Summary of Bioinformatic Analyses and Approaches for Transposon Insertion Sequencing

Author: Douglas Rusch[a] and Emily Knebel[b]

[a]Center for Genomics and Bioinformatics, Indiana University, Bloomington, Indiana

[b]Division of Veterinary Pathobiology, University of Missouri, Columbia, Missouri

Library Demultiplexing. Raw reads were demultiplexed with bcl-convert[1] (v3.8.2-12-g85770e0b) to generate the reads files R1 and R2 for each Tn library. Transposon sequences spanning genomic DNA were limited to R2 and thus further processed for analysis.

Adapter and Quality Trimming, Transposon Sequencing Filtering: Low-quality bases and adapter sequences were removed using fastp[2] (v0.23.2; parameter: --dont_eval_duplication). The first 26 bps of read2 were verified to match internally to the transposon and were trimmed using a custom Perl script. Specifically, sequences with the following parameters and matching the below sequence were determined to be valid and the bolded bases were trimmed:

```
>edge /gc=68.7 /offset=0 /full_length=33 /nlength=1
NCCTGGGCACGCGACGACGCTCTTCCGATCTGG
```

After trimming, if read2 was smaller than 15 bps, it was discarded.

Read Alignment to the Reference Genome: High-quality reads were mapped to the latest *Brucella abortus* S19 reference genome (currently chromosomes I (NC_010740.1) and II (NC_010742.1)) and *Agrobacterium fabrum* str. C58 genome (circular chromosome (AE007869.2), linear chromosome (AE007870.2), plasmid Ti (AE007871.2), and plasmid AT (AE007872.2)) using bowtie2[3] (v2.3.5.1; parameters: --local --no-unal). The bowtie2 alignments were then parsed into stranded insertion sites using a custom Perl script. This information was then aggregated to get the total number of insertions, or total number of unique insertions by position, for each gene by strand using a custom Perl script.

Tn Insertion Data Output: Output files were generated containing the number of Tn-insertions associated with every base of the genome shifting. They were organized with the following different parameters:

60per	Only Tn-inserts in the 60 percent of the body of a gene are counted (so Tn's in the first and last 20% of the gene are excluded).
100per	All Tn-inserts are counted, regardless of where in the gene they are found.
Unique	If multiple Tn-inserts are seen at the same location, they are only counted once. This avoids potential effects of PCR-bias. The orientation of the Tn matters, so any position could have a value of 0, 1, or 2 (two indicating that a Tn was detected in both orientations).
All	All Tn-inserts are treated as independent and counts.

Data Shaping to Assess Tn Insertions and Effective Density: Custom R scripts were used to shape the transposon insertion data by library sample per gene per molecule position every 2 base pairs with extracted gff3 annotation information of locus tag, gene name, sequence type (e.g. coding, riboswitch), protein ID, and product description. This enabled aggregation of total coding

sequence sites by molecule position and counting the total Tn inserts within each site for each library. As described by Sharma et al. 2023[4], the unique Tn read counts per 60% of each gene were transformed into Shannon diversity indices, then taking the exponent of the index to identify the total number of effective sites, representing the number of equally abundant or frequent transposon insertion sites one would need to have the same resulting Shannon index values. From this, Tn effective density was calculated as effective sites / total sites to account for distribution of transposon insertions within a gene. Packages used included tidyverse, vegan, dplyr, permute, lmPerm, tidyR, stringR, and ggRepel. Full repository of R script from Sharma et al. is available on https://github.com/gayatri-101/TnDivA.

Statistical Analyses of Tn Effective Densities: Statistical analyses were conducted in R using custom scripts. Tn effective density values for each library condition were divided in pairs of a given condition over an appropriate control to determine a ratio of Tn effective density per gene. This was log2 transformed. Genes with a log2 value that is < -1 are considered to have significantly lower Tn effective density in the condition as compared to the control. Additionally, the effective density values per gene were regressed against each other in a linear regression model of the two conditions, followed by Cook's Distance outlier analysis which identified genes for which their removal significantly impacted the regression coefficient and overall fit of the regression model. A threshold value of (>4 / n), where n is the total number of genes, identified these outlier genes as influential on the regression fit, suggesting a strong influence of that gene in the comparison of effective densities. Packages used included tidyverse, dplyr, permute, lmPerm, tidyR, stringR, and ggRepel. Full repository of R script from Sharma et al. is available on https://github.com/gayatri-101/TnDivA.

References:

1. BCL Convert Support. https://support.illumina.com/sequencing/sequencing_software/bcl-convert.html.
2. Chen, S., Zhou, Y., Chen, Y., & Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 2018; 34(17), i884-i890.
3. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012; 9:357-359.
4. Sharma G, Zee PC, Zea L, Curtis PD. Whole genome-scale assessment of gene fitness of Novosphingobium aromaticavorans during spaceflight. *BMC Genomics*. 2023; 24(1):782. doi:10.1186/s12864-023-09799-z