pRESTO Report: IgSeqBX1

Contents

1 Overview of Data

Date: Today

Author: Olga Botvinnik

Description: B cells

Run: Run1

Sample: IgSeqBX1

pRESTO Version: 0.5.4

2 Summary of Processing Steps

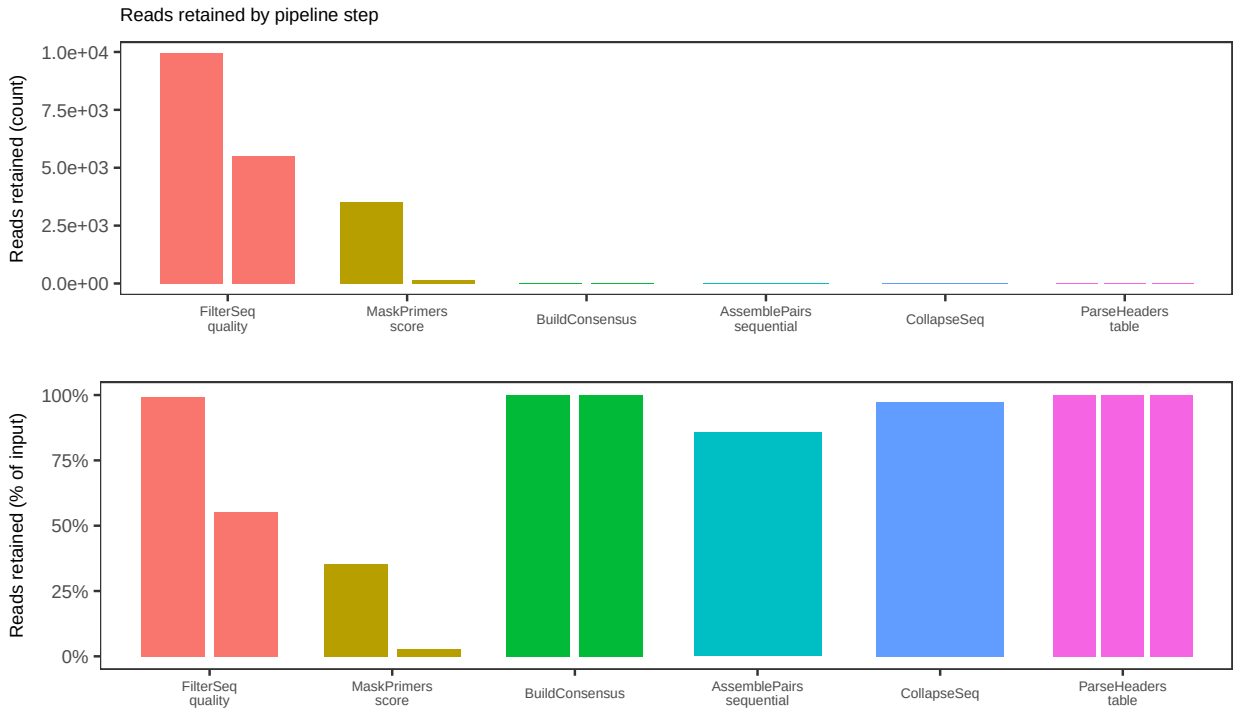

Figure 1: The number of reads or read sets retained at each processing step. Shown as raw counts (top) and percentages of input from the previous step (bottom). Steps having more than one column display individual values for read 1 (first column) and read 2 (second column).

Step	Task	Input	Passed	Failed
1	FilterSeq-quality	10000	9935	65
2	FilterSeq-quality	10000	5513	4487
3	MaskPrimers-score	9935	3511	6424
4	MaskPrimers-score	5513	160	5353
6	BuildConsensus	42	42	0
7	BuildConsensus	42	42	0
9	AssemblePairs-sequential	42	36	6
10	CollapseSeq	36	35	1
12	ParseHeaders-table	36	36	0
13	ParseHeaders-table	36	36	0
14	ParseHeaders-table	1	1	0

Table 1: The count of reads that passed and failed each processing step.

3 Quality Scores

NA FilterSeq tool remove reads with low mean Phred quality scores. Phred quality scores are assigned to each nucleotide base call in automated sequencer traces. The quality score (`Q`) of a base call is logarithmically related to the probability that a base call is incorrect (`P`): $Q = -10log_{10}P$. For example, a base call with

`Q=30` is incorrectly assigned 1 in 1000 times. The most commonly used approach is to remove read with average `Q` below 20.

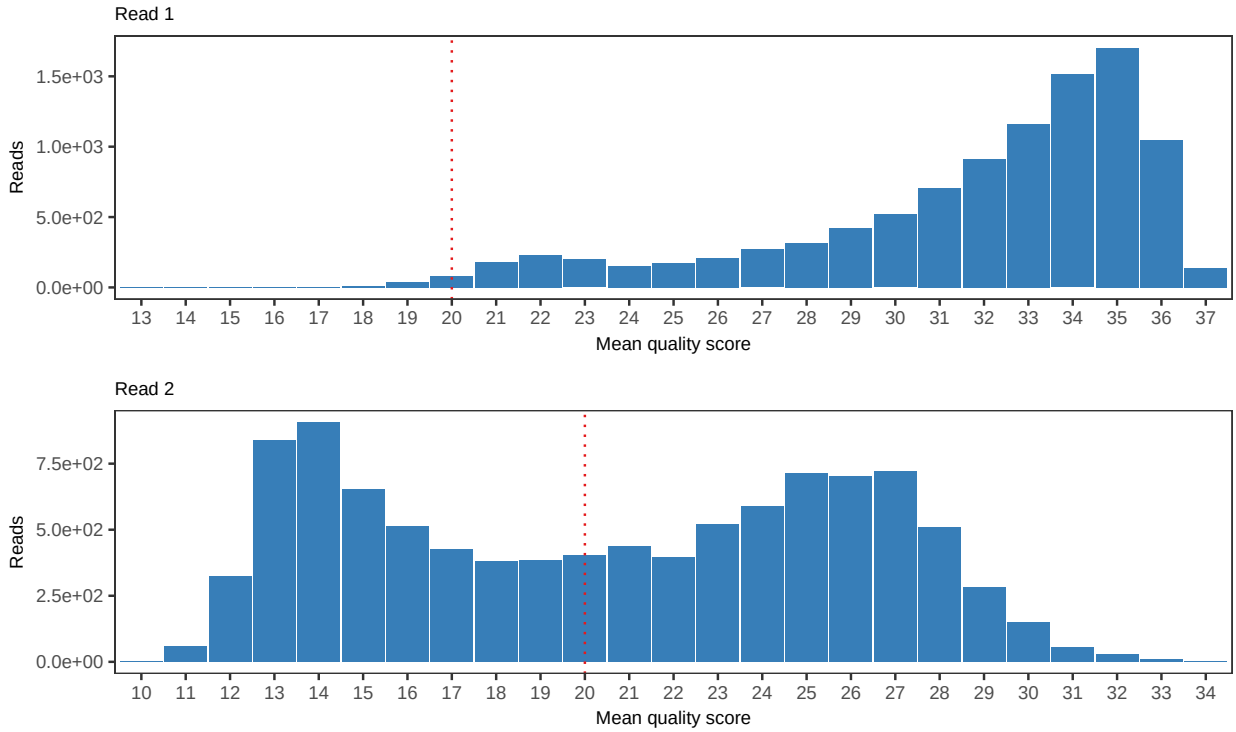

Figure 2: Mean Phred quality scores for read 1 (top) and read 2 (bottom). The dotted line indicates the average quality score under which reads were removed.

4 Primer Identification

The MaskPrimers tool supports identification of multiplexed primers and UMIs. Identified primer regions may be masked (with Ns) or cut to mitigate downstream SHM analysis artifacts due to errors in the primer region. An annotion is added to each sequences that indicates the UMI and best matching primer. In the case of the constant region primer, the primer annotation may also be used for isotype assignment.

4.1 Count of primer matches

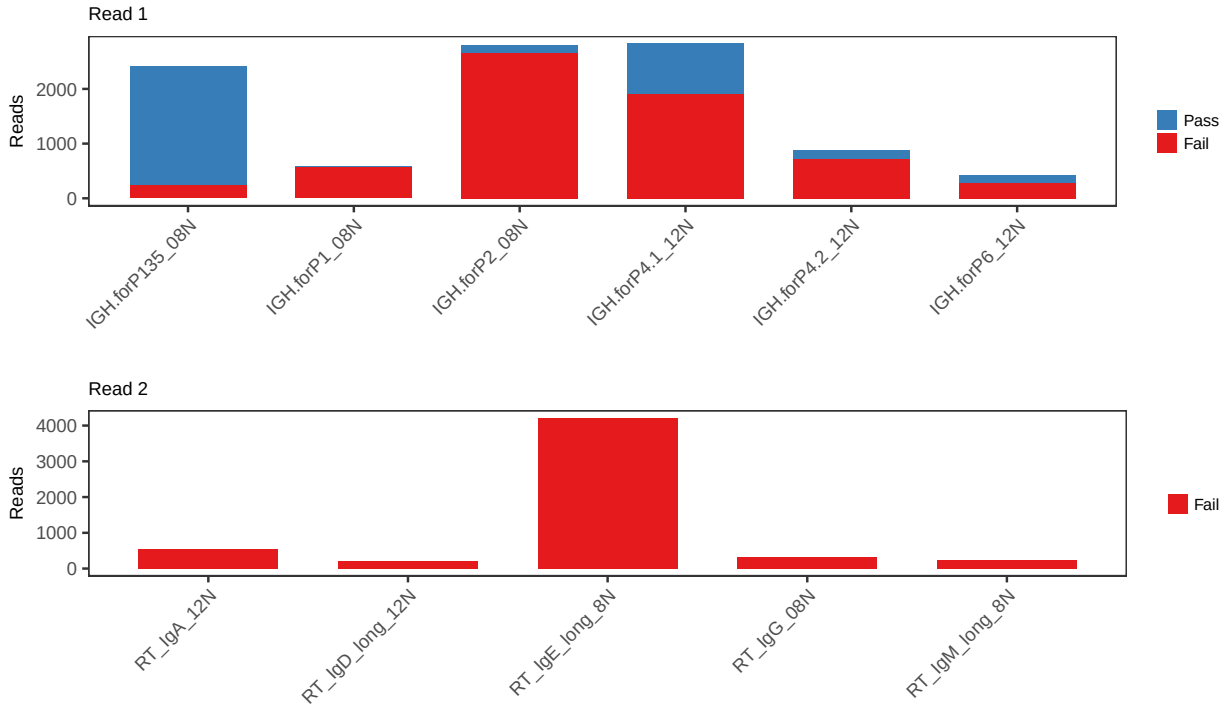

Figure 3: Count of assigned primers for read 1 (top) and read 2 (bottom). The bar height indicates the total reads assigned to the given primer, stacked for those under the error rate threshold (Pass) and over the threshold (Fail).

4.2 Primer match error rates

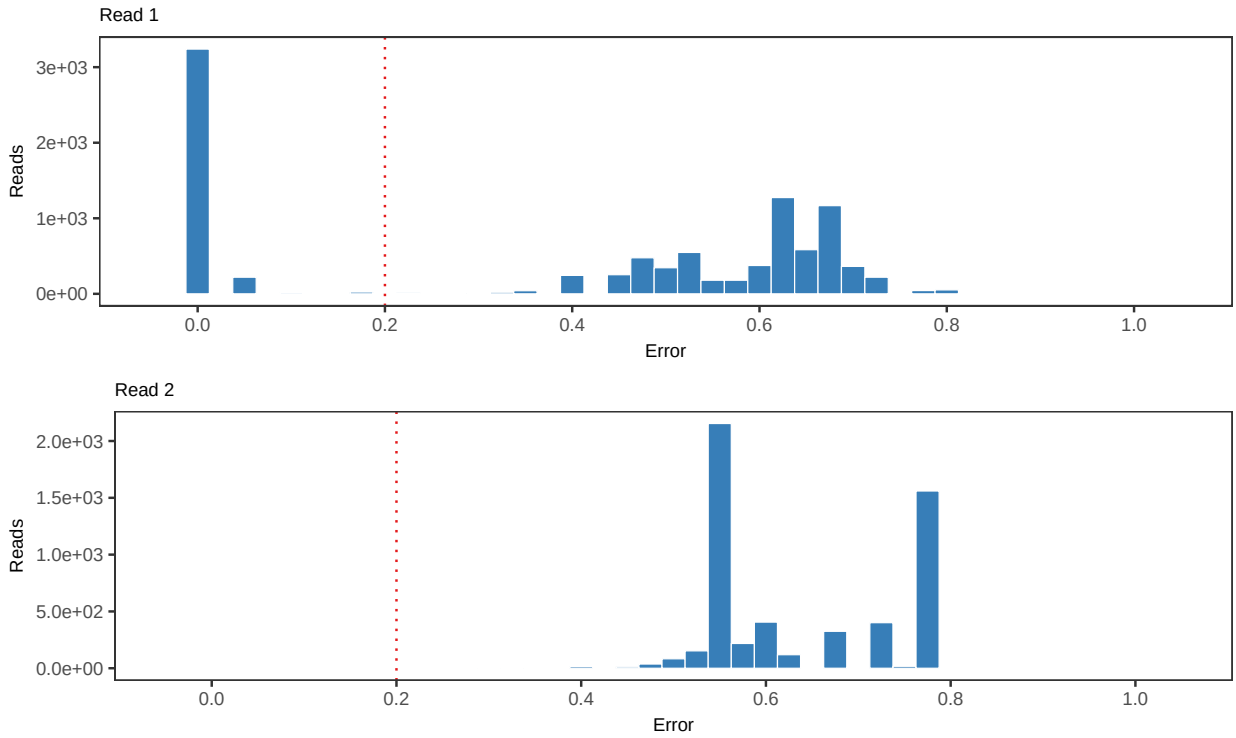

Figure 4: Distribution of primer match error rates for read 1 (top) and read 2 (bottom). The error rate is the percentage of mismatches between the primer sequence and the read for the best matching primer. The dotted line indicates the error threshold used.

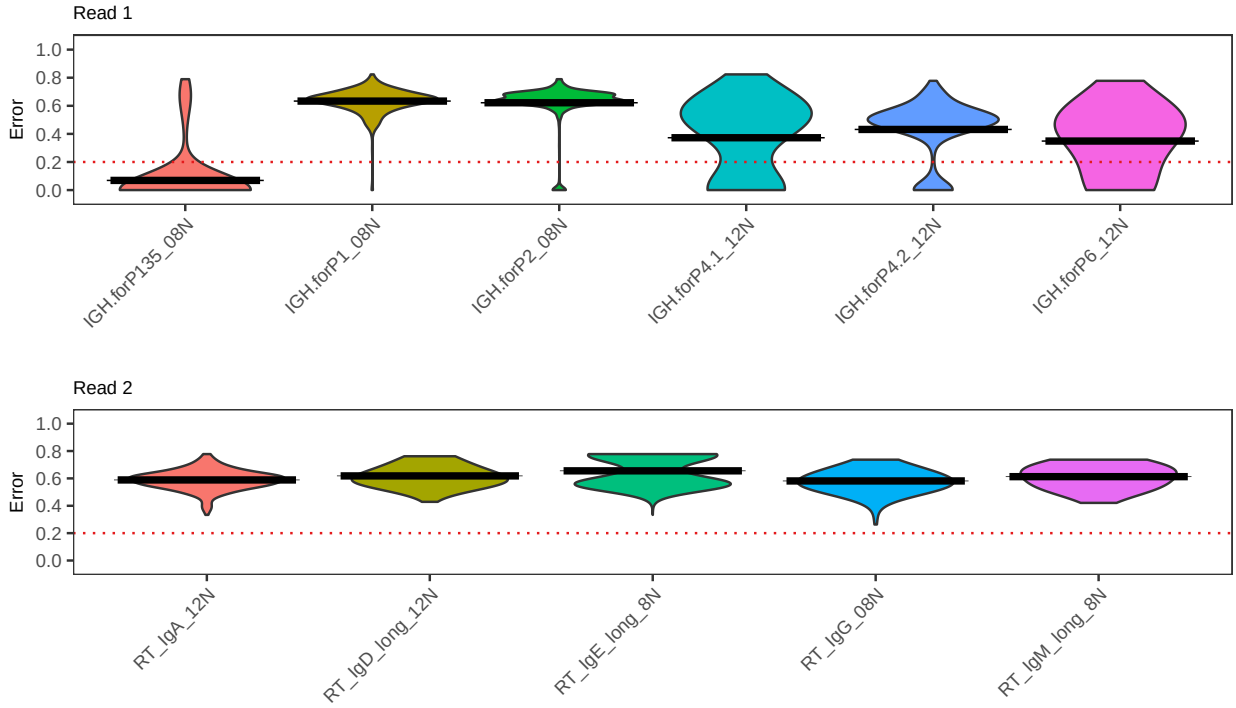

Figure 5: Distribution of primer match error rates for read 1 (top) and read 2 (bottom), broken down by assigned primer. The error rate is the percentage of mismatches between the primer sequence and the read for the best matching primer. The dotted line indicates the error threshold used.

5 Generation of UMI Consensus Sequences

Reads sharing the same UMI are collapsed into a single consensus sequence by the BuildConsensus tool. BuildConsensus considers several factors in determining the final consensus sequence, including the number of reads in a UMI group, Phred quality scores (`Q`), primer annotations, and the number of mismatches within a UMI group. Quality scores are used to resolve conflicting base calls in a UMI read group and the final consensus sequence is assigned consensus quality scores derived from the individual base quality scores. The numbers of reads in a UMI group, number of matching primer annotations, and error rate (average base mismatches from consensus) are used as strict cut-offs for exclusion of erroneous UMI read groups. Additionally, individual reads are excluded whose primer annotation differs from the majority in cases where there are sufficient number of reads exceeding the primer consensus cut-off.

5.1 Reads per UMI

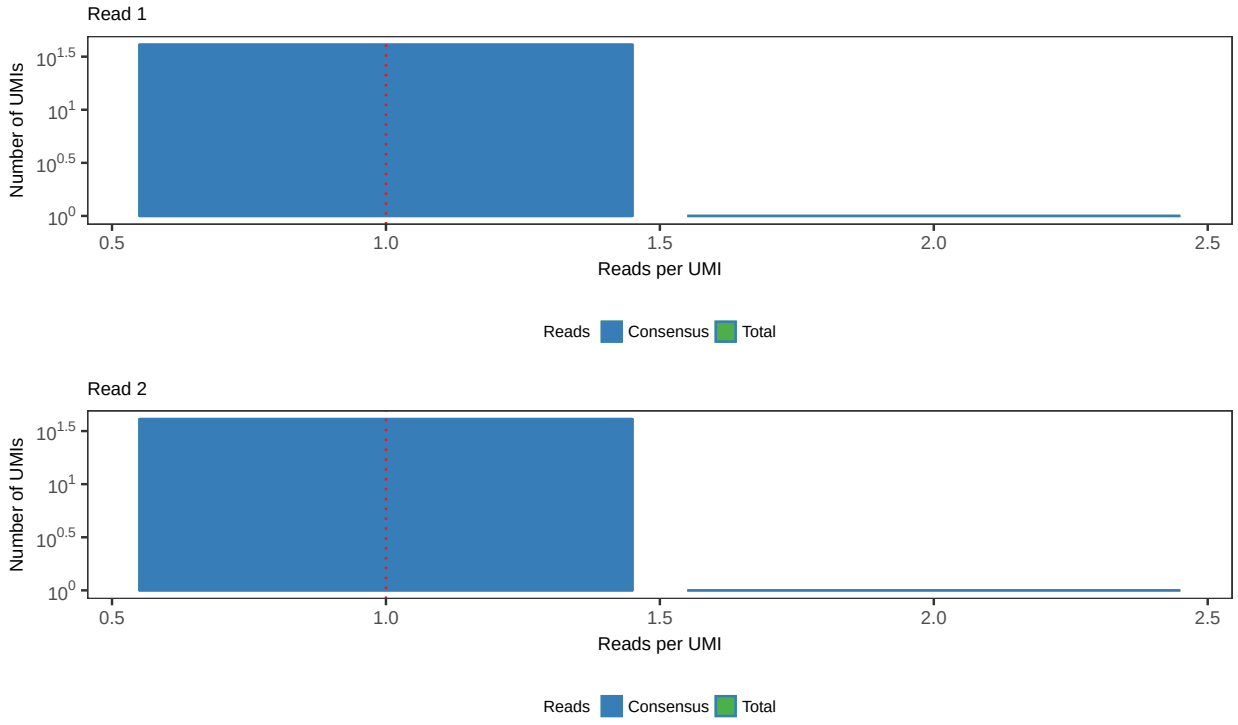

Figure 6: Histogram of UMI read group sizes (reads per UMI) for read 1 (top) and read 2 (bottom). The x-axis indicates the number of reads in a UMI group and the y-axis is the number of UMI groups with that size. The Consensus and Total bars are overlayed (not stacked) histograms indicating whether the distribution has been calculated using the total number of reads (Total) or only those reads used for consensus generation (Consensus).

5.2 UMI read group primer frequencies

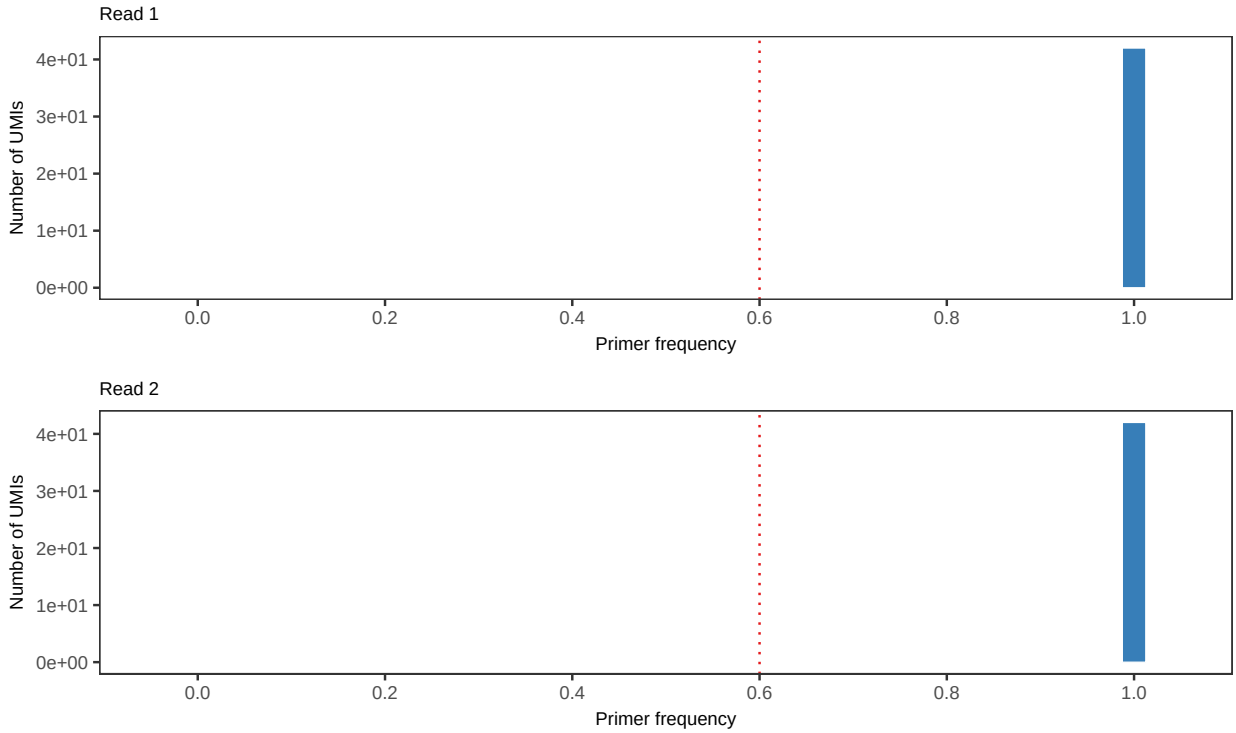

Figure 7: Histograms shoing the distribution of majority primer frequency for all UMI read groups for read 1 (top) and read 2 (bottom).

```
## Warning in plotBuildConsensus(consensus_log_1, consensus_log_2, titles
## = c("Read 1", : Not enough data points for violin plot. Falling back on
## boxplot.

## Warning in plotBuildConsensus(consensus_log_1, consensus_log_2, titles
## = c("Read 1", : Not enough data points for violin plot. Falling back on
## boxplot.
```

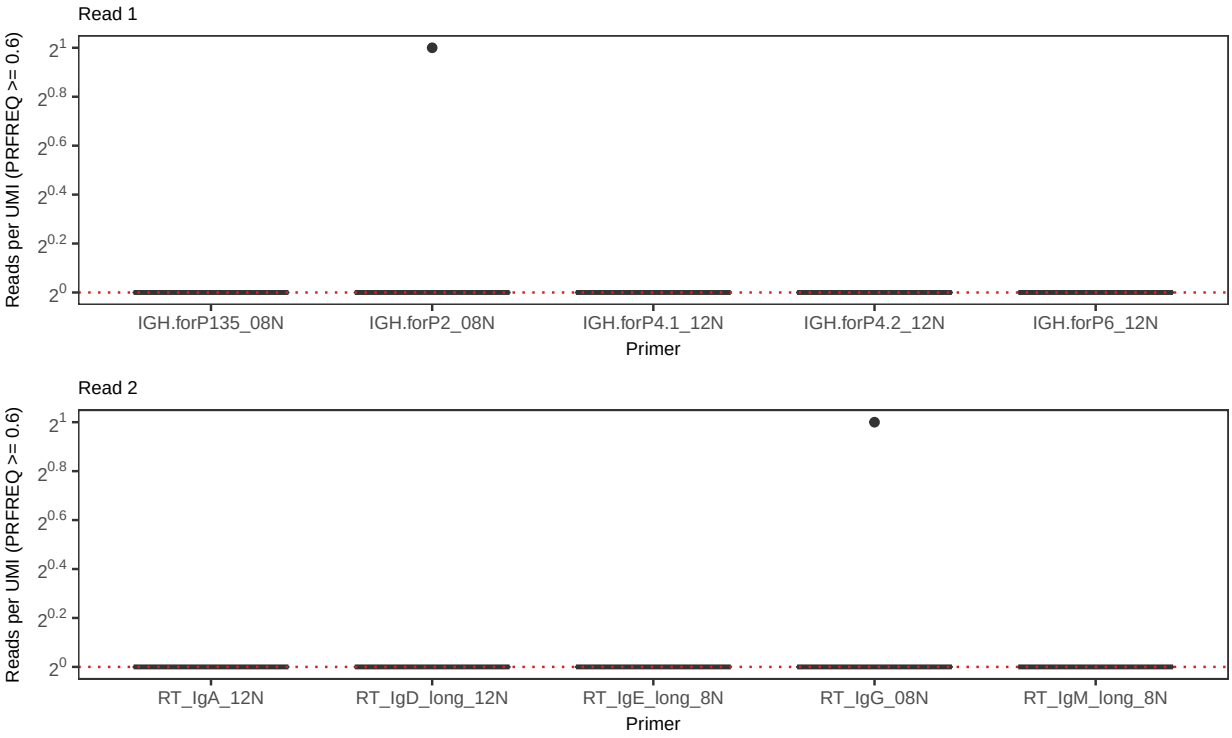

Figure 8: Violin plots showing the distribution of UMI read group sizes by majority primer for read 1 (top) and read 2 (bottom). Only groups with majority primer frequency over the PRFREQ threshold set when running BuildConsensus. Meaning, only retained UMI groups.

5.3 UMI read group error rates

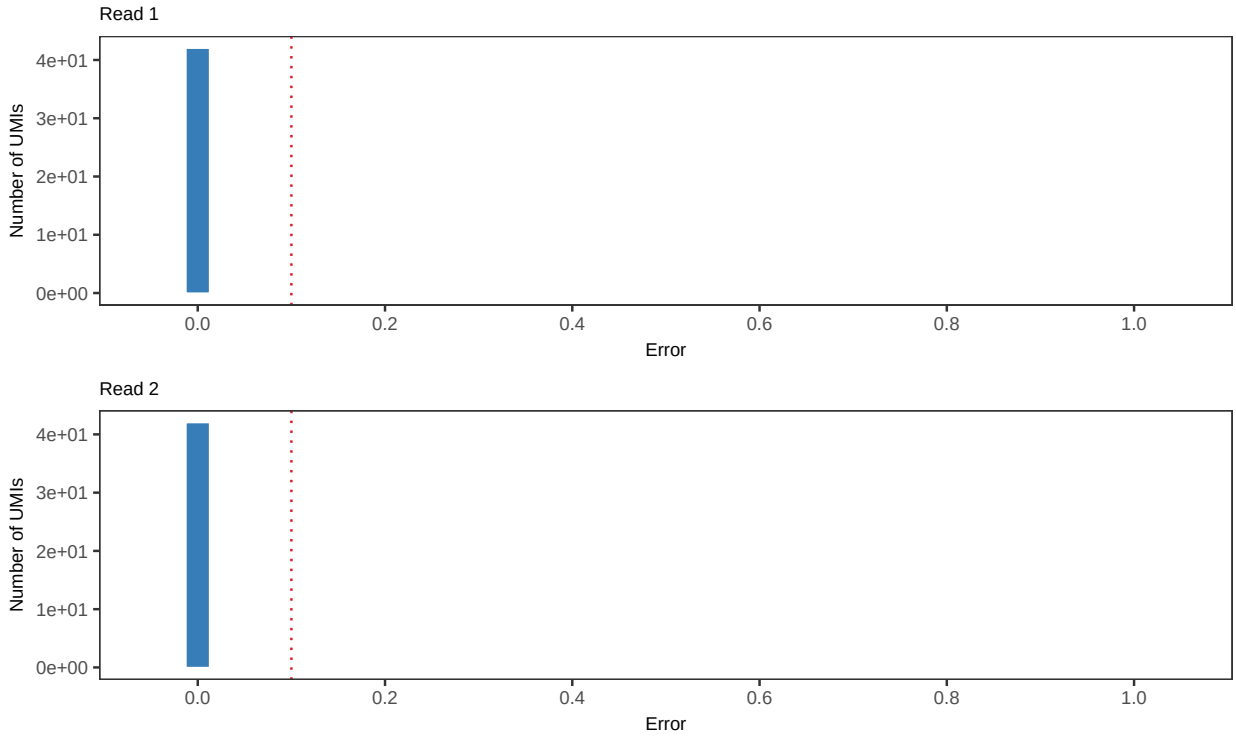

Figure 9: Histogram showing the distribution of UMI read group error rates for read 1 (top) and read 2 (bottom).

```
## Warning in plotBuildConsensus(consensus_log_1, consensus_log_2, titles
## = c("Read 1", : Not enough data points for violin plot. Falling back on
## boxplot.

## Warning in plotBuildConsensus(consensus_log_1, consensus_log_2, titles
## = c("Read 1", : Not enough data points for violin plot. Falling back on
## boxplot.
```

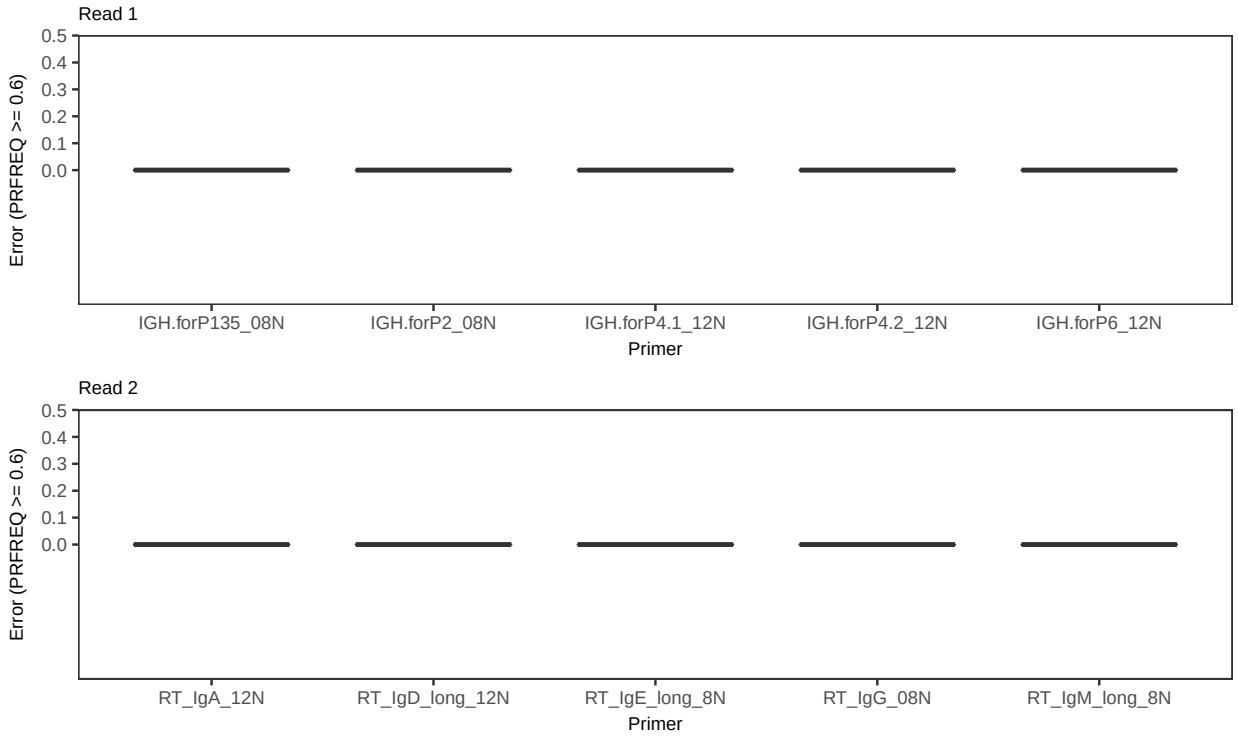

Figure 10: Violin plots showing the distribution of UMI read group error rates by majority primer for read 1 (top) and read 2 (bottom). Only groups with majority primer frequency over the PRFREQ threshold set when running BuildConsensus. Meaning, only retained UMI groups.

6 Paired-End Assembly

Assembly of paired-end readsis performed using the AssemblePairs tool which determines the read overlap in two steps. First, de novo assembly is attempted using an exhaustive approach to identify all possible overlaps between the two reads with alignment error rates and p-values below user-defined thresholds. This method is denoted as the `Align` method in the following figures. Second, those reads failing the first stage of de novo assembly are then mapped to the V-region reference sequences to create a full length sequence, padding with Ns, for any amplicons that have insufficient overlap for de novo assembly. This second stage is referred to as the `Reference` step in the figures below.

6.1 Counts of passing and failing assemblies

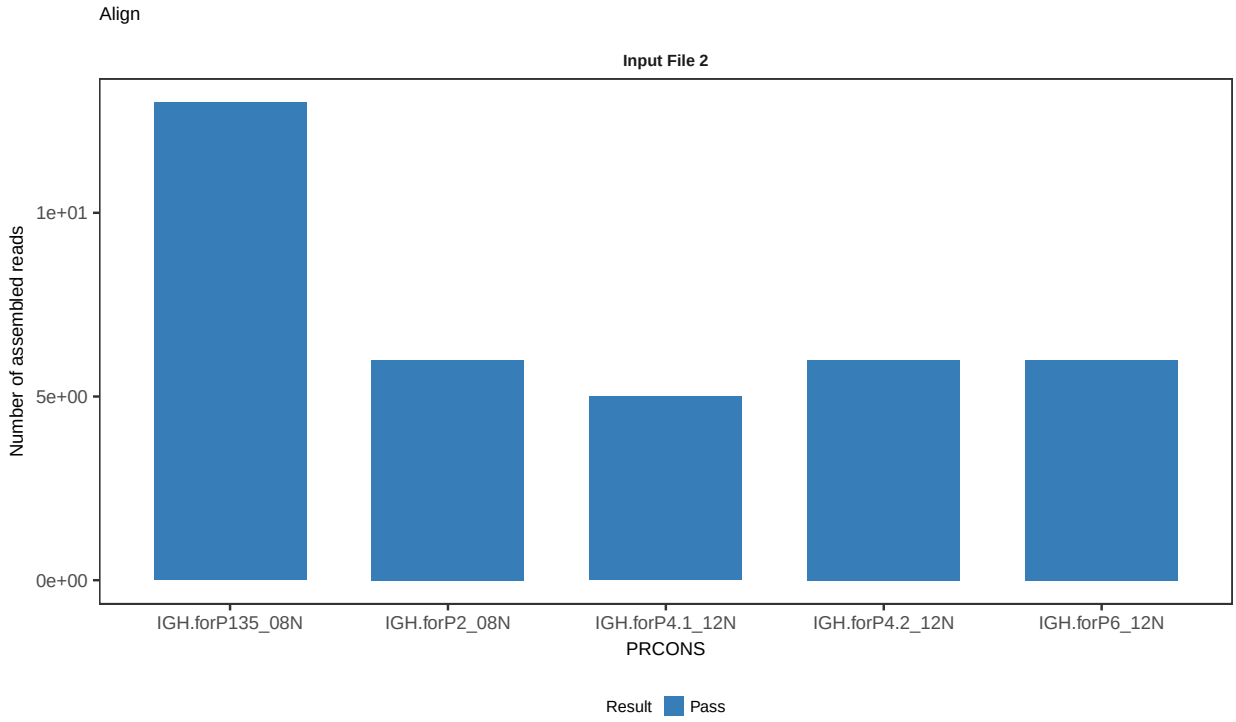

Figure 11: Count of mate-pairs passing and failing paired-end assembly are shown by by C-region primer annotation. The height of the bar is stacked reflecting the number of passing mate-pairs (blue) and failing mate-pairs (red). Results for the Align (top) and Reference (bottom) steps are indicated separately. The title of each panel indicates the input file that contains the PRCONS annotation.

6.2 Assembled sequence lengths

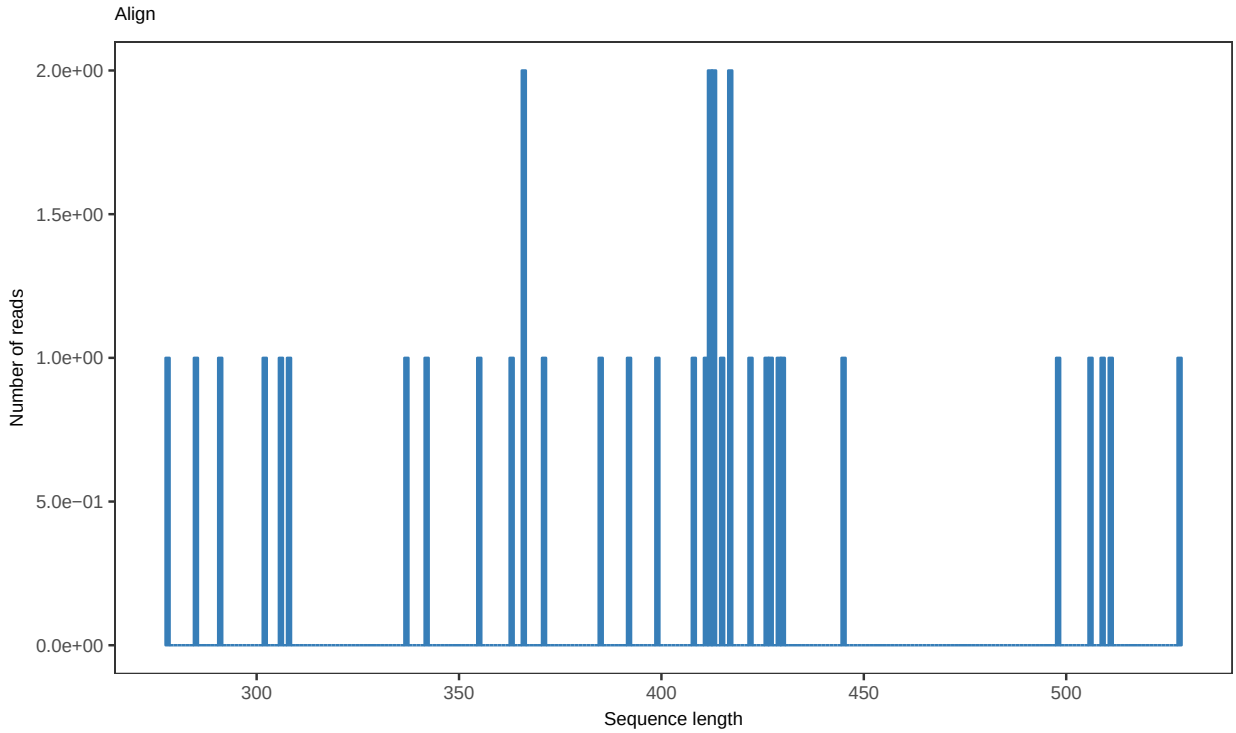

Figure 12: Histogram showing the distribution assembled sequence lengths in nucleotides for the Align step (top) and Reference step (bottom).

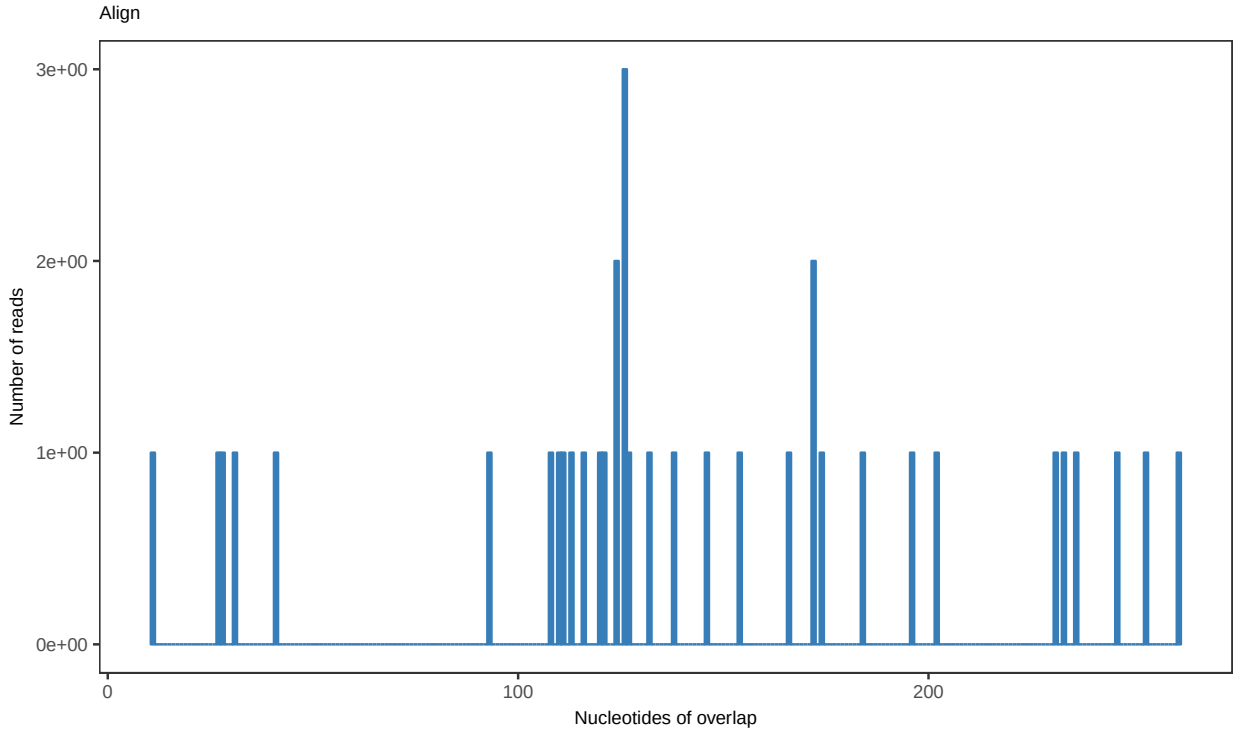

Figure 13: Histogram showing the distribution of overlapping nucleotides between mate-pairs for the Align

step (top) and Reference step (bottom). Negative values for overlap indicate non-overlapping mate-pairs with the negative value being the number of gap characters between the ends of the two mate-pairs.

6.3 Alignment error rates and significance

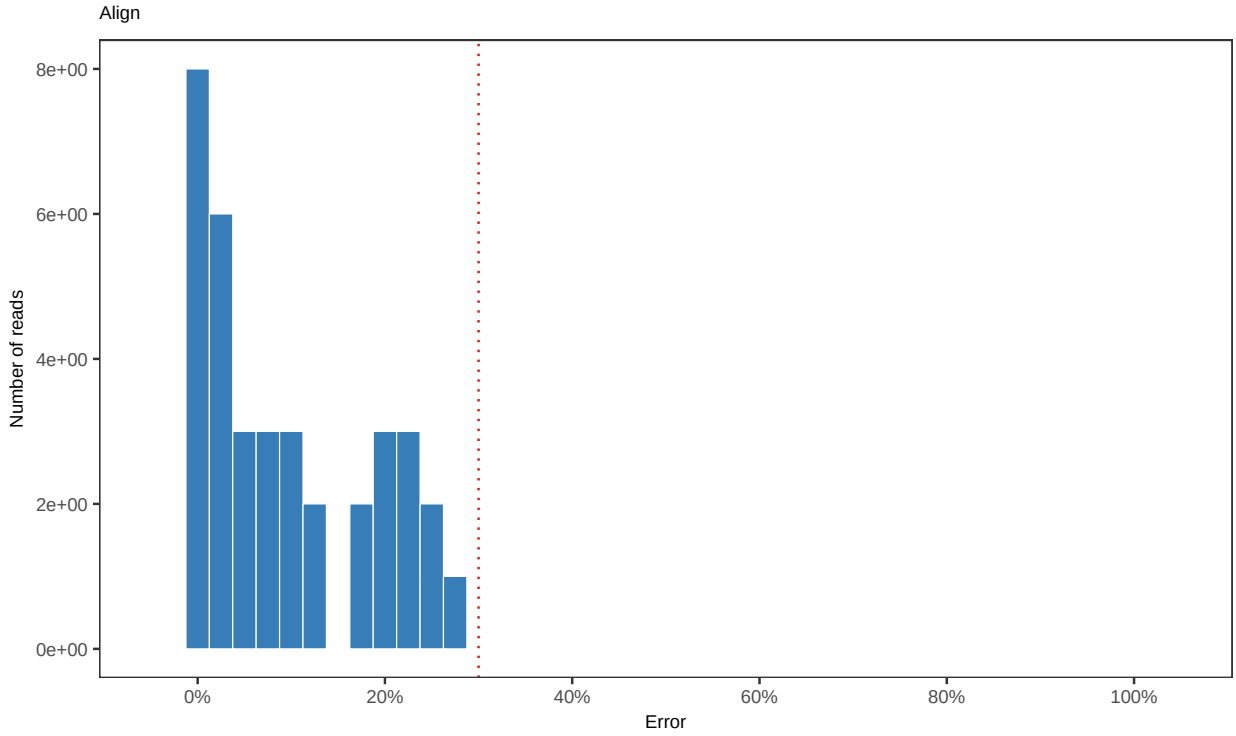

Figure 14: Histograms showing the distribution of paired-end assembly error rates for the Align step (top) and identity to the reference germline for the Reference step (bottom).

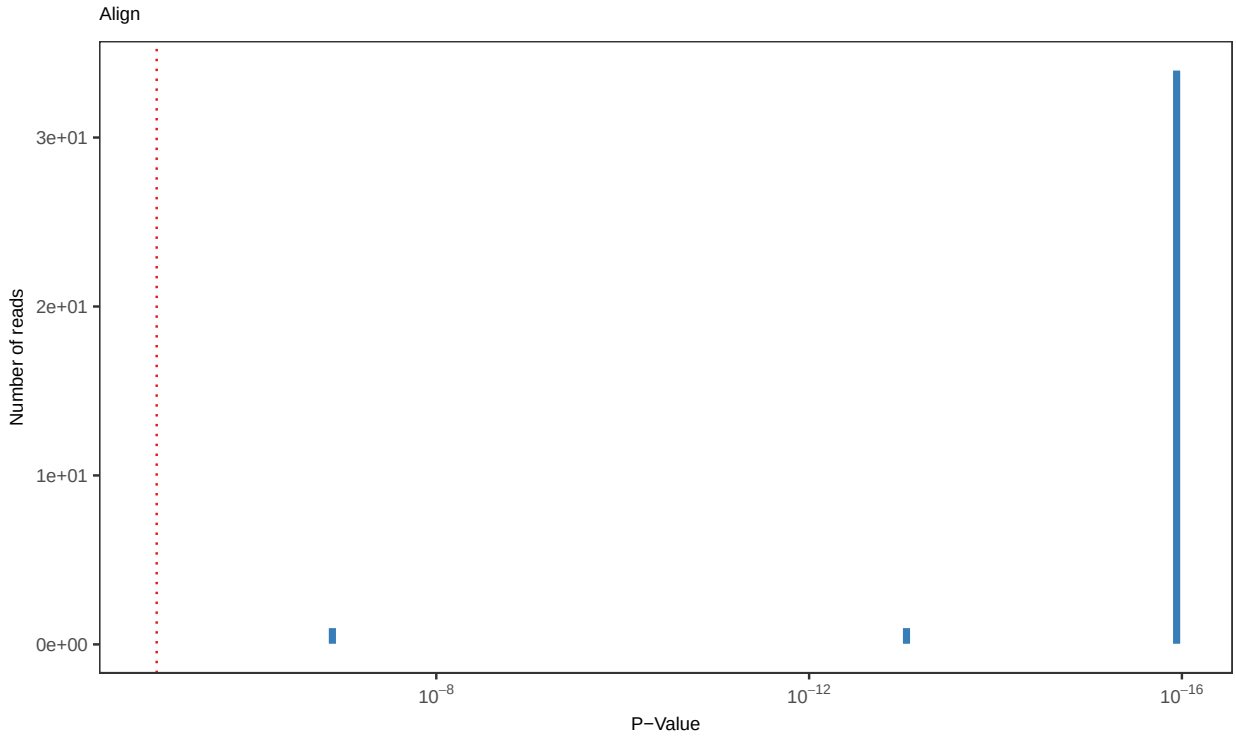

Figure 15: Histograms showing the distribution of significance scores for paired-end assemblies. P-values for the Align mode are shown in the top panel. E-values from the Reference step's alignment against the germline sequences are shown in the bottom panel for both input files separately.

7 Summary of Final Output

Final processed output is contained in the `total`, `unique`, and `unique-atleast-2` files, which contain all processed sequences, unique sequences, and only those unique sequences represented by at least two raw reads, respectively. The figures below shown the distributions of annotations for these final output files.

7.1 Distribution of read and UMI counts

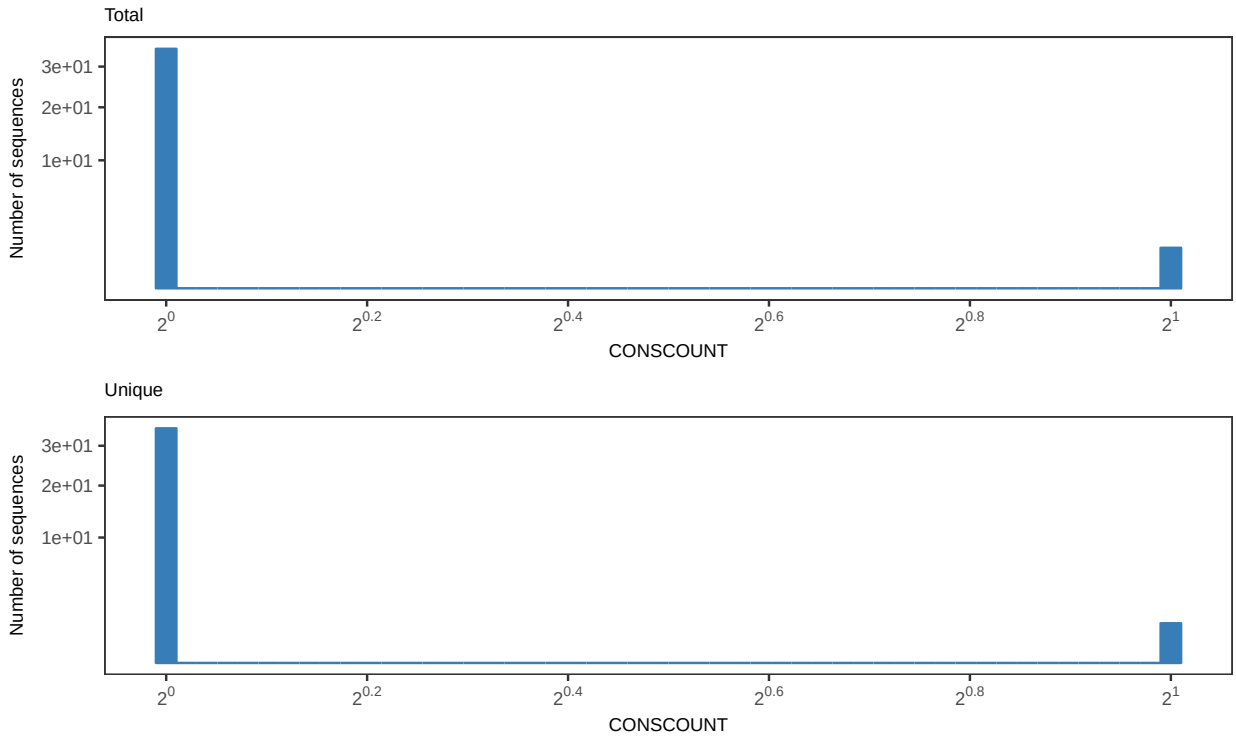

Figure 16: Histogram showing the distribution of read counts (CONSCOUNT) for total sequences (top) and unique sequences (bottom).

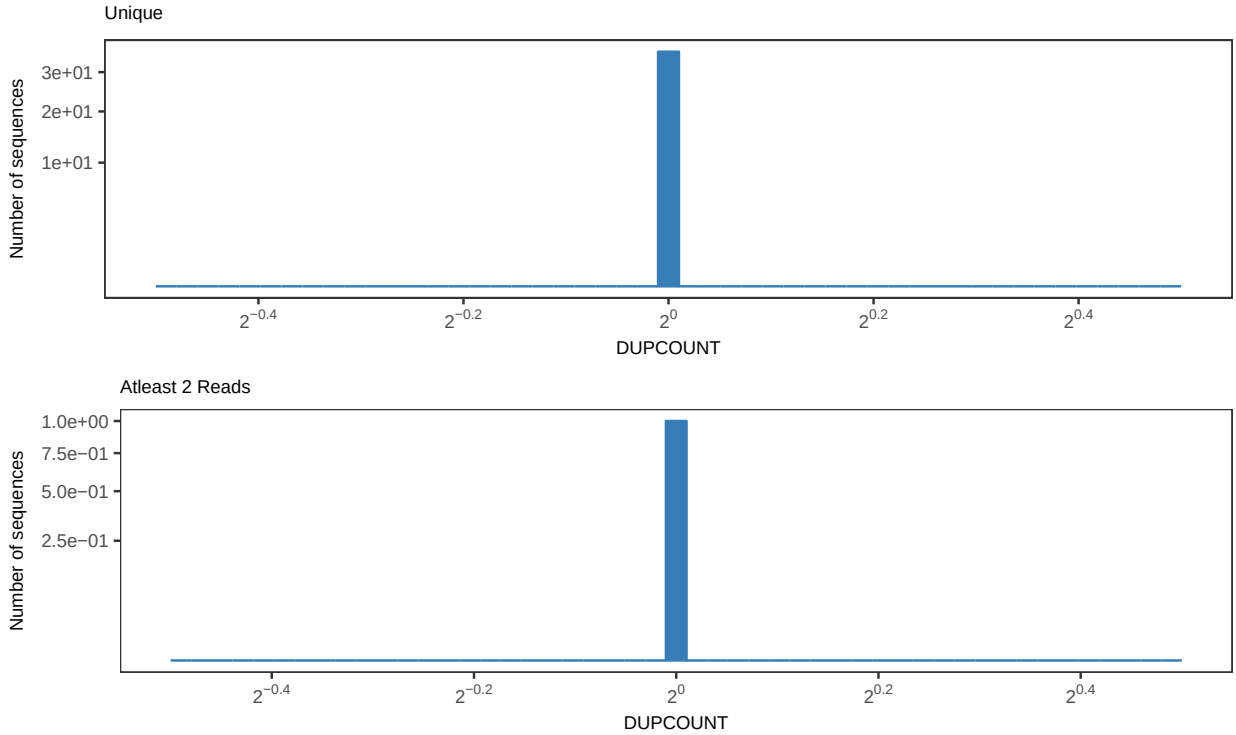

Figure 17: Histogram showing the distribution of unique UMI counts for all unique sequences (top) and

unique sequences represented by at least two raw reads (bottom).

7.2 C-region annotations

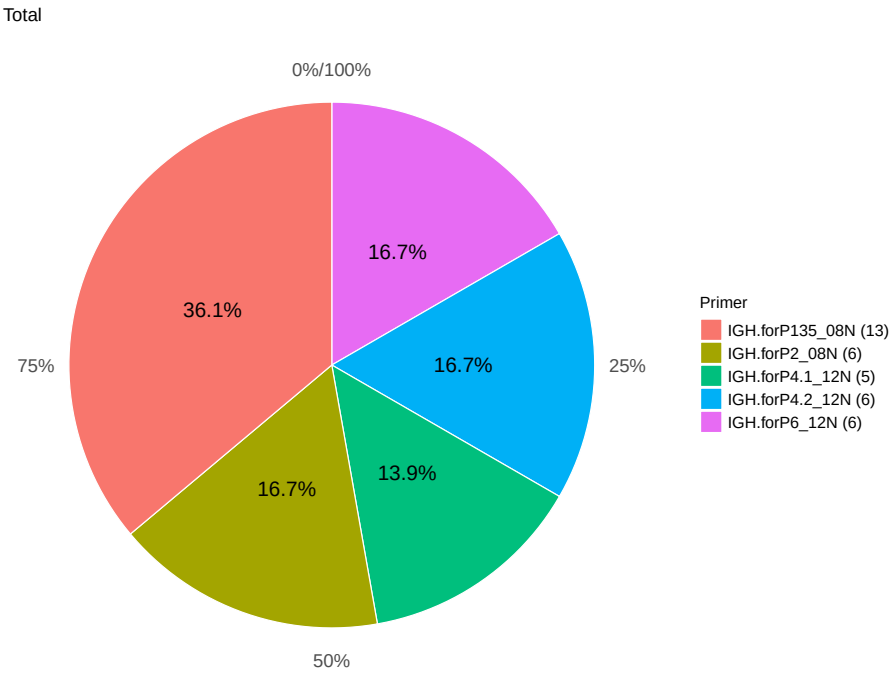

Figure 18: Percentage internal C-region annotations for total sequences. Parenthetical numbers in the legend are the number of sequences.

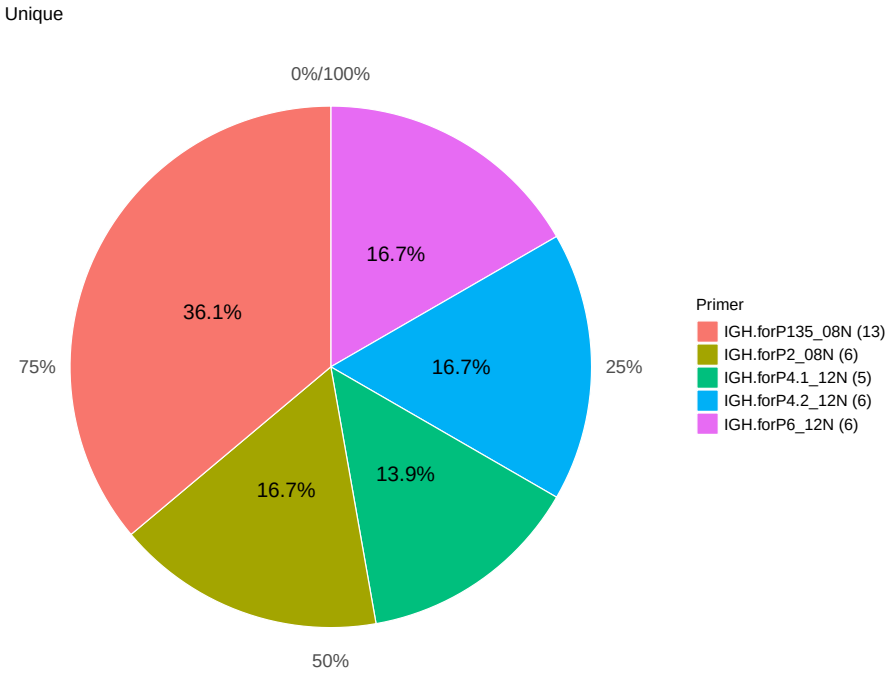

Figure 19: Percentage internal C-region annotations for all unique sequences. Parenthetical numbers in the legend are the number of sequences.

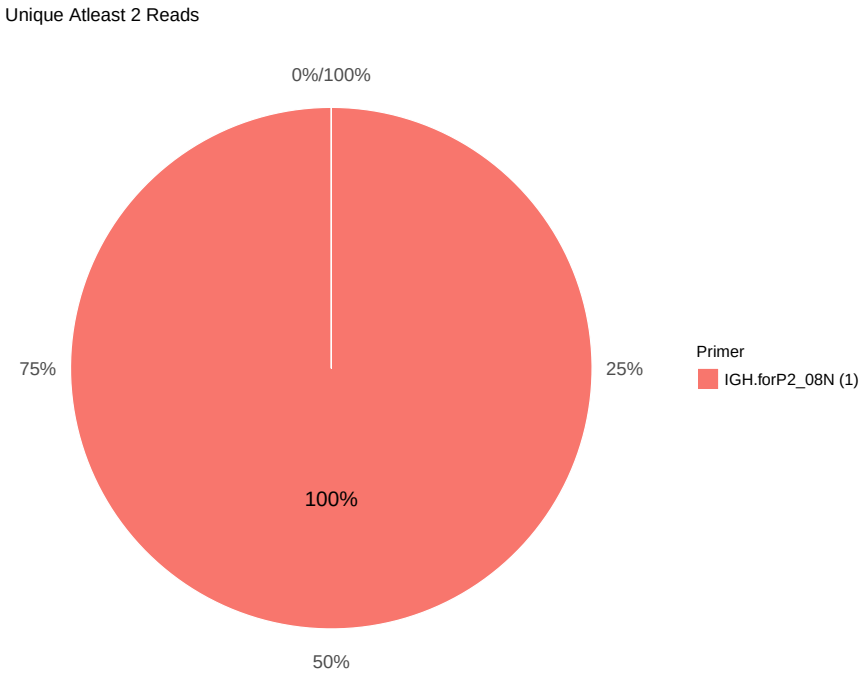

Figure 20: Percentage internal C-region annotations for unique sequences represented by at least two raw reads. Parenthetical numbers in the legend are the number of sequences.