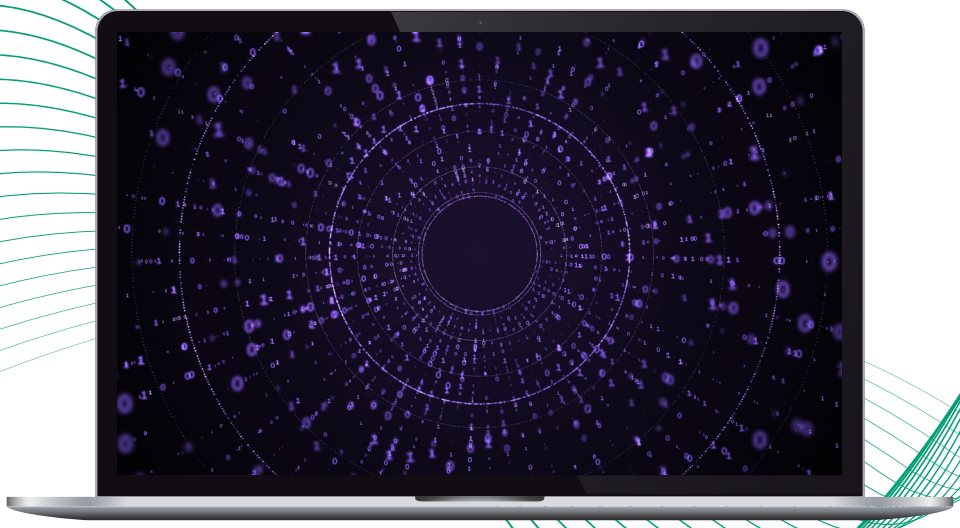
GENEWIZ®
From Azenta Life Sciences

GUIDE

Plasmid-EZ Quick Start Guide



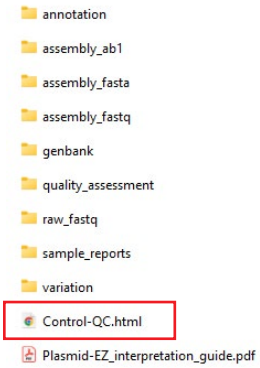
AZENTA
LIFE SCIENCES

Table of Contents

1 Getting Started

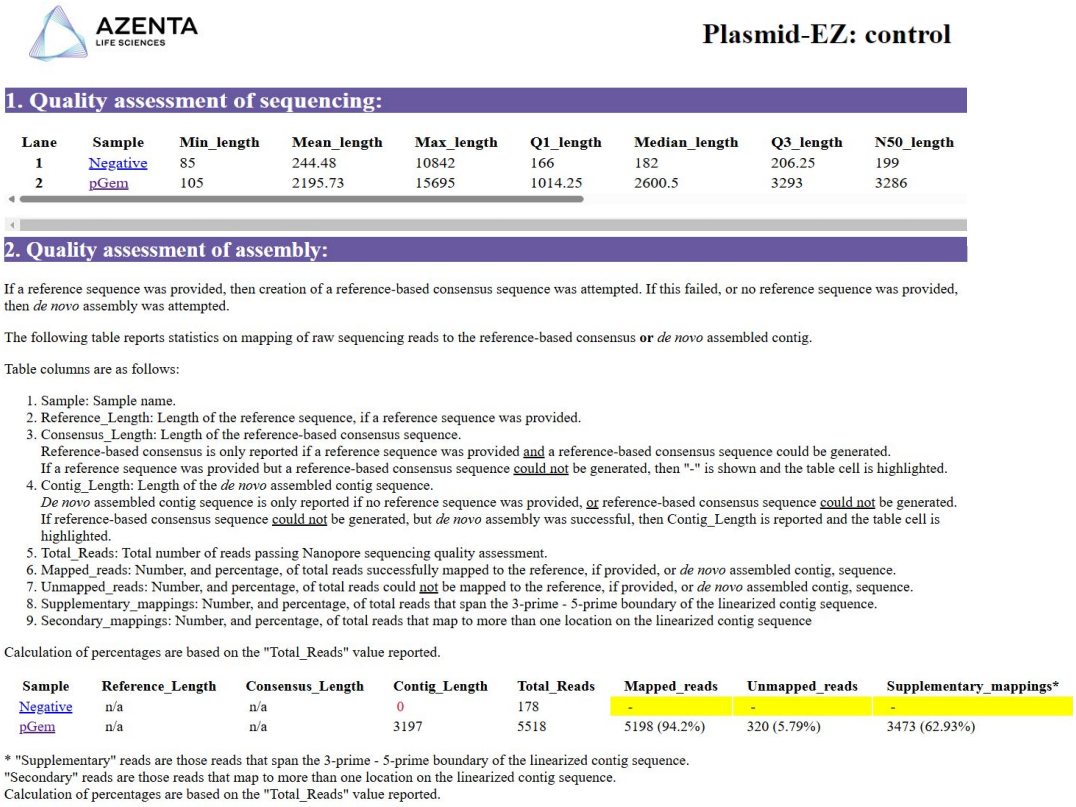
VIEWING PROJECT REPORT

To start, click on the *30-xxxxxxxxx.QC.html* file. This file provides an overall QC report, has links for individual sample reports, and acts as a launching point for accessing all reports.



OPENING SAMPLE REPORTS

From the QC report, you can click on the sample name to open the sample report.



AZENTA LIFE SCIENCES

Plasmid-EZ: control

1. Quality assessment of sequencing:

Lane	Sample	Min_length	Mean_length	Max_length	Q1_length	Median_length	Q3_length	N50_length
1	Negative	85	244.48	10842	166	182	206.25	199
2	pGem	105	2195.73	15695	1014.25	2600.5	3293	3286

2. Quality assessment of assembly:

If a reference sequence was provided, then creation of a reference-based consensus sequence was attempted. If this failed, or no reference sequence was provided, then *de novo* assembly was attempted.

The following table reports statistics on mapping of raw sequencing reads to the reference-based consensus **or** *de novo* assembled contig.

Table columns are as follows:

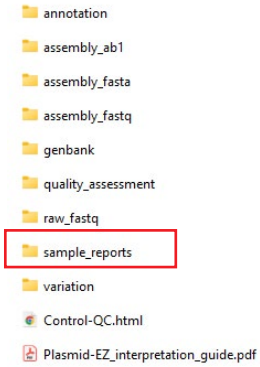
1. Sample: Sample name.
2. Reference_Length: Length of the reference sequence, if a reference sequence was provided.
3. Consensus_Length: Length of the reference-based consensus sequence.
   Reference-based consensus is only reported if a reference sequence was provided and a reference-based consensus sequence could be generated.
   If a reference sequence was provided but a reference-based consensus sequence could not be generated, then "-" is shown and the table cell is highlighted.
4. Contig_Length: Length of the *de novo* assembled contig sequence.
   De novo assembled contig sequence is only reported if no reference sequence was provided, or reference-based consensus sequence could not be generated.
   If reference-based consensus sequence could not be generated, but *de novo* assembly was successful, then Contig_Length is reported and the table cell is highlighted.
5. Total_Reads: Total number of reads passing Nanopore sequencing quality assessment.
6. Mapped_reads: Number, and percentage, of total reads successfully mapped to the reference, if provided, or *de novo* assembled contig, sequence.
7. Unmapped_reads: Number, and percentage, of total reads could not be mapped to the reference, if provided, or *de novo* assembled contig, sequence.
8. Supplementary_mappings: Number, and percentage, of total reads that span the 3-prime - 5-prime boundary of the linearized contig sequence.
9. Secondary_mappings: Number, and percentage, of total reads that map to more than one location on the linearized contig sequence

Calculation of percentages are based on the "Total_Reads" value reported.

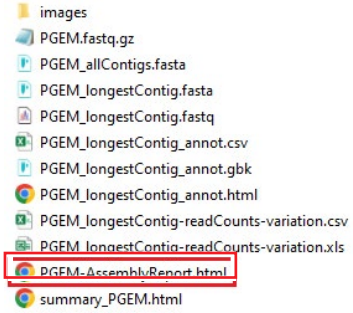
Sample	Reference_Length	Consensus_Length	Contig_Length	Total_Reads	Mapped_reads	Unmapped_reads	Supplementary_mappings*
Negative	n/a	n/a	0	178	-	-	-
pGem	n/a	n/a	3197	5518	5198 (94.2%)	320 (5.79%)	3473 (62.93%)

* "Supplementary" reads are those reads that span the 3-prime - 5-prime boundary of the linearized contig sequence.
"Secondary" reads are those reads that map to more than one location on the linearized contig sequence.
Calculation of percentages are based on the "Total_Reads" value reported.

Individual sample reports can also be accessed by going to the sample folder.

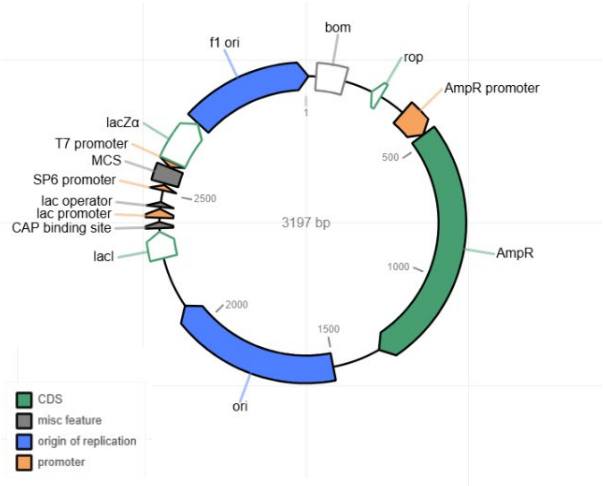


Then, click on the *AssemblyReport.html* file.



2 Viewing Your Plasmid

ANNOTATION MAP OF LONGEST CONTIG



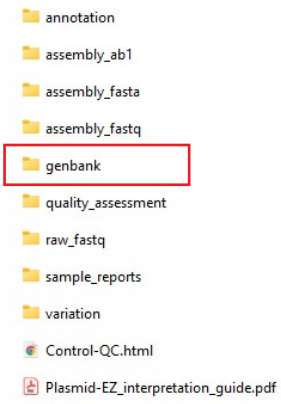
In the sample report, the first thing you will see is an annotated plasmid map for the longest contig assembled. Hovering over the map will pull up a summary of the region, which is also listed in the table below the map.

Data extracted from /output/pGem_contig_annotation.csv

Feature	Type	percent identity	percent match length	Description
f1 ori	rep_origin	100.0	100.0	f1 bacteriophage origin of replication; arrow indicates direction of (+) strand synthesis
AmpR promoter	promoter	100.0	100.0	bla
AmpR	CDS	99.76	100.0	β-lactamase; bla; confers resistance to ampicillin
ori	rep_origin	99.83	100.0	high-copy-number ColE1/pMB1/pBR322/pUC origin of replication
MCS	misc_feature	100.0	100.0	pUC18/19 multiple cloning site
lac promoter	promoter	100.0	100.0	promoter for the E. coli lac operon
CAP binding site	protein_bind	100.0	100.0	CAP binding activates transcription in the presence of cAMP. E. coli catabolite activator protein
T7 promoter	promoter	100.0	100.0	promoter for bacteriophage T7 RNA polymerase
SP6 promoter	promoter	100.0	100.0	promoter for bacteriophage SP6 RNA polymerase
lac operator	protein_bind	100.0	100.0	The lac repressor binds to the lac operator to inhibit transcription in E. coli. This inhibition can be relieved by adding lactose or isopropyl-β-D-thiogalactopyranoside (IPTG). lac repressor encoded by lacI
lacZα	CDS	100.0	90.80	LacZα fragment of β-galactosidase; lacZ fragment
bom	misc_feature	100.0	70.92	basis of mobility region from pBR322
lacI	CDS	100.0	8.58	lac repressor; lacI; The lac repressor binds to the lac operator to inhibit transcription in E. coli. This inhibition can be relieved by adding lactose or isopropyl-β-D-thiogalactopyranoside (IPTG).
rop	CDS	100.0	18.75	Rop protein

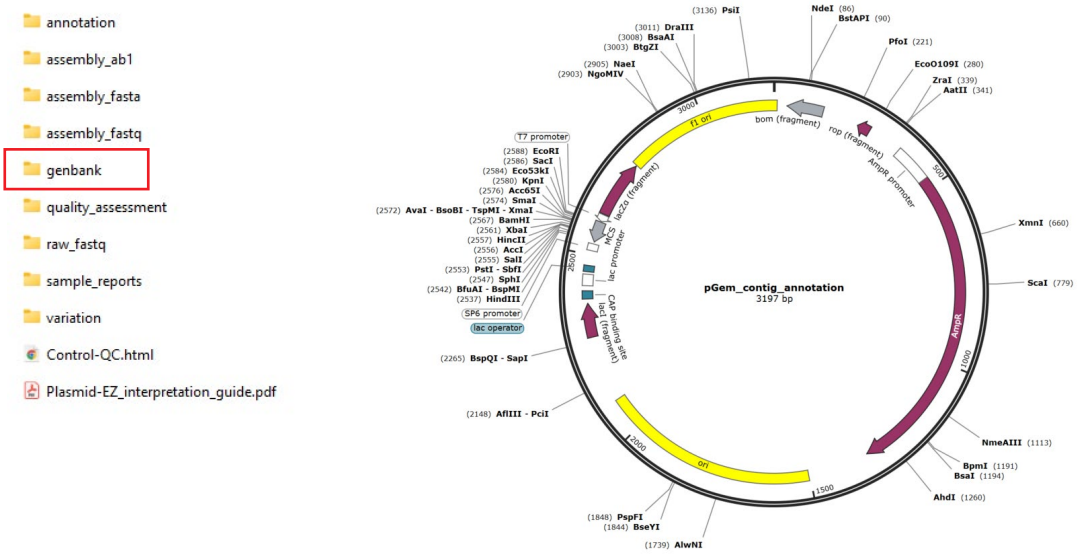
OPENING THE GENBANK FILE IN SNAPGENE VIEWER

A copy of the annotation map and sequence are provided in the sample folder in GenBank format. This file can be opened in any plasmid viewer program like SnapGene Viewer.



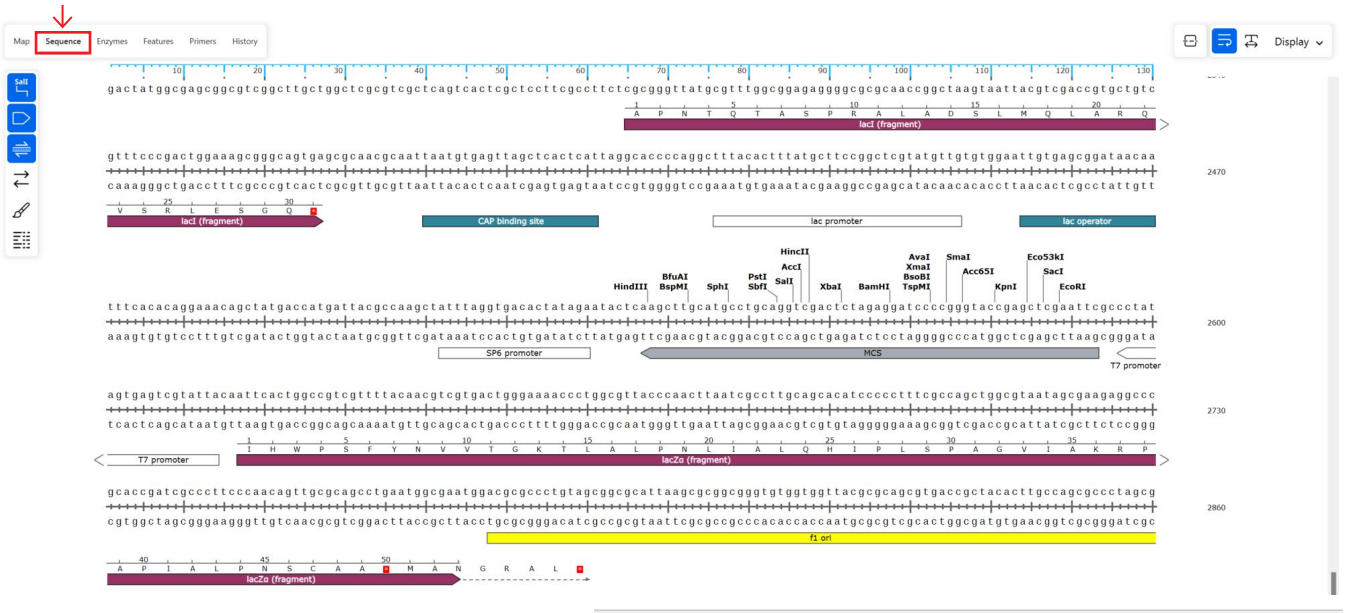
Viewing Your Plasmid (Continued)

Opening the GenBank file (.gbk) in SnapGene Viewer will present you with the following screen upon opening. This shows you a map from the report as well as all restriction enzyme sites.

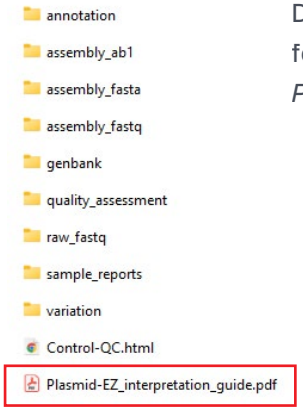


VIEWING NUCLEOTIDE SEQUENCE, AMINO ACID SEQUENCE AND ANNOTATION

Clicking the sequence tab at the top of the screen (highlighted in a red box below) will open the nucleotide sequence, along with the annotation and amino acid sequence for all coding regions.



Viewing Your Plasmid (Continued)

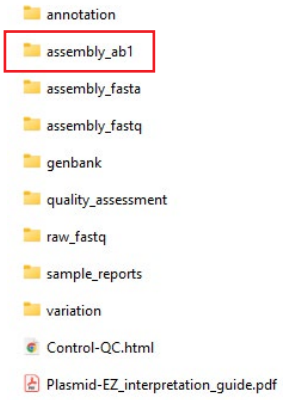


Details for the meaning of the headers for this and all tables can be found in the *Plasmid-EZ_interpretation_guide.pdf* file.

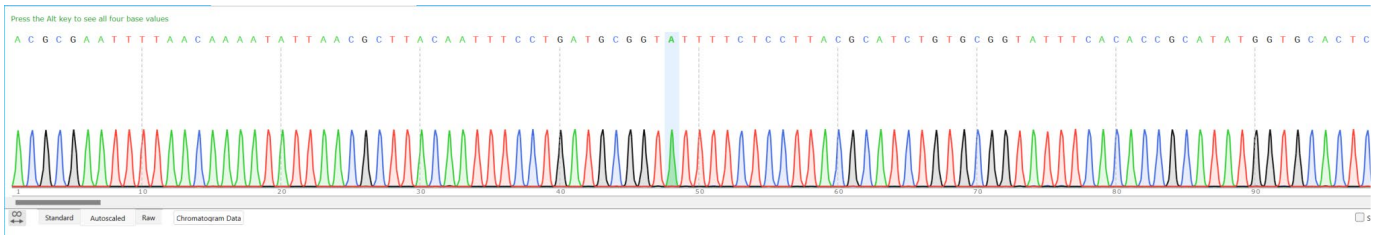
VIEWING ASSEMBLY_AB1 FILES

The ab1 files can be utilized to assess quality and identify positions with mixed base-calling.

The ab1 files are fragmented into 2kB length files; therefore, each sample will be composed of multiple ab1 files.



Our ab1 files allow you to validate sample quality with a quick scroll through the file. The ab1 file is a great tool to identify any mixed base calling and validate the sample quality.

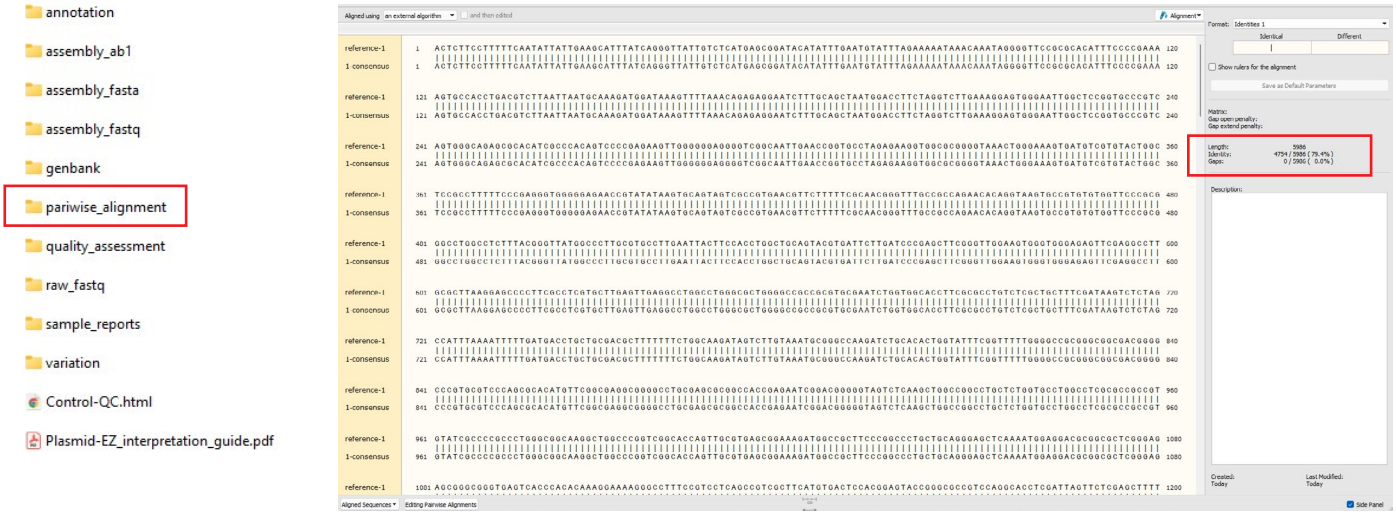


As the ab1 example shows below, mixed base calling can be easily identified and verified from this single file.

Viewing Your Plasmid (Continued)

REVIEWING PAIRWISE_ALIGNMENT OF CONSENSUS FILES

If reference file is provided on the order form, we will run reference-based analysis and provide a pairwise alignment to compare the assembled contig to the reference file provided. The example below shows alignment between the reference and consensus file with the highlighted region showing percentage match between the files.



- Why are NNNs in the consensus file? The Ns indicates a mismatched or missing region in the consensus file compared to the reference file. Evaluation of the ab1 or variation file for the positions with NNNs will determine whether the region is truly missing from the consensus file or of too low quality to match the reference file.
- The reference-based analysis is started right after sequencing, with the missing region during the analysis denoted by N and tallied into the total length count. Therefore, the AssemblyReport will show a plasmid size same as the reference file. The sample consensus Genbank file will show the total size to also be same as the reference file, but with the region showing the Ns (or the missing region) to be grayed out.

3 Assessing the Quality of Your Data

VIEWING THE READ-LENGTH DISTRIBUTION AND QUALITY SCORES

Going back to the sample report (AssemblyReport.html), you will see the sequencing quality metrics, including the read-length distribution (red graph), the Q-score distribution (blue graph), and the percentage map reads (section 3).

2. Read Lengths

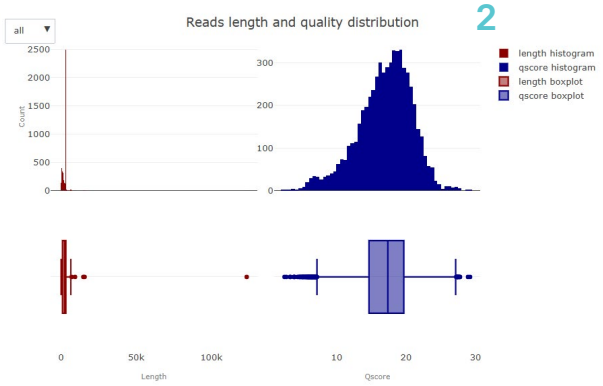
Summary statistics of read lengths.

read_type	min_length	mean_length	max_length	q1_length	median_length	q3_length	n50_length
all	104	2182.27	123546	972	2497	3292	3285
pass	105	2195.73	15695	1014.25	2600.5	3293	3286

WHAT YOU WANT TO SEE:

1. A clear plasmid peak matching your assembly length

1



WHAT YOU WANT TO SEE:

2. Most of the reads having a Q > 10

3. Mapping of sequencing reads to the assembly

If a reference sequence was provided, then creation of a reference-based consensus sequence was attempted. If this failed, or no reference sequence was provided, then *de novo* assembly was attempted.

The following table reports statistics on mapping of raw sequencing reads to the reference-based consensus **or** *de novo* assembled contig.

1. Sample: Sample name.
2. Reference_Length: Length of the reference sequence, if a reference sequence was provided.
3. Consensus_Length: Length of the reference-based consensus sequence.
   Reference-based consensus is only reported if a reference sequence was provided and a reference-based consensus sequence could be generated.
   If a reference sequence was provided but a reference-based consensus sequence could not be generated, then "-" is shown and the table cell is highlighted.
4. Contig_Length: Length of the *de novo* assembled contig sequence.
   De novo assembled contig sequence is only reported if no reference sequence was provided, or reference-based consensus sequence could not be generated.
   If reference-based consensus sequence could not be generated, but *de novo* assembly was successful, then Contig_Length is reported and the table cell is highlighted.
5. Total_Reads: Total number of reads passing Nanopore sequencing quality assessment.
6. Mapped_reads: Number, and percentage, of total reads successfully mapped to the reference, if provided, or *de novo* assembled contig, sequence.
7. Unmapped_reads: Number, and percentage, of total reads could not be mapped to the reference, if provided, or *de novo* assembled contig, sequence.
8. Supplementary_mappings: Number, and percentage, of total reads that span the 3-prime - 5-prime boundary of the linearized contig sequence.
9. Secondary_mappings: Number, and percentage, of total reads that map to more than one location on the linearized contig sequence.

Calculation of percentages are based on the "Total_Reads" value reported.

3

Sample	Reference_Length	Consensus_Length	Contig_Length	Total_Reads	Mapped_reads	Unmapped_reads	Supplementary_mappings
pGem	n/a	n/a	3197	5518	5198 (94.2%)	320 (5.79%)	3473 (62.93%)

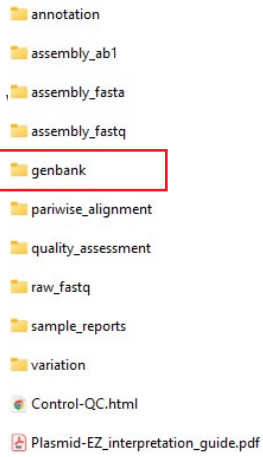
WHAT YOU WANT TO SEE:

3. Most of your reads mapping to the assembly

3 Assessing the Quality of Your Data (Continued)

VIEWING PER BASE CONFIDENCE SCORES

We also provide a FASTQ file with a confidence Q score per base that can be viewed in SnapGene similar program.

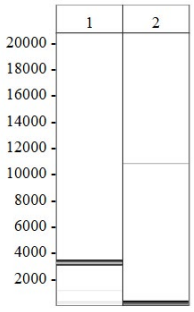


The higher the bar, the higher the confidence for the base call at this position. If a bar is lower, it could indicate either low-quality sequence data or the presence of a polymorphism at the site.

VIEWING THE VIRTUAL GEL

The virtual gel is also available to quickly visualize the sample quality.

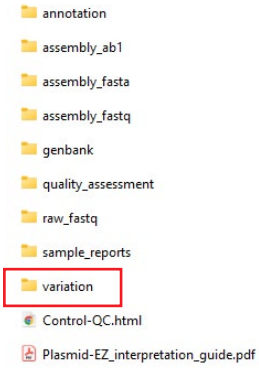
The virtual gel provides a visual representation of the read lengths and the intensity of each found in the sample. Please note the gel image may not represent all read lengths in the sample and should only be used as a visual representation and not for analysis.



4 Variant Calling

VIEWING VARIANTS BY BASE

We provide a variant Excel file (*Contig-readCounts-variation.xls*) that has the number of reads for each base.



Any bases with a second nucleotide represented in >10% of reads are highlighted in yellow. This file also provides you with the number of reads that have insertions or deletions for that base.

AZENTA LIFE SCIENCES - Plasmid-EZ

Color Legend: >= 10% deviation

These data relate to *raw* sequencing reads

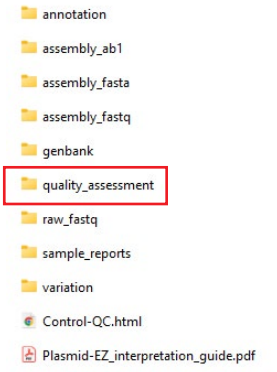
Position	Reference	Coverage	A	T	G	C	N	Insertions	Top Insertion	Deletions
1	A	3435	3435	0	0	0	0	0	-	0
2	C	3440	0	0	0	3440	0	0	-	0
3	G	3454	0	0	3454	0	0	0	-	0
4	C	3463	0	11	0	3452	0	1	G (1)	0
5	G	3482	1	1	3477	3	0	4	C (2)	0
6	A	3490	3488	1	1	0	0	0	-	0
7	A	3492	3479	11	2	0	0	11	T (10)	0
8	T	3499	13	3486	0	0	0	1	A (1)	0
9	T	3502	0	3501	0	1	0	0	-	0
10	T	3506	0	3504	0	2	0	0	-	0
11	T	3506	1	3491	0	13	0	27	A (15)	1
12	A	3514	3475	2	10	3	0	1	AAAGG (1)	24
13	A	3506	3500	0	3	0	0	9	G (4)	3
14	C	3507	7	3	17	3471	0	34	A (21)	9
15	A	3518	3481	1	21	0	0	0	-	15
16	A	3519	3510	0	9	0	0	1	TAC (1)	0
17	A	3524	3519	2	2	1	0	0	-	0
18	A	3525	3516	2	7	0	0	0	-	0
19	T	3527	3	3514	2	7	0	3	G (2)	1
20	A	3528	3504	2	8	8	0	28	T (21)	6

Note: This file is corrected for read quality; an uncorrected raw number of reads with each base can be found in the *readCounts-variantion.csv* file.

5 Common Factors Affecting Plasmid-EZ Assembly

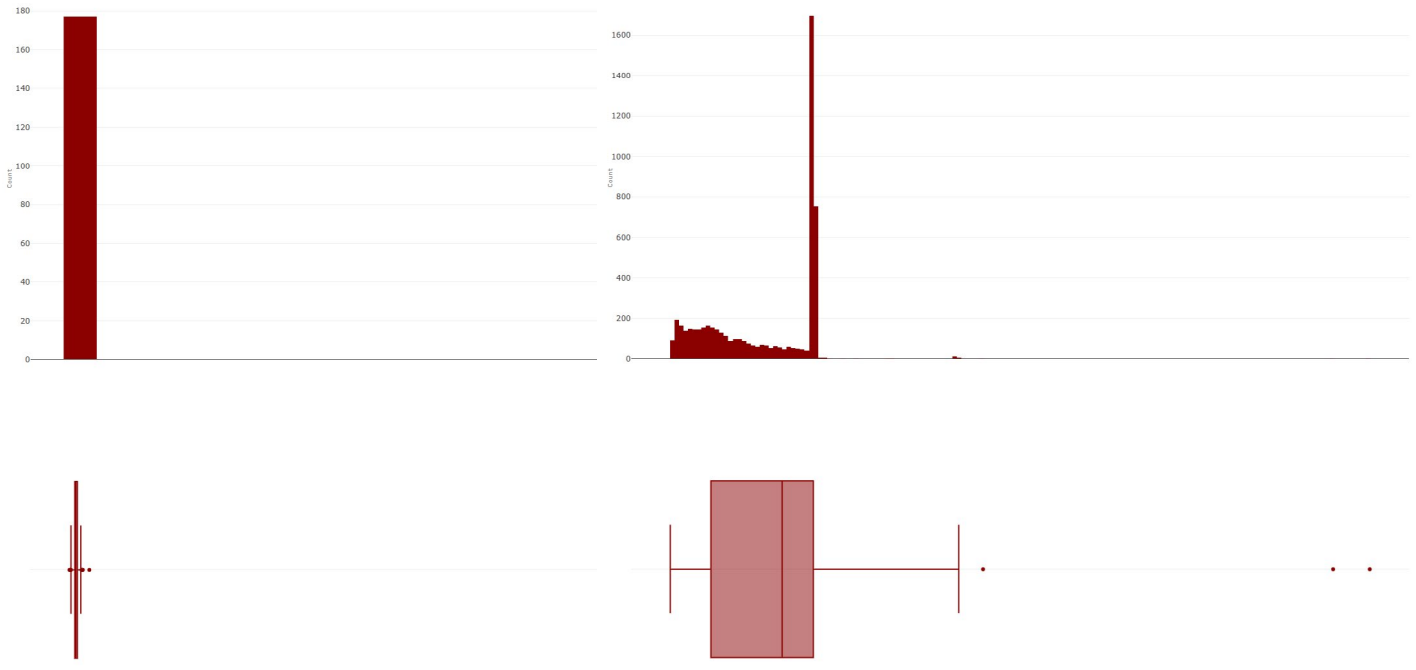
CONCENTRATION AND SAMPLE QUALITY

In the unfortunate event your sample failed to produce an assembly, the sample folder will only contain the raw FASTQ reads and a summary report that includes the read length and quality of the data.



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 without full-length plasmid reads tend to fail assembly.



5 Common Factors Affecting Plasmid-EZ Assembly (Continued)

HOMOPOLYMERIC REPETITIVE REGIONS

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

Position	Reference	Coverage	A	T	G	C	N	Insertions	Top Insertion	Deletions
9095	G	773	0	1	771	0	0	292	T (162)	1
9096	T	774	0	496	0	0	0	0	_	278
9097	T	637	0	636	0	1	0	0	_	0
9098	T	709	0	709	0	0	0	0	_	0
9099	T	740	0	740	0	0	0	0	_	0
9100	T	750	0	750	0	0	0	0	_	0
9101	T	756	0	756	0	0	0	0	_	0
9102	T	769	0	768	0	0	0	0	_	1
9103	T	771	0	771	0	0	0	0	_	0
9104	T	772	0	771	0	1	0	0	_	0

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, as shown in the example below. In such cases, Sanger sequencing follow up may be needed for further confirmation for the base position. The image below shows a GATC methylation sites show mixed base calling with A/G. In this example, nucleotide G was called but for further validation a Sanger sequencing order may be needed.

Position	Reference	Coverage	A	T	G	C	N	Insertions	Top Insertion	Deletions
4240	G	2020	803	6	1149	14	0	63	A (43)	48
4241	A	2070	2056	3	6	2	0	161	T (130)	3
4242	T	2069	73	1977	0	2	0	22	TC (9)	17
4243	C	2057	60	371	5	1615	0	21	T (21)	6