Illumina vs. AVITI

thomas silvers

September 18, 2024

We performed two sequencing experiments on a library prepared from a pool of samples. In one experiment, Illumina sequencing was used; in the other, AVITI. To compare results, we restrict our analysis to *E. coli* samples from a single donor (Baby 2, B002).

1 Number of reads

We used `fastp` to trim, filter, and tally reads with pseudocode:

```
$ fastp {input FASTQs} \
  --cut_front \
  --cut_tail \
  --trim_poly_x \
  --cut_mean_quality 30 \
  --qualified_quality_phred 30 \
  --unqualified_percent_limit 10 \
  --length_required 50
```

Results were collected using MultiQC and parsed using custom code.

Results

- AVITI has $\sim 10^9$ reads, even after filtering (figure 1)

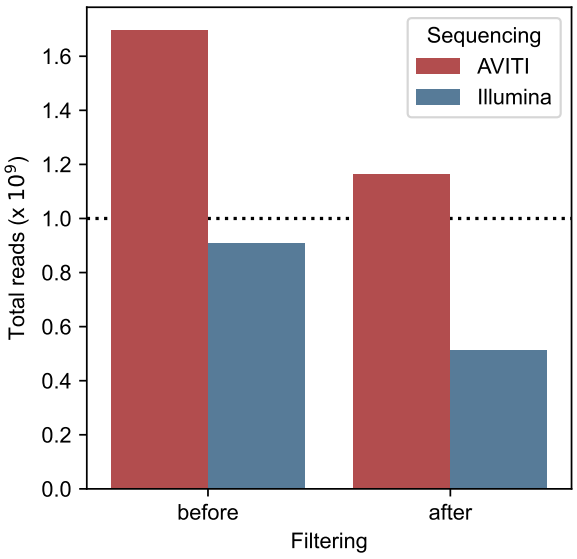

Figure 1: **Total number of reads** for AVITI or Illumina, before and after filtering with `fastp`.

2 Variant quality scores

We used `bcftools` to generate pile-ups and `bcftools stats` to extract variant quality scores with pseudocode:

```
$ bcftools mpileup --fasta-ref {reference} --min-BQ 20 {bam} \
    | bcftools call --output-type v --ploidy 1 --multiallelic-caller \
    | bcftools reheader --samples sample_name.list \
    | bcftools view --output-file {prefix}.vcf.gz --output-type z
$ tabix -p vcf -f {prefix}.vcf.gz
$ bcftools stats {prefix}.vcf.gz > {prefix}.bcftools_stats.txt
```

Results were collected using MultiQC and parsed using custom code.

Results

- AVITI has slightly higher, though comparable, variant quality scores compared with Illumina (figure 2)

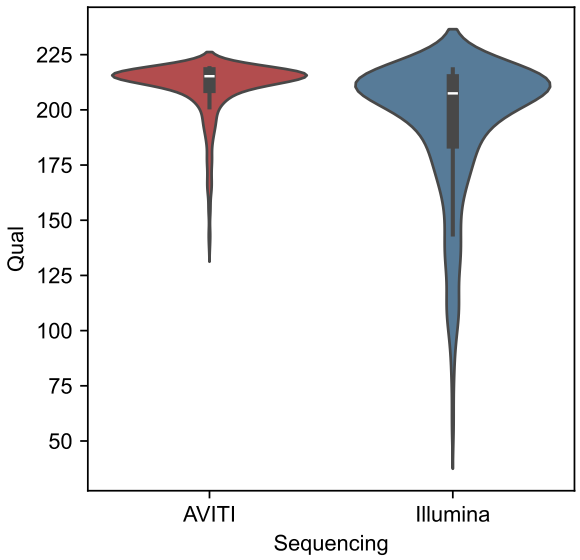

Figure 2: **Variant quality scores** for AVITI or Illumina.

3 Sequence typing

We used `srst2` to perform sequence typing for E. coli (MLST database name `Escherichia coli#1`) with pseudocode:

```
$ srst2 --input_pe {trimmed reads} --mlst_* '{Escherichia_coli#1}'
```

Results

- $\frac{250}{256} \approx 98\%$ agreement between AVITI and Illumina (table 1)
- AVITI ($\diamond$) has higher seq. depth at core genes used for sequence typing than Illumina ($\bullet$) (figure 3)
- (ST)73 is the dominant sequence type (figure 3)

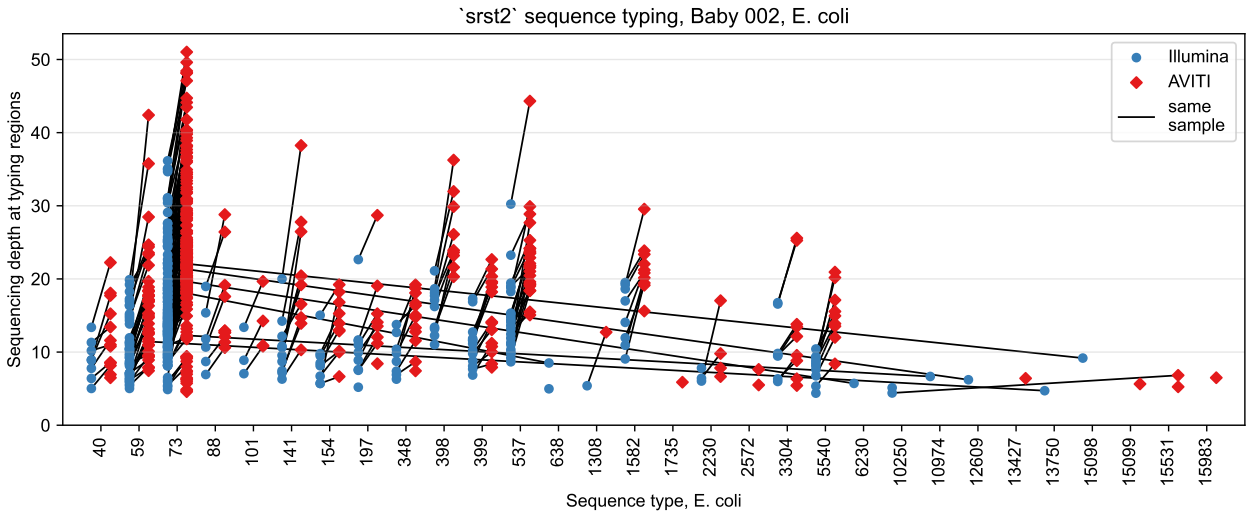

Figure 3: **Sequence typing results** for successful `srst2` sequence typing of identical samples (−) prepared using AVITI (⋄) or Illumina (•).

	Count
AVITI & Illumina (total)	398
AVITI, typed (total)	379
Illumina, typed (total)	331
AVITI & Illumina, typed (total)	312
AVITI & Illumina, typed (agree)	305

Table 1: **Summary of sequence typing results**, tallying the number of samples for different criteria. The top row provides the total number of samples; the bottom row provides the number of samples *successfully* sequence typed for *both* AVITI *and* Illumina *and agree* in the designated sequence type.

4 Reconstructing phylogenies of dominant STs

We used `raxml-ng` to infer the phylogeny of samples from the dominant sequence type with pseudocode:

```
$ raxml-ng --search1 --model GTR+G --outgroup {outgroup} --msa {msa}
```

We used 1000 randomly sampled positions with an ALT allele called in at least 2 samples that met the following criteria:

- No indels
- ≥ 3 reads supporting ALT on each strand
- MAF $\geq .95$
- QUAL ≥ 30

A final phylogeny would use full pseudogenomes and estimate bootstrap intervals. We find that a low depth cut-off is required, otherwise many Illumina (•) calls are filtered out (figure 4)

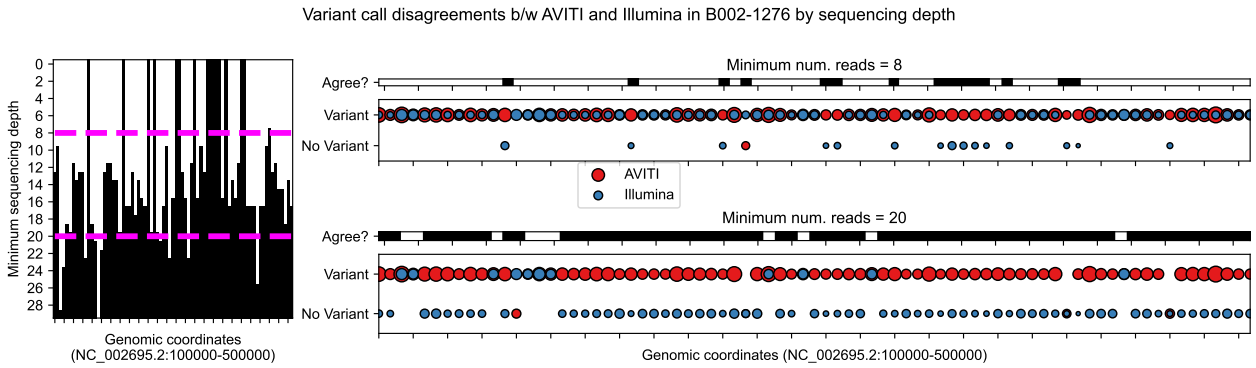

Figure 4: **Variant call disagreements** are mainly caused by combining the higher depth of AVITI and sequencing metrics thresholds. A case study of sample B002−1276 shows that AVITI (•) and Illumina (•) call nearly identical variants (agree $= \square$, disagree $= \blacksquare$) when the minimum required depth of a variant call (number of overlapping reads with the alt allele) is low (*see left plot*). For instance, at the cut-off $= 8$, there is agreement at all but 15 sites (*top-right plots*). Marker size is relative to sequencing depth; note the higher depth for AVITI (•). At a higher cut-off value (20), there are now numerous disagreements (*left and bottom-right plots*).

The two sample phylogenies, from AVITI and Illumina, show agreement (Robinson-Foulds agreement metric between two trees $= 0.658$; figure 5). See a combined tree from a random subset of samples at `https://transfer.mpiib-berlin.mpg.de/f/14297709` (a more legible version with some samples highlighted is at `https://transfer.mpiib-berlin.mpg.de/f/14297708`).

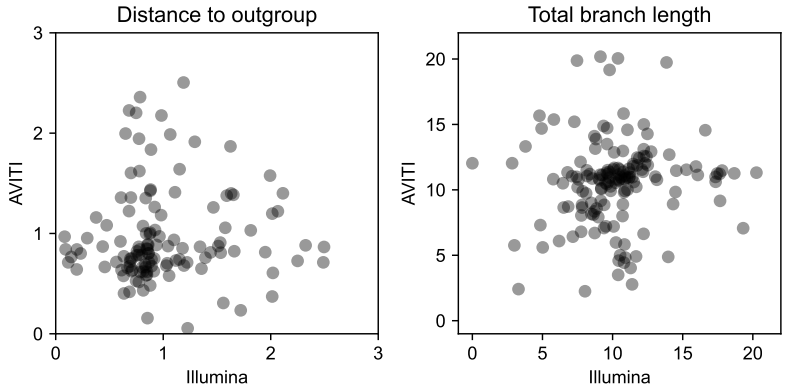

Figure 5: **Comparing tree metrics** for AVITI and Illumina from trees built using `raxml-ng`.

5 Appendix

Code to reproduce is available on GitHub at t-silvers/sequencing-comparison-aviti-illumina.