LTR Parser Results

Kevin McCormick

01 May, 2020

Overview

This report contains the results of attempted primer & LTR identification in the following FASTQ file:
/home/kevin/dev/ltrparser/testdata/demo/fastq.gz/hivPosControl.fastq.gz

Summary Statistics for raw FASTQ file

Total	50000
Distinct	48708
Max	166
TotalLinker	9702
DistinctLinker	9506
MaxLinker	16
TotalHuman	17035
DistinctHuman	15743
MaxHuman	166

Explanation of terms

Total: Number of reads

Distinct: Number of distinct sequences

Max: Number of copies of most abundant sequence

TotalLinker: Number of reads containing R1 linker sequence

DisctinctLinker: Number of distinct sequences containing R1 linker sequence

MaxLinker: Number of copies of most abundant sequence containing R1 linker sequence

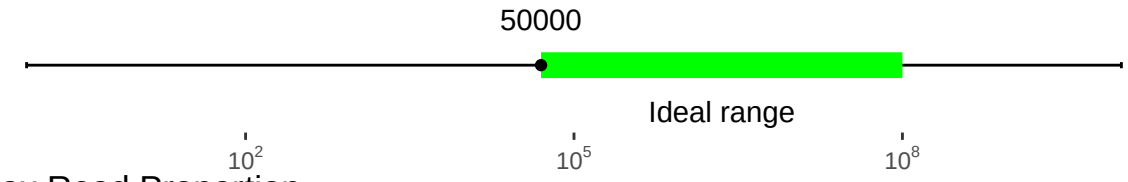
TotalHuman: Number of reads mapping to human genome

DisctinctHuman: Number of distinct sequences mapping to human genome

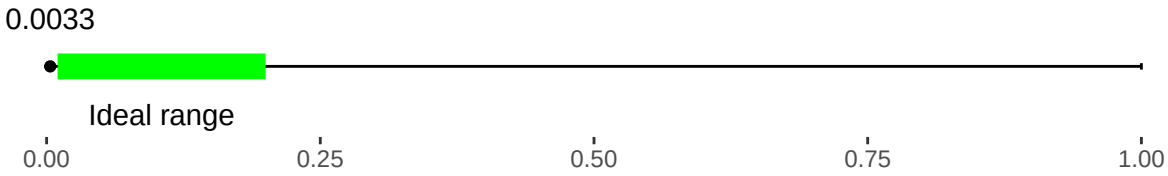
MaxHuman: Number of copies of most abundant sequence mapping to human genome

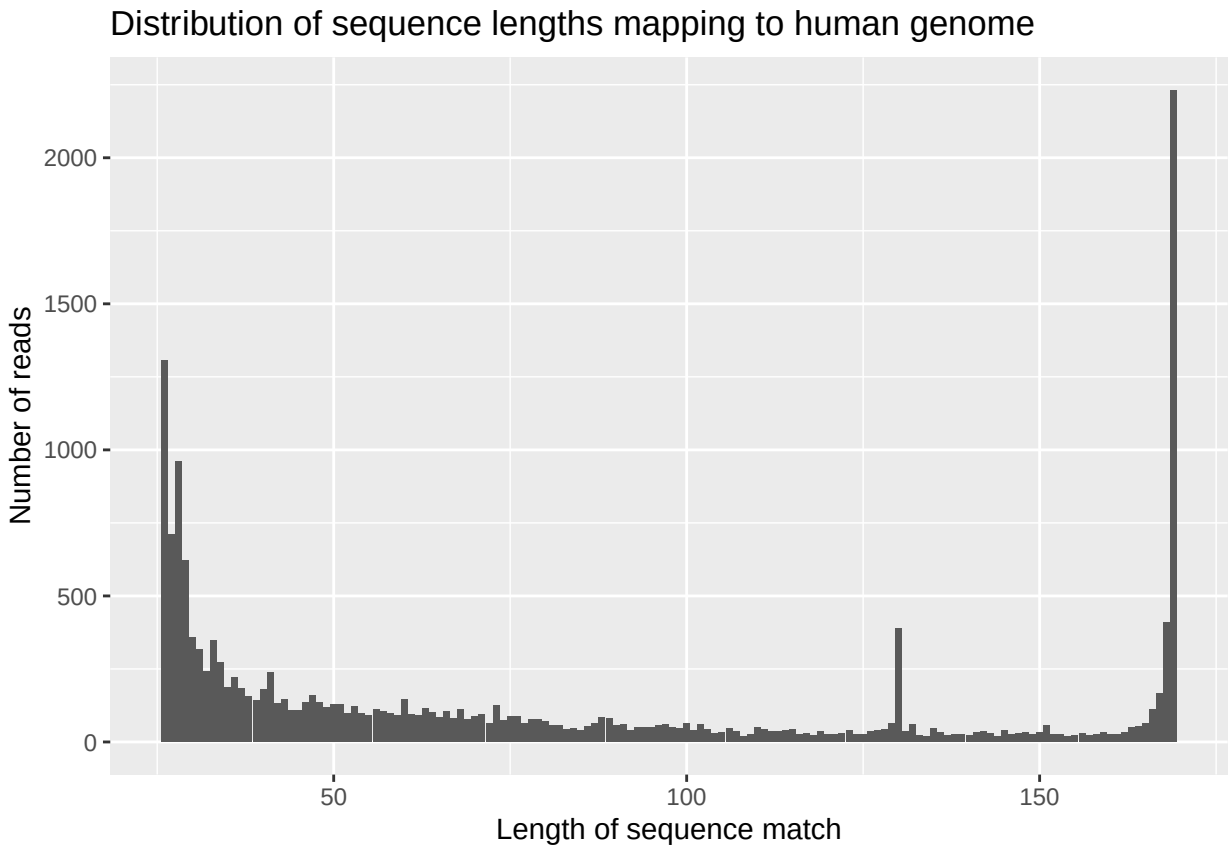
Key metrics

Total Reads



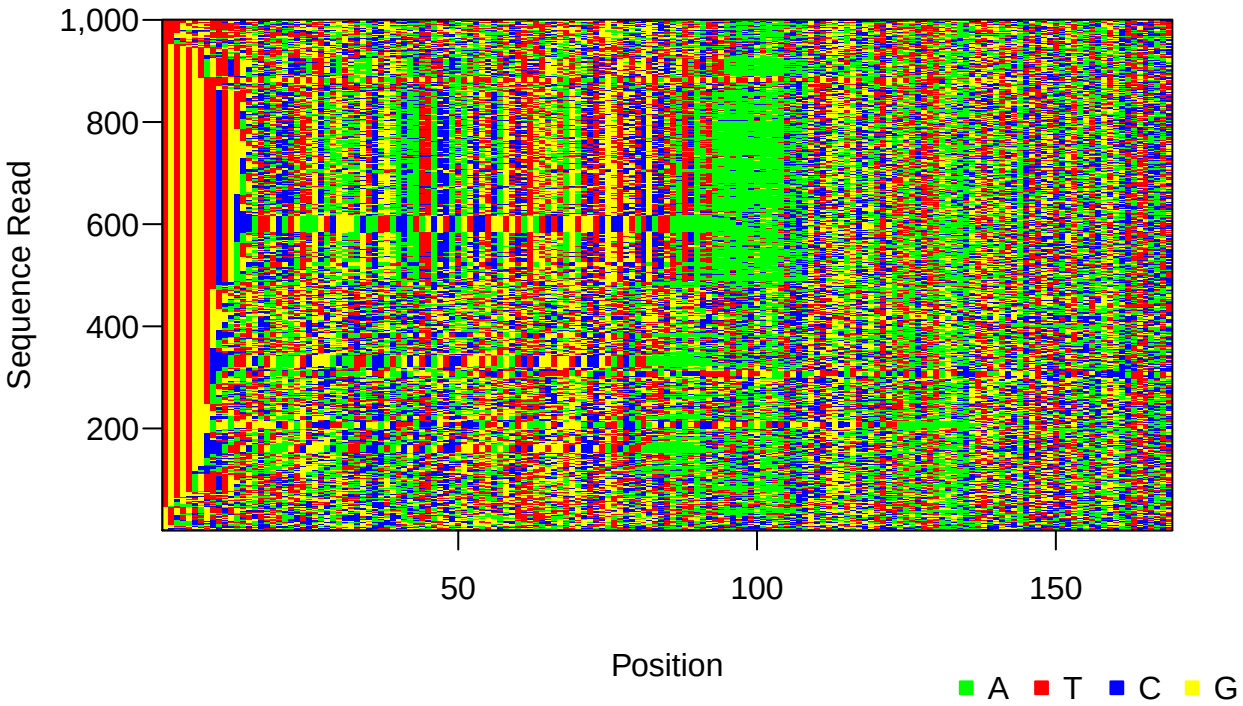
Max Read Proportion



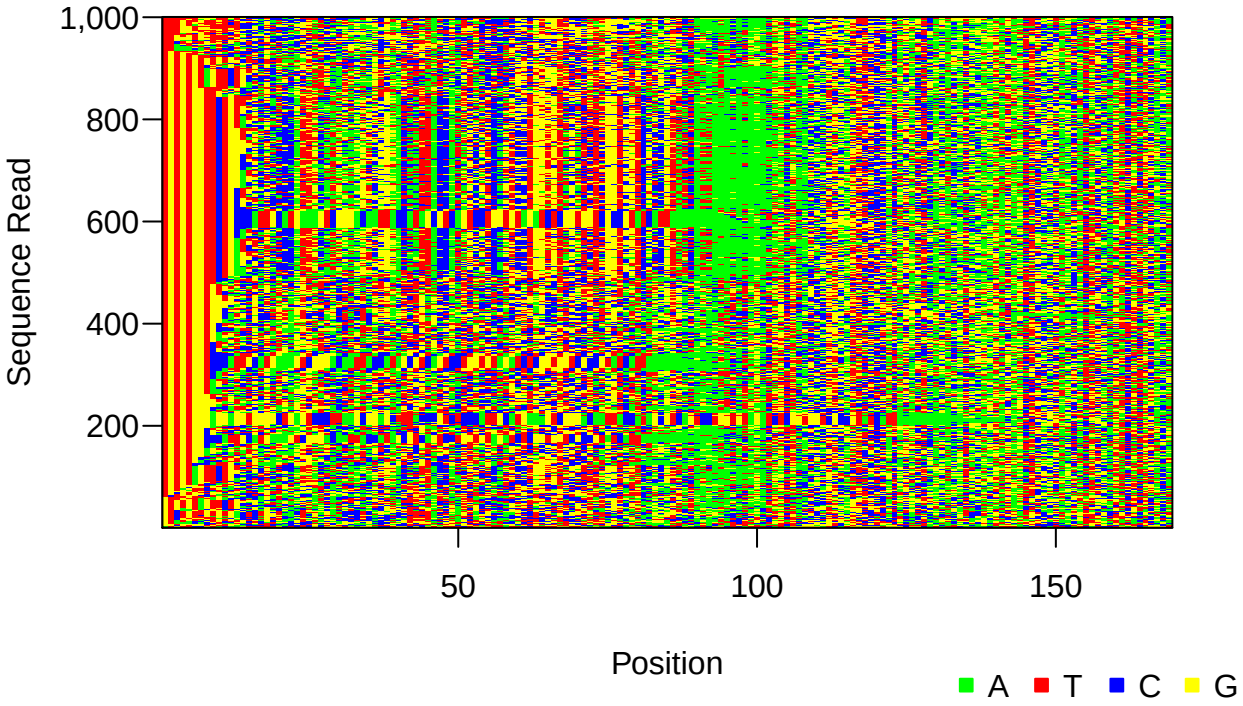

Distribution of sequence lengths mapping to human genome
Number of reads
2000
1500
1000
500
0
50
100
150
Length of sequence match

Visualization of raw FASTQ reads

Random sample of 1000 reads, arranged by nucleotide sequence:

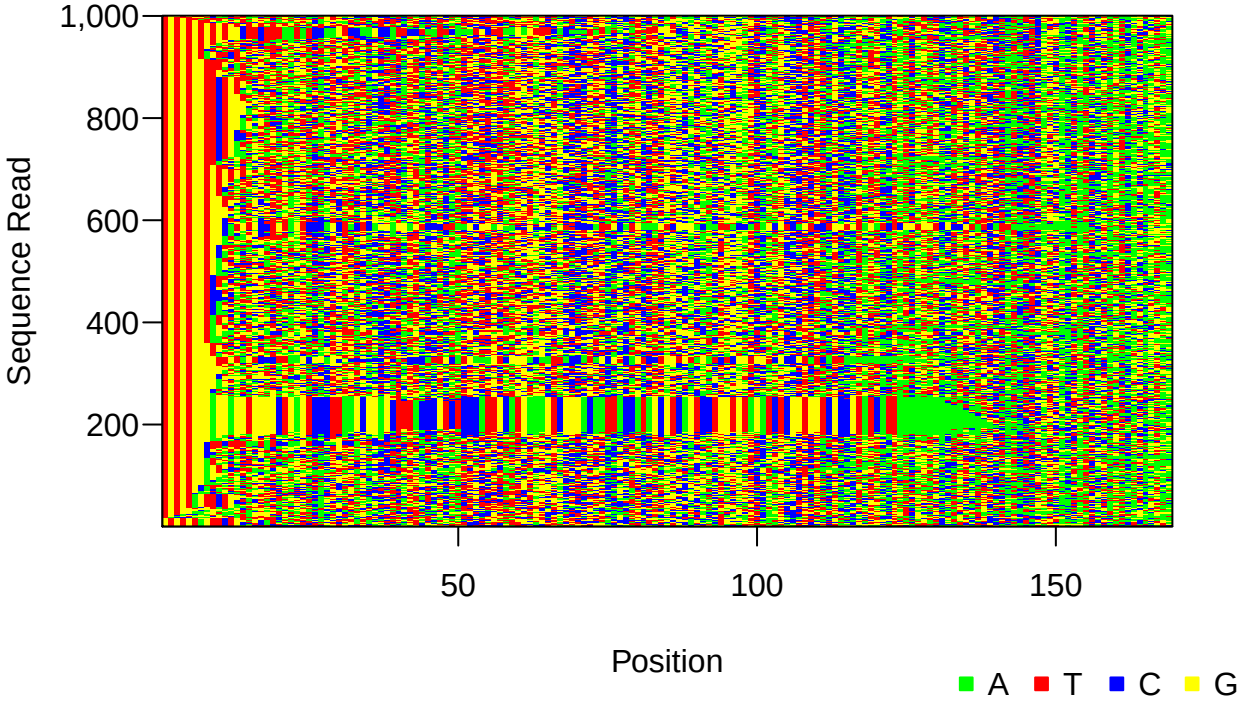


Random sample of 1000 distinct reads (i.e. after removing duplicates), arranged by nucleotide sequence:

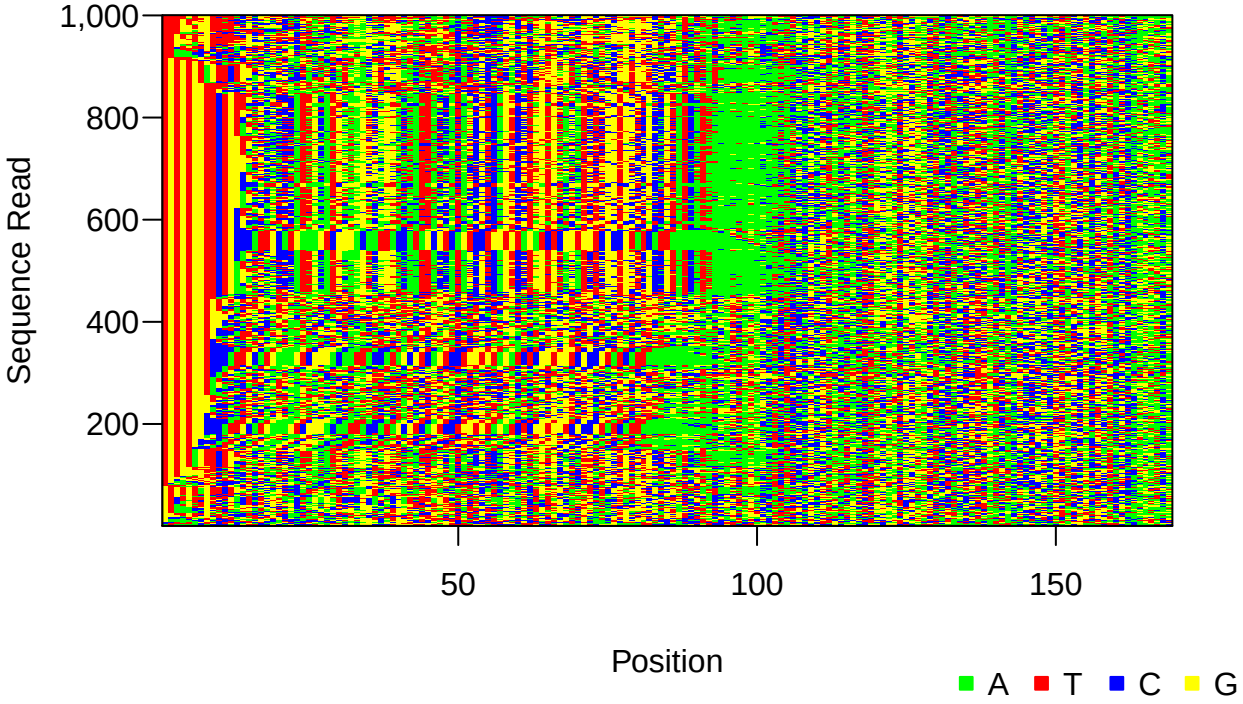


Visualization of raw FASTQ reads with R1 linker sequences

Random sample of 1000 reads containing R1 linker sequence, arranged by nucleotide sequence:



Random sample of 1000 reads that do NOT contain R1 linker sequence, arranged by nucleotide sequence:



Analysis of common primer-LTR pairs

A total of 2 potential primer-LTR pairs were identified.

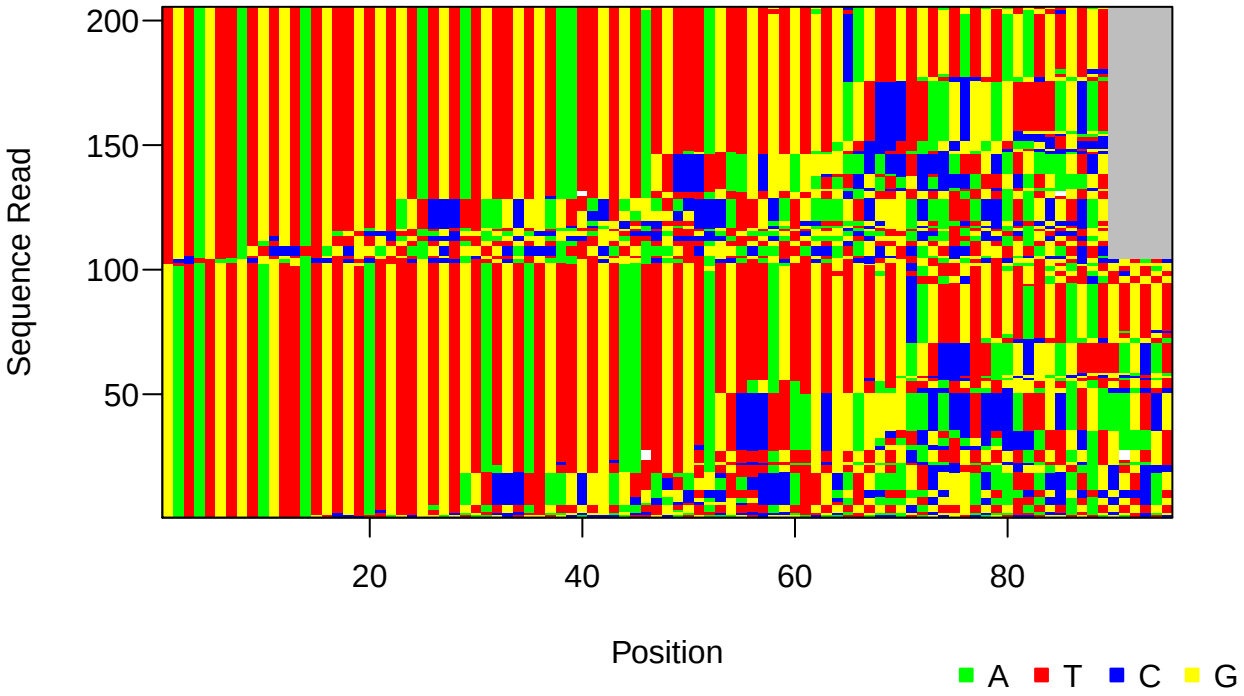
This table shows how the reads captured by each primer-LTR pair compare to each other and to the raw FASTQ file.

primer	LTR	expectedLTR	id
TGTGTGGT	TTTGTGTGTAGGTGTGTGTATGTAGATGTAATTGTGTGTATATGTTTAGTTGTGTGTGTATTAACA	FALSE	1
TGTGTGGT	TTTGTGTGTAGGTGTGTGTATGTAGATGTAATTGTGTGTATATGTTTAGTTGTGTGTGTATTAACAGATACA	FALSE	2

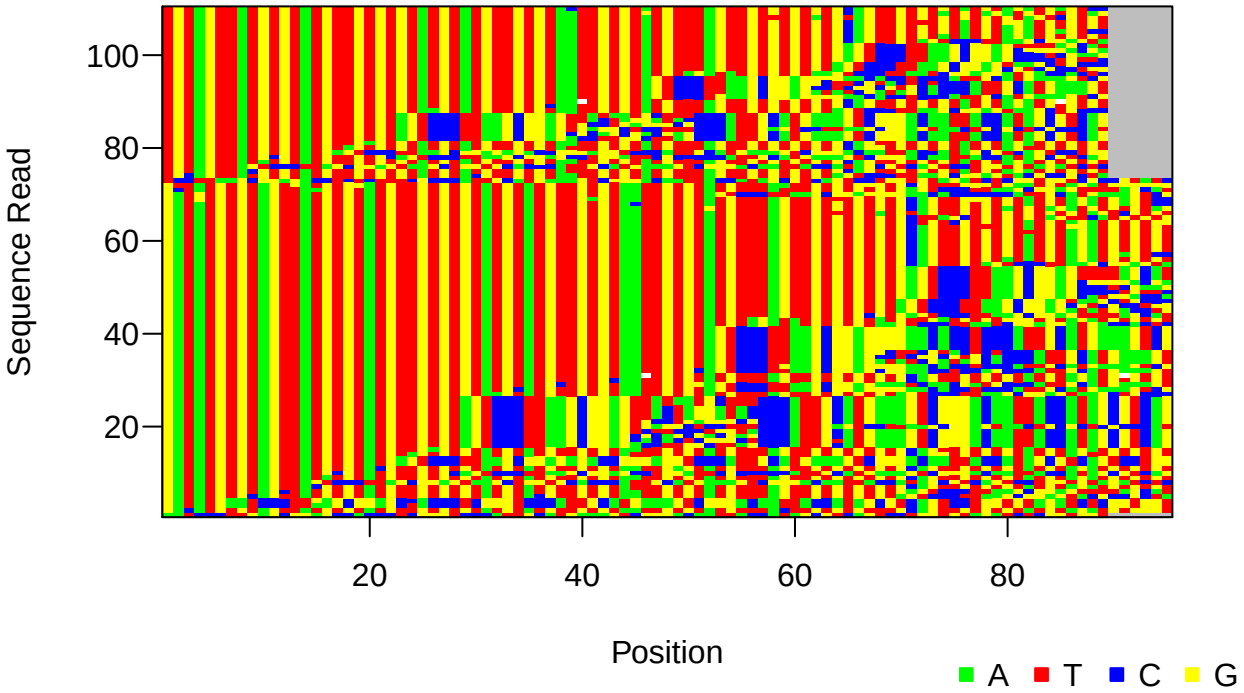
Variable	0	1	2
Distinct	48708	72	38
DistinctHuman	15743	72	38
DistinctLinker	9506	12	7
Max	166	16	16
MaxHuman	166	16	16
MaxLinker	16	6	6
Total	50000	121	84
TotalHuman	17035	121	84
TotalLinker	9702	19	14

Visualization of FASTQ reads, filtered for common primer & LTR

Random sample of 205 reads, filtered for common primer & LTR and with primer & LTR sequences trimmed, arranged by nucleotide sequence:



Random sample of 1000 distinct reads (i.e. after removing duplicates), filtered for common rimer & LTR and with primer & LTR sequences trimmed, arranged by nucleotide sequence:



Analysis of FASTQ reads mapped to LTR regions of HIV genomes in the Los Alamos National Laboratory database

0 candidates have an LTR sequence that mapped to an LTR region of at least one genome in the Los Alamos National Laboratories HIV database.