3

Development of Metagenomic Sequencing Workflow

This chapter describes the development of an enrichment-based metagenomic sequencing workflow for application to plasma samples from individuals with sepsis due to community acquired pneumonia.

3.1 Introduction

We anticipated two main challenges associated with metagenomic sequencing from sepsis plasma samples. Firstly, sensitivity for detecting microbial nucleic acids would be low since human nucleic acids would be present in far greater excess, especially with antimicrobial use prior to sample collection. Secondly, the design of a single library preparation method suitable for sequencing both DNA and RNA when most examples in the literature target only one type of nucleic acid. These challenges were addressed through the application of probe-based enrichment and a combined DNA and RNA library preparation workflow, respectively and we termed this probe-based targeted enrichment method *Castanet*.

3.1.1 Probe-based enrichment

Our starting point for targeted metagenomics using oligonucleotide enrichment probes was work done (Bonsall et al. 2015) by the Stratified Medicine to Optimise the Treatment of Patients with Hepatitis C Virus Infection (STOP-HCV) consortium (led by Professor Ellie Barnes, University of Oxford). The genome of Hepatitis C virus (HCV) is highly diverse with strains subdivided into seven genotypes which differ at approximately 30-35% positions across the 9650 nt genome. This makes work on HCV highly applicable to the range of diverse bacteria and viruses we would need to target in CAP.

Bonsall and colleagues observed greater than 10^3 fold enrichment in mid-range viral load samples (10^4-10^5 IU/ml). This degree of enrichment enabled an increased depth of sequencing to be achieved, enabling more affordable sequencing through multiplexing of larger numbers of samples for the same amount of sequencing capacity.

In addition, the STOP-HCV group were able to exploit a phenomenon they observed, that a 20% divergence between probe and target was tolerated with minimal reduction in enrichment efficiency. They started with a probe panel covering 4 genotypes and augmented it to improve coverage for the four already-included subtypes and six additional subtypes. To do this, a consensus sequence for each HCV subtype was generated. Then, for each whole genome sequence in a reference set, genomic regions with less than 80% identity to a starting panel of probes were identified. For each of these regions, the subtype consensus sequence was considered as a reference if there was $\geq$ 80% similarity. If not, a probe was added to cover that genomic region. In this way, a cost-effective, non-redundant set of probes was generated.

3.1.2 Ribosomal multilocus sequence typing

The ribosomal multilocus sequence typing (rMLST) scheme (Jolley et al. 2012) for combined taxonomy and typing in bacteria is a development from 16S rRNA gene approaches (Woese 1987). This approach indexes variation of the 53 bacterial ribosome protein subunit (*rps*) genes to enable resolution into groups at all taxonomic and most typing levels. This is possible as the *rps* genes are conserved enough to enable taxonomic organisation and yet sufficiently diverse to enable species and type characterisation. As of 23 July 2019, the rMLST database (http://pubmlst.org/rmlst) contained 304,876 genomes and 1,861,484 alleles.

The rMLST scheme can be applied to the generation of an efficient panel of probes which enable species/type characterisation without coverage of the whole bacterial genome. In addition, the targeting of conserved regions means that probes are able to capture bacterial organisms despite intra-species genomic variation.

3.1.3 Library preparation methods for metagenomics

There are few examples in the literature of a single combined library preparation workflow suitable for the metagenomic sequencing of both DNA and RNA. A 2018 review (Forbes et al. 2018) details 65 peer-reviewed studies of diagnostic clinical metagenomics. The majority of studies describe sequencing-based techniques applied only to DNA-based or RNA-based genomes with a few examples describing parallel RNA and DNA workflows (Langelier et al. 2018), (Salzberg et al. 2016). Only one small study (n=6) (Doan et al. 2016) successfully sequenced reads from DNA and RNA viruses (as well as fungi and parasites) from ocular fluid using a single protocol. This involved nucleic acid extraction, cDNA synthesis, and subsequent processing using the Illumina Nextera DNA Library Prep kit.

Other examples of combined library preparation methods include that described in the VirCapSeq-VERT method (Briese et al. 2015) which involved nucleic acid extraction, cDNA synthesis, fragmentation using ultrasonication, and subsequent processing using the KAPA library prepration kit. However, the authors did not trial this method on actual clinical specimens.

3.1.4 Viral multiplex reference

We evaluated our workflow using a Viral Multiplex Reference (VMR) control (11/242) available through the UK National Institute for Biological Standards and Control (NIBSC). The reagent contains 25 infectious viruses covering a range of genome types (dsDNA, dsRNA, ssRNA+, ssRNA-), sizes (6.8-233.7 kb), envelope types and pre-assayed concentrations (Table B.3).

Mee and colleagues (Mee et al. 2016) document a study involving 15 laboratories who were invited to process the VMR control using their own wet-lab and informatics methods. In this study, 6/25 target viruses were detected by all laboratories and two laboratories detected all 25 viruses. We will compare the performance of *Castanet* against these 15 laboratories.

3.1.5 Aims

1. To use probe-based enrichment to increase sensitivity for sequencing organisms relevant to sepsis from CAP
2. To optimise a library preparation method suitable for sequencing both DNA and RNA-based organisms from plasma
3. To evaluate performance of the library preparation with enrichment against a known positive control reference set

3.2 Results

3.2.1 Probe panel development

Development of the probe panel was performed in collaboration with the Childhood Meningitis and Encephalitis Study (ChiMES) group. We compiled a list of viral and bacterial pathogens relevant to paediatric meningitis and adult sepsis from CAP in the UK (Table B.1; Table B.2). We also included several pathogens of current interest during the time of probe set development (Zika virus, Chikungunya virus). Considering the number of distinct entries on our list (116, from 17 virus families and 35 bacterial species) and the criteria for inclusion, we inferred that any omissions of a priori less likely organisms, including relevant fungal or parasite pathogens, would comprise rare (<1% frequency) or novel and therefore unsuspected causes of meningitis or pneumonia and sepsis. We also included probes to 4 spike-in control sequences for methodological evaluation of *Castanet*.

We targeted similar lengths of genomic sequence for each pathogen to achieve a comparable assay sensitivity, optimising the breadth of pathogens we could target and avoiding bias in favour of larger genomes. For each of the viruses, we downloaded from NCBI RefSeq the full set of complete genomes available at 1st August 2015. We constructed genome alignments using MAFFT from which to design the probes. For each of the included herpesviruses, whose genomes exceed 100 kbp, this involved a low-diversity region of 20kb whilst for all other viruses we used the whole genome. For bacterial species, we took advantage of the ribosomal multilocus sequence typing (rMLST) scheme, which targets 53 genes encoding ribosomal proteins present in all bacteria and resolves bacteria to a sub-species level, extracting relevant sequences from the rMLST database on 11 December 2015.

Probe design was carried out by Dr Azim Ansari. In previous work with HCV, it had been observed that sequence capture efficiency is preserved when probe and target sequences differ at up to 20% of positions, and that exploiting sequence similarity to avoid redundancy can make probe design substantially more efficient without sacrificing performance. Accordingly, for each sequence alignment we constructed a tree using pairwise distances, within which we identified clusters such that all sequences were less than 5% divergent from one another. The 5.86 x 10^6 bases of cluster consensus sequences were used to design a panel of 52,101 Agilent SureSelect, 120 nucleotide RNA probes on the complementary strand.

3.2.2 Evaluation of four library prepration methods

We chose to perform this initial library preparation (library prep) development on plasma from patients infected with Hepatitis C Virus due to a previously established workflow with this sample type within the STOP-HCV consortium.

Four different library prep methods were compared in five HCV samples: (i) RNA; (ii) DNA; (iii) Combined with Fragmentation (CF. RNA method followed by DNA method); and (iv) Combined with no Fragmentation (CnoF; DNA method preceded by reverse transcription with random primers). To assess the suitability of each method for DNA and RNA-based pathogens, samples were spiked with an RNA (External RNA Controls Consortium Spike-In Mix, ERCC) and DNA positive control (dsDNA plasmid fragments). Following total nucleic acid extraction, the DNA and RNA content of each sample was assayed (Agilent 2100 Bioanalyser platform). This enabled us to spike in the plasmid DNA and ERCC RNA and 3% and 1% concentration by mass respectively. There was no enrichment stage in these experiments.

Sequencing yielded a mean total read count of 1.5x10^6 (range 0.9-2.6x10^6) across the five HCV samples and four library preps with the majority of sequences aligning to the human reference genome (mean 95.6%; range 91.5-99.1%). Libraries following RNA prep were significantly lower in cDNA concentration, requiring volumes for equimolar pooling in excess of the other three methods by an average of 16-fold. Unsurprisingly, following the DNA prep method, no reads aligning to RNA sequences (HCV or ERCC) were identified in any sample (Figure 3.1). These observations indicate that neither the RNA nor DNA prep in isolation would be suitable for the sequencing of sepsis samples. Thus, the remainder of this chapter concentrates on comparing the two combined library preps.

HCV yield was highest in the CnoF prep, superior even to the standard RNA library prep method used by the STOP-HCV group (Figure 3.1c and d). The CnoF prep also yielded a higher percentage of reads mapping to ERCC than the other three methods (Figure 3.1a). However, the CF prep was superior to the CnoF prep in yield of DNA plasmid reads (Figure 3.1b).

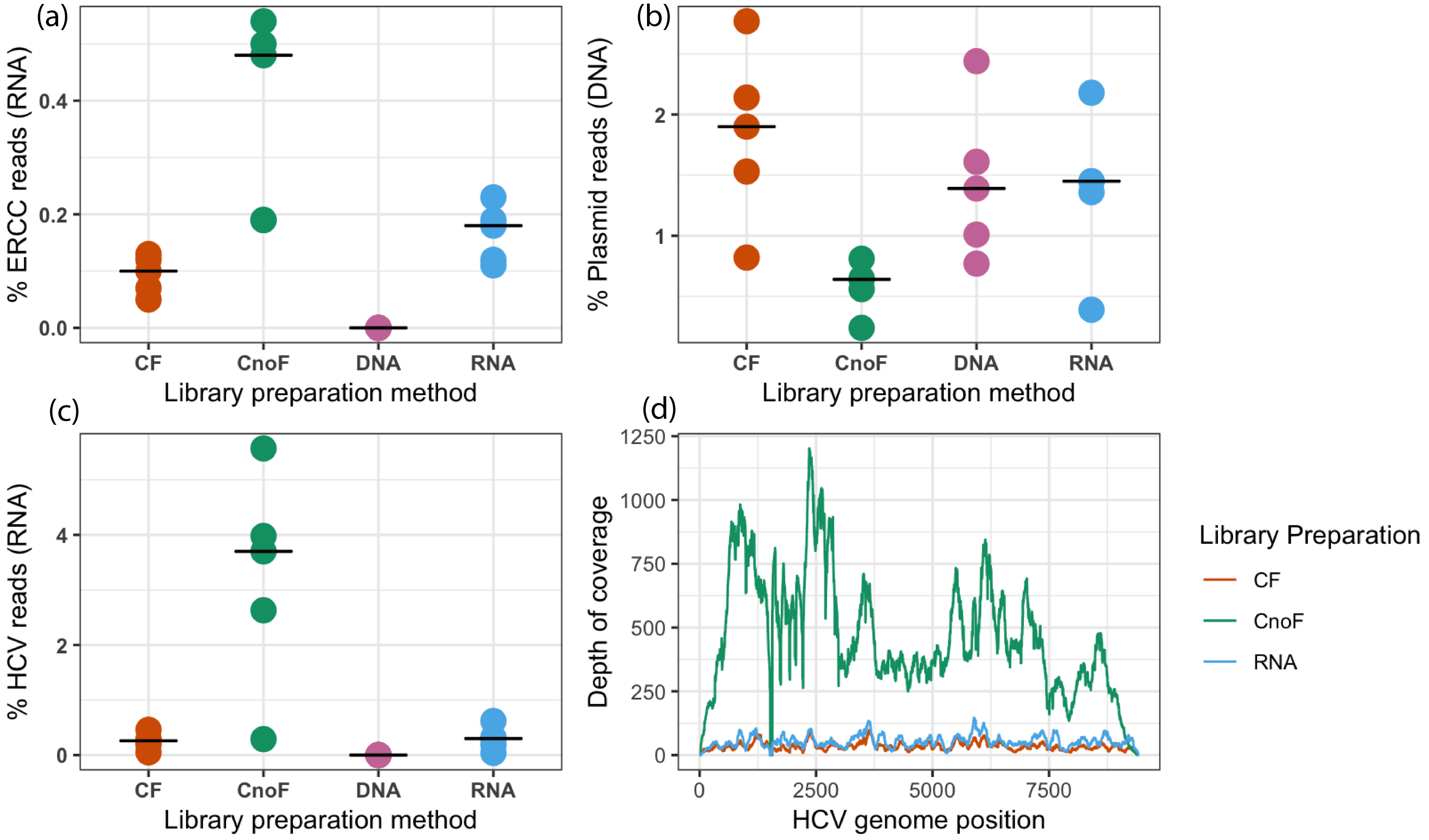

Figure 3.1: Comparison of four library preparation methods Performance of the different library preparation methods was evaluated with regards to: (a) ERCC RNA yield; (b) Plasmid DNA yield; (c) HCV RNA yield; (d) HCV genome coverage. (ERCC=External RNA Controls Consortium Spike-in Mix; CF=Combined with Fragmentation; CnoF=Combined with no Fragmentation; HCV=Hepatitis C Virus

For each sample, we calculated the input DNA:RNA spike-in ratio and compared this against the ratio of DNA:RNA (plasmid:ERCC) reads recovered (Table 3.1). We observed that the yield of DNA:RNA reads closely matched the input spike-in concentrations of DNA:RNA with the CnoF prep, indicating that this method method was similarly efficient for recovery of both DNA and RNA when compared to the CF prep.

		HCV146	HCV371	HCV953	HCV958	HCV972
Input (Plasmid:ERCC mass)	Both	1.5	5.9	1.4	1.8	1.8
Output (Plasmid:ERCC reads)	CF CnoF	12.0 1.1	53.6 4.2	11.0 1.3	21.3 1.3	16.4 1.2

Table 3.1: Sequencing yield of plasmid:ERCC reads relative to plasmid:ERCC spike-in mass. The combined with no fragmentation (CnoF) and combined with fragmentation (CF) library preps are compared.

3.2.3 Evaluation of the Combined no Fragmentation protocol

A subsequent experiment was performed to further evaluate the CnoF library prep protocol (Figure 3.2), with the following specific aims: (i) to compare relative yield of reads originating from RNA and DNA; (ii) to evaluate the relationship of sequencing yield with read length and concentration for RNA and DNA; and (iii) to trial the CnoF protocol in sepsis patients.

Sequencing yielded a mean total read count of 24.8x10^6 with minimal variation between HCV and sepsis samples (range 22.7-28.3x10^6) (Table 3.2).

	HCV371	HCV958	Sepsis4	Sepsis6
Total reads (x10^6)	23.9	22.7	25.7	25.1
% Human	95.2	96.4	96.4	99.1
% HCV	1.98	0.44	0	0
% ERCC	0.23	0.27	0.20	0.18
% Plasmid	0.38	0.50	0.23	0.19

Table 3.2: Percentage of reads aligning to each of the relevant reference genomes. Results are for samples with dual ERCC and plasmid spike-ins

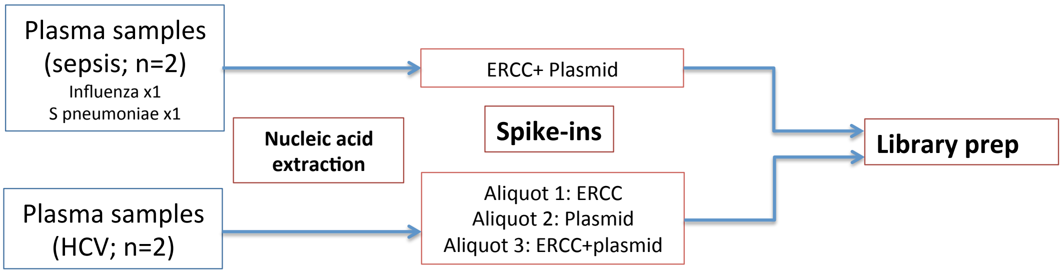

Figure 3.2: Workflow for CnoF Evaluation Experiment. Each HCV sample was divided into three aliquots following nucleic acid extraction and spiked with (i) ERCC only, (ii) plasmid only, or (iii) both ERCC and plasmid. Each sepsis sample was spiked with both ERCC and plasmid.

The proportion of reads mapping to the ERCC and plasmid references remained consistent whether the controls were spiked in individually or in combination, confirming that RNA was not impacting recover of DNA or vice versa (Figure 3.3).

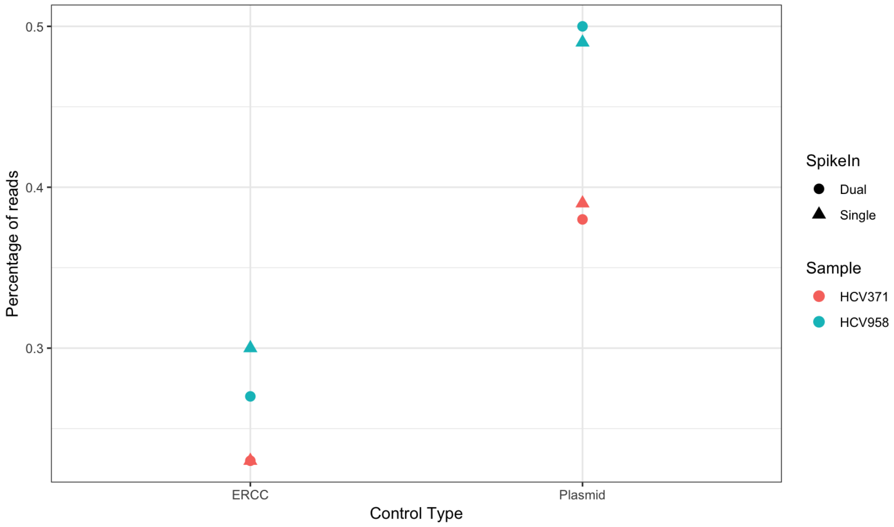

Figure 3.3: Sequencing yield of controls Yield of ERCC and plasmid compared between aliquots receiving single and dual spike-ins

The plasmid and ERCC controls were spiked-in at 1% by mass of the initial sample DNA and RNA concentrations respectively. Thus, the relative mass ratio of plasmid to ERCC spike-ins varied between the samples. However, the ratio of plasmid to ERCC reads yielded reflected the input mass ratio fairly consistently

Sample	Input ratio (mass)	Output ratio (reads)
HCV371	1.94	1.65
HCV958	2.40	1.86
Sepsis4	1.12	1.15
Sepsis16	0.70	1.03

Table 3.3: Sequencing yield of plasmid:ERCC (reads) relative to spike-in input (mass)

between samples, indicating minimal bias towards RNA or DNA (Table 3.3).

There was no association between sequencing yield and fragment size for either plasmid or ERCC (Figure 3.4), indicating that the CnoF protocol performs consistently across the range of fragment sizes studied (plasmid 379-3190bp; ERCC 250-2000nt).

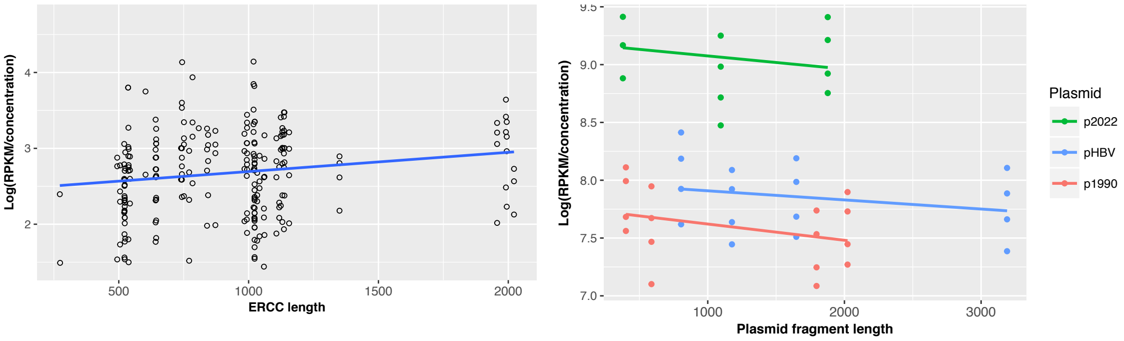

Figure 3.4: Relationship between read yield and fragment length. (a) ERCC, (b) Plasmid. The y-axis displays read count normalised for fragment length, total read count per sample, and fragment concentration. RPKM = reads per kilobase of fragment per million reads.

For the ERCCs, sequencing yield was proportional to input concentration (Figure 3.5). This association was less clear for the plasmid spike-ins. Although sequencing yield was highest for the p1990 plasmid (which was spiked-in at the highest concentration), the yield of p2022 was higher than that of pHBV despite a higher spike-in concentration of pHBV relative to p2022 (Figure 3.5). The differences observed between ERCC and plasmid probably reflect inaccuracies in plasmid nucleic acid quantification rather than differences between RNA and DNA processing in the library prep method.

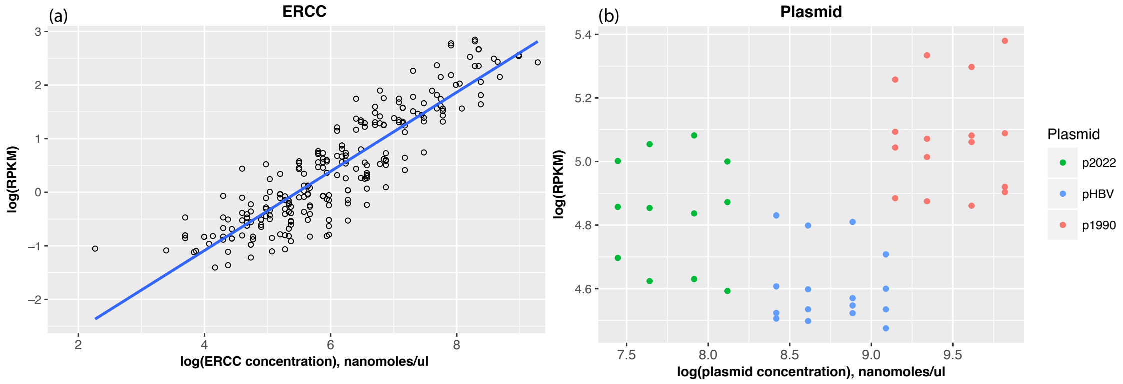

Figure 3.5: Relationship between read yield and input concentration (molarity). (a) ERCC, (b) Plasmid. RPKM = reads per kilobase of fragment per million reads.

Finally, we did not observe significant differences in total read count or recovery of spike-ins between plasma from sepsis patients compared to HCV patients.

3.2.4 Evaluation of performance using a Viral Multiplex Reference control

We combined four 1ml aliquots of reagent and made two replicates of a series of five dilutions (neat, 1:10, 1:100, 1:500, 1:1000) in phosphate-buffered saline solution, forming 500ul aliquots for extraction and library preparation.

We used dilutions of a commercially available mixture of viruses (NIBSC Viral Multiplex Reference 11/242) to assess the quantitative range of detection of our method. For two undiluted VMR replicates, *Castanet* sequencing yielded 9.1 and 10.9x10^7 reads. We detected all 21 viruses for which we had enrichment probes, with at least 8.65 reads per million in enriched samples, as well as two viruses not included (Sapovirus and Astrovirus; Figure 3.6). Norovirus GI and GII were not detected; we did not have probes to capture this virus. A likely reason for this difference is that Sapovirus and Astrovirus were present in high enough concentrations to be sequenced without enrichment whilst Norovirus was present in lower concentrations. This is consistent with the 80% failure rate in sequencing one or both Norovirus species in an evaluation of 15 laboratories with the same VMR control.

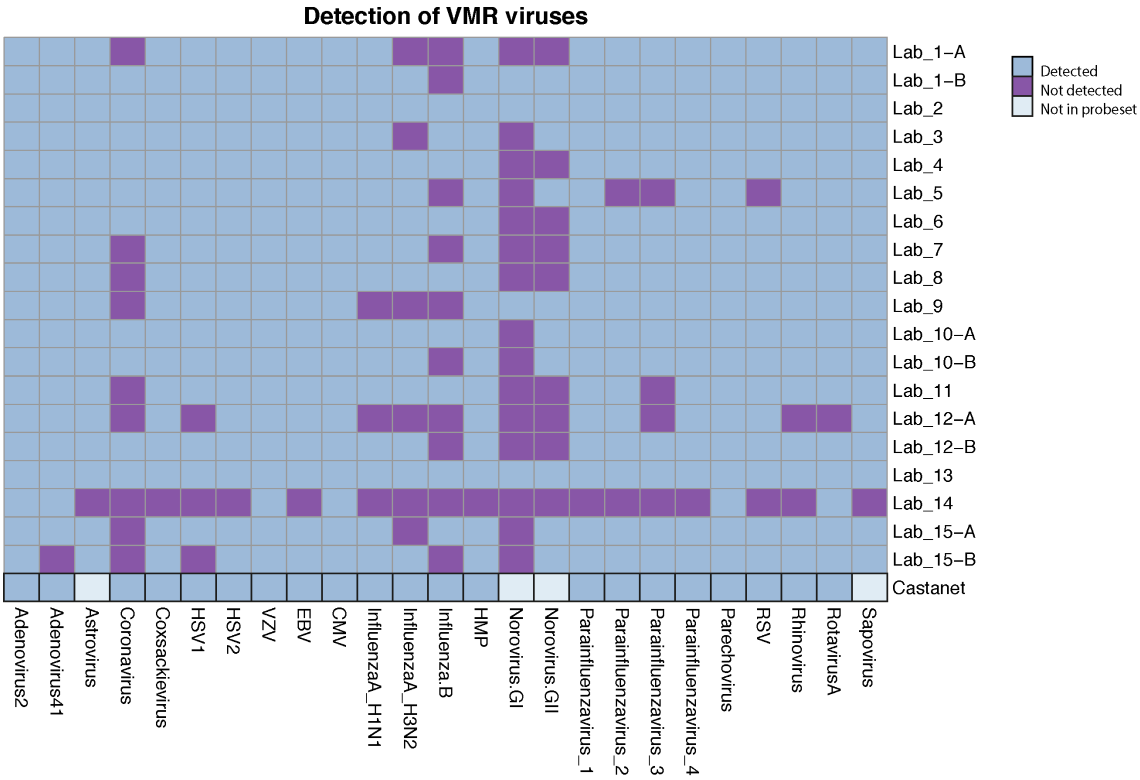

Figure 3.6: Comparison of Castanet with 15 other laboratories for sequencing of viruses in the NIBSC Viral Multiplex Reference 11/242. Mee and colleagues document the performance of 15 laboratories in sequencing the 25 viruses in the VMR. HSV=herpes simplex virus; VZV=varicella zoster virus; EBV=Epstein-Barr virus; CMV=cytomegalovirus; HMP=human metapneumovirus; RSV=respiratory syncytial virus

For individual microorganisms, we observed a linear relationship between organism load and sequencing yield. The VMR included five viruses where viral load had been quantified by the NIBSC using qPCR. For each virus, the number of deduplicated reads was proportional to input concentration across the dilution series (neat, 1 in 10, 1 in 100, 1 in 500, 1 in 1000) (Figure 3.7). However, the relationship between input viral load and yield of deduplicated reads differed between viruses. We observed a 10^2-10^3-fold enrichment for the five quantified viruses (Figure 3.8).

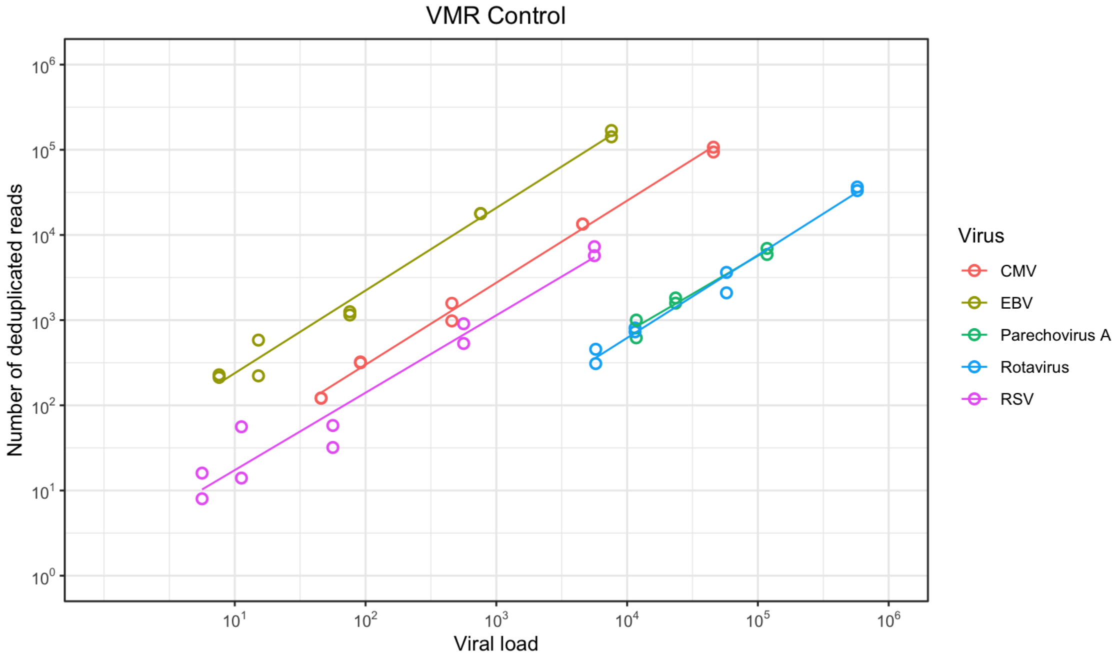

Figure 3.7: Relationship between viral load and sequencing yield in Viral Multiplex Reference (VMR) samples. The VMR was sequenced at a range of dilutions in two replicates. For the five viruses in the VMR that had been quantified by the NIBSC using qPCR, the relationship between viral load and sequencing yield is plotted. (CMV=cytomegalovirus; EBV=Epstein-Barr Virus; RSV=Respiratory Syncytial Virus)

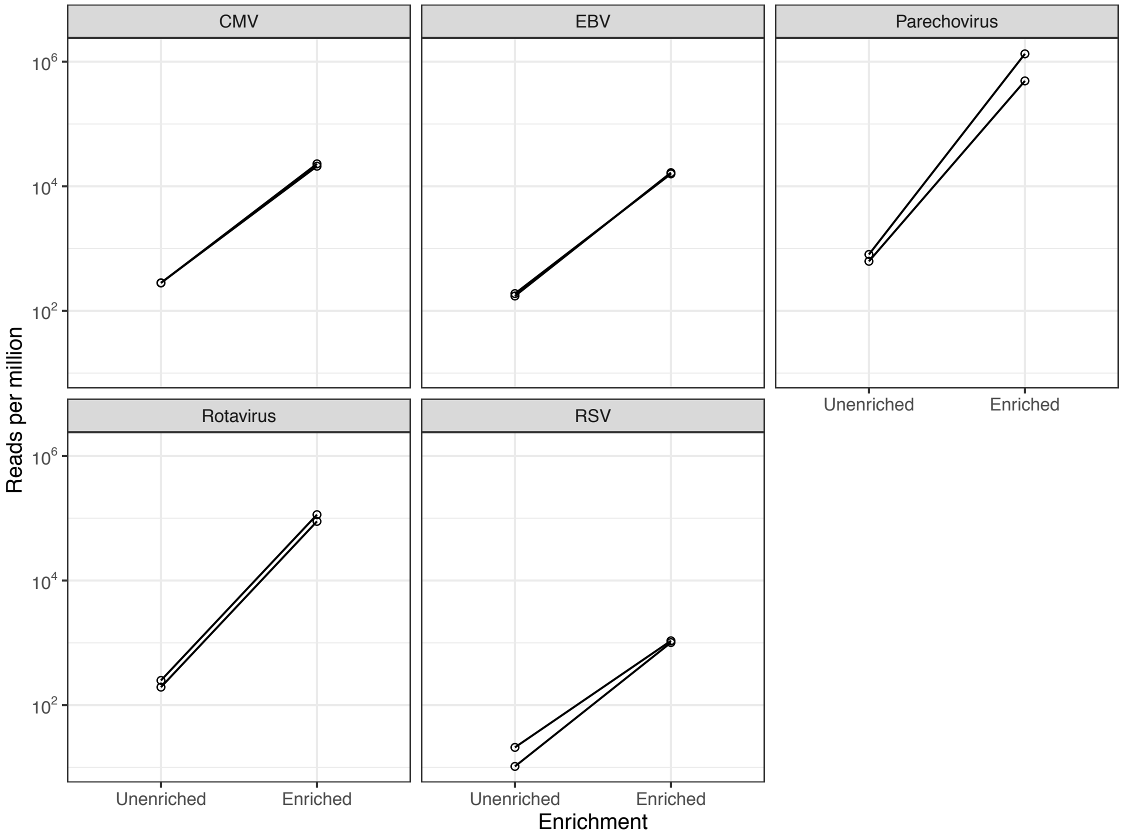

Figure 3.8: Increase in sequencing yield with enrichment The VMR was sequenced in two replicates at undiluted concentration. The sequencing yield (reads per million total reads) is plotted for each of the five viruses quantified by the NIBSC using qPCR. (CMV=cytomegalovirus; EBV=Epstein-Barr Virus; RSV=Respiratory Syncytial Virus)

3.2.5 Data processing

A data processing pipeline (Figure 3.9) was developed in collaboration with Dr Tanya Golubchik from the ChiMES project.

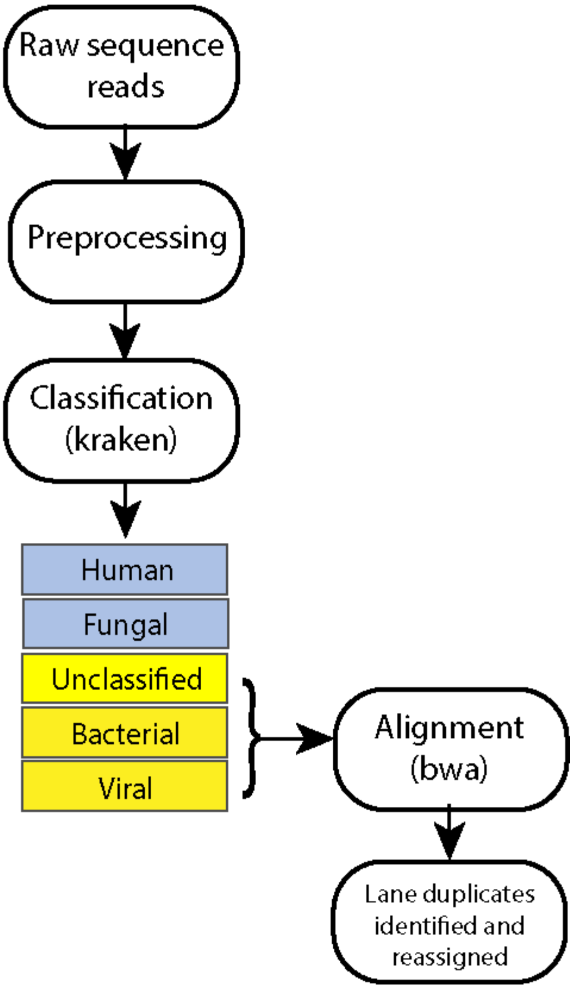

Figure 3.9: Data processing pipeline

De-multiplexed sequence read-pairs were trimmed of adapter sequences using Trimmomatic v0.36, with the ILLUMINACLIP options set to 2:10:7:1:true MINLEN:50, using the set of Illumina adapters supplied with the software (Bolger et al. 2014). The trimmed reads were then classified using Kraken v1 (Wood and Salzberg 2014) using a custom database containing the human genome (GRCh38 build), all RefSeq bacterial and viral genomes, and a selection of fungal genomes that were most likely to be associated with cases of meningitis (Cuomo 2017). These were: *Aspergillus fumigatus*, *Candida* spp., *Coccidioides* spp., *Cryptococcus* spp., *Histoplasma capsulatum*, *Paracoccidioides brasiliensis*, and *Pneumocystis* spp. Reads identified as bacterial or viral were aligned using BWA v0.7.1243 with default settings to a multi-fasta reference of consensus sequences corresponding to the enrichment probe targets, augmented with sequences of known or suspected contaminants. These included (i) reagent contaminants (*Alteromonas* and *Achromobacter* spp.), (ii) genomes of two viruses known to have been sequenced on the same flow cell: MVMPCG spike-in control and Echovirus 7; and (iii) the rMLST sequences of commensal Streptococcus species (*S. mitis, S. oralis, S. pseudopneumoniae*) that were thought to be likely contaminants in clinical samples.

Following alignment, we corrected our sequencing results for index misassignment (index hopping). This well-recognised phenomenon occurs when a small proportion of reads belonging to a sample gets misassigned to a different index on a nearby optical cluster. For each sequencing pool, we identified PCR-duplicated reads and reassigned all reads in each duplicate cluster to the sample with the highest number of reads in that cluster.

After duplicate reassignment, we calculated a set of descriptive statistics for each sample and target organism. These included sequencing depth with and without deduplication, and coverage of target sequences at various depth thresholds. The collected statistics were combined with available laboratory data and ddPCR results where available and the resulting data frame used to train a random forest model.

3.3 Discussion

In this chapter, I have described *Castanet*, a versatile probe-based enrichment sequencing method that combines the analysis of RNA and DNA from the same starting material in a single protocol and enriches for pathogens of interest using a modestly sized panel of probes.

3.3.1 Combined library preparation method

We developed a single library preparation workflow which combines separate RNA and DNA workflows in a single streamlined protocol, enabling the successful sequencing of both spike-in controls (ERCC and plasmid) as well as the VMR control. The experiments involving the spike-in controls were performed without enrichment whilst those involving the VMR control were performed with enrichment.

The CnoF protocol enabled sequencing of targets originating from RNA and DNA without bias, as evidenced by the yield of DNA:RNA closely matching input ratios of DNA:RNA for the plasmid (DNA) and ERCC (RNA) spike-in controls. This lack of bias is particularly important for detecting cases of co-infection in CAP, e.g. co-infection with DNA-based *Streptococcus pneumoniae* and RNA-based influenza A virus.

We also observed that there was no association between sequencing yield and fragment size for either plasmid or ERCC. This is important as it demonstrates a lack of bias towards particular fragment sizes, which could in theory preferentially favour the sequencing of one organism over another. However, the analysis is limited to the range of fragment sizes studied (plasmid 379-3190bp; ERCC 250-2000nt).

For the five quantified viruses in the VMR, we observed a linear relationship between input concentration and sequencing yield. However, we noted that this relationship differed between viruses, presumably because of differences in the enriched (genomic) sequence length, the efficiency of sequencing library formation and capture and, perhaps the calibration of qPCR assays. Nevertheless, our results indicate that for a particular organism, deduplicated read counts can be compared between samples to provide information about relative organism loads.

3.3.2 Enrichment for targets of interest

We developed a probe panel covering 116 organisms from 17 virus families and 35 bacterial species. Considering the extent of coverage, the size of the probe panel was modest at 5.86 x 10^6 bases, with minimal redundancy and associated cost savings.

To my knowledge, this is the first published example of a probe panel that includes both bacteria and viruses as well as one targeting specific diseases. Other examples of probe panels include the separate bacterial BacCapSeq (Allicock et al. 2018) and viral VirCapSeq-Vert (Briese et al. 2015) panels which target all human pathogenic bacteria and vertebrate viruses respectively.

As our probe panel was designed based on conserved *rps* genes of the bacterial genome, there are several limitations. Firstly, we are unable to distinguish between different species of the Enterobacteriaceae family because of sequence homology within this region. Secondly, we did not enrich for regions encoding for virulence genes and antimicrobial resistance genes. Although we did not set out with the latter aim, future iterations of the probe panel would benefit from extending beyond the rMLST system.

We observed between a 10^2 to 10^3-fold enrichment for the five quantified VMR viruses. The two higher viral load viruses (Parechovirus, Rotavirus) showed higher fold-change enrichment (10^3-fold vs 10^2-fold) compared to the lower viral load viruses (CMV, EBV, RSV). This is in keeping with the observations of (Bonsall et al. 2015) who noted a 10^3 fold enrichment for mid-range viral load samples but lower fold-change enrichment for lower viral loads.

3.3.3 Data processing pipeline

Our aim with the data processing was to implement a computationally efficient pipeline that would enable accurate alignment of microbial sequences. One challenge was the low signal to noise ratio, with the majority of samples containing >95% human reads. We dealt with this by using kraken for classification prior to bwa alignment. By building a custom database with human and microbial reference sequences, we could identify the human sequences and discard them before BWA alignment, making the latter stage substantially faster.

Another challenge was contaminants, including those from kit reagents, patient skin/mucosa, and simultaneously sequenced samples. We dealt with this by adding these sequences to the multi-fasta file for alignment so that the contaminants would map to the correct references rather than mis-mapping to a closely related pathogenic organism.

One area for future work is the diagnosis of fungal infections. Currently, we do not enrich for fungal reads and any sequences classified as fungal by kraken are discarded prior to BWA alignment. Fungal causes of CAP are currently underappreciated (Chen et al. 2001) and may represent a proportion of the cases which remain diagnosed after routine bacteriology/virology.

3.4 Conclusions

In this chapter, I have described *Castanet*, a targeted metagenomic approach using enrichment probes for the sequencing of DNA and RNA-based bacteria and viruses from patient samples. Here, I have evaluated *Castanet* in terms of its performance for sequencing positive controls (plasmid and ERCC spike-ins; VMR control). In the next chapter, *Castanet* is applied to a cohort of patient samples and further evaluated.

B

Application of Metagenomic Sequencing to Sepsis Samples

Virus family	Virus species
Adenoviridae	Human adenovirus
Arenaviridae	Lassa virus Lymphocytic choriomeningitis virus
Coronaviridae	Human coronavirus HKU1, NL63, OC43, 229E Middle East respiratory syndrome coronavirus Severe acute respiratory syndrome coronavirus
Flaviviridae	Dengue virus Japanese encephalitis virus Murray Valley encephalitis virus St Louis encephalitis virus Tick-borne encephalitis virus West Nile virus Yellow fever virus Zika virus
Herpesviridae	Human herpesvirus 3 (Varicella zoster virus) Human herpesvirus 4 (Epstein Barrvirus) Human herpesvirus 5 (Cytomegalovirus) Human herpesvirus 6-7 (Roseolovirus) Human herpesvirus 8 (Kaposi's sarcoma-associated herpesvirus) Herpes simplex virus 1-2
Orthomyxoviridae	Influenza virus A-C
Paramyxoviridae	Hendra henipavirus Human metapneumovirus Humanparainfluenzavirus1-5 Measles morbilivirus Mumps rubulavirus Nipah henipavirus Respiratory syncytial virus Sosugavirus
Parvoviridae	Human bocavirus Human parvovirus B19 Human parvovirus 4 Primate erythroparvovirus 1 Primate tetraparvovirus 1
Peribunyaviridae	California encephalitis virus
Phenuiviridae	Rift valley fever virus Sandfly fever Naples virus Sandfly fever Sicilian virus
Picornaviridae	Cardiovirus A-B Coxsackie A virus ECHO virus Enterovirus A, B, D Hepatitis A virus Parechovirus A-B Rhinovirus A-C Rosavirus A Salivivirus
Polyomaviridae	BK virus JC polyomavirus
Reoviridae	Rotavirus A-C
Rhabdovirus	Australian bat lyssavirus Duvenhage lyssavirus European bat lyssavirus 1-2 Lagos bat lyssavirus Mokola lyssavirus Rabies lyssavirus
Togaviridae	Chikungunya virus Eastern equine encephalitis virus Rubella virus Venezuelan equine encephalitis virus Western equine encephalitis virus

Table B.1: Viruses included in enrichment probe set

Bacterial genus	**Bacterial species**
Acinetobacter	*baumanii* *calcoaceticus*
Bartonella	*henselae*
Bordetella	*pertussis*
Borrelia	*burgdorferi*
Brucella	
Burkholderia	*cepacia*
Chlamydophila	*pneumoniae* *psittaci*
Coxiella	*burnetii*
Enterobacter	*aerogenes* *cloacae*
Escherichia	*coli*
Haemophilus	*influenzae* *parainfluenzae*
Klebsiella	*pneumoniae* *oxytoca*
Legionella	*pneumophila*
Leptospira	
Listeria	*monocytogenes*
Moraxella	*catarrhalis*
Mycobacterium	*avium* *intracellulare* *tuberculosis*
Mycoplasma	*pneumoniae*
Neisseria	*meningitidis*
Nocardia	
Pseudomonas	*aeruginosa*
Serratia	*marcescens*
Staphylococcus	*aureus*
Stenotrophomonas	*maltophilia*
Streptococcus	*agalactiae* *pneumoniae* *pyogenes*
Treponema	*pallidum*

Table B.2: Bacteria included in enrichment probe set

Table B.3: Viral Multiplex Reference reagent 11/242 (UK NIBSC). This reference set included 25 viruses of various nucleic acid types, envelope types and genome sizes. 21/25 viruses had corresponding enrichment probes in our probe panel. (Adapted from Mee et al.)

Group	Family	Species/serotype	Envelope	Genome size	Included in probeset	Concentration (log10 copies/ml)
dsDNA	Adenoviridae	Adenovirus 2	No	35.9	Yes	NA
		Adenovirus 41		34.2	Yes	NA
	Herpesviridae	Human herpesvirus 1	Yes	151.2	Yes	NA
		Human herpesvirus 2		154.7	Yes	NA
		Human herpesvirus 3 (VZV)		124.8	Yes	NA
		Human herpesvirus 4 (EBV)		171.7	Yes	3.88
		Human herpesvirus 5 (CMV)		233.7	Yes	4.66
dsRNA	Reoviridae	Rotavirus A	No	18.5	Yes	6.76
ssRNA (+)	Astroviridae	Astrovirus	No	6.8	No	NA
	Caliciviridae	Norovirus GI	No	7.6	No	NA
		Norovirus GII		7.5	No	NA
		Sapovirus C12		7.5	No	NA
	Coronaviridae	Coronavirus 229E	Yes	27.2	Yes	NA
	Picornaviridae	Coxsackievirus B4	No	7.4	Yes	NA
		Rhinovirus A39		7.1	Yes	NA
		Parechovirus 3		7.2	Yes	7.07
ssRNA (-)	Orthomyxoviridae	Influenza A virus H1N1	Yes	13.2	Yes	NA
		Influenza A virus H3N2		13.6	Yes	NA
		Influenza B virus		14.2	Yes	NA
	Paramyxoviridae	Metapneumovirus A	Yes	13.3	Yes	NA
		Parainfluenzavirus 1		15.5	Yes	NA
		Parainfluenzavirus 2		15.7	Yes	NA
		Parainfluenzavirus 3		15.4	Yes	NA
		Parainfluenzavirus 4		17.4	Yes	NA
		Respiratory syncytial virus A2		15.2	Yes	3.75

References

Allicock, O. M., Guo, C., Uhlemann, A.-C., et al. (2018). “BacCapSeq: a Platform for Diagnosis and Characterization of Bacterial Infections.” eng. *mBio* 9 (5).

Bolger, A. M., Lohse, M., and Usadel, B. (2014). “Trimmomatic: a flexible trimmer for Illumina sequence data”. *Bioinformatics* 30: pp. 2114–20.

Bonsall, D., Ansari, M. A., Ip, C., et al. (2015). “ve-SEQ: Robust, unbiased enrichment for streamlined detection and whole-genome sequencing of HCV and other highly diverse pathogens”. *F1000Res* 4: p. 1062.

Boomer, J. S., To, K., Chang, K. C., et al. (2011). “Immunosuppression in patients who die of sepsis and multiple organ failure.” eng. *JAMA* 306 (23): pp. 2594–605.

Briese, T., Kapoor, A., Mishra, N., et al. (2015). “Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis.” eng. *mBio* 6 (5): e01491–15.

Chen, K. Y., Ko, S. C., Hsueh, P. R., et al. (2001). “Pulmonary fungal infection: emphasis on microbiological spectra, patient outcome, and prognostic factors.” eng. *Chest* 120 (1): pp. 177–84.

Cohen, J. I. and Lekstrom, K. (1999). “Epstein-Barr virus BARF1 protein is dispensable for B-cell transformation and inhibits alpha interferon secretion from mononuclear cells.” eng. *Journal of virology* 73 (9): pp. 7627–32.

Cowley, N. J., Owen, A., Shiels, S. C., et al. (2017). “Safety and Efficacy of Antiviral Therapy for Prevention of Cytomegalovirus Reactivation in Immunocompetent Critically Ill Patients: A Randomized Clinical Trial.” eng. *JAMA internal medicine* 177 (6): pp. 774–783.

Cuomo, C. A. (2017). “Harnessing Whole Genome Sequencing in Medical Mycology”. *Curr Fungal Infect Rep* 11: pp. 52–59.

Davenport, E. E., Burnham, K. L., Radhakrishnan, J., et al. (2016). “Genomic landscape of the individual host response and outcomes in sepsis: a prospective cohort study.” eng. *The Lancet. Respiratory medicine* 4 (4): pp. 259–71.

Daviaud, F., Grimaldi, D., Dechartres, A., et al. (2015). “Timing and causes of death in septic shock.” eng. *Annals of intensive care* 5 (1): p. 16.

Doan, T., Wilson, M. R., Crawford, E. D., et al. (2016). “Illuminating uveitis: metagenomic deep sequencing identifies common and rare pathogens”. *Genome Med* 8: p. 90.

Forbes, J. D., Knox, N. C., Peterson, C. L., et al. (2018). “Highlighting Clinical Metagenomics for Enhanced Diagnostic Decision-making: A Step Towards Wider Implementation”. *Comput Struct Biotechnol J* 16: pp. 108–120.

Gao, L., Zhong, J.-C., Huang, W.-T., et al. (2017). “Integrative analysis of BSG expression in NPC through immunohistochemistry and public high-throughput gene expression data.” eng. *American journal of translational research* 9 (10): pp. 4574–4592.

Jolley, K. A., Bliss, C. M., Bennett, J. S., et al. (2012). “Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain”. *Microbiology* 158: pp. 1005–15.

Langelier, C., Zinter, M. S., Kalantar, K., et al. (2018). "Metagenomic Sequencing Detects Respiratory Pathogens in Hematopoietic Cellular Transplant Patients". *Am J Respir Crit Care Med* 197: pp. 524–528.

Li, L., Deng, X., Mee, E. T., et al. (2015). "Comparing viral metagenomics methods using a highly multiplexed human viral pathogens reagent". *J Virol Methods* 213: pp. 139–46.

Libert, N., Bigaillon, C., Chargari, C., et al. (2015). "Epstein-Barr virus reactivation in critically ill immunocompetent patients." eng. *Biomedical journal* 38 (1): pp. 70–6.

Mee, E. T., Preston, M. D., Minor, P. D., et al. (2016). "Development of a candidate reference material for adventitious virus detection in vaccine and biologicals manufacturing by deep sequencing". *Vaccine* 34: pp. 2035–2043.

Moore, K. W., Waal Malefyt, R. de, Coffman, R. L., et al. (2001). "Interleukin-10 and the interleukin-10 receptor." eng. *Annual review of immunology* 19: pp. 683–765.

Morrison, T. E., Mauser, A., Wong, A., et al. (2001). "Inhibition of IFN-gamma signaling by an Epstein-Barr virus immediate-early protein." eng. *Immunity* 15 (5): pp. 787–99.

Munford, R. S. and Pugin, J. (2001). "Normal responses to injury prevent systemic inflammation and can be immunosuppressive." eng. *American journal of respiratory and critical care medicine* 163 (2): pp. 316–21.

Ong, D. S. Y., Bonten, M. J. M., Spitoni, C., et al. (2017). "Epidemiology of Multiple Herpes Viremia in Previously Immunocompetent Patients With Septic Shock." eng. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 64 (9): pp. 1204–1210.

"Proposed methods for testing and selecting the ERCC external RNA controls" (2005). *BMC Genomics* 6: p. 150.

Rol, M.-L., Venet, F., Rimmele, T., et al. (2017). "The REAnimation Low Immune Status Markers (REALISM) project: a protocol for broad characterisation and follow-up of injury-induced immunosuppression in intensive care unit (ICU) critically ill patients." eng. *BMJ open* 7 (6): e015734.

Salzberg, S. L., Breitwieser, F. P., Kumar, A., et al. (2016). "Next-generation sequencing in neuropathologic diagnosis of infections of the nervous system". *Neurol Neuroimmunol Neuroinflamm* 3: e251.

Satoh JK N; Yamamoto, Y. (2013). "Molecular network of chromatin immunopreciptation followed by deep sequencing-based (ChIP-Seq) Epstein-Barr virus nuclear antigen 1-target cellular genes supports biological implications of Epstein-Barr virus persistence in multiple sclerosis." *Clinical and Experimental Neuroimmunology*: pp. 181–192.

Scicluna, B. P., Vught, L. A. van, Zwinderman, A. H., et al. (2017). "Classification of patients with sepsis according to blood genomic endotype: a prospective cohort study." eng. *The Lancet. Respiratory medicine* 5 (10): pp. 816–826.

Walton, A. H., Muenzer, J. T., Rasche, D., et al. (2014). "Reactivation of multiple viruses in patients with sepsis." eng. *PloS one* 9 (2): e98819.

Woese, C. R. (1987). "Bacterial evolution." eng. *Microbiological reviews* 51 (2): pp. 221–71.

Wong, A. M. G., Kong, K. L., Chen, L., et al. (2013). “Characterization of CACNA2D3 as a putative tumor suppressor gene in the development and progression of nasopharyngeal carcinoma.” eng. *International journal of cancer* 133 (10): pp. 2284–95.

Wood, D. E. and Salzberg, S. L. (2014). “Kraken: ultrafast metagenomic sequence classification using exact alignments”. *Genome Biol* 15: R46.