admera health

Admera Health
126 Corporate Boulevard • South Plainfield, NJ 07080
+1-908-222-0533 • custom-services@admerahealth.com • www.admerahealth.com
CLIA ID 31D2038676

CRISPRseq Indel Calling Analysis Report

Analysis Schema

1. Quality Assessment and Sequence Alignment
2. Alignment Summary
3. CIGAR Calling (Edit Detection)
4. Nucleotide Composition and Indel Distribution
5. Summary Table
6. Appendix
7. Citation
8. Contact Us

Quality Assessment and Sequence Alignment

We utilized the crisprseq pipeline from nf-core, implemented in Nextflow, for our analysis. Initially, the paired fastq files were merged using the Pear tool, and adapter trimming was performed using Cutadapt. As part of the preprocessing step, low-quality bases were masked using the seqtk tool. The resulting clean reads were then mapped to the provided reference sequence using Minimap2. The resulting alignment files are available in the format <sample_ID>.bam under directory **02.BamFiles**. Subsequently, quality assessment was conducted using FastQC tools. For a comprehensive understanding of the FastQC report, please refer to http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. The FastQC visualization and summary files are available in the format <sample_ID>.fastqc.zip and <sample_ID>.fastqc.html, respectively. The QC reports can be found under directory "**01.FastqQualityCheck**"

Alignment Summary

An example count summary for different preprocessing steps can be found in Table 1. For individual samples, similar tables are available as <sample_ID>_alignment_summary.csv. Additionally, Figure 1 presents these statistics in a plot format. To explore an interactive Pie chart for each individual sample, please refer to the file <sample_ID>_reads.html under directory **03.Results**. The summary of the edits for the project can be identified as "alignment_summary_combined.csv"

Table 1 Read counts after various steps.

Type	count
raw-reads	159075 (100.0%)
merged-reads	158680 (99.8%)
reads-with-adapters	157503 (99.3%)
quality-filtered-reads	156767 (98.8%)
aligned-reads	152756 (97.3%)

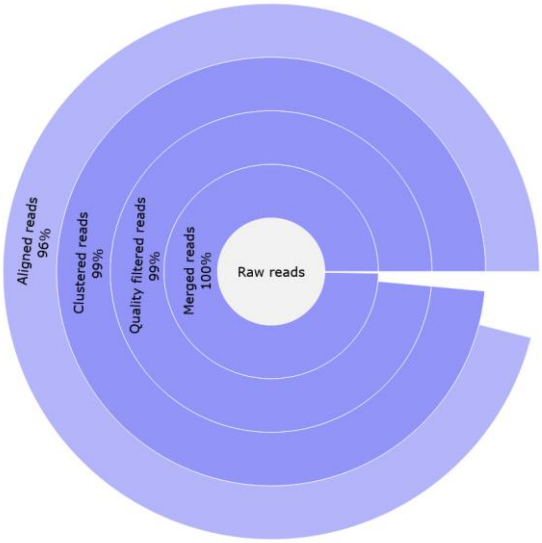

Figure 1 Pie charts with Read counts after various pre-processing steps.

CIGAR Calling (Edit Detection)

The crisprseq pipeline utilizes a custom R-script to identify different types of edits. Figure 2 presents a pie chart depicting the distribution of edit types. Reads that do not exhibit any edits are classified as wild type (WT), while those with edits are categorized as indels. The indels are further categorized into deletions, insertions, and delins (deletion + insertion). Deletions and insertions can occur in either the in-frame or out-of-frame regions. Within the files <sample_ID>_edition.html, you can explore interactive Pie charts displaying the distribution of different edit types across all samples. The corresponding data providing more detailed information on the edits can be accessed in <sample_ID>_edits.csv files under directory **03.Results**. The summary of the edits for the project can be identified as "edit_combined.csv"

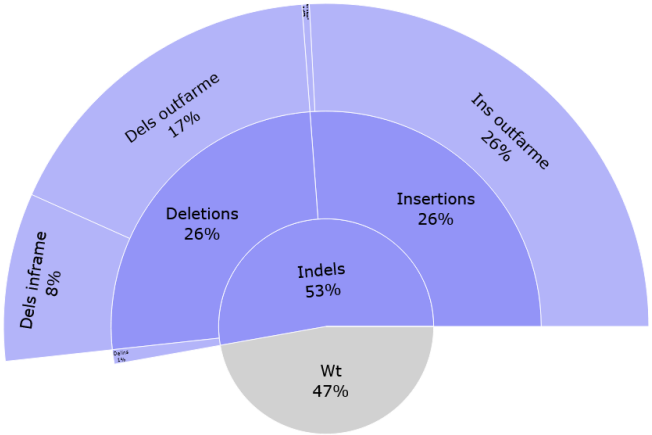

Figure 2 Pie chart depicting the distribution of edit types

Nucleotide Composition and Indel Distribution

<sample_ID>_subs-perc.csv is the csv file containing the percentage of each nucleotide found for each reference position, Figure 3 below shows the snippet of the nucleotide composition. The full version of the plot can be found <sample_ID>_nucleotide_comp.pdf under directory 03.Results These plots serve as a visual guide for identifying and understanding patterns and variations in the nucleotide composition of the CRISPRseq data.

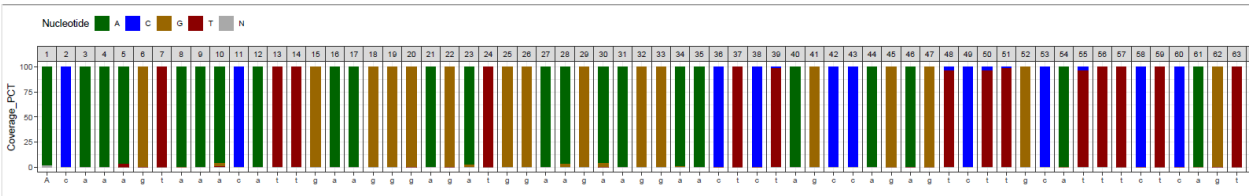

Figure 3 Nucleotide composition plot.

Figure 4 presents the length distribution of Indels, and the corresponding csv file containing information for all reads can be found as <sample_ID>_indels.csv. This file provides comprehensive details for each read related to Indel, including the edit type, edit start position and length, presence within the common edit window, frequency, percentage, pattern, surrounding nucleotides in case of insertions, protospacer cut site, sample ID, number of aligned reads, and the count of reads with and without a template modification. The distribution plots for the individual sample can be identified as <sample_ID>_indel_dist.pdf under directory **03.Results**

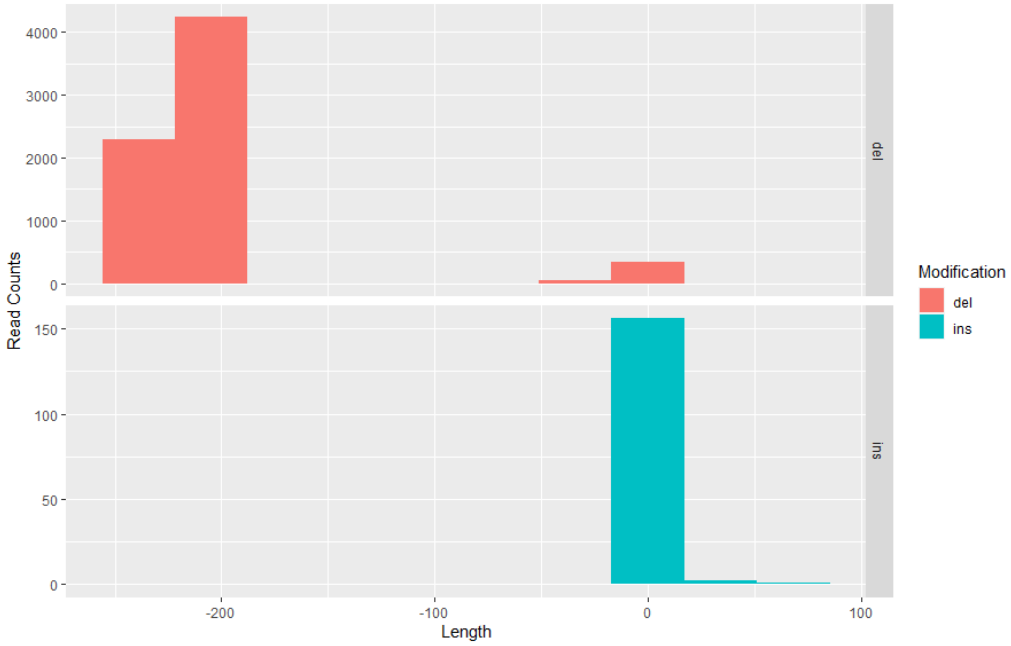

Figure 4 Indel Length distribution Plot.

Summary Table

In conclusion, as illustrated in Table 2, data from multiple samples from Project are consolidated into a unified table. This table encompasses information regarding the percentage of reads aligned to the given reference, the count of wild-type reads, and the corresponding wild-type and Indel percentages. Additionally, Moreover, Indels are classified into different categories, encompassing Delins, Insertions, and their respective locations in both the in-frame and out-of-frame contexts.

Table 2 : Indel Summary

Sample	aligned_reads	wt_reads	Wt	Indels (%)	Delins	Insertions	Ins inframe	Ins outfarme	Deletions	Dels inframe	Dels outfarme
Sample-01	152756 (97.3%)	68112	47.2	52.8	2.04	49.62	1.41	98.59	48.34	32.98	67.02
Sample-02	157962 (98.5%)	155042	98.4	1.6	3.17	47.25	16.25	83.75	49.58	7.34	92.66
Sample-03	146203 (97.4%)	72865	52.28	47.72	1.92	49.96	1.41	98.59	48.12	30.44	69.56
Sample-04	143798 (97.2%)	97365	69.85	30.15	1.91	49.15	2.28	97.72	48.94	26.71	73.29
Sample-05	146899 (99.6%)	144685	98.72	1.28	4.59	40.13	16.89	83.11	55.28	5.98	94.02
Sample-06	154504 (99.8%)	152117	98.68	1.32	4.83	38.92	17.85	82.15	56.26	5.34	94.66
Sample-07	140995 (99.8%)	138753	98.67	1.33	4.07	35.9	22.35	77.65	60.03	7.31	92.69
Sample-08	130015 (99.3%)	98911	78.36	21.64	3.2	50.41	3.14	96.86	46.39	27.69	72.31
Sample-09	137083 (99.4%)	109846	82.04	17.96	4.18	51.91	3.37	96.63	43.92	29.58	70.42
Sample-10	143725 (99.2%)	117940	84.06	15.94	3.37	51.47	3.93	96.07	45.17	28.81	71.19
Sample-11	128901 (99.4%)	121539	94.7	5.3	1.69	52.57	4.09	95.91	45.73	14.83	85.17
Sample-12	138701 (99.5%)	133372	96.47	3.53	2.29	58.15	5.77	94.23	39.55	14.12	85.88
Sample-13	140422 (99.5%)	138382	98.78	1.22	4.14	39.28	24.93	75.07	56.59	7.42	92.58
Sample-14	137941 (99.5%)	135884	98.73	1.27	3.54	40.22	25.82	74.18	56.25	4.97	95.03
Sample-15	146329 (99.7%)	144151	98.71	1.29	3.76	39.02	20.11	79.89	57.21	4.82	95.18
Sample-16	132445 (99.5%)	108445	83.87	16.13	2.83	50.17	3.18	96.82	47	33.18	66.82
Sample-17	143509 (99.4%)	116574	82.87	17.13	3.93	53.45	3.24	96.76	42.63	26.85	73.15
Sample-18	142459 (99.5%)	126454	89.96	10.04	3.56	52.13	3.49	96.51	44.32	27.36	72.64
Sample-19	148234 (98.9%)	145785	98.61	1.39	3.02	41.29	19.34	80.66	55.7	4.02	95.98
Sample-20	148882 (98.8%)	146240	98.44	1.56	3.59	41.76	14.7	85.3	54.65	6.72	93.28

Appendix

The following software used in the analysis pipeline

Software	Version
cutadapt	3.4
Minimap2	2.24-r1122
PEAR	0.9.6
Samtools	1.17
seqtk	1.3-r106
Nextflow	23.04.1
nf-core/crisprseq	2.0.0
Reference Genome	Custom (Provided)
gRNA targeting sequence	Custom (Provided)