**SUPPLEMENTARY RESULTS**

**1 – SUPPLEMENTARY RESULTS for de-novo assembly**

**Supplementary Result figures 1 to 3**

**2 – SUPPLMENTARY RESULTS for inStrain optimisation**

**Supplementary Result Figures 4 to 8 and Supplementary Result Table 1**

**1 – SUPPLEMENTARY RESULTS: De-novo Assembly**

One aim of this study was to perform de-novo assembly of viral bait capture sequence data, to assess whether high-quality contigs and genomes could be assembled. Additionally, we aimed to identify differences in contig/genome quality between the three viral bait capture methods (BCMs) employed in Pilot studies 1 and 2.

For the de-novo assembly, we used the MetaSPAdes function of SPAdes (v3.14) to assemble high-quality (HQ), short, paired end reads (minimum 100 bp) that had been filtered using KneadData. The quality of the resulting MetaSPAdes contigs and genomes was assessed using the MetaQUAST function of QUAST (v4.4). Several metrics were analysed using MetaQUAST, including:

1. The total number of contigs generated at various minimum length thresholds: no minimum length, 1 kb, 5 kb, 10 kb, and 25 kb, for the three viral BCMs in the Pilot 1 study (see **Supplementary Result Figure 1**).
2. The mean length differences for contigs ≥500 bp at the single-sample level (see **Supplementary Result Figure 2**).
3. Key alignment metrics such as the largest alignment, total aligned length, mismatches per 100 kb, indels per 100 kb, and genome fraction sequenced for clinical and Twist SARS-CoV-2 samples across the three viral BCMs in the Pilot 1 study (see **Supplementary Result Figure 3**).

“In Pilot 1, three viral BCMs were compared: veSEQ 8-plex, RVI-seq 8-plex (in-house), and Illumina 3-plex. References to four viral BCMs additionally include the Illumina 8-plex method, which was incorporated in subsequent analyses after Pilot 1.”

**Contig numbers across different viral BCMS (Supplementary Result Figure 1)**

De-novo assembly using MetaSPAdes revealed significant differences in the total number of contigs generated across the three distinct viral BCMs. When contig numbers per sample were compared between the three BCMs using a non-parametric paired Friedman test with Dunn's multiple comparisons adjustment, significant differences were observed. (**a**) Specifically, the comparison between the RVI-seq and veSEQ methods (p = 0.009) and between the veSEQ and Illumina methods (p = 0.0295) were statistically significant for all contigs. (**b**) For the ≥1 kb contig dataset, the only significant difference was found between the RVI-seq and Illumina methods (p = 0.0201). (**c**) No significant differences were detected between the three BCMs for the ≥5 kb contig dataset. (**d**) Notably, the veSEQ and RVI-seq methods produced the highest number of contigs ≥25 kb for six of the eight Heron-positive COVID-19 samples, while the Illumina method generated only two contigs corresponding to the SARS-CoV-2 genome.

**Comparison of 500bp minimum contig lengths across different viral BCMs**

A statistical comparison of all contigs within a sample across the three distinct viral BCMs was performed using a minimum contig lengths of 500 bp. The analysis revealed significant differences among the three viral BCMs in the Pilot 1 study (see **Supplementary Result Figure 2**). This comparison, assessed with the Kruskal-Wallis test, indicated that the median contig lengths varied significantly across the BCMs between most samples. A total of seven SARS-CoV-2 positive samples were analysed across the three BCMs. Notably, two veSEQ samples showed a statistically higher median contig length compared to the Illumina and in-house (RVI-seq) methods. In contrast, one RVI-seq sample exhibited a higher median contig length compared to the Illumina samples. Furthermore, the median length of contigs generated from hCoV samples using the Illumina and RVI-seq BCMs was significantly higher than that generated by the veSEQ method. For most other samples, including the H3N2 dilution series and hPIV1, H1N1, and their mixtures with SARS-CoV-2, the Illumina method produced contigs with a statistically higher median length than the veSEQ method, and often also compared to the RVI-seq method. Additionally, some RVI-seq samples exhibited a higher median contig length than the veSEQ samples. All statistically significant individual differences from Dunn's multiple comparisons tests are presented in Supplementary Result Figure 2, with p-values indicated as \* (<0.05), \*\* (<0.01), \*\*\* (<0.001), and \*\*\*\* (<0.0001).

**Differences in largest contigs and total aligned length, with comparable mismatches and indels across three viral BCMs (Supplementary Figure 3)**

(**a**) The veSEQ and RVI-seq methods produced larger alignments for five out of the seven SARS-CoV-2 positive samples, compared to the Illumina method. (**b**) However, the total aligned length of SARS-CoV-2-containing samples was found to be comparable across all three BCMs. (**c**) The number of mismatches and (**d**) indels per 100 kb were similar between the BCMs. The non-parametric one-way ANOVA with Dunn’s multiple comparisons test yielded p-values of 0.0854 for mismatches and 0.5045 for indels. (**e**) All three viral BCMs demonstrated the ability to sequence between 99.5% and 100% of the SARS-CoV-2 genome.

**2 – SUPPLEMENTARY RESULT for inStrain optimisation**

Another objective of this study was to optimize inStrain analysis for the taxonomic classification of viral bait captured short reads by comparing three different analysis options in inStrain. The optimal inStrain configuration was then compared with Kraken2/Bracken analysis. Initial optimization was conducted on sequence data from the ZeptoMetrix Respiratory Panel 2.1 (RP2.1) controls, followed by a subsequent test on high-quality paired-end short reads (down sampled to a maximum of 1 million reads) from the Pilot 2 viral BCMs. Three distinct inStrain options were evaluated: the first option utilized database mode with a minimum genome coverage of 1 and a minimum read average nucleotide identity (ANI) of 0.92; the second option did not employ database mode, set genome coverage to zero, and had a minimum read ANI of 0.9; the third option, the default setting, also did not use database mode, but had a minimum genome coverage of zero and a minimum read ANI of 0.95.

The heatmap presented in **Supplementary Result Figure 4** illustrates the short-read abundances from the ZeptoMetrix Respiratory Panel 2.1 (RP2.1) Controls. Below is a detailed explanation of the heatmap, from top to bottom.

Adenovirus Type 1 (or Type A) was identified as Human Adenovirus Type 1 in low abundance (less than 1.6%) across all three inStrain options and all four viral BCMs in ZeptoMetrix (ZM) Control 1. The most prevalent virus identified was Human Mastadenovirus B1, also known as Adenovirus Type 3 (or Type C), with inStrain signals ranging from 68.5% to 72.1% across all three inStrain options and all four viral BCMs. Adenovirus Type 31 was not detected by inStrain due to its absence in the NCBI viral reference database, "viral.1.1.genome 2023-09-09," which was used for inStrain analysis. All three inStrain options identified a substantial signal for human adenovirus 7, with abundances ranging from 23% to 25%.

All eight segments of the H1N1 pandemic A/NY strain were identified at comparably low levels by all three inStrain options and across all four viral BCMs. The H3N2 A/Brisbane strain was identified by all three inStrain options in only three segments (1, 4, and 6), while the remaining five segments were detected in ZMControl 2 as the FLU A H3N2 NY strain.

It is worth noting that NCBI viral RefSeq 1.1.genomic.2023-09-09\_1 only has one influenza B “Lee 1940” reference genome in it.

Metapneumovirus 8 was identified as Human Metapneumovirus by inStrain Option 2 and was only detected by the RVI-seq method. Parainfluenza types 1 and 4, designated as human parainfluenza 1 and 4a in the NCBI virus database, were identified by all three inStrain options and all four viral BCMs. Rhinovirus 1A was identified as human rhinovirus 1 by inStrain Option 1 using the veSEQ and RVI-seq methods, and by inStrain Option 2 across all four viral BCMs. It was not identified by inStrain Option 3.

The final virus detected in the ZeptoMetrix RP2.1 Control 1, SARS-CoV-2, was identified by all four viral BCMs using all three inStrain options. From Control 2, all eight segments of the H1N1 A/Caledonia strain were exclusively identified by inStrain Option 2, the RVI-seq method, and the Illumina method (at both 3-plex and 8-plex). Among the four human coronaviruses, strain 229E was the most prevalent and was detected at comparable frequencies by all three inStrain options and all four viral BCMs. The recombinant HKU1, which was included in the panel as a short sequence, was identified exclusively by inStrain options 2 and 3, and not by the veSEQ method.

All eight segments of the influenza B virus were identified by all three inStrain options. Notably, only inStrain options 1 and 2 could detect all eight segments across all four viral BCMs. Parainfluenza type 3 was consistently identified using inStrain options 2 and 3 for the reference strains (NC\_001796.2 and NC\_07546.1) available in the NCBI viral database. Similarly, parainfluenza type 2 and RSV A were detected at comparable frequencies across all four viral BCMs using all three inStrain options.

A comparison was then made between the three inStrain options and Kraken2 analysis using all samples from Pilot 2. Details of the samples processed, and taxa detected are provided in **Supplementary Result Table 1**. The inStrain Option 1 parameters identified the fewest viral taxa and processed the smallest proportion of samples (76%) compared to the other inStrain options and Kraken analysis. In contrast, inStrain Option 2 processed the highest number of samples (95%) and generated the largest number of viral hits (n=381). Kraken2 analysis was able to process between 93% of the samples. Both inStrain Option 3 (n=356) and Kraken2/Bracken (with 217 viral signals) showed reduced background signal and low-abundance noise compared to the other methods. However, it is important to note that inStrain can distinguish between different influenza strains, such as influenza A H3N2 and H1N1, and identify all influenza segments, resulting in a higher number of viral hits by default.

|  |  |  |
| --- | --- | --- |
| **Supplementary Result Table 1. Number of taxa identified between the three different inStrain methods and Kraken2/Bracken analysis** | | |
| **Method used and specific options** | **# of viral taxa hits** | **Samples processed (n374)** |
| inStrain option 1 “database mode, min genome coverage 1, min read average nucleotide identity 0.92, and skip mm profiling” | 80 | 290/374 (76%) |
| inStrain option 2 “no database mode, min genome coverage 0, min read average nucleotide identity 0.9” | 381 | 356/374 (95%) |
| inStrain option 3 (default) “no database mode, min genome coverage 0, min read average nucleotide identity 0.95” | 356 | 351/374 (94%) |
| Kraken2/Bracken | 217 | 349/374 (93%) |

In conclusion, inStrain option 2 provided the most accurate classification of short reads across multiple instances. As a result, a further comparison was conducted between inStrain option 2 and Kraken2/Bracken analysis, as shown in **Supplementary Result Figure 5**. This comparison revealed that both inStrain (**Supplementary Results Figure 5a**) and Kraken2/Bracken (**Supplementary Result Figure 5b**) analysis yielded similar results. However, Kraken2 failed to detect RSV A, while inStrain successfully identified RSV A across all four bait capture methods. Additionally, Kraken2 was unable to distinguish between different influenza A strains and did not identify signals for the various influenza strain segments. Among the four bait capture methods, the veSEQ method failed to detect certain viruses present in ZeptoMetrix RP2.1 Control 2, which were successfully identified by the RVI-seq method and Illumina method (at both 3-plex and 8-plex).

The next three supplementary result figures present a comparison of inStrain methods and Kraken analysis across various sample types: SARS-CoV-2 positive samples (**Supplementary Result Figure 6**), NIBSC flu samples (**Supplementary Result Figure 7**), and parainfluenza, metapneumovirus, rhinovirus, and their mixtures, as well as nasal swab negative samples with and without spike-ins containing either H1N1, H3N2, RSV B, FLU B, or SARS-CoV-2 (**Supplementary Result Figure 8**).

**SARS-CoV-2 Positive Sample Analysis (Supplementary Result Figure 6)**  
All four analytical methods and BCMs detected the presence of SARS-CoV-2 in the same samples, except for inStrain Option 1, which failed to detect a portion of the SARS-CoV-2 signal in clinically positive samples. Additionally, all three inStrain methods identified a strong signal for ssRNA phage in a subset of samples. A weak signal for the spiked MS2 phage was observed across all four analytical methods and BCMs.

The results for influenza A, influenza B, and RSV B are presented in **Supplementary Result Figure 7**. Both Kraken and inStrain option 2 were able to detect influenza B at a copy number as low as 1k. In the 10k copy number sample, inStrain option 2 detected four segments, option 1 detected two segments, and option 3 did not detect any segments. Samples processed using the veSEQ and RVI-seq methods produced the most consistent and accurate results.

The **RVI-seq**and**veSEQ** methods were the most effective at detecting individual influenza A H3N2 segments (highlighted by black boxes). In the 100-copy number sample, the **RVI-seq** method identified two segments, with inStrain options 2 and 3. None of the Illumina methods were able to detect a signal in the 100-copy number sample. The **RVI-seq** method also successfully detected all eight segments in the 1k copy number sample, regardless of the inStrain option used. All segments were detectable in the 10k, 100k, and 1M copy number samples across all methods.

For **influenza A H1N1** segments (highlighted in red boxes), inStrain options 2 and 3, along with the **RVI-seq** method, were the most effective, particularly at the 100 and 1k copy number samples. However, the enhanced sensitivity of inStrain options 2 and 3 slightly reduced specificity, leading to weak, non-specific detections between different influenza strains.

All four viral BCMs methods detected the influenza A virus (H3N2 and H1N1) in the 1k copy number sample using Kraken. However, only the **RVI-seq** method detected signals from the 100-copy number sample for both H3N2 and H1N1 (Kraken column).

For **RSV B**, both Kraken and inStrain methods classified the virus as Human Orthopneumovirus. The **veSEQ** and **Illumina** methods detected the virus at 1k sample using Kraken and all three inStrain options. Additionally, the **RVI-seq**method successfully detected the virus at 100 copies per sample, but only with Kraken and inStrain options 2 and 3 (highlighted by blue arrows).

Kraken identified **Alphapapillomavirus 7**, while inStrain detected **human papillomavirus** in most samples, with varying signal intensities. The signal was particularly prominent in the low-copy number samples, suggesting potential contamination from papillomavirus, especially in low-biomass samples. Kraken identified **Human Mastadenovirus B**, while inStrain detected **Human Adenovirus B1** as contaminants in the RSV B negative control sample.

All analysis methods detected a weaker signal for **SARS-CoV-2,** with the Illumina methods showing stronger intensity compared to the RVI-seq and **veSEQ**methods. The positioning of the SARS-CoV-2 samples on the plate map suggested potential contamination from neighbouring wells.

**Supplementary Result Figure 8 :** Lastly, all four analytical methods and BCMs successfully detected human metapneumovirus. Detection of human rhinovirus A, either alone or in combination with H1N1, was only possible through the identification of a weak signal for rhinovirus C using Kraken. This was observed exclusively in samples from the RVI-seq BCM (indicated by red arrows). In the multiplex sample containing influenza, RSV, and SARS-CoV-2, only the SARS-CoV-2 signal was detected across all four analytical methods and BCMs.

In the five nasal swab negative samples, the veSEQ method produced no signal, while the RVI-seq method generated only weak signals (shown in black boxes). However, when using the two Illumina bait capture methods, all four analytical methods detected robust signals for Human Orthopneumovirus and Human Respirovirus 1/Human Parainfluenza Virus (indicated by black boxes).

Kraken successfully identified both H1N1 and H3N2 in the negative nasal swab samples spiked with the two influenza strains. However, inStrain options 2 and 3 were only able to detect all eight segments for both strains in the samples prepared with the RVI-seq and Illumina methods (depicted by the black dashed ellipse). The veSEQ method failed to detect certain segments.

In conjunction with Kraken and inStrain options 2 and 3, as well as the veSEQ, RVI-seq, and Illumina 3-plex methods (marked by an asterisk), the presence of respiratory syncytial virus (RSV B), also known as human Orthopneumovirus, was identified in the third nasal swab sample spiked with RSV B. Notably, Kraken and the RVI-seq method were the only methods that successfully identified the influenza B virus in the nasal swab negative 4 sample (marked with \*\*).