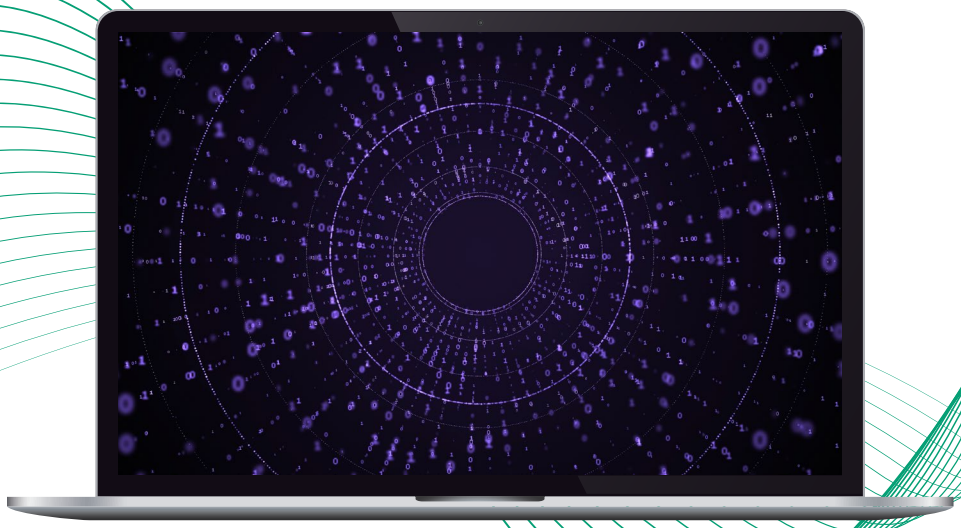
GUIDE

Plasmid-EZ Quick Start Guide



GENEWIZ®
From Azenta Life Sciences

genewiz.com

Table of Contents

1. Summary of Deliverables 03
2. Project Report Review 04
Viewing the Project Report
3. Sample Report Evaluation 05
Sample Report
Analysis Results
Annotation Map
Annotation Table
Data Quality Control
Failed Assembly Report
4. Assembled Contig Review 09
Genbank
Assembly Fasta
Assembly Fastq
Assembly Ab1
5. Sample Coverage Assessment 11
Variation
6. Reference-Based Analysis 13
Coverage Analysis
Variants
7. Common Factors Affecting Plasmid-EZ Assembly 15
Concentration and Sample Quality
Homopolymeric Repetitive Regions
Methylation Sites

1 Summary of Deliverables

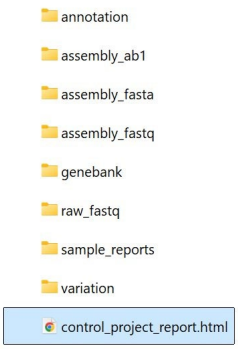
control_project_report.html	Project summary report in HTML format
raw_fastq	Raw data in zipped fastq format
assembly_fasta	Assembled sequence(s) in fasta format
assembly_fastq	Assembled sequence(s) in fastq format
genbank	Annotation in GenBank format
annotation	Annotated features in CSV format
assembly_ab1	Assembled sequence(s) in abI format
variation	• Base count from the alignment of the raw reads to the final consensus sequence • Variant calling results in VCF format • Index file for the VCF file • Variant calling results in tab-delimited text format
sample_reports	Sample report in HTML format

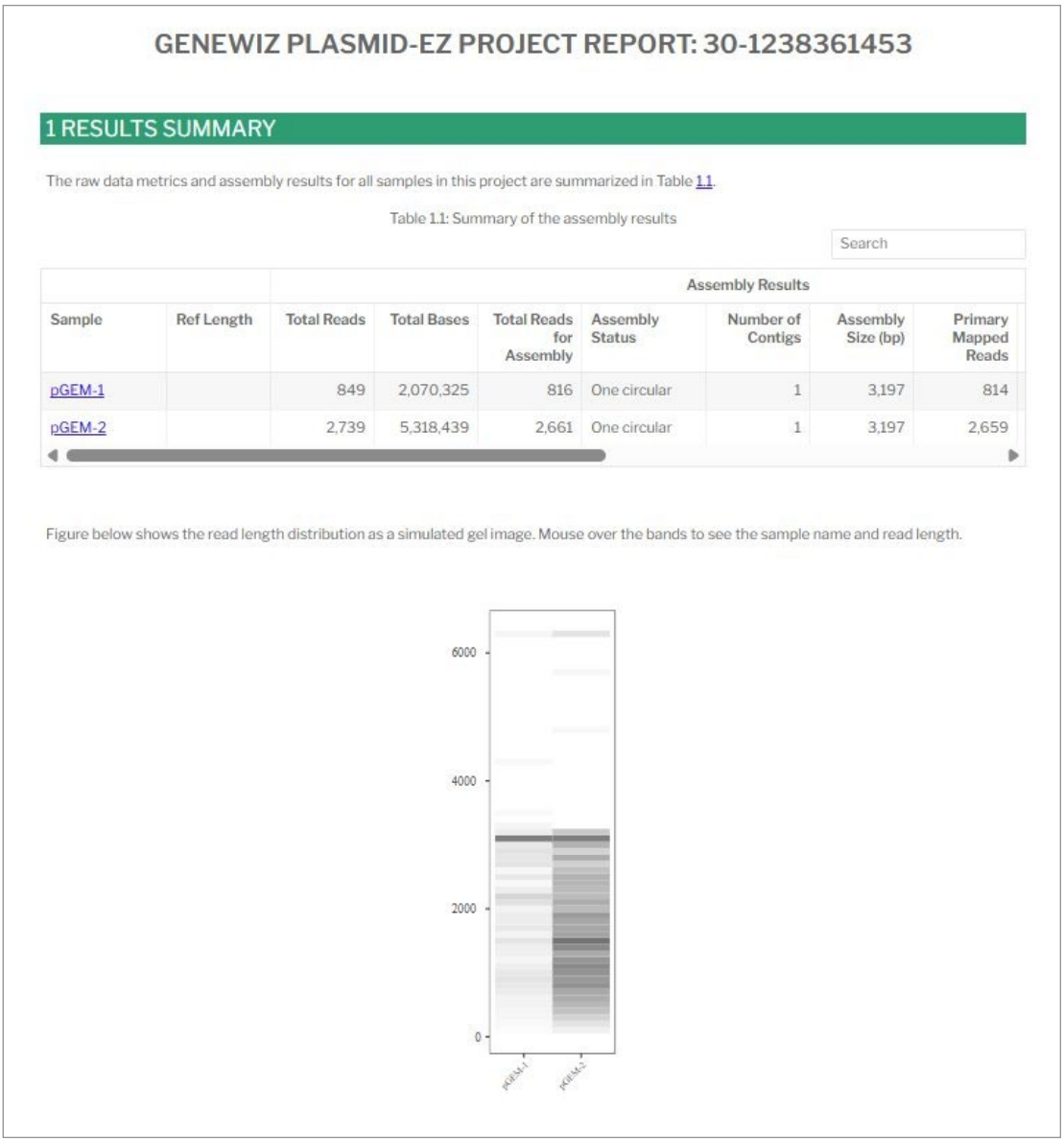
2 Project Report Review

VIEWING PROJECT REPORT

Start by clicking on the 30-xxxxxxxxx_report.html file.
The file provides a results summary highlighting the read counts, assembly status, number of contigs isolated, mapped reads percentage, etc.

A virtual gel is provided to show the read length distribution for each sample.





GENEWIZ PLASMID-EZ PROJECT REPORT: 30-1238361453

1 RESULTS SUMMARY

The raw data metrics and assembly results for all samples in this project are summarized in Table 1.1.

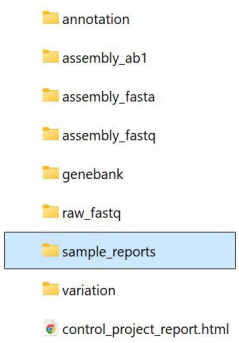
Table 1.1: Summary of the assembly results

Search

		Assembly Results						
Sample	Ref Length	Total Reads	Total Bases	Total Reads for Assembly	Assembly Status	Number of Contigs	Assembly Size (bp)	Primary Mapped Reads
pGEM-1		849	2,070,325	816	One circular	1	3,197	814
pGEM-2		2,739	5,318,439	2,661	One circular	1	3,197	2,659

Figure below shows the read length distribution as a simulated gel image. Mouse over the bands to see the sample name and read length.

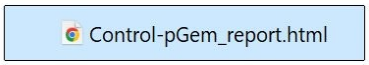
3 Sample Report Evaluation



SAMPLE REPORT

Individual sample reports can be accessed from the sample_reports folder.

Click the individual files to access the reports.



ANALYSIS RESULTS

The sample report lists the number of contigs assembled, their lengths, topology, mapped bases, total mapped base percentage and average coverage of the assembly.

If multiple contigs are identified, separate line items will be provided per assembled contig, including assembly information.

The most common topology will be circular, linear or failed. All possible topology are listed below:

- Circular → closed loop contig assembled
- Linear → two-ended contig assembled
- Multiple circular → 1circular contigs assembled
- Mixed → linear and circular contigs assembled
- Multiple linear → 1 linear contig assembled
- Failed → no assembly

Assembly QC is also provided, which shows the coverage by base position.

Any position of low coverage can easily be identified from the chart for further review in the ab1 or variation files.

GENEWIZ SAMPLE REPORT

1 SAMPLE INFORMATION

Quote 30-1245406959
Sample pGEM

2 ANALYSIS RESULTS

2.1 Results summary

Preprocessed reads (see "Data quality control" section) were used for assembly. If the assembly is successful, the draft assembly is polished using the preprocess reads by using Medaka. The preprocessed reads were then aligned to the polished assembly using Minimap2 and small variants were called using Clair3.

Table 2.1 shows the assembly summary. Note that the `Mapped Bases` are based on the covered bases on the assembly, so it could be slightly more than the total bases in the preprocessed reads.

Table 2.1: Summary of the assembly results

Sample	Number of Contigs	Total Length	Topology	Mapped Bases	% of Total Bases	Average Coverage
pGEM	1	3,197	circular	2,083,144	99.36%	652

2.3 Assembly QC

2.3.1 Coverage along the assembly

Figure 2.1 shows the coverage of the assembly when reads are aligned back to the assembly.

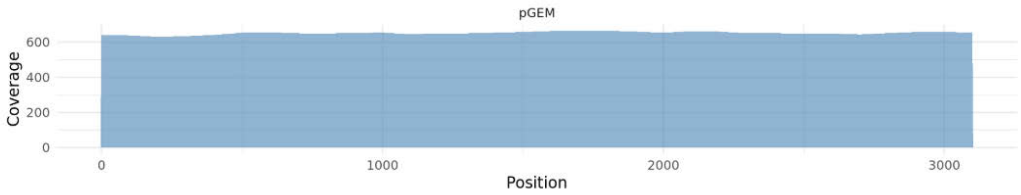


Figure 2.1: Coverage along the assembly

ANNOTATION MAP

The sample report also provides an interactive annotation plasmid map.

ANNOTATION TABLE

Within the report, below the plasmid map, is an annotation table with detailed information on each featured annotation segment. This table is optimal for reviewing the identity and length-matched percentage for each feature.

2.2.2 Annotation Table

Table 2.2 shows the feature annotation details.

Table 2.2: Annotated features in the assembly

Show 10 entries Search:

Contig	Feature	Type	Start	End	Strand	% Identity	% Matched Length	Description
pGEM	f1 ori	rep_origin	853	1,282	-1	100.0	100.0	f1 bacteriophage origin of replication; arrow indicates direction of (+) strand synthesis
pGEM	AmpR promoter	promoter	390	495	-1	100.0	100.0	bla
pGEM	AmpR	CDS	2,726	390	-1	99.8	100.0	β-lactamase; bla; confers resistance to ampicillin, carbenicillin, and related antibiotics
pGEM	ori	rep_origin	1,967	2,556	-1	99.8	100.0	high-copy-number ColE1/pMB1/pBR322/pUC origin of replication

DATA QUALITY CONTROL

The data quality control section in the sample report provides the length and quality distribution of the raw reads.

As seen on the right, the left graph shows the read count by read length. Optimal results will show a single peak for the expected contig length with base level degradation (read length lower than 2 kB). The graph to the right shows the quality for the raw reads. For high quality samples, the peak of the graph will veer toward QS of 20-30.

3 DATA QUALITY CONTROL

3.1 Raw data QC

Raw reads were trimmed to remove adapter and barcode sequences, and reads less than 100 bp were discarded.

E. coli contamination is estimated based on the percentage of reads mapped to the *E. coli* reference genome (NC_000913.3). A high percentage of *E. coli* contamination may indicate issues during sample preparation. However, this estimation may not be accurate since some plasmid sequences may share homology with the *E. coli* genome.

Table 3.1 shows the raw data quality control summary.

Table 3.1: Raw data quality control summary

Sample	Total Bases	Number of Reads	Mean Quality Score	Estimated E. coli Sequences
pGEM	2,119,892	1,019	19.6	1.10%

Figure 3.1 shows the length and quality distribution of the raw reads after adapter trimming.

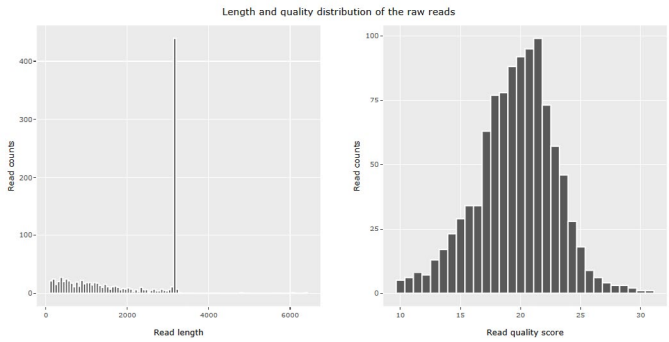


Figure 3.1: Length and quality distribution of the raw reads

FAILED ASSEMBLY REPORT

For samples that fail assembly, only the project report, sample report and raw_fastq folders are provided. The example below shows the data provided for a failed sample report.

The length and quality distribution of raw reads will be provided dependent on the total reads attained for the sample.

As the QC for the failed sample shows below, only 3 reads were attained at varied length with low QC. The total reads, in this case, were insufficient to produce an assembly.

3 DATA QUALITY CONTROL

3.1 Raw data QC

Raw reads were trimmed to remove adapter and barcode sequences, and reads less than 100 bp were discarded.

E. coli contamination is estimated based on the percentage of reads mapped to the *E. coli* reference genome (NC_000913.3). A high percentage of *E. coli* contamination may indicate issues during sample preparation. However, this estimation may not be accurate since some plasmid sequences may share homology with the *E. coli* genome.

Table 3.1 shows the raw data quality control summary.

Table 3.1: Raw data quality control summary

Sample	Total Bases	Number of Reads	Mean Quality Score	Estimated E. coli Sequences
PL40361-001-1	7,121	3	16.7	0.00%

Figure 3.1 shows the length and quality distribution of the raw reads after adapter trimming.

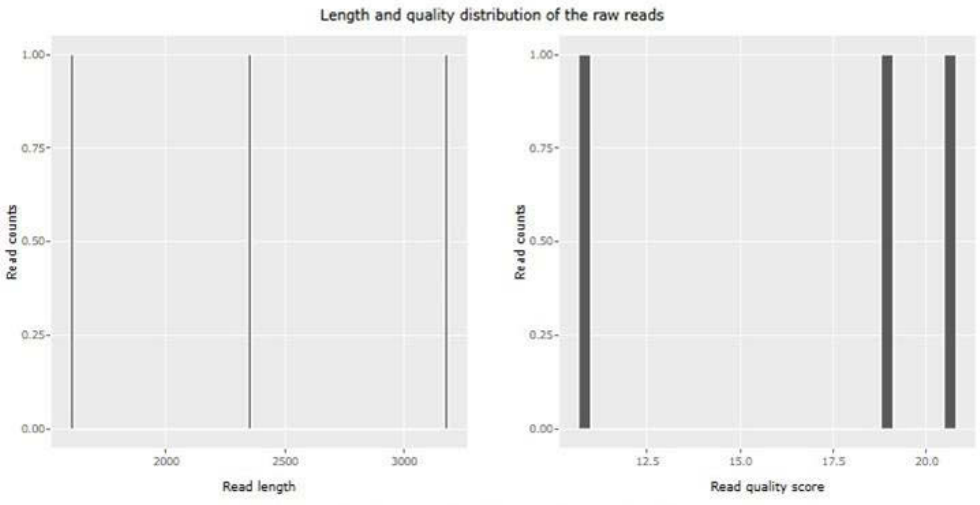


Figure 3.1: Length and quality distribution of the raw reads

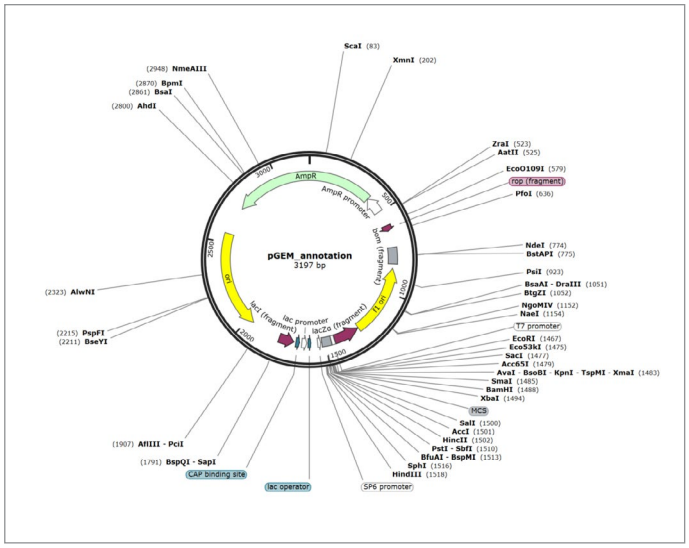
4 Assembled Contig Review

GENBANK

The GenBank folder contains the annotated assembled contig.

ASSEMBLY FASTA

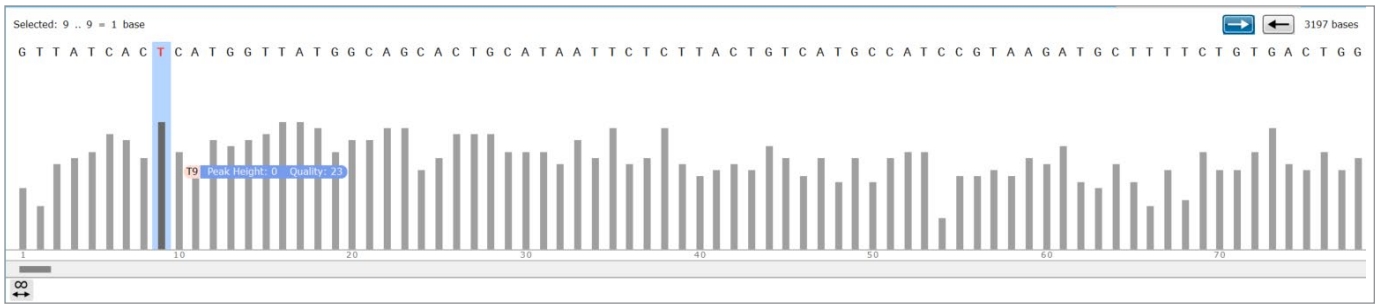
Review the assembled contig from the files contained in the assembly_fasta folder.



ASSEMBLY FASTQ

Scroll though the assembled fastq files to quickly review the quality by base position for the assembled contig.

Hovering over each peak provides the base position and quality score.

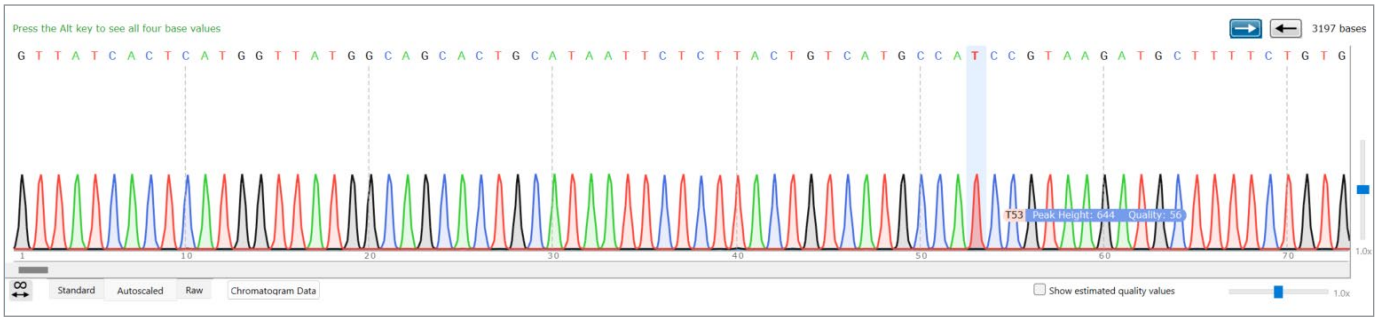


ASSEMBLY AB1

For Sanger-like data, an ab1 file is now available in a single file for ease of review. Mixed base positions can easily be identified by scrolling through and checking for any multiple peaks for single base positions. Easily identify any secondary peak with the color coding provided: A = green, T = red; G = black, C = blue.

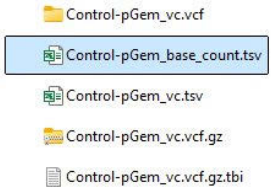
As shown below, hovering over each position provides the called base and quality score.

Mixed base positions identified in the ab1 file can be further reviewed in the variation Excel file by searching for the position to evaluate exact read counts of the mixed bases.



5 Sample Coverage Assessment

VARIATION

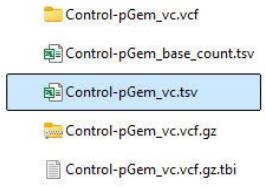


The variation folder provides information on the variation of base calling in several different formats.

The {sample_name}_base_count.tsv file provides the coverage score by base positions. The file also provides the spread of base calling in the assembled contig.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Contig	Position	Reference	Coverage	A	T	G	C	N	Insertions	Top Inserti	Deletions
2	pGEM	1	G	622	0	0	622	0	0	0	-	0
3	pGEM	2	T	622	0	622	0	0	0	0	-	0
4	pGEM	3	T	622	0	622	0	0	0	0	-	0
5	pGEM	4	A	623	623	0	0	0	0	1	T (1)	0
6	pGEM	5	T	625	0	625	0	0	0	0	-	0
7	pGEM	6	C	626	0	0	0	626	0	0	-	0
8	pGEM	7	A	626	626	0	0	0	0	0	-	0
9	pGEM	8	C	626	0	0	1	623	0	0	-	2
10	pGEM	9	T	625	0	624	1	0	0	0	-	0
11	pGEM	10	C	627	0	0	5	622	0	2	T (2)	0
12	pGEM	11	A	629	629	0	0	0	0	0	-	0
13	pGEM	12	T	630	0	630	0	0	0	0	-	0
14	pGEM	13	G	630	0	0	630	0	0	0	-	0
15	pGEM	14	G	631	0	0	630	1	0	0	-	0
16	pGEM	15	T	631	0	631	0	0	0	0	-	0

To easily identify any mixed base, high deletion or high insertion positions, we recommend evaluating the {sample_name}_vc.tsv file. Any base position with more than 5% variability compared to the assembled contig will be highlighted in the file.



	A	B	C	D	E	F	G	H
1	CHROM	POS	REF	ALT	GT	DP	AD	AF
2								
3								

The meaning of each acronym is listed below.

Acronym	
CHROM	Chromosome (sample name)
POS	Position
REF	Reference Base
ALT	Variant Base

Acronym	
GT	Genotype
DP	Total Depth
AD	Allele Depth
AF	Allele Frequency

For the example listed below, the information provided can be interpreted as follows:

Sample-1 shows variations at positions 5970 and 6556. The primary base for both positions are G with alternative base A. For position 5970, there is coverage of 144 for base G and 46 for base A, which is a 22% variation for the position. For position 6556, the coverage for G is 130 and coverage for A is 54, which is about 26% variation for this position.

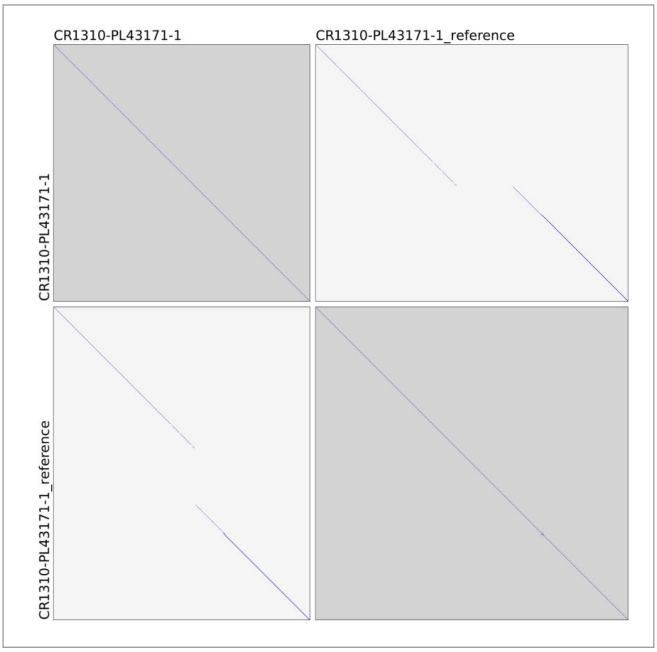
	A	B	C	D	E	F	G	H
1	CHROM	POS	REF	ALT	GT	DP	AD	AF
2	sample-1	5970	G	A	0/1	205	144,46	0.2244
3	sample-2	6556	G	A	0/1	202	130,54	0.2673

6 Reference-Based Analysis

If the reference file is provided on the order form, reference-based analysis will be run after the de novo assembly, to remove bias from the reference file.

The sample report will show a dotplot and a separate coverage plot comparing the alignment between the assembled contig and reference file.

The dotplot and the coverage plot are great tools for quick check on the alignment between the assembled contig and the reference file.



COVERAGE ANALYSIS

The coverage plot provided for reference-based analysis shows alignment between the assembled contig and the reference file.

In the example shown below, there is high coverage between the assembled contig and reference file except for about a 1.5 kB missing region from the assembled contig. Possible interpretation for the results is that the insertion region for the sample did not integrate as expected.

2.4 Reference Based Analysis

A refernce was provided for this sample. Preprocessed reads were aligned to the reference using Minimap2 and coverage along the reference was calculated using Mosdepth. Figure 2.3 shows the coverage of the reference sequence.

Variants were called using Clair3 and results are saved in the ref_based_analysis folder.

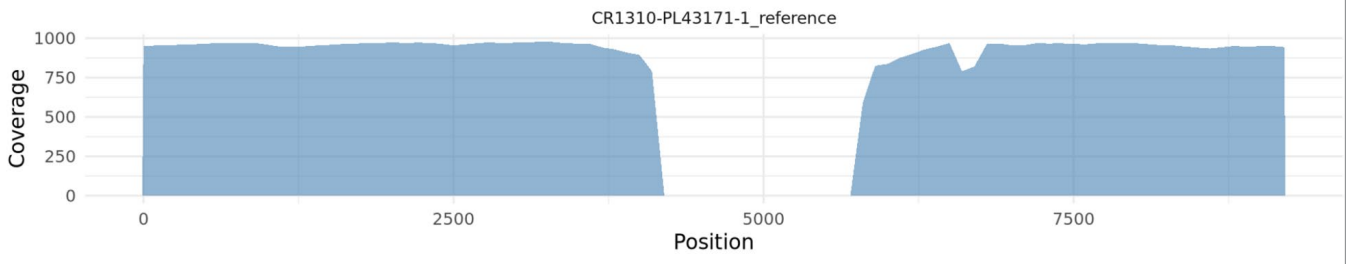
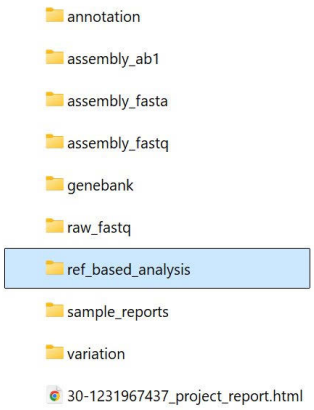


Figure 2.3: Coverage along the reference

VARIANTS

The reference-based analysis folder provides various excel files to compare the assembled contig to the reference file. Below is a snapshot of the all the files provided per sample.



Control_asm_vs_ref.var.txt	Variations between assembly and reference file.
Control_vs_ref_base_count.tsv	Base calling and count comparison between contig and reference file.
Control_vs_ref_vc.tsv	Variant calling results in tab delimited text format.
Control_vs_ref_vc.vcf.gz	Variant calling results based on the reference in VCF format.
Control_vs_ref_vc.vcf.gz.tbi	Index file for the VCF file.

We recommend viewing the vs_ref.var.txt file to review the marked variants between the assembly and the reference file. To open the file in Excel, first extract the folder and then drag the file into the Excel search window.

The image below shows an example variant file and lists the variants by base positions. Interpretation of a few base positions is as follows:

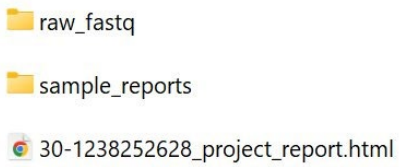
For positions 3484-3485, the reference file shows base calling C but the assembled contig contains G.
For positions 4188-5829, the reference file shows base calling but the assembled contig shows this region as missing.

	A	B	C	D	E	F	G	H	I	J	K	L
1	flag	ref	start	end	depth	mapq	ref_allele	alt_allele	contig_name	contig_start	contig_end	contig_strand
2	R	CR1310-PL43171-1 reference	0	9253								
3	V	CR1310-PL43171-1 reference	3484	3485	1	60	c	g	CR1310-PL43171-1	3484	3485	
4	V	CR1310-PL43171-1 reference	3485	3486	1	60	t	a	CR1310-PL43171-2	3485	3486	
5	V	CR1310-PL43171-1 reference	3486	3487	1	60	g	c	CR1310-PL43171-3	3486	3487	
6	V	CR1310-PL43171-1 reference	3487	3488	1	60	t	a	CR1310-PL43171-4	3487	3488	
7	V	CR1310-PL43171-1 reference	3489	3490	1	60	g	c	CR1310-PL43171-5	3489	3490	
8	V	CR1310-PL43171-1 reference	3733	3734	1	60	a	g	CR1310-PL43171-6	3733	3734	
9	V	CR1310-PL43171-1 reference	3734	3735	1	60	a	g	CR1310-PL43171-7	3734	3735	
10	V	CR1310-PL43171-1 reference	3735	3736	1	60	g	c	CR1310-PL43171-8	3735	3736	
11	V	CR1310-PL43171-1 reference	3736	3737	1	60	c	g	CR1310-PL43171-9	3736	3737	
12	V	CR1310-PL43171-1 reference	4085	4086	1	60	a	g	CR1310-PL43171-10	4085	4086	
13	V	CR1310-PL43171-1 reference	4090	4091	1	60	c	a	CR1310-PL43171-11	4090	4091	
14	V	CR1310-PL43171-1 reference	4091	4092	1	60	g	c	CR1310-PL43171-12	4091	4092	
15	V	CR1310-PL43171-1 reference	4188	5829	1	60	cggcggcga	_	CR1310-PL43171-13	4188	4188	
16	V	CR1310-PL43171-1 reference	6636	6637	1	60	a	t	CR1310-PL43171-14	4995	4996	
17	V	CR1310-PL43171-1 reference	6638	6643	1	60	taaaa	_	CR1310-PL43171-15	4997	4997	
18	V	CR1310-PL43171-1 reference	6667	6674	1	60	aaaaaaaaa	_	CR1310-PL43171-16	5021	5021	
19	V	CR1310-PL43171-1 reference	6706	6718	1	60	aaaaaaaaaaa	_	CR1310-PL43171-17	5053	5053	

7 Common Factors Affecting Plasmid-EZ Assembly

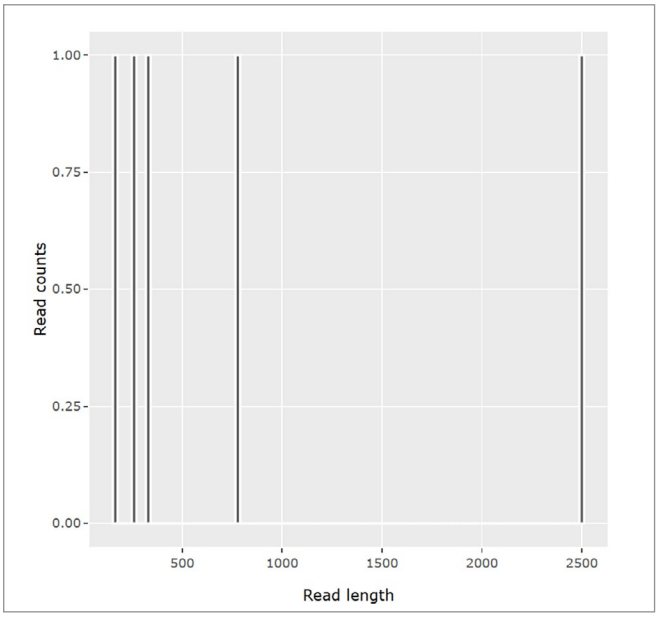
CONCENTRATION AND SAMPLE QUALITY

In the unfortunate event your sample failed to produce an assembly, the sample folder will only contain the Project Report, Sample Report and raw FASTQ reads.

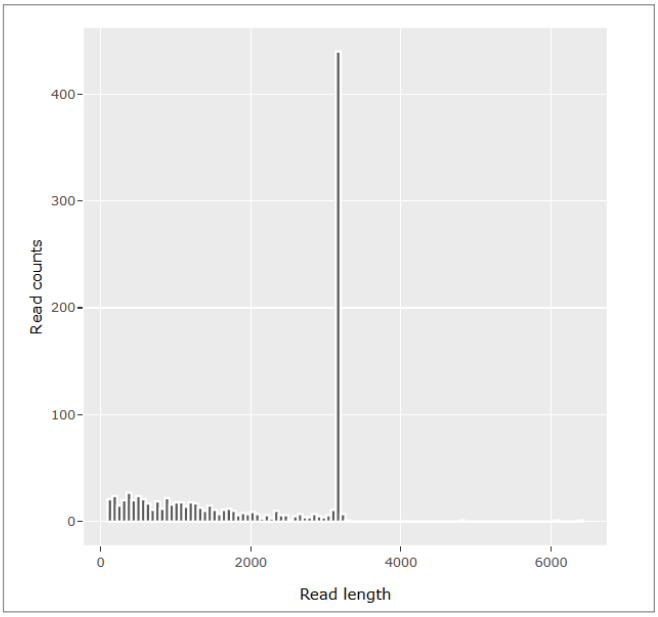


To provide a fast turnaround time at a low cost, we do not perform sample QC to determine why samples failed assembly. However, the most common reason for failure is the sample not meeting the required 50 ng/ul concentration. Low concentrations may lead to increased fragmentation during library preparation and/or a low number of reads generated for the sample. We strongly recommend checking the concentration of your samples on a Qubit or equivalent before sending samples to us to reduce the chance of failure.

Below on the left is the read length graph for a sample that failed and on the right is one that worked. Samples with a clear plasmid peak like on the right tend to assemble, while samples with low read count and no clear contig tend to fail assembly.



The graph above shows presence of three lengths with single read count. The read counts and distribution is inadequate for assembly and suggests low sample concentration.



The length distribution in the graph above shows a primary peak around 3 kB with very low read count or other fragments. This distribution highlights the presence of a single contig at high quality.

HOMOPOLYMERIC REPETITIVE REGIONS

Plasmid-EZ may at times show insertion/deletion following a homopolymeric repetitive region. In such cases, please review the variation file for the position to confirm possible insertion/deletion at the region. The example below showcases insertion of an additional A at the start of the homopolymeric region. For further confirmation, we recommend running a Sanger sequencing order for the regions to ensure complete coverage.

The base_count.tsv example below shows a homopolymeric repetitive region. Though a few positions show possible insertions, only position 5053 contains a significant number of insertions.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Contig	Position	Reference	Coverage	A	T	G	C	N	Insertions	Top Insertion	Deletions
2	CR1310-PL43171-1	5052	T	776	55	718	0	0	0	10	AAA (2)	3
3	CR1310-PL43171-2	5053	T	779	84	682	1	0	0	267	AAA (50)	12
4	CR1310-PL43171-3	5054	A	903	569	3	0	0	0	0	_	331
5	CR1310-PL43171-4	5055	A	619	619	0	0	0	0	0	_	0
6	CR1310-PL43171-5	5056	A	666	666	0	0	0	0	0	_	0
7	CR1310-PL43171-6	5057	A	697	697	0	0	0	0	0	_	0
8	CR1310-PL43171-7	5058	A	736	736	0	0	0	0	1	T (1)	0
9	CR1310-PL43171-8	5059	A	762	761	1	0	0	0	3	TTTT (2)	0
10	CR1310-PL43171-9	5060	A	785	783	2	0	0	0	0	_	0
11	CR1310-PL43171-10	5061	A	798	797	1	0	0	0	0	_	0
12	CR1310-PL43171-11	5062	A	813	811	2	0	0	0	0	_	0
13	CR1310-PL43171-12	5063	A	828	825	3	0	0	0	1	TT (1)	0
14	CR1310-PL43171-13	5064	A	842	836	6	0	0	0	1	TTTTT (1)	0

The vc.tsv for the same sample only highlights position 5053, as this position showed significant insertion. According to this chart, position 5053 is T but an alternative for this position is TAAA with about 5% variation.

	A	B	C	D	E	F	G	H
1	CHROM	POS	REF	ALT	GT	DP	AD	AF
2	CR1310-PL43171-1	5053	T	TAAA	0/1	884	80,46	0.052

METHYLATION SITES

GATC, CCAGG and CCTGG sites are methylation sites in *E. coli*. These sites may show discrepancy and will require review of the variation file. If discrepancy is due to methyl GATC site, the variation file will show half A/half G for the position. In such cases, Sanger sequencing follow-up may be needed for further confirmation of the base position.

HAVE QUESTIONS? Contact us.

Email DNAseq@azenta.com **Call 1-888-229-3682 ext 2** **Or reach out on LiveChat**