viral-ngs Documentation

Release v2.1.33.0-4-g140f3d0

Broad Institute Viral Genomics

2023-03-28

CONTENTS

CHAPTER

ONE

CONTENTS

1.1 Description of the methods

1.1.1 Viral genome analysis

Viral genome assembly

The filtered and trimmed reads are subsampled to at most 100,000 pairs. *de novo* assemby is performed using SPAdes. Reference-assisted assembly improvements follow (contig scaffolding, orienting, etc.) with MUMMER and MUSCLE or MAFFT. Gap2Seq is used to seal gaps between scaffolded *de novo* contigs with sequencing reads.

Each sample's reads are aligned to its *de novo* assembly using Novoalign and any remaining duplicates were removed using Picard MarkDuplicates. Variant positions in each assembly were identified using GATK IndelRealigner and UnifiedGenotyper on the read alignments. The assembly was refined to represent the major allele at each variant site, and any positions supported by fewer than three reads were changed to N.

This align-call-refine cycle is iterated twice, to minimize reference bias in the assembly.

1.2 Command line tools

1.2.1 assembly.py - *de novo* assembly