

Figures For CELL-SYSTEMS-D-22-00268R1

Chunyu Zhao

November 29, 2022

Contents

Data	1
ANI bins	1
Load BAM counts	1
Figure 4A: Reference Bias	2
Figure 6: MAPQ & MAPID Boxplot	2
Figure S6: Read Flow Chat	2
Cross Mapping	2
Figure S2: for Neighbor Genome	2
Figure 5A: for on-target Genome	3
MIDAS2 SNPs summary	3
Figure 5B: horizontal coverage	3
Figure S3: vertical coverage	4
Figure S4: Local Alignment	4
Figure S5: BWA	4
86 NCBI strains	4
Figure 7A	4
Figure 7B	6
Figure 7C	6
Figure 4B & 4C	6
Figure S1	6
Figure S7 Runtime	7

Data

ANI bins

Load BAM counts

- For the neighbor reads, we look at neighbor correctly aligned, and neighbor incorrelctly aligned (on the species level).

Figure 4A: Reference Bias

```
##
## Pearson's product-moment correlation
##
## data: reference_bias_data$intra_species_ani and reference_bias_data$aligned_rate
## t = 32.276, df = 810, p-value < 0.00000000000000022
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## 0.7183175 0.7786677
## sample estimates:
## cor
## 0.7500498
```

Figure 6: MAPQ & MAPID Boxplot

Figure S6: Read Flow Chat

Cross Mapping

- neighbor-incorrect: neighbor reads cross mapped to on-target genome
- neighbor-correct: neighbor reads mapped to off-target genome

Figure S2: for Neighbor Genome

```
##
## Error: on_target_species
##
## Df Sum Sq Mean Sq F value
## intra_species_ani 1 0.799 0.7990 36.158
## between_species_ani 1 2.533 2.5329 114.621
## intra_species_ani:between_species_ani 1 0.095 0.0954 4.318
## Residuals 310 6.850 0.0221
## Pr(>F)
## intra_species_ani 0.00000000511 ***
## between_species_ani < 0.0000000000000002 ***
## intra_species_ani:between_species_ani 0.0385 *
## Residuals
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Error: Within
##
## Df Sum Sq Mean Sq F value
## intra_species_ani 1 0.000 0.000 0.00
## between_species_ani 1 15.774 15.774 3526.30
## intra_species_ani:between_species_ani 1 0.227 0.227 50.63
## Residuals 3224 14.422 0.004
## Pr(>F)
## intra_species_ani 1
## between_species_ani < 0.0000000000000002 ***
## intra_species_ani:between_species_ani 0.000000000000136 ***
```

```
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5A: for on-target Genome

```
##
## Error: on_target_species
##
##      Df Sum Sq Mean Sq F value
## intra_species_ani      1  2.2129   2.2129  339.17
## between_species_ani      1  1.8280   1.8280  280.17
## intra_species_ani:between_species_ani      1  0.2466   0.2466   37.79
## Residuals          310  2.0226   0.0065
##
##                                Pr(>F)
## intra_species_ani          < 0.0000000000000002 ***
## between_species_ani          < 0.0000000000000002 ***
## intra_species_ani:between_species_ani          0.00000000242 ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Error: Within
##
##      Df Sum Sq Mean Sq F value
## intra_species_ani      1   1.922   1.922  1520
## between_species_ani      1 10.630  10.630  8407
## intra_species_ani:between_species_ani      1   4.199   4.199  3321
## Residuals          3224   4.076   0.001
##
##                                Pr(>F)
## intra_species_ani          <0.0000000000000002 ***
## between_species_ani          <0.0000000000000002 ***
## intra_species_ani:between_species_ani          <0.0000000000000002 ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

MIDAS2 SNPs summary

Figure 5B: horizontal coverage

- Only one species in the reads and two species in the database.
- for off-target reads, there is no concept of percentage_aligned_reads because it is all from cross-mapping

```
##
## Error: on_target_species
##
##      Df Sum Sq Mean Sq F value          Pr(>F)
## off_bin      18  9.485  0.5270  61.632 < 0.0000000000000002 ***
## on_bin       5  3.827  0.7654  89.525 < 0.0000000000000002 ***
## off_bin:on_bin 73  1.158  0.0159   1.855      0.00033 ***
## Residuals    217  1.855  0.0086
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Error: Within
##              Df Sum Sq Mean Sq F value           Pr(>F)
## off_bin      17  34.11  2.0065 1224.08 <0.0000000000000002 ***
## on_bin        5   3.34  0.6674  407.15 <0.0000000000000002 ***
## off_bin:on_bin 84   8.50  0.1012   61.72 <0.0000000000000002 ***
## Residuals    3121   5.12  0.0016
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure S3: vertical coverage

```
##
## Error: on_target_species
##              Df Sum Sq Mean Sq F value           Pr(>F)
## off_bin      18   822   45.6   1.296           0.192
## on_bin        5  2018  403.7  11.459 0.000000000791 ***
## off_bin:on_bin 73  1431   19.6   0.556           0.998
## Residuals    217  7645   35.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Error: Within
##              Df Sum Sq Mean Sq F value           Pr(>F)
## off_bin      17   833   49.0   1.193           0.261
## on_bin        5  4211  842.2  20.498 <0.0000000000000002 ***
## off_bin:on_bin 84   866   10.3   0.251           1.000
## Residuals    3121 128225   41.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure S4: Local Alignment

Figure S5: BWA

86 NCBI strains

Figure 7A

```
##
## Error: accession
##              Df Sum Sq Mean Sq F value           Pr(>F)
## intra_species_ani 1  0.3018  0.3018   26.55 0.00000180988 ***
## between_species_ani 1  0.5409  0.5409   47.58 0.00000000112 ***
## Residuals        80  0.9095  0.0114
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Error: accession
##              Df Sum Sq Mean Sq F value           Pr(>F)
## intra_species_ani 1  0.1839  0.1838   57.13 0.0000000000589 ***
```

```
## between_species_ani  1 0.1205 0.12048  37.43 0.0000000330236 ***
## Residuals           80 0.2575 0.00322
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

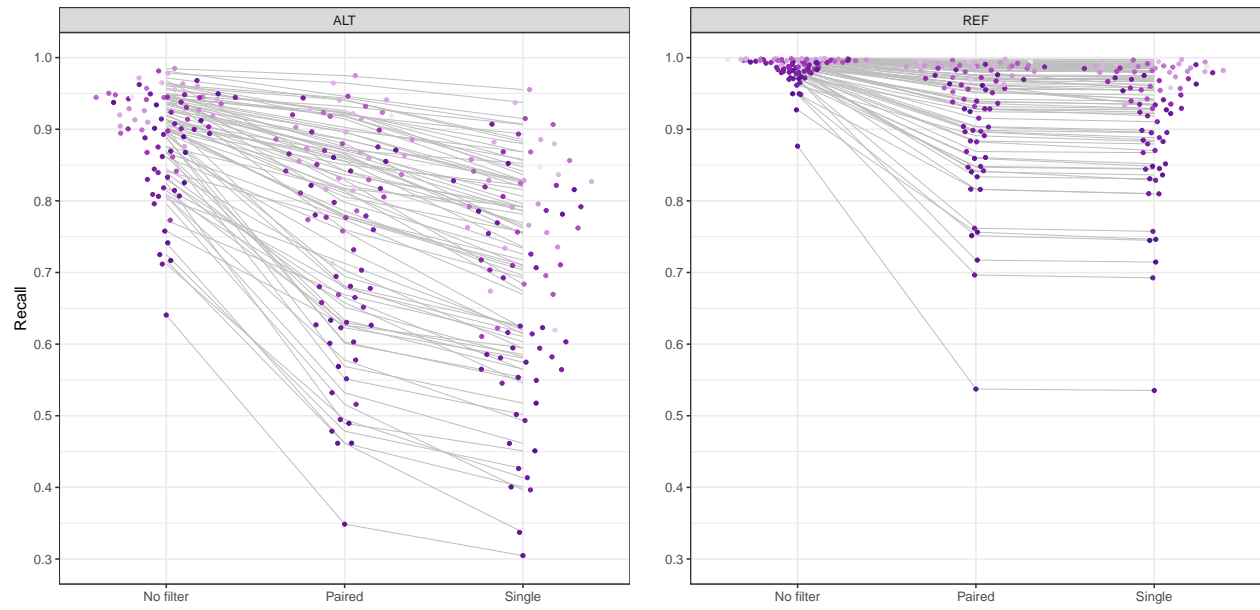
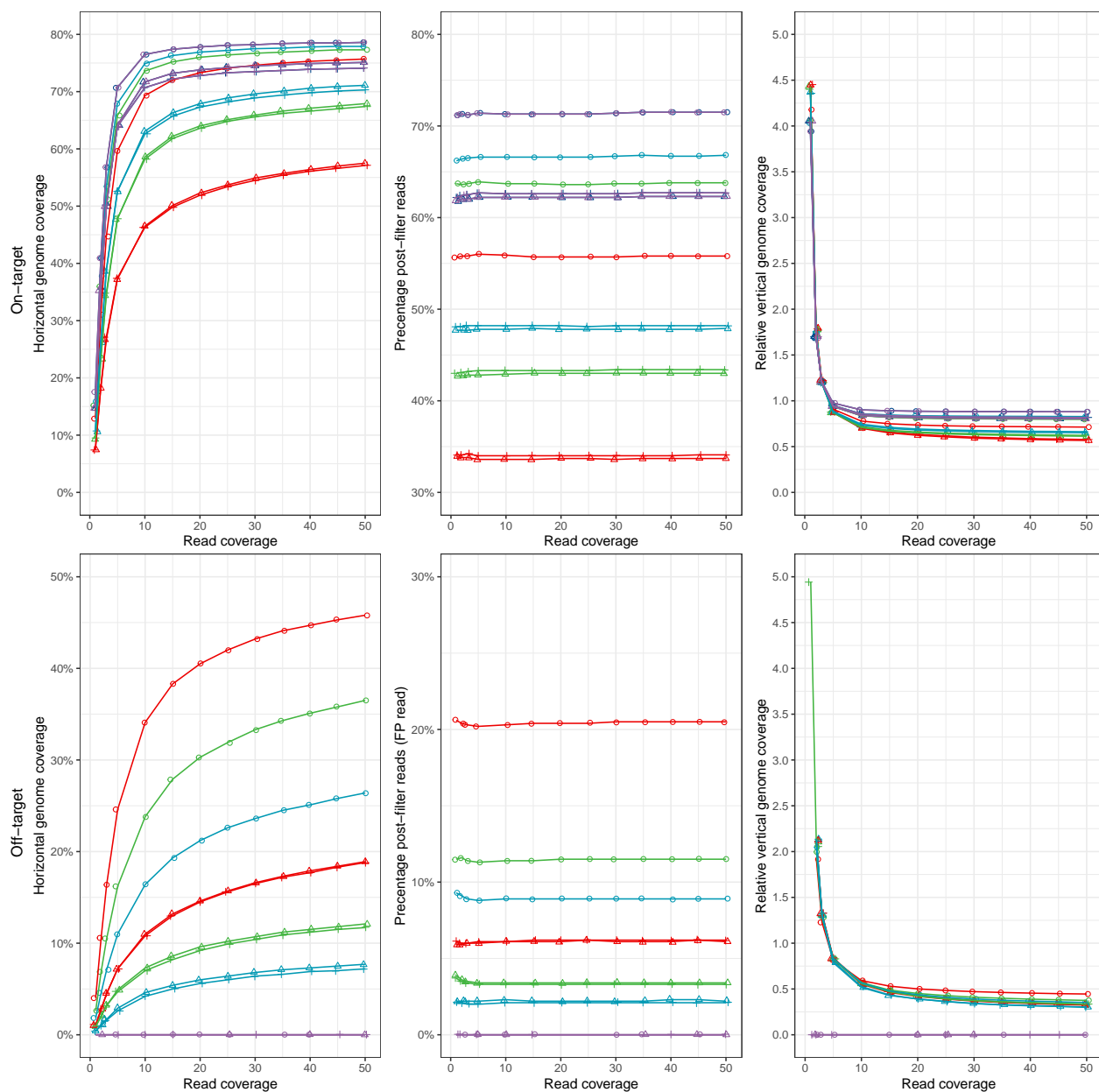


Figure 7B

Figure 7C

Figure 4B & 4C

Figure S1



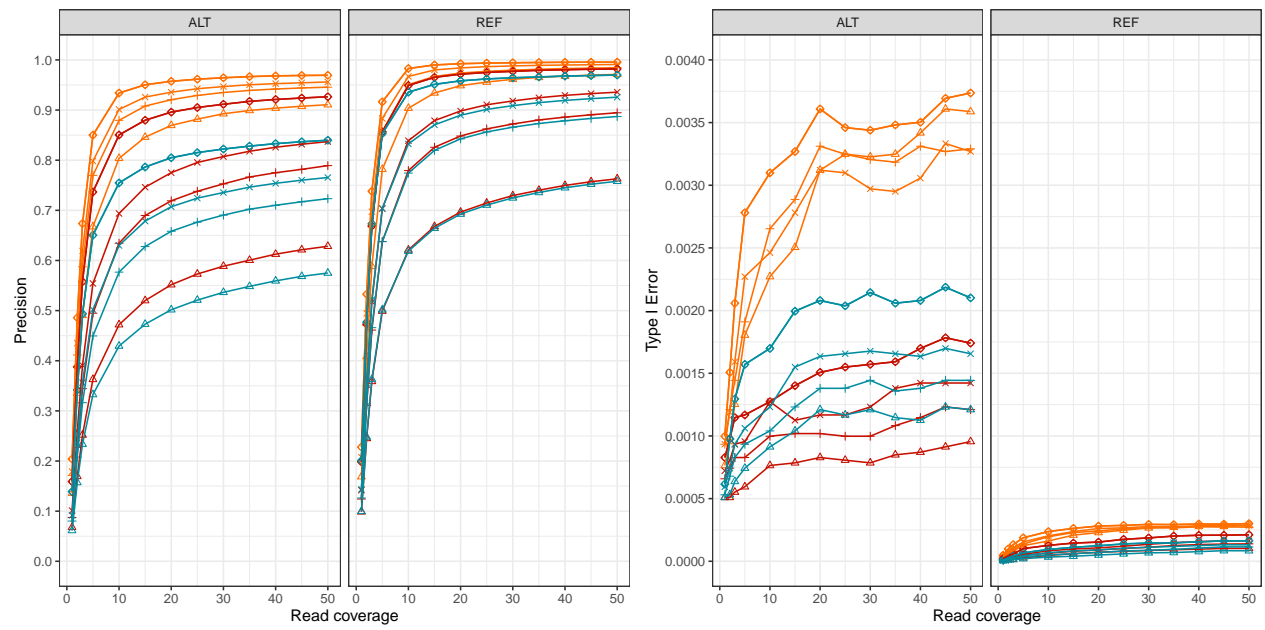


Figure S7 Runtime