AZENTA
LIFE SCIENCES
azenta.com

BIOINFORMATICS REPORT

Pacbio De-novo Hifi Eukaryotic Assembly

AZENTA LIFE SCIENCES
NEXT GENERATION SEQUENCING

115 Corporate Boulevard
South Plainfield, NJ 07080
Email: ngs@azenta.com
Phone: 908-222-0711 Ext. 1

5/23/2024

Table of Contents

1 Project Information

Project	30-742644819
Customer	Thomas Parchman
Institution	University of Nevada - Reno
Email	tparchman@unr.edu
Service	Pacbio Hifi DNA De-novo Assembly Package
Species	*Loxia curvirostra*
Sample	2

2 Description of Workflow

2.1 Description of library preparation workflow

PacBio library was prepared using SMRTbell prep kit per the manufacturer's protocol. Follow the run design guidelines for sequencing. Figure 1 outlines the PacBio library preparation and sequencing workflow.

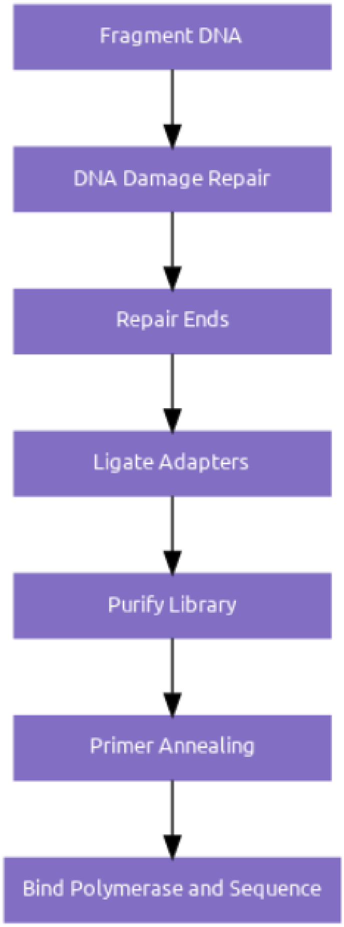
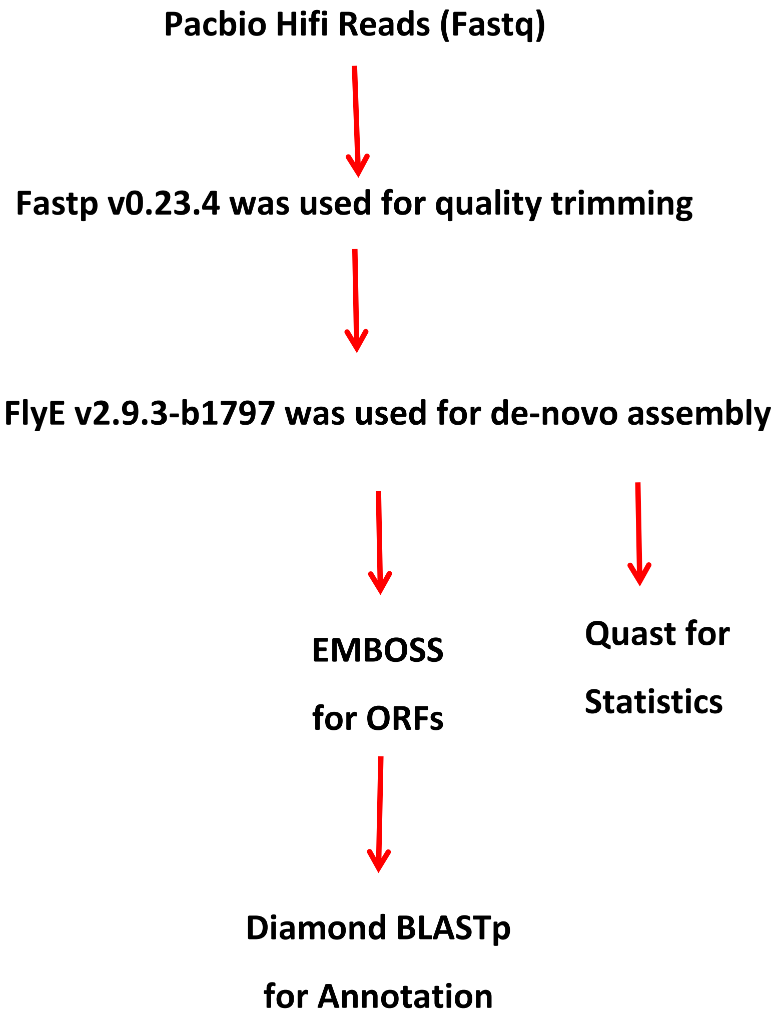


Figure 1: Pacbio library preparation and sequencing workflow 3. CCS ANALYSIS

2.2 Workflow of Data Analysis



Pacbio Hifi Reads (Fastq)

Fastp v0.23.4 was used for quality trimming

FlyE v2.9.3-b1797 was used for de-novo assembly

EMBOSS for ORFs

Quast for Statistics

Diamond BLASTp for Annotation

3 Data Analysis Results

The Hifi data for each of the samples was de-novo assembled to a genome using FlyE v2.9.3-b1797. The raw data was assessed for quality trimming using fastp v0.23.4 setting of length being 2000 bps and quality score of each read being a minimum of 15. FlyE v2.9.3-b1797 was used with default settings with polishing to generate a draft de-novo assembly. We then filtered the draft assembly for small contigs (less than 50Kbps) to a generate a final assembly. The final assembly was analyzed for open reading frames via EMBOSS tools ORF Finder. The identified protein ORF was used for Diamond BLASTp annotation through NCBI's NR database. Also, Quast was then used to generate general assembly statistics.

3.1 Assembly Statistics

Quast Statistics for each assembly.

Assembly Statistics	171203-5	S170996-2
# contigs (>= 0 bp)	9,765	9,014
# contigs (>= 1000 bp)	9,765	9,014
# contigs (>= 5000 bp)	9,765	9,014
# contigs (>= 10000 bp)	9,765	9,014
# contigs (>= 25000 bp)	9,765	9,014
# contigs (>= 50000 bp)	9,765	9,014
Total length (>= 0 bp)	1,495,363,063	1,697,079,343
Total length (>= 1000 bp)	1,495,363,063	1,697,079,343
Total length (>= 5000 bp)	1,495,363,063	1,697,079,343
Total length (>= 10000 bp)	1,495,363,063	1,697,079,343
Total length (>= 25000 bp)	1,495,363,063	1,697,079,343
Total length (>= 50000 bp)	1,495,363,063	1,697,079,343
# contigs	9,765	9,014
Largest contig	3,863,186	5,613,873
Total length	1,495,363,063	1,697,079,343
GC (%)	42.64	43.38
N50	197,710	286,440
N75	101,361	130,809
L50	1,817	1,439
L75	4,493	3,698
# N's per 100 kbp	0.85	0.72

4 Deliverables

- Standard Hifi Sequel IIe
  - Hifireads.bam
  - Hifireads.fastq.gz
- Assembly
  - FlyE v2.9.3-b1797 – filtered for small contigs (50 Kbps)
    - Assembled contigs (fasta)
  - FlyE v2.9.3-b1797 - Unfiltered includes small contigs (50 Kbps)
    - Unfiltered contigs (fasta)
- Annotation
  - EMBOSS ORF Finder
    - Nucleotide ORFs – fasta and table (TSV)
    - Protein ORFs – fasta and table (TSV)
  - Diamond BLASTp
    - Annotation Table (TSV)
- Statistics
  - FlyE v2.9.3-b1797
    - Per contig stats on unfiltered contigs (TXT)
  - Quast – Assembly statistics
    - General assembly statistics on filtered de-novo assembly (TSV)