St. Geme transposon library mapping project

John K. Everett, Ph.D.

October 2019, draft 1

This analysis describes the creation of a sequencing library created from *Kingella kingae* DNA samples provided by the St. Geme research group and the subsequent mapping of identified transpon insertions. A sequencing library was created by shearing genomic DNA and the attachment of adapter sequences followed by a nested PCR where the first set of primers bound within the body of the experimental transposon while the second set of primers bound within the transposon ITR segments. The library was sequenced with the Illumina MiSeq platform and transposon insertions were identified by searching for the 8 terminal ITR nucleotides followed by a TA sequence (CAACCTGTTA). The number of insertions recovered from each sample is shown in Table 1. Sequences were aligned to the representative *Kingella kingae* strain *Vir5453* (NCBI tax id: 1305785).

For the purpose of visualizing the data, the number of recovered insertions were normalized by dividing the number of sites within 10KB genomic blocks by the total number of sites recovered in each sample (Figure 1).

Figure 1. Visualization of recovered insertions within *Kingella kingae*.

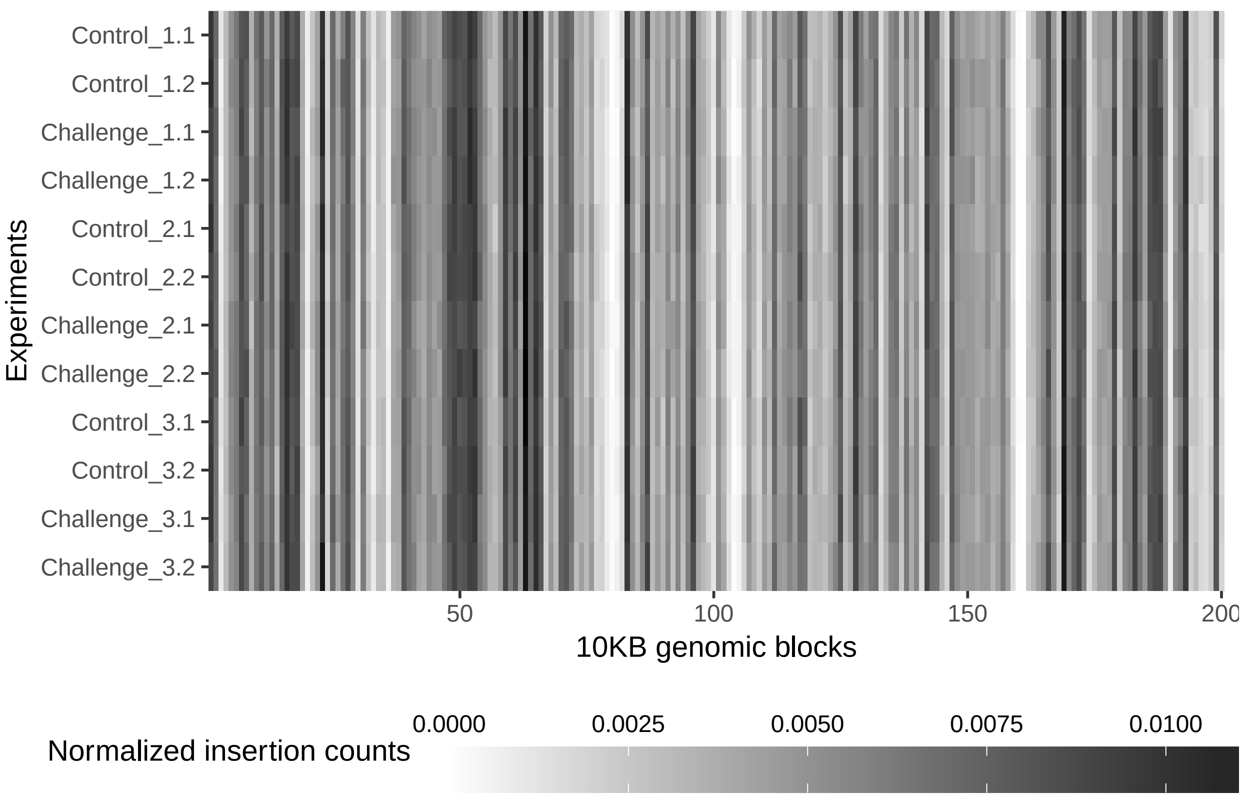


Table 1. Number of recovered insertions.

Sample	Insertions	Sample	Insertions
Control_1.1	12,604	Challenge_1.1	11,432
Control_1.2	11,472	Challenge_1.2	11,574
Control_2.1	12,034	Challenge_2.1	11,941
Control_2.2	11,963	Challenge_2.2	11,708
Control_3.1	11,929	Challenge_3.1	11,048
Control_3.2	11,313	Challenge_3.2	12,121

The number of insertions within transcription units (TUs) was gauged using two approaches. The first approach considered the number of insertions within each TU divided by the total number of insertions recovered in the sample. The second approach considered the total number of inferred cells (unique genomic break points) associated with insertions within each TU divided by the total number of inferred cells in the sample. The site count approach showed a fair degree of variation between techinical replicates (Figure 2) while the abundance method showed less variation between replicates and averaged samples (Figure 3).

Figure 2. Distriubtions of differences between technical replicate and averaged sample TUs using the site count approach.

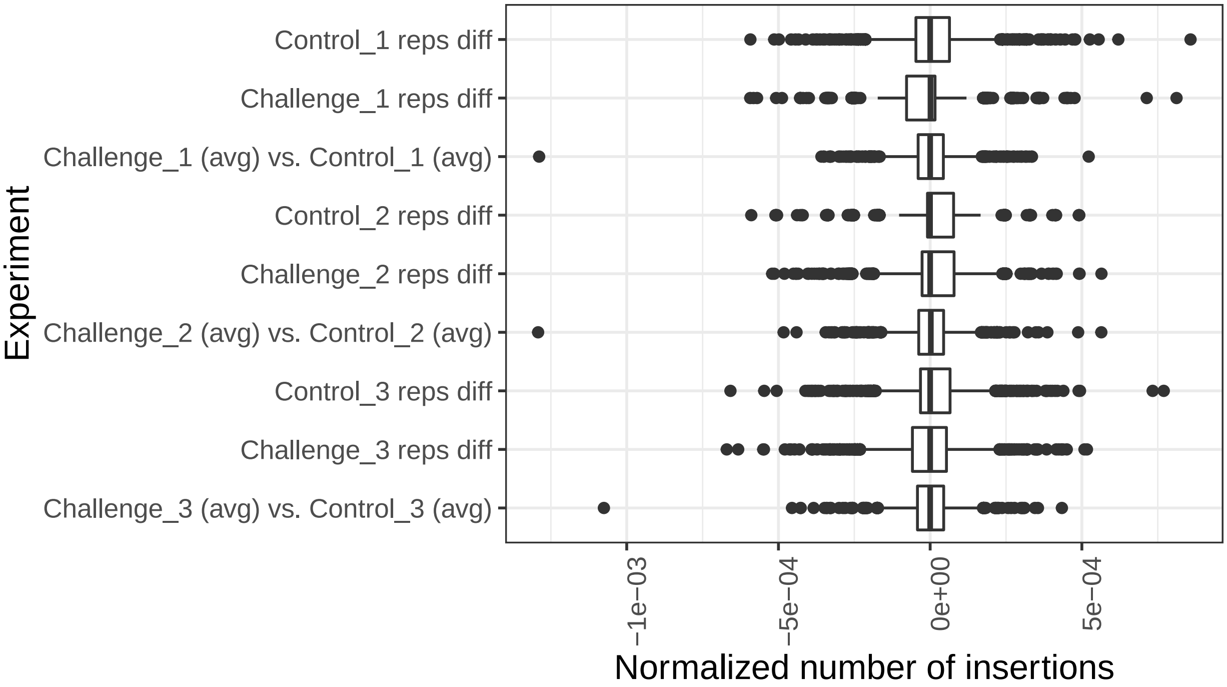
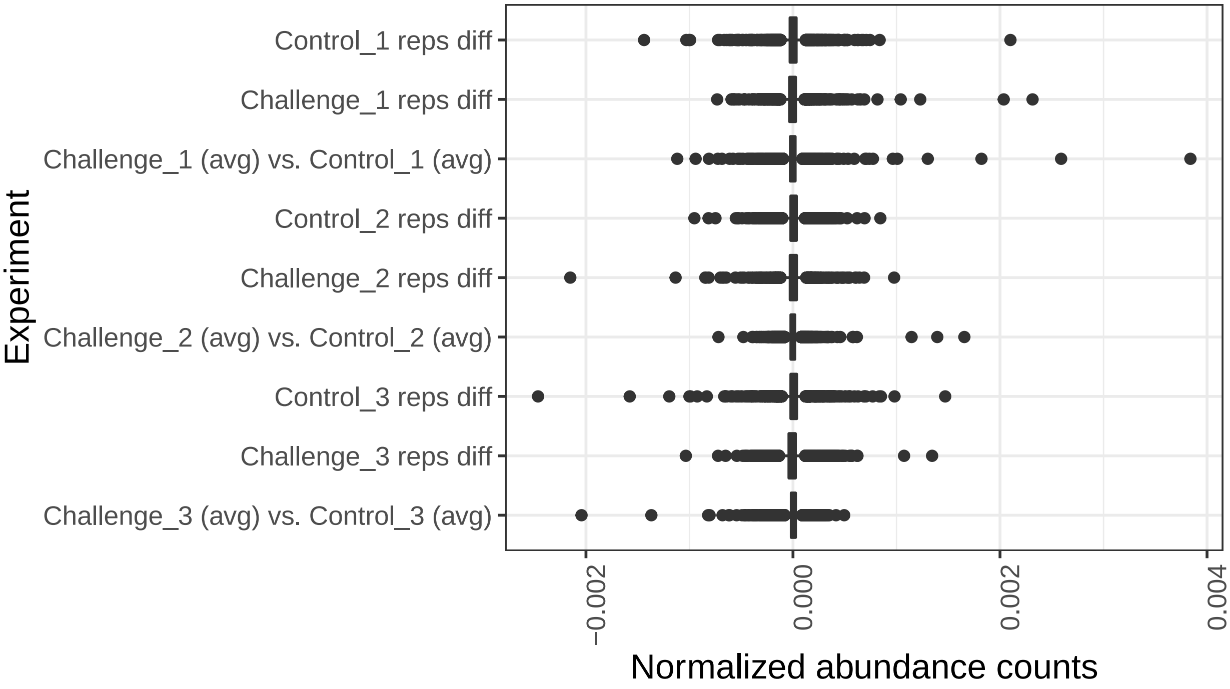


Figure 3. Distriubtions of differences between technical replicate and averaged sample TUs using the abundance approach.



Using the abundance approach, clear clustering of biological samples was found though there was not remarkable separation between control and challenge samples whithin biological sample clusters (Figures 4 & 5). The normalized site count approach provided less distinctive clustering (Supp. Figures S1 & S2).

Figure 4. Principle component analysis of all samples using the abundance normalization approach.

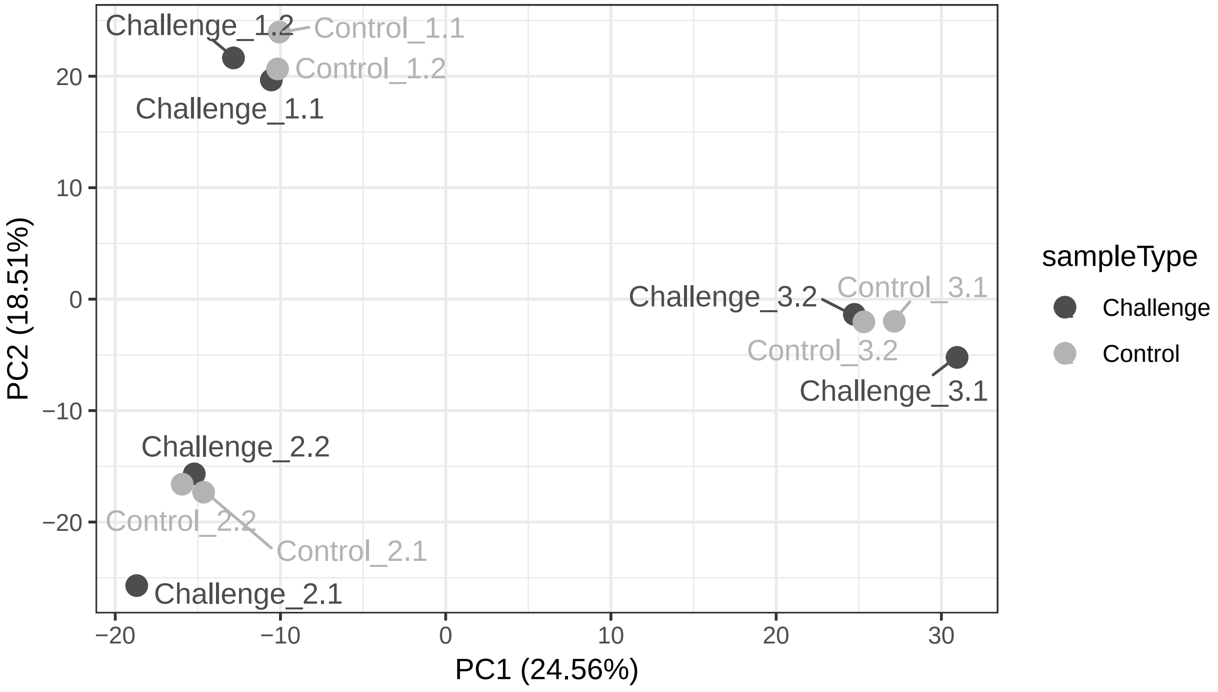
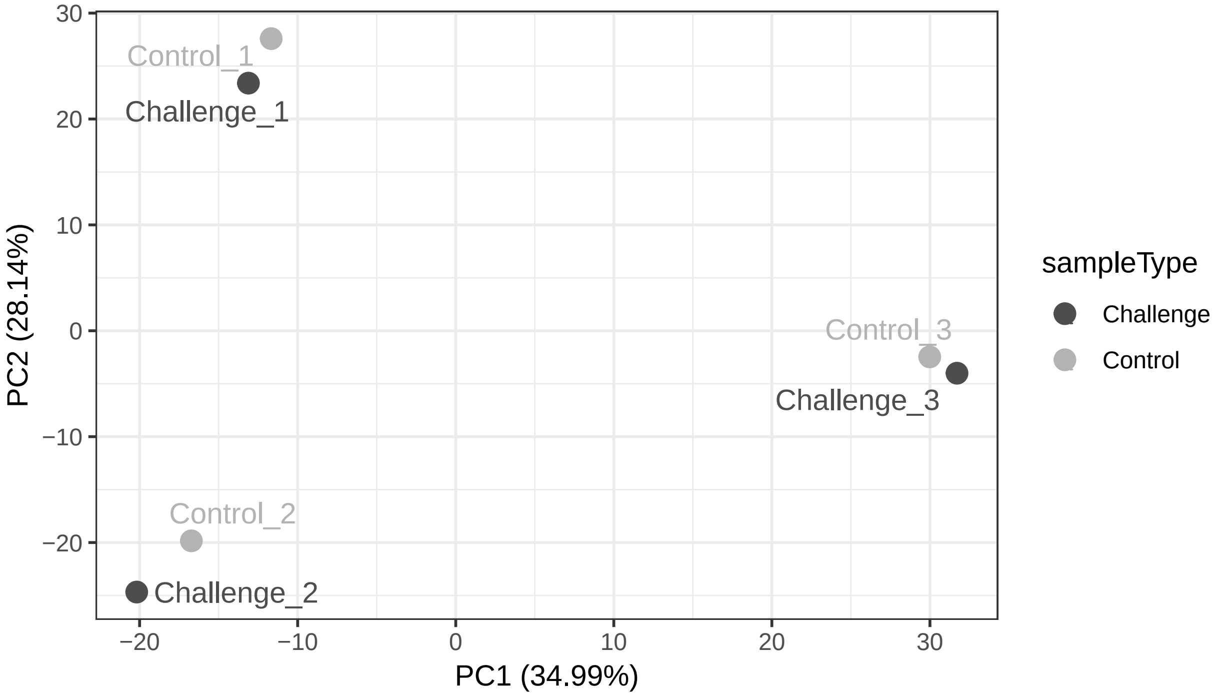


Figure 5. Principle component analysis of averaged technical replicates using the abundance normalization approach.



For each transcription unit, t-tests were used to test for differences between control and challenge insertion frequencies. Transcription units with significant uncorrected p-values are shown in Tables 2 & 3. Full gene tables are available on-line via this **link**.

Table 2. Genes with significant uncorrected p-values using the abundance correction method.

nearestFeature	geneDesc	pVal	pVal.adj	higherInChallenge
GeneID:34400097	hypothetical protein	0.0014811	1	FALSE
GeneID:34400613	ushA	0.0026656	1	TRUE
GeneID:34399909	membrane protein insertase YidC	0.0062998	1	TRUE
GeneID:34399835	hypothetical protein	0.0070110	1	TRUE
GeneID:34400681	hypothetical protein	0.0088456	1	FALSE
GeneID:34399182	TatD family deoxyribonuclease	0.0114130	1	TRUE
GeneID:34400394	type II toxin-antitoxin system death-on-curing family toxin	0.0141995	1	TRUE
GeneID:34399980	DUF2238 domain-containing protein	0.0167213	1	TRUE
GeneID:34400353	membrane protein	0.0172674	1	TRUE
GeneID:34399797	metW	0.0183453	1	TRUE
GeneID:34399888	transcriptional regulator	0.0259565	1	TRUE
GeneID:34400584	hypothetical protein	0.0277087	1	TRUE
GeneID:34399865	glutamate–ammonia ligase]-adenylyl-L-tyrosine phosphorylase...	0.0279244	1	TRUE
GeneID:34399910	DUF2892 domain-containing protein	0.0383931	1	FALSE
GeneID:34400955	hypothetical protein	0.0434916	1	FALSE
GeneID:34399479	DUF4198 domain-containing protein	0.0441051	1	TRUE
GeneID:34400788	IS1595 family transposase	0.0457648	1	FALSE

Table 3. Genes with significant uncorrected p-values using the normalized site count method.

nearestFeature	geneDesc	pVal	pVal.adj	higherInChallenge
GeneID:34400067	hypothetical protein	0.0000166	0.0274923	FALSE
GeneID:34399909	membrane protein insertase YidC	0.0001905	0.3156844	TRUE
GeneID:34399108	ABC transporter substrate-binding protein	0.0019507	1.0000000	TRUE
GeneID:34399950	DUF560 domain-containing protein	0.0025517	1.0000000	FALSE
GeneID:34400097	hypothetical protein	0.0037385	1.0000000	FALSE
GeneID:34400081	cysK	0.0157563	1.0000000	TRUE
GeneID:34400576	histidinol-phosphate transaminase	0.0193252	1.0000000	TRUE
GeneID:34400994	nucleic acid-binding protein	0.0246465	1.0000000	TRUE
GeneID:34400613	ushA	0.0253929	1.0000000	TRUE
GeneID:34400138	MFS transporter	0.0265198	1.0000000	FALSE
GeneID:34399868	DNA translocase FtsK	0.0283402	1.0000000	FALSE
GeneID:34399118	hypothetical protein	0.0347470	1.0000000	TRUE
GeneID:34399663	hypothetical protein	0.0354803	1.0000000	FALSE
GeneID:34399976	type II secretion system protein F	0.0410356	1.0000000	TRUE
GeneID:34400056	ABC transporter ATP-binding protein	0.0448511	1.0000000	TRUE
GeneID:34400788	IS1595 family transposase	0.0477979	1.0000000	FALSE
GeneID:34400353	membrane protein	0.0479632	1.0000000	TRUE
GeneID:34399400	hypothetical protein	0.0485404	1.0000000	TRUE

Supplimental

Figure S1. Principle component analysis of all samples using the normalized site count method.

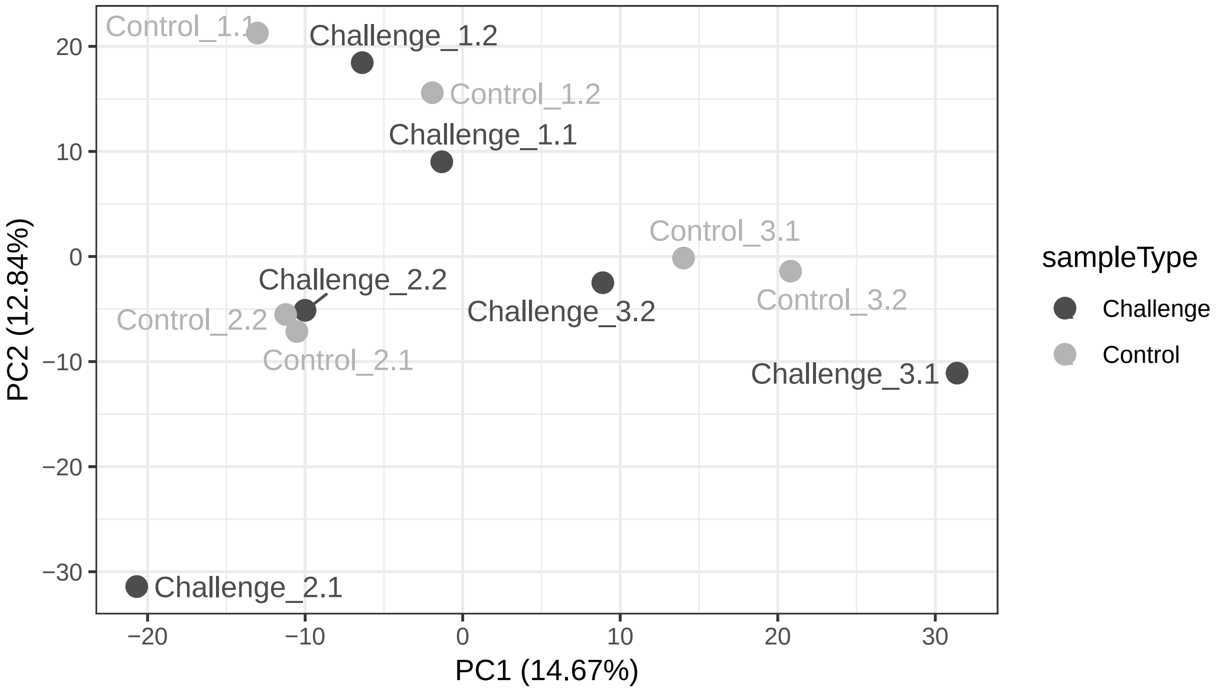
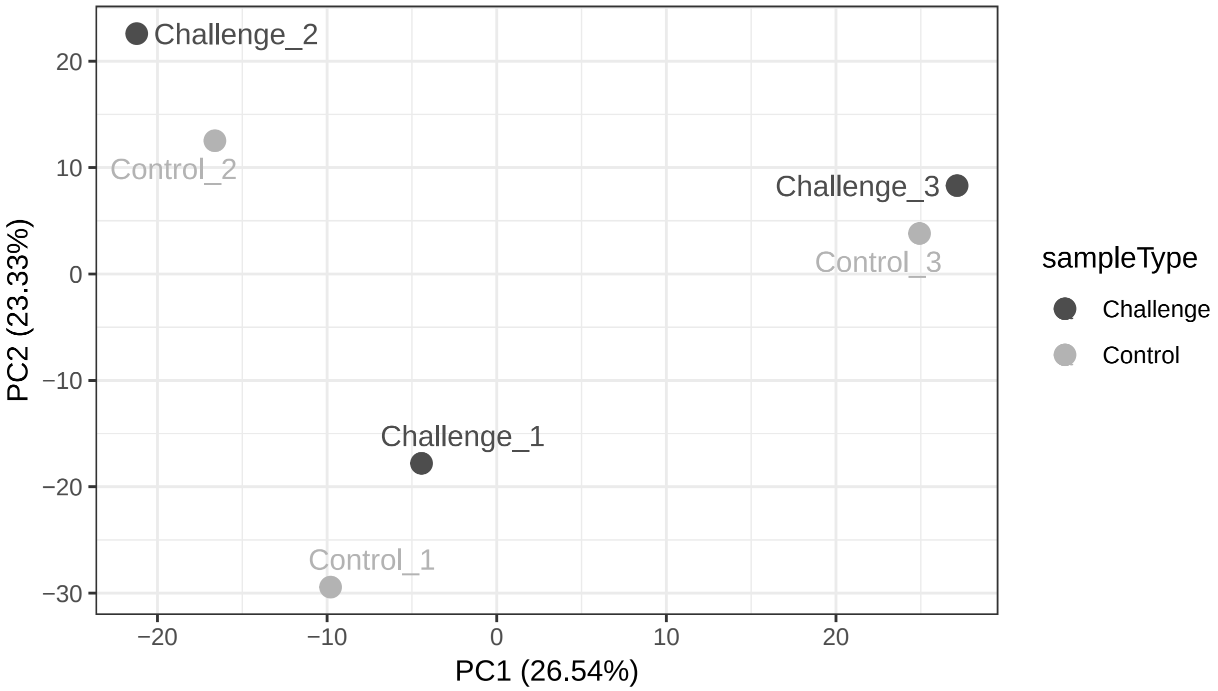


Figure S2. Principle component analysis of averaged technical replicates using the normalized site count method.


Challenge_2
Control_2
Challenge_3
Control_3
Challenge_1
Control_1
PC2 (23.33%)
PC1 (26.54%)
sampleType
Challenge
Control