St. Geme transposon library mapping project

John K. Everett, Ph.D.

November 2019, draft 2

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 *KWG1*.

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*.

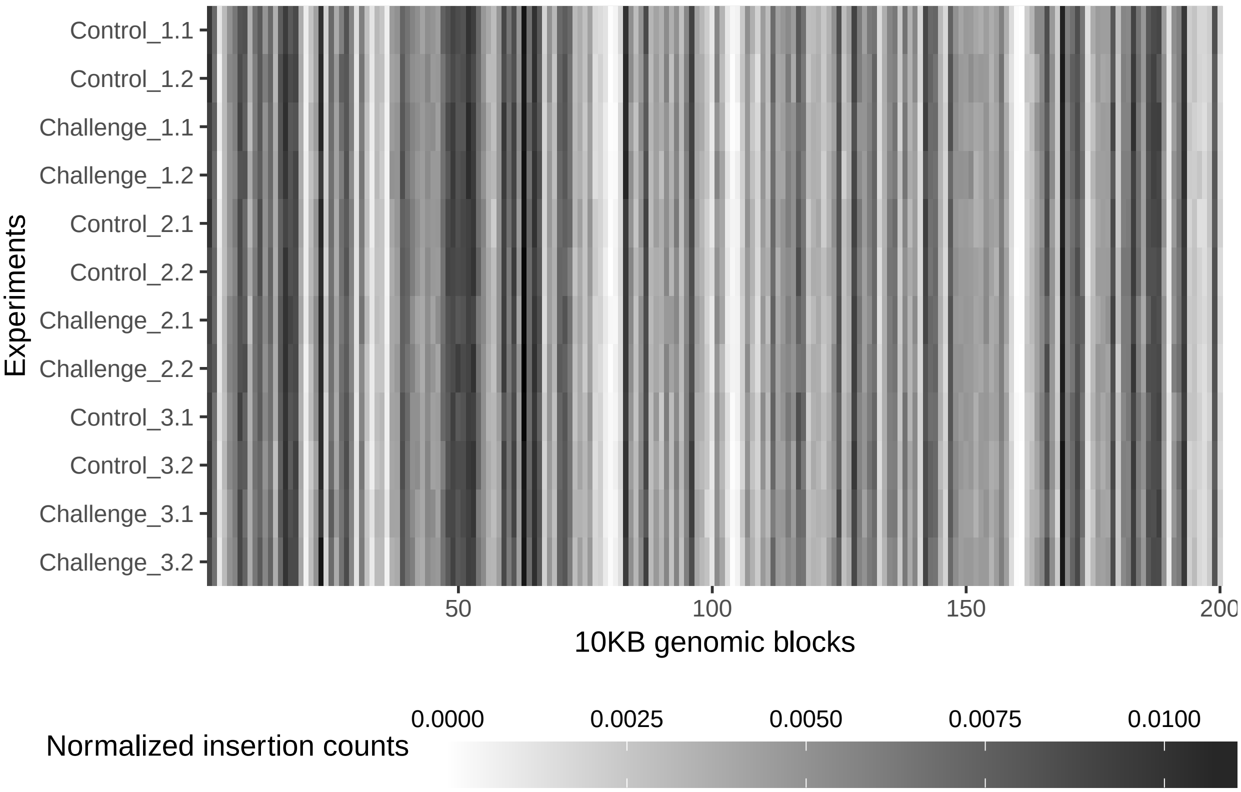


Table 1. Number of recovered insertions.

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.

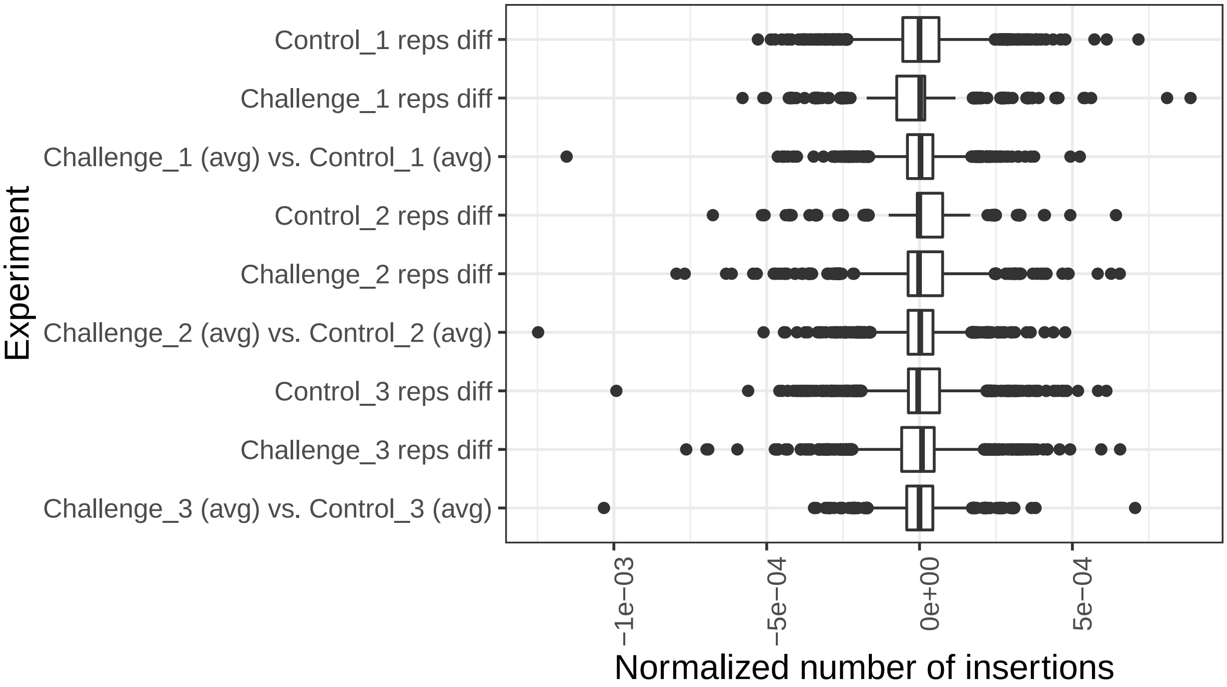
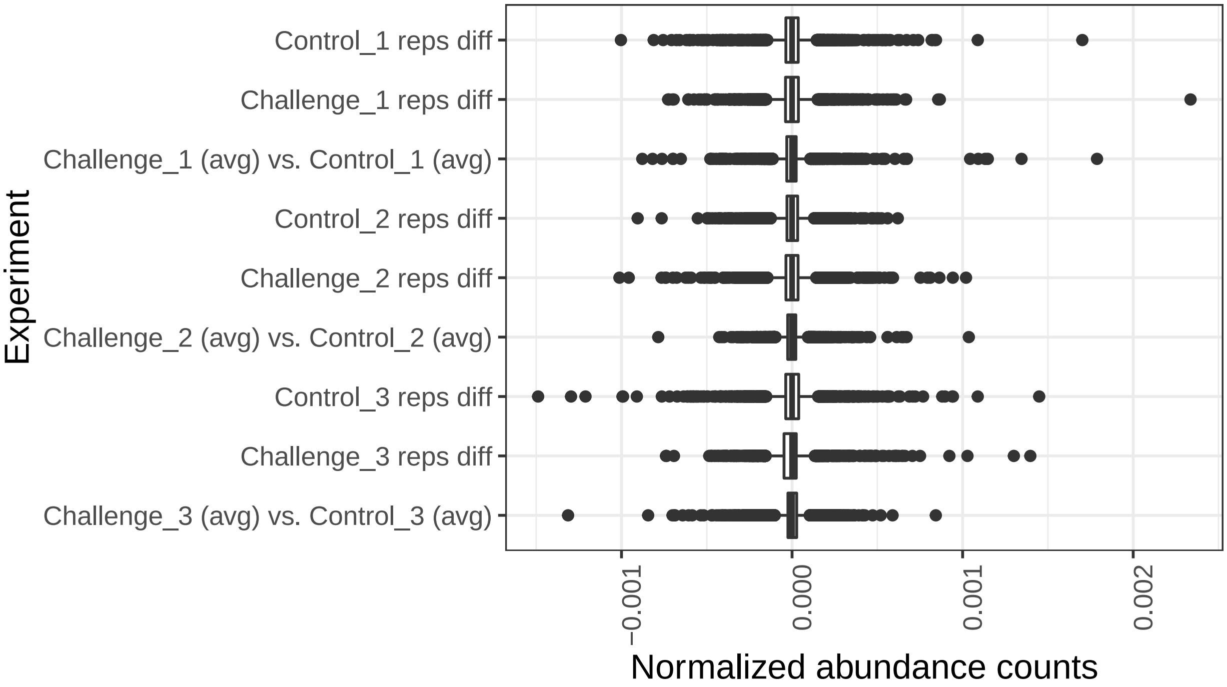


Figure 3. Distriubtions of differences between technical replicate and averaged sample TUs using the abundance approach.



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.

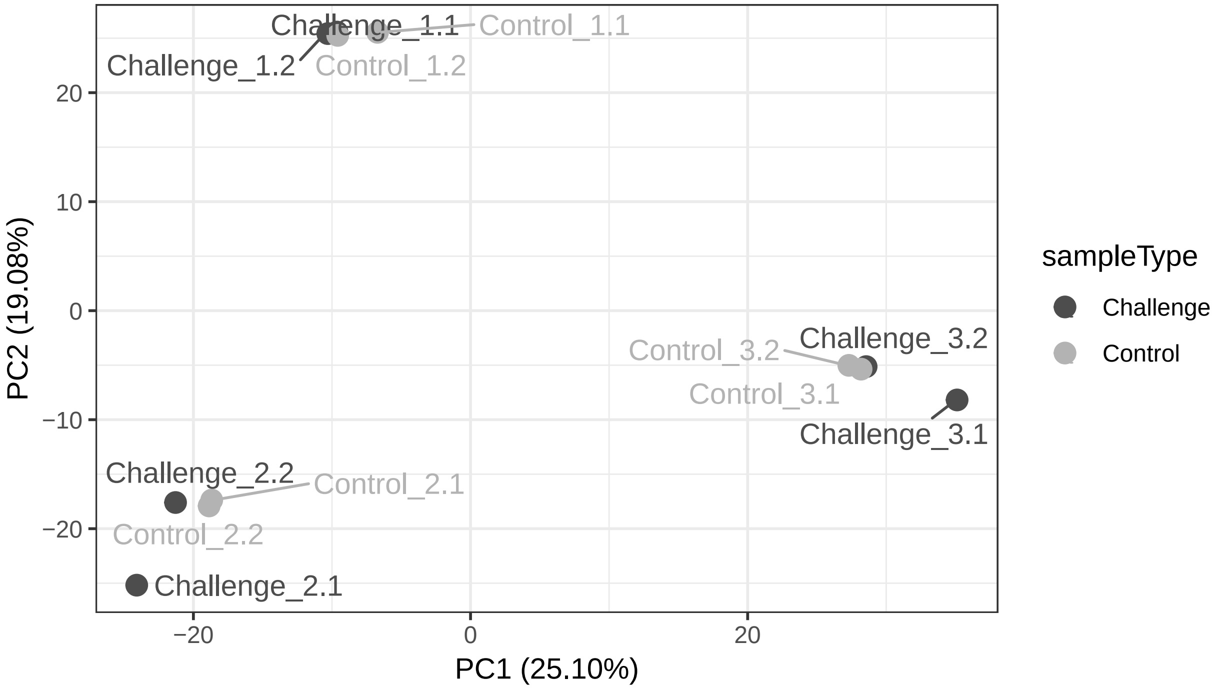
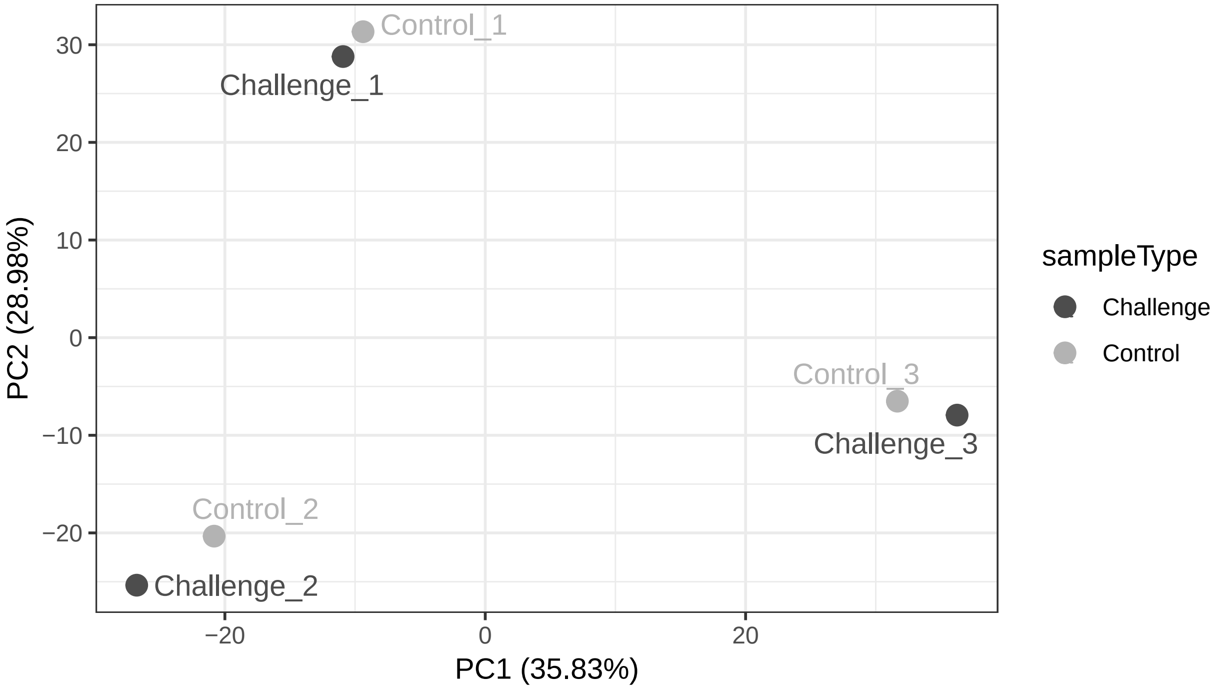


Figure 5. Principle component analysis of averaged technical replicates using the abundance normalization approach.



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. Gene names followed by ‘PRO’ represent potential promotor regions 1-50 NTs upstream of genes. 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
KKKWG1_RS06005	histidinol-phosphate transaminase	0.0006560	1	FALSE
KKKWG1_RS02565	tRNA-Pro	0.0055934	1	TRUE
KKKWG1_RS06150	hypothetical protein	0.0062688	1	TRUE
KKKWG1_RS06130	replicative DNA helicase	0.0086197	1	FALSE
KKKWG1_RS04725	hypothetical protein	0.0098742	1	FALSE
KKKWG1_RS08530	hypothetical protein	0.0159679	1	FALSE
KKKWG1_RS06015	ABC transporter ATP-binding protein	0.0201837	1	TRUE
KKKWG1_RS07690	hypothetical protein	0.0237803	1	TRUE
KKKWG1_RS01295	rpmA	0.0300514	1	TRUE
KKKWG1_RS04105	TrbM protein	0.0355494	1	TRUE
KKKWG1_RS10045 PRO	hypothetical protein PRO	0.0370049	1	FALSE
KKKWG1_RS07225	4-hydroxybenzoate octaprenyltransferase	0.0412672	1	FALSE
KKKWG1_RS03140 PRO	lipopolysaccharide heptosyltransferase II PRO	0.0429499	1	FALSE
KKKWG1_RS07920	type I deoxyribonuclease HsdR	0.0447194	1	FALSE

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

nearestFeature	geneDesc	pVal	pVal.adj	higherInChallenge
KKKWG1_RS06005	histidinol-phosphate transaminase	0.0022873	1	FALSE
KKKWG1_RS03795	hypothetical protein	0.0040045	1	FALSE
KKKWG1_RS08685	restriction endonuclease subunit M	0.0076240	1	FALSE
KKKWG1_RS03420	carbamoyl-phosphate synthase small subunit	0.0087569	1	FALSE
KKKWG1_RS06010	homoserine kinase	0.0098731	1	FALSE
KKKWG1_RS08530	hypothetical protein	0.0129918	1	FALSE
KKKWG1_RS02490	SPOR domain-containing protein	0.0183272	1	FALSE
KKKWG1_RS06015	ABC transporter ATP-binding protein	0.0186272	1	TRUE
KKKWG1_RS02780	tpx	0.0210730	1	FALSE
KKKWG1_RS04200	family 2 glycosyl transferase	0.0215421	1	FALSE
KKKWG1_RS05735	gltB	0.0245428	1	FALSE
KKKWG1_RS06730	homoserine dehydrogenase	0.0257945	1	FALSE
KKKWG1_RS03180	hypothetical protein	0.0268371	1	FALSE
KKKWG1_RS10065	methyltransferase	0.0284097	1	TRUE
KKKWG1_RS09870	DNA polymerase III subunit delta’	0.0316320	1	TRUE
KKKWG1_RS01440	hypothetical protein	0.0347470	1	TRUE
KKKWG1_RS08450	hypothetical protein	0.0351005	1	FALSE
KKKWG1_RS06000 PRO	acyl-CoA dehydrogenase PRO	0.0361090	1	FALSE
KKKWG1_RS07775	hypothetical protein	0.0415417	1	FALSE
KKKWG1_RS02970	xerD	0.0436035	1	TRUE
KKKWG1_RS04105	TrbM protein	0.0436854	1	TRUE
KKKWG1_RS00025	ABC transporter ATP-binding protein	0.0442441	1	FALSE
KKKWG1_RS08445	hypothetical protein	0.0449023	1	TRUE
KKKWG1_RS04700	traS protein	0.0465510	1	TRUE
KKKWG1_RS06065	enterobactin receptor FetA	0.0468414	1	TRUE
KKKWG1_RS00355	lipid II flippase MurJ	0.0469789	1	TRUE
KKKWG1_RS01790	membrane protein	0.0493780	1	FALSE

Supplimental

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

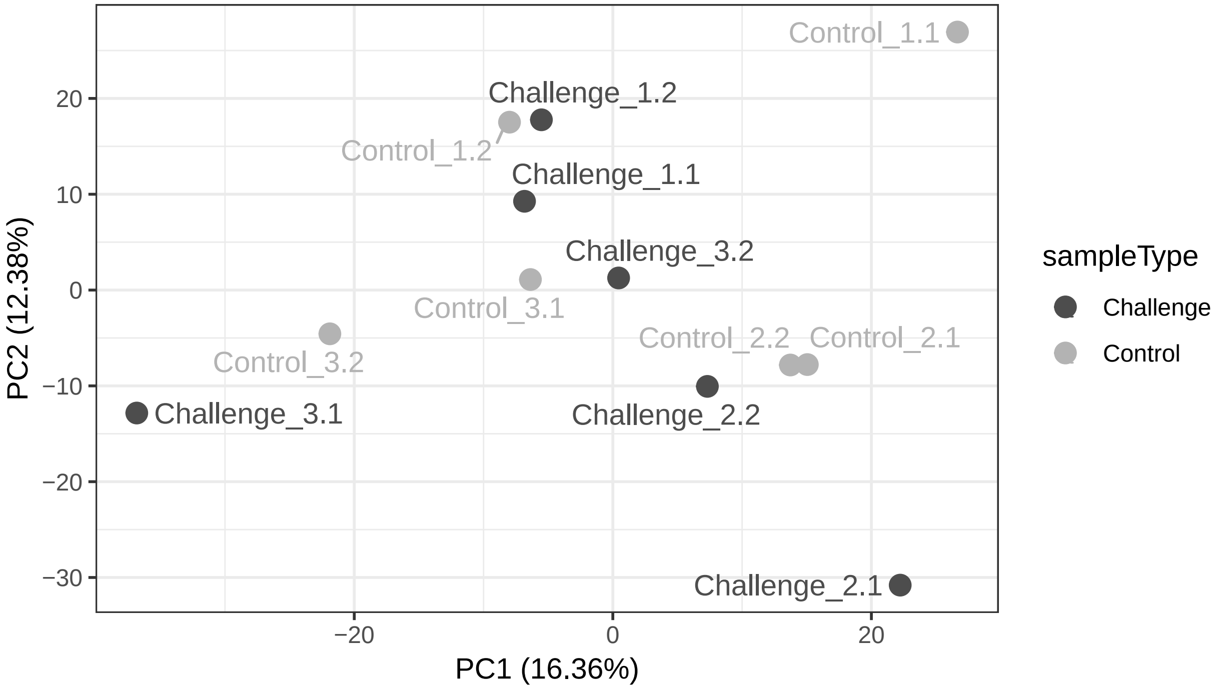
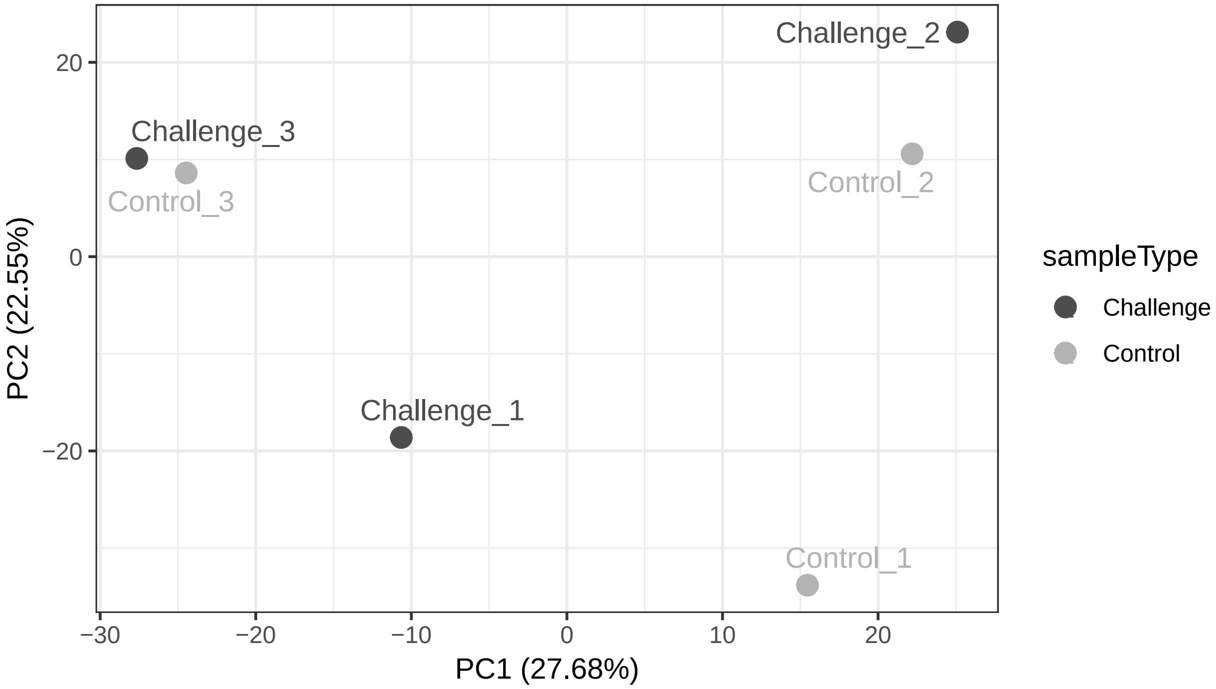


Figure S2. Principle component analysis of averaged technical replicates using the normalized site count method.


Challenge_2
Challenge_3
Control_3
Control_2
Challenge_1
Control_1
PC1 (27.68%)
PC2 (22.55%)
sampleType
Challenge
Control