Analysis of viral barcodes in single-cell RNA-seq data.

David C. King and Fitz Fitzmeyer

Synopsis

The following is an enumeration of distinct viral barcode sequences, taking into account read quality and considering the subsets defined by different samples (mosquito midguts), cells (cell barcode or whitelist), and UMIs (unique molecular identifiers). However, many distinct viral barcodes are attached to the same UMI with the same cell. This observation is unexpected and, at the time of writing, an anomalous observation that we consider unreliable until better understood. We can detect these anomalies by counting how many distinct viral sequences a UMI maps to within cells, and if this number is greater than 1, all reads for that midgut/cell/UMI combination are discarded.

Conclusion

In this dataset, the anomalous UMI/viral barcode sequences do not provide reliable evidence that more than one virus can infect a single cell. The phenomenon occurs at the highest quality standard (37), so cannot be attributed to base calling errors in the sequencing. Until the anomaly is explained, these observations must be filtered from the analysis, leaving few cases of deduplicated reads that indicate a distinct viral origin (ranging from 0 cases to 23 in different midgut samples). This analysis ignores some considerations. First, there may be evidence outside of the viral barcode that two distinct sequences share a UMI. If that is happening, the number of reliable counts would decrease further. Second, we did not evaluate whether this result is expected by UMI collision. However, an 11 base-pair UMI can have 4,194,304 unique sequences, far outnumbering the number of reads captured for a given cell. In conclusion, without being able to account for distinct sequences sharing a UMI in this dataset, we are unable to reliably observe viral diversity in a cell.

Data

There are 7 samples (West Nile virus infested mosquito midguts) represented by read pair data in fastq format. After searching for the viral barcode in the second read, the number of usable reads for this analysis is considerably less than the original file.

After selecting for reads with viral barcode

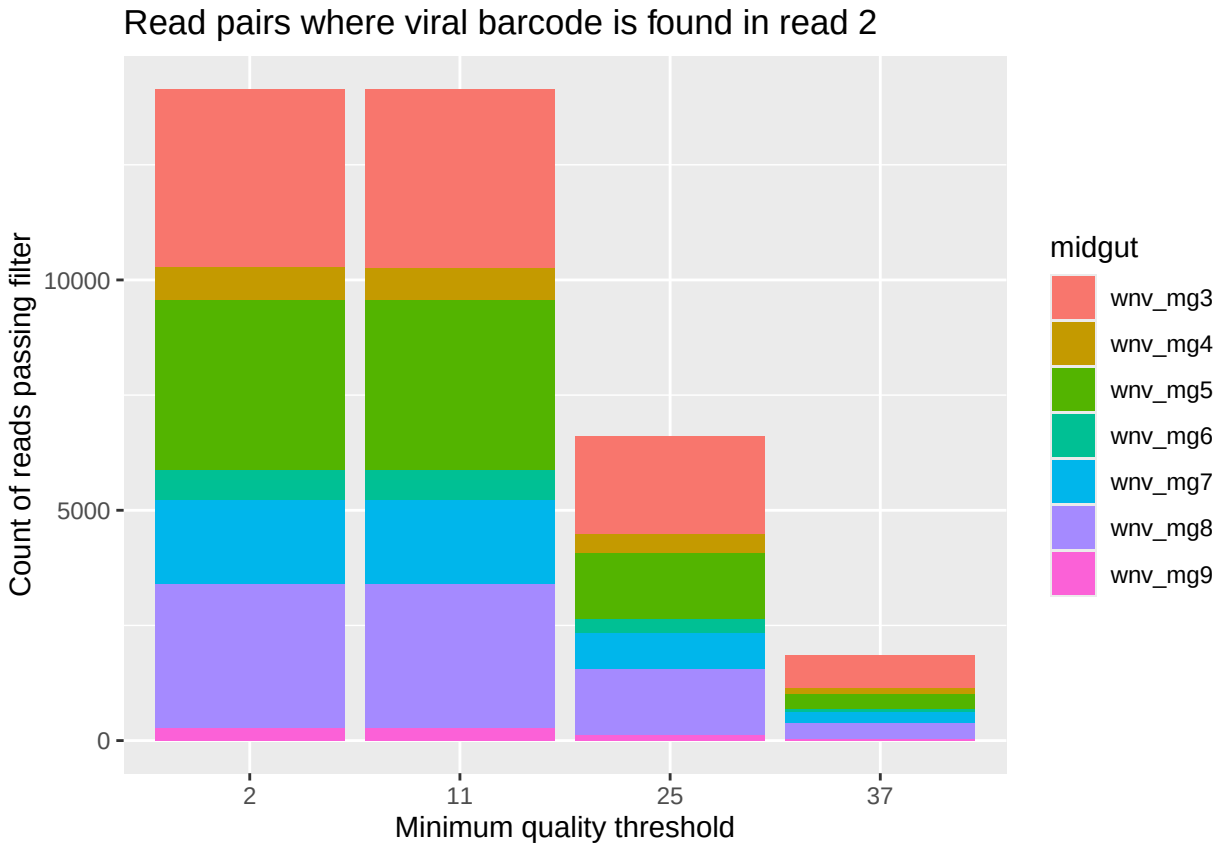
The script `pysam_analyze.py` iterates over the matched read pair files and outputs the extracted barcodes, their quality string, and the minimum quality of each.

Summary of qualities per midgut

Table 1: Number of reads with the following lowest quality score

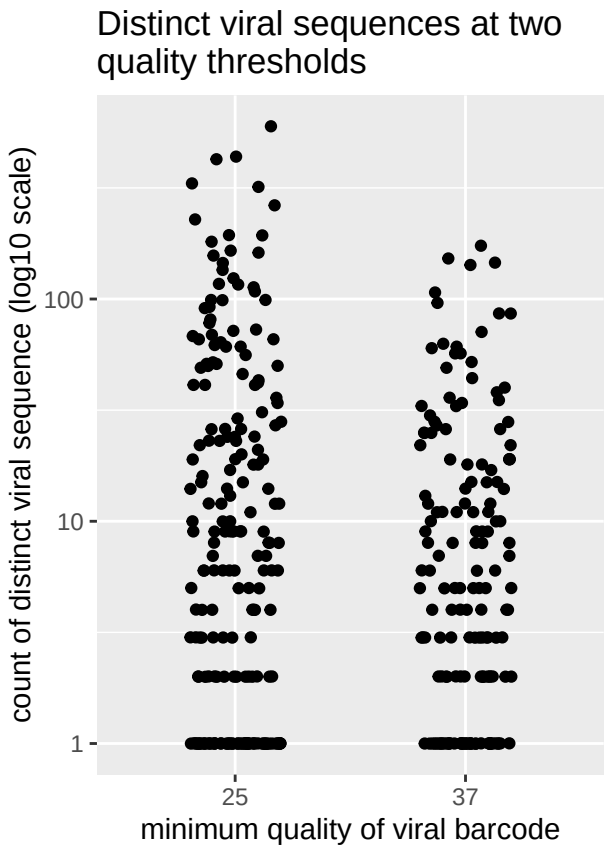
midgut	37	25	11	2
wnv_mg3	722	1413	1748	2
wnv_mg4	137	271	292	1
wnv_mg5	310	1117	2268	5
wnv_mg6	84	219	335	2

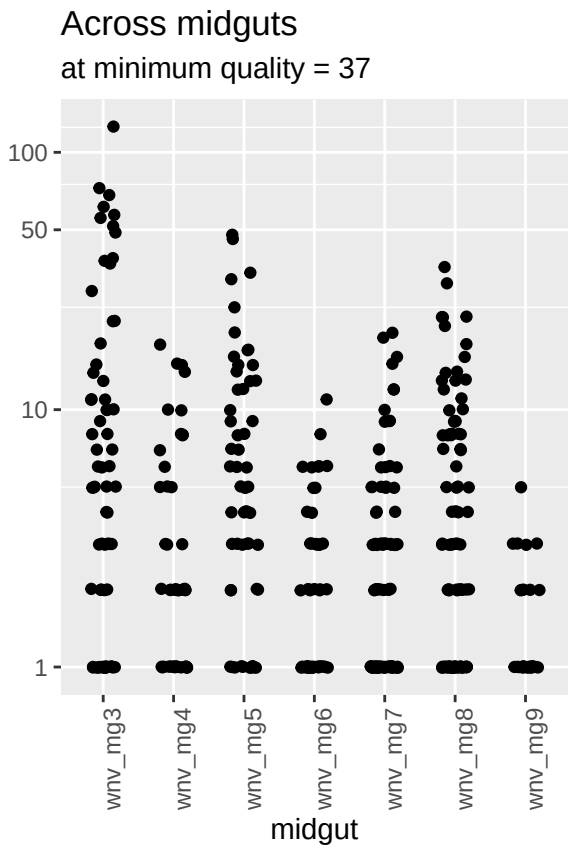
midgut	37	25	11	2
wnv_mg7	220	564	1049	2
wnv_mg8	352	1087	1677	5
wnv_mg9	29	85	158	0



Viral diversity

At the two top quality scores, 25 and 37, how many distinct viral barcode sequences do we see and how often do we see them? For each plot, a single point is a distinct viral barcode, its y value is the number of reads having that barcode, and it is distinct within the groups of the plot (x-axis).





Numbers for the above plots

Table 2: Breakdown of distinct viral counts by min. quality score

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	>= 25
25	57	15	8	7	4	9	3	4	9	3	1	4	1	3	2	1	1	2	3	1	1	1	3	3	59
37	42	16	13	7	8	4	2	5	4	3	5	3	1	2	3	0	1	2	4	0	0	2	0	0	33

Table 3: Breakdown of distinct viral counts by midgut at min. score of 37

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	>= 25
mg3	17	5	6	2	5	3	2	2	1	3	3	0	1	1	1	0	0	1	0	0	0	2	0	12
mg4	16	12	3	0	3	1	1	3	0	2	0	0	0	1	2	0	0	1	0	0	0	0	0	0
mg5	11	4	6	5	3	3	2	2	2	1	0	2	2	1	2	1	2	0	0	1	0	0	0	5
mg6	22	7	7	2	2	5	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
mg7	34	9	19	3	6	4	1	0	3	1	0	2	0	0	1	1	0	0	1	1	0	0	0	0
mg8	27	11	8	4	4	1	3	7	3	2	1	1	3	2	0	1	0	1	0	0	1	0	3	2
mg9	16	5	4	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Deduplication and the UMI

Background

cDNA libraries attempt to capture RNA molecules in a sample, such as mRNA transcripts produced during gene expression. To provide enough material for sequencing, the cDNA is usually amplified by PCR, boosting the "signal" of a molecular by increasing its representation in the isolated cDNA. The production of the RNA molecule is of biological origin. The duplication of the cDNA fragment is of technical origin. When the goal is to count the number of molecules produced by biological means, the technical duplication should be controlled for.

The current approach is to use a unique molecular identifier (UMI), a short oligonucleotide attached to the cDNA of a biologically produced molecule *before* the amplification step. Therefore, identical sequences which include the same UMI are assumed to come from a single molecule and are processed with software to collapse, or deduplicate, the duplicate sequences into a single observation.

In actual sequenced reads, there can be variation in the observed sequence due to sequencing errors or uncertainty in the base call. Therefore, some ambiguity must be taken into account when deduplicating reads. Most methods allow a small number of mismatches between UMIs if the attached sequence maps to the same place in the target genome. Therefore, with the aid of the alignment, UMIs and the read can have variation.

Given any two sequence reads, the following cases are treated this way: - Same UMI; same sequence: single molecule duplicated by PCR - Different UMI; same sequence: two distinct molecules generated biologically (e.g. two mRNAs transcribed from a gene) - Different UMI; different sequence: If UMIs are close (i.e. <= 1 mismatch) AND the attached sequences map to exactly the same place via alignment, the two reads may have been a PCR duplication. - Same UMI; different sequence: ??????????????

Anomalous UMI-sequence combinations

Grouping by distinct combinations of UMI and payload will determine uniqueness.

Filter first by quality >= 37, then count instances after grouping by distinct UMIs. There is no deduplication, so "occurence" includes PCR duplicate read counts. No breakdown by yet midgut.

Table 4: How many different viral barcodes share a distinct UMI?

Distinct viral barcode sequences that share a UMI	occurrence
1	3277
2	102
3	30
4	19
5	13
6	10
7	8
8	3
9	5
10	5
>= 10	106

Break apart by midgut

Viral diversity after deduplicating

After filtering out the anomalous cases, deduplication can be performed by collapsing exact copies of UMI:viral barcode reads into a single row in the table.

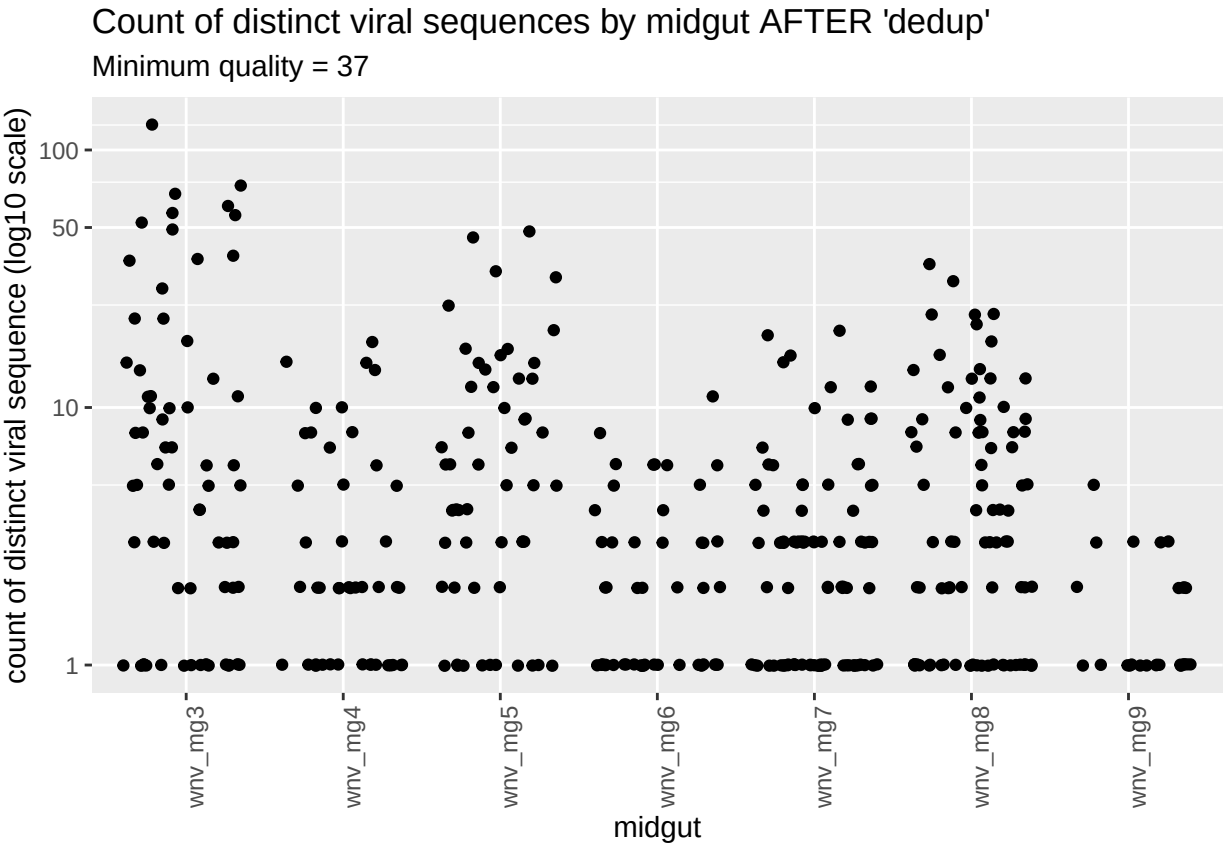


Table 5: data for figure above

	wnv_mg3	wnv_mg4	wnv_mg5	wnv_mg6	wnv_mg7	wnv_mg8	wnv_mg9
1	17	16	11	22	34	27	16
2	5	12	4	7	9	11	5
3	6	3	6	7	19	8	4
4	2	0	5	2	3	4	0
5	5	3	3	2	6	4	1
6	3	1	3	5	4	1	0
7	2	1	2	0	1	3	0
8	2	3	2	1	0	7	0
9	1	0	2	0	3	3	0
>=10	24	6	17	1	7	17	0

Viral diversity within cells

Break out by cell barcode

```
## # A tibble: 6 x 8
## # Groups:   midgut, cellbarcode [5]
##   midgut  cellbarcode     umi         viral_seq     distinct_viral_seque~1     n
##   <chr>   <chr>           <chr>       <chr>                          <int> <int>
## 1 wnv_mg3 AAACCTGCGCTAACC TGTACTGCCTT CTAACCGTAACT~                       1     1
## 2 wnv_mg3 AAACCTGCGCTAACC TGTACTGTCTT CTAACCGTAACT~                       1     1
## 3 wnv_mg3 AAACCTGCTTCTCGT ACTGTCTTCAC CTAACTGTAACC~                       1     1
## 4 wnv_mg3 AAACCTGGTGATAAA CTCGGCCCAGG CTTACCGTAACG~                       1     1
```

```
## 5 wnv_mg3 AAACGGGAACCATGT AGGCATAGTGG CTTACCGTTACT~                      1     1
## 6 wnv_mg3 AAACGGGGTCCATCC TTTCTTCATAG CTAACTGTAACC~                      1     1
## # i abbreviated name: 1: distinct_viral_sequences
## # i 2 more variables: midgut_cellbarcode <chr>, umi_viral_seq <chr>
```

```
## .
##    1    2    3    4    6    7
## 5146   45    5    1    1    1
```

```
##
##    1    2    3    4
## 5108   42    1    1
```

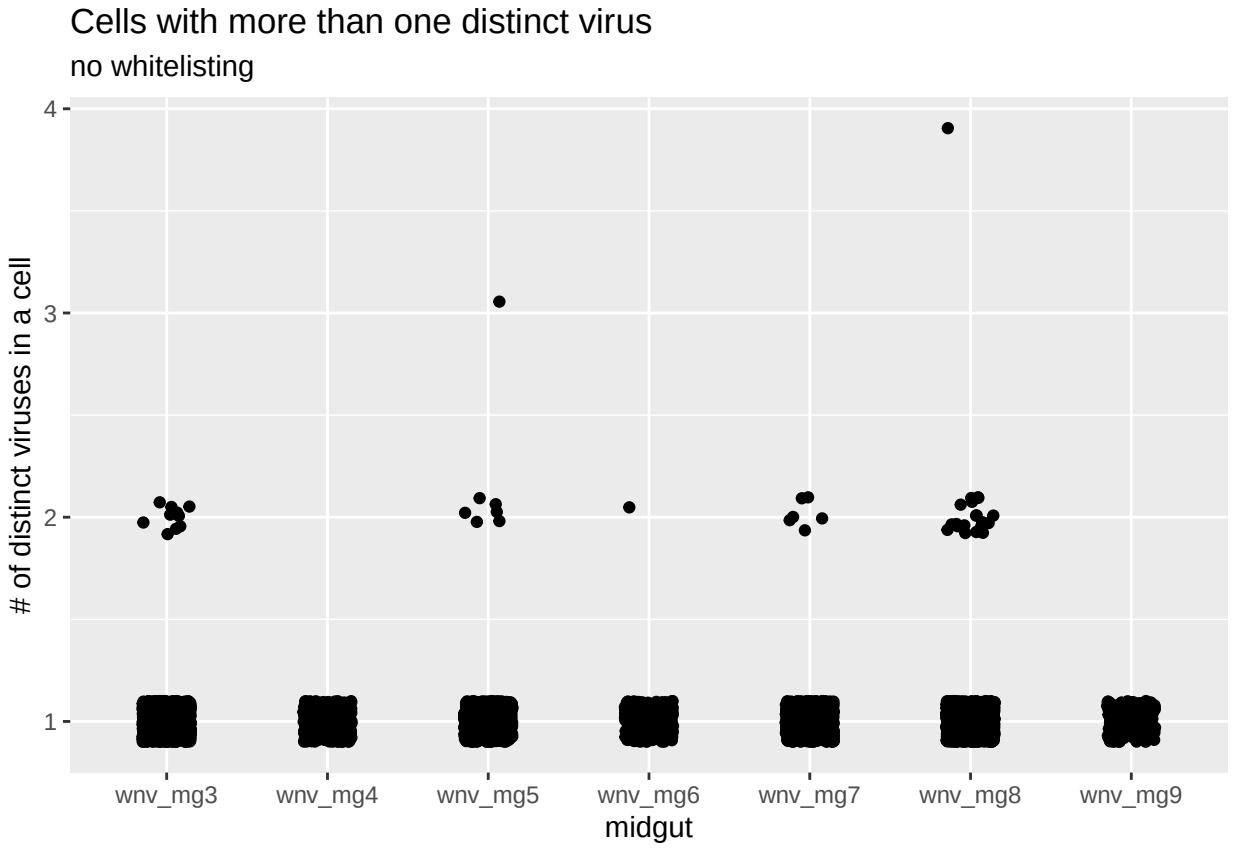


Table 6: Numbers for the plot above

	wnv_mg3	wnv_mg4	wnv_mg5	wnv_mg6	wnv_mg7	wnv_mg8	wnv_mg9
1	1309	415	880	332	794	1194	184
2	10	0	6	1	6	19	0
3	0	0	1	0	0	0	0
4	0	0	0	0	0	1	0

```
## # A tibble: 6 x 8
## # Groups:   midgut, cellbarcode_consensus_1 [5]
##   midgut  cellbarcode_consensus_1 umi     viral_seq distinct_viral_seque~1     n
##   <chr>   <chr>                   <chr>   <chr>                      <int> <int>
## 1 wnv_mg3 AAACCTGCGCTAACC         TGTACT~ CTAACCGT~                      1     1
## 2 wnv_mg3 AAACCTGCGCTAACC         TGTACT~ CTAACCGT~                      1     1
## 3 wnv_mg3 AAACCTGGTGATAAA         CTCGGC~ CTTACCGT~                      1     1
```

```
## 4 wnv_mg3 AAACGGGAACCATGT         AGGCAT~ CTTACCGT~                      1     1
## 5 wnv_mg3 AAACGGGGTCCATCC         TTTCTT~ CTAACTGT~                      1     1
## 6 wnv_mg3 AAACGGGGTTCACTA         CTACTG~ CTGACAGT~                      1     2
## # i abbreviated name: 1: distinct_viral_sequences
## # i 2 more variables: midgut_cellbarcode <chr>, umi_viral_seq <chr>
```

```
## .
##    1    2    3    4    6    7
## 4370   43    6    1    1    1
```

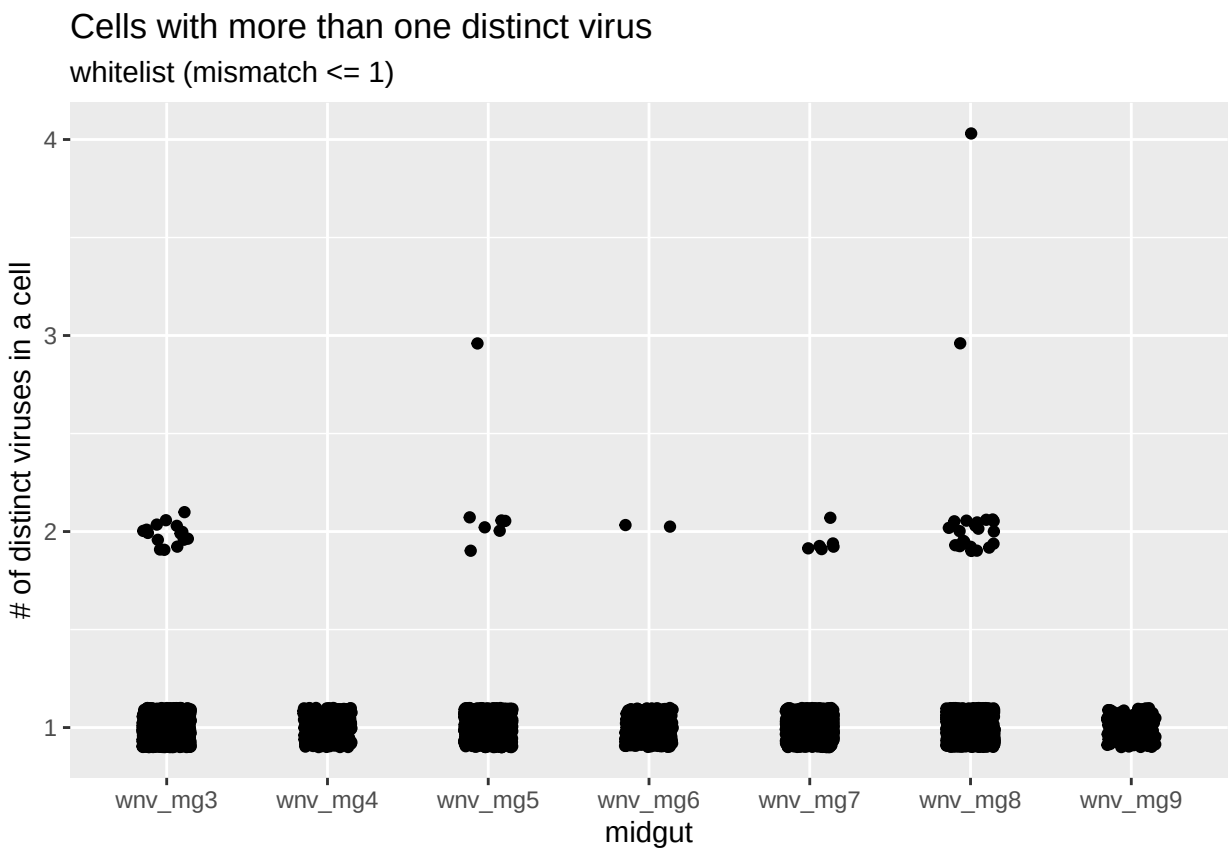


Table 7: Numbers for the plot above

	wnv_mg3	wnv_mg4	wnv_mg5	wnv_mg6	wnv_mg7	wnv_mg8	wnv_mg9
1	1134	346	742	273	663	1003	151
2	15	0	6	2	6	21	0
3	0	0	1	0	0	1	0
4	0	0	0	0	0	1	0

Conclusion

In this dataset, the anomalous UMI/viral barcode sequences do not provide reliable evidence that more than one virus can infect a single cell. The phenomenon occurs at the highest quality standard (37), so cannot be attributed to base calling errors in the sequencing. Until the anomaly is explained, these observations must be filtered from the analysis, leaving few cases of deduplicated reads that indicate a distinct viral origin (ranging from 0 cases to 23 in different midgut samples). This analysis ignores some considerations. First, there may

be evidence outside of the viral barcode that two distinct sequences share a UMI. If that is happening, the number of reliable counts would decrease further. Second, we did not evaluate whether this result is expected by UMI collision. However, an 11 base-pair UMI can have 4,194,304 unique sequences, far outnumbering the number of reads captured for a given cell. In conclusion, without being able to account for distinct sequences sharing a UMI in this dataset, we are unable to reliably observe viral diversity in a cell.