (Rowley et al., 2005) and antigen was detected in the gastrointestinal tract of some deceased KD patients (Miura et al., 2005). Studies of late-stage KD fatalities and non-infant control fatalities showed high prevalence of antigen detection (6/7 KD cases and 7/27 controls), and staining was specific for RNA and not DNA. This would be consistent with an RNA virus.

In a subsequent study (Rowley et al., 2011) electron microscopy (EM) and laser capture was applied to inclusion bodies. EM findings suggested the inclusion bodies are virus-like particles and pyrosequencing revealed no known viral sequences. It was suggested that unknown sequences required further analysis, though no further reports can be found. Most recently, the group report identification of a specific protein epitope from inclusion body-binding monoclonal antibodies (Rowley et al., 2020).

Broader epidemiological studies

The studies described in the previous section focus on potential infectious aetiologies of KD, rendering them epidemiological in nature. However, many other epidemiological studies of KD have been broader, including environmental, meteorological and other phenomena.

Seasonal patterns have been observed worldwide, generally with peaks between January and March in the extra-tropical northern hemisphere and May through June in the tropics and southern hemisphere (Burns et al., 2013). Japan has seen three nationwide KD epidemics between 1979 and 1986, with "wave-like" spread across the country (Rowley & Shulman, 2018). Spatio-temporal clustering has been demonstrated in Japan (Sano et al., 2016), Canada (Hearn et al., 2018) and San Diego (Burns et al., 2021a). Sequential onset of KD within 10 days have been reported in around 1% of siblings in Japan (Fujita et al., 1989).

In a series of studies using decades of KD registry data, Burns and colleagues have demonstrated that periods of high KD incidence in Japan, Hawaii and San Diego correlate with specific tropospheric wind directions (Rodó et al., 2011, 2014). In models of air transport, these winds were shown to originate in north-eastern China, suggesting possible long-distance transport of an airborne agent. Most recently, the group has shown that the KD cases which cluster in space and time in San Diego also share more clinical features than expected by chance alone (Burns et al., 2021a). This raises the hypothesis that distinct triggering agents are associated with differences in clinical features.

Manlhiot *et al.* (2018) ambitiously attempted to integrate demographic, household, environmental, atmospheric and climatological factors utilising data from a single hospital, Canadian registry and worldwide KD incidence. In their hospital cohort, they found independent associations between KD and larger family size, reduced environmental allergy exposure, recent house construction and reduced tree coverage. The also replicated an earlier finding of an association with recent deep carpet cleaning (Patriarca et al., 1982). Children with KD were reported more frequently to be unwell up to a month before diagnosis, and illnesses in household members were also increased. Nationally, associations with wind direction and atmospheric particles from plants and fungi were identified. Internationally the authors were able to explain 84% of the variation in KD incidence using proportion of population of Asian ancestry, urbanisation and gross national product, and distance west and south of North-east China. However, it is unclear how much Asian ancestry alone may have contributed to this model.

Evidence supporting a microbial origin

The greatest efforts in KD aetiology research have gone into identifying genetic risk factors and causative microbial agents. Overall, genetic studies have yielded far more confident causal associations, and provided implications for pathogenesis. It is helpful to consider whether the hypothesis of a microbial cause is a reasonable one.

Superficially, KD has similarities of presentation and laboratory features to many other infectious illnesses, including measles, adenovirus infection and scarlet fever. Prominent and common symptoms of mucosal changes, red lips and cervical adenopathy focus on the upper respiratory tract, which is the main route by which children encounter pathogens. Nonetheless, some other diseases which can appear similar, including Stevens-Johnson syndrome, are not infectious, or represent post-infectious immune-mediated syndromes.

The peak incidence between 6 months and 5 years of age (lower in Japan) coincides with the window of lowest adaptive immunity, between the loss of passive protection of material antibodies, and a time when many infectious diseases have already been encountered for the first time. Coupled with

the relative rarity of recurrent KD (3-4% in Japan and 1.7% in the USA; Medaglia *et al.*, 2021), compared with, for example, IgA vasculitis (around a third), this is consistent with triggering by first exposure to a ubiquitous microbe in a genetically susceptible individual.

Indeed, in Japan the cumulative incidence of KD during childhood is around 1.4%, so the risk of recurrence is only 2-3 times higher than the background rate. Based on the incidence ratio between Japan and the USA, the cumulative incidence of KD in the USA is likely to be around 0.1% (Maddox et al., 2015), suggesting a greater relative risk of recurrence. If multiple ubiquitous triggers could cause disease in susceptible individuals, recurrence would be expected a common phenomenon, unless protection from recurrent KD is not based on specific immunity to a pathogen (e.g., exhaustion of a pool of KD-driving adaptive immune cells).

As described earlier in the chapter, the temporal and spatial dynamics of KD are consistent with an infectious agent. Finally, many of the genetic risk factors identified lie firmly within the adaptive immune system, and more specifically the B cell and antibody response.

It is striking that despite decades of research with no single causative microbial agent demonstrated, the paediatric community mostly remains convinced that KD represents an abnormal immune response to a microbe. It is thus important to consider why it has not yet been possible to confidently identify such an agent.

The studies cited are frequently limited by reliance on small sample sizes, especially for studies based on laboratory analyses. Further, identification of infections frequently depended upon targeted serological, culture or PCR-based tests. Serological tests can lack sensitivity, and those for immunoglobulin M (IgM) are frequently non-specific due to cross-reactivity. Culture methods restrict identifications only to culturable organisms with the selected medium, and frequently only certain organisms are considered.

Heterogeneity

There is debate as to whether KD is a single disease process with single aetiopathogenesis, or multiple diseases with distinct triggers and pathophysiology but overlapping phenotypes. Incomplete KD (where coronary artery aneurysms develop in the absence of the full symptomatic criteria) and KD with myocardial shock represent clearly demarcated subgroups (McCrindle et al., 2017). Both of these atypical phenotypes are seen more frequently in KD patients presenting outside of the main pre-school age range (1-5y).

It is not known why phenotypes may differ at the extremes of age. If the main age range of KD is determined by a period in which the immune system is most primed for the KD response, it could be

expected that a greater genetic and environmental "push" is needed for KD to develop in those outside this window. It is also possible that different triggers predominate at different ages.

Sim et al. (2018) explored the association of two known polymorphisms (in BLK and FCGR2A) among different KD subgroups (age strata, complete vs incomplete, gender, coronary artery aneurysms, family history, recurrent, IVIg responding, though not presence of shock). They report lack of associations in those over 5 years old and those with incomplete KD, suggesting that genetic factors may be distinct. However, formal testing of heterogeneity was not undertaken, and effect sizes overlap.

Jackson et al. (2021) clustered KD patients based on blood proteomic and transcriptomic signatures, finding clusters which each shared more in common with bacterial or viral responses. Burns et al., (2021a) found that KD cases occurring within spatiotemporal clusters shared more clinical features than would be expected by chance. These provide some support for the existence of multiple triggers with variations in pathophysiology.

Sites to seek causative agents

The portal of entry for most systemic infections in children is the upper respiratory tract. Coupled with this, some early features of KD concern the upper airway and associated tissues (i.e. mucosal changes and cervical lymphadenopathy). The upper respiratory tract therefore shows promise as a site to search for possible infectious causes of KD.

Shotgun metagenomics can allow detection of bacteria, archaea, DNA viruses and fungi (the latter with appropriate sample preparation; Ghannoum *et al.*, 2010), as opposed to 16S/18S rRNA-based microbiome profiling, which are limited to bacteria and fungi respectively. Additionally, taxonomic identification is typically more precise, and specific genes and strains can be recognised from reads or assemblies, allowing detection of potential toxins and virulence factors.

My supervisor reported the presence of immune complexes (ICs) in the blood of children with KD (Levin et al., 1985) and multiple other studies have confirmed their presence (Salo et al., 1987; Koike, 1991). Initial studies established that high-molecular-weight (HMW) IC were present in the plasma from the first week of illness and peaked in concentration in the second to third weeks. The complexes were shown to contain IgG, IgA and IgM, and to bind and activate platelets, causing release of inflammatory mediators. The time course suggests that KD may have an initial phase in which an infectious agent or its antigens are present in high concentration, followed by the production of specific antibodies resulting in formation of IC, leading to clearance of the agent.

Additionally, there is interest in understanding what the specific antibody/B-cell receptor and T-cell receptor profiles in infectious and inflammatory disorders can tell us about antigens, superantigens and the degree to which convergent or divergent receptor responses are produced between individuals.

Work on which this study depends

The group designed a study in collaboration with Prof Jane Burns' group at the University of California San Diego (UCSD), in which children presenting at Rady Children's Hospital with KD and contemporaneous febrile controls from the Emergency Department (ED) would have dry throat swabs and plasma or serum collected, with the aim of identifying microbial agents associated with KD within IC and the pharyngeal metagenome.

Throat swabs were shipped to Imperial College and additional swabs from UK KD patients added. My colleague, Dr Stephanie Menikou, managed and undertook all of the necessary laboratory procedures, working up extraction methodologies to optimise DNA recovery, especially from fungi, and achieve optimal library preparation. Sequencing data was obtained from the Imperial Genomics Facility (IGF).

Similarly, Dr Menikou managed and undertook laboratory procedures for plasma and serum samples. Earlier work supported the feasibility and effectiveness of polyethylene glycol (PEG) precipitation to concentrate immune complexes. Precipitates underwent mass spectrometry (MS) on Orbitrap instruments in Bristol and Oxford.

Aims, objectives and hypotheses

My hypothesis is that KD is triggered in genetically susceptible individuals by one or more microbial agents.

The primary aim of my work presented in this thesis is to identify and characterise microbial agents which are associated with KD, and may be involved in triggering the disease. Furthermore I propose that the causative agent may have remained undetected previously because it may be difficult to distinguish from, or represent a component of the normal mucosal flora. By using metagenomic sequencing analysis I will be able to examine both culturable and non-culturable organisms, and distinguish the many organisms that are normally considered commensals.

Key objectives are detailed below. Additional and more specific objectives will be included in relevant chapters.

Pharyngeal metagenome	**Immune complex metaproteome**
Identify organisms in the pharyngeal metagenome associated with Kawasaki Disease: • Considering bacteria, archaea, fungi and DNA viruses as far as possible • Considering species and higher-level taxonomic units, as well as specific strains and organisms carrying accessory genes.	Identify microbial protein antigens associated with KD in circulating immune complexes by shotgun metaproteomics: • Considering bacteria, archaea, fungi and viruses as far as possible • Considering individual proteins, species and higher-level taxonomic units Explore the peptide sequences of antibody variable regions within circulating ICs and the association between locus utilisation and KD

2 | Metagenomic data, quality and contaminant identification

Introduction

The data analysed for this chapter was provided by the Imperial Genomics Facility (IGF) with libraries provided by Dr S Menikou, and library preparation at the Cambridge Genomics Facility. I was involved in the experimental design of the process and in particular the determination of sequencing depth targets and liaision with IGF during sequencing.

The aims of the metagenomic analysis are to identify features whose prevalence is specifically associated with KD, taking into account relevant covariates. However, before such analyses can be undertaken, the quality and depth of sequencing data need to be explored, and taxonomic profiling optimised and finalised. Further, artefactual findings and contaminants need to be accounted for.

In this chapter, the objectives are:

1. Describe the sequencing data quality and relative contributions of bacterial, fungal, viral, archaeal and human DNA
2. Optimise read binning
3. Identify and mitigate effects of contaminated reference genomes
4. Identify potential contaminants and spurious identifications

Methods

Patients and controls

The studies contributing samples to both the metagenomic and metaproteomic (p92 onwards) analyses are described fully in Appendix A (p180) along with details of ethical approval.

Prior to my involvement in the project, samples from children with KD and febrile controls were selected in a single batch from UCSD, where a long-term cohort study of KD is ongoing with extensive clinical data and sample collection. A smaller number of KD cases from children recruited at St Mary's Hospital were also selected. Although precise details of recruitment differ, both studies utilised the American Heart Association guidance to diagnose KD (McCrindle et al., 2017). As described in the Appendix, febrile controls were adjudicated not to have KD, but did have at least one of the KD clinical criteria.

DNA extraction

Dry human throat swab samples were collected in 2 ml Eppendorf tubes and frozen immediately. Two negative control throat swabs were also included. Because the aetiology of KD is unknown, Dr Menikou aimed to develop a DNA extraction method to capture DNA from organisms including

viruses, bacteria and fungi. The QIAmp PowerFecal DNA kit (Qiagen) was used with some modification. A frozen sample of each throat swab was placed into the bead tube (mixture of glass and dry gamet beads) along with the lysis buffer and microbial DNA extraction was performed. The modified protocol involved changes in the beads used, using a FastPrep machine and longer incubation time. This resulted in a sample of 50 µl elution volume. The total DNA concentration, purity and integrity was assessed using Nanodrop (Thermo Scientific), Qubit (Thermo Scientific) and Bioanalyzer (Agilent Technologies, Santa Clara, CA).

All experiments were performed in a microbiological safety cabinet that was UV irradiated before every experiment to offer protection from aerosols and exogenous contamination.

Library preparation

DNA libraries were constructed using KAPA Hyper Prep Kit according to the manufacturer's instructions (Kapa Biosystems, Inc., Wilmington, MA) at Cambridge Genomic services. The workflows from KAPA were used to perform end repair and A-tailing, adapter ligation, post-ligation cleanup, library amplification and post-amplification cleanup.

Libraries were normalised to 20nM for HiSeq using values from Tapestation (Agilent Technologies, Santa Clara, CA). Libraries from batches of 28 samples were pooled in equimolar ratios to a final concentration of 20 nM.

Multiplexing, sequencing and quality control

Samples were sequenced on the HiSeq 4000 platform within the Imperial Genomic Facility of Imperial College London. 2×150 base pair (bp) paired-end reads were requested, with half a lane's depth per sample targeted. Based on estimated throughput of 650-750 giga bp (Gbp) per 8-lane flow cell, this would be expected to achieve an average of 135-156 million paired reads.

Given the uncertain accuracy of the final sample molarity, IGF were requested to multiplex and sequence adaptively, targeting 140 million reads per sample. This was important because of the high variability in expected proportion of microbial DNA per sample (preliminary data not shown), which would be further compounded by variations in sequencing depth.

IGF performed the demultiplexing according to their standard pipeline, which uses bcl2fastq and produces FastQC, MultiQC and FastQScreen reports.

Read binning and reallocation

A custom database was prepared for Kraken2 (Wood & Salzberg, 2014) comprising the National Center for Biotechnology Information (NCBI) RefSeq human genome, viruses and fungal sequences and

taxonomy, and Genome Taxonomy Database (GTDB) (Parks et al., 2018) bacterial and archaeal representative sequences and taxonomy (release 95). NCBI Genome Download (Blin, 2022) was used to obtaining the RefSeq sequences. Flextaxd was used to replace the NCBI bacterial and archaeal taxonomies with those from GTDB, and compile Kraken sequence and taxonomy files. The Kraken2 database was constructed according to the default settings (kmer size 35).

Untrimmed reads were initially binned with Kraken 2 according to the default settings, with no confidence threshold.

Conifer (Silamikelis, 2021) was applied to summarise confidence scores for each non-human species within a single sample, and aid selection of a reasonable confidence threshold for Kraken 2.

Reads were reallocated with Bracken 2 (Lu et al., 2017) at species level. Reads were summarised by organism group (human, bacteria, fungal, viral and archaeal) and proportions presented by sample group. The relationship between extracted sample DNA concentration and organism groups was explored graphically.

It is known that there is widespread contamination of bacterial reference genome sequences with human genome sequences (Breitwieser et al., 2019; Steinegger & Salzberg, 2020; Merchant, Wood & Salzberg, 2014; Kryukov & Imanishi, 2016). These contaminant sequences are frequently poorly represented in the human reference genome, and therefore less likely to be filtered out by host-removal processes.

In order to identify species containing contaminant sequences, species whose abundances had a strong positive linear correlation with human abundances were identified by generalised linear regression. Kraken read counts both with and without a confidence threshold were used, since the lack of confidence threshold may increase sensitivity. Those with adjusted R^2 greater than 0.1 and a positive correlation co-efficient were removed, and the reads reallocated to *Homo sapiens*.

These species were cross-checked with a list of reference genomes known to contain human sequences (Breitwieser et al., 2019), both with and without direct reference human genome matches.

Contaminant detection

Sample-species count matrices were generated for each organism group (bacteria, fungi, viruses and archaea) and converted to relative abundance (RA) matrices. Decontam (Davis et al., 2018) was applied to identify potential contaminants using the frequency-based approach. In this process, taxa whose abundance increases with diminishing sample biomass are considered as potential contaminants.

Multiple approaches were considered. The input DNA concentration was taken either to be: (a) the measured DNA concentration in the swab extraction multiplied by the proportion of reads allocated to the relevant group, or (b) the proportion of reads allocated to the relevant group. The first scenario accounts for contaminants from the swab and to the point of DNA extraction. The second scenario accounts for contaminants from the point of making equimolar solutions after DNA extraction. Potential contaminants were identified at a threshold $p \leq 0.05$.

All organisms which ever achieved relative abundance over 10% were manually reviewed as potential contaminants. A list of known proteobacterial genera contributing common laboratory contaminants was also included (Salter et al., 2014), excluding genera which include known constituents of the oral flora.

Indexing reagent contamination was sought by regressing species relative abundance (within the whole sample) against the P5 and P7 index tags. Thresholds were reviewed manually.

Contaminant identification methods were compared through weighted Euler diagrams. The proportion of a sample's reads within a group contributed by potential contaminants was plotted against the estimated input DNA concentration and human DNA proportion.

Human genome alignment

Sample reads were aligned against the human genome (GRCh37) using Bowtie 2 with maximum insert size of 1000 and other default settings. Read pairs for which neither partner was aligned were extracted with samtools.

GC content bias estimation

GC content was measured for each sample's reads which aligned to the first million bases of chromosome 1 and plotted against the number of PCR cycles to check for potential amplification bias.

Coverage estimation

Nonpareil was used to generate curves for unaligned reads from each sample, and estimate effective metagenome coverage.

Taxonomic survey

The phyla with the greatest median relative abundance were summarised.

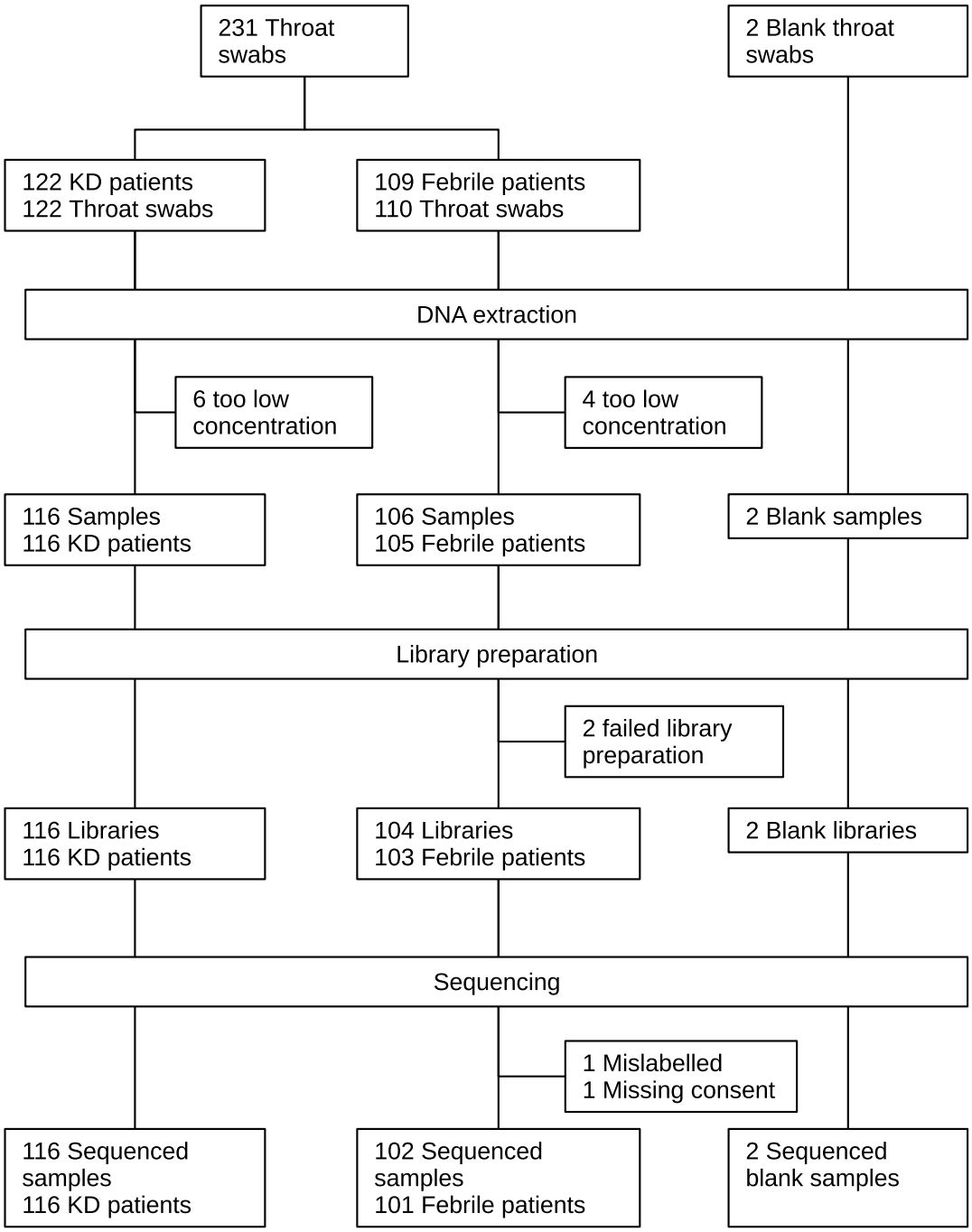

Figure 3 Flow diagram from throat swab samples to metagenomic sequencing

Results

Samples and data

231 throat swabs and 2 blank control throat swabs underwent DNA extraction (Figure 3). Six swabs provided insufficient DNA or failed library preparation. Within the febrile group, one febrile sample was mislabelled and corresponded to an unknown patient, one sample had missing consent, and two swabs were taken 10 days apart from the same febrile patient. Figure 3 displays a flowchart of the process from DNA extraction to sequencing. DNA sequencing was undertaken over 12 months,

interrupted due to the COVID pandemic, and ultimately sequencing data was available for 116 KD patients and 101 Febrile patients.

Demographic and clinical data is shown in Table 1.

	Febrile	**KD**	
	USA (N=101)	**UK** (N=19)	**USA** (N=97)
Sex			
Female	45 (44.6%)	6 (31.6%)	33 (34.0%)
Male	56 (55.4%)	13 (68.4%)	64 (66.0%)
Age (y)			
Mean (SD)	4.19 (3.00)	4.04 (4.70)	4.19 (3.17)
Median [Min, Max]	3.60 [0.100, 14.0]	2.80 [0.300, 17.5]	3.10 [0.300, 15.3]
Ethnicity			
American Indian/Alaska Native	1 (1.0%)		1 (1.0%)
Asian	9 (8.9%)	1 (5.3%)	8 (8.2%)
Black/African American	3 (3.0%)	7 (36.8%)	2 (2.1%)
Caucasian	22 (21.8%)	7 (36.8%)	23 (23.7%)
Hispanic	33 (32.7%)		39 (40.2%)
Multiple	20 (19.8%)	2 (10.5%)	21 (21.6%)
Unknown	8 (7.9%)	0 (0%)	
Other		2 (10.5%)	2 (2.1%)
Missing	5 (5.0%)		1 (1.0%)
Day of illness			
Mean (SD)	6.33 (3.53)	7.06 (3.83)	6.53 (4.03)
Median [Min, Max]	6.00 [1.00, 27.0]	7.00 [2.00, 14.0]	5.00 [2.00, 25.0]
Missing	5 (5.0%)	2 (10.5%)	0 (0%)
CRP (mg/dL)			
Mean (SD)	57.7 (82.5)	119 (91.3)	93.7 (83.1)
Median [Min, Max]	31.0 [5.00, 524]	90.0 [5.00, 301]	60.0 [5.00, 363]
Missing	7 (6.9%)		1 (1.0%)
WBC (10^9/L)			
Mean (SD)	10.2 (6.92)	17.1 (6.31)	13.8 (5.17)
Median [Min, Max]	9.05 [2.60, 48.9]	16.0 [7.90, 28.8]	12.7 [5.00, 32.4]
Missing	5 (5.0%)		
PMNs (10^9/L)			
Mean (SD)	5.32 (5.88)	11.1 (4.82)	8.07 (4.15)
Median [Min, Max]	3.89 [0.114, 45.0]	11.6 [2.00, 18.2]	7.32 [0.700, 24.9]
Missing	5 (5.0%)	1 (5.3%)	
Antibiotics before sampling			
Yes	45 (44.6%)		52 (53.6%)
Coronary Artery Aneurysm			
Yes	21 (21.6%)		3 (15.7%)

Table 1 Demographic and clinical details of patients with available metagenomic data

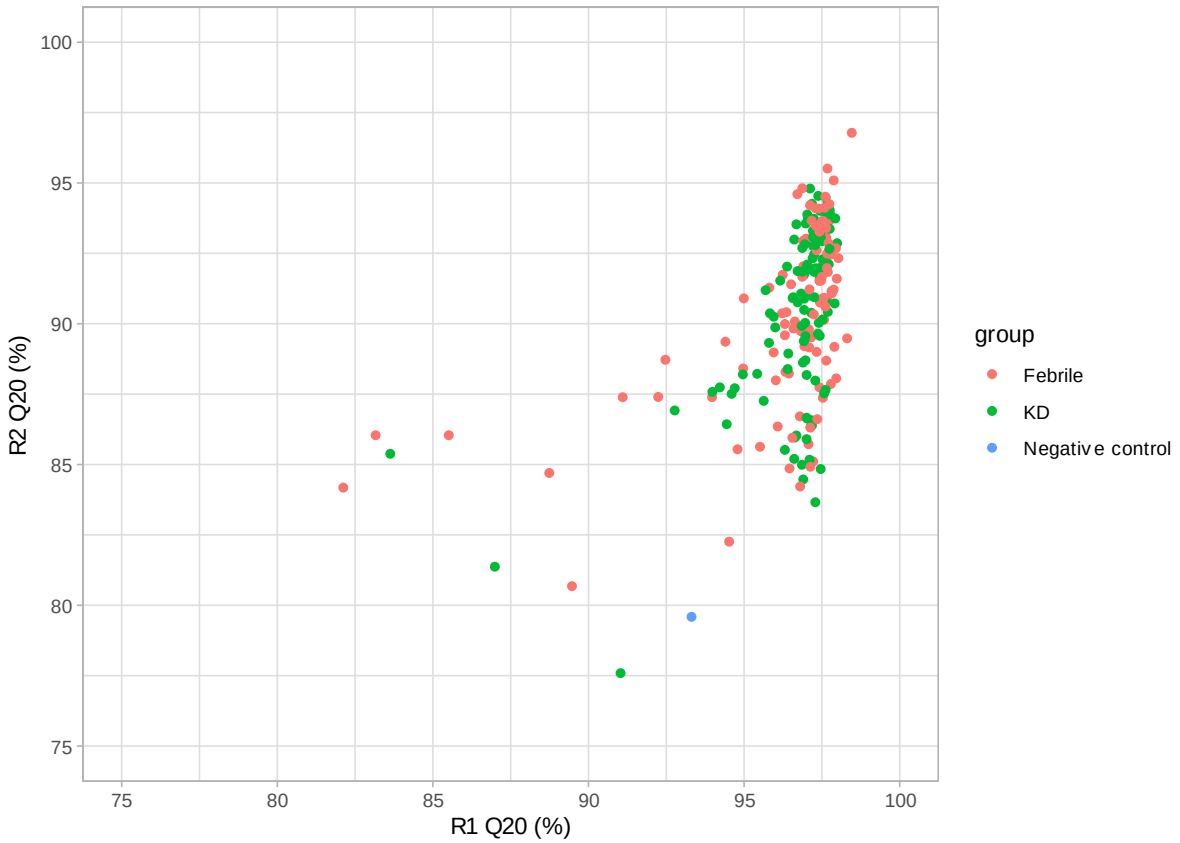

Figure 4 Proportion of bases exceeding Q20 quality score in read 1 (R1) and read 2 (R2)

A total of 37.3 billion read pairs were obtained, with a mean of 171 million read pairs per non-control sample and median of 158 million (interquartile range [IQR] 149-180 million). Eight of 210 non-control samples provided fewer than 135 million read pairs.

The proportion of bases with a quality score of 20 and above (first 100 000 reads) exceeded 95% in read one and 85% in read two for the majority of samples (Figure 4). Read two is expected to have lower quality.

Based on reads aligned to the first million bases of chromosome 1, there was no association between PCR cycles and GC proportion (p=0.93) with a median of 47.7% for each of 8, 10 and 12 cycles.

Read binning and reallocation

Overall, Kraken classified 96% of reads: 74.8% of reads were classified as human, 20.9% of as bacterial, 0.2% as fungal, 0.02% as viral (including bacteriophage) and 7 in 100 000 as archaeal.

The median proportion of human reads was higher in the KD group, and correspondingly the proportion of bacterial reads lower by more than 50%.

Conifer was applied to a single sample to explore the confidence with which reads were assigned to each microbial species. The confidence score for each read is calculated as the proportion of

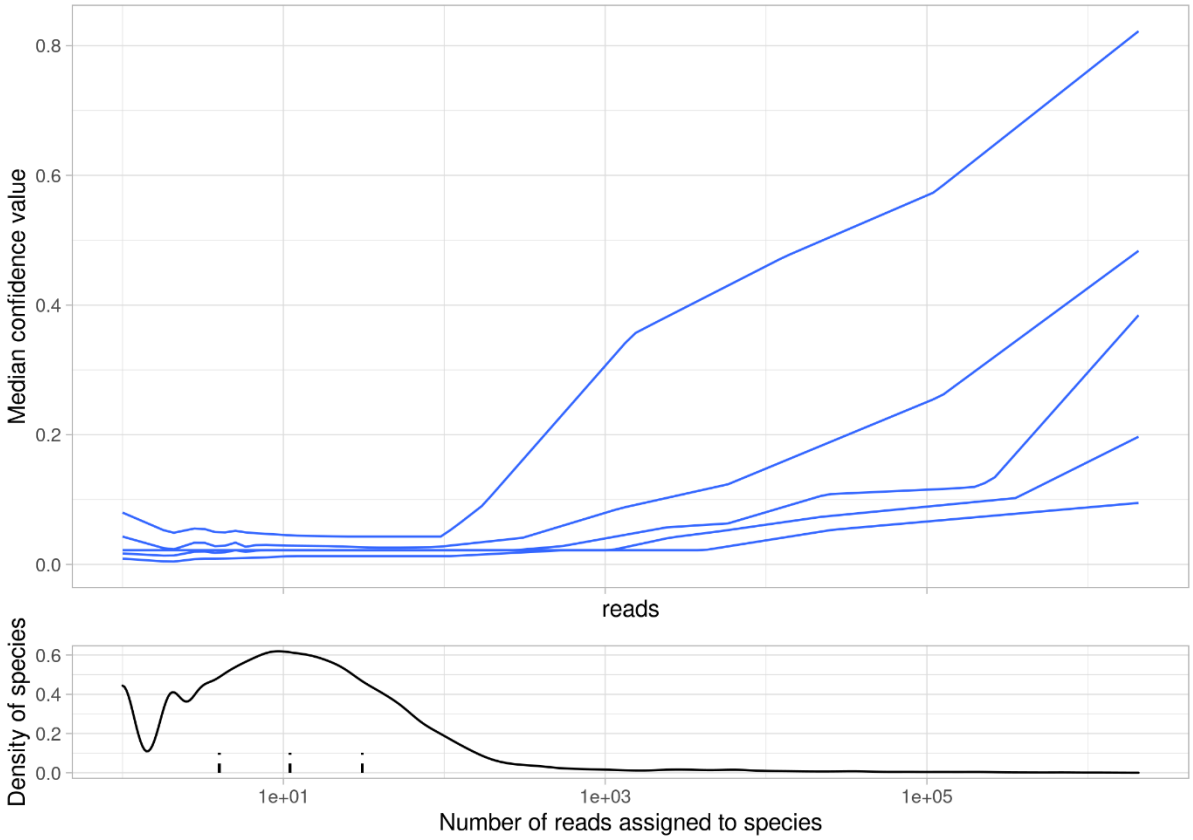

Figure 5 Analysis of species-level read binning within a single sample. The top panel shows a quantile regression plot (5th, 25th, 50th, 75th and 95th percentiles) of median read binning confidence value by species read count. The lower panel shows the distribution of species by read count, with vertical marks indicating 25th, 50th and 75th percentiles.

assignable *k*-mers assigned to the selected node for the read. The median confidence score per non-human read was 0.19. A confidence value of 0.05 was exceeded by 84% of reads.

Only 1.4% of the 33 328 non-human taxa identified exceed 0.01% of identified reads, though 96% of non-human assigned reads belong to these taxa. The median of median confidence scores only begins to rise appreciably at the 1 000-read threshold – reaching 0.035 for taxa with between 950 and 1 050 reads (Figure 5). Confidence scores are expected to be lowered when not applying any base quality threshold with Kraken, since erroneous nucleotides may render *k*-mers unidentifiable.

The same sample was rerun with a confidence threshold of 0.05 (corresponding to minimum 12 of 232 possible 35-mers per read pair assigned to the reported node) and quality threshold of 20 (corresponding to 1% chance of error – if every base in a 35-mer has this quality score, there would be a 30% chance of one or more erroneous bases). 94.5% of bases exceed this quality score in this sample (first 100 000 reads); low quality bases are likely to cluster at the beginning and end of reads. From the first 100 000 read pairs, a median of 168 35-mers were provided (IQR 115-229) by each pair. No 35-mers were available from 3 845 (4%) of read pairs.

Unclassified reads and those assigned to Opisthokonta or Cellular Organisms increased from 3.8 to 9.4%. Human reads fell from 71.3 to 67.9% and bacterial reads from 24.8 to 22.7%. Fungal reads fell by 62%, viral reads by 46% and archaeal reads by 91%. The number of taxa fell to 27 559, with 0.8% exceeding 0.01% of identified reads, and an increased 99% of assigned reads to these taxa.

The proportion of bacterial reads assigned at species level fell from 88 to 70%, with genus level assignments rising from 11 to 27%. Assignments above genus level increased from 1.8 to 3.5%.

When reads were reallocated with Bracken, the total number of species identified dropped, with the largest reduction among bacterial and archaeal species: 2 933 of 16 174 bacterial species remained, 17 of 378 archaeal species, 155 of 324 fungi and 32 of 68 viruses. The preferential removal of low abundance species (<1000 reads, ~2 in 100 000 bacterial reads) prompted selection of this approach (Figure 6).

Human-contaminated references

Ideally, the microbial genome sequences used in database construction would be accurate and complete. However, the known contamination of reference bacterial genomes with human sequences

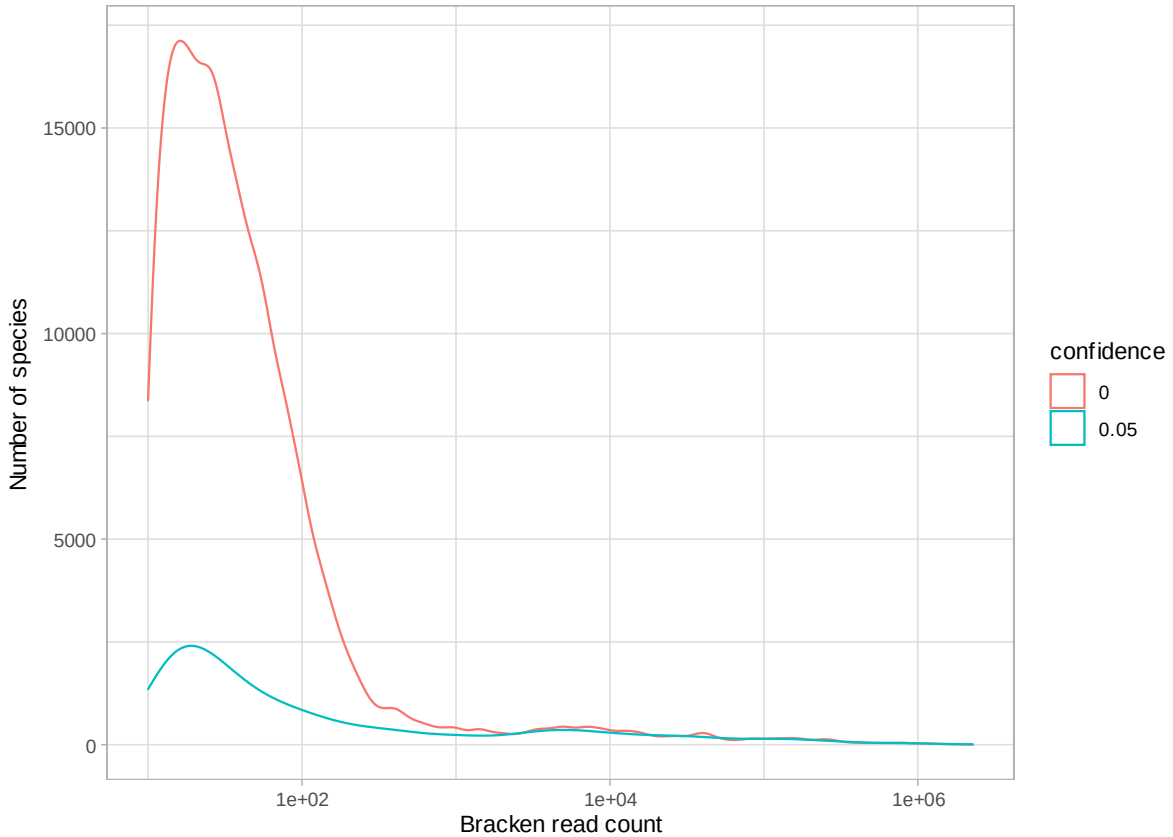

Figure 6 Exploration of the effects on species read count distributions by applying a Kraken confidence threshold. Bracken-reallocated read counts are shown, to ensure the reduced proportion of reads allocated at species level does not drive the difference.

(Breitwieser et al., 2019), coupled with the high and variable proportion of human DNA in these samples raised a severe risk of spurious signal if human reads could be mislabelled as bacterial.

It is implausible that a true pharyngeal organism would increase in relative abundance within the sample as the proportion of human reads increases (since the total proportion of microbial reads will diminish). In the case that specific human sequences can be misclassified as belonging to one or more species, those species should increase linearly in abundance with human sequences.

Using a generalised linear model (to protect from outlier influence) I identified 90 species with significant positive correlation with human read proportion ($R^2 \geq 0.1$) from either Kraken run (data not shown). Of these, 40 organisms have reference genomes identified to be contaminated with human sequences by Breitwieser *et al.* (2019). These species were flagged for removal from Bracken-adjusted read counts and reads were reallocated to *Homo sapiens*.

Reallocation

Bracken reallocated reads to species level. The species reallocated to *Homo sapiens* accounted for a median of 0.1% of reads per sample (range 0.02 to 0.4%), and 1% of bacterial reads (range 0.02 to 86%).

84% of bacterial species had some reads reallocated. The median proportionate increase in reads by bacterial species was 23% (IQR 4-63%). Read allocations by group are shown in Table 2.

Contaminant detection

Extracted DNA concentration was available for 212 samples, recorded as below the limit of detection for 5 samples, and missing for three. Values below the limit of detection were replaced with the minimum value of 2 ng mL^{-1}. The effective concentration of bacterial DNA was calculated by multiplying the sample DNA concentration by the proportion of reads allocated to bacteria.

Organism group	**Febrile**	**KD**
Homo sapiens	67.4% (45.9-84.3)	82.0% (59.2-89.3)
Bacteria	16.4% (5.81-36.0)	7.50% (1.85-25.4)
Fungi	51 (43-61)	49 (43-57)
Viruses	44 (20-150)	17 (9-36)
Archaea	3 (1-8)	2 (0.5-5)

Table 2 Summary of read allocations to top-level organism groups. Numbers shown are percentages or reads per 100 000, with interquartile range.

Since the abundance of contaminant DNA is unlikely to be related to the quantity of non-contaminant DNA transferred onto the swab and extracted in the laboratory, Decontam uses a linear model to detect features whose relative abundance increases with reducing DNA input (Davis et al., 2018).

Per-group sample-species matrices were processed by Decontam in frequency mode, with two approaches. Firstly, bacterial DNA proportion was used as DNA input, to account for contamination during library preparation, when samples are equimolar. Secondly, the estimated bacterial DNA concentration was used (both with and without previously identified contaminants removed), to account for contamination arising onto the swab or during extraction. In this way the possible two stages of contaminant introduction were considered.

Common proteobacterial laboratory contaminant genera were taken from Salter *et al.* (2014) excluding genera known to be part of the normal oral flora (Kingella, Pseudomonas and Acinetobacter) and also used to annotate potential contaminants.

Associations between relative abundance of each bacterial species (of total sample reads) and P5 and P7 tags were also assessed by generalised linear regression. A q-value threshold of 0.0005 was selected so as to identify organisms with strong tag-related patterns by visual inspection of matrix plots (data not shown).

Dominant residual contaminant organisms in some samples will drive the apparent relative abundance of other organisms down (e.g. a contaminant with 10% RA will reduce other organisms' RAs by the same proportion). Fifty-four species were ever present at 10% or greater RA (1-13 instances) following removal of other contaminants. Likely contaminants in this list comprised only *Sphingobacterium multivorum*, *Bifidobacterium breve* and *Syntrophomonas methylbutyratica*. These organisms were also annotated as potential contaminants.

Contaminant genera were reviewed for potential false positives, since the human DNA proportion potentially reflects throat inflammation and may therefore be associated with pathological changes in the microbiome. The 87 of 1 406 genera with 10 or more species identified as contaminants were manually reviewed. Five genera were identified as part of the normal pharyngeal flora (Table 3). Most species (240/259) were redundantly identified by Decontam using bacterial proportion as the biomass

Genus	Frequent contaminants	Number of species identified as contaminants	Median proportion of genus relative abundance	R^2 total RA (contaminant RA)
Acinetobacter	Yes	49 of 130	85%	0.43 (0.51)
Pseudomonas	Yes	99 of 444	65%	0.27 (0.44)
Corynebacterium	Yes	60 of 146	6%	0.11 (0.39)
Staphylococcus	No	31 of 68	99%	0.33 (0.37)
Moraxella	No	20 of 24	81%	0.22 (0.31)

Table 3 Known pharyngeal-resident genera with 10 or more species identified as contaminants by Decontam. For each genus and sample, the proportion of relative abundance (RA) classified as belonging to contaminant species is calculated, and the mean presented per genus. The R^2 value is shown for a simple linear regression model of genus RA (and genus contaminant species' RA) versus log10(bacterial read proportion).

measure. All relative abundance relationships with bacterial proportion were highly significant (linear regression of log(RA) against log(bacterial proportion); $p<0.001$), whether utilising the total genus abundance or only contaminant species.

Euler diagrams were generated to demonstrate the contribution of each method to identifying sets of contaminant species (Figure 7) using a variety of weights: species count, species median RA and species mean RA. The approach with Decontam using the microbial DNA proportion identifies the largest number of potential contaminants. The candidate genus approach adds the next largest number, followed by the concentration approach. With mean abundance weighting, the dominant approach remains Decontam proportion though there is more balance. The Decontam concentration approach is dominated by the Decontam proportion approach when considering median abundance weighting. The genus approach adds considerable contaminant abundance when the median approach is considered, and other means contribute little. In total 4 707 potential contaminant species were identified.

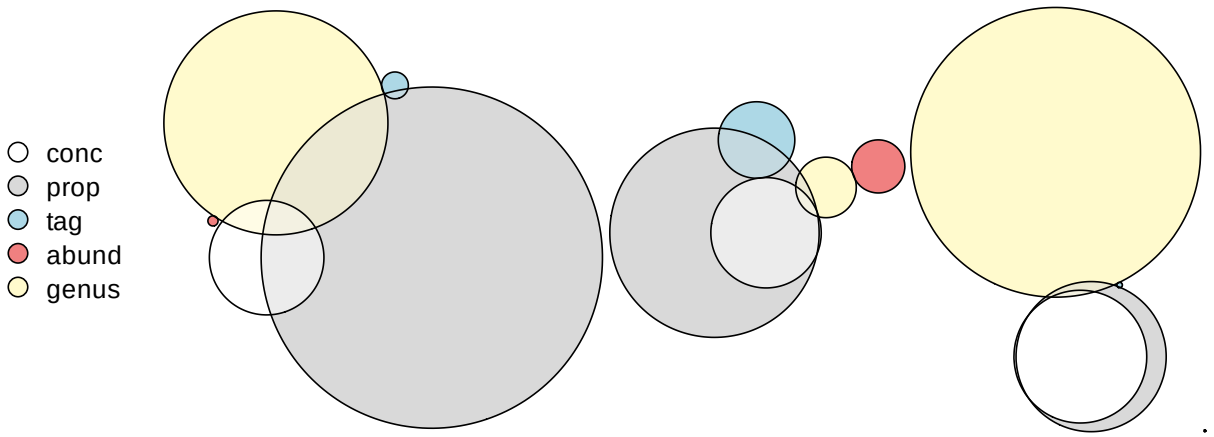

Figure 7 Three Euler diagrams summarising the median amount of contaminant identification within samples using different approaches. The left block shows number of species; the middle block shows mean relative abundance proportion; the right block shows median relative abundance proportion (conc=Decontam concentration/biomass approach; prop=Decontam proportion approach; tag=Index tag reagent identified contaminants; abund=highly abundant organisms identified as likely contaminants; genus=Abundant proteobacterial laboratory contaminant genera from Salter et al [2014]).

In the samples with the lowest bacterial DNA proportions (<2%), the proportion of reads contributed by potential contaminants can exceed 50%. Eighteen of 218 clinical samples exceed 25% bacterial contaminant proportion (14 KD and 4 febrile).

Bacterial contaminants contributed a median 0.05% of reads per sample (IQR 0.03-0.1%), with a small difference in favour of KD patients (median 0.04 vs. 0.06%). However, due to the lower bacterial proportion in KD samples, the proportion of bacterial reads attributed to contaminants is higher in KD patients (median 0.5 vs 0.3%).

Samples with absolute bacterial contaminant proportion (of total reads) ≥ 0.5% were inspected to ensure dominant contaminants were unlikely to represent usual flora. Fifteen samples were represented. Seven were dominated by obvious contaminants: two by Lactobacilli and Bifidobacteria; one by a Ruminococcus; another by Acinetobacteria and Enterobacteria; one by Stenotrophomonas; one by Pseudomonas and Cutibacterium; another by Mycobacteria. Three samples were dominated by Staphylococci and five by Moraxella.

Contaminants cannot be resolved with certainty, and some of the above may represent true Moraxella- or Staphylococcus-dominated microbiomes. Nonetheless, there is potential to re-include contaminants at later stages of analysis.

Potential contaminants were shared between 41 phyla (NCBI definitions) and represented a mean of 7% of bacterial relative abundance. 95% of mean contaminant relative abundance was provided by four phyla: Firmicutes (43%), Proteobacteria (33%), Actinobacteriota (16%), and Bacteroidota (2%). Within these phyla, 6, 14, 6 and 2% of phylum relative abundance (respectively) was contributed by contaminant species.

The same four phyla contributed 95% of non-contaminant relative abundance, but the order differed: Firmicutes (48%), Actinobacteriota (19%), Bacteroidota (16%) and Proteobacteria (14%).

If contaminants and non-contaminants are well separated, no correlation would be expected between the paired relative abundances of phyla within contaminants and non-contaminants. Although all four phyla demonstrate positive correlations between contaminant and non-contaminant abundances, adjusted R^2 values are 0.07 or below. This suggests that contaminants and non-contaminants may be imperfectly separated.

With contaminants flagged, the depth of metagenomic sequencing per sample can be summarised. Median non-contaminant read depth is 17.2 (IQR 4.7-49.3) million for bacteria, 1 798 (IQR 1 010-4 053) for fungi, 12 311 (IQR 4 513-46 469) for viruses and 172 (IQR 34-492) for archaea. This is broken down by disease group in Table 4, revealing the difference in microbial read depths between KD and

Organism group	Febrile	KD
Bacteria	28 667 (9 860-60 324)	10 619 (2 342-37 962)
Fungus	2.4 (1.4-5.4)	1.5 (0.8-2.8)
Virus	20.6 (9-104.6)	7.5 (2.8-21.4)
Archaea	0.3 (0.1-0.7)	0.1 (0.0-0.4)

Table 4 Median and interquartile range of non-contaminant read counts (×1 000) by organism group and sample group.

febrile patients (due to the greater human proportion in the KD group). The most striking difference is for bacteria, where the median depth was nearly three times as high for febrile patients.

Residual effects of contamination

Potential residual effects from unidentified contaminants could be considered by summing the total proportion of non-contaminant bacterial reads contributed by the most abundant genera overall – excluding those genera known to colonise skin. Thus, for each sample the total relative abundance of Prevotella, Veillonella, Rothia, Neisseria, Pauljensenia, Haemophilus, Gemella and Granulicatella was summarised. A transition was evident at around 3% contamination with a drop in the proportion of reads from dominant genera from median 57% (IQR 51-64%) to 41% (29-51%, p<0.001 by Wilcoxon test; Figure 8).

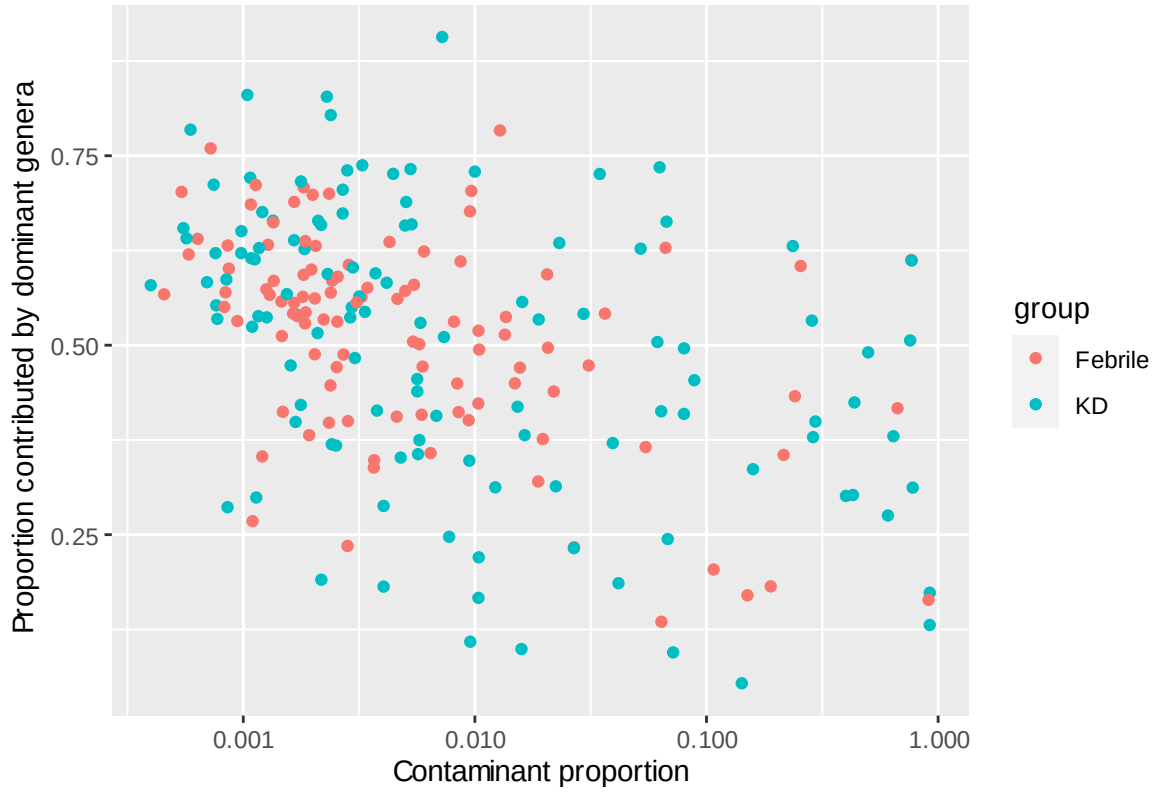

Figure 8 The proportion of each sample's non-contaminant relative abundance contributed by the eight dominant genera.

This suggests the possibility that above 3% identified contaminant proportion there could still be significant unidentified contaminants with the potential to falsely reduce the RA of non-contaminant species. Given that both the potential effect and the disparity in proportions of KD vs febrile samples is most apparent above contaminant proportion of 10% (15% of KD samples vs. 9% of febrile samples), this gives a potential group to exclude in sensitivity analyses.

Drivers of variation in bacteria DNA proportion

The variable balance of human and microbial DNA between samples and groups could be explained by variation in either or both of human DNA and microbial DNA load on throat swabs.

Considering only non-contaminants, the estimated concentrations of bacterial and human DNA in samples was explored. Bacterial and human DNA were present at median concentrations of 2.7 (IQR 0.8-6.5) and 17 (IQR 7.1-51) ng mL^{-1} respectively. Median bacterial DNA concentrations were similar between KD and Febrile groups at 2.2 and 2.8 ng mL^{-1} respectively (p=0.33, Wilcoxon rank sum). In contrast human DNA concentrations were different at 24.9 and 11.2 ng mL^{-1} respectively (p=0.001). There was no relationship between the human and bacterial DNA concentrations (p=0.56 by linear regression).

The proportion of a sample's non-contaminant reads represented by bacteria was associated with both the estimated human and bacterial concentrations in multivariable linear regression (adjusted R^2 0.42). The bacterial proportion fell as the human concentration rose (-0.11% mL ng^{-1}) and rose as the bacterial concentration rose (+1.0% mL ng^{-1}). Sequential analysis of variance shows the bacterial DNA concentration explains more of the variance (25 vs 17%) overall.

Negative control samples

The two negative control samples comprised few reads had very few reads (6.9 million and 735 000 in sample A and B respectively). The proportion of human reads was typical at 82 and 90% of reads respectively, and the remainder mostly bacterial. In sample A, the most abundant bacterial species was classified as a contaminant (*Syntrophomonas methylbutarica*) but subsequent organisms were part of the oral flora. In sample B, most abundant bacterial species were part of the oral flora.

Human genome alignment

The proportion of read pairs aligning at least once to the human genome approached the proportion assigned as human by Kraken 2 (median difference -4.2%, range -1 to -34%).

Read pairs with neither pair aligned were extracted ("unaligned reads") and used in further analyses.

Coverage estimation

Nonpareil 3 (Rodriguez-R et al., 2018) is a database-independent *k*-mer based tool which estimates metagenomic coverage. Coverage estimates are abundance weighted, therefore relate to coverage of organisms, not the range of species represented.

Nonpareil proceeds through two stages. The first is redundancy estimation, in which sequences are subsampled and *k*-mer redundancy measured each time. Abundance-weighted average coverage is estimated at different depths by fitting a sigmoidal model to the data.

Nonpareil was run on the first pair of each sample's reads, with alignment length of 100 (maximum). Reads were first trimmed with trimmomatic (parameters LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50). 103 of 116 samples produced warnings that the curves "reached near saturation and coverage estimations could be unreliable" – the only step would be to increase the minimum similarity value beyond the tested 0.95. Fitted Nonpareil curves are shown in Figure 9. The median coverage of samples was estimated to be 96% (IQR 92-99%), with 129 of 174 samples having coverage of 90% or above. One KD sample (1510181) has a dramatically low estimated average coverage.

Within-sample diversity estimates were a median of 16.5 (IQR 15.6-17.0, range 10.8-27.9). The medians differed slightly by group, with slightly greater diversity in the febrile group (KD 16.2, Febrile

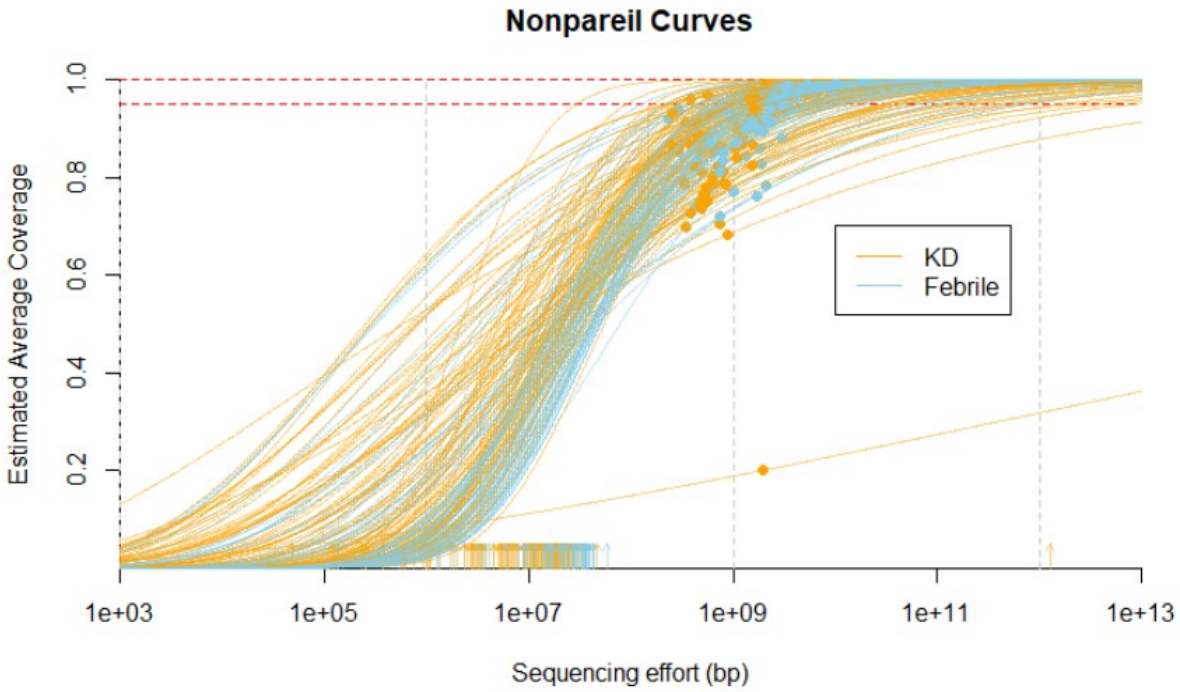

Figure 9 Nonpareil curves (Rodriguez-R et al., 2018) coloured by sample group. Average coverage (estimated by k-mer redundancy) is shown against sequenced bases, with the points representing the actual sequence available for the sample.

16.8, p<0.001 Kruskal-Wallis). The outlying low-coverage sample (1510181) has outlying high diversity and is unusual in other aspects. 23% of reads are assigned directly to the root node (next largest 8%; median 0.2%). Once human-associated species have been reallocated, 99.9% of reads are human – only two other samples exceed this. Read quality is low: fifth lowest for the first mate (14% of bases Q20 or below) and second lowest for the second mate (19% of bases Q20 or below). Only 72 364 non-contaminant bacterial reads are allocated. This sample is excluded from further analyses.

Dominant phyla

Firmicutes are the dominant phylum both by median and mean relative abundance (48 and 47% respectively). However, there is a very wide spread of distribution of RA, with IQR 35-59% and range 0.7-93%.

Figure 10 shows the top 10 phyla by median abundance with other phyla grouped. Phyla from 5th to 10th most abundant represent likely contaminants, since Campylobacterota are gut organisms and the others predominantly environmental. Median relative abundance of these five phyla together is 0.5% (IQR 0.2-1.2%; maximum 8.8%). Adding the remaining minority phyla does not materially change these estimates. Assuming these represent residual contamination, the impact upon relative abundance of non-contaminant organisms should be minimal.

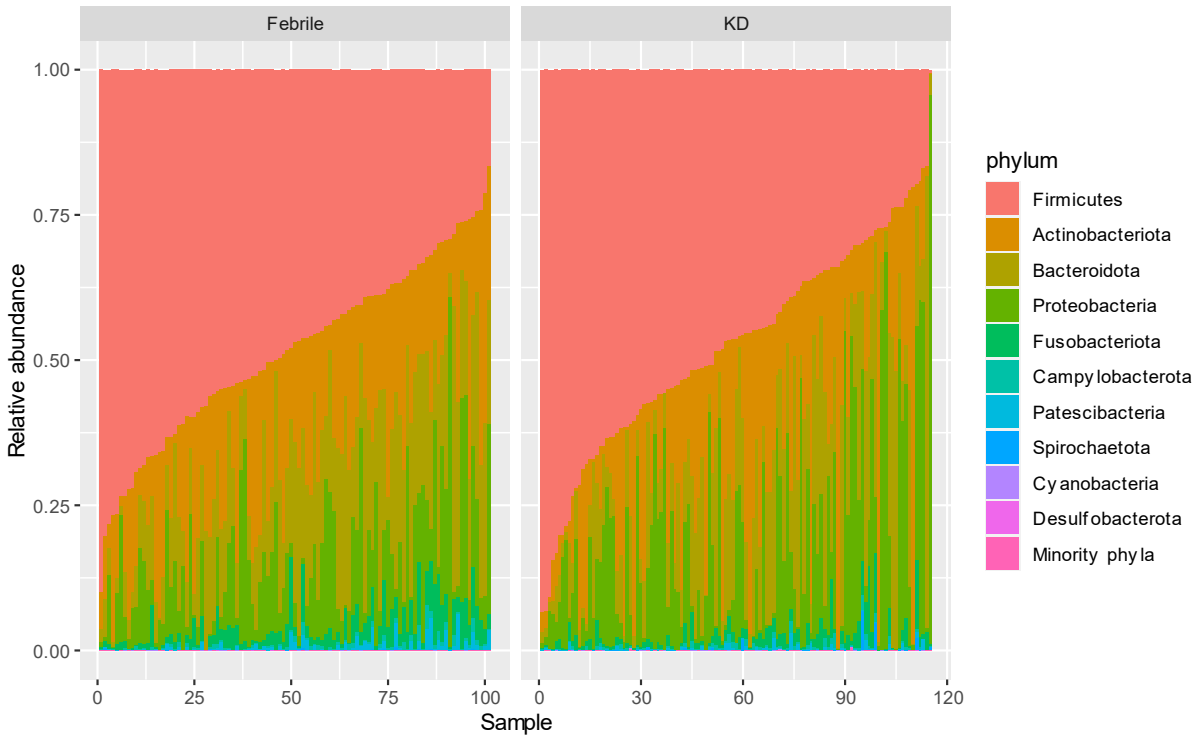

Figure 10 Distribution of relative abundances of the top 10 phyla across all samples ordered by decreasing abundance of firmicutes.

Discussion

In this chapter I have described the large quantity of shotgun metagenomic data obtained from our cohort of over 200 individuals, assessing quality and taxonomic classifications. Critically, I have undertaken extensive analyses to identify spurious identifications due to contaminated reference sequences, and the small amount of likely laboratory contamination.

Quality control

The adaptive method of multiplexing undertaken by the sequencing facility has ensured that the sampling depth was sufficiently evenly distributed between samples, and the vast majority of samples exceeded the average depth targeted.

Sequencing quality is adequate for the majority of samples. Downstream tools including Kraken 2, Bowtie 2 and MEGAHIT are quality aware, and adapter removal was included as part of the in-house demultiplexing pipeline. For this reason, I selected not to trim reads for this analysis.

PCR is a helpful step to facilitate successful library preparation, especially for low concentration samples. However, PCR introduces a risk of guanine-cytosine (GC) bias. With 8-12 cycles of PCR per sample, dependent on DNA concentration, it was important to ascertain whether this introduced any GC bias.

Since GC content varies throughout the human genome in a predictable manner, I opted to measure bias within reads aligning to a region of the human genome. Selecting bulk bacterial (or non-human) reads could confuse any signal due to inter-sample variation in bacterial average GC content. The GC content was remarkably consistent between samples with 8, 10 and 12 cycles of PCR. This is consistent with the findings of Chafee, Maignien & Simmons (2015) who showed only a marginal reduction in GC content (59.6 to 58.6%) when increasing from 8 to 24 cycles of PCR.

Taxonomic identification

Various methods exist for measuring the taxonomic profile of samples by shotgun metagenomics. These can broadly be classified into those based on sequence alignment, *k*-mer composition or marker-gene identification. *K*-mer based approaches, exemplified by Kraken and Centrifuge, are computationally efficient (though memory hungry) and usually sensitive. Marker gene-based approaches, like MetaPhlAn, are more specific, though the trade-off is a higher coverage threshold to detect organisms.

In work conducted during this fellowship, I reanalysed data from a study investigating the effects of variable sequencing depth and host DNA contribution on taxonomic profiling (Pereira-Marques et al., 2019). This study used a known mixture of bacterial DNA spanning a wide range of concentrations,

and varying proportions of murine DNA. The authors applied MetaPhlAn 2 and found that the lowest abundance organisms became undetectable when host DNA content was high (90-99%). Motivated by the requirements for this study, I reanalysed their data with Kraken and found that all organisms could be detected in all samples, with a stable taxonomic profile (McArdle & Kaforou, 2020; see Appendix B). I noted that likely contaminating organisms were detected at greater relative abundances (among bacteria) in the samples with the highest host DNA content, and applied Decontam to successfully identify a large proportion of these organisms.

Since any microbial trigger of KD may not be present in high abundance (or may only be residually present following recent active infection), sensitivity is a strong requirement for this study. Kraken 2 (Wood, Lu & Langmead, 2019) builds upon Kraken 1, with *k*-mer minimisers reducing the database size. It has reduced memory usage (compared to Kraken 1), allows custom databases and provides control over base-quality and identification confidence thresholds. Additionally, with Bracken 2 (Lu et al., 2017), reads can be reallocated from higher taxonomic levels to provide more accurate relative abundances and the most appropriate denominator.

KrakenUniq extends the approach of Kraken by counting unique *k*-mers identified per taxon, which provides additional confidence in identifications (Breitwieser, Baker & Salzberg, 2018). However, it is built around the approach of Kraken 1, requiring greater amounts of random-access memory (RAM), and without confidence and base-quality thresholds. Usefully, *k*-mer minimiser counting has been incorporated into Kraken 2.

Ganon (Piro et al., 2020) allows for continuous, quick updating of sequence databases, however in my tests the memory required exceeded 200 Gb and would not be suitable for high-throughput operation on Imperial's High Performance Computing (HPC) infrastructure.

Kraken 2 has favourable results in recent benchmarks on both short reads and assembled contigs (Meyer et al., 2021). It was the fastest read binner, and the top performer for contigs. At the time of writing in the LEMMI continuous benchmark of metagenomic classifiers (Seppey, Manni & Zdobnov, 2020) Kraken 2 with Bracken obtains the highest score for species detection (>100 reads) and for relative abundance of organisms. For low abundance detections and read binning accuracy, Ganon and MetaCache overtake. However, once computational resources are weighted at all, Kraken 2 also leads for read binning accuracy.

Based on its efficiency, flexibility and favourable evaluations, I implemented Kraken 2 for read binning, with Bracken for reallocation. In line with the authors' recommendations, I found that a confidence threshold of 0.05 (selecting the lowest taxonomic level with 5% or more of a read's *k*-mers in a read,

or marking the read as unidentifiable if this is not possible) selectively reduced very low count identifications of a large number of likely spurious identifications. Additionally, I implemented a liberal quality threshold corresponding to 95% confidence of base identification.

With this configuration, over 90% of reads were classified below the level of cellular organisms or Opisthokonta, and over two-thirds of bacterial reads were assigned at species level.

Human genome contamination of reference genomes

A small set of studies from the last decade have identified and quantified the problem of contamination of reference genomes in public databases. Merchant, Wood & Salzberg (2014) showed that the draft genome assembly of *Bos taurus* contained multiple contigs of bacterial origin, and that a *Neisseria gonorrhoeae* genome contained cow and sheep sequences. Breitwieser *et al.* (2019) identified human contamination in 2 250 RefSeq bacterial genomes. Most recently, Steinegger & Salzberg (2020) reported more than 2 million sequences in GenBank with cross-kingdom contamination, including 51 thousand bacterial sequences, and nearly 100 thousand RefSeq bacterial sequences.

In designing this analysis, I had little awareness of this problem. However, in refining the approach I noted a *Marinomonas* species lost over 95% of its reads when implementing the Kraken 2 thresholds. This species had been moderately abundant when no Kraken thresholds were applied, and was associated with KD. Its abundance was strongly positively correlated with human abundance (r^2=0.15; $p<10^{-8}$).

This prompted me to identify species with strong positive correlations with human read proportions – something biologically unexpected given that the total proportion of bacterial reads declines as the human proportion rises. I identified 90 species with such suspicious associations – reassuringly 40 of them had previously been identified in the 2019 study.

To my knowledge, no other post-analytic mitigations against host-contaminated reference genomes has been demonstrated in the literature. The ideal approach would have been to clear the reference database of human contaminated sequences before construction. However, release 95 of GTDB post-dates the 2019 publication and therefore contamination in newer genomes would be missed. Conterminator (Steinegger & Salzberg, 2020) could be applied to identify cross-kingdom contamination in the database, however there was not sufficient time within the fellowship to apply this, rebuild the database and rerun analyses. Unfortunately, there are no plans for GTDB to identify and remove such contamination (Donovan Parks, personal communication). Work is underway to mitigate the problem at NCBI (Terence Murphy, personal communication).

Contaminant detection

Laboratory and reagent contamination of samples undergoing amplicon and metagenomic analyses, especially those with low biomass, has received increasing attention. A recent review (Eisenhofer et al., 2019) covered both laboratory/reagent contamination and cross-contamination, and highlighted the frequent omission of analyses to identify and account for contamination. They review instances where contamination is suspected to drive significant findings in published studies, including studies of the placental microbiome (where it is feasible that the placenta is sterile since blank controls are indistinguishable), and raise concern about a profusion of studies of low-biomass samples without appropriate controls. They make a series of recommendations, reproduced in Figure 11.

In our case, it is important to note that samples underwent extraction in disease groups, due to concerns by group members to minimise potential between-group cross-contamination. PCR and library preparation also occurred in systematic layouts on 96-well plates. For this reason, identification and mitigation of contamination is of extreme importance in this study.

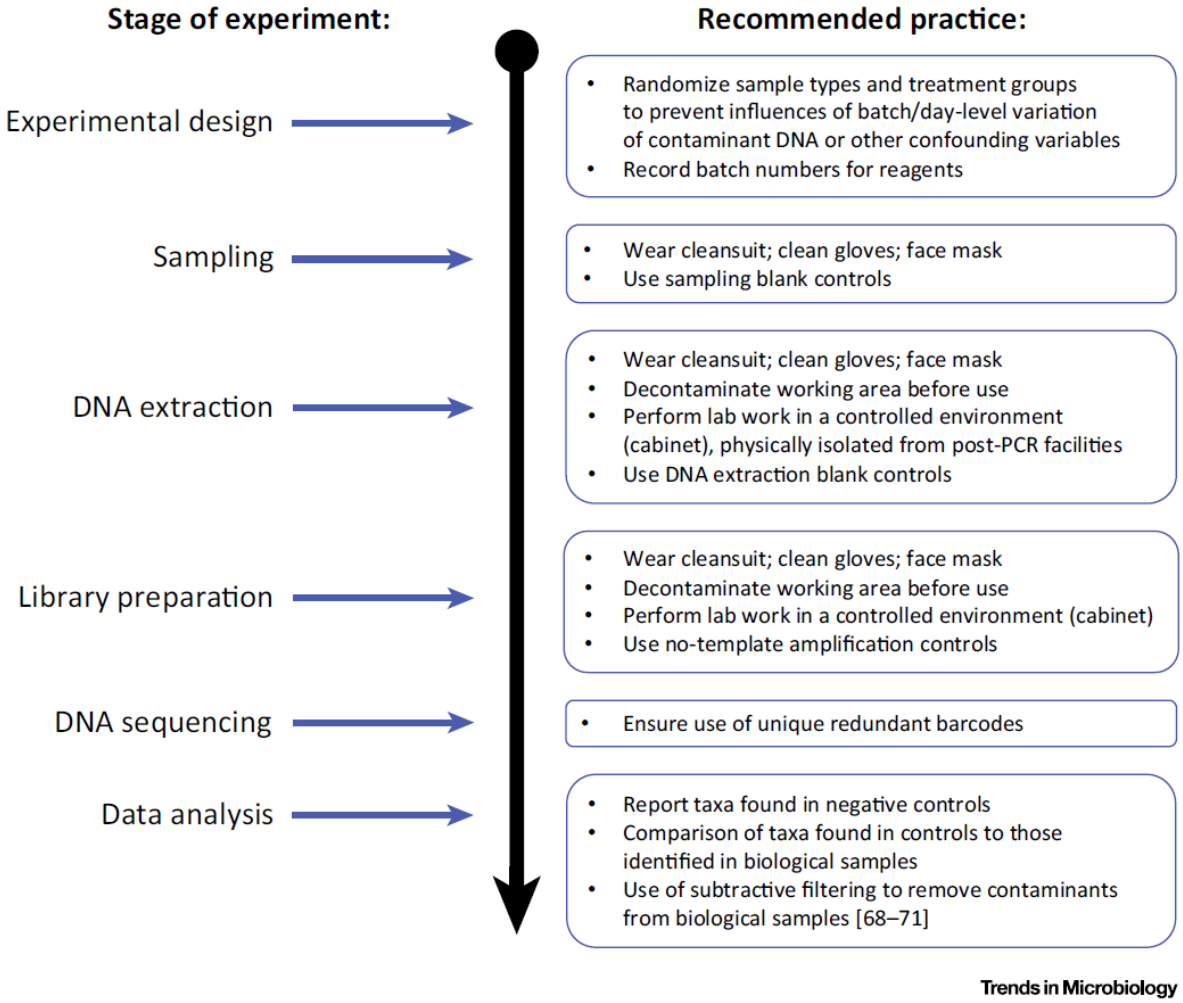

Figure 11 Recommendations for reducing the impact of contaminants in microbiome and metagenomic experiments, reproduced from Eisenhofer et al. (2019)

Karstens *et al.* (2019) benchmarked four computational approaches to identifying contaminants in 16S amplicon experiments, using a mock microbial community across four orders of magnitude of dilution. The simplest were filters based on either presence in a negative control or abundance below specified thresholds. Two published tools were also tested.

SourceTracker (Knights et al., 2011) utilises community profiles from known sources (whether internal controls, or external samples) to predict the probability of taxonomic units arising from defined sources. Decontam (Davis et al., 2018) uses one or both of a frequency (*i.e.* abundance) and prevalence approach. The frequency approach identifies taxonomic units whose abundance increases with reducing sample biomass. The prevalence approach requires negative controls, and identifies taxonomic units whose prevalence is higher in negative controls than samples. In the benchmark, SourceTracker performed well when source environments were well defined. However, with poor definition, performance was poor. Decontam, tested in frequency mode, did not misclassify mock community members, whilst correctly classifying 70-90% of contaminant Operational Taxonomic Units (OTUs). A large proportion of contamination by relative abundance remained in the lowest concentration samples (1:729-1:6 561) at the standard threshold ($p \leq 0.1$). However, up to 1:27 dilution, the residual contaminant proportion was less than 5%.

In our study, contamination must be considered for these reasons:

1. **Contaminants may contribute a high proportion of bacterial reads in some samples**. The bacterial biomass in our samples varies across four to six orders of magnitude, whether one considers the estimated concentration of extracted bacterial DNA (2 pg mL-1 to 113 ng mL-1) or the proportion of each sample which is contributed by bacterial DNA (range 0.08 to 87%).
2. **Contaminants may be spuriously associated with KD**. Bacterial biomass (by proportion) is strongly associated with group (median 8 vs 19% for KD and febrile respectively), thus contamination arising following equimolar dilution of extracted DNA would be proportionately greater for KD samples. Additionally, the systematic extraction and library preparation renders samples vulnerable to contamination in a biased pattern whether by reagent batches, or environmental contamination varying in time and space.
3. **High abundance contamination will artificially lower the relative abundance of non-contaminant organisms and could mask true signal**. For example, if contaminants account for 50% of the bacterial reads in a sample, non-contaminant relative abundances will be reduced by 50%. This greater abundance of contaminants in KD samples could mask true signals from organisms associated with KD.

4. **A large number of contaminant organisms will reduce power.** Even were contamination randomly distributed between KD and febrile samples, the added statistical tests for these organisms will reduce power to detect true associations at a fixed false-discovery rate.

In addition to the encouraging results of Karstens *et al.* (2019), I obtained favourable results applying Decontam (in frequency mode) to the aforementioned host-associated mock community data (McArdle & Kaforou, 2020, as presented in Appendix B). Decontam was able to remove 61% of off-target species and 79% of off-target reads (92% in the lowest biomass sample). I therefore selected Decontam as a means of detecting contaminants. In our case, only the frequency method was applicable as only two negative control swabs were sequenced, and their taxonomic profiles were remarkably similar to human samples.

The likely scenario is that these reads mostly resulted from "index hopping", due to free index tags in these samples with minimal input DNA (Sinha et al., 2017). Sample A shared its indices with five other samples and sample B with six, and both shared an index tag. Because these negative controls will potentially have been enriched for free index tags, index hopping may be a greater problem for them. Nonetheless, on the HiSeq 4000, Sinha *et al.* (2017) measured 5-7% rates of index hopping with both negative controls and cDNA-containing samples.

Decontam was applied using two biomass measures. Contaminants were considered as those with p values ≤ 0.05, more stringent than the 0.1 default. Contaminants were also sought through association with specific index tags, by applying a shortlist of common proteobacterial laboratory contaminants, excluding those typical of pharyngeal flora and manual searching of high abundance organisms.

Together these methods identified over 4 000 potential contaminants, accounting for a small proportion of reads per sample, but a modest, and sometimes very high, proportion of bacterial reads.

Notably, contaminant reads are similar in number to those from likely human-contaminated reference genomes (46 vs. 42 million).

The lower total abundance of prevalent genera in samples with over 3% of bacterial reads assigned to contaminants suggests that residual contamination could still contribute a significant bulk in these samples. However, given the likely underlying cause of high contamination is throat inflammation, biological causes are also possible. The small contributions from phyla outside the top five likely represents residual contamination (median total relative abundance 0.5%).

There remains a possibility of falsely rejecting true organisms of the oropharynx, as well as incorrectly including contaminants. Caution was taken to avoid over-aggressive rejection by selecting a more

conservative threshold with Decontam ($p \leq 0.05$ rather than 0.1). Species which represent a mix of both true pharyngeal colonisers and laboratory contaminant pose the greatest challenge.

Within the Staphylococci, *S. epidermidis* is confidently identified as a contaminant ($p = 1 \times 10^{-8}$), whereas *S. aureus* despite sometimes occurring at high abundance, is identified with less confidence ($p = 0.002$). *S. epidermidis* is not known to colonise the oropharynx, and as a coloniser of skin, is a common laboratory contaminant. In contrast, *S. aureus* is known to colonise both moist skin and the oropharynx (Mertz et al., 2009; Sollid et al., 2014) – the lower confidence identification as a contaminant may signify this.

Steps to mitigate against incorrect assignment of contaminant status will be considered in further analyses.

Conclusions

This chapter described preparatory work to explore the quality and high-level composition of the sequencing data, and mitigate against computational artefacts and contamination.

This extensive work is necessary to reduce the chances of spurious associations with KD, prevent distortion of true signals, and increase statistical power. The lack of clear consensus and ready-made pipelines to achieve these goals meant much laborious and creative work was required. Using findings from the literature and my own published investigation, I have applied published tools to these data and generated some of my own solutions to recognised problems.

3 | Metagenomics – identifying species associated with Kawasaki Disease

Introduction

Having analysed sequencing quality and developed a process to mitigate against sample and reference database contamination, the primary analyses could be undertaken. These comprise descriptive analysis of the microbiomes, with exploration of microbiome variation between KD and controls, and the search for species associated with KD.

Objectives are:

1. Summarise microbiome profiles of bacteria, community-level metrics and top-level associations with KD
2. Search for statistically significant associations between KD and organism prevalence and abundance across taxonomic levels within bacteria

Methods

Top-level analyses

Highly abundant non-contaminant bacterial species were summarised across samples and described. Organisms with the greatest absolute spread of relative abundances (5-95th percentiles) were described. Diversity and evenness metrics were calculated using the vegan package.

Non-metric multidimensional scaling (MDS) and hierarchical clustering were used to visualise the overall similarity between KD and febrile samples. MDS is an approach to dimensionality reduction and is commonly applied in ecology and psychometrics (Groenen & van de Velden, 2004). Relationship with sequencing depth was investigated, to allow consideration of potential bias. Weighted UniFrac distances (Lozupone & Knight, 2005) based on logged relative abundances were used, with zero abundances replaced with half the species' minimum. Only species with maximum relative abundance 0.001 or greater were included. MDS was applied as implemented by vegan with two dimensions and 500 maximum iterations.

Analysis of covariance (ANCOVA) as implemented within the vegan package (Dixon, 2003) was used to test the significance of inter-group differences and estimate the proportion of variance explained by group and other covariates.

Search for bacteria associated with KD

Sample-species relative abundance matrices were loaded into a TreeSummarizedExperiment object in R, along with the relevant taxonomy. Non-species nodes' relative abundances were calculated as a sum of the child species. Zero counts were replaced with half the taxonomic units' minimum.

The standard modelling approach of MaAsLin 2 (Mallick et al., 2021) was re-implemented and applied across all nodes of the taxonomy with maximum abundance 0.1% or greater and present in 10% or more samples. This comprised a fixed-effects generalised linear model with outcome of logarithm of relative abundance, and group and other clinical features as covariates. Two sensitivity analyses were undertaken, either adding log(bacterial proportion) to the model, or excluding samples with greater than 10% contamination. These complementary approaches sought to mitigate against spurious signal from unidentified contaminants.

A separate dichotomised analysis was undertaken, with dynamic thresholding of OTUs. For each feature, potential thresholds were considered at any nadirs in the density curve (against log[RA]) and the 75th percentile. Any threshold which gave a positive proportion between 12.5 and 50% was considered as acceptable. The nadir threshold with positive proportion closest to 25% was selected. If no nadir thresholds were acceptable, then the 75th percentile was used. For each feature a binomial family generalised linear model was fitted modelling presence as a function of the same covariates as above, and using the same sensitivity analyses.

In both cases, treeclimbR was applied to p values for each predictor to select the taxonomic level at which to report values from each lineage, and apply false-discovery rate (FDR) control, presented as Q values. Sensitivity analyses underwent independent processing by treeclimbR, though Q values were also reported for each node selected by the primary analyses.

Associations with age were analysed separately, and models adjusted where non-linear relationships were encountered. The relative contributions of predictors to explaining variance was explored with the relaimpo package (Groemping & Matthias, 2021).

Additional analyses were undertaken with dichotomous data on antibiotic exposure for children recruited at UCSD.

Results

Bacterial microbiome

Non-contaminant Bracken species counts were used to describe the microbiome.

Phylum	This study	Human Microbiome Project	Nasopharyngeal microbiome
Firmicutes	47 (35-59)	38 (21-62)	21
Actinobacteria	17 (9-25)	14 (5-27)	3
Bacteriodota	14 (7-23)	12 (3-20)	11
Proteobacteria	10 (2-22)	18 (11-28)	67
Fusobacteriota	2 (0.5-3)	3 (1-6)	1

Table 5 Comparison of relative abundance (%; median, and where available, interquartile range) of five major phyla in this study, oral metagenomic samples of the Human Microbiome Project (Huttenhower et al., 2012; Oliveira et al., 2018) and nasopharyngeal samples from an amplicon approach (Bogaert et al, 2011).

19 555 bacterial species from 5 325 GTDB genera (5 274 when GTDB genus splits merged) were identified by one or more reads. Only 637 species ever had a relative abundance 0.1% or higher, and 305 species exceed this in at least 10% of samples.

The median Shannon diversity index is 4.5 (IQR 4.3-4.8) with a small but significantly greater diversity in the Febrile group (median 4.7 vs. 4.4, $p<0.001$, Wilcoxon test). Evenness (Shannon-Weaver, ignoring zeroes) was similar between groups ($p=0.80$, Wilcoxon test), with overall median 0.63 (IQR 0.57-0.70).

Estimates by phylum show considerable agreement between this study and the shotgun metagenomic arm of the Human Microbiome Project (HMP; Huttenhower *et al.*, 2012; Oliveira *et al.*, 2018), though firmicutes have greater dominance and proteobacteria and fusobacteria demonstrate somewhat lower abundance (Table 5). There is considerable disagreement with the nasopharyngeal microbiome of children aged 18 months, where a much higher proportion of proteobacteria was found (Bogaert et al., 2011).

Table 6 shows the most abundant genera by median relative abundance compared with HMP.

Genus	Median relative abundance (RA%)	Human Microbiome Project (RA%)
Streptococcus	27	19
Prevotella	8.8	3
Veillonella	7.2	5
Rothia	5.6	3
Neisseria	3.8	5
Pauljensenia	3.4	-
Haemophilus	2.5	5
Gemella	1.6	2
Actinomyces	1.3	5
Granulicatella	1.2	0.1

Table 6 Relative abundance (RA) of most abundant genera in this study compared with oral metagenomic samples in the Human Microbiome Project (Huttenhower et al., 2012; Oliveira et al., 2018).

Rarefaction analysis shows that the median number of species which account for at least 75% of bacterial reads is 55 (IQR 40-70; Figure 12). Top species by median relative abundance are shown in Table 7, with and without merging of species and genera split by GTDB.

Species	Median relative abundance (%)
Rothia mucilaginosa	1.4
Prevotella melaninogenica	1.2
Rothia sp001808955	1.0
Veillonella atypica	0.8
Rothia mucilaginosa_B	0.7
Streptococcus salivarius	0.6
Rothia mucilaginosa_A	0.5
Streptococcus sp001556435	0.5
Veillonella dispar_A	0.4
Prevotella histicola	0.4

Species	Median relative abundance (%)
Streptococcus mitis	6.4
Rothia mucilaginosa	3.8
Streptococcus pseudopneumoniae	2.0
Prevotella melaninogenica	1.5
Streptococcus parasanguinis	1.3
Streptococcus oralis	1.1
Rothia sp001808955	1.0
Streptococcus infantis	0.9
Haemophilus parainfluenzae	0.9
Streptococcus pneumoniae	0.8

Table 7 Most abundant species by relative abundance. Left table shows original GTDB species and right table with split taxa merged.

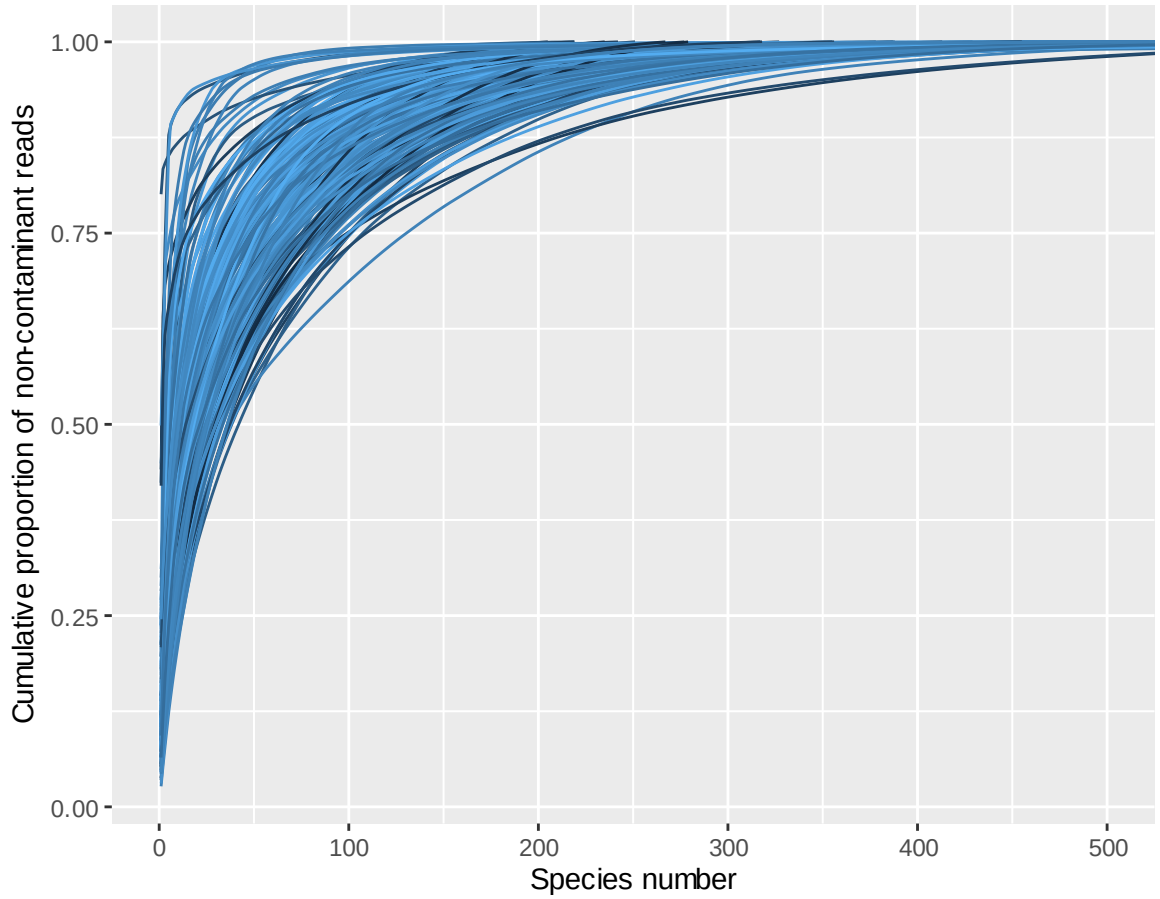

Figure 12 Rarefaction curves showing the cumulative proportion of relative abundance comprised by species from most abundant to least abundant.

Merging subdivided GTDB species reveals dominance by streptococcal species (*Strep. mitis*, *pseudopneumoniae*, *parasanguis*, *oralis*, *pnuemoniae*, *infantis* and *pneumoniae*), *Rothia mucilaginosa*, *Prevotella melaninogenica* and *Haemophilus parainfluenzae*. Indeed, at genus level, median relative abundance of *Streptococcus* is the highest at 27%, followed by *Prevotella* (9%), *Veillonella* (7%), *Rothia* (6%) and *Neisseria* (4%) (see Table 6, Chapter 3, p62).

The species with the greatest spreads (5-95th percentile) comprise *Streptococcus mitis* and other species, *Rothia mucilaginosa*, *Prevotella melaninogenica* and *histicola* and *Gemella haemolysans*, ranging between 7 and 32 percentage points.

MDS of weighted UniFrac distances (excluding species which never exceeded 0.1% relative abundance) revealed broad overlap between KD and Febrile samples (Figure 13 left). Two dimensions and 500 iterations were sufficient to allow convergence. KD and Febrile samples broadly overlap, and distance from the centroid appears to have no relationship with the bacterial proportion (Figure 13 right).

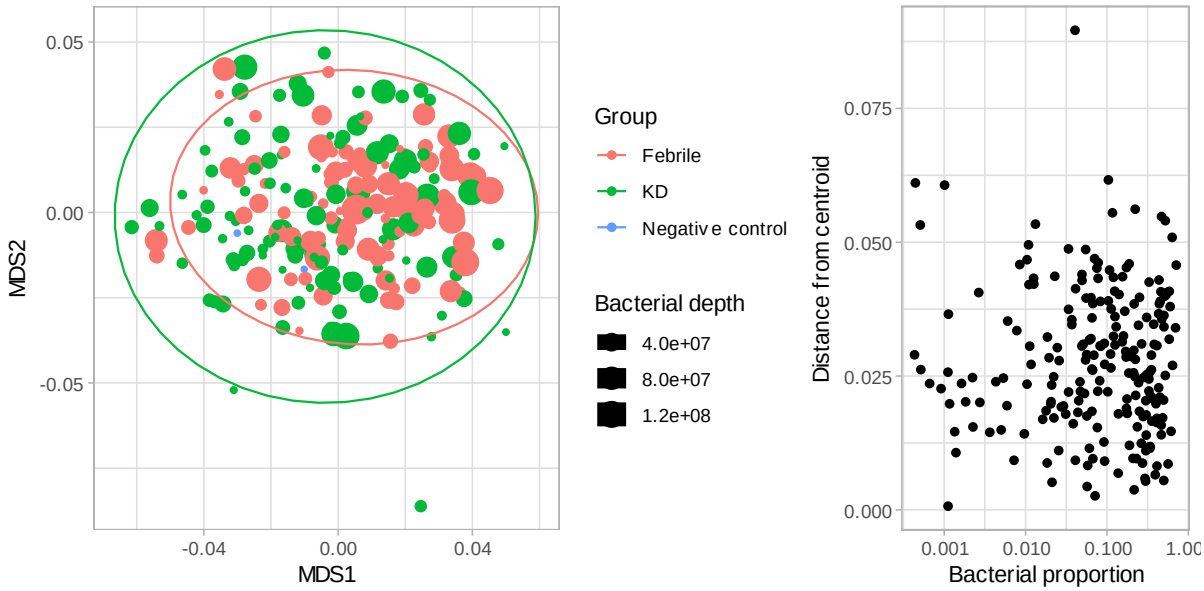

Figure 13 Non-metric multidimensional scaling (MDS) plot (left) of weighted UniFrac distances between samples showing no clustering by disease group. On the right, each sample's distance from the MDS centroid is shown against the proportion of bacterial reads in the sample.

Eight of the 11 samples exceeding the 95% percentile for distance from the centroid belonged to the KD group. Three had non-contaminant bacterial depth below 1 million reads. Reviewing the most abundant five species present in these 11 samples revealed samples dominated by known oral flora, except for one low-depth sample which was dominated by *Pseudomonas helleri* (first identified in raw cows' milk) and *Acinetobacter junii* (which has been reported to cause infections in humans).

Notably the two negative control samples cluster close to the centroid and are dominated by similar organisms to other samples. Their mean distance to the centroid of samples sharing indices is 42-44% of the mean distance for other samples. This is consistent with "index hopping" where free tag from control samples with minimal library DNA has been spliced with other samples' DNA.

The weighted UniFrac distance between the samples from the same febrile patient (10 days apart) was 40% of the mean pairwise distance.

Sources of variation

Univariable ANCOVA of weighted UniFrac distances showed KD to explain 2.7% of the variance in community structure (p=0.001). Multivariable ANCOVA incorporating country, sex, age and group did not materially alter this, but country, gender and age were also associated (respectively p=0.03, 0.007 and 0.001; 1%, 1% and 5% of variance; R^2=0.10). Within UCSD samples (n=199) prior antibiotic exposure explained 2.8% of variance (p=0.001), with R^2 of 0.13 combining the clinical covariates. Neither predicted ancestry (as determined from human sequences by Evangelos Bellos, personal communication) nor season were significantly associated with microbiome (p=0.28 and 0.80 respectively).

Adding the logarithm of bacterial DNA proportion to the model increased R^2 to 0.12 and 3.5% of the variance was explained by this (p=0.001). Adding the logarithm of the ratio between contaminant and non-contaminant bacterial DNA increased R^2 to 0.14, with this variable explaining 1.0% of the variance (p=0.046).

Given the clear relationship between species-level microbiome and the bacterial DNA proportion, independent of sample group, it was considered relevant to incorporate these into sensitivity analyses of multivariable analyses of KD association.

Pathogens

Six bacterial respiratory tract pathogens were identified in the sequencing data across a range of abundance and prevalence values (Table 8).

Agreement with clinical classifications of febrile control children as adenovirus or *Streptococcus pyogenes* infection was explored. *Streptococcus pyogenes* was present at relative abundance ≥ 0.1% in all seven patients with this clinical classification, 18/95 other febrile patients and 7/115 KD cases.

Total mastadenovirus abundance (using non-human reads as denominator, since the virome is dominated by bacteriophage reads) meets or exceeds 0.1% in nine of eleven patients with a positive RT-PCR reported, 17/91 without and 8/115 KD cases.

Identification of organisms associated with KD

Relative abundances of species were aggregated through ascending levels of the GTDB taxonomy, and the default statistical approach of MaAsLin 2 (Mallick et al., 2021) applied by operational taxonomic unit (OTU) with a generalised linear model measuring the association of group, age, country and sex.

Species	**Present (one or more read)**		**Median non-zero relative abundance (×10 000) (max, %)**	
	Febrile	**KD**	**Febrile**	**KD**
Streptococcus pneumoniae	102/102 (100%)	113/115 (98.3%)	81.1 (43%)	72.5 (82%)
Haemophilus influenzae	100/102 (98%)	106/115 (92.2%)	3.84 (6%)	4.36 (7%)
Neisseria meningitidis	101/102 (99%)	108/115 (93.9%)	2.17 (5%)	2.74 (3%)
Mycoplasmoides pneumoniae	2/102 (2%)	1/115 (0.9%)	0.332 (<0.01%)	1.16 (0.01%)
Streptococcus pyogenes	97/102 (95.1%)	98/115 (85.2%)	0.230 (4%)	0.113 (0.03%)
Legionella pneumophila	2/102 (2%)	1/115 (0.9%)	0.006 (<0.01%)	0.001 (0.01%)

Table 8 Presence and abundance of bacterial pathogens.

GTDB OTUs include conventional taxonomic levels (e.g. species and genus) as well as hierarchical groups of taxa between levels. Two sensitivity analyses were undertaken. The first added log(bacterial DNA proportion) as a parameter (similar to the Decontam model). The second excluded samples with greater than 10% contamination.

OTUs which never exceeded 0.1% relative abundance or were present in fewer than 10% of samples were disregarded, but still contributed to abundances higher up the taxonomy. Zero abundances were replaced with half the taxon's minimum. A dynamic thresholding approach was also taken, prioritising nadirs in the relative abundance distribution curves with between 12.5 and 50% of samples above this threshold.

For each approach and predictor, the treeclimbR algorithm was applied to determine the specific level at which to report p values in every part of the taxonomy, and apply false-discovery rate control. For the primary analyses, corresponding sensitivity analysis p values were reported. However, sensitivity analysis p values were also aggregated independently by treeclimbR and presented separately.

In the differential RA approach, positive associations with KD were found for 10 species (negative associations for 243) at a false-discovery rate of 0.2, and one at 0.05 (Table 9). No p values were reported from higher taxonomic levels. Coefficients correspond to the logarithm of relative abundance, thus a coefficient of 0.78 corresponds to a 6-fold increase in mean RA.

Organism	**Coefficient** (log_{10} scale)	**Q value**	**Median relative abundance** (%)	**Q value sensitivity 1**	**Q value sensitivity 2**
Lautropia mirabilis	0.78	0.06	0.66	0.32	0.06
Rothia dentocariosa	0.76	**0.05**	0.28	0.32	0.08
Neisseria sp000090875	0.65	0.11	0.05	0.20	0.32
Prevotella oris	0.62	0.06	0.21	0.24	0.14
Abiotrophia defectiva	0.62	0.13	0.13	0.34	**0.03**
Abiotrophia sp001815865	0.58	0.18	0.19	0.49	0.06
Streptococcus sp000831085	0.50	0.07	0.06	0.08	**0.03**
Streptococcus pseudopneumoniae O	0.49	0.11	0.55	0.42	**0.04**
Streptococcus sanguinis G	0.47	0.15	0.03	0.23	0.21
Streptococcus mitis AD	0.43	0.14	0.30	0.25	0.06

Table 9 Species with relative abundance positively associated with KD in MaAsLin 2 generalised linear model (Mallick et al., 2021) accounting for age, sex and country as additional covariates (FDR=0.2). TreeclimbR was applied to allow reporting of signal at higher taxonomic levels, though no aggregation occurred (Huang et al., 2021). Two sensitivity analyses are presented. The first includes an additional contaminant signal covariate of log10(bacterial proportion). The second excludes samples with greater than 10% of bacterial reads assigned to identified contaminants. Q values under 0.2 are underlined and those under 0.05 are emboldened.

Analysis	Organisms	Coefficient	Q value	Median relative abundance (%)
Sensitivity 1	Neisseria sp000090875	0.61	0.20	0.05
Sensitivity 1	Streptococcus sp000831085	0.54	0.08	0.06
Sensitivity 1	Streptococcus oralis W	0.43	0.15	0.02
Sensitivity 1	Streptococcus oralis AC	0.43	0.14	0.03
Sensitivity 1	*Albidiferax, Hylemonella, C04, Rhodoferax*	0.34	0.20	0.0006
Sensitivity 2	Lautropia, *UBA4615, UBA3064, SCN-69-89*	0.85	0.12	0.68
Sensitivity 2	Abiotrophia, *Dolosicoccus paucivorans, Aerococcus, Facklamia, Globicatella, Eremococcus coleocola, Ignavigranum ruoffiae, Facklamia A, Facklamia B*	0.82	0.08	0.33
Sensitivity 2	Streptococcus pseudopneumoniae O, Streptococcus mitis BE	0.57	0.11	0.89
Sensitivity 2	Streptococcus oralis, Streptococcus oralis (AB, R, D, M, N, S), Streptococcus sp000831085, Streptococcus sp900546335	0.46	0.12	0.54
Sensitivity 2	Streptococcus oralis W	0.42	0.12	0.02

Table 10 Species and groups of species with relative abundance positively associated with KD in MaAsLin 2 generalised linear model (Mallick et al., 2021) sensitivity analyses with TreeclimbR applied to allow reporting of signal at higher taxonomic levels (Huang et al., 2021) with FDR = 0.2. Species in italics fell below the 0.1% relative abundance threshold and were not tested individually. Both analyses accounted for age, sex and country as additional covariates. Sensitivity 1 includes an additional contaminant signal covariate of log10(bacterial proportion). Sensitivity 2 excludes samples with greater than 10% of bacterial reads assigned to identified contaminants.

With both sensitivity analyses, 5 positive associations with KD were reported at Q value threshold of 0.2, with some aggregation of p values at higher taxonomic levels (Table 10).

In the differential prevalence (dichotomous) approach (Table 11) 13 species or groups of species, genera or orders have positive associations with KD reported at a Q value threshold of 0.05 (22 negative). In sensitivity analyses, Q values fell between 0.08 and 0.22.

Organisms	Coefficient	Q value	Median relative abundance (%)	Q value sensitivity 1	Q value sensitivity 2
Streptococcus mitis (AC, AD, AF, AG, AH, AI, AK, BE, BH, S), Streptococcus pseudopneumoniae (G, O), Streptococcus sp001650315, Streptococcus sp003627155	1.4	**0.02**	3.2	0.08	0.12
Abiotrophia, *Dolosicoccus paucivorans, Aerococcus, Facklamia, Globicatella, Eremococcus coleocola, Ignavigranum ruoffiae, Facklamia A, Facklamia B*	1.3	**0.02**	0.33	0.09	0.10
Lautropia, *UBA4615, UBA3064, SCN-69-89*	1.2	**0.03**	0.68	0.09	0.13
Cardiobacteriales, *PS1, GCA-002705445, UBA11654, SAR86*	1.2	**0.03**	0.023	0.12	0.23
Streptococcus oralis BH	1.2	**0.03**	0.022	0.08	0.13
Streptococcus oralis (H, L, V)	1.2	**0.03**	0.18	0.14	0.17
Streptococcus oralis (E, F, Y)	1.2	**0.03**	0.099	0.14	0.17
Streptococcus oralis W	1.1	**0.04**	0.016	0.09	0.16
Streptococcus oralis Z	1.1	**0.04**	0.017	0.09	0.13
Streptococcus oralis, Streptococcus oralis (AB, D, M, N, R, S) Streptococcus sp000831085, Streptococcus sp900546335	1.1	**0.04**	0.54	0.14	0.14
Streptococcus pseudopneumoniae J, Streptococcus sp002355895, Streptococcus mitis AZ, Streptococcus oralis O	1.1	**0.04**	0.13	0.17	0.19
Streptococcus oralis AC	1.0	**0.04**	0.031	0.14	0.22
Neisseria sp001809325	1.0	**0.04**	0.11	0.13	0.20

Table 11 Species with dynamically dichotomised relative abundance associated with KD in generalised linear model accounting for age, sex and country as additional covariates (FDR=0.05). TreeclimbR was applied to allow reporting of signal at higher taxonomic levels (Huang et al., 2021). Species in italics fell below the 0.1% relative abundance threshold and were not tested individually. Two sensitivity analyses are presented. The first includes an additional contaminant signal covariate of log10(bacterial proportion). The second excludes samples with greater than 10% of bacterial reads assigned to identified contaminants. Q values under 0.05 are emboldened.

With sensitivity analyses, six positive associations were reported when bacterial proportion was taken into account, and twelve when samples with 10% or more contamination were excluded (Table 12).

Analysis	Organisms	Taxonomic level	Coefficient	Q value	Median relative abundance (%)
Sensitivity 1	Streptococcus mitis BE	Species	1.3	0.04	0.33
Sensitivity 1	Streptococcus sp001650315	Species	1.3	0.04	0.20
Sensitivity 1	Abiotrophia sp001815865	Species	1.3	0.05	0.19
Sensitivity 1	Abiotrophia defectiva	Species	1.2	0.05	0.13
Sensitivity 1	Streptococcus mitis S	Species	1.2	0.04	0.23
Sensitivity 1	Lautropia mirabilis	Species	1.2	0.05	0.66
Sensitivity 2	Abiotrophia, *Dolosicoccus paucivorans, Aerococcus, Facklamia, Globicatella, Eremococcus coleocola, Ignavigranum ruoffiae, Facklamia (A, B)*	Genus	1.6	0.012	0.33
Sensitivity 2	Streptococcus mitis (AC, AD, AF, AG, AH, AI, AK, BE, BH, S), Streptococcus pseudopneumoniae (G O), Streptococcus sp001650315, Streptococcus sp003627155	Species	1.5	0.013	3.2
Sensitivity 2	Lautropia, *UBA4615, UBA3064, SCN-69-89*	Genus	1.3	0.017	0.68
Sensitivity 2	Streptococcus oralis (H, L, V)	Species	1.3	0.032	0.18
Sensitivity 2	Streptococcus oralis, Streptococcus oralis (AB D M N R S), Streptococcus sp000831085, Streptococcus sp900546335	Species	1.3	0.026	0.54
Sensitivity 2	Streptococcus oralis Z	Species	1.2	0.032	0.017
Sensitivity 2	Streptococcus pseudopneumoniae J, Streptococcus sp002355895, Streptococcus mitis AZ, Streptococcus oralis O	Species	1.2	0.032	0.13
Sensitivity 2	Streptococcus oralis BH	Species	1.2	0.043	0.022
Sensitivity 2	Streptococcus mitis BJ, Streptococcus oralis (C, E, F, T, W, Y), Streptococcus sp900550895, Streptococcus halitosis	Species	1.1	0.046	0.26
Sensitivity 2	Streptococcus sp900104285	Species	1.1	0.043	0.026
Sensitivity 2	Neisseria sp001809325	Species	1.1	0.047	0.11
Sensitivity 2	Streptococcus oralis AC	Species	1.1	0.046	0.031

Table 12 Species with dynamically dichotomised relative abundance positively associated with KD in generalised linear model sensitivity analyses. TreeclimbR was applied to allow reporting of signal at higher taxonomic levels, though no aggregation occurred (Huang et al., 2021). FDR = 0.05. Both analyses accounted for age, sex and country as additional covariates. Sensitivity 1 includes an additional contaminant signal covariate of log10(bacterial proportion). Sensitivity 2 excludes samples with greater than 10% of bacterial reads assigned to identified contaminants.

Consistent results were found in favour of *Lautropia mirabilis* and *Abiotrophia defectiva*. The differences in relative abundance by group are shown in Table 13. The differences in means are greater than the medians, indicating a greater positive skew in the KD abundances.

Species/genus	Mean		Median	
	KD	Febrile	KD	Febrile
Abiotrophia defectiva	0.2%	0.05%	0.05%	0.03%
Abiotrophia	0.5%	0.1%	0.1%	0.07%
Lautropia mirabilis	1%	0.3%	0.2%	0.1%
Lautropia	1%	0.3%	0.2%	0.1%

Table 13 Mean and median relative abundances of *Abiotrophia defectiva*, *Lautropia mirabilis* and parent genera in disease groups.

Antibiotic associations

Antibiotic exposure was not included in the main model because of missing data for the UK KD patients (n=18). Antibiotic exposure was slightly more frequent in the KD group (54 vs 45%, see Table 1, page 40). The differential RA analysis was rerun with only samples from UCSD (n=199) adding antibiotic exposure as a covariate.

191 OTUs had negative associations with antibiotic exposure, and 3 positive (FDR=0.05). The positive associations comprised *Lautropia mirabilis*, *Lautropia sp003892345* and *Streptococcus sp001808705* (coefficients=1.4, 0.99 and 0.55, Q values<0.001, 0.03 and 0.02 respectively). *Abiotrophia sp001815865* had a weaker positive association with antibiotic usage (coefficient=0.59, Q value=0.12). *Abiotrophia defectiva* had less evidence of any positive association (coefficient=0.47, Q value=0.21). Neither association with *Abiotrophia* species had nominal significance (p=0.054 and 0.11 respectively).

The genus Streptococcus possessed the largest number of organism groups with significant negative associations, with 72 of 179 tested (FDR 0.05). Eleven other genera with smaller numbers of organism groups tested had high rates of negative associations and accounted for most of the negative associations (Table 14). Aggregating these genera showed overall a modest reduction in mean RA from 37 to 29%.

Genus	**Number of species groups tested**	**Number of species groups with negative coefficients**	**Number of species groups with negative associations** (FDR=0.05)	**Mean (Median) RA with no antibiotic exposure**	**Mean (Median) RA with antibiotic exposure**
TM7x	8	8 (100%)	8 (100%)	0.6% (0.1%)	0.2% (0.02%)
UMGS1907	3	3 (100%)	3 (100%)	0.2% (0.02%)	0.08% (0.01%)
Actinomyces	17	17 (100%)	15 (88%)	2.9% (1.8%)	1.9% (0.8%)
F0422	8	8 (100%)	6 (75%)	0.7% (0.09%)	0.3% (0.03%)
Rothia	7	7 (100%)	6 (86%)	8.6% (6.0%)	6.6% (4.4%)
Fusobacterium	14	14 (100%)	9 (64%)	1.5% (0.7%)	1.0% (0.4%)
Leptotrichia	12	12 (100%)	7 (58%)	0.9% (0.3%)	0.6% (0.2%)
Campylobacter	21	21 (100%)	10 (48%)	0.5% (0.3%)	0.4% (0.2%)
Granulicatella	5	4 (80%)	1 (20%)	1.6% (1.3%)	1.5% (1.1%)
Pauljensenia	22	22 (100%)	4 (18%)	7.0% (5.3%)	5.3% (3.0%)
Prevotella	45	41 (91%)	5 (11%)	12% (10%)	11% (7.5%)
Total	162	158 (97%)	74 (46%)	37% (38%)	29% (29%)

Table 14 Genera with high prevalence or species groups having reduced relative abundance (RA) with antibiotic exposure.

Five positive associations with KD were found (FDR=0.2), reduced from 10 in the earlier model. The KD association with *Lautropia mirabilis* remained, though now reported at genus level and weaker (Q value 0.19, coefficient 0.68). Antibiotic exposure explained more variance than disease group (8.6 vs 3.0%). The association with *Abiotrophia defectiva* was also reported at genus level and weakened (Q value 0.24, coefficient 0.57). For both organisms nominal significance remained (p=0.024 and 0.042 respectively). *Rothia dentocariosa*, *Prevotella oris* and *Streptococcus sp000831085* also remained at the FDR threshold. *Parvimonas micra* (grouped with *Parvimonas sp000214475*) arose as a new finding (coefficient 0.91, Q value 0.09).

Age associations

In the original differential RA analysis, 74 OTUs had significant associations with age (Q value ≤ 0.05) from species up to order level. The strongest positive associations were with a group of *Prevotella* species, the genus Alloscardovia, family Dialisteraceae, *Lancefieldella rimae*, Tannerellacaea and the genera *Alloprevotella* and *Prevotellamassilia*. To illustrate, the group of *Prevotella* species had a median relative abundance of 0.2% under 2 years of age, and 0.6% from age 2 to 9 and 1.7% in children 10 years of age and older.

The genera *Abiotrophia* and *Lautropia* also have positive associations with age (Q=0.04 for both). TreeclimbR aggregated the *Parvimonas* species' positive association to the level of family (Helococcacaea; Q=0.01), though *P. micra* itself had a stronger positive signal (coefficient 0.27 vs. 0.14). Median quantile regression shows RA of both *A. defectiva* and *L. mirabilis* rising exponentially from three months to around 18 months for both organisms (Figure 14). For *A. defectiva*, median RA

below 1 year is 0.001% and from 1 year is 0.04%. For *L. mirabilis,* median RA below 1 year is 0.002% and from 1 year is 0.2%. For the *Parvimonas* species there is a slower exponential rise to age 5, with RA 0.0005% below 1 year, 0.003% from 1 year to below 5 years, and 0.02% age 5 and above.

The associations with age, especially of *A. defectiva* and *L. mirabilis*, would not have been well-modelled by a simple linear relationship, nor by any simple transformation. In the primary relative abundance model, transforming age to have a ceiling of 18 months increases the strength of association between KD and *Lautropia mirabilis* (unadjusted p value changed from 0.02 to 0.004 and coefficient from 0.78 to 0.83) and KD and *Abiotrophia defectiva* (unadjusted p value changed from 0.04 to 0.02 and coefficient from 0.62 to 0.66). For the *Parvimonas* species, a ceiling of 5 years increases the strength of association with KD (unadjusted p value changes from 0.008 to 0.004 and coefficient from 0.91 to 0.96). In each case, age accounts for higher proportions of the variance (11% vs 3.5% for *Parvimonas* species, 23 vs. 1.9% for *A. defectiva*, and 27% vs 2.1% for *L. mirabilis*).

The strongest negative associations were *Streptococcus peroris*, two other individual Streptococcal species, a *Porphyromonas* species and the F0422 genus (Veillonellacaea). To illustrate, median relative abundance of F0422 is 0.3% under 2 years of age, 0.04% from age 2 to 9 and 0.03% in children 10 years of age and older.

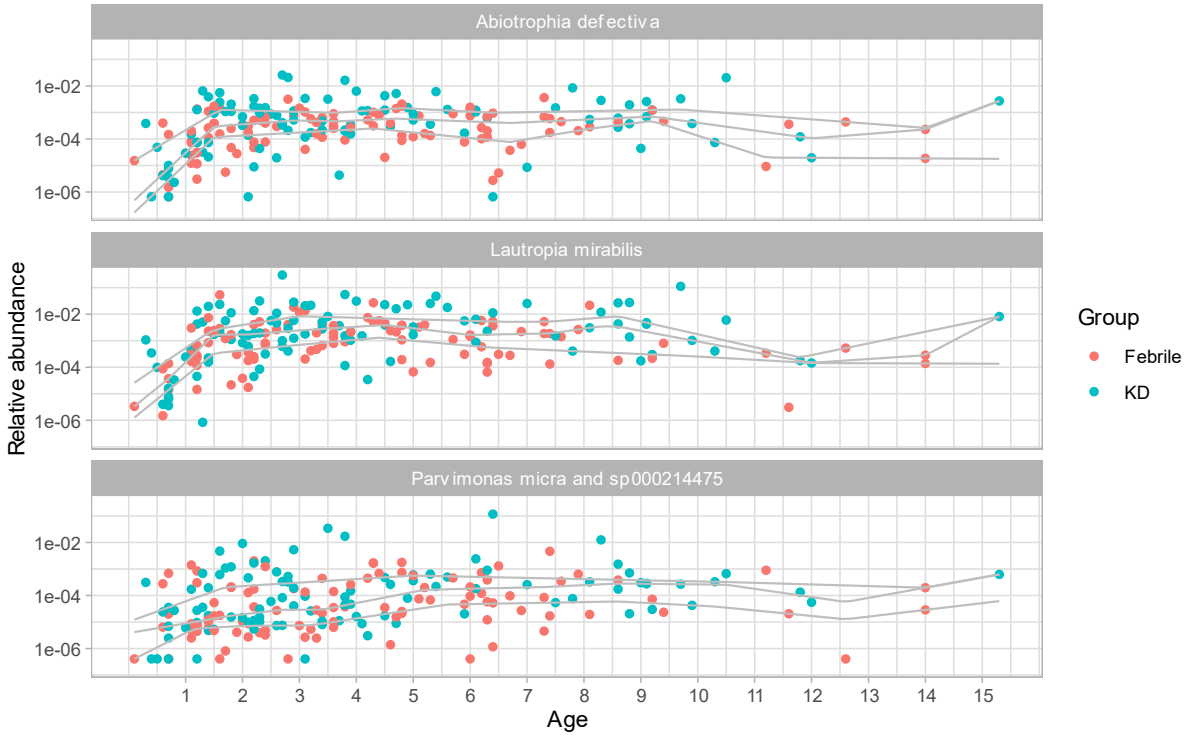

Figure 14 Relative abundance versus age of three organisms positively associated with Kawasaki Disease and age. Overlaid lines indidate 25th, 50th and 75th percentiles by quantile regression (smoothed spline).

Covariance of KD-associated taxa

A. defectiva and *L. mirabilis* RAs were correlated (R^2=34%), and most of this correlation persisted when the associations with age were taken into account (R^2 of residuals=0.20). The *Parvimonas* species' total RA was less strongly correlated with *A. defectiva* and *L. mirabilis*, and this was lost when age was taken into account (R^2=0.07 and 0.10, R^2 of residuals=0.01 and 0.03).

Using the residuals of RAs to predict KD in a logistic generalised linear model, all three were significant individually (p=0.006, 0.008, 0.02 for *A. defectiva*, *L. mirabilis* and *Parvimonas* species respectively). In a multivariable model all three lost significance.

Scaled principal component analysis of species and groups with associations with KD showed strong covariation with 43% of variance explained by the first principal component (data not shown).

Discussion

In this chapter, I have described the metagenomic data by species and genus composition, analysed overall variation with clinical covariates and sought species and higher-level taxonomic units associated with KD. The previous chapter illustrated the wide variation in samples by phylum, and here I show that species abundances vary by up to 32 percentage points (5-95th percentile).

Comparisons with existing knowledge of the pharyngeal microbiome

The top genera by median relative abundance comprise typical oral flora. Direct comparison with data from oral samples in the shotgun metagenomic arm of the HMP (Huttenhower et al., 2012) using MicrobiomeDB (Oliveira et al., 2018) revealed similarity, though it should be noted that these were adult oral samples, including throat, dental, tongue and palate samples.

The majority of paediatric pharyngeal microbiome studies use amplicon sequencing-based approaches. These cannot be used for direct comparison with our shotgun metagenomic approach, since genome size and 16S operon count weight the read counts in each approach. Despite much effort, current approaches to normalise 16S counts for copy number have not proven effective (Starke, Pylro & Morais, 2021).

Comparison with a study of the nasopharyngeal microbiome of healthy 18-month old children showed great differences, with a clear dominance of Proteobacteria instead of Firmicutes. Mean 16S operon copy number is higher in Firmicutes (6.8 vs 5.4) and the mean genome size is smaller (3.3 vs 4.6 million) (Stoddard et al., 2015; Westoby et al., 2021). Both of these differences would be expected to lead to 16S-based studies to estimate higher abundances for Firmicutes and lower for Proteobacteria.

Man *et al.* (2019) studied children up to 6 months of age and show that the nasopharyngeal and oral (buccal and sublingual) microbiomes are distinct from a week of age and that the oral microbiome is more stable over time. They show that the oral microbiome is consistently dominated by Firmicutes. In contrast, the nasopharyngeal dominance of firmicutes rapidly wanes by two months, with proteobacteria supplanting.

A search of the European Bioinformatics Institute's MGnify database provides only one study allocated to the pharyngeal biome – a 16S amplicon study of children with Cystic Fibrosis. 133 studies are allocated to the mouth biome – searching for "throat" and "pharynx" adds only one set of assemblies including throat samples among a wider range of oral samples. A PubMed search[1] identifies only one published shotgun metagenomic study of the paediatric oropharynx, but this included only neonates.

Although relative abundances cannot easily be compared directly between shotgun metagenomic and amplicon studies, differences in relative abundances should remain comparable. However, in regards of the changes shown with age, few studies provide comparative data. Mortensen *et al.* (2016) studied the establishment of the hypopharyngeal microbiome of infants using 16S amplicon sequencing at one week, one month and three months. However, with only one patient in our study under 3 months of age, this is of limited relevance. Streptococcacaea increased in relative abundance from 17 to around 30% by a month and Moraxellacaea rose from 9 to 24% by 3 months, while Staphylococcacaea fell from 49 to 10%.

Clustering and covariation

No clustering was evident by disease group, nor within the febrile group, and only a small portion of the variation could be explained by clinical covariates (R^2=10% rising to 13% including antibiotic exposure).

This is an interesting finding given the clear clustering of infant hypopharyngeal microbiomes (Mortensen et al., 2016). Nonetheless the nasopharyngeal microbiome did not demonstrate such clear clustering in children with Respiratory Syncytial Virus illness, though there were clear associations with disease severity (de Steenhuijsen Piters et al., 2016).

Associations with KD

Multiple approaches were considered for identifying organisms associated with KD. It would be most ideal to have unequivocal dichotomous results for presence or absence of organisms. However, the

[1] ((child* OR paediatr* OR pediatr*) AND (throat OR oropharyn* OR pharyn*) AND shotgun AND metagenom*)

nature of metagenomic sequencing and read binning gives rise to potential false classifications of reads due to genetic relatedness of organisms and errors in sequencing and databases.

A conventional differential relative-abundance approach was applied, as this would detect signal from both differential abundance and prevalence, with noisy read classification. The modelling approach selected is the default implemented by MaAsLin 2, since it performed best in benchmarking approaches (Mallick et al., 2021) in spite of not accounting for the compositional nature of the data (Gloor et al., 2017).

A dichotomous approach was also implemented, recognising that a static threshold (e.g. "non-zero relative abundance") would not be expected to perform well. For example, of 637 species which ever exceed 0.1% relative abundance, only 150 are undetected in 25% or more samples. Potential thresholds considered for each species included zero, the 75th percentile (fixing the proportion positive at 25%) and any troughs in the relative abundance distribution. These latter distributional measures were prioritised, though only 33 of 536 nodes reported in the primary analysis utilised these thresholds.

Many microbiome and metagenomic studies are limited by selecting a single taxonomic level at which to undertake differential analyses. When multiple levels are analysed, methods to control false discovery rates have been limited, and it is unclear how to select which results to report. TreeclimbR is a novel approach to select levels of the taxonomic hierarchy to report in a data-dependent manner. Applying it here increases the potential to detect associations with KD, regardless of the breadth or narrowness of organisms involved.

Positive associations with KD were found in both approaches, though the stronger associations were found in the dichotomous approach (lowest Q value 0.02 vs 0.05). *Abiotrophia defectiva* (or genus) was identified in both the continuous and dichotomous approaches (Q=0.13 and 0.02 respectively), the coefficients indicating a four-fold increase in mean relative abundance, or 3.7-fold increase in odds of relative abundance over the 75th percentile. The four-fold difference is evident in the crude estimates of mean relative abundance by group for both *Abiotrophia* and *Abiotrophia defectiva*.

Abiotrophia was identified by relative abundance in the independent sensitivity analysis excluding samples with ≥ 10% contamination (Q=0.08) with an increased co-efficient. However, in the sensitivity analysis adding bacterial proportion to the model, its significance diminished, though coefficient remained positive (Q=0.34, coefficient=0.45). It should be considered that this analysis is intended to account for organisms whose association with KD may be spuriously mediated by contamination,

through a shared relationship with sample bacterial proportion. *Abiotrophia defectiva* convincingly does not have an inverse association with bacterial proportion in the Decontam model (p=0.90).

In the sensitivity analyses for the dichotomous approach, *Abiotrophia defective* (or genus) retains its positive association both when bacterial proportion is included in the model or samples with 10% or more contamination are included (Q=0.05 and 0.08).

Abiotrophia defectiva is a known cause of infectious endocarditis and has been shown to adhere to vascular endothelium (Senn, Entenza & Prod'hom, 2006; Sasaki et al., 2020). More recently it has been shown that the protein DnaK assists in endothelial adhesion and is pro-inflammatory (Sasaki et al., 2021). One case of the inflammatory disorder haemophaghocytic lymphohistiocytosis has been associated with *A. defectiva* endocarditis (Kiernan et al., 2008). *Abiotrophia* is identified at very low abundance in infants from one week to three months (average relative abundance 0.002% compared to 0.1% in this study) (Mortensen et al., 2016).

Lautropia mirabilis (or genus) also demonstrated a positive association with KD (Q=0.06 and 0.03 in continuous and dichotomous approaches respectively), with coefficients corresponding to a six-fold increase in mean relative abundance or three-fold increase in probability of exceeding the 75th percentile. The crude estimates of mean relative abundance by group however showed only a three-fold difference.

The association with relative abundance was similarly maintained when excluding samples with high contamination (Q=0.06) and though reduced when adding bacterial proportion to the model (Q=0.32, coefficient=0.49), *Lautropia mirabilis* was again an unlikely contaminant (p=0.89). Its dichotomous association was maintained when bacterial proportion was added to the model or high-contaminant samples excluded (Q=0.09 and 0.13 respectively).

Lautropia mirabilis has no known disease associations, but its isolation from the oral cavity was noted to be strongly associated with human immunodeficiency virus (HIV) infection in children born to HIV-infected mothers in the USA (Rossmann et al., 1998). It is not identified in the study of infants up to three months (Mortensen et al., 2016).

The non-linear relationship of both of these organisms with age – with median relative abundance plateauing at 18 months – suggests that abundance of these organisms may relate to primary dentition. It is notable that the strength of the association with KD only increases when the relationship with age is better modelled.

That KD may be caused by exposure to ubiquitous micro-organisms in genetically susceptible hosts is a well-established hypothesis (Rowley & Shulman, 2018). The age-related incidence of KD could then

relate to both age at first exposure, and an immunological window of susceptibility, whether related to temporal development of the immune system, immune training or cross-reactive adaptive immune responses. Recent evidence through the pandemic adds credence to immunological windows of susceptibility and these results will be considered in context in the final discussion (p142).

Antibiotic associations

The analysis of associations with antibiotic exposure provides biologically plausible evidence of broad reductions in RA across a range of taxa, with some genera being particularly impacted. Although the fold decreases for some organism groups is very large, the aggregate effect is modest – this fits with the small proportion of microbiome variance explained by antibiotic exposure.

It is interesting to see three species with moderate signals of increased RA in association with antibiotic exposure – with coefficients corresponding to increases in mean RA of 73 to 322%. The strongest signal was from *Lautropia mirabilis,* which is intriguing since two studies show susceptibility to penicillin, ampicillin, erythromycin and gentamicin (Gerner-Smidt et al., 1997; Dekhil et al., 1997).

Positive causal associations with antimicrobial exposure would give rise to potential confounding given the more frequent exposure to antibiotics in the KD group, and negative associations could mask true associations. Reassuringly, the positive KD associations are not abrogated by the inclusion of this covariate, and nominal significance is maintained. One new potential association is found with the dental organism *Parvimonas micra*.

Cohort

The majority of patients providing samples were from the USA, and all febrile controls. The AHA criteria were followed for both US and UK recruitment. Clinical characteristics of US and UK KD patients are similar, though ethnic compositions differ (Table 1, p40). Reassuringly, country explained only 1% of the variance in microbiome composition, less than age, disease status or antibiotic exposure. This supports the inclusion of UK patients, though it would have been ideal to also include UK febrile controls. Geographical diversity has some potential to impede the identification of causative agents if they differ between regions, since statistical associations may be weaker when shared across more agents.

Negative associations with KD

Far more species had negative associations with KD, with 166 species at an FDR of 0.05 (data not shown). The most significant associations included Streptococcus, Roseburia, Pauljensenia, Veillonella and Lachnoaerobaculum species. Of greatest clinical relevance is *Streptococcus pyogenes* (coef=-2.0,

q=0.001), which was already noted to have greater abundance in febrile patients, and likely represents the contribution of this organism to a proportion of non-KD febrile presentations.

Conclusions

In this chapter I present extensive exploration of the oropharyngeal microbiome in our children with KD and other febrile illnesses, and a range of approaches to identify organisms associated with KD, whether through increased prevalence or abundance.

No signal is identified from an organism which is low prevalence in febrile controls and higher prevalence in children with KD. This would have been represented by significant results from organisms dynamically thresholded at zero, or at a distributional trough representing a boundary between low-level false positive identifications, and true identifications. Indeed the minimum unadjusted p values from 33 trough-thresholded organisms is 0.88, and from 61 zero-thresholded organisms is 0.11.

The signals identified for *Abiotrophia defectiva* and *Lautropia mirabilis* are driven by differential abundance of these highly prevalent organisms, independent of the strong non-linear association with age. The pathological relevance of these differences requires further exploration.

Epidemiologically an organism which is almost universally prevalent after 3 months and acquired early in life seems unlikely to be causatively associated with KD. However, abrupt and dramatic changes in abundance relating to primary dentition could be hypothesised as a trigger of the immune system. Alternatively, the abundance signal could relate to acquisition of distinct strains of these organisms which harbour genes driving virulence.

In the next chapter I present pangenome and strain analyses of these two organisms to elucidate subspecies differences. I also produce *de novo* assemblies and metagenome assembled genomes from samples, in order to more broadly characterise subspecies differences across organisms.

4 | Metagenomics – identifying genes and strains associated with Kawasaki Disease

Introduction

The former chapter presented analyses focusing on identifying organisms associated with KD at the level of species and above. A key advantage of the metagenomic approach here, as compared with amplicon-based approaches, is the ability to look at higher specificity, including strains and accessory genes.

In this chapter I apply off-the-shelf tools to search for strain and pangenome differences in candidate organisms between KD patients and controls. I also present a bespoke approach to look for bacterial genes over-represented in KD patients.

While KD might be caused by a novel bacterial species, there are a number of examples of diseases where a common species or strain, acquires or evolves a single gene or group of genes which drive specific diseases. Examples of this include Diarrhoea-associated Haemolytic Uraemic Syndrome, which is now known to be caused by *E. coli* possessing a verotoxin gene (Karmali et al., 1983). Another example is the staphylococcal Toxic Shock Syndrome, caused by *S. aureus* possessing superantigen toxin genes in non-immune individuals (Kulhankova, King & Salgado-Pabón, 2014).

In this Chapter I attempted to explore the metagenomic data to identify a novel gene or toxin that might be enriched in KD vs febrile controls.

Objectives are:

1. Search for statistically significant associations between strain profiles and pangenomes of *Lautropia mirabilis* and *Abiotrophia defectiva*
2. Identify clusters of highly homologous bacterial genes from assembled sequence data and search for associations between KD and gene prevalence using unique kmers

Methods

Assembly and binning of unaligned reads

Unaligned reads generated following human genome alignment were assembled *de novo* using MEGAHIT (Li et al., 2015) and the meta-large preset, which is optimised for large metagenomes. Each sample's reads were aligned to the corresponding contigs. MetaBAT 2 (Kang et al., 2019) was run per sample with default settings to bin contigs into metagenome assembled genomes (MAGs) according to tetranucleotide frequencies and coverage.

MAG quality was assessed with CheckM (Parks et al., 2015). Completeness of ≥90% and contamination of <5% were considered to represent high quality MAGs and ≥50% and <10% to represent medium quality.

Taxonomic identification of contigs and MAGs

MAGs were identified using sourmash (Pierce et al., 2019), which is a *k*-mer hash-based indexing and serarching tool. Sourmash was run in lowest common ancestor mode, using pre-built database of 31-mers for GTDB release 202 (release 95 unavailable).

Contigs were mostly too short for identification by sourmash, and were instead identified by Kraken against the custom database, with confidence threshold 0.05.

Contigs and MAG taxonomic identifications were cross-referenced with potential contaminants.

Generation of species-specific pangenomes

MAGs identified by sourmash corresponding to *Lautropia mirabilis* and *Abiotrophia defectiva* (or identified at genus level) and of at least medium quality were identified. Reference genomes for the species were also obtained from NCBI. The PanPhlAn genome exporter was used to predict, identify and cluster genes to form the pangenome for each species.

PanPhlAn analyses

Non-human reads from each sample were analysed with PanPhlAn 3 (Beghini et al., 2021) against each pangenome, and profiled in sensitive mode (min_coverage 1, left_max 1.70, right_min 0.30, th_non_present 0.25, th_present 0.5) according to the author's advice (Dubois L, personal communication).

Gene presence-absence matrices were analysed by Principal Components Analysis (PCA). Genes with differential prevalence between KD and Febrile patients were sought by Fisher's exact test, and by Firth logistic regression, accounting for total gene count.

A permutation-based approach was used to detect significant contiguous gene sets with concordant signal. Sample labels were permuted 500 times. Across each permutation, and the original labelling, clusters of contiguous genes with nominally significant p values and similar direction of effect were identified. Cluster mass was defined as the sum of -log10(p) values. The 95th percentile of maximum cluster mass was defined as the threshold for reporting significant clusters.

StrainPhlAn analysis

Non-human reads from each sample were analysed with MetaPhlAn 3 (Beghini et al., 2021) to produce bowtie2 alignment files against the marker gene database. Consensus reference sequences were

extracted, and alignments generated for *Lautropia mirabilis* and *Abiotrophia defectiva*. Distance matrices were created using distmat from the EMBOSS package with Kimura scoring (Rice, Longden & Bleasby, 2000). Distances were subject to principal co-ordinates analysis (PCoA).

Global gene prediction

Prodigal was applied to unbinned contigs to predict genes and write their corresponding nucleotide sequences. Bedtools was used to write nucleotide sequences for predicted genes from MAGs, according to genome feature files produced by CheckM above, also using Prodigal.

Predicted gene clustering and unique kmer generation

All predicted genes were clustered with mmseqs2 (Steinegger & Söding, 2017) to 95% identity using cluster mode 2, minimum coverage of 80%, coverage mode 1 and cluster reassignment.

Unique 32-mers were identified from cluster representative genes using Jellyfish. Overlapping 32-mers were filtered by aligning them against the source sequences to 100% identity (minimum score zero) with bowtie 2 and filtering any aligned 32-mers <33 nucleotides 3' from a previous alignment. A maximum of two alignments were allowed, to confirm uniqueness. 32-mers were written into a tab separated values (TSV) file with one column for the 32-mer and one for the source sequence. A numbered 32-mer fasta file was also generated.

Representative gene detection

Jellyfish was applied to count the shortlist of unique kmers per sample, and the resulting Javascript Object Notation file converted to tab separated values. Further bash scripts appended the corresponding representative genes and summarised 32-mer counts per representative gene. Each sample's file was read into R separately and combined into a matrix of 32-mer detection counts by sample and representative gene, dichotomised to presence or absence.

Since the odds of detection of low abundance sequences will vary linearly with sequencing depth, logistic regression by gene was applied with presence as the outcome and covariates comprising group (KD or Febrile) and the logarithm of the number of genes detected. Benjamini-Hochberg FDR-control was applied. Model-fitting was accelerated across multiple cores with the parallel package.

Results

Assembly of unaligned reads

Unaligned reads were assembled on a sample-wise basis with MEGAHIT. For contigs over 2500 bp, the median N50 (contig size threshold above which 50% of bases in the assembly reside) was 13 150 (IQR 9 767-18 292) and assembled length 66 million (IQR 25-111 million) bp. Median N50 was 12 832 and

13 476 in Febrile and KD groups respectively (p=0.30 by Wilcoxon test) and assembly length 90 million and 44 million (p<0.001).

A median of 64% of each sample's reads were aligned to corresponding assemblies (IQR 37-77%, range 3-91%) – 69% in the febrile group and 61% in KD (p=0.02 by Wilcoxon test).

Binning of contigs

Contigs from each sample were binned with metaBAT 2 with default settings. 12 870 metagenome assembled genomes (MAGs) were generated of any quality. By checkm, 2 825 were of at least medium quality (completeness ≥ 50% and contamination < 10%) and 604 of high quality (completeness ≥ 90% and contamination < 5%).

Taxonomic identification of contigs and MAGs

Contigs over 2500 bp numbered 2.1 million. Kraken gave a taxonomic label to 92% of these. Of identified contigs, 97% were bacterial, 3% human, 0.1% fungal and an even smaller proportion viral or archaeal. The most numerous bacterial contigs by genus were Streptococcus, Prevotella, Pauljensenia and Veillonella, together comprising 47% of bacterial contigs.

Where human contigs were binned, these primarily made purely human MAGs (n=134). Only 22 MAGs mixed human and bacterial contigs.

MAGs were identified by sourmash in lowest common ancestor mode, with 82% receiving an identity. Of these, 66% were identified at species level and 31% at genus level. The most numerous MAGs by genus were Prevotella, Streptococcus, Pauljensenia and Rothia, comprising 46% of MAGs.

Contaminants

Previously identified contaminant species, and genera with over 50% of mean relative abundance contributed by contaminants were identified. Contigs were identified as potential contaminants, and MAGs containing potential contaminant contigs, or identified by sourmash as potential contaminants. Only 0.4% of contigs and 2.6% of MAGs were identified as potential contaminants.

Generation of species-specific pangenomes

Abiotrophia defectiva and *Lautropia mirabilis* MAGs of at least medium quality were identified by sourmash taxonomic labelling. Genus-level matches were also included.

Seventeen *Abiotrophia defectiva* and *Abiotrophia* MAGs were identified, of which 26% of contigs were classified as *Abiotrophia defectiva,* 73% as *Abiotrophia sp001815865* and the remainder at higher levels or unclassified. *A. sp001815865* is a closely-related species – distance 0.008. Two *A. defectiva*

reference genomes (GCF_000160075.2 and GCF_013267415.1) and one of the closely-related GTDB species (GCF_001815865.1) were added.

Fifty-four *Lautropia mirabilis* and *Lautropia* MAGs were identified, of which 99% of contigs were classified as *Lautropia mirabilis*, the remainder as a moderately-closely related species (distance 0.09), higher-level taxonomic labels or unclassified. Two *L. mirabilis* reference genomes (GCF_000186425.1 and GCF_900637555.1) were added.

PanPhlAn pangenomes were generated for both genome collections. The *A. defectiva* pangenome comprised 3 390 clusters of orthologous genes (COGs), of which 1 823 were present within any of the three genome references and 1 302 were unknown COGs. The two *A. defectiva* reference genomes were highly similar with only one COG distinct to each genome. The close relative was missing 291 COGs and provided no additional COGs. Of 285 known COGs not derived from the reference genomes, 12 derived from strains of *A. defectiva* and 232 from *A. sp. HMSC24B09*. The largest proportion of the remainder derived from *Streptococcus sp*.

The *L. mirabilis* pangenome comprised 11 765 COGs, of which 2 500 were present within the two genome references. All COGs were shared between the two references. Of 205 known COGs not derived from the reference genomes, 66 derived from strain ATCC 51599 of *L. mirabilis,* 16 from *L. mirabilis* and 51 from *L. dentalis*.

Pangenome comparisons

In order to identify whether the gene content of *A. defectiva* and *L. mirabilis* differs between KD and febrile children, PanPhlAn 3 was run with non-human reads from each sample. PanPhlAn identifies COGs as present or absent within a sample, based on coverage. For both species, 64 KD and 65 febrile patients gave results.

For *A. defectiva*, a median of 1 339 of 1 823 (maximum 1 571) reference genome COGs could be detected per sample. Detection rates clearly varied between reference genome COGs, with some detected in all samples and some in no samples (median 84%). Ninety-one percent of MAG COGs could be identified by PanPhlAn in the sample reads, with similar numbers of additional COGs detected which had not been assembled or binned together. No difference was observed between KD and Febrile patients by PCA (11% of variance explained by first two components).

For *L. mirabilis*, a median of 1 933 of 2 500 (maximum 2 077) reference genome COGs could be detected per sample. Detection rates again varied widely between reference genome COGs (median 93%). Eight-four percent of MAG COGs could be identified by PanPhlAn in the sample reads. No

difference was observed between KD and Febrile patients by PCA (9% of variance explained by first two components).

For both species, the number of COGs detected varied between samples (*L. mirabilis* 1 870 to 2 611; *A. defectiva* 1001 to 1 829). It is notable how close the maximum is in each case to the number of reference genes. More COGs were detected in KD patients than febrile patients (median 1602 vs 1494 for *A. defectiva* and 2 447 vs. 2 377 for *L. mirabilis*).

COGs significantly over-represented in KD were sought by Fisher's exact test with KD numerators divided by the ratio of median COG counts by group, to account for depth effects. Only COGs which were detected 10 or more times were included. For *A. defectiva*, no over-represented COGs were found at FDR 0.2. Eight COGs were under-represented. Four of these were known COGs: an uncharacterised protein from strain ATCC 49176, ltrA from *Streptococcus cristatus* ATCC 51100, mco from *Oribacterium sp.* oral taxon 108 str. F0425 and cadC from *Kurthia sp.* 11kri321. Three of these COGs were also identified by Firth logistic regression with adjustment for the logarithm of core gene count.

For *L. mirabilis*, no over- or under-represented COGs were found by Fisher's exact test at an FDR of 0.2.

Since accessory gene differences within a species are often driven by loss or acquisition of regions of DNA, signals from genes may be better resolved across the series of genes within genomes. Thus permutation testing was used to identify the presence of significant clusters of genes with differential prevalence between KD and febrile patients. For *L. mirabilis* no significant gene clusters were identified at p=0.05, and for *A. defectiva* the only significant result was the single most under-represented COG (above).

Strain-level comparisons

For *A. defectiva*, 28 febrile and 35 KD patients could be analysed, with minimum sample marker threshold reduced to 15 and marker prevalence threshold reduced to 60%. Samples were a median of 8.1 (IQR 5.9-10.9) units apart, and 9.9 (IQR 8.2-11.9) units from the nearest reference genome. PCA explained 67% of the variance and no relationship with sample group was apparent.

For *L. mirabilis*, 55 febrile and 53 KD patients could be analysed, with one patient sampled twice, using default settings. Samples were a median of 5.2 (IQR 4.0 to 6.4) units apart and 7.0 (IQR 5.6-8.2) units from the reference genomes. The paired samples were very close (0.07). PCoA explained 22% of the variance and no relationship with sample group was apparent (PERMANOVA p=0.97 *A. defectiva*,

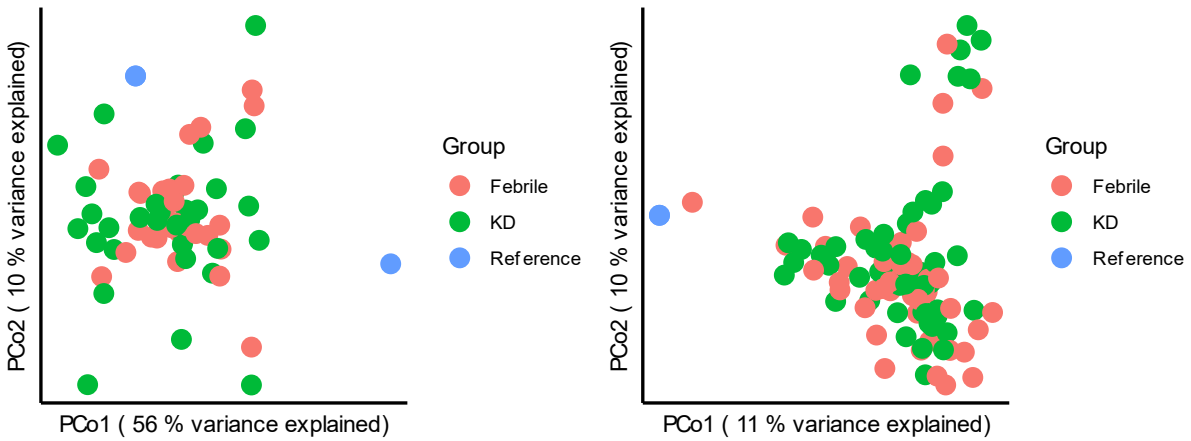

Figure 15 Principal component analysis of *Abiotrophia defectiva* and *Lautropia mirabilis* (left and right, respectively).

p=0.21 *L. mirabilis;* Figure 15). In neither case did samples from the UK cluster distinctly to the USA (data not shown), although for *L. mirabilis* PERMANOVA was significant at p=0.03 (p=0.25 for *A. defectiva*).

Gene prediction and clustering

Combining genes predicted for MAGs as part of the checkm pipeline and genes predicted from unbinned contigs, almost 17.4 million individual genes were available. Just over 2.3 million clusters were generated at 95% sequence identity and 80% coverage, with the longest sequence used as representative. Clusters ranged in size from one to 901 sequences, with median 2 and 27% of clusters with 5 or more sequences.

Unique 32-mers were reduced from 1.2 billion to 7.3 million when overlapping sequences were removed. Only 3% of cluster representatives did not have a unique 32-mer.

Clusters were cross-referenced with kraken taxonomic labels of contigs. Only 73 clusters mixed human and non-human sequences, and 144 511 clusters comprised human sequences alone. Cross-referencing contaminant classifications, 59 652 clusters had 90% or more sequences classified as contaminants. Over 2.1 million clusters remained for further analysis.

Representative gene detection

Jellyfish (Marçais & Kingsford, 2011) was used to count the specific 32-mers in each sample, which were then summed per representative gene as a proxy for gene presence. Exclusively human clusters were removed from further analyses.

Fifty-six percent of representative genes could be detected at least once, and 43% of representative genes in 10 or more samples. Representative genes could be detected in their corresponding samples 55% of the time among a random sample of 10 000 representatives.

A mean of 232 684 representative genes were detected per sample, 284 943 in febrile samples and 190 027 in KD samples. This difference represents the greater bacterial sequencing depth in febrile samples, and potentially the slightly greater diversity.

Representative genes with differential prevalence

In a Firth logistic regression model with the logarithm of number of genes detected, to account for varying diversity and sequencing depth, no representative genes were significantly over-represented in KD patients at FDR 0.2, and only two genes over-represented in febrile patients.

One of these representative genes was from a singleton cluster, and a contig identified as deriving from *Catonella morbi* (37/102 Febrile vs 7/115 KD; q=0.16). Another was from a cluster of 16 genes, with each source contig identified as deriving from *Moraxella catarrhalis* (31/102 febrile vs 2/115 KD; q=0.002); the representative gene shares 99% homology with the MscS mechanosensitive ion channel family gene of *M. catarrhalis*. *Catonella morbi* relative abundance has a significant crude association with febrile patients (coef=1.1, p=0.004) though reduced once age, gender and country are considered (coef=-0.77, p=0.04). *Moraxella catarrhalis* relative abundance has very little association with febrile patients (coef=0.28, p=0.58).

Discussion

The former two chapters focused on taxonomic composition of samples at species-level and above, for the purpose of descriptive analysis, contaminant identification and identification of species with differential abundance and prevalence in KD.

L. mirabilis and *A. defectiva* have significantly higher RA in the oropharyngeal microbiome of children with KD compared to febrile controls. It is highly relevant to explore whether there are genetic differences between the genomes of respective strains in different individuals which could influence susceptibility and pathogenesis. Such analyses can be broadly considered under the concept of *pangenomics*.

Pangenomics

Pangenomics and strain-resolved metagenomics are somewhat overlapping terms. Pangenomics is concerned with the total genetic content of a species, across strains, and resolving the specific structure of species by sample. Most frequently, this refers to presence or absence of orthologous gene groups, although some approaches consider sequence-level diversity (Iranzadeh & Mulder, 2019). As outlined in the introduction, pathogenic organisms sometimes comprise all members of a species, but sometimes only specific members harbouring one or more specific virulence factors. Segata (2018) argues that strain-level resolution is needed to achieve the full potential of comparative

metagenomics, since “many microbial phenotypes are strain specific” and strains are the “building blocks of a microbial community.” The possibility that strain-level genotypes and phenotypes matter in any microbial aetiology of KD is real, and therefore my metagenomic analysis must account for this.

A variety of methods exist to interrogate gene composition differences and sequence level diversity within species in metagenomes. Compositional approaches are exemplified by PanPhlAn, part of the bioBakery suite of metagenomic tools (Beghini et al., 2021). A PanPhlAn pangenome species reference comprises orthologous UniRef clusters of orthologous (predicted) genes (COGs) from reference genomes, with data on the composition of each reference, and ready-made indices for sequence alignment. The PanPhlAn pipeline aligns sample reads against all reference genomes and sums aligned reads by COG. Samples which provide enough depth of coverage to report composition are determined, and a matrix of COG presence by sample produced. StrainPhlAn exemplifies a sequence-based approach to strain typing, producing sequence alignments and phylogenetic trees of species-specific marker genes.

A limitation of both approaches is the modelling of a single dominant strain per sample – PanPhlAn adjudicates COGs as present or absent, but does not allow for heterogeneity with multiple distinct species genomes. Similarly, StrainPhlAn extracts a single consensus marker gene sequence for each sample. Nonetheless, StrainPhlAn analysis of gut metagenomes (Truong et al., 2017) supported the presence of dominant strains for a large majority of species. Whether this is replicated in the oral microbiome is unknown.

Approaches which consider multiple strains within a species have been developed. ConStrains (Luo et al., 2015) uses MetaPhlAn 2 bacterial marker gene reference sequences and detects and phases single nucleotide variants to identify and quantify strains within samples, however it has not been under active development since 2016. pStrain achieves the same and is compatible with MetaPhlAn 3 (Wang, Jiang & Li, 2020). DESMAN (Quince et al., 2017) extends this approach to MAGs from *de novo* assemblies. Most recently, STRONG (Quince et al., 2021) has been developed to resolve and phase strain-level variation directly upon the assembly graphs of co-assembled samples. Both of these approaches depend upon co-assembly, which is computationally intractable with a dataset of this size and standard HPC infrastructure. Finally, inStrain (Olm et al., 2021) detects and quantifies single nucleotide variants based on sequence alignments to reference genomes and/or dereplicated MAGs. Its utility is limited by the lack of phasing of variants.

Strain and pangenome analyses

StrainPhlAn and PanPhlAn were applied here because together they consider two important dimensions of strain-level variation within species, and are established, supported tools under active development.

The application of PanPhlAn to *A. defectiva* and *L. mirabilis* was limited by the paucity of reference genome sequences. Fortunately, it was still possible to generate pangenome references using sample-derived MAGs, with the attendant caveat that both the assembly and binning processes could include sequences from outside of the species. Indeed, based on contig classifications, the Abiotrophia MAGs were dominated by contigs assigned to a very closely related species, whose median RA modestly exceeds that of *A. defectiva*, but shares the same association with KD. It is unclear to what extent this genuinely reflects the presence of two closely-related species with the same biological association, or simply sharing of read and contig classifications from a single organism (per sample) between two closely related references.

The pangenome sizes (3 320 and 11 765 COGs respectively) are not unusual. In a study of bacterial pangenomes across a range of species, Maistrenko et al. (2020) identified core genome sizes from 443 to 5 946 genes and pangenomes from 959 to 17 739 genes, with a strong pangenome size dependence on the number of genomes sequenced.

Pangenome compositions showed no clustering by disease group, and COG-wise prevalence identified no genes significantly over-represented in KD patients, and only a small number of genes under-represented in the *Abiotrophia* pangenome. Those which were from identifiable COGs were labelled as originating from non-*Abiotrophia* species, thus the signals may be driven by differential abundance of other organisms.

StrainPhlAn analysis of marker gene consensus sequences showed no clustering of disease groups by PCA, and no significant differences by PERMANOVA.

Gene-level analysis

The above approach is limited to the consideration of candidate organisms, here selected on the basis of associations of RA with KD. However, this would be blind to a scenario where a coloniser is similarly abundant/prevalent in KD and febrile controls, but a toxin gene is over-represented in KD and has an aetiological role.

I sought to develop an approach which could identify over-represented genes (defined by sequence similarity) in KD vs febrile controls. This was algorithmically and computationally challenging, requiring *de novo* assembly of each sample and consideration of a vast number of predicted genes.

A quick and scalable method of clustering genes was required. This need was met by the linear time clustering algorithm of mmseqs2 (Steinegger & Söding, 2017). There remained the problem of determining presence of absence of gene clusters in samples. The presence of genes in assemblies could not be taken as a fair representation since *de novo* assembly is typically only possible for well-covered sequences. Sequence alignment against 2.1 million representative gene sequences (or the 17 million underlying genes) is intractable with present tools.

An alternative approach applied here was to identify unique *k*-mers within representative genes and then count the occurrence of these *k*-mers within sample reads. Such an approach has been developed and applied before in a tool named fastv (Chen et al., 2021). However, my own testing was discouraging, with unique *k*-mers undercounted and no response to an issue raised on GitHub.

For this reason, I developed a novel pipeline using jellyfish (Marçais & Kingsford, 2011) as a fast *k*-mer counter both to identify unique *k*-mers in representative genes and to count their occurrences in sample reads. It was surprising that representative genes could be detected in corresponding samples only 55% of the time –unique k-mers belonging to an assembled gene should also be present in the corresponding sample reads. Absences could be explained by a *k*-mer only occurring in parts on reads and only in full when assembled, or the *k*-mer being present with sequencing errors each time, never corresponding to the consensus sequence, and therefore going uncounted. However, both of these should be unlikely occurrences given the depth of coverage necessary for assembling sequences *de novo*. This deserves further investigation.

Despite adjusting for sample sequencing depth, no gene groups could be found with significant over-representation among KD patients.

Conclusions

In this chapter, I continued the search for differences in the KD microbiome at finer resolution, considering two candidate organisms whose RA is increased in KD, as well as developing a gene-based approach across organisms. The approaches leveraged the additional strengths of shotgun metagenomics over amplicon-based approaches because of the availability of sequences from whole bacterial genomes.

The two complementary approaches to candidate organisms presented in this chapter do not demonstrate plausible differences between strains of these two organisms.

The StrainPhlAn analysis suggests that there is no evolutionary distance between the core genomes of *A. defectiva* and *L. mirabilis* species in children with and without KD. The only significant results

from the PanPhlAn analysis may be confounded by inclusion of COGs from outside the genus No COGs were found to be over-represented in KD for either species.

Since acquisition of these organisms is likely to occur in infancy and predate KD for the majority of patients, the lack of significant differences by disease group is potentially unsurprising. There is no literature on the long-term strain-level stability of the oropharyngeal microbiome, though one study demonstrates stability of dental plaque strains in adults over three months (Utter, Mark Welch & Borisy, 2016).

In the absence of overt strain-level differences between these KD-associated organisms of the healthy microbiome, conclusions about any role they may play in KD are limited. I consider here a number of hypotheses:

1. **The increased RA of these organisms in KD is a consequence of disease or its treatment**. For example, pharyngeal inflammation or altered oral intake and dental hygiene driving increased RA of some bacteria. Additionally, for *L. mirabilis* I found evidence of an increased abundance with antibiotic exposure among these patients, though the mechanism is unclear. Although this is mostly independent of the relationship with KD, differences in characteristics of antibiotic exposures (e.g. timing, dosing and drug choice) could account for more of the relationship than currently observed.
2. **These organisms (though not their acquisition) are part of the causal pathway of KD**, with environmental or microbial factors triggering proliferation and a phenotype which triggers KD in susceptible individuals.
3. **Primary acquisition of these organisms triggers KD in a subset of susceptible infants.**
4. **Acquisition of secondary strains of these organisms triggers KD in a subset of susceptible individuals.**

The increased RA of these organisms in KD is not specific to the infant population, which provides evidence against hypothesis 3. Further, the coefficients obtained correspond to more than a doubling in the mean RA of these organisms in KD – if this occurred due to acquisition of a new strain, it would be the dominant strain and should be detected with the tools applied in this chapter. Investigating hypothesis 2 would require study of the biological activity of these organisms in the oropharynx, whether through transcriptomic or proteomic approaches.

The representative gene detection approach also did not show genes significantly over-represented in KD when considering all microbial genes. The approach developed clearly warrants further optimisation. An ideal approach would have identified unique *k*-mers from orthologous gene groups

rather than representative genes alone – where possible *k*-mers which are present in all members and no non-members. This would have required much more development work, and would not be as amenable to shell scripting solutions with jellyfish.

This chapter marks the end of the metagenomic analyses and a switch to the metaproteomics of ICs.

5 | Metaproteomics – analysis of pilot data

Introduction

As described in the introduction, IC provide a tantalising window into the aetiology and pathogenesis of KD. In my project, I am primarily interested in the potential carriage of microbial antigens within IC, which could help to explain the aetiology of the disease.

This work commenced years before my fellowship and involved my colleague, Dr Stephanie Menikou, in painstaking experiments to test and compare techniques for purifying IC from serum and plasma, and preliminary proteomic experiments with artificially generated ICs (Menikou, 2016).

Aside from the experimental challenge of purifying ICs at scale, multiple challenges could be anticipated in metaproteomic analysis. These include the challenge of searching large metaproteomic databases while maintaining power to confidently detect valid peptide and protein matches and managing the computational cost.

In this chapter I undertake detailed analysis of these proteomic data to describe the constituents of IC-enriched fractions, and test a novel database-reduction strategy for improved metaproteomic searching. The first part of this has been published as part of a recent manuscript (Menikou et al., 2020).

Objectives are:

1. Describe protein and immunoglobulin content of purified IC
2. Quantify the specific detection and abundance of spiked influenza proteins in purified IC

Methods

Sample collection and laboratory methods

Laboratory experimental work was designed and undertaken by Dr Menikou, comparing two methods of immune complex extraction.

For testing size exclusion chromatography and affinity purification (SEC-AP), serum from seven healthy adult donors (Central Office for Research Ethics Committees, Imperial College Healthcare NHS Trust, reference number REC 12/WA/0196, Imperial College Healthcare Tissue Bank project R13062) was obtained. All volunteers had been routinely vaccinated with influenza vaccine Split Virion BP 2014/2015 (strains: A/California/7/2009(H1N1), A/Texas/50/2012(H3N2) and B/Massachusetts/2/2012 from Sanofi Pasteur) 6–10 weeks earlier. Ten patients with KD recruited within 7 days of onset of fever were recruited at University of California San Diego (UCSD) with

parental consent (as described in Appendix A), with acute and convalescent serum combined in equal ratios.

For testing polyethylene glycol (PEG) precipitation, serum was obtained from two healthy adults immunised in the previous six months with the influenza vaccine Split Virion BP 2015/2016 (strains: A/California/7/2009(H1N1) virus, A/Switzerland/9715293/2013(H3N2) virus, B/Phuket/3073/2013 virus from Sanofi Pasteur).

In vitro ICs were created from healthy serum by adding 40 µl of the corresponding influenza vaccine to 160 µl of serum. Serum and influenza vaccine were mixed, incubated at 37°C for 1 hour to allow antibody-antigen binding, and then processed by SEC-AP on Protein G columns, or by PEG precipitation. KD samples had identical volumes of phosphate-buffered saline added and underwent the same processing. For PEG precipitation, healthy samples were processed twice, with influenza vaccine or saline added.

SEC-AP resulted in two high-molecular weight (HMW) fractions, with the first (HMW 1) corresponding to IgM and its complexes and second (HMW 2) to IgG. Each fraction was separated into a protein G eluant and wash-through. The process is illustrated in Figure 16.

Following protein quantification, the SEC-AP purified and PEG precipitated samples underwent in-gel digestion and liquid chromatography tandem mass spectrometry (LC-MS/MS) at the University of Bristol Proteomics Facility on an Orbitrap Elite instrument. Precursors (MS1) were analysed on the Orbitrap and fragments (MS2) on the ion trap.

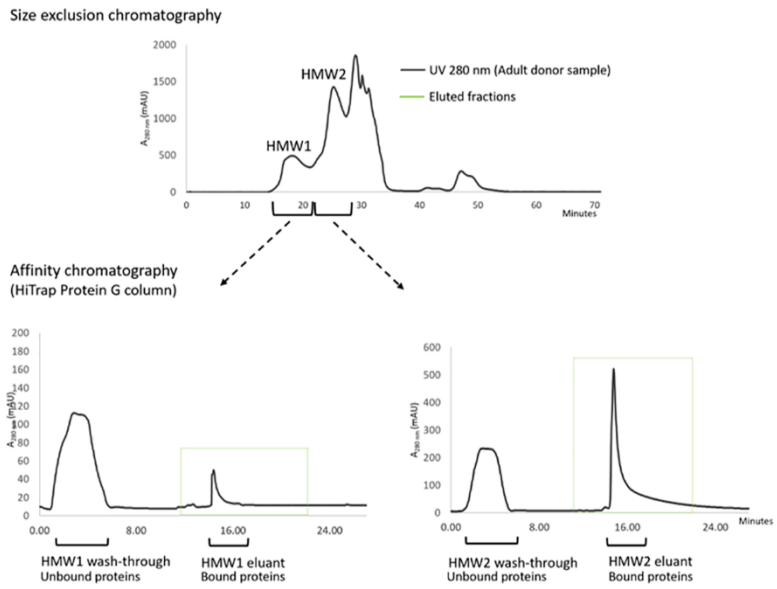

Figure 16 Fractionation of precipitated immune complexes by size-exclusion chromatography and affinity purification with Protein G. Reproduced from Menikou et al. (2020).

Descriptive bioinformatic analysis

Mass spectrometry data were analysed using MaxQuant 1.6.6.10 (Tyanova, Temu & Cox, 2016). Spectra were searched against the UniProt human proteome (February 2019), immunoglobulin sequences (January 2018 Kabat and NCBI databases) from the abYisis database (Prof Andrew Martin, Personal Communication and Swindells et al., [2017]), and MaxQuant's standard contaminant database. Trypsin was selected as the enzyme with full specificity and up to two missed cleavages. Peptide precursor mass tolerance was set at 10 ppm, and MS/MS tolerance was set at 0.5 Da. Search criteria included carbamidomethylation of cysteine (+57.0214) as a fixed modification and oxidation of methionine (+15.9949) as a variable modification.

Intensity based absolute quantification (iBAQ) was enabled and match-between-runs activated with default settings. False discovery rate (FDR) was set at 0.01 for peptides. FDR was set at 1 for proteins to prevent exclusion of immunoglobulin variable peptides which would be distributed between many protein sequences.

Downstream analysis was undertaken in R for Windows 3.5.1. Full gene names for UniProt proteins were determined by InterMineR. Individual proteins were classified using string matching as keratin, trypsin, complement, albumin, fibrinogen, immunoglobulin, other, contaminant and influenza. Individual peptides were then classified according to their corresponding protein(s). Where multiple classes matched, the first in the list preceding was selected.

iBAQ normalises total MS1 intensity per protein based on the number of theoretical tryptic peptides, approximating molar ratios of proteins within samples. I sought to calculate mass ratios and recalculate immunoglobulin results using constant region peptides only (due to anticipated more reliable detection). I therefore calculated modified iBAQ-based values at peptide level.

Immunoglobulin peptides were classed as belonging to the constant region based on edit distance of two or below to a UniProt human constant immunoglobulin sequence (approximate Levenshtein distance algorithm as implemented in the base R function adist).

For each protein, including immunoglobulin constant sequences, the number of tryptic peptides per kDa was calculated. Normalised peptide intensities were calculated by dividing the original intensities by the corresponding ratios. Non-constant region immunoglobulin peptide intensities were set to zero. Normalised constant immunoglobulin peptide intensities were then multiplied by the ratio between the molecular weight (MW) of the (sub)class and its constant region. Where a peptide corresponded to multiple immunoglobulin subclasses, the mean of each correction was taken.

Normalised intensities were summarised by class and sample, with the total for each sample normalised to correspond to the proportion of MS1 intensity belonging to identified features. In this way, differences in identification rates between samples remain informative.

Database reduction testing

MetaNovo (Potgieter et al., 2019) is a database reduction tool designed for metaproteomics as part of a two-stage approach. It uses sequence tags generated from fragment spectra to search a sequence database, respecting enzymatic cleavage rules. Proteins are ranked according to joint protein and organism posterior probabilities based on size-normalised match counts. Proteins are sequentially inspected, and all those which match at least one unmatched spectrum are included in a final reduced database. This can then be used with any conventional proteomic search engine (though x!tandem is included within the default workflow).

I constructed a metaproteomic database comprising UniProt reference proteomes for bacteria, fungi, viruses and archaea, the UniProt human proteome, and influenza vaccine sequences for the strains included in the Northern hemisphere in 2014-15 and 2015-16 as detailed earlier. The database was 9.7 billion amino acids in size, comprising 28 045 072 proteins from 12 443 microbial organisms, and *Homo sapiens*.

Serum from two additional healthy individuals who had been immunized with the Split Virion BP 2015/2016 vaccine (A/California/7/2009 (H1N1)pdm09, A/Switzerland/971593/2013(H3N2), B/Phuket/3073/2013) at most 6 months prior to serum collection was used for PEG precipitation experiments, with addition of 40 µl influenza vaccine or PBS. These samples underwent proteomic analysis as part of the first batch of febrile and KD samples (as described in the subsequent chapter).

The metaproteomic database was reduced using MetaNovo and searched against using MaxQuant 1.6.17

Results

Protein classes

The proportion of MS1 intensity belonging to identified features is lowest (around 30%) in the samples with the highest proportion of immunoglobulin (Control & V HMW2 eluant and Febrile HMW2 eluant

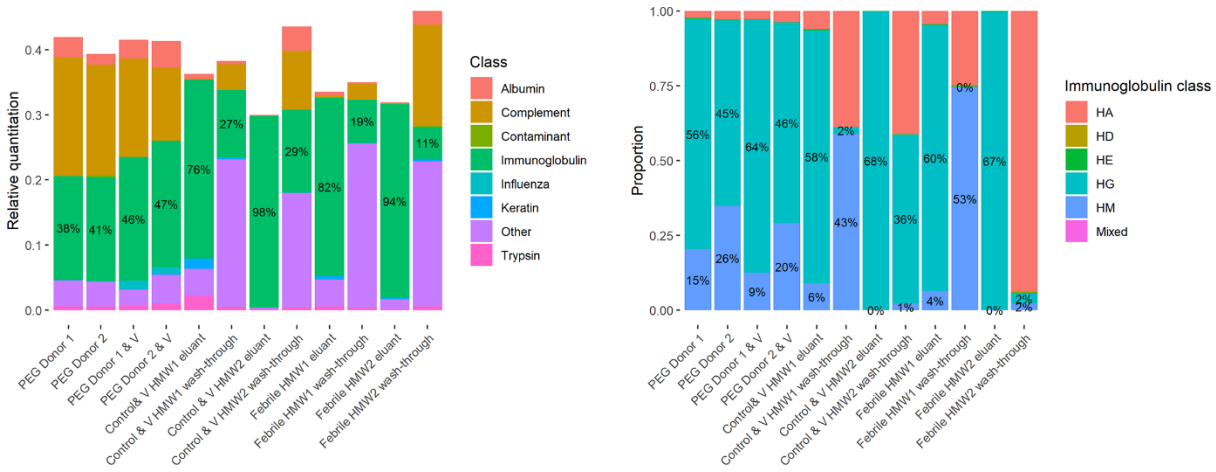

Figure 17 Quantification of different protein classes and identification of immunoglobulins. V indicates samples to which influenza vaccine was added. Febrile samples are not spiked with vaccine. (Left) Relative quantities of different classes of protein as calculated through a modified iBAQ-based approach. The total height of each bar is the proportion of MS1 intensity belonging to each identified feature. Samples come from two donors (Donor 1 and Donor 2). (Right) Proportion of immunoglobulin mass by class as estimated using a modified iBAQ-based approach and extrapolated from constant peptides only. HA: IgA heavy chain, HD: IgD heavy chain, HE: IgE heavy chain, HG: IgG heavy chain and HM: IgM heavy chain; iBAQ: intensity-based absolute quantification; MS: mass spectrometry.

columns in Figure 17 left), which is expected due to the limited feature identification expected for variable immunoglobulin peptides.

Although no serum comparison is included here, results could be compared to laboratory reference ranges. The normal range of serum albumin concentration is 3.5-5 g dL^{-1} and of immunoglobulin is 2.3-3.4 g dL^{-1} (Anon, 2021). In contrast, modified iBAQ estimates in PEG precipitates were 5-10 times greater for immunoglobulin than albumin.

Alpha-2-macroglobulin is a very large serum protein with molecular weight similar to IgM (725 kDa vs. ~900 kDa) and serum concentration of around 270 mg/dL, around ten times lower than immunoglobulin (Yoshino et al., 2019). Modified iBAQ estimates in PEG-precipitates were 100-300 times lower, suggesting alpha-2-macroglobulin was much less preferentially precipitated.

Complement content

The complement fraction in PEG samples is dominated by three proteins which contribute 60-70% of modified iBAQ estimates: Complement Factors 3 (C3; 185 kDa) and 4 (C4; 203 kDa), and Complement Factor 4 Binding Protein A (CF4BPA; 570 kDa). These three proteins typically have a total serum concentration of around 1.5 g/L, around half the concentration of immunoglobulin. The total modified iBAQ estimates in PEG precipitates were 40-74% relative to immunoglobulin, suggesting these components were precipitated in a similar manner.

Within SEC-AP samples, the largest quantities of complement proteins were seen in the HMW2 wash-throughs, and dominated by C3 and C4 (81 and 83% in each wash-through). Complement was not

retained well in eluant, suggesting very little stable complex formation with immunoglobulins, even in the presence of viral antigen.

Immunoglobulin content

The highest proportion of immunoglobulin is identified in the HMW2 peak from SEC that was eluted from the affinity column (94% and 98%). The unbound fractions have low proportions of IgG (11% and 29%) indicating that the HiTrap Protein G column efficiently bound IgG. Complement effectively passed into the wash-through (0.2% and 0.5% in eluant vs. 21% and 34% in wash-through).

The HMW1 eluants, which should be enriched for IgM and ICs also had high proportions of immunoglobulin (76% and 82%), and the wash-throughs were similarly depleted (19% and 27%). Complement components were in low abundance overall and mostly passed into the eluant (0.4% and 0.9% vs. 7% and 10%).

The PEG precipitates contained intermediate proportions of immunoglobulin (38–47%) with complement being the other major constituent (27–44%). The proportion of immunoglobulin was increased in the vaccine-spiked samples compared to controls (46% and 47% vs. 38% and 41%).

Immunoglobulin classes and subclasses

IgG dominated in the HMW2 eluants (99.4% and 99.7% of immunoglobulin). It was nearly absent in the febrile wash-through (2%), though in the influenza spiked wash-through a larger proportion of immunoglobulin was IgG (36%), suggesting that the column may have been saturated (Figure 17 right). The remainder in both cases was predominantly IgM. IgG also dominated in the HMW1 eluants (85% and 89%), with IgM (6% and 9%), and IgA (4% and 6%) present. The wash-throughs were dominated by IgM and IgA.

The identification of IgG subclasses (HG1, HG2, HG3, HG4) was further explored (Figure 18). Some peptides could not be uniquely assigned to a subclass (accounting for 51–80% of estimated mass). IgG1, as expected, was in the greatest abundance across all samples. Notably, in the PEG precipitates the proportion of IgG1 was increased in the influenza vaccine spiked samples vs the PEG Donor 1 and 2 samples that were not spiked (83% and 90% vs. 68% and 58% respectively).

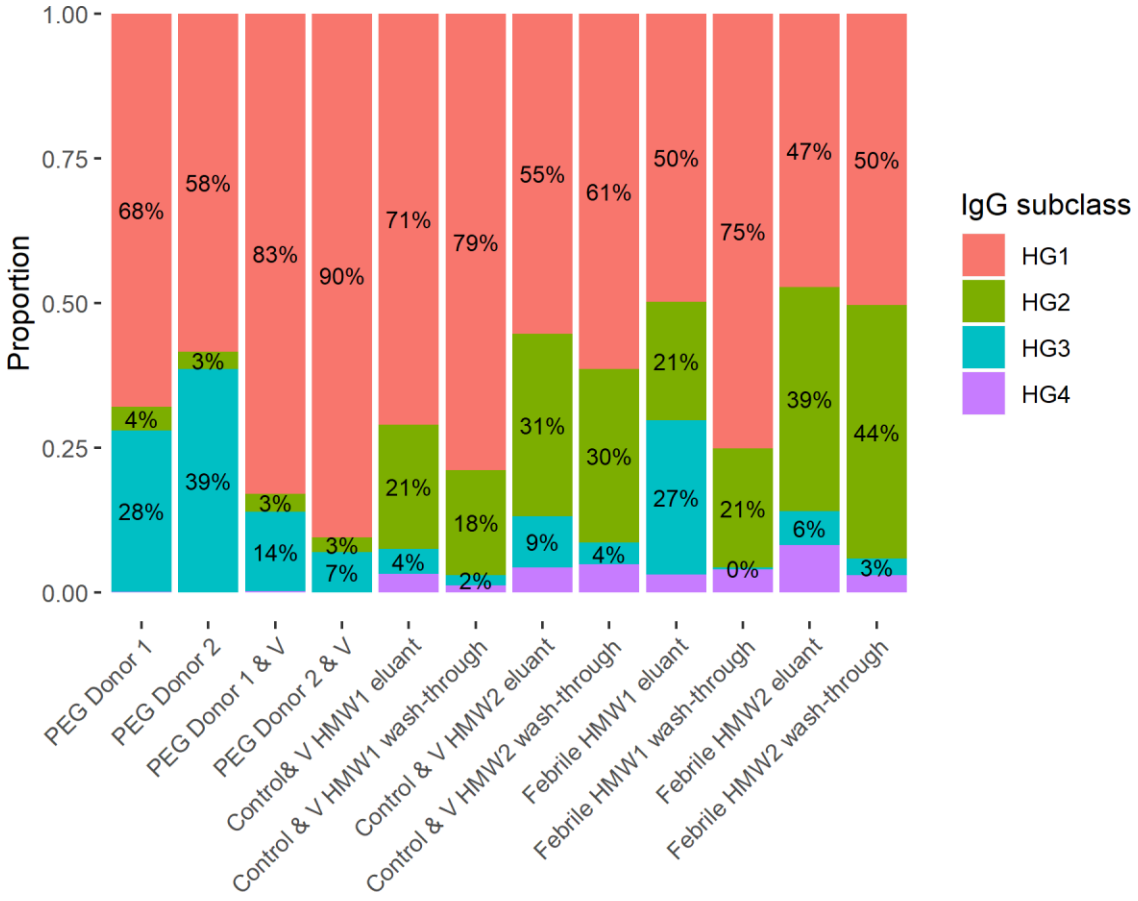

Figure 18 Proportion of estimated IgG mass by subclass (excluding contributions by non-specific peptides). PEG precipitated samples come from two donors (D1 and D2) and V indicates samples to which influenza vaccine was added. Febrile samples were not spiked with vaccine. HG1: heavy chain of IgG subclass 1, HG2: heavy chain of IgG subclass 2, HG3: heavy chain of IgG subclass 3 and HG4: heavy chain of IgG subclass 4; PEG: polyethylene glycol.

Identification of Influenza proteins.

Influenza peptides were detected in all vaccine-spiked samples (Figure 19), the greatest number of spectra being in the PEG precipitates. More spectra mapping to influenza peptides were identified in the eluant of the HMW1 fraction. The estimated mass contribution of influenza protein was also higher in the spiked PEG samples (2.4% and 3.2%) than the HMW1 eluant (0.02%). The five spectra that matched to influenza in unspiked samples corresponded to high abundance influenza peptides from spiked samples, suggesting some potential sample crossover.

Within the SEC-AP samples, vaccine strain proteins were selected as the leading razor protein by MaxQuant for 15 of the 57 PSMs (A/California/07/2009 and A/Texas/50/2012). Within the PEG samples, vaccine strain proteins comprised 25 of 38 PSMs (B/Phuket/3073/2013 and A/California/07/2009).

Database reduction

Using LC-MS/MS data from two individuals' samples, with and without spiked influenza vaccine, MetaNovo reduced the 9.7-billion amino acid metaproteomic database to 41 million amino acids

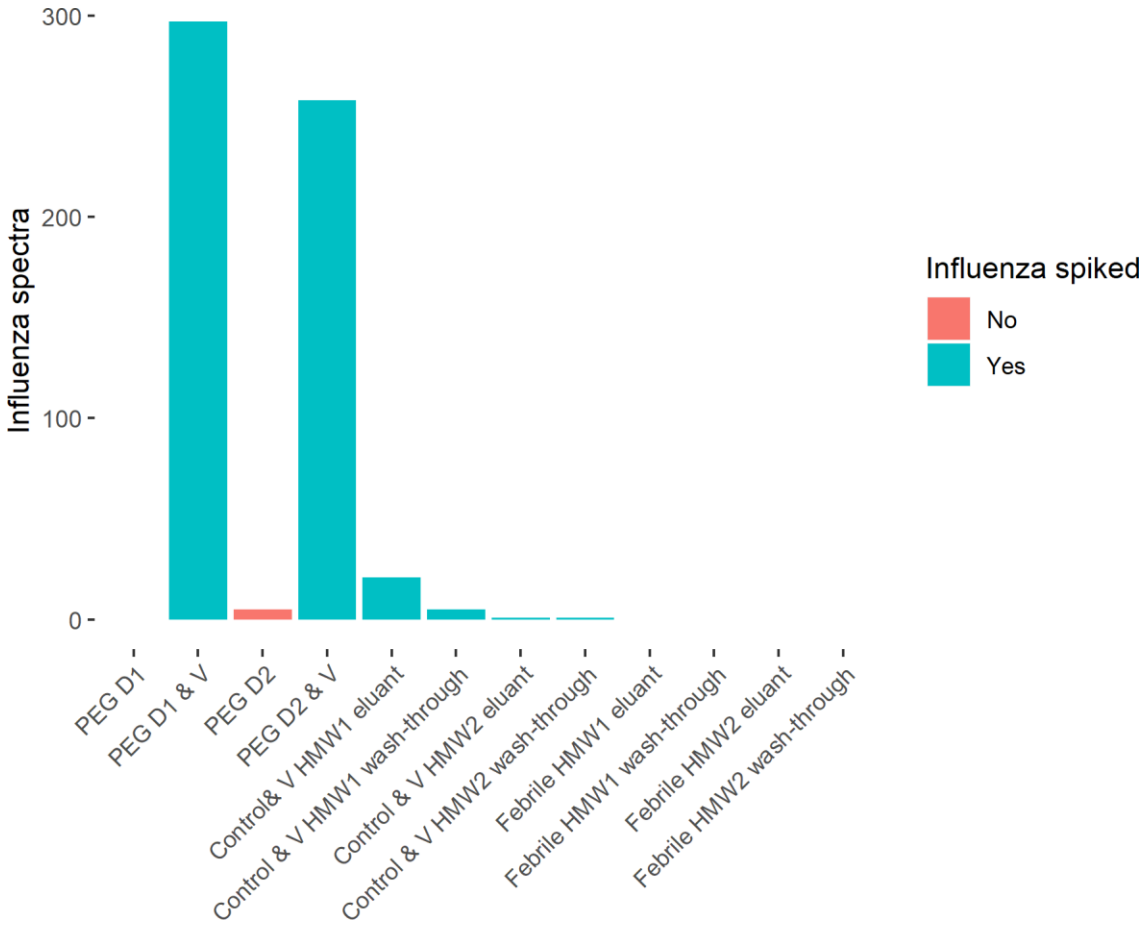

Figure 19 Number of MS2 spectra assigned to influenza peptides in each sample. MS1 features matched by run are excluded. Samples come from two donors (D1 and D2). V indicates samples to which influenza vaccine was added. SEC-AP V HMW1 eluant, wash-through, and SEC-AP HMW2 eluant and wash-through are the control samples spiked with vaccine. SEC-AP Febrile HMW1 eluant, wash-through and SEC-AP Febrile HMW2 eluant and wash-through are not spiked with vaccine. MS: mass spectrometry; SEC-AP: size exclusion chromatography affinity purification; HMW: high molecular weight.

(0.4% of the original database). Human proteins were enriched, increasing from 0.07% to 1.2% of protein entries (1 211 of 20 585 proteins retained). 18 of 128 influenza proteins were incorporated. 97 442 proteins were incorporated from other organisms. The largest number of proteins from a single organism was 597 from a strain of *Rhizophagus irregularis*.

Following search with MaxQuant using a taxonomy-restricted database comprising only human and influenza sequences, 239 influenza PSMs were identified at PSM and protein FDR of 1% (with an additional 43 MS1 features matched by run). Using the metaproteomic database reduced by metanovo, 421 influenza PSMs were identified (and an additional 77 features matched by run). This improvement was mostly due to FDR control at protein level, since without applying the protein FDR the reduced database identified only 9 extra PSMs (423 vs. 414).

Without protein FDR control, the taxonomy-restricted database gave 55 248 human PSMs, falling to 36 691 with protein FDR control. The reduced database gave 43 523 non-influenza PSMs, only falling

a small amount with protein FDR control, to 43 460. There were 1 269 PSMs to other species, with 267 passing protein FDR control. These proteins comprised 51 genera, with only Streptomyces, Aspergillus, Clostridium, Colletotrichum and Escherichia contributing two or more individual proteins. No *Rhizophagus irregularis* PSMs were reported.

Discussion

IC extraction and antigen identification

The pipeline of SEC followed by affinity purification and LC-MS/MS led to successful recovery and identification of spiked influenza peptides. Reassuringly, influenza peptides were not identified in the non-spiked febrile controls aside from five peptides in one fraction, which may represent carry-over from the spiked sample.

In SEC-AP samples, the recovery of influenza vaccine antigen preferentially in HMW1 eluant supports their presence within large immune complexes, since Protein G specifically binds IgG, whose molecular weight is not within the HMW1 range. The far higher recovery of influenza peptides by PEG precipitation does not definitively support the superiority of the process for recovering antigen within ICs, as this may represent non-specific purification of unbound influenza virions. Nonetheless, in Western blotting experiments, PEG was not shown to recover influenza virions alone (Menikou et al., 2020).

PEG appears to preferentially precipitate immunoglobulin and complement regardless of the presence of ICs. This likely relates to their high molecular weight – though it is interesting that the very high molecular weight alpha-2 macroglobulin does not appear to be preferentially precipitated. Complement Factor 3, when hydrolysed, binds to immunoglobulin (as C3b). Complement Factor 4 (isoform A; C4A) can be covalently bound to ICs and protein antigens. However, there is no indication that here that they were precipitated in complex with immunoglobulin. Notably, in the SEC-AP experiments, complement is enriched in the HMW2 wash-through, and not retained in the eluant.

The addition of influenza vaccine resulted in greater precipitation of immunoglobulin and increased proportion of IgG1. This is consistent with the formation of ICs, since it is known that the response to IIV (inactivated influenza vaccine) mostly comprises the IgG1 subclass of antibodies (Manenti et al., 2017; El-Madhun, Cox & Haaheim, 1999; Pedersen et al., 2014).

Background to metaproteomic searching

The specific recovery and identification of protein antigen from a single species within human proteomic samples using a targeted database is encouraging. However, it does not represent the

challenge of identifying potential protein antigens from unknown organisms within a large number of samples.

Metaproteomics is the study of proteins from communities of organisms. The development and application of techniques began with environmental communities (Wilmes & Bond, 2004; Kan et al., 2005) and this remains a major focus of application. However, quickly techniques were applied to the study of host-associated microbiomes, principally the gut (Klaassens, Vos & Vaughan, 2007).

Shotgun mass spectrometry-based metaproteomics faces a number of challenges above proteomics as applied to single organisms. The field itself remains a niche, with PubMed containing 904 articles containing “metaproteom*” since first occurrence in 2004. Challenges have been well-reviewed by others (Muth et al., 2013, 2015; Muth, Renard & Martens, 2016; Heyer et al., 2017) and are summarised here. The protein sequence databases used must provide good coverage of the diversity of proteins in the sample. Since protein sequence repositories may represent only a small fraction of the proteins in a given sample, optimal analysis frequently requires sample- or experiment-specific databases of predicted proteins derived from metagenomic sequencing. These databases can be very large, and comprise many proteins which will likely never be observed (due to incorrect predictions, lack of expression or low abundance).

Regardless of the source of sequences, metaproteomic searches typically employ protein sequence databases orders of magnitude larger than conventional proteomic databases. The magnitude of computation required increases with the size of the database since each spectrum acquired must be compared with predicted spectra derived from the entire sequence database to obtain PSMs.

A further, and more significant challenge, is the control of false-discovery rate (FDR), which becomes more challenging as databases grow (Jagtap et al., 2013). Mass spectra are typically identified based on the thresholding of best-scoring matches to predicted spectra. Scores are calculated in different ways between software packages, but encompass some measure of how closely the experimental and theoretical spectra match.

Target-decoy based approaches have become the most commonly applied type of thresholding for PSM and protein identifications, since these methods aim to control the FDR (Jeong, Kim & Bandeira, 2012). They depend upon searching the experimental spectra against both the target database and a decoy database. The decoy database is most frequently a sequence-reversed target database with preservation of cleavage sites; it comprises proteins which do not exist, and whose tryptic peptides should therefore be unobserved. Matches between experimental spectra and decoy peptides are thus taken to represent false matches. The distribution of scores against the target and decoy databases

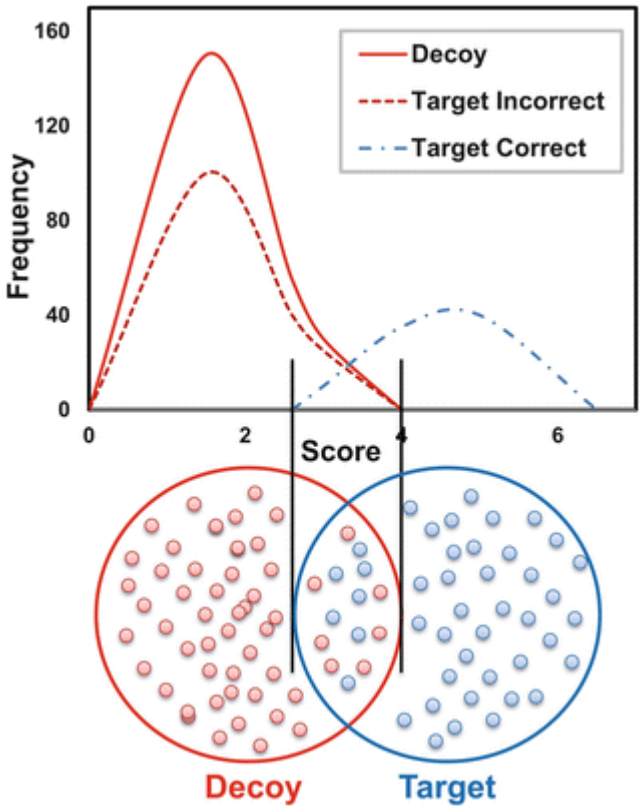

Figure 20 Hypothetical Illustration of the overlapping score distributions from the target-decoy database and how this can be used to estimate the false discovery rate at given score thresholds (Figure reproduced from Aggarwal & Yadav, 2016)

can be used to determine a threshold which achieves a given FDR (most frequently 1%) for PSMs. The process is illustrated graphically in Figure 20.

Since false discoveries are chance events, the number of false discoveries will grow as the database size increases. Simultaneously, the decoy score distribution may include higher scores, due to the competitive selection of best-scoring matches. If the yield of target matches does not also grow commensurately, then the FDR will increase at a given score threshold and conversely the number of identifications will decrease at a given FDR (see Figure 21).

As expected, Tanca et al. (2013) showed in a mock microbial community that a taxonomy-restricted database provided more PSMs at a given FDR. Similarly, Rechenberger et al. (2019) in a large study of human stool metaproteomes, showed over a ten-fold reduction in PSMs at the same FDR when replacing the SwissProt bacterial database with the more comprehensive Integrated Gene Catalog of the human stool microbiome. In contrast, when re-scoring PSMs with a machine learning tool (Percolator) there was a near 7-fold increase in identifications at the same FDR.

Myriad approaches have been proposed for overcoming the computational and inferential challenges of metaproteomic searching. Some have focused on targeted database restriction on a per-experiment or per-sample basis, for example through taxonomy restriction (based on amplicon

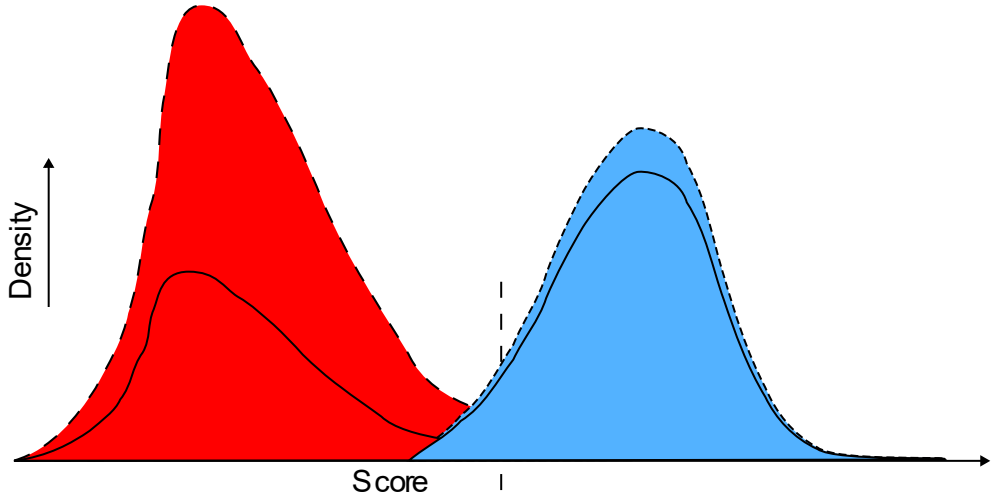

Figure 21 Hypothetical illustration of the score distribution of target and decoy matches with two databases (blue and red respectively). The first database (bold fill and solid outline) is small but contains a large proportion of potential target matches. The second database (lighter fill and dotted outline) is larger. Decoy matches grow proportionately, with some additional growth in the right tail. Target matches grow by a smaller proportion since most potential target matches have already been identified. The dotted vertical line indicates a fixed score threshold. For database 1, the FDR is low at the given threshold. For database 2, the FDR is higher. FDR: false discovery rate.

sequencing, for example) (Xiao et al., 2018) or use of predicted open-reading frames (ORFs) rather than 6 frame translations (Tanca et al., 2016). Others have advocated for the potential benefits of more advanced machine-learning-based PSM rescoring approaches, utilising for example fragment intensity and/or retention time predictions (Verbruggen et al., 2021; Gessulat et al., 2019). *De novo* sequencing based approaches have also been promoted, since they dispense with the need for database searching by spectral matching, though additional problems arise, including the challenge of confidently identifying error-prone *de novo* sequences (Muth et al., 2013, 2015; Muth, Renard & Martens, 2016) and availability of only one tool to estimate FDR, which does not appear to be under active development (Leprevost et al., 2014).

Two-stage approaches combine a first search for database restriction followed by a second search to identify PSMs. Jagtap et al. (2013) presented encouraging results showing two-stage approaches can preserve similar numbers of host-associated PSMs to a host-only single-stage database search, while identifying more bacterial PSMs than a single-stage approach with a comprehensive database. Two-stage approaches have been implemented in several metaproteomic search engines (Cheng et al., 2017; Zhang et al., 2016; Muth et al., 2018; Potgieter et al., 2019).

Metaproteomic application to these data

The nature of the samples being processed and the aims in this project differ vastly from those of most published metaproteomic studies. Metaproteomic studies generally seek to functionally characterise communities, whether free-living or host associated, based on their protein content. The samples

utilised are typically dominated by the organisms of interest, even if host proteins are also present. The species present are typically well known, through amplicon or metagenomic sequencing.

In contrast, I seek to identify a potentially small number of organisms across many samples based on the likely very low amounts of protein antigen which may be captured within precipitated ICs. My samples are not only dominated by host protein, but also enriched for antibodies, and therefore the vast sequence diversity of the antibody variable region. Although a hypothesis-driven approach could focus on a small number of candidates, there is little justification for excluding any micro-organisms from a primary approach and the starting sequence database is necessarily large. This search for low abundance microbial proteins could be called “minority metaproteomics.”

Following expert recommendation (Prof D Tabb, personal communication), I elected to test a two-stage database reduction approach utilising Metanovo, a tool being developed under his supervision (Potgieter et al., 2019). It was encouraging to find that Metanovo could reduce the large UniProt reference proteome database amino acid size by a factor of around 250, and concentrate human proteins by a factor of 17. The preservation of 17 influenza proteins in the reduced database demonstrated the success of the approach.

The inclusion of such a large number of proteins from a single strain of *Rhizophagus*, a fungus of plant roots, is striking (Morin et al., 2019). It should be noted that *Rhizophagus* species have large genomes (over 100 Mbp) and the four strains included among UniRef sequences possess between 25 and 43 000 predicted proteins, compared to 1 500 to 3 000 for Streptococcal species. The included proteins represent 1.4% of those in the original database.

The final test of Metanovo was to compare the performance of its reduced database to a taxonomy-restricted database containing only human and influenza sequences. I observed comparable detection of influenza PSMs with only PSM FDR control, and much improved detection with more stringent protein FDR control.

Conclusions

Both SEC-AP and PEG precipitation can concentrate antibody and corresponding antigen. Influenza antigens can be identified by proteomic database searching with high specificity. Metanovo allows a non-taxonomy restricted database to be reduced, allowing effective proteomic analyses with stringent FDR control.

In the next chapter, I will present the analysis of the primary study data from children with KD and febrile controls using Metanovo, whilst also piloting another complementary approach.

6 | Metaproteomics – analysis of study data

Introduction

The prior chapter described metaproteomic method development using pilot data. In this chapter, I analyse data from a large cohort of children with KD and three sets of controls in order to identify potential microbial protein antigens within immune complex-enriched PEG-precipitated samples, and associations with KD.

I add a novel "spectra-first" approach to the data, using Quandenser (The & Käll, 2020). This tool reverses the typical approach of proteomic searches, which seek first to identify peptides represented by MS1 spectra, then quantify peptides and proteins. Instead, Quandenser identifies, quantifies and matches MS1 peptide features across all samples. MS2 fragment spectra are clustered, and matched peptide features undergo FDR control based on shared fragment clusters.

Downstream processes can then be applied to filter features based on differential abundance or prevalence and extract consensus spectra from corresponding fragment spectral clusters. These higher-quality consensus spectra can be subjected to identification by database searches or *de novo* sequence generation. Clustering and filtering reduces the number of spectra which need to be searched.

All laboratory method development and work was undertaken by Dr Menikou.

Objectives are:

1. Generate a reduced metagenomic database suitable for proteomic searching
2. Identify database peptides and proteins in the samples and corresponding taxa
3. Identify taxa with differential prevalence between KD and controls
4. Filter peptide features with differential prevalence between KD and controls and identify corresponding consensus spectra

Methods

Patients and controls

The studies contributing samples to both the metaproteomic and earlier metagenomic analyses are described fully in Appendix A (p180) along with details of ethical approval.

Prior to my involvement in the project, samples from children with KD and febrile controls were selected in three batches from UCSD, where a long-term cohort study of KD is ongoing with extensive clinical data and sample collection. The American Heart Association guidance was used to diagnose

KD (McCrindle et al., 2017). As described in the Appendix, febrile controls were adjudicated not to have KD, but did have at least one of the KD clinical criteria.

Healthy children were recruited from outpatient clinics and as unrelated contacts of children with meningococcal disease at St Mary's Hospital, London. Adults with TB were recruited from the multi-centre IDEA study.

Samples

Immune complexes are known to be detected in KD from as early as 10 days into the illness, however most patients present and are treated earlier than this (Levin et al., 1985). Samples earlier than 10 days of illness (acute) may not contain ICs, though microbial antigen may be present. These samples were mixed in equal ratios with paired convalescent samples, which would potentially contain host-derived antibodies to microbial agents implicated in the cause of KD, as well as pooled immunoglobulin from IVIg administered during the illness. This was anticipated to promote the formation of ICs. Samples from patients presenting later than 10 days are more likely to contain circulating ICs, and were used without mixing.

Both serum and plasma samples were used, due to uncertainty about potential limitations of each. Proteolytic enzymes are released and platelets activated during the clotting process in serum samples, and these could sequester ICs. Using plasma would avoid this limitation, but adds the disadvantage of retaining a large mass of clotting factors, which will occupy the mass spectrometer during analysis.

Febrile controls were similarly classified as acute and subacute. However, due to the limited availability of convalescent samples from children with febrile illnesses, pooled convalescent samples were used for mixing with acute samples.

Serum was recovered from venous blood of 31 healthy children. Samples from 162 patients with KD and febrile controls were obtained (Table 15).

Sample preparation

Patients' and healthy controls' whole blood was collected in serum-separating tube II advance bottles (Becton Dickinson), kept at room temperature (RT) for 30 minutes and centrifuged at 4° C at 3000 g for 10 minutes for serum separation and recovery.

Isolation of ICs from human sera

Equal 200 µl volumes of serum and 6% PEG 6000 (BDH) were mixed and dissolved in borate buffer comprising 100 mM boric acid (Sigma Aldrich), 75 mM NaCl (BDH), 25 mM sodium tetraborate (Sigma Aldrich), and incubated statically at 4° C overnight. The resulting precipitate was centrifuged at 2000 g

for 20 minutes at 4° C. After decanting the supernatant, precipitates were washed once with a 400 µl PEG solution of the same final concentration (3%). Supernatant was decanted and the pellet dissolved in 200 µl of 1× phosphate buffer solution (PBS) (Sigma Aldrich) which was left to stand at 4° C overnight and stored at -80° C until analysed.

Protein sequencing

For total protein quantification of the precipitated samples, bicinchoninic acid (BCA) protein assays was used (ThermoFisher Scientific). Following protein quantification, the proteins were digested at 70° C (denaturing condition) by thermostable trypsin using the SMART digestion kit (ThermoFisher Scientific). After peptide clean up, 0.5 µg peptide was injected into an Orbitrap LC-MS/MS (Lumos Fusion, Thermo Fisher scientific) at the Oxford proteomic facility. Samples were ordered according to disease group (febrile ahead of KD group). To avoid bias by system error, the performance of LC-MS/MS (sensitivity of detection, retention time shifting, number of proteins identified) were monitored by running a quality control sample (pool of all individual samples) every 10 samples within batch. A blank sample was processed between each run to avoid sample carryover.

An initial subset of samples (N=54) were sequenced by LC-MS/MS at the University of Bristol Proteomics Facility (UK) by in-gel digestion instead of in-solution digest (Calvopiña et al., 2017).

Database reduction

Metanovo was reimplemented as a nextflow pipeline (Di Tommaso et al., 2017), allowing much greater flexibility in parallelisation of tasks within nodes, and also allowing for parallelisation across multiple nodes. This was necessary because of bottlenecks which were encountered in writing to the sqlite database when scaling metanovo to hundreds of samples. During implementation, the Compomics PeptideMapping API was upgraded to a more recent version which handles unknown amino acids in reference sequences.

In view of the narrower fragment mass tolerances in the first batch of samples analysed in Oxford, these were not used in the metanovo database reduction step. Instead, results from batch 1 samples analysed at Bristol, and batch 2 samples analysed at Oxford were run together.

The starting database comprised the human and UniProt reference proteomes as described in Chapter 6.

Peptide and protein identification

Mass spectrometry data were analysed using MaxQuant 1.6.10.43 (Tyanova, Temu & Cox, 2016) on the Imperial HPC environment. Data from Bristol and the second batch processed in Oxford were analysed separately. Spectra were searched against the reduced database and MaxQuant's standard

contaminant database. Trypsin was selected as the enzyme with full specificity and up to two missed cleavages. Peptide precursor mass tolerance was set at 10 ppm, and MS/MS tolerance was set at 0.5 Da. Search criteria included carbamidomethylation of cysteine (+57.0214) as a fixed modification, with oxidation of methionine (+15.9949) and N-terminal carbamylation (+43.0058) as variable modifications.

Intensity based absolute quantification (iBAQ) was enabled and match-between-runs activated with default settings. False discovery rate (FDR) was set at 0.01 for peptide spectrum matches (PSM). FDR was set at 1 for proteins, allowing the effect of protein FDR control to be explored subsequently.

Analysis

The reduced database was described by contributions from human proteins and proteins from other species. Abundant human and contaminant proteins identified by MaxQuant with PSM and protein level FDR control were described.

Non-human non-contaminant proteins were annotated by species and genus-level annotations, and the overall contribution by species and genus was summarised.

Prevalence of any proteins from each genus and species within sample groups were explored by group using Fisher's exact test. FDR was controlled with the method of Benjamini-Hochberg.

Quandenser

Raw files were converted to mzML format using msconvert (Adusumilli & Mallick, 2017) and Quandenser (v0.02) was run with default settings except for preventing exclusion of rare features (max-missing=500). MS1 matched peptide features and MS2 fragment spectral clusters were described. Further development of the software was required to able to successfully process the samples, and a development version was used (M. The, personal communication).

Fisher's exact test was applied at 5% FDR (Benjamini-Hochberg) to identify differentially prevalent feature groups and corresponding MS2 clusters. Highly-prevalent feature groups (present in 190 or more samples) were also identified and the largest MS2 cluster for each was extracted to provide identifiable control spectra.

These spectra were together searched against the UniProt reference proteome database using comet and x!tandem wthin SearchGUI 4.0.0-beta. A further search was also carried out following database reduction with MetaNovo.

Results

Samples and data

Demographic and clinical data are shown in Table 15.

	Febrile		Healthy	KD		TB
	Acute (N=25)	**Subacute** (N=35)	(N=31)	**Acute** (N=56)	**Subacute** (N=46)	(N=37)
Sex						
Female	10 (40.0%)	16 (45.7%)	12 (38.7%)	21 (37.5%)	13 (28.3%)	
Male	15 (60.0%)	19 (54.3%)	13 (41.9%)	35 (62.5%)	33 (71.7%)	
Missing			6 (19.4%)			37 (100%)
Age (y)						
Mean (SD)	3.12 (2.27)	3.93 (3.02)	6.57 (4.49)	4.61 (3.65)	4.25 (2.87)	
Median [Min, Max]	2.30 [0.100, 8.70]	3.00 [0.700, 12.0]	6.17 [0.250, 15.2]	2.87 [0.222, 15.3]	3.80 [0.300, 11.9]	
Missing			6 (19.4%)			37 (100%)
Ethnicity						
American Indian/Alaska Native				1 (1.8%)		
Asian	2 (8.0%)	2 (5.7%)		7 (12.5%)	6 (13.0%)	
Black/African American	1 (4.0%)	1 (2.9%)		2 (3.6%)	2 (4.3%)	
Caucasian	7 (28.0%)	8 (22.9%)		9 (16.1%)	13 (28.3%)	
Hispanic	7 (28.0%)	10 (28.6%)		27 (48.2%)	17 (37.0%)	
Multiple	3 (12.0%)	10 (28.6%)		8 (14.3%)	5 (10.9%)	
Other				1 (1.8%)	1 (2.2%)	
Unknown	5 (20.0%)	4 (11.4%)			1 (2.2%)	
Missing			31 (100%)	1 (1.8%)	1 (2.2%)	37 (100%)
Day of illness						
Mean (SD)	4.76 (1.36)	12.2 (2.58)	-	6.45 (3.57)	15.3 (4.68)	-
Median [Min, Max]	5.00 [2.00, 6.00]	13.0 [7.00, 16.0]	-	6.00 [2.00, 24.0]	14.5 [3.00, 25.0]	-
Missing						
Convalescent day of illness						
Mean (SD)	-	-	-	48.9 (16.4	-	-
Median [Min, Max]	-	-	-	48.0 [26, 124]	-	-
Missing				3 (5.4%)		
CRP (mg/dL)						
Mean (SD)	46.6 (56.7)	37.6 (49.5)	-	138 (150)	63.8 (66.1)	-

	Febrile		Healthy	KD		TB
	Acute (N=25)	**Subacute** (N=35)	(N=31)	**Acute** (N=56)	**Subacute** (N=46)	(N=37)
Median [Min, Max]	25.5 [3.00, 212]	16.0 [3.00, 202]	-	78.0 [3.00, 990]	43.5 [3.00, 291]	-
Missing	1 (4.0%)	4 (11.4%)		1 (1.8%)	2 (4.3%)	
WBC (10^9/L)						
Mean (SD)	10.7 (7.86)	13.8 (8.16)	-	15.2 (5.53)	14.7 (6.35)	-
Median [Min, Max]	8.90 [1.60, 43.0]	12.8 [2.10, 31.0]	-	15.6 [5.70, 27.9]	12.9 [8.00, 38.0]	-
Missing		1 (2.9%)		2 (3.6%)	1 (2.2%)	
PMNs (10^9/L)						
Mean (SD)	3.80 (2.28)	6.49 (5.82)	-	9.02 (4.89)	8.46 (5.42)	-
Median [Min, Max]	3.70 [0.288, 9.03]	5.35 [0, 19.9]	-	8.95 [0.900, 23.2]	6.90 [1.69, 24.7]	-
Missing		2 (5.7%)		2 (3.6%)	1 (2.2%)	
Lab						
Bristol		20		10	10	
Oxford	25	15	31	46	36	37
Sample batch						
06/04/2013			-		17*	-
04/08/2016	25	35	-	32	27	-
08/11/2017			-	24	2	-

Table 15 Demographic and clinical data for patients included within metaproteomic analyses. Samples from acute febrile patients were mixed with pooled convalescent serum. *The KD patients in the first batch of samples were late samples from patients already treated with IVIG. Other subacute samples came from patients presenting late and are taken before IVIG.

Twenty-eight samples come from individuals also providing metagenomic data, 26 KD and 2 Febrile. Fourteen additional samples were present in the batch analysed at Bristol – comprising ten healthy adult controls and the two pairs of adult samples with and without spiked influenza vaccine (see preceding chapter). They were processed in the pipelines but not included in the presentation of results.

Proteomic data

The samples run at Bristol provided a median 78 074 MS2 scans per sample (IQR 72 987-86 703), and Oxford 27 308 (22 985-30 875).

Database reduction

Metanovo reduced the unrestricted database from 28 million sequences and 9.7 billion amino acids to 94 267 sequences and 45.9 million amino acids. These comprise 1 010 human sequences, and 6 752 from other species. The greatest number of non-human sequences are provided by *Rhizophagus clarus* (324). 3 676 sequences are shared with the database produced in the preceding chapter.

Peptide and protein identification

At 1% PSM FDR, 7.9% and 4.1 % of spectra were identified from samples at Bristol and Oxford respectively. There was a median of 3 968 (IQR 4 454-5072) and 749 (IQR 427-1 594) PSMs available per sample.

More protein groups were identified in Bristol samples than Oxford (5 095 vs 2 432). This remained true after applying protein FDR control at 1% (589 vs 433). The retained proteins comprised 93% and 88% of PSMs at Bristol and Oxford respectively.

Match between runs increased the identification of MS1 features, with 58% of identifications based upon matching.

The default MaxQuant contaminants database includes serum proteins from humans and other mammals (mostly *Bos taurus*). Of 79 contaminant protein groups identified across the two runs, 20 were deemed likely true sample proteins (e.g. albumin, immunoglobulin and complement related proteins).

Remaining likely true contaminants accounted for 2.9% of PSMs in Bristol and 1.6% of PSMs at Oxford. Within each site there was no evidence for difference in contaminant PSM proportion by group (Bristol p=0.18 by Wilcoxon rank sum; Oxford p=0.18 by Kruskal-Wallis).

Human proteins

Human protein groups passing FDR control were ranked within samples based on iBAQ values. Table 16 shows the top 10 protein groups based upon mean ranking, stratified by laboratory (due to different experimental techniques which are likely to impact upon quantitation).

Rank	**Bristol**	**Oxford**
1	Immunoglobulin Kappa Chain	Immunoglobulin G1
2	Albumin	Immunoglobulin Lambda Chain 2
3	Immunoglobulin G1	Immunoglobulin Kappa Chain
4	Immunoglobulin G3	Complement Factor 1QB
5	Immunoglobulin Lambda Chain 2	Apolipoprotein A1
6	Immunoglobulin M	Apolipoprotein C1
7	Complement Factor 4 Binding Protein Subunit A	Immunoglobulin G2
8	Complement Factor 4B	Apolipoprotein C3
9	Complement Factor 1QC	Immunoglobulin Kappa Variable 320
10	Complement Factor 1QB	Haemoglobin subunit alpha

Table 16 The most abundant 10 proteins per laboratory by mean iBAQ ranking. Proteins present in both rankings are underlined.

Level	Number identified (FDR)	
	No protein FDR	1% Protein FDR
Protein	3 660 (52%)	121 (6%)
Organism	2 275 (60%)	115 (6%)
Species	2 221 (60%)	113 (6%)
Genus	1 216 (69%)	92 (8%)

Table 17 Number of proteins, organisms, species and genera identified with and without protein false discovery rate (FDR) control. For each number, the corresponding FDR is given.

When grouped into protein families, the 5 top-ranked in Bristol were Immunoglobulin, Complement, Albumin, Apolipoprotein and coagulation factors. In Oxford, Immunoglobulin, Apolipoprotein, Complement, Haemoglobin and Amyloid were top.

Non-human protein identifications

Protein groups were classified at organism, species and genus levels. Where protein groups included multiple taxa at any level, the taxon with the larger number of razor peptide matches was retained (with alphabetical precedence if any tie). Identifications at each level are summarised in Table 17. Notably the protein FDR for microbial proteins exceeds 1% in this subset of proteins since the FDR in the larger number of human proteins is lower (0.4%).

Each microbial protein group was identified in a median of 8 samples (IQR 2-22, maximum 226), rising to 21 (IQR 6-40) among those passing protein FDR. Match between runs was important for identification of microbial proteins, with 66% of protein-sample identifications by matching.

Most microbial protein groups were identified in one laboratory only (3 928, 90%). This proportion was lower at species and genus levels, without protein FDR control (72 and 58% respectively). Perhaps surprisingly, the proportion was higher after protein FDR control (85 and 80%).

The majority of microbial protein identifications depended upon only one distinct peptide (97%). As expected, this proportion was lower with protein FDR control (47%). Microbial proteins failing FDR control almost all had only one identified peptide (3863/3898). Microbial protein coverage was low, with median 2.6% (IQR 1.5-4.5%, maximum 64%). Twenty-six *Rhizophagus* proteins were detected, though none passed protein FDR control.

Twenty-one microbial protein groups were identified in at least half of samples and passed protein FDR control. All but one were identified from both laboratory's samples, however only one protein was identified by more than two peptides. Each originated from distinct species, though two genera possessed two proteins each (*Phialocephala* and *Streptomyces*). Other genera comprised *Bacillus*,

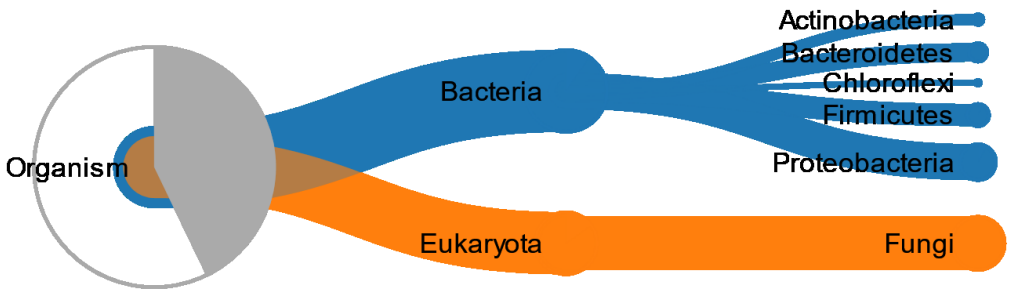

Figure 22 UniPept analysis (Mesuere et al, 2015) of taxonomic profile of microbial peptides identified at 1% protein FDR in patient samples by at least 1 PSM. The area of the circle is proportional to the number of peptides contained within each taxon. The segment shows the proportion of those peptides assigned to the taxon. FDR: false discovery rate; PSM: peptide spectral match.

Bdellovibrio phage, *Chaetomium*, *Elizabethkingia*, *Fonsecaea*, *Insolitispirillium*, *Paenibacillus*, *Penicillium*, *Polaribacter*, *Pseudomonas*, *Rozella*, *Sphingobacteriales*, *Sphingobacterium*, *Syntrophobotulus*, *Trichoderma* and *Variovorax*.

Twelve microbial protein groups surpassed 25% coverage in one laboratory and passed protein FDR control. Seven of these were influenza proteins with MS1 features matched between runs to spiked samples in at most six samples. The others included a thioredoxin protein with razor species identified as *Klebsiella pneumoniae* (though razor protein identification is to *E. coli*) identified from 5 peptides across 40 samples and a *Shigella flexneri* glutaredoxin identified by 4 peptides across 18 samples. Both were identified in Bristol alone. Other proteins were identified by one or two peptides only.

All microbial peptides belonging to protein groups identified at 1% FDR with at least one PSM in the samples were submitted to UniPept (Mesuere et al., 2015) to explore the taxonomic profile of lowest common ancestor identifications (Figure 22). Of 199 peptides, 189 could be identified.

Under half of identified peptides were specific to genus level or lower (n=75; 40%), while 81 (43%) were assigned to the root node. Three genera possessed four distinct peptides: *Penicillium*, *Tenacibaculum* and *Tilletiopsis*. The remainder possessed two or fewer.

Microbial protein counts were similar between KD and febrile samples in Bristol (median 9 vs. 10; p=0.26 by Wilcoxon rank sum). Differences were evidence between groups in Oxford, with KD highest,

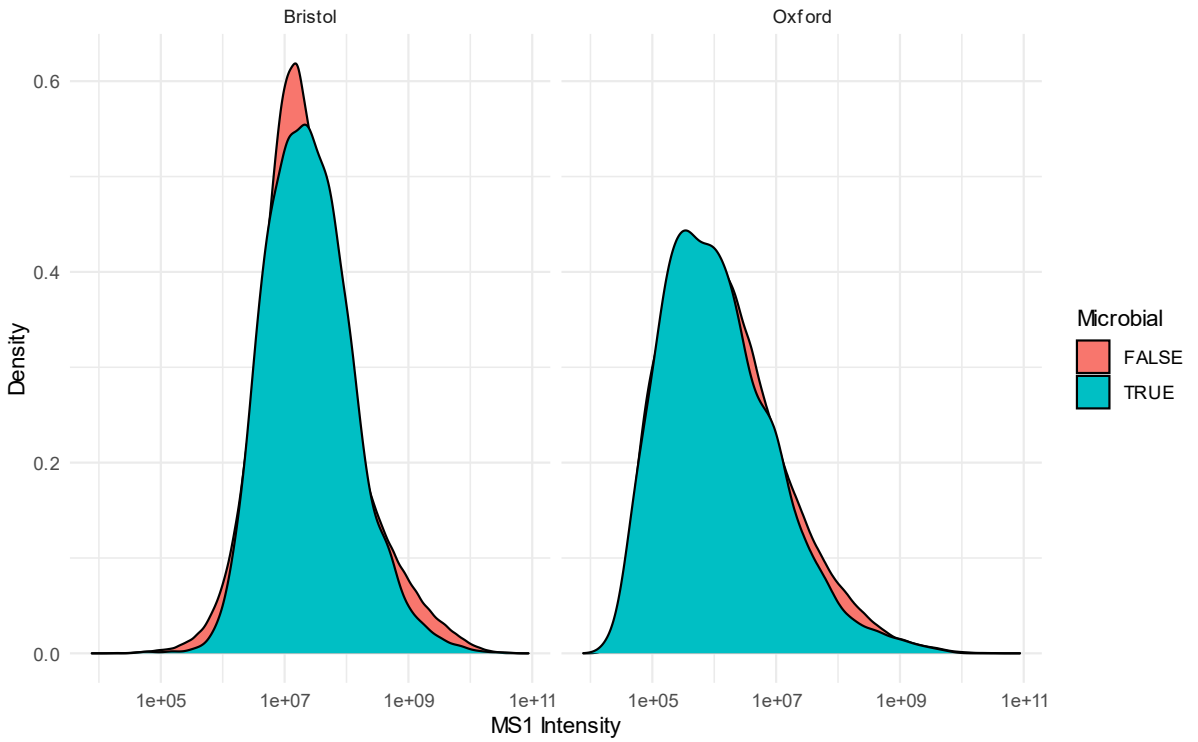

Figure 23 MS1 intensity of peptides from proteins at 1% false discovery rate by microbial status. MS: mass spectrometry.

followed by febrile, healthy then TB samples (29, 23.5, 14.5 and 11.5 respectively; p=0.09, <0.001 and <0.001 for KD compared to each).

Microbial peptides had similar intensities to non-microbial peptides from proteins passing FDR control (Figure 23).

Microbe-KD associations

Microbes with prevalence associated with KD were sought at species and genus level. Analyses were conducted within laboratory groups for taxa restricted to a single laboratory. Comparisons were made by Fisher's exact test and adjusted (Benjamini-Hochberg) between KD and febrile and healthy controls separately and combined, and with the TB group included.

Four species demonstrated increased prevalence in KD compared to healthy controls at 1% protein FDR. However, no species, including these, were significant when KD was compared to febrile controls, or both controls together. Table 18 shows that the prevalence of these species is similar or higher among febrile controls and TB samples. Identical results were recapitulated at genus level.

Species	Lab	KD	Healthy	Febrile	TB	KD vs healthy adjusted p value
Hyaloscypha variabilis	Both	59/102 (57.8%)	2/31 (6.5%)	39/60 (65%)	37/37 (100%)	<0.001
Polaribacter sp. ALD11	Both	67/102 (65.7%)	7/31 (22.6%)	39/60 (65%)	37/37 (100%)	0.003
Pseudomassariella vexata	Oxford	29/82 (35.4%)	0/31 (0%)	26/40 (65%)	4/37 (10.8%)	0.002
Sphingobacteriales bacterium	Both	58/102 (56.9%)	5/31 (16.1%)	56/60 (93.3%)	37/37 (100%)	0.007

Table 18 Species with significantly higher prevalence in KD versus healthy controls. Results for other comparisons are not shown, as there are no species significantly more prevalent in KD. P values are adjusted by the method of Benjamini Hochberg.

Quandenser

242 samples could be processed by this "spectra first" approach, as one sample could not be converted to mzML format and another consistently prevented the software from running to completion. MS1 feature groups numbered 3.0 million. From 9.4 million MS2s, 4.9 million MS2 clusters were identified, 81% of which were singletons.

A significant proportion (55%) of MS2 clusters were not assigned to MS1 features, preventing their detection by matching in other samples. These mostly originated from the Bristol samples.

MS1 features were frequently specific to laboratory (only 23% of features present in 10% or more of either labs' samples were shared), presumably due to loading differences and the challenge of aligning retention times. MS2 clusters similarly were frequently lab-specific (4% of clusters of size 10 or above included MS2s from both labs).

Fisher's exact test at 5% FDR identified 1 488 feature groups with increased prevalence in KD with 29 646 corresponding MS2 clusters. 3 544 highly-prevalent feature groups (190 or more samples) were identified and the largest MS2 cluster for each was extracted (n=3 484) to include MS2s with a high likelihood of being identifiable.

KD-associated features were detected in a median of 73 samples (IQR 61-86) and largest MS2 clusters contained a median of 3 spectra (IQR 2-5). Control features were detected in a median of 202 samples (IQR 191 to 218) and largest MS2 clusters contained a median of 9 spectra (IQR 3-39). Mean intensities of KD-associated features was low compared to control features – median 8.4×10^5 vs. 9.2×10^7.

Consensus spectra were extracted for all 33 130 clusters and searched against the reference proteome database with comet and x!tandem using SearchGUI 4.0.0-beta (Eng, Jahan & Hoopmann, 2013; Fenyö & Beavis, 2003; Barsnes & Vaudel, 2018). This tool was selected since it is not possible to search

spectra alone with MaxQuant. The distribution of PSM scores within target and decoy databases were similar, likely due to the size of the database (data not shown).

To mitigate against this, database reduction was carried out using MetaNovo, resulting in a database of 179 972 amino acids, of which human proteins comprised 26%. Only one organism (*Hanseniospora osmophila*) contributed more than 2 proteins.

567 of 3 484 (16%) consensus MS2 spectra from highly prevalent feature groups had PSMs at 1% FDR, of which 559 were confident. 96 of 126 proteins were human and comprised known constituents of PEG eluant (including immunoglobulin and complement).

In contrast, only 45 of 29 646 (0.2%) consensus MS2 spectra from differentially-prevalent feature groups had PSMs, 39 of which were confident. Five of 14 proteins were human (expected proteins, as above). Nine were single peptides from distinct organisms, whose genera comprised *Aspergillus*, *Cohaesibacter*, *Coniochaeta*, *Dendrosporobacter*, *Gloebacter*, *Kitasatospora*, *Phenylobacterium*, *Pseudogymnoascus* and *Xenorhabdus*.

Discussion

In this chapter, I undertook the challenge of identifying potential microbial protein antigens in a large cohort of samples from children with KD and multiple sets of controls. Experimental design and data acquisition preceded my involvement in the project.

Healthy controls provide samples in which circulating microbial antigen will not be expected. Febrile samples provide alternative controls, where antigen from pathogens may be expected though KD-related organisms should not be enriched. Subacute KD and febrile cases (≥7 days from fever onset) were included to cover to the period during which circulating immune complexes are detected in KD. It was considered that in acute cases (<7 days) microbial antigens may be more abundant both in febrile illnesses and KD, thus increasing the potential to detect them. Convalescent serum was added (paired, in the case of KD) to encourage formation of ICs for precipitation. TB samples were included because the finding of BCG scar inflammation in some children with KD raises the possibility of a shared antigen.

In the preceding chapter I explored the particular challenges posed by the “minority metaproteomics” analysis to identify microbial proteins over-represented in KD ICs. Metanovo was able to reduce a large reference database, concentrating both human sequences and those from the influenza viruses known to be present. Indeed, the performance of the subsequent MaxQuant search was better with this database than an *a priori* taxonomy-restricted database.

Database reduction and conventional search approach

Metanovo required reimplementation to run at the scale required for these samples. Nextflow (Di Tommaso et al., 2017) was critical to the feasibility of this, allowing the bash scripts and Python code to be orchestrated across multiple nodes of the HPC infrastructure, taking advantage of almost 200 machines.

The reduced database was of a similar size to that from the preceding chapter with a similar number of human proteins. Metanovo adds a protein to the database when it matches a sequence tag from a spectrum which does not have a higher-ranking match (Potgieter et al., 2019). This suggests that saturation may have occurred with a relatively small number of spectra.

Rhizophagus proteins were again abundant within the database but not detected after protein FDR control. It is likely that their strong representation relates to the large number of proteins in the reference database.

Only 3% of microbial proteins with peptide identifications passed protein FDR control. This contrasts with a rate of 86% for human proteins. This is consistent with the much higher protein FDR estimated within the microbial protein subset (52% without protein FDR, and 6% with).

Contaminants are a ubiquitous problem in metagenomics and metaproteomics. Contaminant databases in proteomics typically focus on high-abundance proteins present in the laboratory environment, mostly proteins used in experimental processes and skin and hair proteins. Microbial contamination is not typically considered. This may be because typical microbial contaminants are present at very low abundances, and are unlikely to be detectable. Even should microbial contaminant proteins be detected and fragmented within the mass spectrometer, they are unlikely to be represented within single-species databases typically used.

It is possible that some of the microbial protein signal, which is detected in all sample groups, represents background contamination. However, there is little overlap between the spectrum of organisms detected and typical metagenomic laboratory contaminants, as considered in chapter 3, This also applies when considering proteins with high sequence coverage or present across a high proportion of samples. The strikingly similar peptide intensity profile for microbial and non-microbial proteins is unexpected if the majority of this signal represents low abundance contaminants and/or circulating microbial antigen.

Taken together, this raises the concern that a sizable proportion of putative microbial protein identifications could be erroneous labels given to human peptides not identified by the sequence database. Concerns have been raised that two-step approaches to proteomic searching can invalidate

FDR control (Everett, Bierl & Master, 2010; Bern & Kil, 2011). Estimates presented here may be best taken as a lower bound.

Even should true microbial protein signals be buried among a range of spurious identifications, the stochastic nature of the latter should allow detection of valid signals – provided a sufficiently large sample for the strength of the signal. Thus, the prevalence of species and genera by sample group (largely on the basis of single proteins) was compared.

Significantly over-represented organisms were identified only when comparing KD samples to healthy controls. These organisms were each identified by single proteins, and detections were similarly prevalent among febrile controls and samples from children with TB. This illustrates the importance of selecting appropriate controls, and the potential for multiple controls to assist in interpretation of results.

Any causative agent of KD is unlikely to be highly prevalent among both children with other febrile illnesses and TB. The possibility looms large of spurious signal linked to misidentified human peptides from proteins which are at low abundance among healthy controls.

The organisms identified are environmental, and with no known association with human infection or disease from literature searches. Two are fungal. *Hyloscypha* is a genus of dimorphic fungi, of family Hyaloscyphacaea and order Heliotiales. Heliotales are described as "morphologically very variable and [exhibiting] saprobic, parasitic as well as mycorrhizal life strategies" (Kosonen, Huhtinen & Hansen, 2021). *Pseudomassariella vexata* is the only recorded species of *Pseudomassariella*, also a fungus of the clade Leotiomyceta, and no specific literature on its ecology can be identified.

Polaribacteria are marine organisms with important environmental roles (Bowman, 2018). Sphingobacteriales are an order of Sphingobacteriia, which were determined to be a major component of lake water and sediment in a study in China (Qu et al., 2015).

Spectra-first approach with Quandenser

Traditional database searches are computationally expensive because of the large numbers of experimental and theoretical spectra that must be compared. This is exacerbated in metaproteomic searches. Quandenser proposes to "focus on the spectra that matter" by implementing advanced MS1 feature identification, quantification and matching between runs (with FDR control informed by MS2 spectral clustering). In this way relevant features and their corresponding spectral clusters can be selected based on signals of differential abundance or prevalence, only then undergoing identification.

Processing the large number of samples through this novel approach was challenging due both to computational resources required and the need for further development by the author. Limitations

were imposed by the dependence on matching features between two groups of samples, which underwent preparation and mass spectrometry in different laboratories, with different procedures and mass spectrometers. The segregation of most features present in 10% or more samples to either laboratory suggests that feature matching between laboratories is limited.

The high prevalence of MS2 spectra (mostly from Bristol samples) which could not be paired with MS1 features is a further challenge. Notably, a much lower proportion of MS2 spectra identified by MaxQuant (12%) lacked a corresponding MS1 feature, though again this was similarly higher in Bristol (19 vs 4%). The author of Quandenser suggested the feature detection algorithm may have struggled due to higher peptide loading and consequent detector saturation in Bristol (M The, personal communication).

Nonetheless, it was possible to identify nearly 1 500 MS1 features which were significantly over-represented among KD samples and extract corresponding consensus MS2s. The consensus-spectrum approach is intended to result in the highest quality spectrum being used for identification. A control group of consensus spectra from high prevalence features was also included for comparison.

The massive reduction achieved in the number of spectra to be identified made it computationally tractable to search an unreduced reference database. However, this led to very poor discrimination between quality of matches to the target and decoy databases – likely relating to the increased probability of better quality matches to the large decoy database.

Consequently, database reduction became a necessary step. Disappointingly, following this reduction a very low proportion of KD-associated MS2s could be identified, as compared to the MS2s from high-prevalence features. Among those few identified MS2s were human peptides and an array of isolated matches to single organisms.

The reasons for the low identification rate of these spectra remains to be determined. Spectral quality is a likely reason. The control spectra are derived from high prevalence and intensity peptides, which give rise to many MS2 fragment spectra which are expected to have high signal-to-noise ratio. This large number of high-quality spectra should produce high quality consensus spectra. In contrast, KD-associated spectra are derived from peptides with lower prevalence and intensity, and with fewer MS2 spectra, likely with lower signal-to-noise ratio, from which to generate consensus spectra. It is also possible that they represent peptides which are not found in the reference database.

Conclusions

Despite the extensive depth of proteomic sequencing undertaken over a large number of samples, significant computational and inferential challenges were faced in the search for microbial proteins associated with KD.

Results from two complementary computational approaches were distinct and neither gave rise to consistent or strong signals towards one or more organisms.

There is considerable potential in further analysis of these data. Experimental approaches could also be optimised. Regarding sample preparation, methods to remove abundant immunoglobulins without depleting cognate antigen would allow the mass spectrometer to focus on fragmenting and identifying non-antibody peptides. Further potential improvements, including mass spectrometry and analytic methods will be considered in detail in chapter 9 (p142).

7 | Antibody proteomics

Introduction

My work in the preceding chapters has focused on direct identification of microbes at the levels of protein and DNA sequence. Both of the datasets furnish data which can be used to answer different questions.

Firstly, as illustrated in chapter 3, the metagenomic data provides more human whole genome sequence (WGS) data than microbial data. There are no published WGS analyses in KD and only one published whole exome sequencing study of 159 KD patients with controls (Kim et al., 2021) and two small WGS studies of a family (Kim et al., 2017) and a child (Kanda et al., 2021). At the time of writing, a colleague has begun to explore Human Leukocyte Antigen (HLA) haplotypes based on these data (Evangelos Bellos, personal communication).

Secondly, the PEG-precipitates are enriched for immunoglobulins (see Figure 17, p96), as would be expected. Thus, the diversity of variable region tryptic peptides, as well as those from the constant region can be anticipated within the proteomic data. These data are perhaps more relevant to the search for microbial antigens which could trigger KD, given the genetic evidence supporting a role for the B-cell response and immune complexes, as reviewed in the introduction.

Existing data from a range of illnesses supports the association of distinct patterns of antibody variable region responses, usually at the level of bulk sequencing of the B-cell receptor (BCR), or complementarity determining region (CDR) 3. Convergent antibody signatures have been demonstrated in human Dengue fever (Parameswaran et al., 2013), hepatitis B and influenza vaccination (Galson et al., 2015; Jackson et al., 2014; Adamson et al., 2017).

Proteomic analysis of antibodies is challenging due to the high diversity of the variable region and poor representation in sequence databases. The most frequent application of proteomics is to monoclonal antibodies, where post-translational modifications, including glycosylation, can be explored (Guthals et al., 2017; Cheung et al., 2012; Sato et al., 2012). Database-independent *de novo* approaches can be used, with specialised approaches in commercial software (Tran et al., 2016).

Some groups have made progress in the proteomic analysis of bulk or affinity-purified immunoglobulin, frequently in tandem with BCR sequencing in the context of vaccination. In one example, Lee et al. (2016) obtained peripheral blood samples from four individuals prior to influenza vaccination and at three subsequent timepoints. IgG antibody binding fragments (F(ab')2 were extracted and underwent affinity purification with influenza antigen. Sample-specific antibody sequence databases were generated for proteomic searching. They demonstrated that the post-

vaccination repertoire was dominated by pre-existing antibody clonotypes, and found a number of clonotypes with cross-reactivity for haemagglutinin 1 and 3. Several other groups have studied the antibody response to vaccination using proteomics (VanDuijn et al., 2017; Adamson et al., 2017; Lavinder et al., 2014)

No proteomic studies have been conducted of antibodies in KD, though one group studied BCR sequences and identified clonotypes shared by IVIG-resistant KD cases and not present in controls (Ko et al., 2018). More recently Chang et al. (2022) explored the clonotypes underlying a large expansion in IGHV4-34 usage among plasmablasts in a single KD patient. They were unable to identify a specific antigen bound by the antibodies.

In this chapter, I undertake secondary analyses of the proteomic data from PEG-precipitated samples used in chapters 6 and 7, with the aim of detecting variable region immunoglobulin peptides in KD patients and controls and exploring quantitative differences in V and J segment usage.

Objectives are:

1. Detect antibody heavy and light chain peptides in PEG-precipitates
2. Structurally align antibody peptides to the immunoglobulin molecule and classify according to isotype and loci
3. Describe coverage of the antibody molecule
4. Compare V and J locus usage between KD cases and controls

Methods

Preprocessing of abYsis database

In the first step, structural numbering of antibody sequences required processing. The abYsis XML databases gives amino acid sequences, structural numbering and region co-ordinates for human and non-human antibodies according to Kabat, Chothia and modified Chothia (Martin) definitions (Swindells et al., 2017; Abhinandan & Martin, 2008).

The abYsis Kabat and EMBLIg databases were processed in R. Human antibodies with heavy and/or light chain amino acid sequences available were selected. Residue numbering was processed to identify insertions (indicated with alphabetical termination, e.g. H52A) and deletions (indicated by non-consecutive numbering (e.g. H29-H32). These were summarised in an edit string (e.g. “6[-1]:95[2]” signifying one amino acid deletion following residue 6 and 2 insertions following residue 95). Residues exceeding the standard range of heavy and light chain amino acids (i.e. H114, L110-111) were renamed to become insertions (H113A, L09A-B) to prevent inconsistency in canonical length.

Proteomic database searching

As described, 40 samples together with 14 method development samples underwent mass spectrometry at both Bristol and Oxford. The remaining 153 samples were analysed only in Oxford. For this analysis, all Raw files were included and analysed together in MaxQuant 1.6.10.43. The human reference proteome and abYsis database was used, with MBR enabled. Other settings were as described in Chapter 7.

For downstream analysis, peptide identifications in Oxford for samples analysed in Bristol were disregarded – in this way, these mass spectrometry runs acted as a source of additional identifications through MBR.

Structural alignment and sequence classification of antibody peptides

Peptides identified by MaxQuant and matching to the abYsis database sequences (including both constant and variable regions) were identified, herein referred to as abYsis peptides. Individual abYsis peptides could match with sequences from one or more references. Sequences were classified as heavy or light chain based on at least a two-fold excess of matches to one chain.

Each abYsis peptide was positionally located within matching reference sequences to determine Martin numbering of the first and last residues. The consensus value was taken. Some peptides could not be given structural numbering because no reference sequence was numbered, due to failure of AbNum (Abhinandan & Martin, 2008). These reference sequences were structurally numbered with ANARCI where possible (Dunbar & Deane, 2016) and the process repeated.

Classification of antibody peptides

abYsis peptides including at least 7 amino acids belonging to the constant region of heavy or light chains were identified. Constant region sequences were extracted and partial edit distances calculated to UniProt reference sequences for IgG1-4, IgM, IgE, IgA1-2, IgD, and kappa and lambda light chain constant regions. abYsis peptides were classified according to the closest class(es).

abYsis peptides covering the variable region were queried by BLAST against the Immunogenetics (IMGT) RefSeq amino acid database of Variable, Diversity and Joining (V, D and J) regions. Peptides were classified according to the best match(es) (lowest E value), including multiple assignments where necessary. Assignments were summarised at locus and locus group levels (e.g. “IGHV4-4,IGHV4-59” and “IGHV4”). Locus groups were further simplified by grouping lettered groups (e.g. “IGKV3,IGKV3D” and “IGKV3/D”).

Relative contributions of constant peptides by antibody class and subclass and variable peptides by locus were summarised. For the purpose of summarising protein and protein class contributions by sample, modified iBAQ intensities were calculated as described in the preceding chapter.

Analysis of antibody coverage

Coverage of the heavy and light chains were summarised by unique peptides and MS1 intensity.

Differential analysis

Variable region peptides were quantile normalised. Separately for V and J loci, normalised intensities were summarised by sample and locus and scaled to achieve identical sums per sample. Summarised intensities were $\log_2$ transformed.

Two-dimensional principal co-ordinates analysis (PCoA) was used to explore overall differences between sample groups, concatenating the normalised V and J locus matrices. ANCOVA (as implemented within vegan) was used to explore factors explaining variance (Dixon, 2003).

Limma (Ritchie et al., 2015) was used to test for differential abundance by group for V and J loci separately. Contrasts were established to compare pre-IVIg KD samples with Febrile and Healthy samples, and with post-IVIG KD samples. FDR control was implemented with the method of Benjamini and Hochberg at 5%. Unadjusted confidence intervals for log2-fold changes were presented.

Results

abYsis database processing

The EBMLIg and Kabat databases (dated 25 and 17 January 2018 respectively) comprised 156 104 antibody sequences. These provided 72 935 unique heavy and light chain amino acid sequences from human antibodies with sequences between 70 and 627 amino acids, which provided the sequence database to be searched. Of these sequences, 62 384 (86%) had structural numbering applied by abNum.

Peptides identified

Almost 1.5 million peptide identifications were made across the 244 samples, comprising 31 253 peptides and accounting for just over 700 thousand MS2 fragment spectra (of nearly 10 million obtained). 54% of identifications were made by MBR. 38% of identifications were antibody sequences, but these comprised 19 130 (61%) of the unique peptides.

The PSM FDR was 1.0%. However, decoy identifications were all from non-antibody sequences, leading to an estimated 1.6% FDR for non-antibody peptides and 0% for antibody peptides.

Cysteine-containing peptides

Importantly, it was noted that cysteine-containing peptides were few in data obtained in Oxford (0.6% of peptides vs 30% in Bristol). This led to the recognition that the SMART Digest Kit (Thermo) used for sample processing in Oxford did not include reducing or alkylating agents. Thus, disulfide bonds will be expected to be largely intact, and peptides including these bonds unidentifiable unless some reduction has occurred. Additionally, cysteines not participating in disulfide bonds will be unidentifiable, since they will not be carbamidomethylated. The FDR in the subset of peptides containing cysteine residues was estimated at 1% in Bristol, and 13% in Oxford.

Three samples with the highest numbers of PSMs were re-run without carbamidomethylated cysteine as a fixed modification to explore whether this would materially increase the numbers of peptides identified. The number of cysteine-containing peptides increased from 61 to 221 (55 to 158 unique peptides). For antibody peptides alone, the increase was from 44 to 104 (39 to 66 unique peptides), corresponding to a gain of less than 1% in both peptides and unique peptides. The total intensity of cysteine-containing peptides was 0.6% of the intensity of peptides identified in the original analysis.

Due to the limited gains in identifications and intensity, the analysis was not re-rerun across all samples. Further, since any reduction of disulfide bonds was likely to be incomplete and variable, intensity-based comparisons between sample groups may be uninformative even could these peptides be identified.

Based on this and the empirical finding of a high FDR among cysteine-containing peptides, the 328 distinct cysteine-containing peptides identified in Oxford, and 387 additional peptides based on MBR were censored from further analysis.

Identification and structural alignment of antibody peptides

The majority of abYsis peptides could be numbered with reference to abNum structural numbering (17 035; 89%) from one or more matching antibody sequences. Corresponding antibody sequences for unnumbered abYsis peptides were extracted and numbered with ANARCI. This allowed most remaining peptides to be numbered (1 654; 78%), with only 465 remaining unnumbered.

Peptides with 7 or more constant region peptides (n=443) were identified by partial distance measurement to constant region reference sequences. The majority (91%, accounting for 98% of occurrences) could be classified with a maximum edit distance of two. IgG subclasses provided the largest number of distinct peptides and total identifications (Table 19).

Immunoglobulin class	Distinct peptides	Number of identifications
IgG	157	56 263
1	31	13 543
2	15	3 848
3	23	3 833
4	14	1 321
Mixed	74	33 718
IgM	66	18 785
IgA	73	5 232
IgD	22	1 254
IgE	2	511
Kappa constant	65	24 786
Lambda constant	75	9 858
Mixed	19	523
Total	464	96 609

Table 19 Classification of constant region peptides and number of occurrences across all samples.

Peptides including part of the variable region (n=18 233) and unlocated abYsis peptides (n=465) were used as queries against the IMGT RefSeq amino acid database with BLAST. 17 582 received identifications with variable (V) loci, and 824 with junctional (J) loci, leaving 292 without identifications. Identifications predominantly corresponded to a single locus group – 95% of V loci and 90% of J loci. These are detailed in Table 20.

Heavy chain locus	Unique peptides
IGHJ1	35
IGHJ2	43
IGHJ3	86
IGHJ4	189
IGHJ5	73
IGHJ6	98
IGHV1	1 804
IGHV2	179
IGHV3	6 740
IGHV4	1 672
IGHV5	590
IGHV6	127
IGHV7	89
Total	**11 725**
Identifications	**255 662**

Kappa light chain locus	Unique peptides
IGKJ1	41
IGKJ2	35
IGKJ3	23
IGKJ4	53
IGKJ5	17
IGKV1/D	1 886
IGKV2/D	382
IGKV3/D	1 507
IGKV4	260
IGKV5	6
IGKV6/D	27
IGKV7	31
Total	**4 268**
Identifications	**154 346**

Lambda light chain locus	Unique peptides
IGLJ1	14
IGLJ3	33
IGLJ5	1
IGLJ7	4
IGLV1	728
IGLV2	322
IGLV3	763
IGLV4	73
IGLV5	56
IGLV6	118
IGLV7	44
IGLV8	24
IGLV9	17
IGLV10	7
IGLV11	5
Total	**2 209**
Identifications	**71 252**

Table 20 Numbers of unique peptides assigned specifically to individual immunoglobulin locus groups with overall numbers of identifications by chain.

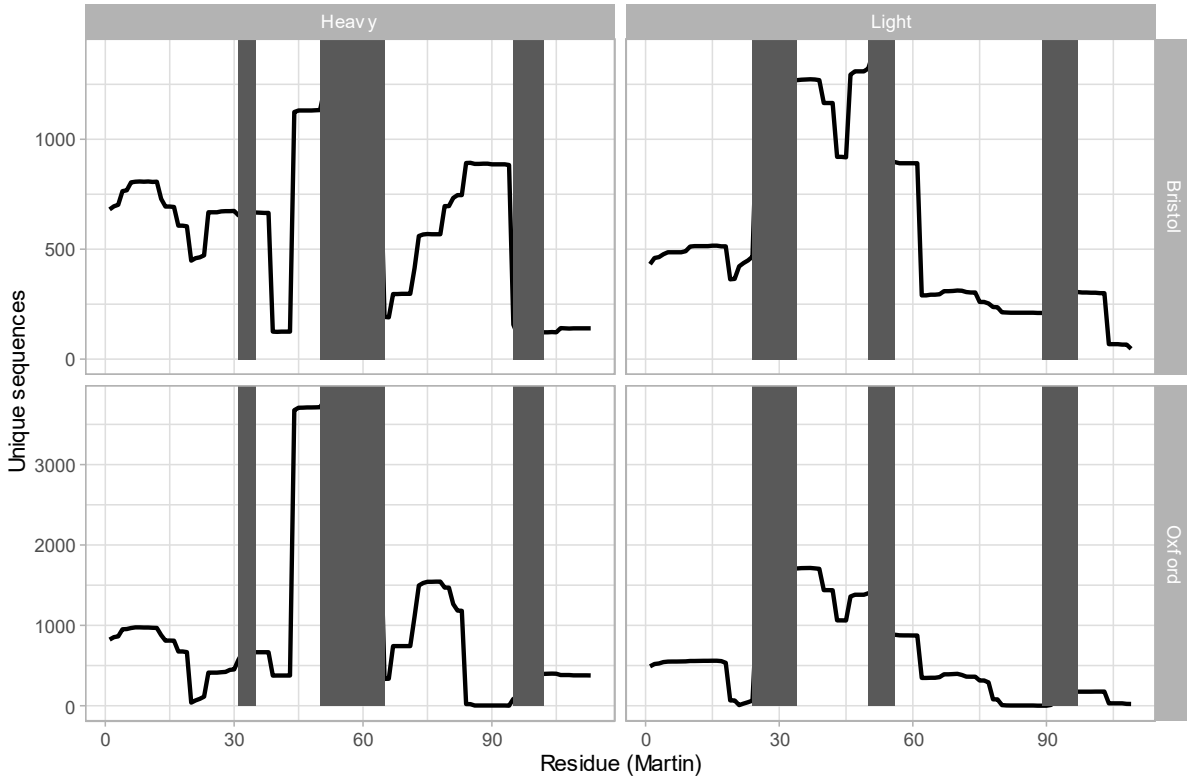

Figure 24 Unique peptides per residue of the immunoglobulin molecule. Results for heavy and light chains and each laboratory set are shown separately. Complementarity determining regions (CDR) are highlighted in grey.

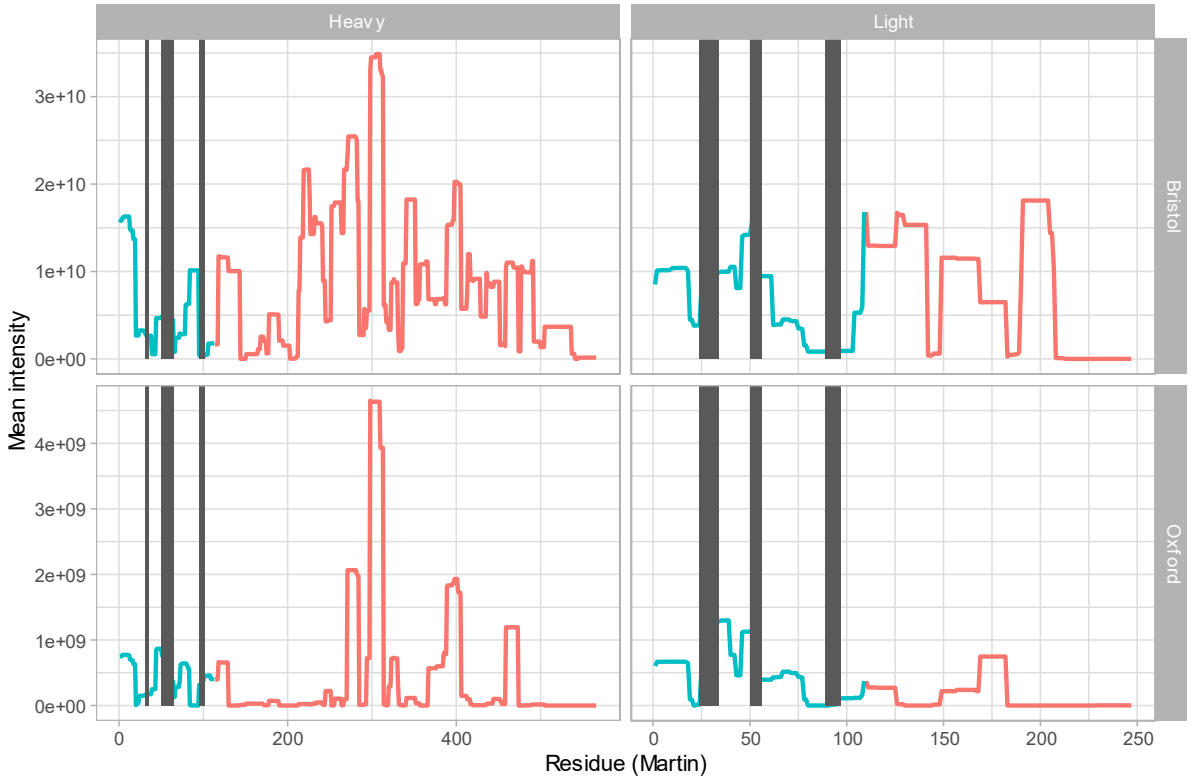

Figure 25 Mean total MS1 feature intensity per sample by residue over the immunoglobulin molecule. Results for heavy and light chains in each laboratory's sample set are shown separately. Complementarity determining regions (CDR) are highlighted in grey. The variable region is shown in blue and constant region in red. MS: mass spectrometry.

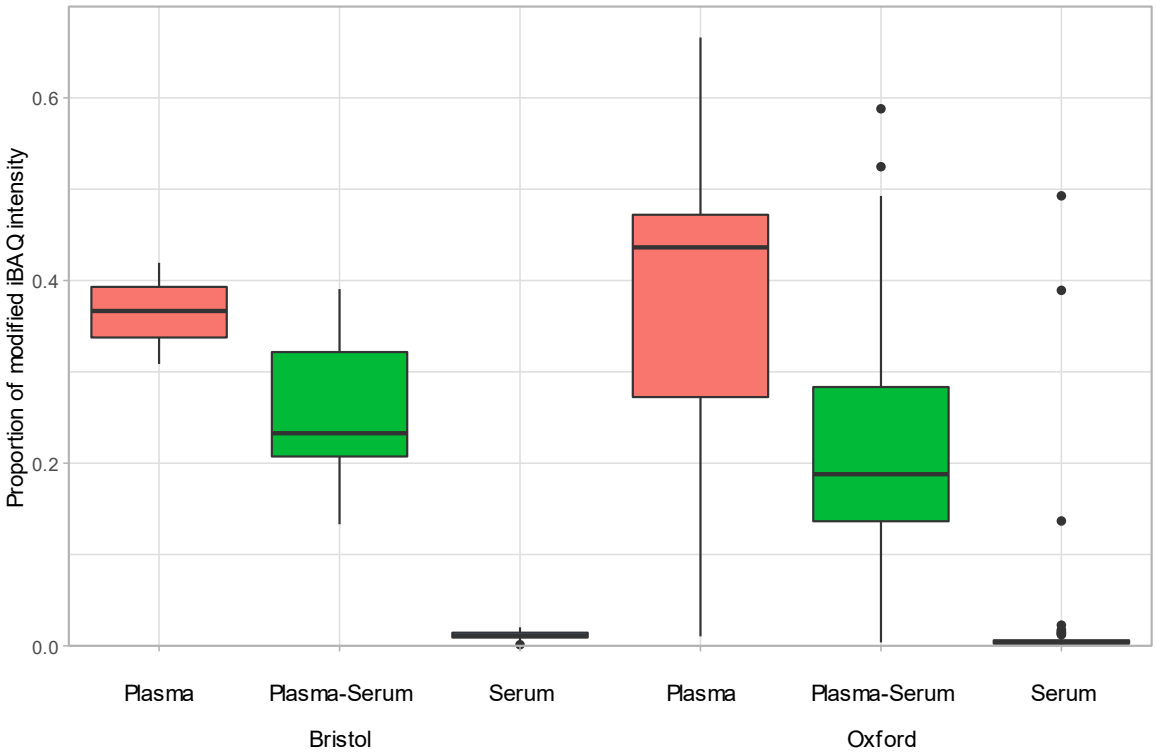

Figure 26 Proportion of coagulation proteins in samples estimated by modified iBAQ approach, shown by laboratory and sample type. Contaminants are excluded. iBAQ: intensity-based absolute quantification.

Coverage

Regions of the immunoglobulin molecule exhibited variable depths of coverage, both by unique peptides (Figure 24) and MS1 feature intensity per sample, a proxy for peptide quantity (Figure 25). CDR2 of the heavy chain, and CDR1 and 2 of the light chain had high numbers of unique peptides identified in both laboratories. Far lower numbers of unique peptides were identified in the region of CDR3.

A disulfide bond is expected in the antibody variable region between a cysteine residue N-terminal of CDR1 (H22/L23) to another N-terminal of CDR3 (H92/L88). Lower coverage in Oxford over H22, H92, L23 and L88 is seen, as expected. There is great disparity between intensity-based coverage of the constant regions between Oxford and Bristol, likely related to the peptides containing disulfide bonds.

Class	**Proportion of modified iBAQ intensity**	
	Bristol	**Oxford**
Albumin	11.4% (8.7-16.5)	1.0% (0.4-11.2)
Complement	36.0% (27.1-42.2)	22.0% (12.1-35.2)
Immunoglobulin	34.4% (30.8-40.1)	37.8% (28.9-48)
Other	13.9% (11-19.2)	29.6% (20.5-37.8)

Table 21 Proportion of protein abundance estimated by modified iBAQ approach in each lab, excluding contaminant and coagulation proteins.

Protein classes

Overall, a median of 39 and 25% of MS1 feature intensity was identified in samples analysed at Bristol and Oxford respectively (IQR 36-41% and 16-38%).

Discounting contaminants and haemoglobin, coagulation proteins represented higher proportions of modified iBAQ estimates in plasma samples and mixed plasma and serum than in serum samples, as expected (Figure 26). Setting aside coagulation proteins, immunoglobulins represented around a third of modified iBAQ intensity (Table 21).

Immunoglobulin classes and subclasses

Across all samples, IgG and IgM accounted for a median of 98.8% (IQR 97.0-99.8%) of estimated heavy chain abundance, with slightly lower estimates at Bristol (median 96.2 vs 99.5%). Kappa chains accounted for a median 54.7% of estimated light chain abundance (IQR 38.3-65.6%) with higher estimates at Bristol (median 70.4 vs 48.8%).

The majority of IgG subclass abundance originated from non-subclass specific peptides (median 71.4%, IQR 62.8-85.3%). Subclass specific breakdowns are shown in Table 22.

IgG3 is detected at much lower abundance in Oxford samples as compared to Bristol. This corresponds to a low number of peptides detected (7/23 total for IgG3) and the non-detection of the most abundant peptide in Bristol. IgG3 contains 18 cysteine residues (other IgG isotypes 9-11), and inspection of the peptides showed the peptides identified in Oxford did not include cysteine residues (excluding low-intensity matched features already excluded).

Reviewing the three samples from Oxford run with optimised modification settings, PSMs mapping to the canonical IgG3 increased from 120 to 132, and unique peptides from 7 to 14. Six of the additional

Sub-class	Bristol			Oxford						
	Febrile	KD		Febrile		Healthy (n=31)	KD			TB (n=37)
	S (n=20)	A+C (n=7)	S (n=13)	A+C (n=25)	S (n=15)		A+C (n=46)	S pre-IVIG (n=20)	S post-IVIG (n=16)	
IgG1	72.1% (67.3-73.7)	76.7% (73.4-79)	79.8% (68.8-84.4)	76.2% (68.2-80.4)	78.6% (68.9-92.5)	36.4% (25.7-55.7)	78.4% (48.8-84.2)	63.9% (53.2-77.9)	52% (44.3-58.3)	60% (40.8-77.8)
IgG2	4.2% (3.4-5.2)	10.2% (7-13.4)	3.7% (2.6-6.5)	22.1% (17.6-31.5)	19.7% (5.9-23.9)	56.4% (40.3-66.6)	21.4% (14.4-45.2)	31.6% (21.4-44.3)	46.5% (39.6-55.6)	34.3% (19.6-50.7)
IgG3	24.7% (21.8-28.1)	14.4% (6.7-16.5)	13.8% (10.5-27.9)	0.2% (0.1-0.3)	0.2% (0-0.8)	0.2% (0.1-0.3)	0.1% (0-0.1)	0.1% (0-0.2)	0.1% (0-0.2)	0% (0-0.1)
IgG4	0.5% (0.3-0.7)	0.4% (0.3-1.2)	0.6% (0.3-0.8)	1.5% (0.8-2.1)	2% (0.9-4)	4.3% (1.9-11)	1.3% (0.4-3.2)	2.3% (0.9-5.5)	2.2% (0.8-3.4)	1.6% (0.4-4.7)

Table 22 Median (interquartile range) modified iBAQ abundances of IgG subclasses expressed as a proportion of subclass-specific IgG peptides. Median subclass abundances do not need to sum to 100%.

peptides contained unmodified cysteine residues. However, the total intensity of matched peptides only increased by 13%, suggesting only a small proportion of disulfide bonds were reduced.

IgG2 is detected at highest abundance in in healthy samples, and KD samples post-IVIG.

Overall variation in immunoglobulin locus usage

Intensities of peptides with V and J locus identifications were normalised. Normalised intensities were summarised by V locus group and J locus. Factors associated with aggregate differences in abundance patterns were explored between and within laboratory datasets using principal co-ordinates analysis (PCoA) of Euclidean distance matrices.

Of note, subacute KD samples (n=49) include 17 from children already treated with IVIg, and 32 pre-treatment samples from children presenting late. A single 2 g kg^{-1} dose of IVIg potentially more than triples the baseline circulating mass of immunoglobulin,[2] thus the post-treatment samples likely contain large contributions from IVIg.

Acute pre-treatment KD samples (n=53) were mixed with paired convalescent serum in an equal volume. Convalescent samples were taken at a median of 48 days (range 26 to 124) after symptom onset. Since the half-life of IgG (except for IgG3) is around 21 days (Vidarsson, Dekkers & Rispens, 2014), almost a quarter of any IVIg given would be expected to remain in the circulation at 48 days. Assuming a 66% starting proportion, this would correspond to ~16%, further halved in the mixing of acute and subacute serum.

The two principal co-ordinates show some separation of groups in samples processed in Bristol and Oxford (Figure 27 and Figure 28 respectively). In Bristol, excluding the two influenza-spiked samples, and pooling unspiked samples with other adult controls, the four groups explained 10% of the variance ($p=0.002$) by ANCOVA. No separation was evident between acute and subacute KD samples – only one subacute sample here was post-IVIG.

In Oxford, segregation of groups was more clearly evident, with separation within KD samples between post-IVIg subacute and other samples (Figure 28). By ANCOVA, 16% of variance is explained by group ($p<0.001$) and an additional 2% by prior IVIg ($p<0.001$). Figure 29 shows only KD samples and indicates the expected proportion of sample immunoglobulin contributed by IVIg. The segregation of subacute post-IVIg samples is shown more clearly. No pattern is seen among mixed acute and convalescent samples dependent on estimated IVIg proportion.

[2] Children's circulating volume is estimated at 70 mL kg^{-1} therefore 2 g kg^{-1} of IVIg would contribute 28.6 g L^{-1} of immunoglobulin. The reference range for total serum immunoglobulin is 7 to 16 g L^{-1}.

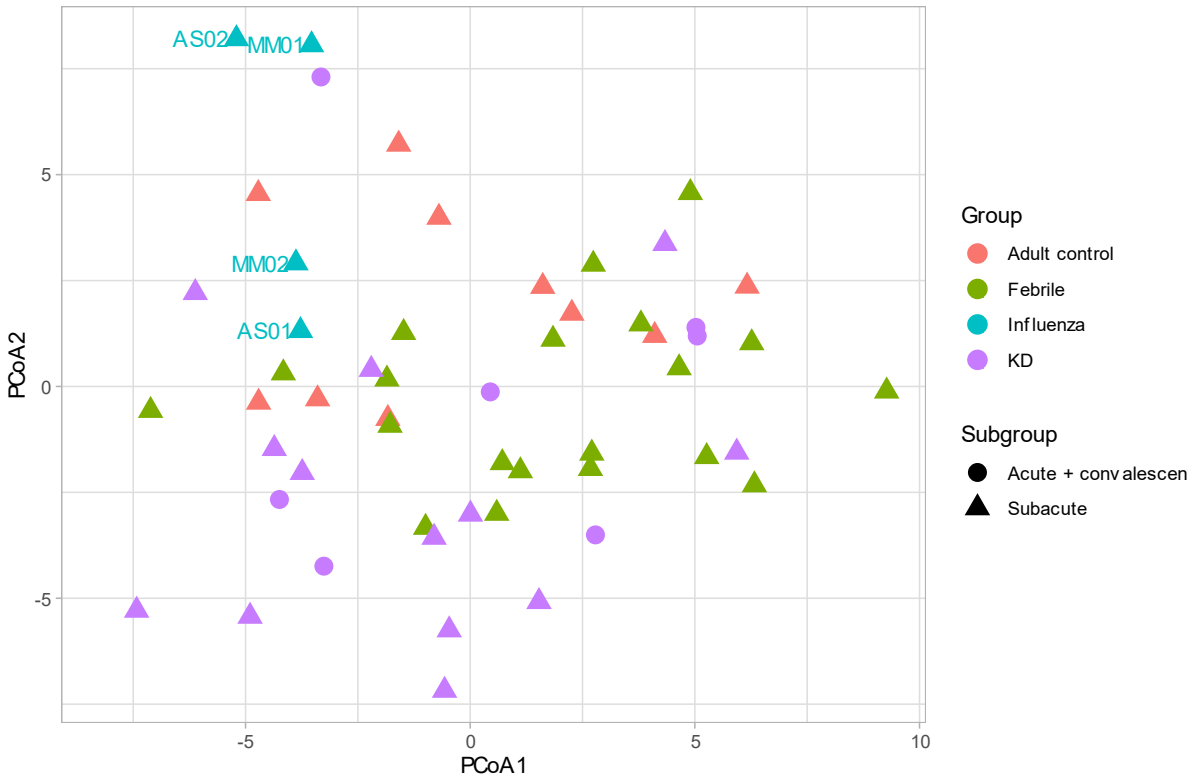

Figure 27 Principal co-ordinates analysis (PCoA) plot of Euclidean distances between Variable locus group and J locus relative abundances in samples processed at Bristol. The four Influenza samples comprise two individuals (AS and MM) with samples with (01) and without (02) added influenza vaccine.

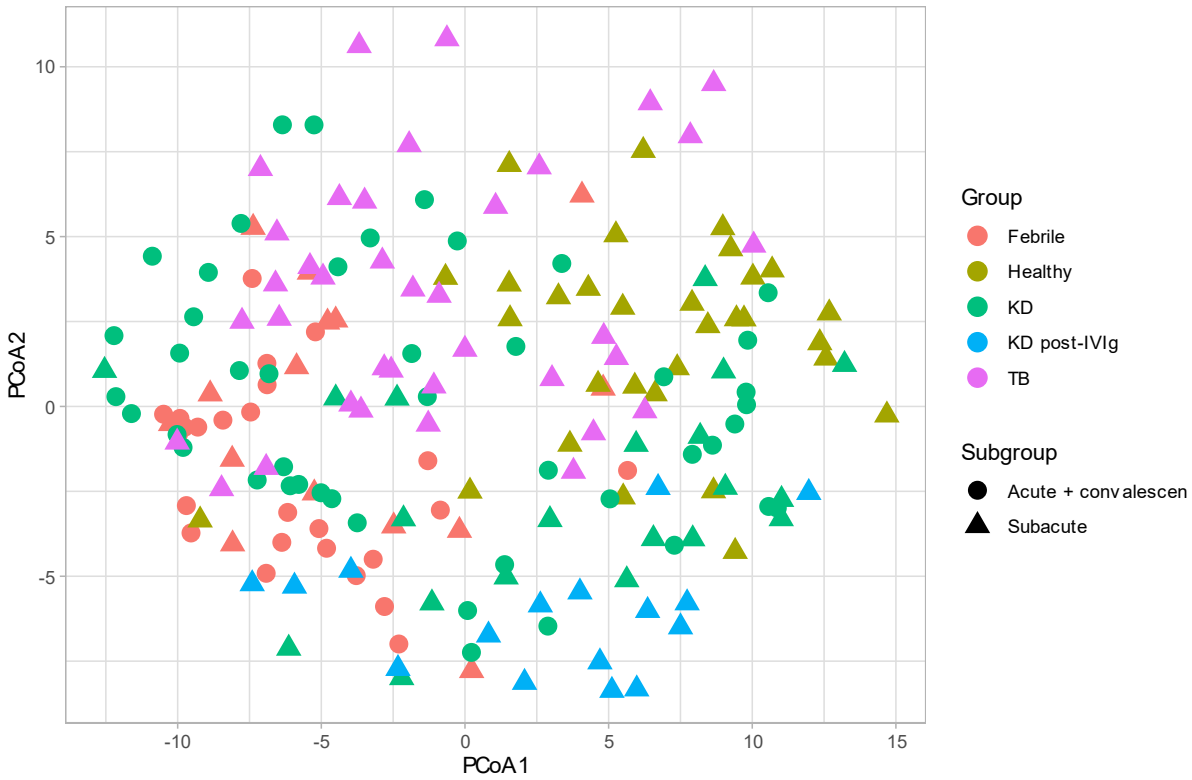

Figure 28 Principal co-ordinates analysis (PCoA) plot of Euclidean distances between Variable locus group and J locus relative abundances in samples processed at Oxford. KD patients who received intravenous immunoglobulin prior to sampling are indicated as a separate group. KD: Kawasaki disease.

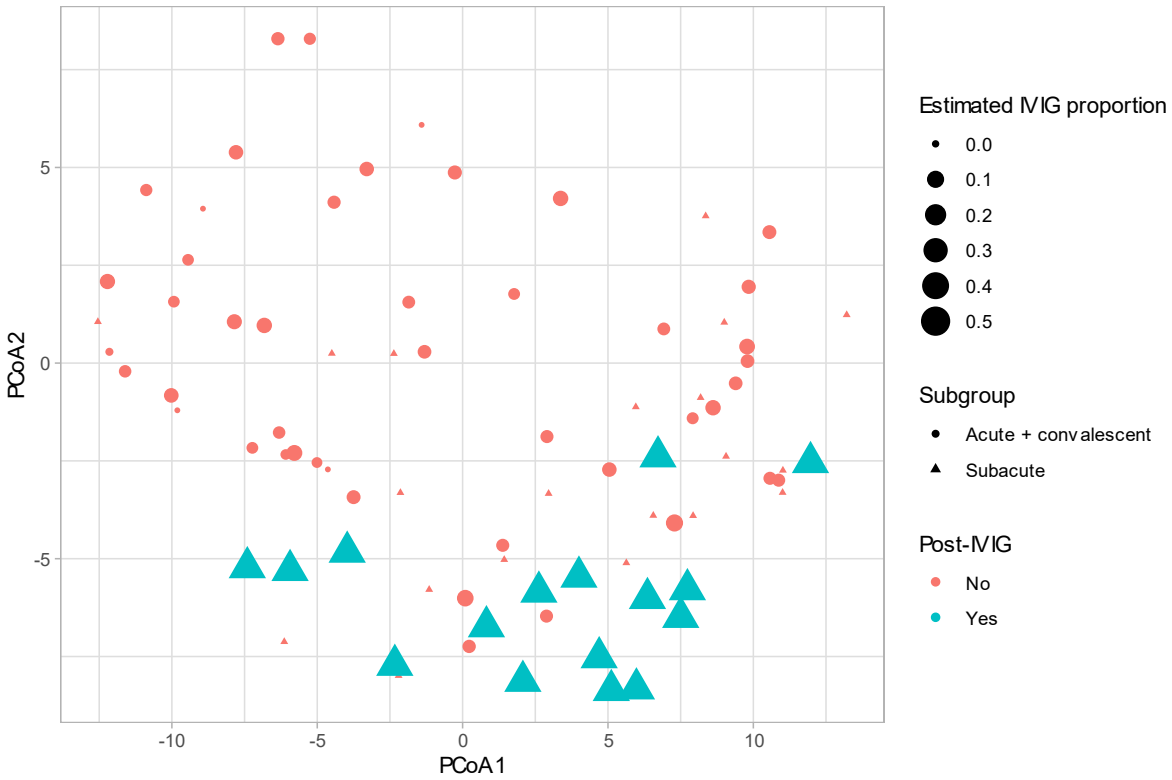

Figure 29 Principal co-ordinates analysis (PCoA) plot of Euclidean distances between Variable locus group and J locus relative abundances in KD samples processed at Oxford. Subacute samples from patients after receiving intravenous immunoglobulin (IVIg) are indicated. The size of points is proportional to the estimated proportion of immunoglobulin contributed by IVIg. This corresponds to 50% for subacute samples post-IVIg. For acute with convalescent samples, estimate is 25% with a 21-day half-life from day 7 of illness to the day of convalescent sampling. KD: Kawasaki disease.

Differential analysis

Based on the overall patterns above, subacute samples post-IVIg are treated as a separate group since differences can likely be ascribed to the bulk of IVIg contained. Samples mixing acute and convalescent blood are analysed with others, since the estimated IVIg proportion is low and no differences are apparent by PCoA. Relative abundances for multiple comparisons are shown for the variable and junctional loci (Figure 30 and Figure 31, respectively).

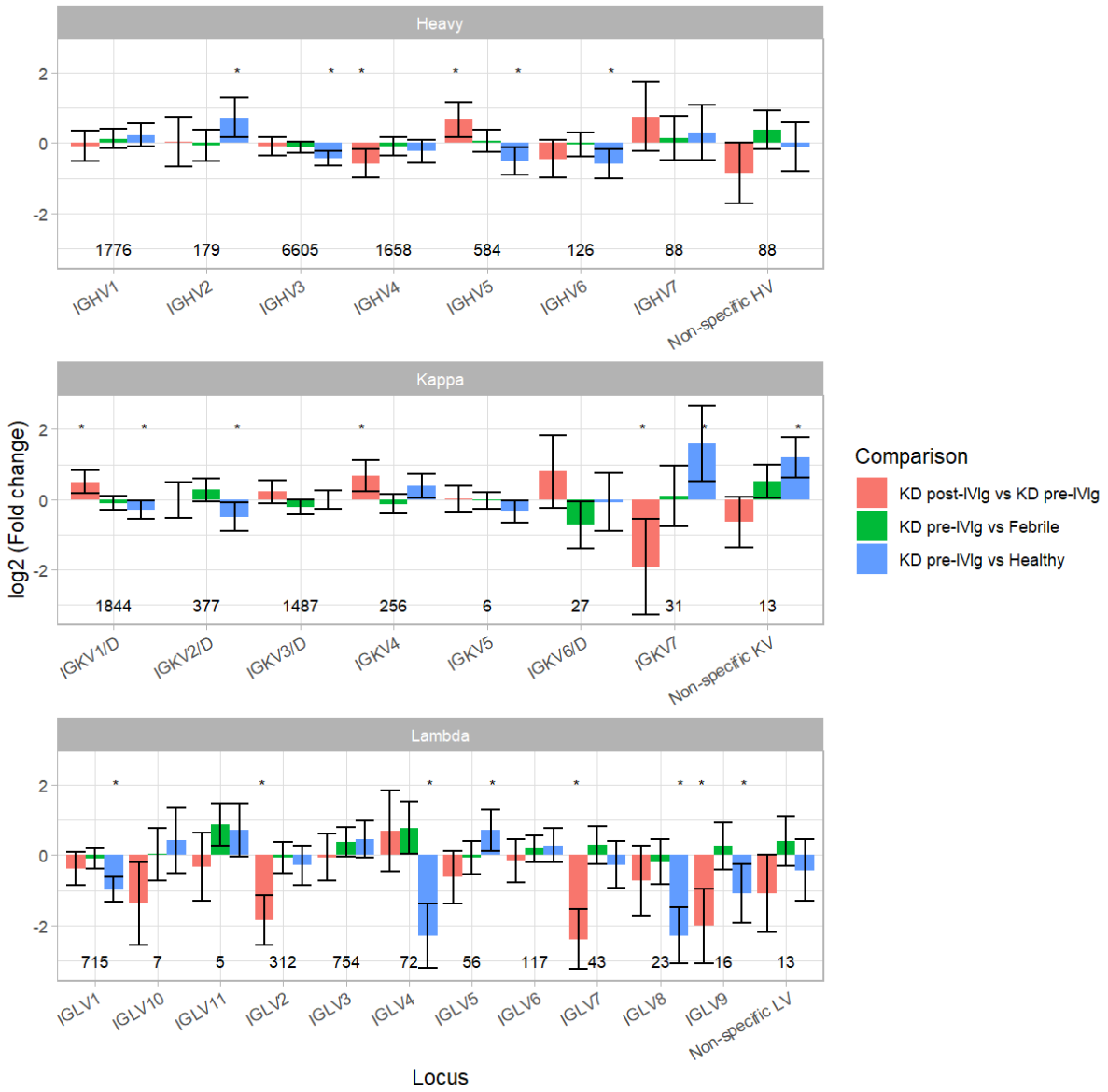

Figure 30 Limma comparisons of quantile-normalised variable locus peptide relative abundance by immunoglobulin locus group. Pre-IVIG KD samples (n=85) are compared to post-IVIG (n=17), febrile (n=60) and healthy samples (n=31). Additional covariates are sample type (plasma, serum or plasma and serum mixture), peptide count, laboratory, age and gender. Significance (false discovery rate 0.05) is indicated by asterisks and unadjusted confidence intervals are shown. Unique peptide counts by locus are indicated. KD: Kawasaki disease.

Discussion

In this chapter I have sought to explore further potential of the extensive proteomic data from PEG-precipitated immune complexes by analysing and comparing abundances of antibody peptides between KD and control samples.

There is limited prior literature on bulk antibody proteomics and I do not have any reference data on which to confirm the accuracy of quantitative measures applied to either constant or variable region immunoglobulin peptides. The approaches employed should be viewed as exploratory and demonstrating proof-of-concept. The in-depth analysis of antibody peptides incorporating structural

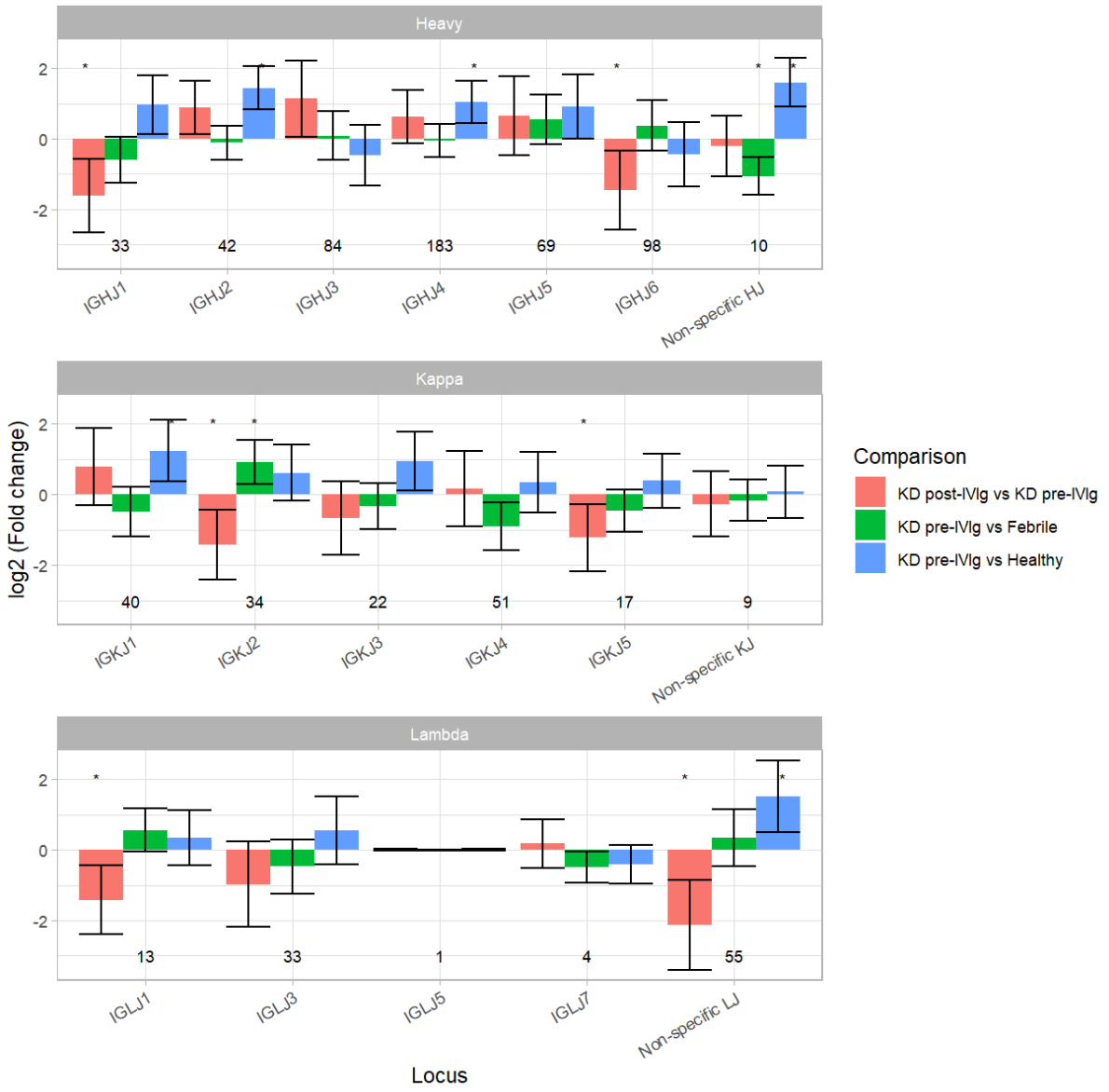

Figure 31 Limma comparisons of quantile-normalised junctional locus peptide relative abundance by immunoglobulin locus group. Pre-IVIG KD samples (n=85) are compared to post-IVIG (n=17), febrile (n=60) and healthy samples (n=31). Additional covariates are sample type (plasma, serum or plasma and serum mixture), peptide count, laboratory, age and gender. Significance (false discovery rate 0.05) is indicated by asterisks and unadjusted confidence intervals are shown. Unique peptide counts by locus are indicated. KD: Kawasaki disease; IVIG: intravenous immunoglobulin.

alignment and sequence similarity-based identifications has provided an excellent opportunity to develop bioinformatic skills.

Encouragingly, despite the abYsis antibody sequence database contributing fewer amino acids than the human sequence database (8.6 vs 11.4 million, with potentially high non-redundancy among abYsis proteins), antibody peptides accounted for a large proportion of identifications (39%) and most of the unique peptides (60%). It is reassuring that the peptide FDR is exceedingly well controlled for antibody proteins, with decoy matches all within non-antibody proteins. It was unexpected that the abYsis database should provide *no* decoy matches, and this finding is worthy of further exploration.

The paucity of cysteine-containing peptides is likely due to the lack of a reduction step to break disulfide bonds. Alkylation also did not take place. Since the search included carbamidomethyl cysteine as a fixed modification, peptides with free cysteine residues would not be identified. I tested a search without this fixed modification. The gains in peptide identifications and intensity were low especially when compared to the high proportion of cysteine-containing peptides in samples analysed in Bristol. Taken together, this supports the majority of cysteine-containing peptides retaining disulfide bonds.

Intensity-based quantification of cysteine-containing peptides has the potential to be uninformative due to limited and likely variable reduction of disulfide bonds. Both for this reason, and due to the limited gains, the whole search was not repeated. Additionally, since it could be empirically demonstrated that cysteine-containing peptides in Oxford have a high FDR, they were censored from further analysis.

MBR drives a large proportion of identifications (56%) and it should be noted that in the version of MaxQuant used, matching is not subject to FDR control. In a two-proteome experiment, 40% of identifications depended upon MBR, and "false transfers" were shown at the peptide level (Lim, Paulo & Gygi, 2019). Without MBR, 11 yeast peptides were identified in human samples, rising to 79 with MBR. However, this corresponded to only 2% of yeast peptides identified overall and typically the matches were to low intensity features. Thus, the unconstrained MBR FDR still appears low and the impact upon quantitation limited. FDR control of matches has recently been implemented in another software package, though too late for inclusion here (Yu, Haynes & Nesvizhskii, 2021).

Since any cysteine-containing peptide identified in Bristol and matched with features in Oxford can be assumed to be a false match, this provides an opportunity to estimate the frequency of false matches in these data between the two laboratories. There are 6 999 unique cysteine-containing peptide species (unique sequence, modifications and charge state) identified in Bristol alone by MS2 fragment spectra. Together, these have 5 044 matches with features in the samples processed in Oxford. This corresponds to fewer than one match in 190 samples per unique peptide species, and seems encouragingly low.

It is reassuring that coagulation proteins have such a clear stepwise association with increasing plasma component in both labs. Immunoglobulin represents a high proportion of estimated abundance as expected.

Most antibody peptides could be structurally annotated (97%), with ANARCI as a fallback where abNum numbering was unavailable. As expected based on the high proportion of IgG in circulating

immunoglobulin, constant region peptides were dominated by IgG both in terms of unique peptides, number of identifications and proportion of estimated IgG abundance.

IgG subclasses are more difficult to clearly describe due to the large proportion of peptide intensity corresponding to peptides shared between subclasses, and striking differences between laboratory in detection of IgG3. IgG3 contains many cysteine residues which take-part in inter- and intra-chain disulphide bonds (Schroeder & Cavacini, 2010). On manual review of the peptides detected by MS2 in Bristol and Oxford it was confirmed that Bristol's peptides included many cysteine residues, but the few Oxford peptides did not (data not shown).

Li (1989) studied immunoglobulins and immune complexes in KD, demonstrating elevated serum IgG1 and IgG3 compared to healthy controls, with IgG1 and IgG3 predominating in immune complexes. Failure to detect appreciable levels of IgG3 in samples at Oxford precludes comparison with healthy controls, however IgG1 is more abundant in KD samples (pre-IVIG; median 64-78.6%) than healthy controls and patients with TB (36.4 and 60% respectively). The rise in IgG2 relative abundance in post- compared to pre-IVIG samples is consistent with the relative paucity of IgG2 in younger children (Bayram et al., 2019), with median IgG2 levels rising from less than a quarter of IgG1 in the third year of life to over a half. IVIg is obtained from pooled adult serum and Gammagard (used at Rady Children's Hospital; Jane Burns, personal communication) has an IgG subclass distribution similar to adult serum.

Among variable region peptides, a majority could be identified based on top-scoring BLAST match or matches as belonging to a specific group of V loci, or single J locus. The heavy chain variable region had many more unique peptides identified than the light chain (11 209 vs. 6 128), though the total number of peptide identifications were similar. The lambda light chain J region had the fewest unique peptides identified (48), with other chain-regions having 161 to 10 694 unique peptides.

The abYsis database was selected since it contains directly sequenced monoclonal antibodies. These represent antibodies with functional significance – e.g. antigen or allergen specificity, pathological, immunological or scientific relevance. As an added benefit, the database sometimes includes data on antigen-specificity of antibodies (~2%), which provides a possibility of adding further significance to the data obtained.

Larger databases of untargeted BCR and immunoglobulin repertoires now exist, including from children (e.g. Ghraichy et al., 2020). These are more likely to represent a greater breadth of the immunoglobulin repertoire. However a study marrying rearranged B-cell sequences and shotgun proteomics of circulating immunoglobulin demonstrated little overlap (Chen et al., 2017). Indeed in

my own test using the Bristol proteomic data, adding cluster representative sequence data from Ghraichy et al. (2020) (Johannes Trück, personal communication) to the human and abYsis sequence databases led to only a 3% increase in antibody peptide identifications and a 10% increase in unique antibody peptides[3].

Coverage, both by number of unique peptides and mean peptide intensity are highly uneven. The constant region has, by definition, a limited range of unique peptides at any given residue given the limited classes, subclasses and polymorphisms, however by intensity frequently exceeds the variable regions especially in the heavy chain. Importantly, equal intensities of different peptides cannot be taken to imply equimolarity since elution profiles and efficiency of electrospray ionisation vary, and matrix effects give rise to interactions between co-eluted peptides (Taylor, 2005).

The most diverse range of peptides is found between the N-termini of CDR2 and CDR3 on both heavy and light chains. The patterns are broadly consistent between samples analysed in Bristol and Oxford, with differences likely related to the second cysteine residue on both heavy and light variable chains.

CDR3 coverage is low in both heavy and light chains. This was also noted by VanDuijn et al. (2017) who ascribed it to the long length of predicted CDR3 peptides. However, there are further factors likely to contribute. CDR3 is the region of greatest diversity, due to nucleotide contributions from V, D and J loci as well as junctional diversity. It can thus be expected that the 77 619 abYsis amino acid sequences cover only a small proportion of the CD3 sequences existing in circulating immunoglobulin in any individual.

Furthermore, because of the diversity of CDR3 peptide sequences in each sample, CDR3-containing tryptic peptide species can be expected to be very diverse and each individually of lower abundance than the less diverse non-CDR3 peptides. This would result in less frequent selection for fragmentation by the mass spectrometer. Lower abundance also corresponds to lower charge accumulation in the mass spectrometer and lower quality MS2 fragment spectra.

By PCoA, V and J locus abundances show clear patterns of difference between each group of samples, and a clear signal relating to recent IVIg treatment. Notably, there is a lack of segregation of acute-only and acute-with-convalescent subgroups within KD and Febrile patients. Remaining IVIg in the acute with convalescent samples could be expected to drive differences, but no clear pattern is seen with the crude estimated remaining IVIg.

[3] 3 042 additional identifications of 1 048 peptides, compared to 105 413 identifications of 10 709 peptides

It is interesting to note that paired adult samples with and without spiked influenza do not cluster in their pairs – although there are no biological replicates available to assess variance. The addition of influenza should preferentially draw influenza-specific antibodies into high molecular weight complexes and was already shown to lead to elevated IgG1 proportion (chapter 6), thus a shift in profiles of V and J segment peptide intensities may be well expected.

When comparing the V locus group and J locus intensities between sample groups, considering the small group sizes, striking differences are seen between pre- and post-IVIg KD samples, with 9 of 28 loci showing significant differences (all but one reduced intensity). Since IVIg is formed by pooled serum from many individuals, diversity and evenness are likely to be increased.

When pre-IVIg KD samples are compared to febrile and healthy controls, distinct sets of differences are observed. In comparison with febrile samples, only reduced non-specific heavy chain J segment peptides, increased kappa chain J2 and reduced J4 are shown, with approximately one-fold change for all. In comparison with healthy samples, a greater number of differences are evident. Only small differences are observed in heavy chain V locus group intensities, but more marked increases in J1, J2, J4 and non-specific J locus segments. In the light chains, increased kappa V7 and J1 and reduced lambda V4 and V8 are observed.

There is limited literature on B-cell receptor and antibody variable locus selection in disease states. Over-representation of IGHJ4 has been demonstrated in combination with a range of V segments. For example, with IGHV4-28 in primary immune thrombocytopenia, with IGHV1-5-7 in immune thrombocytopenia in chronic lymphocytic leukaemia, with IGHV3 loci in CMV infection, and with IGHV4-31 in expanded EBV infected lymphoblastic cell lines (Wen et al., 2020; Hirokawa et al., 2019; Watson & Breden, 2012; Watson, Glanville & Marasco, 2017; Visco et al., 2012; McLean et al., 2005). Recently, expansion of IGHV4-34 plasmablasts has been shown in a child with KD

In these data, the smaller apparent difference between KD and febrile samples as compared to KD patients and healthy controls suggests that some of the observed differences in KD may be more general features of inflammatory and/or infectious diseases. No distinct signature has been extracted or demonstrated.

There are significant limitations to the interpretation of analyses presented here. Proteomic data is more typically analysed at protein level with multiple levels of false discovery rate control. Here, only PSM-level FDR control can be applied since individual antibody variable peptides can map to an unlimited number of antibody proteins.

FDR control focuses on controlling the rate of spurious matches. However, on the spectrum from spurious to perfect matches, there also exist matches which are partial and imperfect. These occur where a theoretical peptide possesses near-identical mass to an MS1 feature, and matches sufficient fragments within the mass error of the ion trap (on the order of 1 Dalton) to generate a good score, despite not possessing the identical amino acid sequence.

Most trivially, this can occur when isoleucine is replaced with leucine, as they share the same exact mass. However, within the fragment mass error, many single amino acids and two-amino acid pairs share similar masses. Additionally, typically not all possible fragments of a predicted peptide are observed. Missing fragments can render spectra consistent with multiple possible arrangements of certain residues. The existence of chimeric spectra (where multiple peptide precursors are fragmented together) can further confound accurate identification.

Although the absence of matches to antibody decoy sequences indicates an extremely low FDR, this does not guarantee that all matches are correct for exact amino acid composition and ordering. MaxQuant will report the best match for each spectrum, and is limited by sequences represented in the database. Search strategies are available which re-search spectra to identify possible divergences from database sequences, designed for the detection of protein variants which may not be represented in databases and a wide range of post-translational modifications (e.g. dependent peptide search in MaxQuant and SPIDER in PEAKS; Tyanova, Temu & Cox, 2016; Han et al., 2011). However, these add significant processing time, add extra complexity to downstream analyses, and are limited to identifying variants of already-identified peptides.

Beyond this, the identified peptides only scratch the surface of the complexity of peptides in the samples. This relates firstly to the standard intensity-based selection method of Data Dependent Acquisition (DDA) in which the most intense peptide precursors are selected for fragmentation (with time dependent exclusions to prevent repeated fragmentation of the same peptides). This is illustrated in Figure 32. Of features within the fifth decile of intensity, 20% are fragmented, and of these 3% identified. In the top decile, 66% are fragmented, and of these 16% identified. Thus, within a sample, identified peptides are heavily biased towards those which are high intensity. With MBR, identified peptides will be expected to be biased towards those which are high intensity across multiple samples.

Taking this into account, it is encouraging that differences can be seen between groups, and that there is a biologically plausible signal of difference between post-IVIG and pre-IVIG KD samples, with IVIG-containing samples distanced further from controls in PCoA.

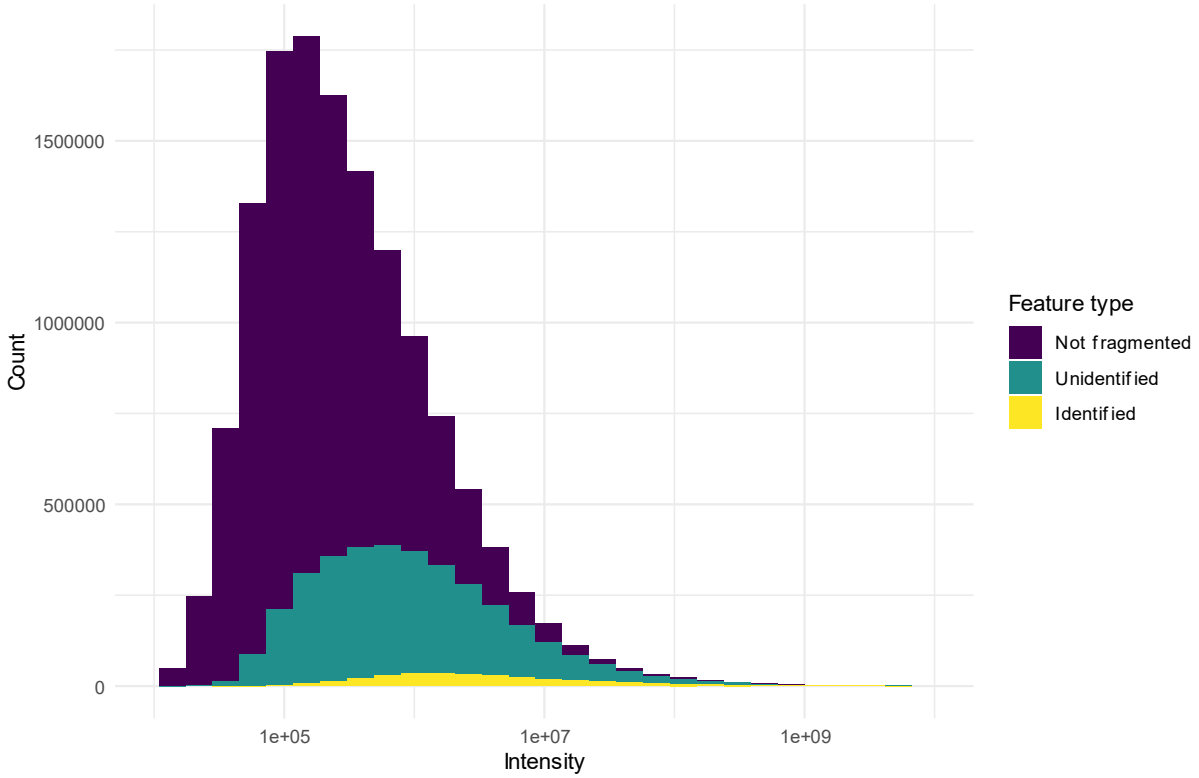

Figure 32 Stacked histogram of MS1 feature intensity distributions in samples processed at Oxford, classified by whether the precursor was selected for fragmentation, and whether fragmented features were identified. MS: mass spectrometry.

Importantly, identified antibody peptides can't be joined into full antibody sequences, nor linked with specific isotypes. This limits how the results can inform antigen specificity or effector functions. CDR3 is of most importance to antigen specificity and is least well covered (see Figure 24 and Figure 26). Even if it were possible to reconstruct longer antibody sequences, our ability to make inferences about function using bioinformatics are limited, though growing (Kovaltsuk et al., 2017, 2018; Krawczyk et al., 2018; Kwong et al., 2017).

In this sense, the identification of differences in antibody peptide signatures between phenotypic groups of children, with further delineation based on the presence of recent IVIg treatment, can be taken as a proof-of-concept for more detailed investigations.

There are several ways in which this work could be taken forward for the investigation of antibody profiles in KD:

1. Performance of proteomic methods for identifying and quantifying antibody peptides could be assessed in a controlled manner using individual monoclonal antibodies and mixtures thereof, with known sequences and relative abundances.
2. Validation in human serum samples or precipitates could be undertaken with spike-in experiments, where known monoclonal antibodies are added in varying concentrations.
3. Assessment of variance with biological and sample replicates

4. Optimisation of sample preparation, for example cleavage and purification of antigen binding fragments, as in Liu et al. (2011).
5. Optimisation of data acquisition on the mass spectrometer – to be considered in the following chapter
6. Peptide identification could be optimised through alternative analytic approaches – to be considered in the following chapter.

8 | Discussion

This work has as its central focus the aim to find microbial triggers of KD. The extensive literature review (p26) evidences the widespread interest in this question, which is of more than academic significance. Knowing the triggers of KD opens the door to better exploration of aetiopathogenesis, especially when integrated with knowledge of genetic and other environmental factors. It also allows means of prevention to be considered.

The study data analysed here are derived from over four hundred individual children, comprising 202 KD patients and 159 febrile controls providing metagenomic or metaproteomic data (28 both), and a further 68 healthy controls and TB patients providing additional metaproteomic data. Organisms associated with KD have been sought in the oropharyngeal microbiome, and antigens within precipitated immune complexes from the blood.

Both metagenomic and metaproteomic analyses have been highly involved and challenging. Careful method and tool selection has been required. There have been many hurdles to overcome, both in the application of software packages, which are frequently novel, niche and/or utilised in atypical ways, and in the downstream analysis.

The key output of these analyses is evidence of a higher abundance of certain bacterial species in the oropharynx of patients with KD as compared to febrile controls, and no convincing evidence for any specific microbial proteins within immune complexes.

Before reviewing the findings, I will describe the impact of the SARS-CoV-2 pandemic on this work while also introducing a new pandemic-related condition of potential relevance to understanding the aetiology of KD.

SARS-CoV-2 pandemic – impact and information

At the onset of the global SARS-CoV-2 pandemic, metagenomic sequence data was available from only one flow cell. It took until December 2020 to obtain the sequence data for the full cohort, due to both facility shutdown and the prioritisation of pandemic-related studies. Metaproteomic data was available at the outset of this project.

The pandemic has provided new knowledge of relevance to the understanding of KD. In April 2020, clinicians in London (including colleagues from within Imperial College and Imperial College Healthcare NHS Trust) recognised a wave of children, mainly teenagers, and almost all over 5 years of age, presenting with a constellation of protracted fever, myocardial dysfunction and shock, diarrhoea, abdominal pain, rash and conjunctivitis (Whittaker et al., 2020). A significant proportion required

intensive care admission for cardiorespiratory support. Presentations appeared to be temporally related to SARS-CoV-2, with high rates of positive serology, and the incidence curve following the community epidemic of SARS-CoV-2 by a delay of around a month. Some children developed CAA. The condition has been termed Paediatric Inflammatory Syndrome Temporally Associated with SARS-CoV-2 (PIMS-TS) in the UK, and multi-system inflammatory syndrome in children (MIS-C) in the USA.

Similarities with KD were noted immediately, due to the shared occurrence of CAA and features of fever, rash and conjunctivitis. Notably, a minority of KD patients also develop myocardial dysfunction and shock (Kanegaye et al., 2009). However, very different ethnic preponderances were noted, with under-representation of children of East Asian ethnicity, and over-representation of children of Black and South Asian ethnicity. Further, the behaviour of CAA appears different with greater spontaneous resolution among children with MIS-C (Felsenstein et al., 2021). Children with MIS-C are typically several years older than children with KD, most frequently adolescents, with pre-school children and adults rarely affected, despite exposures to SARS-CoV-2 occurring across all age groups.

The differences suggest that although both conditions could represent post-infectious self-limiting inflammatory disorders in genetically predisposed individuals, they may be aetiologically and pathologically somewhat distinct. It is hoped that the developing understanding of MIS-C aetiopathogenesis may also shed more light on KD (Sancho-Shimizu et al., 2021). Already, investigators have shown modest prevalence of monogenic susceptibility to inflammation (Chou et al., 2021). The concentration of MIS-C among adolescents suggests an age-related immune susceptibility in combination with genetic predisposition and SARS-CoV-2 exposure.

In the first months of the pandemic, I contributed to paediatric clinical care through a half-time placement at St Mary's Hospital, allowing more junior trainees to be redeployed to adult COVID care. In the early part of 2021, I took on a short-term surge role at Great Ormond Street Hospital to assist in the management of MIS-C patients and recruitment into a range of studies.

I was also involved from inception in May 2020 in a global retrospective cohort study of children with MIS-C, the Best Available Treatment Study. The primary focus was to compare the efficacy of different immunomodulatory treatments. I was seconded to lead the analysis of this study for six months from January 2021, leading to publication in the New England Journal of Medicine (McArdle et al., 2021).

As SARS-CoV-2 approached global endemicity and restrictions were relaxed in most parts of the world, outbreaks of hepatitis were noted, predominantly among pre-school children, and not explained by the typical infectious or non-infectious causes. Significant numbers of children required liver transplantation. Extensive investigation, much undertaken at Great Ormond Street Hospital, and

including metagenomic analyses of liver biopsies, showed strong association with adenovirus and adenovirus-associated virus 2 (UK Health Security Agency, 2022). Intriguingly, these viruses did not seem to represent novel strains, and direct viral-mediated lysis of hepatocytes was not demonstrated. A strong human leukocyte antigen allele association was also shown. This leads to the hypothesis that delayed exposure to these ubiquitous viruses (due to reduced transmission during the pandemic) can result in this rare outcome among predisposed children because it hits a window where a pathogenic immune response can be triggered. Conversely, during the pre-pandemic era, exposures would typically be at an earlier age when the immune system of susceptible individuals is not primed for such a response.

The lessons from both MIS-C and this paediatric hepatitis outbreak are of great potential relevance to the search for a cause of KD, highlighting both the potential for windows of immunological susceptibility, and the potential for multiple microbial causes to be involved.

In summary, the SARS-CoV-2 pandemic, while disrupting and delaying this research, has provided additional knowledge of relevance to the aetiology and pathogenesis of KD and broadened my experience during this fellowship.

Metagenomics

The metagenomic analyses depended upon considerable preparatory bioinformatics, including that presented in Appendix B. The analyses themselves followed a stepwise approach through quality control, identification of contaminants and spurious identifications and exploratory analyses (chapter 2, p35) before proceeding to species level (chapter 3, p59) and strain/pangenome analyses (chapter 4, p79).

Read quality and depth was shown to be adequate and significant GC bias absent. Read binning with Kraken 2 was fine-tuned through adjustment of the confidence parameter and base quality threshold. Bracken was applied to reallocate reads at species level. I recognised and corrected for the under-appreciated problem of human-contaminated reference sequences (Breitwieser et al., 2019; Kryukov & Imanishi, 2016), and identified potential contaminants using a range of methods, applying a publicly available tool (Davis et al., 2018), self-developed modelling and manual curation. Ultimately, contamination, although typically at very low levels in each sample (median 0.05% of reads) was sometimes very high as a proportion of bacterial reads. The extensive consideration of contaminants and database errors I present exceeds that frequently seen in the metagenomic and microbiome literature (Eisenhofer et al., 2019) and can be considered as a strength of this study.

I explored the factors underlying the low bacterial DNA proportion in KD samples, concluding this is driven by overabundance of human DNA, likely due to pharyngeal inflammation. This is important to understand for future studies of the pharyngeal metagenome. I also detected evidence of "index hopping" to negative control samples – since this depends upon free index tags remaining after clean-up, it is likely to affect less the samples with extracted DNA where index tags should be consumed.

I showed broad comparability of these microbiome data with existing data, with unsurprising divergence from nasopharyngeal data. In contrast with typical microbiome studies, I did not expect (indeed *hoped* not to find) large differences between the microbiome in KD and febrile controls. Rather, I intended to seek differences at the scale of only a small number of species. Reassuringly, KD explained only a small proportion of microbiome variation (2.7%) – less than age, though more than sex and country. Antibiotic exposure explained a similar proportion of variation.

In the multiple approaches and sensitivity analyses undertaken, a diverse range of streptococcal species (consistent with normal oral flora) showed positive associations with KD, along with *Rothia dentocariosa*, *Prevotella oris*, unnamed *Neisseria*, *Abiotrophia defectiva* and *Lautropia mirabilis.* The taxonomic levels at which results were reported differed by approach, due to the treeclimbR algorithm. The most specific associations, *A. defectiva*, *L. mirabilis*, *R. dentocariosa* and *P. oris*, were only slightly weakened when antibiotic exposure was taken into account, despite the reduced sample size, and *Parvimonas micra* was added as another organism with relative abundance positively associated with KD. Notably, accounting for the specific pattern of exponential increase in RA of *L. mirabilis* and *A. defectiva* through early childhood increased the strength of association with KD.

Given the high prevalence of these organisms, identifying genetic differences between organisms in KD patients and controls could provide additional support to any aetiological role which would be consistent with the epidemiology. With two complementary approaches, considering both genome composition (pangenomics) and nucleotide-level strain variation, no meaningful differences could be identified. Nonetheless, the pangenome approach was limited by the paucity of reference metagenomes available, leading to a dependence on potentially erroneous metagenome assembled genomes. The strain-level approach was limited by coverage – not all samples containing each organism could be included. A separate analysis looking at all predicted microbial genes from the assembled reads found no genes significantly over-represented in KD.

The potential for ubiquitous and commensal organisms such as these to be involved in triggering KD leads requires further consideration. The lack of clear sub-species level differences reduces the relevance of hypotheses involving toxin genes or specific strains. Were a ubiquitous non-commensal (or transient) organism triggering KD and present at the time of presentation, a statistical association

of prevalence (or abundance) should be easy to demonstrate, whether with healthy or febrile controls. This stands even if genetic predisposition and a temporal window of susceptibility were also necessary.

The situation becomes more challenging to study if the causation relates to the acquisition of a ubiquitous commensal in predisposed individuals within a window of susceptibility. In this instance, the organism may be highly prevalent and reasonably abundant in febrile and healthy children and those with KD. It may be expected that when a commensal is first acquired it could reach a high abundance until immune regulation and/or adaptive competition from other components of the flora restrict its growth. Therefore, one may expect that children with KD would have more recently acquired one or more of these organisms and have them in high abundance, whereas acquisition by age-matched controls would be likely more distant (or yet to occur) and thus abundance would be lower.

The findings presented here are consistent with such a model, but any hypothesis for a causal role of such commensals would also need to explain seasonal, temporal, spatio-temporal, geographic and meteorological patterns in KD. Spatio-temporal clustering of cases could still be explained if acquisition of species of the oral flora occur appreciably outside the household. Seasonality could then be explained if such transmission occurred more readily in the Winter (in the Northern hemisphere). Temporal trends in incidence (rising in many regions) could be explained if the age of acquisition of triggering microbes was shifting over time (whether earlier or later) in such a way as to align more with the window of immunological susceptibility. Associations with tropospheric winds would be harder to fit into this framework unless it could be shown that components of the normal flora could be acquired by long-range airborne transport.

The approach developed here has several strengths, including the large sample size, extensive consideration of contaminants, sensitive approach to organism identification and abundance estimation and analysis at multiple taxonomic resolutions, including strains and genes.

However, important limitations remain. There was very limited identification of fungi, archaea and phage and non-phage DNA viruses, and RNA viruses could not be considered. DNA viruses causing clinical upper respiratory infections are few: adenovirus and herpes simplex virus being the main examples. Given the prior implication of coronaviruses in KD aetiology (Patra et al., 2022) and the novel SARS-CoV-2-associated inflammatory disorder described above, this is a significant limitation.

The emergence of MIS-C as a SARS-CoV-2-triggered inflammatory disorder, occurring with a delay between infection and disease of about a month suggests an additional complexity in searching for

the causative pathogen. If KD were similarly triggered by a virus or bacteria, which infected the child a month earlier, our metagenomic study may have missed identifying the pathogen.

Archaea have not been specifically studied in the oropharynx, though a single study developed archaea-specific primers and applied them to the nasopharynx, among other sites (Koskinen et al., 2017). Most human and animal studies of archaea have focused on the gut (Borrel et al., 2020). Regarding fungi, *Candida albicans* can colonise and cause infection in the oropharynx. In the normal microbiome, fungi are estimated to contribute <0.1% by colony forming units. Nonetheless, a targeted study detected over 75 genera of fungi (Ghannoum et al., 2010).

Future attempts to seek oropharyngeal organisms associated with KD could gain further strength in a number of ways. Given the lack of GC bias introduced by the low cycles of PCR applied, this could be increased leading to greater DNA abundance in further processing steps, and less sensitivity to laboratory contamination. A separate pipeline with reverse transcription could allow detection of RNA viruses, though again human RNA has been shown to dominate (90-95% in a study by Nakamura et al. [2009]).

Approaches based on assembled metagenomes are attractive and reduce database dependence, however sensitivity to low-abundance organisms is likely to be very low. The sensitivity of the metagenomic analyses here will have been impacted by the high, variable and biased proportion of human DNA in samples. As mentioned in Appendix B, although laboratory methods exist to deplete human DNA prior to sequencing these may adversely impact detection or quantitation of some organisms, especially those dying or under immune attack (Oechslin et al., 2018). Perhaps more promising is the dynamic exclusion of host reads from the Nanopore sequencing platform, as made possible with tools like Readfish (Payne et al., 2021).

Metaproteomics

Metagenomics is an established field with 27 200 publications indexed by PubMed and nearly 5 000 publications in 2021. By contrast, metaproteomics is a small niche, with 904 publications and only 144 in 2021.

This aspect of the study provided considerable challenges (chapters 5 and 6; p92 and 105 respectively), since it required methods to identify likely low abundance microbial antigens within exceedingly complex, antibody-rich samples with limited justification for prior database restriction. In this way, the study represents an atypical application of niche processes.

Much time was invested in identifying and testing potential computational approaches to database-dependent peptide and protein identification. However, the field is dominated by tools focused on the gut microbiome and many packages are orphaned after first publication.

I chose to apply two promising and complementary approaches. Metanovo provided an effective and computationally tractable means of reducing a large database prior to conventional proteomic searching (Potgieter et al., 2019). Quandenser represents a radical departure from typical proteomic analyses, similar in many ways to metabolomics – identifying and quantifying features first, then selecting those with relevant biological signals before moving to identification (The & Käll, 2020).

These two approaches together could not provide evidence of any specific microbe being associated with KD. Indeed, the near identical intensity distribution of microbial and non-microbial peptides was not consistent with origins in background contamination and low abundance signal as expected.

Antibody proteomics

Comparing the profiles of antibody sequences in precipitated immune complexes between disease states can provide a window into the KD immune response and potentially, pathogenesis. In the secondary analysis presented in chapter 7 (p121) I showed clustering of KD, febrile, TB and healthy samples when V and J locus relative abundances were compared, and a striking signal evident with sampling post-IVIg treatment. Differential abundance analysis highlighted specific V and J loci with significant signals, with the strongest signals unsurprisingly corresponding to IVIg treatment. Differences were also evident between pre-IVIg KD samples and both Febrile and Healthy controls, though the functional interpretation of these is limited at present.

Limitations and improvements to proteomic methods

In both the metaproteomic and antibody proteomic chapters, I deferred discussion of most improvements that could be considered in the sample processing, mass spectrometry and bioinformatic analyses due to the shared considerations. I highlighted that both approaches could be improved by fractionation; that is, purification of antibody variable fragments for antibody analyses, and depletion of the same for metaproteomic analyses.

The higher raw feature intensities and the high frequency of “orphaned” MS2 fragment spectra without identified MS1 isotopic feature clusters in samples processed in Bristol suggests that peptide loading may have been excessive. Overloading can widen peaks (“peak tailing”) and increase ionisation competition (a matrix effect) whereby some peptide species are preferentially ionised, constraining the ionisation of other peptide species (Maia et al., 2020). These effects can reduce the number of identifiable peptide species. Overloading can also impair quantitative accuracy by leading to detector

saturation. Nonetheless, unlike in studies where relative quantitation is vital, in this metaproteomic study detection is key. Improved identification of low abundance peptides would be an acceptable trade-off against loss of quantitative accuracy.

For complex samples like these, depth of identification could be improved with better separation through peptide fractionation or two-dimensional liquid chromatography (the latter has been shown to increase identifications in metaproteomics, e.g. Hinzke et al. [2019]). However, both add considerable time, cost and material resource requirements, whilst potentially reducing reproducibility and simplicity.

On the mass spectrometer, acquisition of high-resolution MS2 fragment spectra would likely provide a major improvement (Mann & Kelleher, 2008; Scherl et al., 2008). In this approach, MS2 fragments are measured with accuracies on the order of a few parts per million in the Orbitrap, rather than almost a Dalton on the linear ion trap. The cost of this is slower acquisition of spectra, since the Orbitrap is employed to scans both precursors and fragments, rather than the Orbitrap and ion trap operating in parallel.

A key advantage of high fragment mass accuracy is that near-isobaric amino acids can be distinguished. Novel fragment-index searches can be applied, eschewing the typical search for peptides within the precursor mass tolerance (Yu et al., 2020; Chi et al., 2018). Rather, predicted fragments can be indexed and experimental fragments searched against these indices. An experimental spectrum may have a very good fragment-based match to a theoretical peptide, even though there is a large precursor mass differential ("delta mass"). This allows unanticipated modifications to be identified. High resolution fragments also allow for much more accurate *de novo* sequence generation, which can mitigate against database incompleteness (Muth & Renard, 2018).

A further computational improvement is FDR-controlled matching between runs, as implemented within FragPipe (Yu, Haynes & Nesvizhskii, 2021) or the standalone IceR (Kalxdorf et al., 2021). The process is similar to that used for peptide-spectrum matches (describe on page 100) and described for FragPipe in Figure 33. This would improve the confidence of peptide identity propagation between samples. Finally, peptide-spectrum match FDR control incorporating retention time and fragment ion

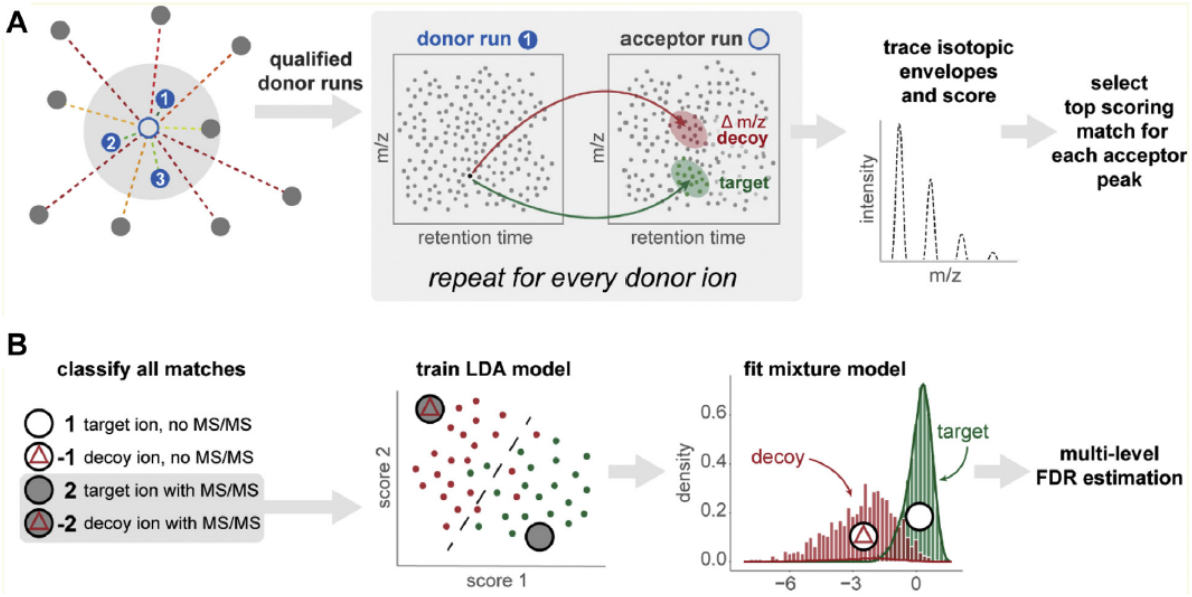

Figure 33 The IonQuant component of FragPipe implements FDR control for feature matching between runs (reproduced from Yu, Haynes & Nesvizhskii [2021]). (A) Samples (runs) are paired with the most similar runs for matching. Target (green circle) matches are identified within specified tolerance, and decoys with an offset m/z. Matches are scored, and the top-scoring match selected for each acceptor peak, which could be a decoy or target peak. (B) A model is trained to discriminate target from decoy matches and allow FDR estimation and thresholding. FDR; False discovery rate.

intensity prediction with machine learning has been shown to drive large gains in confident identifications (Verbruggen et al., 2021; Guan, Moran & Ma, 2019).

Limitations – the challenge of confidently identifying causative agents

I have presented organisms with increased abundance in the oropharynx in KD, with associations that are robust to control for likely confounders. However, important limitations remain. The identified organisms are near ubiquitous, and their presence is not specific to or associated with KD. No differences can be found to suggest transmission of specific subspecies or strains.

The finding of increased abundance of ubiquitous organisms does not lead to straightforward explanations in the context of the known epidemiology (e.g. spatial clustering and strong meteorological associations), unless an environmental factor can be shown to stimulate proliferation and activation of certain organisms.

Beyond the technical challenges of metaproteomics and metagenomics, as encountered and discussed at length in the respective chapters, there remain many reasons why finding a causative agent of KD might be difficult in samples taken after the onset of disease (Table 23 row 5).

Limitation	Example/explanation	Pre-analytic mitigations	Analytic mitigations
1. KD is a syndromic diagnosis which shares features with other infectious and infection-related diseases.	Cohorts of patients diagnosed with KD could be an admixture of a (hard-to-define) true KD and other disorders. Examples of other diagnoses include: measles, adenovirus infection and scarlet fever (caused by group A streptococcus)	• Detailed clinical data on participants. ✓ • Well-defined inclusions and exclusions for cases and controls. ✓ • Multiple controls sets. ✓	• Sensitivity analyses: e.g. restrict cases to those with more KD-specific features, e.g. CAA ×
2. There may not be single final microbial cause of KD for all patients.	Multiple agents may be able to cause KD in isolation, and could vary geographically, seasonally, temporally and between individuals for reasons of exposure, immunity or genetics.	• Application of techniques which can identify many microbes in parallel. ✓ • Large sample size to strengthen signals from small subpopulations. ×	• Epidemiologically-informed analysis: e.g. incorporation of temporal/symptomatic clustering ×
3. There may not be a single microbial cause for each individual patient.	Multiple microbial agents may interact, or a single agent may interact with additional environmental or other factors, whether simultaneously or sequentially.	• As above	• Exploration of microbiological covariation / dimensionality reduction and associations with KD ×
4. As necessary but not sufficient causes, causative microbes could be prevalent in temporally-matched controls	If host genetics is the largest causative factor and causative agent(s) are common, they may be similarly prevalent between cases and controls. Many potential causative agents are frequent parts of the normal flora in healthy individuals, or cause asymptomatic infections.	• Mitigation here is challenging, since avoiding temporal matching would give rise to confounding signals from seasonal organisms.	• As indicated to the left.
5. The microbial cause may be absent or undetectable at the time of presentation	Since KD is a self-limiting inflammatory disorder, the arrival and departure of any microbial	• Sample processing and data acquisition techniques to increase sensitivity to low	• Analytic techniques to increase sensitivity to low abundance microbes or antigen,

Limitation	Example/explanation	Pre-analytic mitigations	Analytic mitigations
	trigger could precede illness or presentation, due to an immunological "incubation period," as clearly exhibited in MIS-C (Nakra et al., 2020)	abundance microbes or antigen, e.g. optimise dynamic exclusion of already-fragmented features. ± • Prospective cohort study design recruiting infants at high risk for KD and sampling repeatedly. Target sufficiently large sample size and prolonged follow-up to ensure high numbers of KD cases occur. ×	e.g. increase identification of low abundance MS1 features. ±
6. The microbial cause may not be present in the oropharynx or ICs.	For example, if the causative agent enters through the lower respiratory tract, it may be difficult or impossible to detect in the pharynx. ICs could be composed purely of self-antigen and antibody, even if specific antibody against a microbial antigen is involved.	• More extensive sampling, although collection of sputum from children is challenging, and invasive lower respiratory sampling cannot be undertaken for research purposes alone. ×	• Irremediable
7. The level of specificity at which any microbial cause is defined is not known.	For example, a set of causative agents could be an entire genus of organisms, a group of related species, a single species, strain, or simply an organism or group of organisms carrying a particular pathogenic gene. For example, diarrhoeal diseases can be caused by *Escherichia coli* with a single enterotoxin gene, and haemolytic-uraemic syndrome by the same with a verotoxin gene.	• Irremediable	• Multilevel analyses with appropriate taxonomies and effective FDR control. ✓

Limitation	Example/explanation	Pre-analytic mitigations	Analytic mitigations
8. Causative agents may be unrepresented in existing sequence databases.	Novel species of micro-organisms are being identified at a high rate. For example, the Genome Taxonomy Database grew from 23 458 bacterial and 1 248 archaeal species in August 2019, to 45 555 bacterial and 2 339 archaeal species in April 2021.	• Irremediable	• Future reanalysis with contemporary databases × • Less database dependent methods, e.g. *de novo* peptide sequencing, and metagenome assembly based analyses and incorporation of metagenomic predicted protein sequences into proteomic databases. ±
9. Technologies have blind spots and limits of detection may be too high.	For example, the metagenomic analyses will not detect RNA viruses. Data-dependent bottom-up proteomics acquisitions typically direct the mass spectrometer to fragment the most intense MS1 features, leaving many features unfragmented (see Figure 32, p140).	• Incorporate parallel reverse transcriptase PCR into pharyngeal analyses × • Sample processing and data acquisition techniques to increase sensitivity to low abundance microbes or antigen, e.g. optimise dynamic exclusion of already-fragmented features. ×	• Analytic techniques to increase sensitivity to low abundance microbes or antigen, e.g. increase identification of low abundance MS1 features. ±

Table 23 Limitations affecting the study of microbes aetiologically linked to KD and potential pre-analytic and analytic mitigations. × indicates mitigations not applied or not possible in this analysis; ✓ indicates mitigations applied, and ± indicates mitigations partially applied.

The lack of RNA sequencing in the oropharynx is a significant limitation. Analysis of the ongoing BATS recruitment shows skewing to a younger demographic and a higher prevalence of KD-like features in recent variant waves (Mike Levin, personal communication). Evolution of the clinical features of MIS-C over time to become closer to those of KD could add support to the existing evidence for a coronavirus trigger of KD.

The emergence of MIS-C as a SARS-CoV-2-triggered inflammatory disorder, occurring with a delay between infection and disease of about a month, suggests an additional complexity in searching for the causative pathogen. If KD was similarly triggered by a virus or bacteria which infected the child a month earlier, our metagenomic study may have missed identifying the pathogen. Some reassurance is at least provided by the RT-PCR-positivity of 24% in the first analysis of BATS (McArdle et al., 2021).

As indicated, many of these issues cannot be mitigated in the analysis of these data. Of those which are amenable to mitigation, not all have yet been applied. For example, Burns et al. (2021) have shown that temporal clusters of KD cases in San Diego share clinical features more than expected by chance. Many patients in this study are members of these temporal clusters, which could be classified into Rotated Empirical Orthogonal Function (REOF) groups based on clinical features. This opens the possibility of re-analysis focused on differences in organism abundance/prevalence between KD cases from distinct REOF clusters.

Given the respiratory tract as the dominant portal of entry for agent(s) causing KD and the potential for public databases to miss relevant organisms, it could be beneficial to use metagenomic sequence data in the construction of protein sequence databases for proteomics. With the limited overlap of individuals with both metagenomic and metaproteomic data, this could not be applied on a *per sample* basis. Instead, predicted protein sequence data from all metagenomic samples would be pooled, most likely predicted open reading frames with deduplication of redundant sequences, as applied in the strain-level analyses (page 86). This is a common approach in metaproteomics, and significant gains in identifications have been shown as compared to public sequence databases (Tanca et al., 2016).

Alternative approaches

The challenges encountered here, especially in the metaproteomic approach, raise more fundamental questions about alternative approaches to identify causative agents in KD.

Given the potential lag between exposure to/infection with a triggering agent, any means of sampling children *before* they develop KD could increase the potential to detect causative agents. This could be achieved if a large enough high-risk cohort could be obtained, for example children with strong family history in Japan. Respiratory samples could be taken regularly through childhood and archived.

Samples could then be selected for processing from the months prior to KD development, with potential for patients to act as their own controls, as well as availability of independent controls who do not develop KD during the study period. Such a study would come with considerable expense with human and logistical challenges. Notably, in a retrospective study of multiple sclerosis which used longitudinal periodic assessments for blood viruses, this approach recently led to the identification of Epstein-Barr virus (EBV) infection as a significant triggering factor (Bjornevik et al., 2022).

Additional alternative approaches could leverage the immune response for clues to triggers, infectious or otherwise, as commenced in the preceding chapter. Immunopeptidomics involves the elution and sequencing of peptides from circulating T cells, and could allow resolution of sufficient distinct peptides to conclusively identify microbial antigen. This approach has been applied to help identify non-canonical candidate tumour antigens (Chong, Coukos & Bassani-Sternberg, 2022). T-cell receptor sequencing can also be used to predict cognate antigen (Lanzarotti, Marcatili & Nielsen, 2019), though this would likely only be practical with a small search space. Inference of antigen specificity from B-cell receptor sequencing is in its infancy (Krawczyk et al., 2018; Kovaltsuk et al., 2017). The advent of high-accuracy deep-learning based structural modelling of proteins (Jumper et al., 2021) shows promise for aiding progress in this area. The complementarity determining regions of immunoglobulin regions prove difficult for current models, though focused training on antibody structures is reported to yield better accuracy (Peng et al., 2023).

Key learning and retrospectives

Taking into account all of the above, confidently identifying microbial triggers of KD remains highly challenging. Broad, untargeted approaches as I have undertaken with my colleagues account for our inability to exclude organisms *a priori*. However, they suffer from significant analytic and inferential challenges. Further, no methods exist for testing power in studies such as these, unless one uses a simulation-based approach with countless assumptions. More focused approaches have a high risk of missing relevant organisms, though it must be acknowledged that targeted studies of coronaviruses could be fruitful, given recent experience.

If faced with the potential to repeat and redesign this study, I would prioritise the incorporation of methods to remove human DNA, and include RT-PCR to allow detection of RNA viruses. I would also consider the application of broad-range serology screening, using a technique like VirScan (Xu et al., 2015), though no differences were shown among 37 cases and matched controls in a recent study (Quiat et al., 2020).

Future work

As demonstrated, potentially fruitful opportunities for further analysis of the data remain through optimising analytic approaches. This could qualify or strengthen existing findings, and lead to new discoveries within these data.

The metagenomic findings in this study should be subjected to further analysis. Within the research group we plan to validate the quantitative association of candidate organisms with KD by quantitative PCR in existing and additional samples. We also plan to measure the antibody response in KD patients and controls to these candidate organisms. Elevated antibodies to these prevalent organisms could support a role in the aetiopathogenesis.

Regarding metaproteomics, further experiments have been conducted with samples from children with PIMS-TS, searching for SARS-CoV-2 antigen within extracted immune complexes.

The proteomics of bulk antibody remains a little-explored area, and my own work here is novel. Plans are underway to further explore the contribution of the antibody response to KD and PIMS-TS by sequencing the germline immunoglobulin loci (Ford et al., 2020).

Concluding remarks

KD is potentially a new disease of the modern era and its aetiology remains unexplained despite extensive study from first discovery. I utilised data generated from throat swabs and ICs extracted from the blood with distinct sequencing approaches across hundreds of children with and without KD. I sought to identify potential aetiological agents with minimal restriction on organisms that could be identified and considered.

I present evidence that some bacteria in the oropharynx are significantly more abundant in KD patients at presentation compared to children with other febrile illnesses. The significance of these findings remains to be determined.

Bibliography

Abhinandan, K.R. & Martin, A.C.R. (2008) Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. *Molecular Immunology*. 45 (14), 3832–3839. doi:10.1016/j.molimm.2008.05.022.

Adamson, P.J., Al Kindi, M.A., Wang, J.J., Colella, A.D., Chataway, T.K., Petrovsky, N., Gordon, T.P. & Gordon, D.L. (2017) Proteomic analysis of influenza haemagglutinin-specific antibodies following vaccination reveals convergent immunoglobulin variable region signatures. *Vaccine*. 35 (42), 5576–5580. doi:10.1016/j.vaccine.2017.08.053.

Adusumilli, R. & Mallick, P. (2017) Data Conversion with ProteoWizard msConvert. In: L. Comai, J.E. Katz, & P. Mallick (eds.). *Proteomics: Methods and Protocols*. Methods in Molecular Biology. New York, NY, Springer. pp. 339–368. doi:10.1007/978-1-4939-6747-6_23.

Anon (2021) *Serum Protein Electrophoresis: Reference Range, Interpretation, Collection and Panels*. https://emedicine.medscape.com/article/2087113-overview.

Bajolle, F., Meritet, J.-F., Rozenberg, F., Chalumeau, M., Bonnet, D., Gendrel, D. & Lebon, P. (2014) Markers of a recent bocavirus infection in children with Kawasaki disease: "A year prospective study". *Pathologie Biologie*. 62 (6), 365–368. doi:10.1016/j.patbio.2014.06.002.

Barsnes, H. & Vaudel, M. (2018) SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines. *Journal of Proteome Research*. 17 (7), 2552–2555. doi:10.1021/acs.jproteome.8b00175.

Bayram, R.O., Özdemir, H., Emsen, A., Türk Dağı, H. & Artaç, H. (2019) Reference ranges for serum immunoglobulin (IgG, IgA, and IgM) and IgG subclass levels in healthy children. *Turkish Journal of Medical Sciences*. 49 (2), 497–505. doi:10.3906/sag-1807-282.

Beghini, F., McIver, L.J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A.M., Valles-Colomer, M., Weingart, G., Zhang, Y., Zolfo, M., Huttenhower, C., Franzosa, E.A. & Segata, N. (2021) Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3 P. Turnbaugh, E. Franco, & C.T. Brown (eds.). *eLife*. 10, e65088. doi:10.7554/eLife.65088.

Belay, E.D., Erdman, D.D., Anderson, L.J., Peret, T.C.T., Schrag, S.J., Fields, B.S., Burns, J.C. & Schonberger, L.B. (2005) Kawasaki Disease and Human Coronavirus. *The Journal of Infectious Diseases*. 192 (2), 352–353. doi:10.1086/431609.

Bern, M. & Kil, Y.J. (2011) Comment on "Unbiased Statistical Analysis for Multi-Stage Proteomic Search Strategies". *Journal of Proteome Research*. 10 (4), 2123–2127. doi:10.1021/pr101143m.

Bjornevik, K., Cortese, M., Healy, B.C., Kuhle, J., Mina, M.J., Leng, Y., Elledge, S.J., Niebuhr, D.W., Scher, A.I., Munger, K.L. & Ascherio, A. (2022) Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis. *Science*. 375 (6578), 296–301. doi:10.1126/science.abj8222.

Blin, K. (2022) *NCBI Genome Downloading Scripts*. https://github.com/kblin/ncbi-genome-download.

Bogaert, D., Keijser, B., Huse, S., Rossen, J., Veenhoven, R., van Gils, E., Bruin, J., Montijn, R., Bonten, M. & Sanders, E. (2011) Variability and diversity of nasopharyngeal microbiota in children: a metagenomic analysis. *PloS One*. 6 (2), e17035. doi:10.1371/journal.pone.0017035.

Borrel, G., Brugère, J.-F., Gribaldo, S., Schmitz, R.A. & Moissl-Eichinger, C. (2020) The host-associated archaeome. *Nature Reviews Microbiology*. 18 (11), 622–636. doi:10.1038/s41579-020-0407-y.

Bowman, J.P. (2018) Polaribacter. In: *Bergey's Manual of Systematics of Archaea and Bacteria*. John Wiley & Sons, Ltd. pp. 1–21. doi:10.1002/9781118960608.gbm00333.pub2.

Breitwieser, F.P., Baker, D.N. & Salzberg, S.L. (2018) KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biology*. 19 (1), 198. doi:10.1186/s13059-018-1568-0.

Breitwieser, F.P., Pertea, M., Zimin, A. & Salzberg, S.L. (2019) Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Research*. gr.245373.118. doi:10.1101/gr.245373.118.

Burgner, D. & Harnden, A. (2005) Kawasaki disease: What is the epidemiology telling us about the etiology? *International Journal of Infectious Diseases*. 9 (4), 185–194. doi:10.1016/j.ijid.2005.03.002.

Burns, J.C., DeHaan, L.L., Shimizu, C., Bainto, E.V., Tremoulet, A.H., Cayan, D.R. & Burney, J.A. (2021a) Temporal Clusters of Kawasaki Disease Cases Share Distinct Phenotypes That Suggest Response to Diverse Triggers. *The Journal of Pediatrics*. 229, 48-53.e1. doi:10.1016/j.jpeds.2020.09.043.

Burns, J.C., DeHaan, L.L., Shimizu, C., Bainto, E.V., Tremoulet, A.H., Cayan, D.R. & Burney, J.A. (2021b) Temporal Clusters of Kawasaki Disease Cases Share Distinct Phenotypes That Suggest Response to Diverse Triggers. *The Journal of Pediatrics*. 229, 48-53.e1. doi:10.1016/j.jpeds.2020.09.043.

Burns, J.C., Herzog, L., Fabri, O., Tremoulet, A.H., Rodó, X., Uehara, R., Burgner, D., Bainto, E., Pierce, D., Tyree, M., Cayan, D., & Kawasaki Disease Global Climate Consortium (2013) Seasonality of Kawasaki disease: a global perspective. *PloS One*. 8 (9), e74529. doi:10.1371/journal.pone.0074529.

Burns, J.C., Kushner, H.I., Bastian, J.F., Shike, H., Shimizu, C., Matsubara, T. & Turner, C.L. (2000) Kawasaki Disease: A Brief History. *Pediatrics*. 106 (2), e27–e27. doi:10.1542/peds.106.2.e27.

Burns, J.C., Newburger, J.W., Sundell, R., Wyatt, L.S. & Frenkel, N. (1994) Seroprevalence of Human Herpesvirus 7 in Patients with Kawasaki disease. *The Pediatric Infectious Disease Journal*. 13 (2), 168.

Calvopiña, K., Hinchliffe, P., Brem, J., Heesom, K.J., Johnson, S., Cain, R., Lohans, C.T., Fishwick, C.W.G., Schofield, C.J., Spencer, J. & Avison, M.B. (2017) Structural/mechanistic insights into the efficacy of nonclassical β-lactamase inhibitors against extensively drug resistant Stenotrophomonas maltophilia clinical isolates. *Molecular Microbiology*. 106 (3), 492–504. doi:10.1111/mmi.13831.

Carmona, E.G., García-Giménez, J.A., López-Mejías, R., Khor, C.C., Lee, J.-K., et al. (2021) Identification of a shared genetic risk locus for Kawasaki disease and immunoglobulin A vasculitis by a cross-phenotype meta-analysis. *Rheumatology*. (keab443). doi:10.1093/rheumatology/keab443.

Catalano-Pons, C., Quartier, P., Leruez-Ville, M., Kaguelidou, F., Gendrel, D., Lenoir, G., Casanova, J.-L. & Bonnet, D. (2005) Primary Cytomegalovirus Infection, Atypical Kawasaki Disease, and Coronary Aneurysms in 2 Infants. *Clinical Infectious Diseases*. 41 (5), e53–e56. doi:10.1086/432578.

Cazzaniga, M., Baselli, L.A., Cimaz, R., Guez, S.S., Pinzani, R. & Dellepiane, R.M. (2020) SARS-COV-2 Infection and Kawasaki Disease: Case Report of a Hitherto Unrecognized Association. *Frontiers in Pediatrics*. 0. doi:10.3389/fped.2020.00398.

Cebey-López, M., Herberg, J., Pardo-Seco, J., Gómez-Carballa, A., Martinón-Torres, N., Salas, A., Martinón-Sánchez, J.M., Gormley, S., Sumner, E., Fink, C., Martinón-Torres, F. & Network, G. (2015) Viral Co-Infections in Pediatric Patients Hospitalized with Lower Tract Acute Respiratory Infections. *PLOS ONE*. 10 (9), e0136526. doi:10.1371/journal.pone.0136526.

Cebey-López, M., Herberg, J., Pardo-Seco, J., Gómez-Carballa, A., Martinón-Torres, N., Salas, A., Martinón-Sánchez, J.M., Justicia, A., Rivero-Calle, I., Sumner, E., Fink, C., Martinón-Torres, F. & Network, G. (2016) Does Viral Co-Infection Influence the Severity of Acute Respiratory Infection in Children? *PLOS ONE*. 11 (4), e0152481. doi:10.1371/journal.pone.0152481.

Chafee, M., Maignien, L. & Simmons, S.L. (2015) The effects of variable sample biomass on comparative metagenomics. *Environmental Microbiology*. 17 (7), 2239–2253. doi:10.1111/1462-2920.12668.

Chang, A.J., Baron, S., Hoffman, J. & Hicar, M.D. (2022) Clonal expansion and markers of directed mutation of IGHV4-34 B cells in plasmablasts during Kawasaki disease. *Molecular Immunology*. 145, 67–77. doi:10.1016/j.molimm.2022.03.011.

Chang, L.-Y., Lu, C.-Y., Shao, P.-L., Lee, P.-I., Lin, M.-T., Fan, T.-Y., Cheng, A.-L., Lee, W.-L., Hu, J.-J., Yeh, S.-J., Chang, C.-C., Chiang, B.-L., Wu, M.-H. & Huang, L.-M. (2014) Viral infections associated with Kawasaki disease. *Journal of the Formosan Medical Association*. 113 (3), 148–154. doi:10.1016/j.jfma.2013.12.008.

Chen, J., Yue, Y., Wang, L., Deng, Z., Yuan, Y., Zhao, M., Yuan, Z., Tan, C. & Cao, Y. (2020) Altered gut microbiota correlated with systemic inflammation in children with Kawasaki disease. *Scientific Reports*. 10 (1), 14525. doi:10.1038/s41598-020-71371-6.

Chen, J., Zheng, Q., Hammers, C.M., Ellebrecht, C.T., Mukherjee, E.M., Tang, H.-Y., Lin, C., Yuan, H., Pan, M., Langenhan, J., Komorowski, L., Siegel, D.L., Payne, A.S. & Stanley, J.R. (2017) Proteomic Analysis of Pemphigus Autoantibodies Indicates a Larger, More Diverse, and More Dynamic Repertoire than Determined by B Cell Genetics. *Cell Reports*. 18 (1), 237–247. doi:10.1016/j.celrep.2016.12.013.

Chen, S., He, C., Li, Y., Li, Z. & Melançon, C.E., III (2021) A computational toolset for rapid identification of SARS-CoV-2, other viruses and microorganisms from sequencing data. *Briefings in Bioinformatics*. 22 (2), 924–935. doi:10.1093/bib/bbaa231.

Cheng, K., Ning, Z., Zhang, X., Li, L., Liao, B., Mayne, J., Stintzi, A. & Figeys, D. (2017) MetaLab: an automated pipeline for metaproteomic data analysis. *Microbiome*. 5 (1), 157. doi:10.1186/s40168-017-0375-2.

Cheung, W.C., Beausoleil, S.A., Zhang, X., Sato, S., Schieferl, S.M., Wieler, J.S., Beaudet, J.G., Ramenani, R.K., Popova, L., Comb, M.J., Rush, J. & Polakiewicz, R.D. (2012) A proteomics approach for the identification and cloning of monoclonal antibodies from serum. *Nature Biotechnology*. 30 (5), 447–452. doi:10.1038/nbt.2167.

Chi, H., Liu, C., Yang, H., Zeng, W.-F., Wu, L., et al. (2018) *Open-pFind enables precise, comprehensive and rapid peptide identification in shotgun proteomics*. doi:10.1101/285395.

Choe, S.-A., An, H.S. & Choe, Y.J. (2021) No temporal association between human coronavirus and Kawasaki disease: National data from South Korea. *Journal of Medical Virology*. 93 (2), 585–587. doi:10.1002/jmv.26435.

Chong, C., Coukos, G. & Bassani-Sternberg, M. (2022) Identification of tumor antigens with immunopeptidomics. *Nature Biotechnology*. 40 (2), 175–188. doi:10.1038/s41587-021-01038-8.

Chou, J., Platt, C.D., Habiballah, S., Nguyen, A.A., Elkins, M., et al. (2021) Mechanisms underlying genetic susceptibility to multisystem inflammatory syndrome in children (MIS-C). *Journal of Allergy and Clinical Immunology*. 148 (3), 732-738.e1. doi:10.1016/j.jaci.2021.06.024.

Davis, N.M., Proctor, D.M., Holmes, S.P., Relman, D.A. & Callahan, B.J. (2018) Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*. 6 (1), 226. doi:10.1186/s40168-018-0605-2.

Dekhil, S.M.B., Peel, M.M., Lennox, V.A., Stackebrandt, E. & Sly, L.I. (1997) Isolation of Lautropia mirabilis from sputa of a cystic fibrosis patient. *Journal of Clinical Microbiology*. doi:10.1128/jcm.35.4.1024-1026.1997.

Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E. & Notredame, C. (2017) Nextflow enables reproducible computational workflows. *Nature Biotechnology*. 35 (4), 316–319. doi:10.1038/nbt.3820.

Dixon, P. (2003) VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science*. 14 (6), 927–930. doi:10.1111/j.1654-1103.2003.tb02228.x.

Dunbar, J. & Deane, C.M. (2016) ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics*. 32 (2), 298–300. doi:10.1093/bioinformatics/btv552.

Ebihara, T., Endo, R., Ma, X., Ishiguro, N. & Kikuta, H. (2005) Lack of Association between New Haven Coronavirus and Kawasaki Disease. *The Journal of Infectious Diseases*. 192 (2), 351–352. doi:10.1086/430797.

Edlinger, E.A., Benichou, J.J. & Labrune, B. (1980) POSITIVE EHRLICHIA CANIS SEROLOGY IN KAWASAKI DISEASE. *The Lancet*. 315 (8178), 1146–1147. doi:10.1016/S0140-6736(80)91603-7.

Eisenhofer, R., Minich, J.J., Marotz, C., Cooper, A., Knight, R. & Weyrich, L.S. (2019) Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends in Microbiology*. 27 (2), 105–117. doi:10.1016/j.tim.2018.11.003.

El-Madhun, A.S., Cox, R.J. & Haaheim, L.R. (1999) The effect of age and natural priming on the IgG and IgA subclass responses after parenteral influenza vaccination. *The Journal of Infectious Diseases*. 180 (4), 1356–1360. doi:10.1086/315003.

Eng, J.K., Jahan, T.A. & Hoopmann, M.R. (2013) Comet: An open-source MS/MS sequence database search tool. *PROTEOMICS*. 13 (1), 22–24. doi:10.1002/pmic.201200439.

Esper, F., Weibel, C., Ferguson, D., Landry, M.L. & Kahn, J.S. (2005) Evidence of a Novel Human Coronavirus That Is Associated with Respiratory Tract Disease in Infants and Young Children. *The Journal of Infectious Diseases*. 191 (4), 492–498. doi:10.1086/428138.

Everett, L.J., Bierl, C. & Master, S.R. (2010) Unbiased Statistical Analysis for Multi-Stage Proteomic Search Strategies. *Journal of Proteome Research*. 9 (2), 700–707. doi:10.1021/pr900256v.

Felsenstein, S., Duong, P., Lane, S., Jones, C., Pain, C.E. & Hedrich, C.M. (2021) Cardiac pathology and outcomes vary between Kawasaki disease and PIMS-TS. *Clinical Immunology*. 229, 108780. doi:10.1016/j.clim.2021.108780.

Fenyö, D. & Beavis, R.C. (2003) A Method for Assessing the Statistical Significance of Mass Spectrometry-Based Protein Identifications Using General Scoring Schemes. *Analytical Chemistry*. 75 (4), 768–774. doi:10.1021/ac0258709.

Ford, M., Haghshenas, E., Watson, C.T. & Sahinalp, S.C. (2020) Genotyping and Copy Number Analysis of Immunoglobin Heavy Chain Variable Genes Using Long Reads. *iScience*. 23 (3). doi:10.1016/j.isci.2020.100883.

Fujita, Y., Nakamura, Y., Sakata, K., Hara, N., Kobayashi, M., Nagai, M., Yanagawa, H. & Kawasaki, T. (1989) Kawasaki Disease in Families. *Pediatrics*. 84 (4), 666–669.

Fuse, S., Fujinaga, E., Mori, T., Hotsubo, T., Kuroiwa, Y. & Morii, M. (2010) Children With Kawasaki Disease Are Not Infected With Epstein-Barr Virus. *The Pediatric Infectious Disease Journal*. 29 (3), 286–287. doi:10.1097/INF.0b013e3181c3f111.

Galson, J.D., Trück, J., Fowler, A., Clutterbuck, E.A., Münz, M., Cerundolo, V., Reinhard, C., van der Most, R., Pollard, A.J., Lunter, G. & Kelly, D.F. (2015) Analysis of B Cell Repertoire Dynamics Following Hepatitis B Vaccination in Humans, and Enrichment of Vaccine-specific Antibody Sequences. *EBioMedicine*. 2 (12), 2070–2079. doi:10.1016/j.ebiom.2015.11.034.

Gerner-Smidt, P., Keiser-Nielsen, H., Dorsch, M., Stackebrandt, E., Ursing, J., Blom, J., Christensen, A.C., Christensen, J.J., Frederiksen, W., Hoffmann, S., Holten-Andersen, W. & Ying, Y.T.Y. 1994 (1997) Lautropia mirabilis gen. nov., sp. nov., a Gram-negative motile coccus with unusual morphology isolated from the human mouth. *Microbiology*. 140 (7), 1787–1797. doi:10.1099/13500872-140-7-1787.

Gessulat, S., Schmidt, T., Zolg, D.P., Samaras, P., Schnatbaum, K., Zerweck, J., Knaute, T., Rechenberger, J., Delanghe, B., Huhmer, A., Reimer, U., Ehrlich, H.-C., Aiche, S., Kuster, B. & Wilhelm, M. (2019) Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods*. 16 (6), 509. doi:10.1038/s41592-019-0426-7.

Ghannoum, M.A., Jurevic, R.J., Mukherjee, P.K., Cui, F., Sikaroodi, M., Naqvi, A. & Gillevet, P.M. (2010) Characterization of the Oral Fungal Microbiome (Mycobiome) in Healthy Individuals. *PLOS Pathogens*. 6 (1), e1000713. doi:10.1371/journal.ppat.1000713.

Ghraichy, M., Galson, J.D., Kovaltsuk, A., von Niederhäusern, V., Pachlopnik Schmid, J., Recher, M., Jauch, A.J., Miho, E., Kelly, D.F., Deane, C.M. & Trück, J. (2020) Maturation of the Human Immunoglobulin Heavy Chain Repertoire With Age. *Frontiers in Immunology*. 11. https://www.frontiersin.org/article/10.3389/fimmu.2020.01734.

Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V. & Egozcue, J.J. (2017) Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*. 8, 2224. doi:10.3389/fmicb.2017.02224.

Groemping, U. & Matthias, L. (2021) *relaimpo: Relative Importance of Regressors in Linear Models*. https://CRAN.R-project.org/package=relaimpo.

Groenen, P. & van de Velden, M. (2004) Multidimensional scaling. *Report / Econometric Institute, Erasmus University Rotterdam*. https://repub.eur.nl/pub/1274/.

Guan, S., Moran, M. & Ma, B. (2019) Prediction of LC-MS/MS properties of peptides from sequence by deep learning. *Molecular & Cellular Proteomics*. doi:10.1074/mcp.TIR119.001412.

Guleria, S., Jindal, A.K., Pandiarajan, V., Singh, M.P. & Singh, S. (2018) Dengue-Triggered Kawasaki Disease: A Report of 2 Cases. *JCR: Journal of Clinical Rheumatology*. 24 (7), 401–404. doi:10.1097/RHU.0000000000000704.

Guthals, A., Gan, Y., Murray, L., Chen, Y., Stinson, J., Nakamura, G., Lill, J.R., Sandoval, W. & Bandeira, N. (2017) De Novo MS/MS Sequencing of Native Human Antibodies. *Journal of Proteome Research*. 16 (1), 45–54. doi:10.1021/acs.jproteome.6b00608.

Hamada, H., Sekizuka, T., Oba, K., Katano, H., Kinumaki, A., Terai, M., Mizutani, T. & Kuroda, M. (2016) Comprehensive pathogen detection associated with four recurrent episodes of Kawasaki disease in a patient during a single year using next-generation sequencing. *JMM Case Reports*. 3 (1). doi:10.1099/jmmcr.0.005019.

Han, X., He, L., Xin, L., Shan, B. & Ma, B. (2011) PeaksPTM: Mass Spectrometry-Based Identification of Peptides with Unspecified Modifications. *Journal of Proteome Research*. 10 (7), 2930–2936. doi:10.1021/pr200153k.

Hearn, J., McCrindle, B.W., Mueller, B., O'Shea, S., Bernknopf, B., Labelle, M. & Manlhiot, C. (2018) Spatiotemporal clustering of cases of Kawasaki disease and associated coronary artery aneurysms in Canada. *Scientific Reports*. 8 (1), 17682. doi:10.1038/s41598-018-35848-9.

Herberg, J.A., Kaforou, M., Gormley, S., Sumner, E.R., Patel, S., Jones, K.D.J., Paulus, S., Fink, C., Martinon-Torres, F., Montana, G., Wright, V.J. & Levin, M. (2013) Transcriptomic Profiling in Childhood H1N1/09 Influenza Reveals Reduced Expression of Protein Synthesis Genes. *The Journal of Infectious Diseases*. 208 (10), 1664–1668. doi:10.1093/infdis/jit348.

Heyer, R., Schallert, K., Zoun, R., Becher, B., Saake, G. & Benndorf, D. (2017) Challenges and perspectives of metaproteomic data analysis. *Journal of Biotechnology*. 261, 24–36. doi:10.1016/j.jbiotec.2017.06.1201.

Hinzke, T., Kouris, A., Hughes, R.-A., Strous, M. & Kleiner, M. (2019) More Is Not Always Better: Evaluation of 1D and 2D-LC-MS/MS Methods for Metaproteomics. *Frontiers in Microbiology*. 10. doi:10.3389/fmicb.2019.00238.

Hirokawa, M., Fujishima, N., Togashi, M., Saga, A., Omokawa, A., Saga, T., Moritoki, Y., Ueki, S., Takahashi, N., Kitaura, K. & Suzuki, R. (2019) High-throughput sequencing of IgG B-cell receptors reveals frequent usage of the rearranged IGHV4–28/IGHJ4 gene in primary immune thrombocytopenia. *Scientific Reports*. 9 (1), 8645. doi:10.1038/s41598-019-45264-2.

Hoang, L.T., Jain, P., Pillay, T.D., Tolosa-Wright, M., Niazi, U., Takwoingi, Y., Halliday, A., Berrocal-Almanza, L.C., Deeks, J.J., Beverley, P., Kon, O.M. & Lalvani, A. (2021) Transcriptomic signatures for diagnosing tuberculosis in clinical practice: a prospective, multicentre cohort study. *The Lancet Infectious Diseases*. 21 (3), 366–375. doi:10.1016/S1473-3099(20)30928-2.

Hoffman, G.S. & Calabrese, L.H. (2014) Vasculitis: determinants of disease patterns. *Nature Reviews Rheumatology*. 10 (8), 454–462. doi:10.1038/nrrheum.2014.89.

Hoggart, C., Shimizu, C., Galassini, R., Wright, V.J., Shailes, H., et al. (2021) Identification of novel locus associated with coronary artery aneurysms and validation of loci for susceptibility to Kawasaki disease. *European Journal of Human Genetics*. 1–11. doi:10.1038/s41431-021-00838-5.

Horinouchi, T., Nozu, K., Hamahira, K., Inaguma, Y., Abe, J., Nakajima, H., Kugo, M. & Iijima, K. (2015) Yersinia pseudotuberculosis infection in Kawasaki disease and its clinical characteristics. *BMC Pediatrics*. 15 (1), 177. doi:10.1186/s12887-015-0497-2.

Hu, B., Li, Y., Wang, G. & Zhang, Y. (2020) The Blood Gene Expression Signature for Kawasaki Disease in Children Identified with Advanced Feature Selection Methods. *BioMed Research International*. 2020, e6062436. doi:10.1155/2020/6062436.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J.H., et al. (2012) Structure, function and diversity of the healthy human microbiome. *Nature*. 486 (7402), 207–214. doi:10.1038/nature11234.

Iranzadeh, A. & Mulder, N.J. (2019) Bacterial Pan-Genomics. In: V. Tripathi, P. Kumar, P. Tripathi, & A. Kishore (eds.). *Microbial Genomics in Sustainable Agroecosystems: Volume 1*. Singapore, Springer. pp. 21–38. doi:10.1007/978-981-13-8739-5_2.

Jackson, H., Menikou, S., Hamilton, S., McArdle, A., Shimizu, C., et al. (2021) Kawasaki Disease Patient Stratification and Pathway Analysis Based on Host Transcriptomic and Proteomic Profiles. *International Journal of Molecular Sciences*. 22 (11), 5655. doi:10.3390/ijms22115655.

Jackson, K.J.L., Liu, Y., Roskin, K.M., Glanville, J., Hoh, R.A., et al. (2014) Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. *Cell host & microbe*. 16 (1), 105–114. doi:10.1016/j.chom.2014.05.013.

Jaggi, P., Kajon, A.E., Mejias, A., Ramilo, O. & Leber, A. (2013) Human Adenovirus Infection in Kawasaki Disease: A Confounding Bystander? *Clinical Infectious Diseases*. 56 (1), 58–64. doi:10.1093/cid/cis807.

Jaggi, P., Mejias, A., Xu, Z., Yin, H., Moore-Clingenpeel, M., Smith, B., Burns, J.C., Tremoulet, A.H., Jordan-Villegas, A., Chaussabel, D., Texter, K., Pascual, V. & Ramilo, O. (2018) Whole blood transcriptional profiles as a prognostic tool in complete and incomplete Kawasaki Disease. *PLOS ONE*. 13 (5), e0197858. doi:10.1371/journal.pone.0197858.

Jagtap, P., Goslinga, J., Kooren, J.A., McGowan, T., Wroblewski, M.S., Seymour, S.L. & Griffin, T.J. (2013) A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *PROTEOMICS*. 13 (8), 1352–1357. doi:10.1002/pmic.201200352.

Jain, A., Misra, D.P., Sharma, A., Wakhlu, A., Agarwal, V. & Negi, V.S. (2018) Vasculitis and vasculitis-like manifestations in monogenic autoinflammatory syndromes. *Rheumatology International*. 38 (1), 13–24. doi:10.1007/s00296-017-3839-6.

Jeong, K., Kim, S. & Bandeira, N. (2012) False discovery rates in spectral identification. *BMC Bioinformatics*. 13 (16), S2. doi:10.1186/1471-2105-13-S16-S2.

Joshi, A.V., Jones, K.D., Buckley, A.-M., Coren, M.E. & Kampmann, B. (2011) Kawasaki disease coincident with influenza A H1N1/09 infection. *Pediatrics International*. 53 (1), e1–e2. doi:10.1111/j.1442-200X.2010.03280.x.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., et al. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*. 596 (7873), 583–589. doi:10.1038/s41586-021-03819-2.

Kalxdorf, M., Müller, T., Stegle, O. & Krijgsveld, J. (2021) IceR improves proteome coverage and data completeness in global and single-cell proteomics. *Nature Communications*. 12 (1), 4787. doi:10.1038/s41467-021-25077-6.

Kan, J., Hanson, T.E., Ginter, J.M., Wang, K. & Chen, F. (2005) Metaproteomic analysis of Chesapeake Bay microbial communities. *Saline Systems*. 1 (1), 7. doi:10.1186/1746-1448-1-7.

Kanda, S., Fujii, Y., Hori, S.-I., Ohmachi, T., Yoshimura, K., Higasa, K. & Kaneko, K. (2021) Combined Single Nucleotide Variants of ORAI1 and BLK in a Child with Refractory Kawasaki Disease. *Children (Basel, Switzerland)*. 8 (6), 433. doi:10.3390/children8060433.

Kanegaye, J.T., Wilder, M.S., Molkara, D., Frazer, J.R., Pancheri, J., Tremoulet, A.H., Watson, V.E., Best, B.M. & Burns, J.C. (2009) Recognition of a Kawasaki Disease Shock Syndrome. *Pediatrics*. 123 (5), e783–e789. doi:10.1542/peds.2008-1871.

Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H. & Wang, Z. (2019) MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*. 7, e7359. doi:10.7717/peerj.7359.

Karmali, MohamedA., Petric, M., Steele, BrianT. & Lim, C. (1983) Sporadic cases of Haemolytic-Uraemic Syndrome associated with faecal cytotoxin and cytotoxin-producing Escherichia coli in stools. *The Lancet*. 321 (8325), 619–620. doi:10.1016/S0140-6736(83)91795-6.

Karstens, L., Asquith, M., Davin, S., Fair, D., Gregory, W.T., Wolfe, A.J., Braun, J. & McWeeney, S. (2019) Controlling for Contaminants in Low-Biomass 16S rRNA Gene Sequencing Experiments. *mSystems*. 4 (4). doi:10.1128/mSystems.00290-19.

Kato, H., Inoue, O., Koga, Y., Shingu, M., Fujimoto, T., Kondo, M., Yamamoto, S., Tominaga, K. & Sasaguri, Y. (1983) Variant strain of Propionibacterium acnes: a clue to the aetiology of Kawasaki disease. *The Lancet*. 322 (8364), 1383–1388. doi:10.1016/S0140-6736(83)90921-2.

Kawasaki, T. (1967) Acute febrile muco-cutaneous lymph node syndrome in young children with unique digital desquamation. *Jpn J Allergol*. 16, 178–222.

Khan, I., Li, X., Law, B., U, K.I., Pan, B.Q., Lei, C. & Hsiao, W.W. (2020) Correlation of gut microbial compositions to the development of Kawasaki disease vasculitis in children. *Future Microbiology*. 15 (8), 591–600. doi:10.2217/fmb-2019-0301.

Kiernan, T.J., O'Flaherty, N., Gilmore, R., Ho, E., Hickey, M., Tolan, M., Mulcahy, D. & Moore, D.P. (2008) Abiotrophia defectiva endocarditis and associated hemophagocytic syndrome—a first case report and review of the literature. *International Journal of Infectious Diseases*. 12 (5), 478–482. doi:10.1016/j.ijid.2008.01.014.

Kim, J., Shimizu, C., Kingsmore, S.F., Veeraraghavan, N., Levy, E., Ribeiro Dos Santos, A.M., Yang, H., Flatley, J., Hoang, L.T., Hibberd, M.L., Tremoulet, A.H., Harismendy, O., Ohno-Machado, L. & Burns, J.C. (2017) Whole genome sequencing of an African American family highlights toll like receptor 6 variants in Kawasaki disease susceptibility. *PloS One*. 12 (2), e0170977. doi:10.1371/journal.pone.0170977.

Kim, J.H., Yu, J.J., Lee, J., Kim, M.-N., Ko, H.K., Choi, H.S., Kim, Y.-H. & Ko, J.-K. (2012) Detection rate and clinical impact of respiratory viruses in children with Kawasaki disease. *Korean Journal of Pediatrics*. 55 (12), 470–473. doi:10.3345/kjp.2012.55.12.470.

Kim, J.-J., Hong, Y.M., Yun, S.W., Lee, K.-Y., Yoon, K.L., Han, M.-K., Kim, G.B., Kil, H.-R., Song, M.S., Lee, H.D., Ha, K.S., Jun, H.O., Choi, B.-O., Oh, Y.-M., Yu, J.J., Jang, G.Y., Lee, J.-K., & Korean Kawasaki Disease Genetics Consortium (2021) Identification of rare coding variants associated with Kawasaki disease by whole exome sequencing. *Genomics & Informatics*. 19 (4), e38. doi:10.5808/gi.21046.

Kinumaki, A., Sekizuka, T., Hamada, H., Kato, K., Yamashita, A. & Kuroda, M. (2015) Characterization of the gut microbiota of Kawasaki disease patients by metagenomic analysis. *Frontiers in Microbiology*. 6. doi:10.3389/fmicb.2015.00824.

Klaassens, E.S., Vos, W.M. de & Vaughan, E.E. (2007) Metaproteomics Approach To Study the Functionality of the Microbiota in the Human Infant Gastrointestinal Tract. *Applied and Environmental Microbiology*. doi:10.1128/AEM.01921-06.

Knights, D., Kuczynski, J., Charlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., Bushman, F.D., Knight, R. & Kelley, S.T. (2011) Bayesian community-wide culture-independent microbial source tracking. *Nature Methods*. 8 (9), 761–763. doi:10.1038/nmeth.1650.

Ko, T.-M., Kiyotani, K., Chang, J.-S., Park, J.-H., Yin Yew, P., Chen, Y.-T., Wu, J.-Y. & Nakamura, Y. (2018) Immunoglobulin profiling identifies unique signatures in patients with Kawasaki disease during intravenous immunoglobulin treatment. *Human Molecular Genetics*. 27 (15), 2671–2677. doi:10.1093/hmg/ddy176.

Koike, R. (1991) The effect of immunoglobulin on immune complexes in patients with Kawasaki disease (MCLS). *Acta Paediatrica Japonica: Overseas Edition*. 33 (3), 300–309.

Koné-Paut, I., Tellier, S., Belot, A., Brochard, K., Guitton, C., Marie, I., Meinzer, U., Cherqaoui, B., Galeotti, C., Boukhedouni, N., Agostini, H., Arditi, M., Lambert, V. & Piedvache, C. (2021) Phase II Open Label Study of Anakinra in Intravenous Immunoglobulin–Resistant Kawasaki Disease. *Arthritis & Rheumatology*. 73 (1), 151–161. doi:10.1002/art.41481.

Koskinen, K., Pausan, M.R., Perras, A.K., Beck, M., Bang, C., Mora, M., Schilhabel, A., Schmitz, R. & Moissl-Eichinger, C. (2017) First Insights into the Diverse Human Archaeome: Specific Detection of Archaea in the Gastrointestinal Tract, Lung, and Nose and on Skin. *mBio*. 8 (6), e00824-17. doi:10.1128/mBio.00824-17.

Kosonen, T., Huhtinen, S. & Hansen, K. (2021) Taxonomy and systematics of Hyaloscyphaceae and Arachnopezizaceae. *Persoonia - Molecular Phylogeny and Evolution of Fungi*. 46 (1), 26–62. doi:10.3767/persoonia.2021.46.02.

Kovaltsuk, A., Krawczyk, K., Galson, J.D., Kelly, D.F., Deane, C.M. & Trück, J. (2017) How B-Cell Receptor Repertoire Sequencing Can Be Enriched with Structural Antibody Data. *Frontiers in Immunology*. 8. doi:10.3389/fimmu.2017.01753.

Kovaltsuk, A., Leem, J., Kelm, S., Snowden, J., Deane, C.M. & Krawczyk, K. (2018) Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires. *The Journal of Immunology*. 201 (8), 2502–2509. doi:10.4049/jimmunol.1800708.

Krawczyk, K., Kelm, S., Kovaltsuk, A., Galson, J.D., Kelly, D., Trück, J., Regep, C., Leem, J., Wong, W.K., Nowak, J., Snowden, J., Wright, M., Starkie, L., Scott-Tucker, A., Shi, J. & Deane, C.M. (2018) Structurally Mapping Antibody Repertoires. *Frontiers in Immunology*. 9. doi:10.3389/fimmu.2018.01698.

Kryukov, K. & Imanishi, T. (2016) Human Contamination in Public Genome Assemblies. *PLOS ONE*. 11 (9), e0162424. doi:10.1371/journal.pone.0162424.

Kuijpers, T.W., Herweijer, T.J., Schölvinck, L., Wertheim-Van Dillen, P.M. & Van De Veer, E.M.A. (2000) Kawasaki disease associated with measles virus infection in a monozygotic twin. *The Pediatric Infectious Disease Journal*. 19 (4), 350–353.

Kulhankova, K., King, J. & Salgado-Pabón, W. (2014) Staphylococcal toxic shock syndrome: superantigen-mediated enhancement of endotoxin shock and adaptive immune suppression. *Immunologic Research*. 59 (1), 182–187. doi:10.1007/s12026-014-8538-8.

Kushner, H.I., Bastian, J.F., Turner, C.L. & Burns, J.C. (2008) The Two Emergences of Kawasaki Syndrome and the Implications for the Developing World. *The Pediatric Infectious Disease Journal*. 27 (5), 377–383. doi:10.1097/INF.0b013e318166d795.

Kwong, P.D., Chuang, G.-Y., DeKosky, B.J., Gindin, T., Georgiev, I.S., Lemmin, T., Schramm, C.A., Sheng, Z., Soto, C., Yang, A.-S., Mascola, J.R. & Shapiro, L. (2017) Antibodyomics: bioinformatics technologies for understanding B-cell immunity to HIV-1. *Immunological Reviews*. 275 (1), 108–128. doi:10.1111/imr.12480.

Lanzarotti, E., Marcatili, P. & Nielsen, M. (2019) T-Cell Receptor Cognate Target Prediction Based on Paired α and β Chain Sequence and Structural CDR Loop Similarities. *Frontiers in Immunology*. 10. https://www.frontiersin.org/articles/10.3389/fimmu.2019.02080.

Lavinder, J.J., Wine, Y., Giesecke, C., Ippolito, G.C., Horton, A.P., Lungu, O.I., Hoi, K.H., DeKosky, B.J., Murrin, E.M., Wirth, M.M., Ellington, A.D., Dörner, T., Marcotte, E.M., Boutz, D.R. & Georgiou, G. (2014) Identification and characterization of the constituent human serum antibodies elicited by vaccination. *Proceedings of the National Academy of Sciences*. 111 (6), 2259–2264. doi:10.1073/pnas.1317793111.

Lee, D.-H. & Huang, H.-P. (2004) Kawasaki disease associated with chickenpox: report of two sibling cases. *Acta Paediatrica Taiwanica = Taiwan Er Ke Yi Xue Hui Za Zhi*. 45 (2), 94–96.

Lee, J., Boutz, D.R., Chromikova, V., Joyce, M.G., Vollmers, C., et al. (2016) Molecular-level analysis of the serum antibody repertoire in young adults before and after seasonal influenza vaccination. *Nature Medicine*. 22 (12), 1456–1464. doi:10.1038/nm.4224.

Lee, M.N., Cha, J.H., Ahn, H.M., Yoo, J.H., Kim, H.S., Sohn, S. & Hong, Y.M. (2011) Mycoplasma pneumoniae infection in patients with Kawasaki disease. *Korean Journal of Pediatrics*. 54 (3), 123–127. doi:10.3345/kjp.2011.54.3.123.

Lehmann, C., Klar, R., Lindner, J., Lindner, P., Wolf, H. & Gerling, S. (2009) Kawasaki Disease Lacks Association With Human Coronavirus NL63 and Human Bocavirus. *The Pediatric Infectious Disease Journal*. 28 (6), 553. doi:10.1097/INF.0b013e31819f41b6.

Leprevost, F.V., Valente, R.H., Lima, D.B., Perales, J., Melani, R., Yates, J.R., Barbosa, V.C., Junqueira, M. & Carvalho, P.C. (2014) PepExplorer: A Similarity-driven Tool for Analyzing de Novo Sequencing Results *. *Molecular & Cellular Proteomics*. 13 (9), 2480–2489. doi:10.1074/mcp.M113.037002.

Levin, M., Holland, P.C., Nokes, T.J., Novelli, V., Mola, M., Levinsky, R.J., Dillon, M.J., Barratt, T.M. & Marshall, W.C. (1985) Platelet immune complex interaction in pathogenesis of Kawasaki disease and childhood polyarteritis. *British Medical Journal (Clinical research ed.)*. 290 (6480), 1456–1460.

Li, C.R. (1989) [IgG subclasses in the serum and circulating immune complexes in patients with Kawasaki disease. *Zhonghua yi xue za zhi*. 69 (10), 578–581, 40.

Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*. 31 (10), 1674–1676. doi:10.1093/bioinformatics/btv033.

Lim, M.Y., Paulo, J.A. & Gygi, S.P. (2019) Evaluating False Transfer Rates from the Match-between-Runs Algorithm with a Two-Proteome Model. *Journal of Proteome Research*. 18 (11), 4020–4026. doi:10.1021/acs.jproteome.9b00492.

Lin, C.-Y., Chen, I.-C., Cheng, T.-I., Liu, W.-T., Hwang, B. & Chiang, B.N. (1992) Virus-like particles with reverse transcriptase activity associated with kawasaki disease. *Journal of Medical Virology*. 38 (3), 175–182. doi:10.1002/jmv.1890380305.

Liu, H., Manuilov, A.V., Chumsae, C., Babineau, M.L. & Tarcsa, E. (2011) Quantitation of a recombinant monoclonal antibody in monkey serum by liquid chromatography–mass spectrometry. *Analytical Biochemistry*. 414 (1), 147–153. doi:10.1016/j.ab.2011.03.004.

Lozupone, C. & Knight, R. (2005) UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Applied and Environmental Microbiology*. doi:10.1128/AEM.71.12.8228-8235.2005.

Lu, J., Breitwieser, F.P., Thielen, P. & Salzberg, S.L. (2017) Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science*. 3, e104. doi:10.7717/peerj-cs.104.

Luo, C., Knight, R., Siljander, H., Knip, M., Xavier, R.J. & Gevers, D. (2015) ConStrains identifies microbial strains in metagenomic datasets. *Nature Biotechnology*. 33 (10), 1045–1052. doi:10.1038/nbt.3319.

Maddox, R.A., Holman, R.C., Uehara, R., Callinan, L.S., Guest, J.L., Schonberger, L.B., Nakamura, Y., Yashiro, M. & Belay, E.D. (2015) Recurrent Kawasaki disease: USA and Japan. *Pediatrics International*. 57 (6), 1116–1120. doi:10.1111/ped.12733.

Maia, T.M., Staes, A., Plasman, K., Pauwels, J., Boucher, K., Argentini, A., Martens, L., Montoye, T., Gevaert, K. & Impens, F. (2020) Simple Peptide Quantification Approach for MS-Based Proteomics Quality Control. *ACS Omega*. 5 (12), 6754–6762. doi:10.1021/acsomega.0c00080.

Maistrenko, O.M., Mende, D.R., Luetge, M., Hildebrand, F., Schmidt, T.S.B., Li, S.S., Rodrigues, J.F.M., von Mering, C., Pedro Coelho, L., Huerta-Cepas, J., Sunagawa, S. & Bork, P. (2020) Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. *The ISME Journal*. 14 (5), 1247–1259. doi:10.1038/s41396-020-0600-z.

Makino, N., Nakamura, Y., Yashiro, M., Sano, T., Ae, R., Kosami, K., Kojo, T., Aoyama, Y., Kotani, K. & Yanagawa, H. (2018) Epidemiological observations of Kawasaki disease in Japan, 2013–2014. *Pediatrics International*. 60 (6), 581–587. doi:10.1111/ped.13544.

Mallick, H., Rahnavard, A., McIver, L.J., Ma, S., Zhang, Y., et al. (2021) *Multivariable Association Discovery in Population-scale Meta-omics Studies*.p.2021.01.20.427420. doi:10.1101/2021.01.20.427420.

Man, W.H., Clerc, M., de Steenhuijsen Piters, W.A.A., van Houten, M.A., Chu, M.L.J.N., Kool, J., Keijser, B.J.F., Sanders, E.A.M. & Bogaert, D. (2019) Loss of Microbial Topography between Oral and Nasopharyngeal Microbiota and Development of Respiratory Infections Early in Life. *American Journal of Respiratory and Critical Care Medicine*. 200 (6), 760–770. doi:10.1164/rccm.201810-1993OC.

Manenti, A., Tete, S.M., Mohn, K.G.-I., Jul-Larsen, Å., Gianchecchi, E., Montomoli, E., Brokstad, K.A. & Cox, R.J. (2017) Comparative analysis of influenza A(H3N2) virus hemagglutinin specific IgG subclass and IgA responses in children and adults after influenza vaccination. *Vaccine*. 35 (1), 191–198. doi:10.1016/j.vaccine.2016.10.024.

Manlhiot, C., Mueller, B., O'Shea, S., Majeed, H., Bernknopf, B., Labelle, M., Westcott, K.V., Bai, H., Chahal, N., Birken, C.S., Yeung, R.S.M. & McCrindle, B.W. (2018) Environmental epidemiology of Kawasaki disease: Linking disease etiology, pathogenesis and global distribution. *PLOS ONE*. 13 (2), e0191087. doi:10.1371/journal.pone.0191087.

Mann, M. & Kelleher, N.L. (2008) Precision proteomics: The case for high resolution and high mass accuracy. *Proceedings of the National Academy of Sciences*. 105 (47), 18132–18138. doi:10.1073/pnas.0800788105.

Marçais, G. & Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. 27 (6), 764–770. doi:10.1093/bioinformatics/btr011.

Matsubara, K. & Fukaya, T. (2007) The role of superantigens of group A Streptococcus and Staphylococcus aureus in Kawasaki disease. *Current Opinion in Infectious Diseases*. 20 (3), 298–303. doi:10.1097/QCO.0b013e3280964d8c.

McArdle, A.J. & Kaforou, M. (2020) Sensitivity of shotgun metagenomics to host DNA: abundance estimates depend on bioinformatic tools and contamination is the main issue. *Access Microbiology,*. 2 (4), e000104. doi:10.1099/acmi.0.000104.

McArdle, A.J., Vito, O., Patel, H., Seaby, E.G., Shah, P., et al. (2021) Treatment of Multisystem Inflammatory Syndrome in Children. *New England Journal of Medicine*. 385 (1), 11–22. doi:10.1056/NEJMoa2102968.

McCrindle, B.W., Rowley, A.H., Newburger, J.W., Burns, J.C., Bolger, A.F., Gewitz, M., Baker, A.L., Jackson, M.A., Takahashi, M., Shah, P.B., Kobayashi, T., Wu, M.-H., Saji, T.T., Pahl, E., & On behalf of the American Heart Association Rheumatic Fever, Endocarditis, and Kawasaki Disease Committee of the Council on Cardiovascular Disease in the Young; Council on Cardiovascular and Stroke Nursing; Council on Cardiovascular Surgery and Anesthesia; and Council on Epidemiology and Prevention (2017) Diagnosis, Treatment, and Long-Term Management of Kawasaki Disease: A Scientific Statement for Health Professionals From the American Heart Association. *Circulation*. 135 (17). doi:10.1161/CIR.0000000000000484.

McLean, G.R., Olsen, O.A., Watt, I.N., Rathanaswami, P., Leslie, K.B., Babcook, J.S. & Schrader, J.W. (2005) Recognition of Human Cytomegalovirus by Human Primary Immunoglobulins Identifies an Innate Foundation to an Adaptive Immune Response. *The Journal of Immunology*. 174 (8), 4768–4778. doi:10.4049/jimmunol.174.8.4768.

Medaglia, A.A., Siracusa, L., Gioè, C., Giordano, S., Cascio, A. & Colomba, C. (2021) Kawasaki disease recurrence in the COVID-19 era: a systematic review of the literature. *Italian Journal of Pediatrics*. 47 (1), 95. doi:10.1186/s13052-021-01041-4.

Menikou, S. (2016) *Identification of the causative agent of Kawasaki Disease through isolation and characterization of antigens within immune complexes*. London, UK, Imperial College London.

Menikou, S., McArdle, A.J., Li, M.-S., Kaforou, M., Langford, P.R. & Levin, M. (2020) A proteomics-based method for identifying antigens within immune complexes. *PLOS ONE*. 15 (12), e0244157. doi:10.1371/journal.pone.0244157.

Merchant, S., Wood, D.E. & Salzberg, S.L. (2014) Unexpected cross-species contamination in genome sequencing projects. *PeerJ*. 2, e675. doi:10.7717/peerj.675.

Mertz, D., Frei, R., Periat, N., Zimmerli, M., Battegay, M., Flückiger, U. & Widmer, A.F. (2009) Exclusive Staphylococcus aureus Throat Carriage: At-Risk Populations. *Archives of Internal Medicine*. 169 (2), 172–178. doi:10.1001/archinternmed.2008.536.

Mesuere, B., Debyser, G., Aerts, M., Devreese, B., Vandamme, P. & Dawyndt, P. (2015) The Unipept metaproteomics analysis pipeline. *PROTEOMICS*. 15 (8), 1437–1442. doi:10.1002/pmic.201400361.

Meyer, F., Fritz, A., Deng, Z.-L., Koslicki, D., Gurevich, A., et al. (2021) *Critical Assessment of Metagenome Interpretation - the second round of challenges*.p.2021.07.12.451567. doi:10.1101/2021.07.12.451567.

Miura, M., Garcia, F.L., Crawford, S.E. & Rowley, A.H. (2005) Detection of Kawasaki-disease associated antigen in inflamed gastrointestinal tract in acute Kawasaki disease. *The Pediatric Infectious Disease Journal*. 24 (10), 927–929. doi:10.1097/01.inf.0000180973.98862.50.

Moreira, A., Leite, I., Baptista, A. & Osório Ferreira, E. (2010) Kawasaki disease associated with parainfluenza type 3 virus infection. *Acta dermatovenerologica Croatica: ADC*. 18 (2), 120–123.

Morin, E., Miyauchi, S., San Clemente, H., Chen, E.C.H., Pelin, A., et al. (2019) Comparative genomics of Rhizophagus irregularis, R. cerebriforme, R. diaphanus and Gigaspora rosea highlights specific genetic features in Glomeromycotina. *New Phytologist*. 222 (3), 1584–1598. doi:10.1111/nph.15687.

Mortensen, M.S., Brejnrod, A.D., Roggenbuck, M., Abu Al-Soud, W., Balle, C., Krogfelt, K.A., Stokholm, J., Thorsen, J., Waage, J., Rasmussen, M.A., Bisgaard, H. & Sørensen, S.J. (2016) The developing hypopharyngeal microbiota in early life. *Microbiome*. 4 (1), 70. doi:10.1186/s40168-016-0215-9.

Muth, T., Benndorf, D., Reichl, U., Rapp, E. & Martens, L. (2013) Searching for a needle in a stack of needles: challenges in metaproteomics data analysis. *Molecular BioSystems*. 9 (4), 578–585. doi:10.1039/C2MB25415H.

Muth, T., Kohrs, F., Heyer, R., Benndorf, D., Rapp, E., Reichl, U., Martens, L. & Renard, B.Y. (2018) MPA Portable: A Stand-Alone Software Package for Analyzing Metaproteome Samples on the Go. *Analytical Chemistry*. 90 (1), 685–689. doi:10.1021/acs.analchem.7b03544.

Muth, T., Kolmeder, C.A., Salojärvi, J., Keskitalo, S., Varjosalo, M., Verdam, F.J., Rensen, S.S., Reichl, U., Vos, W.M. de, Rapp, E. & Martens, L. (2015) Navigating through metaproteomics data: A logbook of database searching. *PROTEOMICS*. 15 (20), 3439–3453. doi:10.1002/pmic.201400560.

Muth, T. & Renard, B.Y. (2018) Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? *Briefings in Bioinformatics*. 19 (5), 954–970. doi:10.1093/bib/bbx033.

Muth, T., Renard, B.Y. & Martens, L. (2016) Metaproteomic data analysis at a glance: advances in computational microbial community proteomics. *Expert Review of Proteomics*. 13 (8), 757–769. doi:10.1080/14789450.2016.1209418.

Nakamura, S., Yang, C.-S., Sakon, N., Ueda, M., Tougan, T., et al. (2009) Direct Metagenomic Detection of Viral Pathogens in Nasal and Fecal Specimens Using an Unbiased High-Throughput Sequencing Approach. *PLOS ONE*. 4 (1), e4219. doi:10.1371/journal.pone.0004219.

Nakra, N.A., Blumberg, D.A., Herrera-Guerra, A. & Lakshminrusimha, S. (2020) Multi-System Inflammatory Syndrome in Children (MIS-C) Following SARS-CoV-2 Infection: Review of Clinical Presentation, Hypothetical Pathogenesis, and Proposed Management. *Children*. 7 (7), 69. doi:10.3390/children7070069.

Newburger, J.W., Takahashi, M. & Burns, J.C. (2016) Kawasaki Disease. *Journal of the American College of Cardiology*. 67 (14), 1738–1749. doi:10.1016/j.jacc.2015.12.073.

Nigro, G., Krzysztofiak, A., Porcaro, M.A., Mango, T., Zerbini, M., Gentilomi, G. & Musiani, M. (1994) Active or recent parvovirus B19 infection in children with Kawasaki disease. *The Lancet*. 343 (8908), 1260–1261. doi:10.1016/S0140-6736(94)92154-7.

Noval Rivas, M. & Arditi, M. (2020) Kawasaki disease: pathophysiology and insights from mouse models. *Nature Reviews Rheumatology*. 16 (7), 391–405. doi:10.1038/s41584-020-0426-0.

Oates-Whitehead, R.M., Baumer, J.H., Haines, L., Love, S., Maconochie, I.K., Gupta, A., Roman, K., Dua, J.S. & Flynn, I. (2003) Intravenous immunoglobulin for the treatment of Kawasaki disease in children. *Cochrane Database of Systematic Reviews*. (4). doi:10.1002/14651858.CD004000.

Oechslin, C.P., Lenz, N., Liechti, N., Ryter, S., Agyeman, P., Bruggmann, R., Leib, S.L. & Beuret, C.M. (2018) Limited Correlation of Shotgun Metagenomics Following Host Depletion and Routine Diagnostics for Viruses and Bacteria in Low Concentrated Surrogate and Clinical Samples. *Frontiers in Cellular and Infection Microbiology*. 8. doi:10.3389/fcimb.2018.00375.

Okano, M., Luka, J., Thiele, G.M., Sakiyama, Y., Matsumoto, S. & Purtilo, D.T. (1989) Human herpesvirus 6 infection and Kawasaki disease. *Journal of Clinical Microbiology*. 27 (10), 2379–2380. doi:10.1128/jcm.27.10.2379-2380.1989.

Oliveira, F.S., Brestelli, J., Cade, S., Zheng, J., Iodice, J., Fischer, S., Aurrecoechea, C., Kissinger, J.C., Brunk, B.P., Stoeckert, C.J., Fernandes, G.R., Roos, D.S. & Beiting, D.P. (2018) MicrobiomeDB: a systems biology platform for integrating, mining and analyzing microbiome experiments. *Nucleic Acids Research*. 46 (D1), D684–D691. doi:10.1093/nar/gkx1027.

Olm, M.R., Crits-Christoph, A., Bouma-Gregson, K., Firek, B.A., Morowitz, M.J. & Banfield, J.F. (2021) inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nature Biotechnology*. 39 (6), 727–736. doi:10.1038/s41587-020-00797-0.

Onouchi, Y. (2018) The genetics of Kawasaki disease. *International Journal of Rheumatic Diseases*. 21 (1), 26–30. doi:10.1111/1756-185X.13218.

Onouchi, Y., Suzuki, Y., Suzuki, H., Terai, M., Yasukawa, K., et al. (2013) ITPKC and CASP3 polymorphisms and risks for IVIG unresponsiveness and coronary artery lesion formation in Kawasaki disease. *The Pharmacogenomics Journal*. 13 (1), 52–59. doi:10.1038/tpj.2011.45.

Paniz-Mondolfi, A.E., Akker, T. van den, Márquez-Colmenarez, M.C., Delgado-Noguera, L.A., Valderrama, O. & Sordillo, E.M. (2020) Kawasaki disease seasonality in Venezuela supports an arbovirus infection trigger. *Journal of Medical Virology*. 92 (12), 2903–2910. doi:10.1002/jmv.26381.

Parameswaran, P., Liu, Y., Roskin, K.M., Jackson, K.K.L., Dixit, V.P., et al. (2013) Convergent Antibody Signatures in Human Dengue. *Cell Host & Microbe*. 13 (6), 691–700. doi:10.1016/j.chom.2013.05.008.

Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.-A. & Hugenholtz, P. (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*. 36 (10), 996–1004. doi:10.1038/nbt.4229.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P. & Tyson, G.W. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*. 25 (7), 1043–1055. doi:10.1101/gr.186072.114.

Patra, P.K., Das, R.R., Banday, A.Z., Singh, M., Goyal, K., Jindal, A.K. & Singh, S. (2022) Non-SARS, non-MERS human coronavirus infections and risk of Kawasaki disease: a meta-analysis. *Future Virology*. 17 (1), 37–47. doi:10.2217/fvl-2021-0176.

Patriarca, P., Morens, D., Rogers, M., Schonberger, L., Kaminski, R., Burns, J. & Glode, M. (1982) Kawasakki syndrome: association with the application of rug shampoo. *The Lancet*. 320 (8298), 578–580. doi:10.1016/S0140-6736(82)90660-2.

Payne, A., Holmes, N., Clarke, T., Munro, R., Debebe, B.J. & Loose, M. (2021) Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nature Biotechnology*. 39 (4), 442–450. doi:10.1038/s41587-020-00746-x.

Pedersen, G.K., Höschler, K., Øie Solbak, S.M., Bredholt, G., Pathirana, R.D., Afsar, A., Breakwell, L., Nøstbakken, J.K., Raae, A.J., Brokstad, K.A., Sjursen, H., Zambon, M. & Cox, R.J. (2014) Serum IgG titres, but not avidity, correlates with neutralizing antibody response after H5N1 vaccination. *Vaccine*. 32 (35), 4550–4557. doi:10.1016/j.vaccine.2014.06.009.

Peng, C., Wang, Z., Zhao, P., Ge, W. & Huang, C. (2023) *AbFold -- an AlphaFold Based Transfer Learning Model for Accurate Antibody Structure Prediction*.p.2023.04.20.537598. doi:10.1101/2023.04.20.537598.

Pereira-Marques, J., Hout, A., Ferreira, R.M., Weber, M., Pinto-Ribeiro, I., van Doorn, L.-J., Knetsch, C.W. & Figueiredo, C. (2019) Impact of Host DNA and Sequencing Depth on the Taxonomic

Resolution of Whole Metagenome Sequencing for Microbiome Analysis. *Frontiers in Microbiology*. 10. doi:10.3389/fmicb.2019.01277.

Pierce, N.T., Irber, L., Reiter, T., Brooks, P. & Brown, C.T. (2019) Large-scale sequence comparisons with sourmash. *F1000Research*. 8, 1006. doi:10.12688/f1000research.19675.1.

Pilania, R.K., Bhattarai, D. & Singh, S. (2018) Controversies in diagnosis and management of Kawasaki disease. *World Journal of Clinical Pediatrics*. 7 (1), 27–35. doi:10.5409/wjcp.v7.i1.27.

Piro, V.C., Dadi, T.H., Seiler, E., Reinert, K. & Renard, B.Y. (2020) ganon: precise metagenomics classification against large and up-to-date sets of reference sequences. *Bioinformatics*. 36 (Supplement_1), i12–i20. doi:10.1093/bioinformatics/btaa458.

Pombert, J.-F. (2021) *PombertLab/MMH: MMH-v0.3.1*. doi:10.5281/zenodo.5532799.

Potgieter, M.G., Nel, A.J., Fortuin, S., Garnett, S., Wendoh, J.M., Tabb, D.L., Mulder, N.J. & Blackburn, J.M. (2019) *MetaNovo : a probabilistic approach to peptide and polymorphism discovery in complex metaproteomic datasets*. doi:10.1101/605550.

Qu, J., Zhang, Q., Zhang, N., Shen, L. & Liu, P. (2015) Microbial Community Diversity in Water and Sediment of an Eutrophic Lake during Harmful Algal Bloom Using MiSeq Illumina Technology. *International Proceedings of Chemical, Biological and Environmental Engineering*. 87, 67–72. doi:10.7763/IPCBEE.2015.V87.12.

Quiat, D., Kula, T., Shimizu, C., Kanegaye, J.T., Tremoulet, A.H., Pitkowsky, Z., Son, M., Newburger, J.W., Elledge, S.J. & Burns, J.C. (2020) High-Throughput Screening of Kawasaki Disease Sera for Antiviral Antibodies. *The Journal of Infectious Diseases*. 222 (11), 1853–1857. doi:10.1093/infdis/jiaa253.

Quince, C., Delmont, T.O., Raguideau, S., Alneberg, J., Darling, A.E., Collins, G. & Eren, A.M. (2017) DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biology*. 18 (1), 181. doi:10.1186/s13059-017-1309-9.

Quince, C., Nurk, S., Raguideau, S., James, R., Soyer, O.S., Summers, J.K., Limasset, A., Eren, A.M., Chikhi, R. & Darling, A.E. (2021) STRONG: metagenomics strain resolution on assembly graphs. *Genome Biology*. 22 (1), 214. doi:10.1186/s13059-021-02419-7.

Rechenberger, J., Samaras, P., Jarzab, A., Behr, J., Frejno, M., Djukovic, A., Sanz, J., González-Barberá, E.M., Salavert, M., López-Hontangas, J.L., Xavier, K.B., Debrauwer, L., Rolain, J.-M., Sanz, M., Garcia-Garcera, M., Wilhelm, M., Ubeda, C. & Kuster, B. (2019) Challenges in Clinical Metaproteomics Highlighted by the Analysis of Acute Leukemia Patients with Gut Colonization by Multidrug-Resistant Enterobacteriaceae. *Proteomes*. 7 (1), 2. doi:10.3390/proteomes7010002.

Rice, P., Longden, I. & Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics: TIG*. 16 (6), 276–277. doi:10.1016/s0168-9525(00)02024-2.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. & Smyth, G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*. 43 (7), e47. doi:10.1093/nar/gkv007.

Rodó, X., Ballester, J., Cayan, D., Melish, M.E., Nakamura, Y., Uehara, R. & Burns, J.C. (2011) Association of Kawasaki disease with tropospheric wind patterns. *Scientific Reports*. 1, 152. doi:10.1038/srep00152.

Rodó, X., Curcoll, R., Robinson, M., Ballester, J., Burns, J.C., Cayan, D.R., Lipkin, W.I., Williams, B.L., Couto-Rodriguez, M., Nakamura, Y., Uehara, R., Tanimoto, H. & Morguí, J.-A. (2014) Tropospheric winds from northeastern China carry the etiologic agent of Kawasaki disease from its source to Japan. *Proceedings of the National Academy of Sciences of the United States of America*. 111 (22), 7952–7957. doi:10.1073/pnas.1400380111.

Rodriguez-R, L.M., Gunturu, S., Tiedje, J.M., Cole, J.R. & Konstantinidis, K.T. (2018) Nonpareil 3: Fast Estimation of Metagenomic Coverage and Sequence Diversity. *mSystems*. 3 (3), e00039-18. doi:10.1128/mSystems.00039-18.

Rossmann, S.N., Wilson, P.H., Hicks, J., Carter, B., Cron, S.G., Simon, C., Flaitz, C.M., Demmler, G.J., Shearer, W.T. & Kline, M.W. (1998) Isolation of Lautropia mirabilis from Oral Cavities of Human Immunodeficiency Virus-Infected Children. *Journal of Clinical Microbiology*. 36 (6), 1756–1760. doi:10.1128/JCM.36.6.1756-1760.1998.

Rowley, A.H., Baker, S.C., Arrollo, D., Gruen, L.J., Bodnar, T., Innocentini, N., Hackbart, M., Cruz-Pulido, Y.E., Wylie, K.M., Kim, K.-Y.A. & Shulman, S.T. (2020) A Protein Epitope Targeted by the Antibody Response to Kawasaki Disease. *The Journal of Infectious Diseases*. 222 (1), 158–168. doi:10.1093/infdis/jiaa066.

Rowley, A.H., Baker, S.C., Shulman, S.T., Fox, L.M., Takahashi, K., Garcia, F.L., Crawford, S.E., Chou, P. & Orenstein, J.M. (2005) Cytoplasmic Inclusion Bodies Are Detected by Synthetic Antibody in Ciliated Bronchial Epithelium during Acute Kawasaki Disease. *The Journal of Infectious Diseases*. 192 (10), 1757–1766. doi:10.1086/497171.

Rowley, A.H., Baker, S.C., Shulman, S.T., Garcia, F.L., Fox, L.M., Kos, I.M., Crawford, S.E., Russo, P.A., Hammadeh, R., Takahashi, K. & Orenstein, J.M. (2008) RNA-Containing Cytoplasmic Inclusion Bodies in Ciliated Bronchial Epithelium Months to Years after Acute Kawasaki Disease. *PLOS ONE*. 3 (2), e1582. doi:10.1371/journal.pone.0001582.

Rowley, A.H., Baker, S.C., Shulman, S.T., Garcia, F.L., Guzman-Cottrill, J.A., Chou, P., Terai, M., Kawasaki, T., Kalelkar, M.B. & Crawford, S.E. (2004) Detection of antigen in bronchial epithelium and macrophages in acute Kawasaki disease by use of synthetic antibody. *The Journal of Infectious Diseases*. 190 (4), 856–865. doi:10.1086/422648.

Rowley, A.H., Baker, S.C., Shulman, S.T., Rand, K.H., Tretiakova, M.S., Perlman, E.J., Garcia, F.L., Tajuddin, N.F., Fox, L.M., Huang, J.H., Ralphe, J.C., Takahashi, K., Flatow, J., Lin, S., Kalelkar, M.B., Soriano, B. & Orenstein, J.M. (2011) Ultrastructural, Immunofluorescence, and RNA Evidence Support the Hypothesis of a "New" Virus Associated With Kawasaki Disease. *The Journal of Infectious Diseases*. 203 (7), 1021–1030. doi:10.1093/infdis/jiq136.

Rowley, A.H. & Shulman, S.T. (2018) The Epidemiology and Pathogenesis of Kawasaki Disease. *Frontiers in Pediatrics*. 6. doi:10.3389/fped.2018.00374.

Rowley, A.H., Wolinsky, S.M., Relman, D.A., Sambol, S.P., Sullivan, J., Terai, M. & Shulman, S.T. (1994) Search for Highly Conserved Viral and Bacterial Nucleic Acid Sequences Corresponding to an Etiologic Agent of Kawasaki Disease. *Pediatric Research*. 36 (5), 567–570. doi:10.1203/00006450-199411000-00003.

Salo, E., Pelkonen, P., Kekomaki, R., Ruuskanen, O., Viander, M. & Wager, O. (1987) Kawasaki disease: circulating immune complexes monitored during the disease. *Progress in Clinical and Biological Research*. 250, 563.

Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P., Parkhill, J., Loman, N.J. & Walker, A.W. (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*. 12 (1), 87. doi:10.1186/s12915-014-0087-z.

Sancho-Shimizu, V., Brodin, P., Cobat, A., Biggs, C.M., Toubiana, J., Lucas, C.L., Henrickson, S.E., Belot, A., MIS-C@CHGE, Tangye, S.G., Milner, J.D., Levin, M., Abel, L., Bogunovic, D., Casanova, J.-L. & Zhang, S.-Y. (2021) SARS-CoV-2–related MIS-C: A key to the viral and genetic causes of Kawasaki disease? *Journal of Experimental Medicine*. 218 (6), e20210446. doi:10.1084/jem.20210446.

Sano, T., Makino, N., Aoyama, Y., Ae, R., Kojo, T., Kotani, K., Nakamura, Y. & Yanagawa, H. (2016) Temporal and geographical clustering of Kawasaki disease in Japan: 2007–2012. *Pediatrics International*. 58 (11), 1140–1145. doi:10.1111/ped.12970.

Sasaki, M., Shimoyama, Y., Ishikawa, T., Kodama, Y., Tajika, S. & Kimura, S. (2020) Contribution of different adherent properties of *Granulicatella adiacens* and *Abiotrophia defectiva* to their associations with oral colonization and the risk of infective endocarditis. *Journal of Oral Science*. 62 (1), 36–39. doi:10.2334/josnusd.19-0021.

Sasaki, M., Shimoyama, Y., Kodama, Y. & Ishikawa, T. (2021) Abiotrophia defectiva DnaK Promotes Fibronectin-Mediated Adherence to HUVECs and Induces a Proinflammatory Response. *International Journal of Molecular Sciences*. 22 (16), 8528. doi:10.3390/ijms22168528.

Sato, S., Beausoleil, S.A., Popova, L., Beaudet, J.G., Ramenani, R.K., Zhang, X., Wieler, J.S., Schieferl, S.M., Cheung, W.C. & Polakiewicz, R.D. (2012) Proteomics-directed cloning of circulating antiviral human monoclonal antibodies. *Nature Biotechnology*. 30, 1039–1043. doi:10.1038/nbt.2406.

Scherl, A., Shaffer, S.A., Taylor, G.K., Hernandez, P., Appel, R.D., Binz, P.-A. & Goodlett, D.R. (2008) On the Benefits of Acquiring Peptide Fragment Ions at High Measured Mass Accuracy. *Journal of the American Society for Mass Spectrometry*. 19 (6), 891–901. doi:10.1016/j.jasms.2008.02.005.

Schroeder, H.W. & Cavacini, L. (2010) Structure and function of immunoglobulins. *Journal of Allergy and Clinical Immunology*. 125 (2), S41–S52. doi:10.1016/j.jaci.2009.09.046.

Segata, N. (2018) On the Road to Strain-Resolved Comparative Metagenomics. *mSystems*. 3 (2), e00190-17. doi:10.1128/mSystems.00190-17.

Senn, L., Entenza, J.M. & Prod'hom, G. (2006) Adherence of Abiotrophia defectiva and Granulicatella species to fibronectin: is there a link with endovascular infections? *FEMS Immunology & Medical Microbiology*. 48 (2), 215–217. doi:10.1111/j.1574-695X.2006.00142.x.

Seppey, M., Manni, M. & Zdobnov, E.M. (2020) LEMMI: a continuous benchmarking platform for metagenomics classifiers. *Genome Research*. 30 (8), 1208–1216. doi:10.1101/gr.260398.119.

Shen, C.T., Wu, H.Y., Wang, N.K. & Huang, C.S. (1990) Reevaluation of streptococcal infection in the pathogenesis of Kawasaki disease. *Zhonghua Minguo Xiao Er Ke Yi Xue Hui Za Zhi [Journal]. Zhonghua Minguo Xiao Er Ke Yi Xue Hui*. 31 (3), 144–150.

Shen, J., Ding, Y., Yang, Z., Zhang, X. & Zhao, M. (2020) Effects of changes on gut microbiota in children with acute Kawasaki disease. *PeerJ*. 8, e9698. doi:10.7717/peerj.9698.

Shimizu, C., Shike, H., Baker, S.C., Garcia, F., Hoek, L. van der, Kuijpers, T.W., Reed, S.L., Rowley, A.H., Shulman, S.T., Talbot, H.K.B., Williams, J.V. & Burns, J.C. (2005) Human Coronavirus NL63 Is Not Detected in the Respiratory Tracts of Children with Acute Kawasaki Disease. *The Journal of Infectious Diseases*. 192 (10), 1767–1771. doi:10.1086/497170.

Shinomiya, N., Takeda, T., Kuratsuji, T., Takagi, K., Kosaka, T., Tatsuzawa, O., Tsurumizu, T., Hashimoto, T. & Kobayashi, N. (1987) Variant Streptococcus sanguis as an etiological agent of Kawasaki disease. *Progress in Clinical and Biological Research*. 250, 571–572.

Shirato, K., Imada, Y., Kawase, M., Nakagaki, K., Matsuyama, S. & Taguchi, F. (2014) Possible involvement of infection with human coronavirus 229E, but not NL63, in Kawasaki disease. *Journal of Medical Virology*. 86 (12), 2146–2153. doi:10.1002/jmv.23950.

Shishido, A. (1979) Failure to confirm the rickettsial etiology of MCLS (Kawasaki disease). *Japanese Journal of Medical Science & Biology*. 32 (4), 250–251.

Silamikelis, I. (2021) *Conifer*. https://github.com/Ivarz/Conifer.

Sim, B.K., Park, H., Kim, J.-J., Yun, S.W., Yu, J.J., Yoon, K.L., Lee, K.-Y., Kil, H.-R., Kim, G.B., Han, M.-K., Song, M.S., Lee, H.D., Ha, K.S., Sohn, S., Hong, Y.M., Jang, G.Y. & Lee, J.-K. (2018) Assessment of the Clinical Heterogeneity of Kawasaki Disease Using Genetic Variants of BLK and FCGR2A. *Korean Circulation Journal*. 49 (1), 99–108. doi:10.4070/kcj.2018.0224.

Singh, S., Vignesh, P. & Burgner, D. (2015) The epidemiology of Kawasaki disease: a global update. *Archives of Disease in Childhood*. 100 (11), 1084–1088. doi:10.1136/archdischild-2014-307536.

Sinha, R., Stanley, G., Gulati, G.S., Ezran, C., Travaglini, K.J., Wei, E., Chan, C.K.F., Nabhan, A.N., Su, T., Morganti, R.M., Conley, S.D., Chaib, H., Red-Horse, K., Longaker, M.T., Snyder, M.P., Krasnow, M.A. & Weissman, I.L. (2017) *Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing*. doi:10.1101/125724.

Sollid, J.U.E., Furberg, A.S., Hanssen, A.M. & Johannessen, M. (2014) Staphylococcus aureus: Determinants of human carriage. *Infection, Genetics and Evolution*. 21, 531–541. doi:10.1016/j.meegid.2013.03.020.

Starke, R., Pylro, V.S. & Morais, D.K. (2021) 16S rRNA Gene Copy Number Normalization Does Not Provide More Reliable Conclusions in Metataxonomic Surveys. *Microbial Ecology*. 81 (2), 535–539. doi:10.1007/s00248-020-01586-7.

de Steenhuijsen Piters, W.A.A., Heinonen, S., Hasrat, R., Bunsow, E., Smith, B., Suarez-Arrabal, M.-C., Chaussabel, D., Cohen, D.M., Sanders, E.A.M., Ramilo, O., Bogaert, D. & Mejias, A. (2016) Nasopharyngeal Microbiota, Host Transcriptome, and Disease Severity in Children with Respiratory Syncytial Virus Infection. *American Journal of Respiratory and Critical Care Medicine*. 194 (9), 1104–1115. doi:10.1164/rccm.201602-0220OC.

Steinegger, M. & Salzberg, S.L. (2020) Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biology*. 21 (1), 115. doi:10.1186/s13059-020-02023-1.

Steinegger, M. & Söding, J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*. 35 (11), 1026–1028. doi:10.1038/nbt.3988.

Stoddard, S.F., Smith, B.J., Hein, R., Roller, B.R.K. & Schmidt, T.M. (2015) rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Research*. 43 (D1), D593–D598. doi:10.1093/nar/gku1201.

Strigl, S., Kutlin, A., Roblin, P.M., Shulman, S. & Hammerschlag, M.R. (2000) Is There an Association between Kawasaki Disease and Chlamydia pneumoniae? *The Journal of Infectious Diseases*. 181 (6), 2103–2105. doi:10.1086/315526.

Swaby, E.D., Fisher-Hoch, S., Lambert, H.P. & Stern, H. (1980) IS KAWASAKI DISEASE A VARIANT OF Q FEVER? *The Lancet*. 316 (8186), 146. doi:10.1016/S0140-6736(80)90026-4.

Swindells, M.B., Porter, C.T., Couch, M., Hurst, J., Abhinandan, K.R., Nielsen, J.H., Macindoe, G., Hetherington, J. & Martin, A.C.R. (2017) abYsis: Integrated Antibody Sequence and Structure—Management, Analysis, and Prediction. *Journal of Molecular Biology*. 429 (3), 356–364. doi:10.1016/j.jmb.2016.08.019.

Tanca, A., Palomba, A., Deligios, M., Cubeddu, T., Fraumene, C., Biosa, G., Pagnozzi, D., Addis, M.F. & Uzzau, S. (2013) Evaluating the Impact of Different Sequence Databases on Metaproteome Analysis: Insights from a Lab-Assembled Microbial Mixture. *PLOS ONE*. 8 (12), e82981. doi:10.1371/journal.pone.0082981.

Tanca, A., Palomba, A., Fraumene, C., Pagnozzi, D., Manghina, V., Deligios, M., Muth, T., Rapp, E., Martens, L., Addis, M.F. & Uzzau, S. (2016) The impact of sequence database choice on metaproteomic results in gut microbiota studies. *Microbiome*. 4 (1), 51. doi:10.1186/s40168-016-0196-8.

Taylor, P.J. (2005) Matrix effects: the Achilles heel of quantitative high-performance liquid chromatography–electrospray–tandem mass spectrometry. *Clinical Biochemistry*. 38 (4), 328–334. doi:10.1016/j.clinbiochem.2004.11.007.

The, M. & Käll, L. (2020) Focus on the spectra that matter by clustering of quantification data in shotgun proteomics. *Nature Communications*. 11 (1), 3234. doi:10.1038/s41467-020-17037-3.

Thissen, J.B., Isshiki, M., Jaing, C., Nagao, Y., Aldea, D.L., Allen, J.E., Izui, M., Slezak, T.R., Ishida, T. & Sano, T. (2018) A novel variant of torque teno virus 7 identified in patients with Kawasaki disease. *PLOS ONE*. 13 (12), e0209683. doi:10.1371/journal.pone.0209683.

Tran, N.H., Rahman, M.Z., He, L., Xin, L., Shan, B. & Li, M. (2016) Complete De Novo Assembly of Monoclonal Antibody Sequences. *Scientific Reports*. 6, 31730. doi:10.1038/srep31730.

Tremoulet, A.H., Jain, S., Jaggi, P., Jimenez-Fernandez, S., Pancheri, J.M., Sun, X., Kanegaye, J.T., Kovalchin, J.P., Printz, B.F., Ramilo, O. & Burns, J.C. (2014) Infliximab for intensification of primary therapy for Kawasaki disease: a phase 3 randomised, double-blind, placebo-controlled trial. *The Lancet*. 383 (9930), 1731–1738. doi:10.1016/S0140-6736(13)62298-9.

Truong, D.T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. (2017) Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Research*. 27 (4), 626–638. doi:10.1101/gr.216242.116.

Tulloh, R.M.R., Mayon-White, R., Harnden, A., Ramanan, A.V., Tizard, E.J., Shingadia, D., Michie, C.A., Lynn, R.M., Levin, M., Franklin, O.D., Craggs, P., Davidson, S., Stirzaker, R., Danson, M. & Brogan, P.A. (2019) Kawasaki disease: a prospective population survey in the UK and Ireland from 2013 to 2015. *Archives of Disease in Childhood*. 104 (7), 640–646. doi:10.1136/archdischild-2018-315087.

Tyanova, S., Temu, T. & Cox, J. (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols*. 11 (12), 2301–2319. doi:10.1038/nprot.2016.136.

Uehara, R., Igarashi, H., Yashiro, M., Nakamura, Y. & Yanagawa, H. (2010) Kawasaki Disease Patients With Redness or Crust Formation at the Bacille Calmette-Guérin Inoculation Site. *The Pediatric Infectious Disease Journal*. 29 (5), 430–433. doi:10.1097/INF.0b013e3181cacede.

UK Health Security Agency (2022) *Investigation into acute hepatitis of unknown aetiology in children in England: technical briefing 4*.p.56.

Utter, D.R., Mark Welch, J.L. & Borisy, G.G. (2016) Individuality, Stability, and Variability of the Plaque Microbiome. *Frontiers in Microbiology*. 7. https://www.frontiersin.org/article/10.3389/fmicb.2016.00564.

VanDuijn, M.M., Dekker, L.J., van IJcken, W.F.J., Sillevis Smitt, P.A.E. & Luider, T.M. (2017) Immune Repertoire after Immunization As Seen by Next-Generation Sequencing and Proteomics. *Frontiers in Immunology*. 8. doi:10.3389/fimmu.2017.01286.

Verbruggen, S., Gessulat, S., Gabriels, R., Matsaroki, A., Van de Voorde, H., Kuster, B., Degroeve, S., Martens, L., Van Criekinge, W., Wilhelm, M. & Menschaert, G. (2021) Spectral Prediction Features as a Solution for the Search Space Size Problem in Proteogenomics. *Molecular & Cellular Proteomics*. 20, 100076. doi:10.1016/j.mcpro.2021.100076.

Vidarsson, G., Dekkers, G. & Rispens, T. (2014) IgG Subclasses and Allotypes: From Structure to Effector Functions. *Frontiers in Immunology*. 5. https://www.frontiersin.org/article/10.3389/fimmu.2014.00520.

Visco, C., Maura, F., Tuana, G., Agnelli, L., Lionetti, M., Fabris, S., Novella, E., Giaretta, I., Reda, G., Barcellini, W., Baldini, L., Neri, A., Rodeghiero, F. & Cortelezzi, A. (2012) Immune Thrombocytopenia in Patients with Chronic Lymphocytic Leukemia Is Associated with Stereotyped B-cell Receptors. *Clinical Cancer Research*. 18 (7), 1870–1878. doi:10.1158/1078-0432.CCR-11-3019.

Wang, S., Jiang, Y. & Li, S. (2020) PStrain: an iterative microbial strains profiling algorithm for shotgun metagenomic sequencing data. *Bioinformatics*. 36 (22–23), 5499–5506. doi:10.1093/bioinformatics/btaa1056.

Wardle, A.J., Connolly, G.M., Seager, M.J. & Tulloh, R.M. (2017) Corticosteroids for the treatment of Kawasaki disease in children. *Cochrane Database of Systematic Reviews*. (1). doi:10.1002/14651858.CD011188.pub2.

Watson, C.T. & Breden, F. (2012) The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. *Genes & Immunity*. 13 (5), 363–373. doi:10.1038/gene.2012.12.

Watson, C.T., Glanville, J. & Marasco, W.A. (2017) The Individual and Population Genetics of Antibody Immunity. *Trends in Immunology*. 38 (7), 459–470. doi:10.1016/j.it.2017.04.003.

Wen, W., Su, W., Tang, H., Le, W., Zhang, X., Zheng, Y., Liu, X., Xie, L., Li, J., Ye, J., Dong, L., Cui, X., Miao, Y., Wang, D., Dong, J., Xiao, C., Chen, W. & Wang, H. (2020) Immune cell profiling of COVID-19 patients in the recovery stage by single-cell sequencing. *Cell Discovery*. 6 (1), 1–18. doi:10.1038/s41421-020-0168-9.

Weng, K.-P., Cheng-Chung Wei, J., Hung, Y.-M., Huang, S.-H., Chien, K.-J., Lin, C.-C., Huang, S.-M., Lin, C.-L. & Cheng, M.-F. (2018) Enterovirus Infection and Subsequent Risk of Kawasaki Disease: A Population-based Cohort Study. *The Pediatric Infectious Disease Journal*. 37 (4), 310–315. doi:10.1097/INF.0000000000001748.

Westoby, M., Nielsen, D.A., Gillings, M.R., Litchman, E., Madin, J.S., Paulsen, I.T. & Tetu, S.G. (2021) Cell size, genome size, and maximum growth rate are near-independent dimensions of ecological variation across bacteria and archaea. *Ecology and Evolution*. 11 (9), 3956–3976. doi:10.1002/ece3.7290.

Whittaker, E., Bamford, A., Kenny, J., Kaforou, M., Jones, C.E., et al. (2020) Clinical Characteristics of 58 Children With a Pediatric Inflammatory Multisystem Syndrome Temporally Associated With SARS-CoV-2. *JAMA*. 324 (3), 259–269. doi:10.1001/jama.2020.10369.

Wilmes, P. & Bond, P.L. (2004) The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environmental Microbiology*. 6 (9), 911–920. doi:10.1111/j.1462-2920.2004.00687.x.

Wood, D.E., Lu, J. & Langmead, B. (2019) Improved metagenomic analysis with Kraken 2. *Genome Biology*. 20 (1), 257. doi:10.1186/s13059-019-1891-0.

Wood, D.E. & Salzberg, S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*. 15 (3), R46. doi:10.1186/gb-2014-15-3-r46.

Wright, V.J., Herberg, J.A., Kaforou, M., Shimizu, C., Eleftherohorinou, H., et al. (2018) Diagnosis of Kawasaki Disease Using a Minimal Whole-Blood Gene Expression Signature. *JAMA Pediatrics*. 172 (10), e182293–e182293. doi:10.1001/jamapediatrics.2018.2293.

Xiao, J., Tanca, A., Jia, B., Yang, R., Wang, B., Zhang, Y. & Li, J. (2018) Metagenomic Taxonomy-Guided Database-Searching Strategy for Improving Metaproteomic Analysis. *Journal of Proteome Research*. 17 (4), 1596–1605. doi:10.1021/acs.jproteome.7b00894.

Xu, G.J., Kula, T., Xu, Q., Li, M.Z., Vernon, S.D., Ndung'u, T., Ruxrungtham, K., Sanchez, J., Brander, C., Chung, R.T., O'Connor, K.C., Walker, B., Larman, H.B. & Elledge, S.J. (2015) Comprehensive serological profiling of human populations using a synthetic human virome. *Science*. 348 (6239), aaa0698. doi:10.1126/science.aaa0698.

Yamaji, N., Lopes, K. da S., Shoda, T., Ishitsuka, K., Kobayashi, T., Ota, E. & Mori, R. (2019) TNF-α blockers for the treatment of Kawasaki disease in children. *Cochrane Database of Systematic Reviews*. (8). doi:10.1002/14651858.CD012448.pub2.

Yoshino, S., Fujimoto, K., Takada, T., Kawamura, S., Ogawa, J., Kamata, Y., Kodera, Y. & Shichiri, M. (2019) Molecular form and concentration of serum α2-macroglobulin in diabetes. *Scientific Reports*. 9 (1), 12927. doi:10.1038/s41598-019-49144-7.

Yu, F., Haynes, S.E. & Nesvizhskii, A.I. (2021) IonQuant Enables Accurate and Sensitive Label-Free Quantification With FDR-Controlled Match-Between-Runs. *Molecular & Cellular Proteomics*. 20, 100077. doi:10.1016/j.mcpro.2021.100077.

Yu, F., Teo, G.C., Kong, A.T., Haynes, S.E., Avtonomov, D.M., Geiszler, D.J. & Nesvizhskii, A.I. (2020) Identification of modified peptides using localization-aware open search. *Nature Communications*. 11 (1), 4065. doi:10.1038/s41467-020-17921-y.

Zhang, X., Ning, Z., Mayne, J., Moore, J.I., Li, J., Butcher, J., Deeke, S.A., Chen, R., Chiang, C.-K., Wen, M., Mack, D., Stintzi, A. & Figeys, D. (2016) MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. *Microbiome*. 4 (1), 31. doi:10.1186/s40168-016-0176-z.

Appendix A – Kawasaki disease and febrile control patient cohorts

University of California San Diego / Rady Children's Hospital Patients

KD patients met the case definition of the American Heart Association for either complete or incomplete KD. In order to avoid the potential for misclassification or confusion with MIS-C, all KD subjects were enrolled from 2004-2017 prior to the SARS-CoV-2 pandemic. All KD subjects were diagnosed and treated by one of two highly experienced KD clinicians. Febrile control subjects were also enrolled from 2002-2017 and met the following case definition: previously healthy child with fever for at least 3 days plus at least one of the clinical criteria for KD. Over 50% of the FC were referred for evaluation because of a clinical suspicion for KD. The final diagnoses for the FC were adjudicated 2-3 months after enrolment by two experienced paediatric clinicians who reviewed the clinical outcomes in the medical record and all available test results. A viral syndrome was defined as a self-limited illness that resolved without treatment and without apparent sequelae. Written consent or assent as appropriate was obtained from parents and subjects and the study was approved by the Institutional Review Board at UCSD (Human Research Protection Program 140220).

Imperial College London / St Mary's Hospital Patients

KD patients, febrile controls and healthy children were previously recruited, with written parental informed consent, under approvals by the research ethics committees of the United Kingdom (St Mary's Hospital 09/H0712/58, 13/LO/0026, EC3263). Febrile controls were drawn from the Immunopathology of Respiratory Infection Study (IRIS) recruiting between 2008 and 2015 (Herberg et al., 2013; Cebey-López et al., 2015, 2016; Wright et al., 2018). Children were recruited from the emergency department of St Mary's Hospital, London if they had fever or were suspected by the clinical team to have an infectious or inflammatory disorder. Healthy controls were recruited from outpatient clinics as part of the IRIS study, and from unrelated contacts of meningococcal disease cases. Control samples from adults with TB (metaproteomics only) were obtained from the IGRAs in Diagnostic Evaluation of Active TB (IDEA) study (11/H0722/8 Hoang et al., 2021) kindly provided by Prof Ajit Lalvani.

Samples from children with KD were drawn from the study "Genetic determinants of Kawasaki Disease for susceptibility and outcome." As per Wright et al. (2018): "Kawasaki disease was diagnosed on the basis of the American Heart Association criteria, with 2-dimensional echocardiography performed soon after presentation (2 and 6 weeks after onset). Patients with fewer than 4 of the 5 classic criteria (bilateral nonpurulent conjunctivitis, oral mucosal changes, peripheral extremity changes, rash, and cervical lymphadenopathy >1.5 cm) were included as having incomplete KD if the maximum coronary artery z score (Z_{max})… at any time during the illness for the left anterior descending or right coronary

arteries was 2.5 or higher or if the patients satisfied the algorithm for incomplete KD in the American Heart Association guidelines."

Appendix B – Preliminary exploration of host DNA and contamination in metagenomics

Summary

This appendix is derived from a finalised manuscript published in Access Microbiology (McArdle & Kaforou, 2020).

Motivated by the need to inform the current study, and prior to the availability of our metagenomic data, I reviewed and reanalysed an experimental study of the impact of sequencing depth and varying host DNA proportion on metagenomic analyses (Pereira-Marques et al., 2019). This study reported that increasing host DNA abundance and reducing read depth impaired the sensitivity to low abundance micro-organisms.

Importantly, the study did not consider contamination and applied a less sensitive marker gene approach for identification and quantification. My reanalysis using Kraken 2 identified contamination as an issue, and reframed the problem within the existing domain of low biomass metagenomics.

The analysis convinced me of the critical need to account for contamination and the need to apply read-binning tools rather than marker gene approaches for maximum sensitivity. Importantly, during this analysis, I was unaware of the issue of eukaryotic contamination of microbial reference genomes. It is therefore possible that some of the apparent contamination detected herein relates to reference database errors.

Introduction

The study of metagenomics and microbiomes has yielded impressive insights into the microbiology of the environment and of multicellular organisms in health and disease (Escobar-Zepeda, Vera-Ponce de León & Sanchez-Flores, 2015).

Although more expensive than amplicon-based microbiome approaches (e.g. 16S), shotgun metagenomics is increasingly gaining prominence. Benefits include no polymerase-chain reaction related bias, greater specificity of identifications and representation of diversity, and ability to detect organisms from all kingdoms (Ranjan et al., 2016). Additionally, metagenomic sequences can be analysed functionally, and whole or partial metagenomes reconstructed with greater depth of sequencing (Quince et al., 2017).

However, high depth sequencing does not guarantee abundant microbial reads. Challenges most frequently arise when microbial biomass is low (Eisenhofer et al., 2019; Weyrich et al., 2019; Karstens et al., 2019). Total DNA will be limited, and few reads may be obtained. Further, the quantity of

contaminant organisms is likely to remain constant, thus their relative contribution will increase. The same problem can arise when samples are dominated by DNA from a host organism – in these cases, host sequencing reads may vastly outnumber those from microbes.

Although, techniques exist to mitigate this by selectively depleting host DNA, usually by removing free DNA before lysis (Marotz et al., 2018; Feehery et al., 2013; Hasan et al., 2016; Nelson et al., 2019), they are in their infancy and could also deplete DNA from dead or damaged organisms, which would include those under immune attack (Oechslin et al., 2018). Depleting host DNA would not reduce the impact of contamination occurring prior to depletion.

In this context, we commend Pereira-Marques et al on their insightful study into the effects of host DNA and read depth on microbial abundance estimates from shotgun metagenomics (Pereira-Marques et al., 2019).

The authors evaluated the impact of a range of amounts of host DNA and sequencing depths on microbiome taxonomic profiling using shotgun metagenomic sequencing, from synthetic samples where bacterial DNA from 20 species of varying abundances was spiked with varying amounts of murine DNA.

The authors showed that increasing proportions of host DNA (10, 90 and 99%) led to decreased sensitivity in detecting very low and low abundant species, increasing the number of undetected species.

Although not stated, we anticipate the authors may have selected MetaPhlAn2 for their analysis because by detecting clade-specific marker genes of known number per organism, relative abundances within a sample can be directly estimated (Segata et al., 2012). Despite this advantage, we are concerned that relying upon a small number of marker genes will render the approach more sensitive to reduced depth than read binning approaches.

Consequently we applied Kraken, a fast and sensitive read binning tool (Wood & Salzberg, 2014), which performed well in recent benchmarks (Lindgreen, Adair & Gardner, 2016; Ye et al., 2019). Advantageously, a partner tool (Bracken) also exists for relative abundance estimation (Lu et al., 2017). We obtained the variable-length trimmed reads from the study (NCBI sequence read archive accession PRJNA521492) and built a Kraken database comprising bacteria, fungi, viruses, archaea and mouse genome sequences with core vector elements. Kraken (version 2.0.8-beta) was then run with default settings, followed by Bracken.

For each sample we categorised reads assigned to any microbial operational taxonomic unit (OTU) as microbial. We follow the sample naming conventions of the original analysis: MS = microbial sample; SS10 = 10% host DNA; SS90 = 90% host DNA; SS99 = 99% host DNA.

Sensitivity

All expected organisms (n=20) were detected in all samples. This contrasts with the results presented in Pereira-Marquez et al where 9 of the 20 species became undetectable in SS99.

Over 75% of microbial reads were allocated to the known species (on target), except in sample SS99 where this fell to 67%. Other species of the expected genera represented much fewer than 1% of microbial reads in all samples. Fewer than 2% of microbial reads were assigned to OTUs outside of the lineage of the expected genera (off target), except for SS99 where this was 12% (data not shown).

Relative abundance

Crude assigned read counts are not a guide to relative abundance because of varying genome size, and because reads from different organisms may be assigned at species level at differing rates due to homology. Bracken was developed to overcome the second limitation by reallocating reads assigned to higher levels. We apply Bracken here at species level to estimate abundance and then correct for genome size. The Bracken database was built for a read length of 150 (the median length of the trimmed reads).

Bracken estimated that over 98% of microbial reads were on-target (species) in MS and SS10. In SS90 this fell to 96.8% and in SS99 to 83.3%.

We normalised abundances by genome size (obtained from NCBI genomes at https://www.ncbi.nlm.nih.gov/genome) for the target species, discounting the small proportion of off-target reads. In MS, the ratios of observed:expected relative abundance was between 0.5 to 2 for

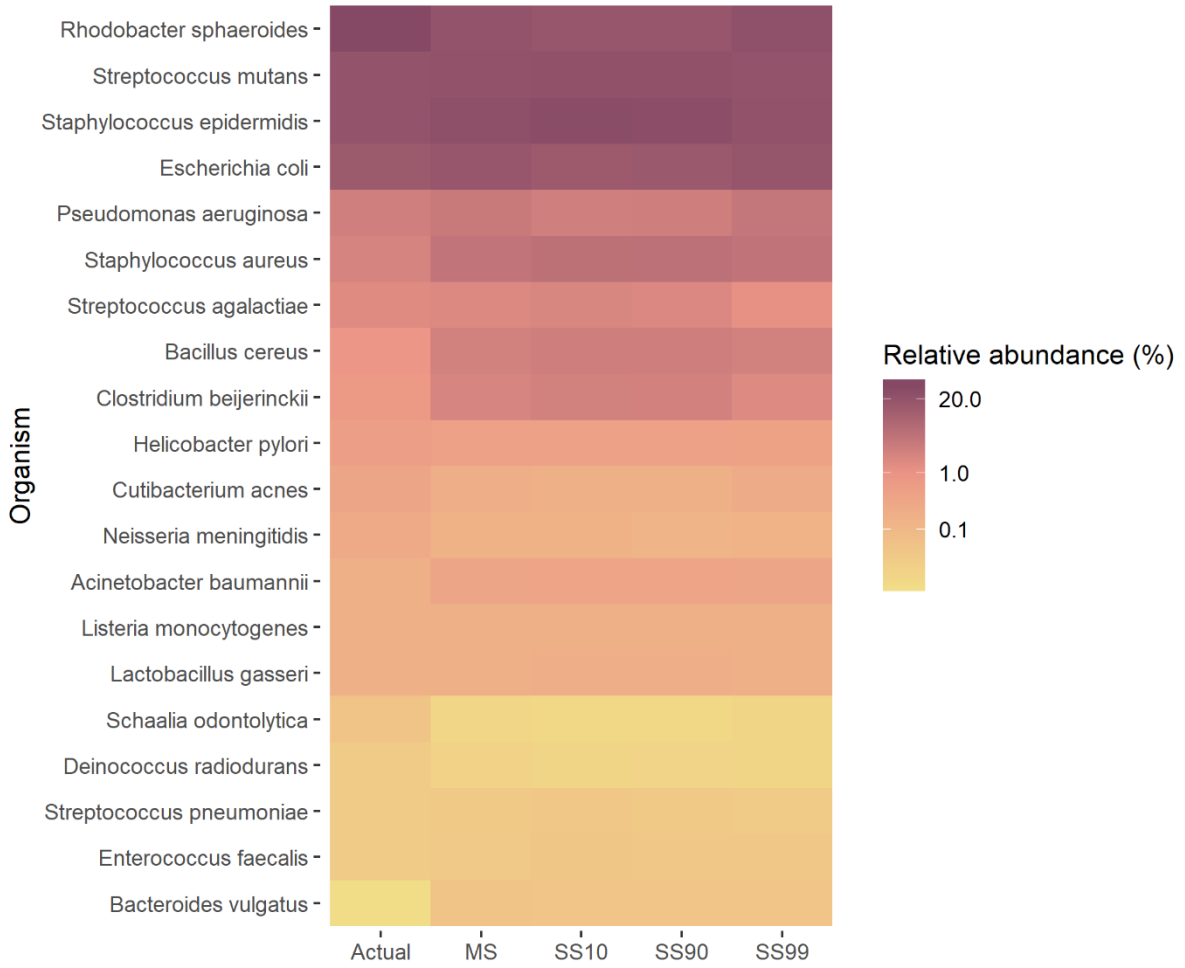

Figure 34 Taxonomic profile of the synthetic metagenome samples determined with Kraken 2, and expressed as relative abundance of species in a heat map. The actual abundances are presented as per the original publication based on the theoretical number of genome copies present. Species are listed from highest to lowest expected relative abundances. MS = microbial sample; SS10 = 10% host DNA; SS90 = 90% host DNA; SS99 = 99% host DNA.

16 of the 20 species, compared to 17 in the published study (Figure 34). Mean squared relative error for MetaPhlAn was 0.3 and for Bracken was 0.45.

Changes in relative abundance due to host DNA abundance were modest, even in SS99 where 12 of 20 organisms were within 10% of the estimate from MS (mean squared relative error 0.02).

We found the association of variation in observed:expected ratio with genome GC content to be similar to the original report (r=-0.74 vs. -0.85; data not shown).

Other species

Using Bracken recalculated reads, off-target genera (n=1 336) could be classified into synthetic-associated (MS:SS99 > 10:1), host-associated (SS99:MS > 10:1) or non-specific. Over 92% of reads were from host- or synthetic-associated genera. Synthetic-associated genera contributed 0.8% of microbial reads in MS, and host-associated genera less than 1:105. Host-associated genera contributed 11.5%

of microbial reads in SS99 (despite being only 0.2% of murine reads), and synthetic-associated genera 1%.

The top four synthetic-associated genera were Shigella, Salmonella, Citrobacter and Klebsiella. These are all likely to represent misclassified E. coli reads. The top four host-associated genera are Pasteurella, Halomonas, Alcanivorax and Mycobacteria. Alcanivorax and Pasteurellaceae have previously been reported to contaminate DNA extraction kits (Glassing et al., 2016). We note that host DNA was extracted in the laboratory whereas the microbial DNA was obtained commercially, thus different contaminants are unsurprising.

The target genera with lowest read counts in SS99 were Schaalia and Deinococcus (36 and 37 reads respectively). Fifty-four off-target genera had 36 or more reads. The most abundant off-target genus (Pasteurella) contributed 11 530 reads, greater than 13 of 17 target genera.

Low microbial biomass

The greater sensitivity of this read binning approach reveals the underlying problem of high relative contamination in the samples with high host DNA content. The problem can now be reframed as one of low (proportionate) microbial biomass and potential mitigations considered.

The challenge of low microbial biomass samples, introduced earlier, has been more extensively studied in rRNA-amplification based approaches than shotgun metagenomics. Nonetheless, many of the problems are shared, and we direct readers to a recent review by Eisenhofer et al. (2019). Pre-analytic mitigations include appropriate controls, as described therein.

Analytic mitigations for 16S rRNA studies were explored in a recent publication (Karstens et al., 2019). The authors investigated filtering based on relative abundance thresholds in negative controls; Decontam (Davis et al., 2018), an approach based on the inverse relationship between relative abundance of contaminants and total microbial DNA, and SourceTracker (Knights et al., 2011) which takes a Bayesian approach using external or internal community references.

In summary, it was found that simple censoring of thresholded negative control OTUs discriminated contaminant and target sequence variants poorly. The Decontam approach discriminated better, correctly classifying all target sequence clusters, and up to 90.4% of contaminant sequence clusters. SourceTracker performed poorly without external references (a typical scenario), identifying less than 1% of contaminant sequence clusters.

Although limited by few samples and no duplicates, we applied Decontam to the Bracken-normalised species counts, using the frequency-based approach. Input DNA concentration was replaced with the total microbial read counts (since all samples had been normalized to 0.2 ng/ml). None of the 20 target

species were classified as contaminants. 2 636 of 4 319 (61%) of off-target species were classified as contaminants, and these accounted for 79% of off-target reads in SS99 and 50% in SS90.

In SS99, the lowest abundant genera, Schaalia and Deinococcus, retained 35 and 34 reads respectively. Only 7 off-target genera had 34 or more reads, comprising the four synthetic-associated genera above, with Cronobacter, Nitrosopumilus and Enterobacter. Shigella had the most reads at 1303, exceeding 10 of 17 target genera.

Interpretation

The marker gene approach employed by MetaPhlAn is very sensitive to read depth, and hence host DNA abundance. In contrast, the read binning approach employed by Kraken 2 detects organisms across the >2,000-fold range of relative abundances even with 99% host DNA content.

Genome-size normalisation of Bracken-estimated read counts provides similarly accurate estimates of relative abundance to MetaPhlAn. The untrimmed reads (not available) may give better results as they would all be of the same length, which is expected by Bracken.

We demonstrate that the large relative contribution of contaminants when microbial reads are in a minority is a greater concern, representing around 10% of microbial reads in SS99 with contaminant genera exceeding the counts of some target genera.

However, the frequency-based Decontam approach allows nearly four-fifths of these off-target reads to be excluded. Further, many of those that remain may represent misclassified target reads.

Concluding remarks

The appropriate selection of analytic tools is vital for accurate and sensitive metagenome analysis. For samples with low microbial biomass, reducing contamination is a priority, though mitigation is possible. Techniques to selectively remove host DNA are required, but thorough benchmarking is awaited.

References

Davis, N.M., Proctor, D.M., Holmes, S.P., Relman, D.A. & Callahan, B.J. (2018) Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*. 6 (1), 226. doi:10.1186/s40168-018-0605-2.

Eisenhofer, R., Minich, J.J., Marotz, C., Cooper, A., Knight, R. & Weyrich, L.S. (2019) Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends in Microbiology*. 27 (2), 105–117. doi:10.1016/j.tim.2018.11.003.

Escobar-Zepeda, A., Vera-Ponce de León, A. & Sanchez-Flores, A. (2015) The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics. *Frontiers in Genetics*. 6. doi:10.3389/fgene.2015.00348.

Feehery, G.R., Yigit, E., Oyola, S.O., Langhorst, B.W., Schmidt, V.T., Stewart, F.J., Dimalanta, E.T., Amaral-Zettler, L.A., Davis, T., Quail, M.A. & Pradhan, S. (2013) A Method for Selectively Enriching Microbial DNA from Contaminating Vertebrate Host DNA. *PLOS ONE*. 8 (10), e76096. doi:10.1371/journal.pone.0076096.

Glassing, A., Dowd, S.E., Galandiuk, S., Davis, B. & Chiodini, R.J. (2016) Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathogens*. 8 (1), 24. doi:10.1186/s13099-016-0103-7.

Hasan, M.R., Rawat, A., Tang, P., Jithesh, P.V., Thomas, E., Tan, R. & Tilley, P. (2016) Depletion of Human DNA in Spiked Clinical Specimens for Improvement of Sensitivity of Pathogen Detection by Next-Generation Sequencing. *Journal of Clinical Microbiology*. 54 (4), 919–927. doi:10.1128/JCM.03050-15.

Karstens, L., Asquith, M., Davin, S., Fair, D., Gregory, W.T., Wolfe, A.J., Braun, J. & McWeeney, S. (2019) Controlling for Contaminants in Low-Biomass 16S rRNA Gene Sequencing Experiments. *mSystems*. 4 (4). doi:10.1128/mSystems.00290-19.

Knights, D., Kuczynski, J., Charlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., Bushman, F.D., Knight, R. & Kelley, S.T. (2011) Bayesian community-wide culture-independent microbial source tracking. *Nature Methods*. 8 (9), 761–763. doi:10.1038/nmeth.1650.

Lindgreen, S., Adair, K.L. & Gardner, P.P. (2016) An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific Reports*. 6, 19233. doi:10.1038/srep19233.

Lu, J., Breitwieser, F.P., Thielen, P. & Salzberg, S.L. (2017) Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science*. 3, e104. doi:10.7717/peerj-cs.104.

Marotz, C.A., Sanders, J.G., Zuniga, C., Zaramela, L.S., Knight, R. & Zengler, K. (2018) Improving saliva shotgun metagenomics by chemical host DNA depletion. *Microbiome*. 6 (1), 42. doi:10.1186/s40168-018-0426-3.

McArdle, A.J. & Kaforou, M. (2020) Sensitivity of shotgun metagenomics to host DNA: abundance estimates depend on bioinformatic tools and contamination is the main issue. *Access Microbiology,*. 2 (4), e000104. doi:10.1099/acmi.0.000104.

Nelson, M.T., Pope, C.E., Marsh, R.L., Wolter, D.J., Weiss, E.J., Hager, K.R., Vo, A.T., Brittnacher, M.J., Radey, M.C., Hayden, H.S., Eng, A., Miller, S.I., Borenstein, E. & Hoffman, L.R. (2019) Human and Extracellular DNA Depletion for Metagenomic Analysis of Complex Clinical Infection Samples Yields Optimized Viable Microbiome Profiles. *Cell Reports*. 26 (8), 2227-2240.e5. doi:10.1016/j.celrep.2019.01.091.

Oechslin, C.P., Lenz, N., Liechti, N., Ryter, S., Agyeman, P., Bruggmann, R., Leib, S.L. & Beuret, C.M. (2018) Limited Correlation of Shotgun Metagenomics Following Host Depletion and Routine Diagnostics for Viruses and Bacteria in Low Concentrated Surrogate and Clinical Samples. *Frontiers in Cellular and Infection Microbiology*. 8. doi:10.3389/fcimb.2018.00375.

Pereira-Marques, J., Hout, A., Ferreira, R.M., Weber, M., Pinto-Ribeiro, I., van Doorn, L.-J., Knetsch, C.W. & Figueiredo, C. (2019) Impact of Host DNA and Sequencing Depth on the Taxonomic Resolution of Whole Metagenome Sequencing for Microbiome Analysis. *Frontiers in Microbiology*. 10. doi:10.3389/fmicb.2019.01277.

Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J. & Segata, N. (2017) Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*. 35 (9), 833–844. doi:10.1038/nbt.3935.

Ranjan, R., Rani, A., Metwally, A., McGee, H.S. & Perkins, D.L. (2016) Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and Biophysical Research Communications*. 469 (4), 967–977. doi:10.1016/j.bbrc.2015.12.083.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O. & Huttenhower, C. (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*. 9 (8), 811–814. doi:10.1038/nmeth.2066.

Weyrich, L.S., Farrer, A.G., Eisenhofer, R., Arriola, L.A., Young, J., Selway, C.A., Handsley-Davis, M., Adler, C.J., Breen, J. & Cooper, A. (2019) Laboratory contamination over time during low-biomass sample analysis. *Molecular Ecology Resources*. 19 (4), 982–996. doi:10.1111/1755-0998.13011.

Wood, D.E. & Salzberg, S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*. 15 (3), R46. doi:10.1186/gb-2014-15-3-r46.

Ye, S.H., Siddle, K.J., Park, D.J. & Sabeti, P.C. (2019) Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell*. 178 (4), 779–794. doi:10.1016/j.cell.2019.07.010.

Appendix C – Tools and databases

I compile below lists of software tools and databases used in the analyses presented.

Tools

Tool	Version	Reference
R	Multiple	
Nextflow	20.01.0.5264	Di Tommaso et al., 2017
Anaconda	4.10.1	
Kraken 2	2.1.0 (compiled)	Wood & Salzberg, 2014; Wood et al., 2019
Bracken 2	2.5.3 (compiled)	Lu et al., 2017
Sourmash	4.2.1 (anaconda)	Brown & Irber, 2016
SAMtools	1.2 (HPC pre-	Li et al., 2009
MMseqs2	13.45111	Steinegger & Söding, 2017
MEGAHIT	1.2.9 (anaconda)	Li et al., 2015b
MetaBAT 2	2.15-	Kang et al., 2015
CheckM	1.1.2 (anaconda)	Parks et al., 2015
Prodigal	2.6.3 (anaconda)	Hyatt et al., 2010
Jellyfish	2.3.0	Marçais & Kingsford, 2011
Flextaxd	0.2.4 (anaconda)	
Newick Utils	1.6 (anaconda)	
Nonpareil	3.3.3 (anaconda)	Rodriguez-R et al., 2018
VCFLib	1.0.2 (anaconda)	
NCBI Genome	0.3.0 (anaconda)	Blin, 2022
MaxQuant	1.6.6.10,	Tyanova, Temu & Cox, 2016
Quandenser	v0.02 (custom)	The & Käll, 2020
Msconvert	3.0.10730	Adusumilli & Mallick, 2017
SearchGUI	4.0.0-beta	Barsnes & Vaudel, 2018
MetaNovo	1.6	Potgieter et al., 2019
Fastv	0.9.0	Chen et al., 2021

Databases

Name	Source	Version or date downloaded
Genome Taxonomy Database Taxonomy and genome files	https://data.gtdb.ecogenomic.org/releases/release95/95.0/	Release 95
Sourmash Genome Taxonomy Database LCA databases	https://osf.io/wxf9z/	Release 202
NCBI RefSeq viral, fungal and human reference genomes	Via NCBI Genome Download tool	8 October 2020
UniProt reference proteomes	Using get_uniprot.pl script of MMH (Pombert, 2021)	March 2019
abYsis database	Prof Andrew Martin, personal communication	6 March 2019