GENEWIZ, Inc.
115 Corporate Boulevard
South Plainfield, NJ 07080
p (877) GENEWIZ (436-3949)
f (908) 333-4511
www.genewiz.com

GENEWIZ BIOINFORMATICS REPORT

de-novo Genome Assembly

Report Summary

Detailed report of de-novo genome assembly and advanced bioinformatic analyses to address biological questions

GENEWIZ Next Generation Sequencing

GENEWIZ, Inc.
115 Corporate Boulevard
South Plainfield, NJ 07080
p (877) GENEWIZ (436-3949)
f (908) 333-4511
www.genewiz.com

GENEWIZ NGS Analysis Report

Client:	Hollie Putnam
Institute/Company:	University of Rhode Island
Project:	30-323686303
Description of Services:	*de-novo* genome assembly
Sample Species:	**Porites astreoides**
Number of Samples:	1
Date:	June 3rd, 2020
GENEWIZ Contact:	Next Generation Sequencing Phone: 908-222-0711 x1 Email: ngs@genewiz.com

GENEWIZ, Inc.
115 Corporate Boulevard
South Plainfield, NJ 07080
p (877) GENEWIZ (436-3949)
f (908) 333-4511
www.genewiz.com

1 Description of workflow

1.1 Bioinformatics analysis workflow

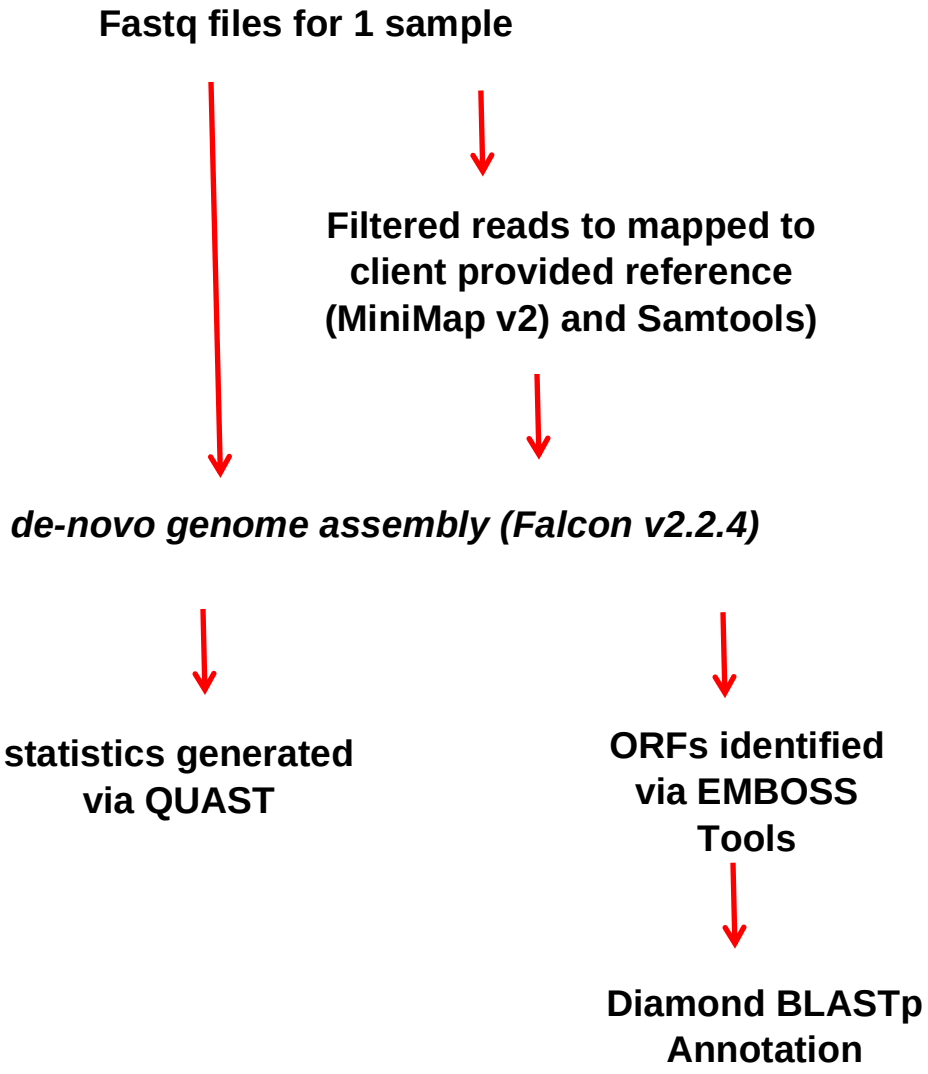


Figure 1.1. Bioinformatics analysis workflow

GENEWIZ, Inc.
115 Corporate Boulevard
South Plainfield, NJ 07080
p (877) GENEWIZ (436-3949)
f (908) 333-4511
www.genewiz.com

2 Data overview

2.1 PacBio sequel sequencing workflow

Purified DNA was fragmented and damage/end repaired. Adapters (containing unique barcode sequences if multiple samples were multiplexed and analyzed on one SMRT cell) were ligated to the DNA fragments. Prepared libraries were purified (samples were pooled together before purification if multiplexed). Sequencing primers were annealed to form polymerase-primer complexes and the resulting complexes were sequenced. Figure 2.1.1 outlines the PacBio library preparation and sequencing workflow.

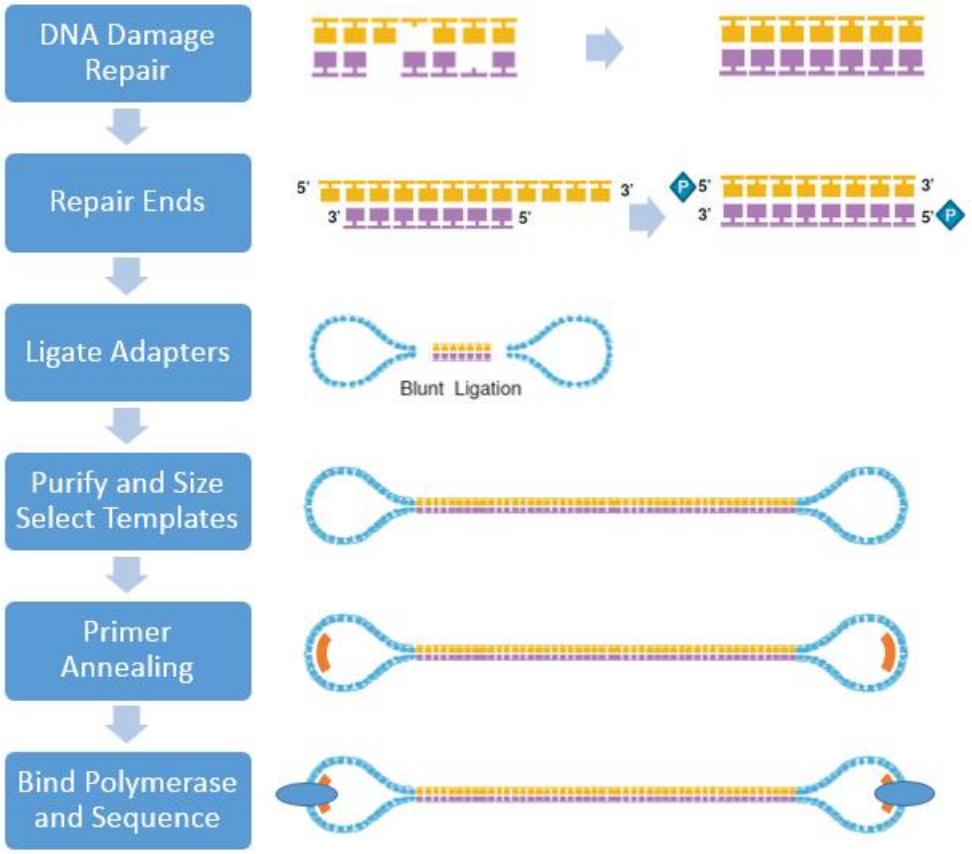


Figure 2.1.1 Pacbio library preparation and sequencing workflow

GENEWIZ, Inc.
115 Corporate Boulevard
South Plainfield, NJ 07080
p (877) GENEWIZ (436-3949)
f (908) 333-4511
www.genewiz.com

2.2 Polymerase read statistics

Meterics	Value
Number of Polymerase Reads	957,670
Total Polymerase Read Length	20,188,576,450
Mean Polymerase Read Length	21,081
Subread Length (mean)	12,850
Insert Length (mean)	12,965

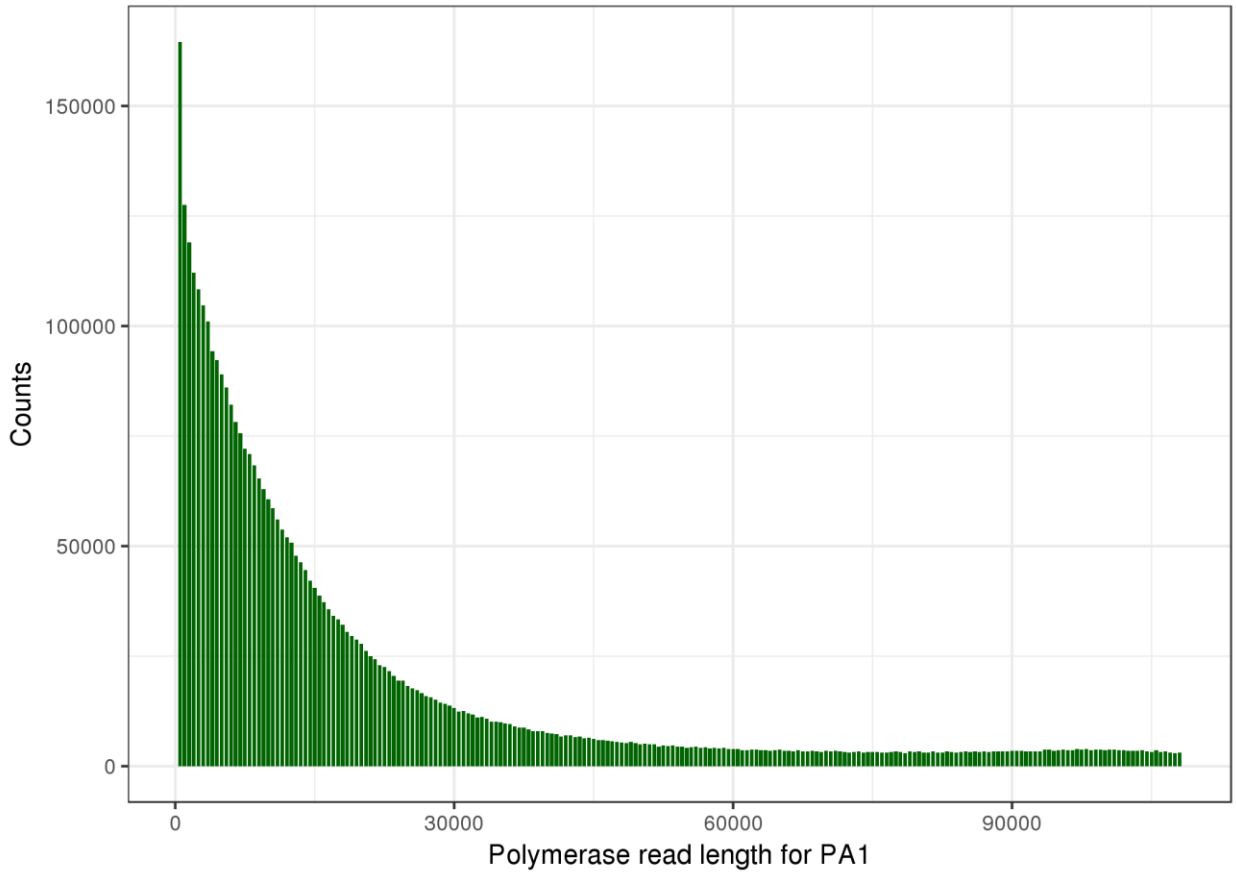
“Polymerase read”: A sequence of nucleotides incorporated by the DNA polymerase while reading a template. A polymerase read can contain several subreads.

“Subread”: Each polymerase read is partitioned to form one or more subreads, which contain sequences from a single pass of a polymerase on a single strand of an insert within a template.

“Insert length”: The length of the double-stranded nucleic acid insert in a SMRTbell template, excluding the hairpin adapters.

GENEWIZ, Inc.
115 Corporate Boulevard
South Plainfield, NJ 07080
p (877) GENEWIZ (436-3949)
f (908) 333-4511
www.genewiz.com

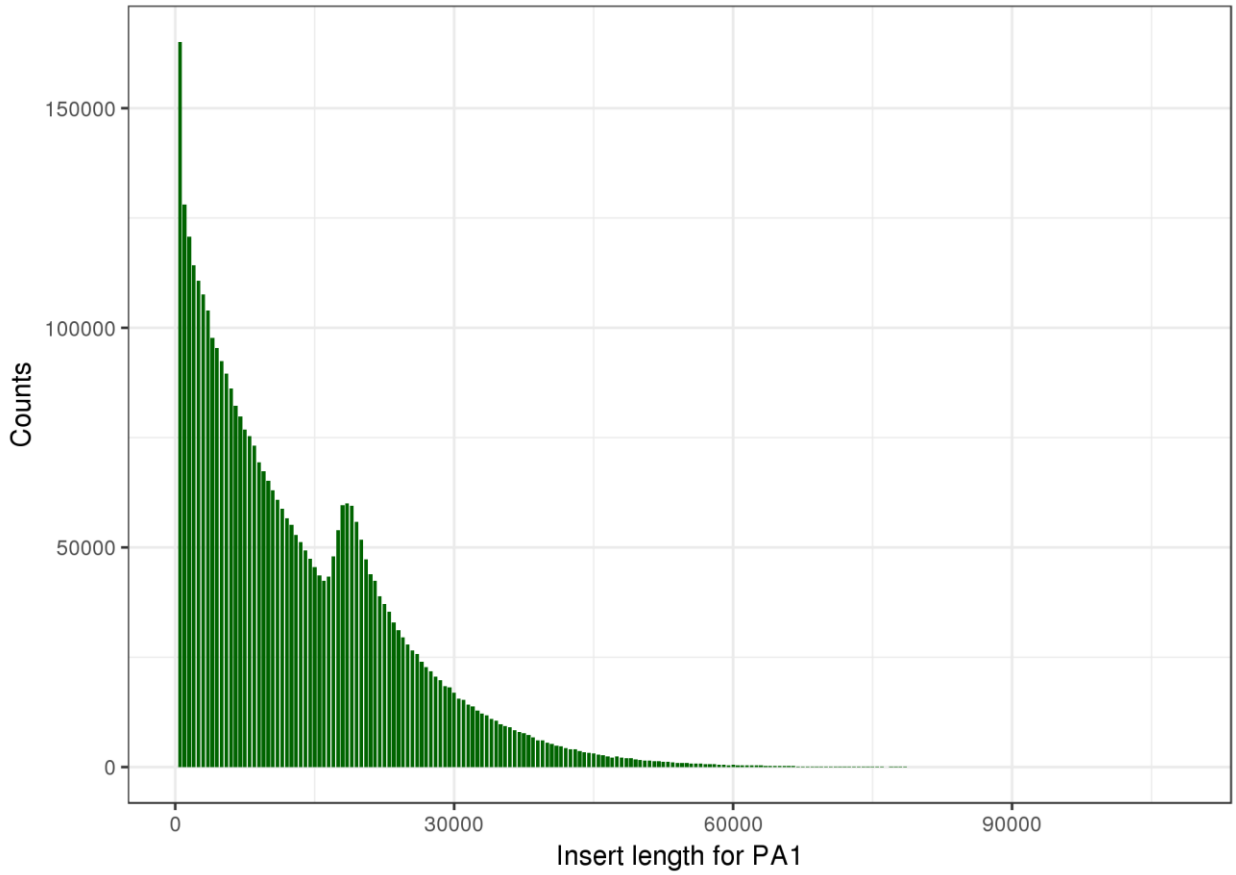
2.3 Polymerase read length distribution



Polymerase read length distribution is a metric related to the sequencing run performance. The adjusted polymerase read length is the total length of all subreads from a polymerase read. Barcode and SMRTbell adapter sequences are not included in the calculation. Therefore, it could be significantly different from the real polymerase read length if the inserts were short. Figure above shows distribution of adjusted polymerase read length.

GENEWIZ, Inc.
115 Corporate Boulevard
South Plainfield, NJ 07080
p (877) GENEWIZ (436-3949)
f (908) 333-4511
www.genewiz.com

2.4 Insert length distribution



Insert length distribution is another metric related to the sequencing run performance.

GENEWIZ, Inc.
115 Corporate Boulevard
South Plainfield, NJ 07080
p (877) GENEWIZ (436-3949)
f (908) 333-4511
www.genewiz.com

3. *De-novo* Genome Assembly

Falcon v2.2.4, de-novo assembler, was used on each of the samples (A. pre-filtered/unfiltered and B. filtered assemblies). The pre/unfiltered assembly used the fastq reads from the four smrt cells. Whereas, the filtered assembly used the reads from the 4 smrtcells that were mapped to client provided reference using MiniMap v2. Then the filtered fastq files were generated using samtools. Overall, de-novo assembled genome was created. QUAST was used to generate statistics for the de-novo assembled genome. EMBOSS tools getorf was then used to find the open reading frames within the de-novo assembled genome. The protein sequences from the open reading frames were then annotated using Diamond BLASTp.

3.1 Contig Length and Assembly Statistics

Statistics	filtered_p_ctg	unfiltered_p_ctg
# contigs	3,051	3,095
# contigs (>= 0 bp)	3,051	3,095
# contigs (>= 1000 bp)	3,051	3,095
# contigs (>= 5000 bp)	3,048	3,093
# contigs (>= 10000 bp)	3,030	3,070
# contigs (>= 25000 bp)	2,866	2,905
# contigs (>= 50000 bp)	2,429	2,432
Largest contig	3,369,715	2,514,820
Total length	677,753,397	680,005,989
Total length (>= 0 bp)	677,753,397	680,005,989
Total length (>= 1000 bp)	677,753,397	680,005,989
Total length (>= 5000 bp)	677,742,019	679,999,105
Total length (>= 10000 bp)	677,611,785	679,826,819
Total length (>= 25000 bp)	674,350,533	676,554,116
Total length (>= 50000 bp)	658,123,657	658,911,805
N50	412,256	408,687
N75	189,769	190,323
L50	437	444
L75	1,049	1,055
GC (%)	39.12%	39.14%
# N's	0	0
# N's per 100 kbp	0	0

GENEWIZ, Inc.
115 Corporate Boulevard
South Plainfield, NJ 07080
p (877) GENEWIZ (436-3949)
f (908) 333-4511
www.genewiz.com

4 Deliverables

1. *De-novo* assembled genome in fasta file format (1 fasta file per assembly)
2. Annotation table of the *de-novo* assembled genome contigs from Diamond BLASTp (1 tsv file per assembly)
3. Open reading frames – nucleotide sequences (1 fasta file per assembly; 1 tsv file per assembly)
4. Open reading frames – protein sequences (1 fasta file per assembly; 1 tsv file per assembly)
5. Data analysis report (1 PDF File)
6. Raw Fastq; Fasta; and BAM files