GUIDE

Plasmid-EZ Quick Start Guide

AZENTA
LIFE SCIENCES

azenta.com

Table of Contents

1. Getting Started 03

Viewing project report

Opening sample reports

2. Viewing Your Plasmid 04

Annotation map of longest contig

Opening the GenBank file in SnapGene Viewer

Viewing the nucleotide sequence, amino acid sequence, and annotation

Viewing the annotation parts in Excel

Opening the FASTA file in a text editor

3. Assessing the Quality of Your Data 07

Viewing the read-length distribution and quality scores

Viewing per base confidence scores

4. Variant Calling 09

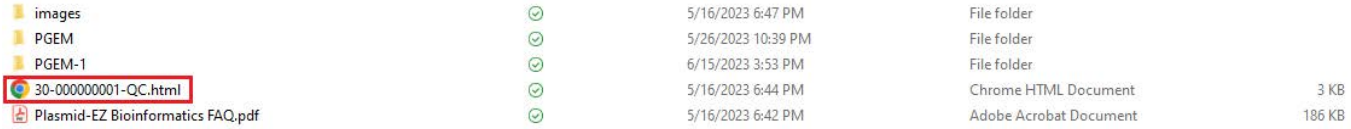
Looking at variants by base

Looking at multiple contigs and generating annotations

5. Why Did My Sample Fail to Produce Assembly? 11

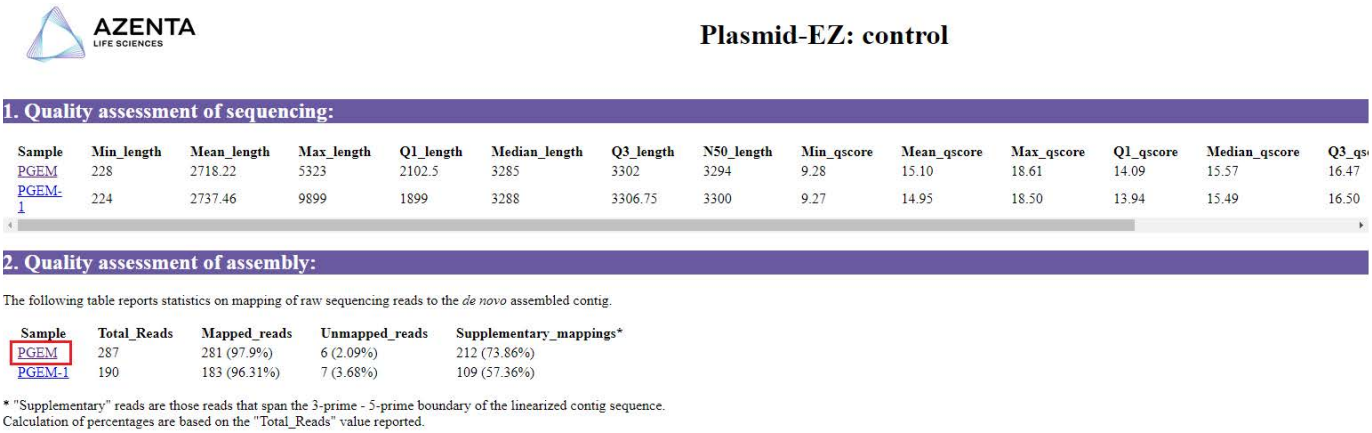
1 Getting Started

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.



images	5/16/2023 6:47 PM	File folder	
PGEM	5/26/2023 10:39 PM	File folder	
PGEM-1	6/15/2023 3:53 PM	File folder	
30-000000001-QC.html	5/16/2023 6:44 PM	Chrome HTML Document	3 KB
Plasmid-EZ Bioinformatics FAQ.pdf	5/16/2023 6:42 PM	Adobe Acrobat Document	186 KB

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:

Sample	Min_length	Mean_length	Max_length	Q1_length	Median_length	Q3_length	N50_length	Min_qscore	Mean_qscore	Max_qscore	Q1_qscore	Median_qscore	Q3_qs
PGEM	228	2718.22	5323	2102.5	3285	3302	3294	9.28	15.10	18.61	14.09	15.57	16.47
PGEM-1	224	2737.46	9899	1899	3288	3306.75	3300	9.27	14.95	18.50	13.94	15.49	16.50

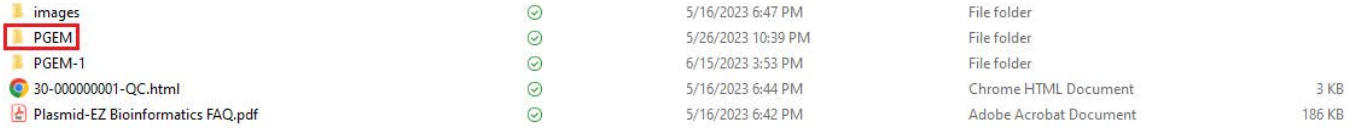
2. Quality assessment of assembly:

The following table reports statistics on mapping of raw sequencing reads to the *de novo* assembled contig.

Sample	Total_Reads	Mapped_reads	Unmapped_reads	Supplementary_mappings*
PGEM	287	281 (97.9%)	6 (2.09%)	212 (73.86%)
PGEM-1	190	183 (96.31%)	7 (3.68%)	109 (57.36%)

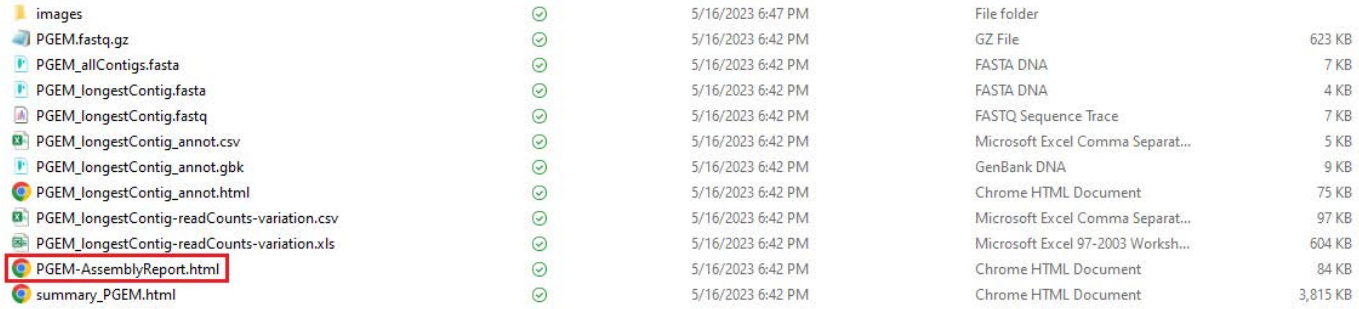
* "Supplementary" reads are those reads that span the 3-prime - 5-prime boundary of the linearized contig sequence.
Calculation of percentages are based on the "Total_Reads" value reported.

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



images	5/16/2023 6:47 PM	File folder	
PGEM	5/26/2023 10:39 PM	File folder	
PGEM-1	6/15/2023 3:53 PM	File folder	
30-000000001-QC.html	5/16/2023 6:44 PM	Chrome HTML Document	3 KB
Plasmid-EZ Bioinformatics FAQ.pdf	5/16/2023 6:42 PM	Adobe Acrobat Document	186 KB

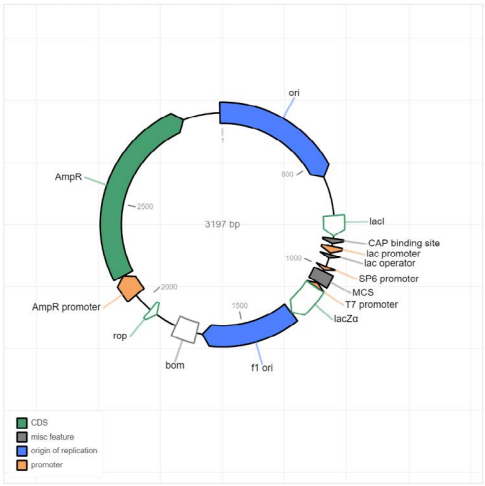
Then, click on the *AssemblyReport.html* file.



images	5/16/2023 6:47 PM	File folder	
PGEM.fastq.gz	5/16/2023 6:42 PM	GZ File	623 KB
PGEM_allContigs.fasta	5/16/2023 6:42 PM	FASTA DNA	7 KB
PGEM_longestContig.fasta	5/16/2023 6:42 PM	FASTA DNA	4 KB
PGEM_longestContig.fastq	5/16/2023 6:42 PM	FASTQ Sequence Trace	7 KB
PGEM_longestContig_annot.csv	5/16/2023 6:42 PM	Microsoft Excel Comma Separat...	5 KB
PGEM_longestContig_annot.gbk	5/16/2023 6:42 PM	GenBank DNA	9 KB
PGEM_longestContig_annot.html	5/16/2023 6:42 PM	Chrome HTML Document	75 KB
PGEM_longestContig-readCounts-variation.csv	5/16/2023 6:42 PM	Microsoft Excel Comma Separat...	97 KB
PGEM_longestContig-readCounts-variation.xls	5/16/2023 6:42 PM	Microsoft Excel 97-2003 Worksh...	604 KB
PGEM-AssemblyReport.html	5/16/2023 6:42 PM	Chrome HTML Document	84 KB
summary_PGEM.html	5/16/2023 6:42 PM	Chrome HTML Document	3,815 KB

2 Viewing Your Plasmid

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 PGEM_longestContig_annot.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

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.

images	5/16/2023 6:47 PM	File folder		
PGEM.fastq.gz	5/16/2023 6:42 PM	GZ File	623 KB	
PGEM_allContigs.fasta	5/16/2023 6:42 PM	FASTA DNA	7 KB	
PGEM_longestContig.fasta	5/16/2023 6:42 PM	FASTA DNA	4 KB	
PGEM_longestContig.fastq	5/16/2023 6:42 PM	FASTQ Sequence Trace	7 KB	
PGEM_longestContig_annot.csv	5/16/2023 6:42 PM	Microsoft Excel Comma Separat...	5 KB	
PGEM_longestContig_annot.gbk	5/16/2023 6:42 PM	GenBank DNA	9 KB	
PGEM_longestContig_annot.html	5/16/2023 6:42 PM	Chrome HTML Document	75 KB	
PGEM_longestContig-readCounts-variation.csv	5/16/2023 6:42 PM	Microsoft Excel Comma Separat...	97 KB	
PGEM_longestContig-readCounts-variation.xls	5/16/2023 6:42 PM	Microsoft Excel 97-2003 Worksh...	604 KB	
PGEM-AssemblyReport.html	5/16/2023 6:42 PM	Chrome HTML Document	84 KB	
summary_PGEM.html	5/16/2023 6:42 PM	Chrome HTML Document	3,815 KB	

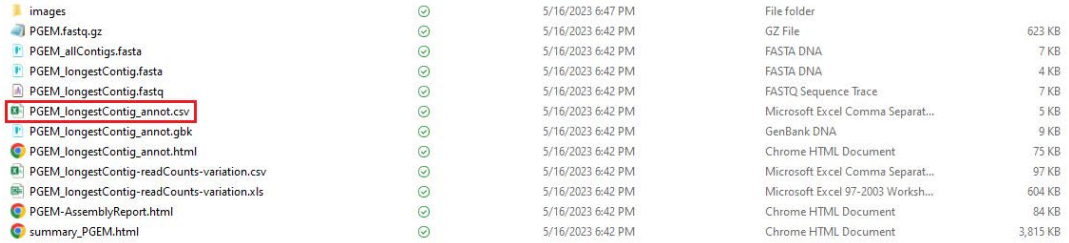
2 Viewing Your Plasmid (Continued)

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.

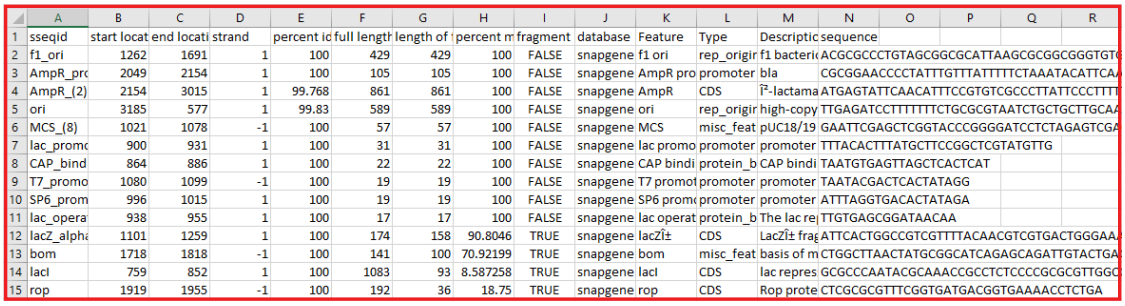
Clicking the sequence tab at the bottom of the screen (highlighted in a red box below) will open the nucleotide sequence, along with the annotation and amino acid sequence for all coding regions.

2 Viewing Your Plasmid (Continued)

A copy of the annotation parts, as well as the sequence for each part, can be found in the *longestContig_annot.csv* file that can be opened in Excel.

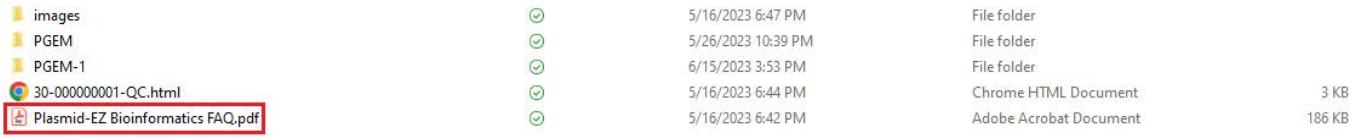


images	5/16/2023 6:47 PM	File folder	
PGEM.fastq.gz	5/16/2023 6:42 PM	GZ File	623 KB
PGEM_allContigs.fasta	5/16/2023 6:42 PM	FASTA DNA	7 KB
PGEM_longestContig.fasta	5/16/2023 6:42 PM	FASTA DNA	4 KB
PGEM_longestContig.fastq	5/16/2023 6:42 PM	FASTQ Sequence Trace	7 KB
PGEM_longestContig_annot.csv	5/16/2023 6:42 PM	Microsoft Excel Comma Separat...	5 KB
PGEM_longestContig_annot.gbk	5/16/2023 6:42 PM	GenBank DNA	9 KB
PGEM_longestContig_annot.html	5/16/2023 6:42 PM	Chrome HTML Document	75 KB
PGEM_longestContig-readCounts-variation.csv	5/16/2023 6:42 PM	Microsoft Excel Comma Separat...	97 KB
PGEM_longestContig-readCounts-variation.xls	5/16/2023 6:42 PM	Microsoft Excel 97-2003 Worksh...	604 KB
PGEM-AssemblyReport.html	5/16/2023 6:42 PM	Chrome HTML Document	84 KB
summary_PGEM.html	5/16/2023 6:42 PM	Chrome HTML Document	3,815 KB



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	sseqid	start locat	end locati	strand	percent id	full length	length of f	percent m	fragment	database	Feature	Type	Descriptic	sequence				
2	f1_ori	1262	1691	1	100	429	429	100	FALSE	snapgene	f1 ori	rep_origir	f1 bacteri	ACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGT				
3	AmpR_pr	2049	2154	1	100	105	105	100	FALSE	snapgene	AmpR pro	promoter	bla	CGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCA				
4	AmpR_(2)	2154	3015	1	99.768	861	861	100	FALSE	snapgene	AmpR	CDS	β-lactama	ATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTT				
5	ori	3185	577	1	99.83	589	589	100	FALSE	snapgene	ori	rep_origir	high-copy	TTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCA				
6	MCS_(8)	1021	1078	-1	100	57	57	100	FALSE	snapgene	MCS	misc_feat	pUC18/19	GAATTCGAGCTCGGTACCCGGGGATCCTCTAGAGTCGA				
7	lac_promc	900	931	1	100	31	31	100	FALSE	snapgene	lac promo	promoter	promoter	TTTACACTTTATGCTTCCGGCTCGTATGTTG				
8	CAP_bind	864	886	1	100	22	22	100	FALSE	snapgene	CAP bindi	protein_b	CAP bindi	TAATGTGAGTTAGCTCACTCAT				
9	T7_promo	1080	1099	-1	100	19	19	100	FALSE	snapgene	T7 promot	promoter	promoter	TAATACGACTCACTATAGG				
10	SP6_prom	996	1015	1	100	19	19	100	FALSE	snapgene	SP6 prom	promoter	promoter	ATTTAGGTGACACTATAGA				
11	lac_opera	938	955	1	100	17	17	100	FALSE	snapgene	lac operat	protein_b	The lac re	TTGTGAGCGGATAACAA				
12	lacZ_alpha	1101	1259	1	100	174	158	90.8046	TRUE	snapgene	lacZ?±	CDS	LacZ?± fra	ATTCACTGGCCGTCGTTTTACAACGTCGTGACTGGGAA				
13	bom	1718	1818	-1	100	141	100	70.92199	TRUE	snapgene	bom	misc_feat	basis of m	CTGGCTTAACTATGCGGCATCAGAGCAGATTGTACTGA				
14	lacI	759	852	1	100	1083	93	8.587258	TRUE	snapgene	lacI	CDS	lac repres	GCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGC				
15	rop	1919	1955	-1	100	192	36	18.75	TRUE	snapgene	rop	CDS	Rop prote	CTCGCGCGTTTCGGTGATGACGGTGAAAACCTCTGA				

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



images	5/16/2023 6:47 PM	File folder	
PGEM	5/26/2023 10:39 PM	File folder	
PGEM-1	6/15/2023 3:53 PM	File folder	
30-000000001-QC.html	5/16/2023 6:44 PM	Chrome HTML Document	3 KB
Plasmid-EZ Bioinformatics FAQ.pdf	5/16/2023 6:42 PM	Adobe Acrobat Document	186 KB

We also provide a simple FASTA-formatted file with the sequence of the longest contig in the *longestContig.fasta* file. This file can be opened using software such as SnapGene Viewer, Geneious, etc., as well as any text viewer including Microsoft Word.

```
>PGEM.tig00000003 len=3197 reads=125 class=contig suggestRepeat=yes
suggestBubble=no suggestCircular=yes trim=0-3197
TTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAA
GAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTTCTTCTAGTGT
AGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACC
AGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCG
CAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGAT
ACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGG
CAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGG
TTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCA
GCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCC
TGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGC
AGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCCGATTC
ATTAATGCAGCTGGCACGACAGGTTTCCCGACTGGAAAGCGGGCAGTGAGCGCAACGCAATTAATGTGAGTTA
```

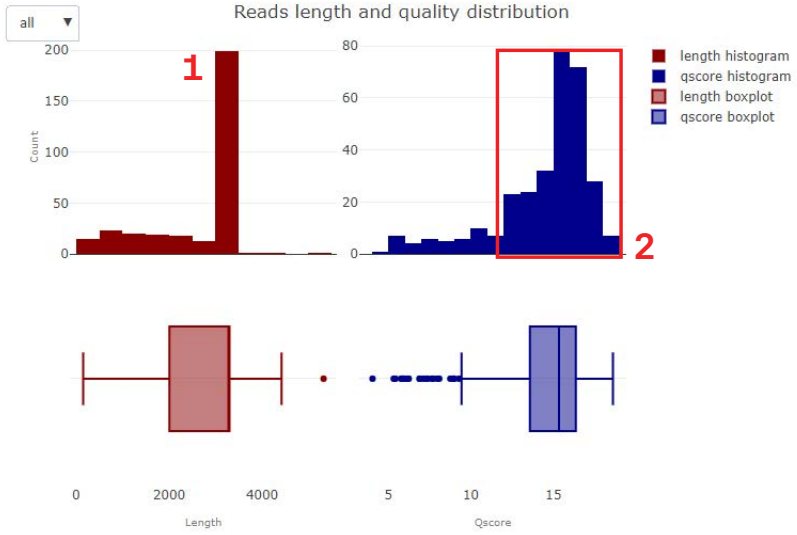
3 Assessing the Quality of Your Data

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	154	2662.15	5323	2011.25	3281	3301.75	3293
pass	228	2718.22	5323	2102.5	3285	3302	3294



3. Mapping of sequencing reads to the assembly

The following table reports statistics on mapping of raw sequencing reads to the *de novo* assembled contig.

Total_Reads	mapped_reads	unmapped_reads	supplementary_mappings*
287	281 (97.9%)	6 (2.09%)	212 (73.86%)

3

* "Supplementary" reads are those reads that span the 3-prime - 5-prime boundry of the linearized contig sequence.
Calculation of percentages are based on the "Total_Reads" value reported.

WHAT YOU WANT TO SEE:

1. A clear plasmid peak matching your assembly length
2. Most of the reads having a Q > 10
3. Most of your reads mapping to the assembly

3 Assessing the Quality of Your Data (Continued)

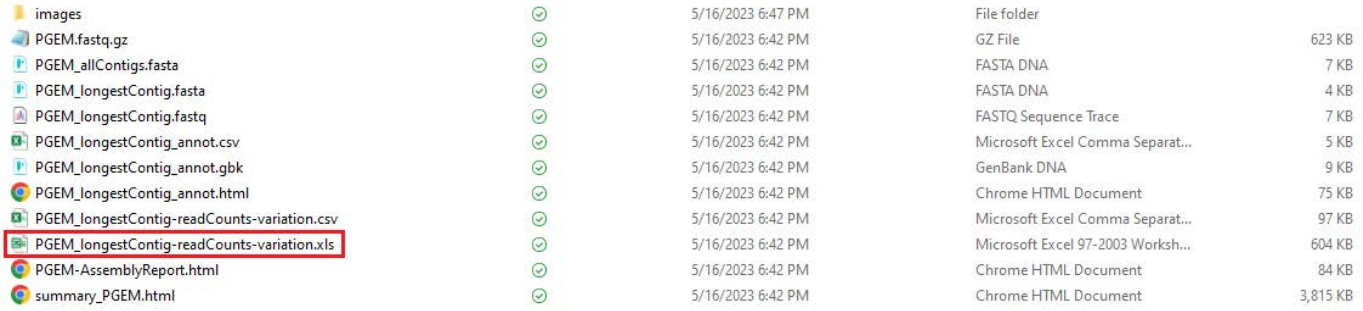
We also provide a FASTQ file with a confidence Q score per base that can be viewed in SnapGene Viewer or a similar program.

images	5/16/2023 6:47 PM	File folder	
PGEM.fastq.gz	5/16/2023 6:42 PM	GZ File	623 KB
PGEM_allContigs.fasta	5/16/2023 6:42 PM	FASTA DNA	7 KB
PGEM_longestContig.fasta	5/16/2023 6:42 PM	FASTA DNA	4 KB
PGEM_longestContig.fastq	5/16/2023 6:42 PM	FASTQ Sequence Trace	7 KB
PGEM_longestContig_annot.csv	5/16/2023 6:42 PM	Microsoft Excel Comma Separat...	5 KB
PGEM_longestContig_annot.gbk	5/16/2023 6:42 PM	GenBank DNA	9 KB
PGEM_longestContig_annot.html	5/16/2023 6:42 PM	Chrome HTML Document	75 KB
PGEM_longestContig-readCounts-variation.csv	5/16/2023 6:42 PM	Microsoft Excel Comma Separat...	97 KB
PGEM_longestContig-readCounts-variation.xls	5/16/2023 6:42 PM	Microsoft Excel 97-2003 Worksh...	604 KB
PGEM-AssemblyReport.html	5/16/2023 6:42 PM	Chrome HTML Document	84 KB
summary_PGEM.html	5/16/2023 6:42 PM	Chrome HTML Document	3,815 KB

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.

4 Variant Calling

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



images	5/16/2023 6:47 PM	File folder	
PGEM.fastq.gz	5/16/2023 6:42 PM	GZ File	623 KB
PGEM_allContigs.fasta	5/16/2023 6:42 PM	FASTA DNA	7 KB
PGEM_longestContig.fasta	5/16/2023 6:42 PM	FASTA DNA	4 KB
PGEM_longestContig.fastq	5/16/2023 6:42 PM	FASTQ Sequence Trace	7 KB
PGEM_longestContig_annot.csv	5/16/2023 6:42 PM	Microsoft Excel Comma Separat...	5 KB
PGEM_longestContig_annot.gbk	5/16/2023 6:42 PM	GenBank DNA	9 KB
PGEM_longestContig_annot.html	5/16/2023 6:42 PM	Chrome HTML Document	75 KB
PGEM_longestContig-readCounts-variation.csv	5/16/2023 6:42 PM	Microsoft Excel Comma Separat...	97 KB
PGEM_longestContig-readCounts-variation.xls	5/16/2023 6:42 PM	Microsoft Excel 97-2003 Worksh...	604 KB
PGEM-AssemblyReport.html	5/16/2023 6:42 PM	Chrome HTML Document	84 KB
summary_PGEM.html	5/16/2023 6:42 PM	Chrome HTML Document	3,815 KB

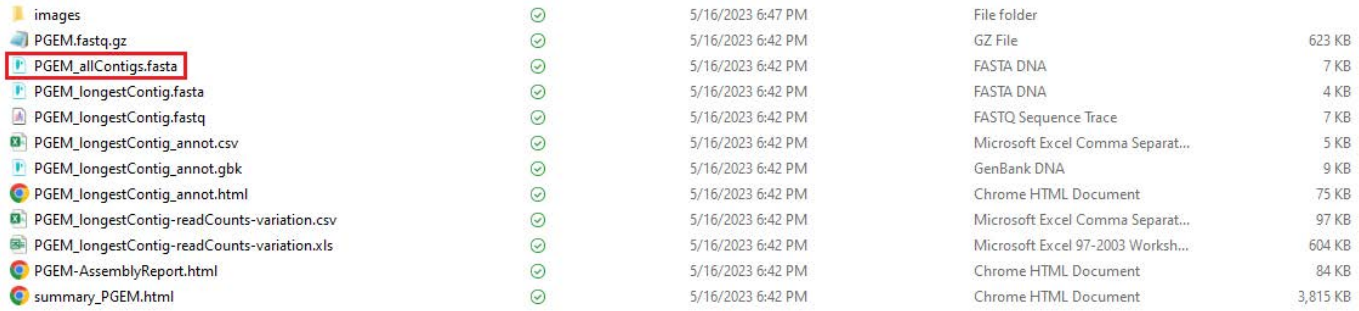
Any bases with a second nucleotide represented in >10% of reads are highlighted in yellow. This file also provides you with the number of reads that have insertions or deletions for that base.

	A	B	C	D	E	F	G	H	I	J	K
1	AZENTA LIFE SCIENCES - Plasmid-EZ										
2											
3	Color Legend:										
4		>= 10% deviation									
5			These data relate to *raw* sequencing reads								
6	**Position**	**Reference**	**Coverage**	**A**	**T**	**G**	**C**	**N**	**Insertions**	**Top Insertion**	**Deletions**
7	**1**	**T**	199	0	199	0	0	0	0	-	0
8	**2**	**T**	202	0	202	0	0	0	0	-	0
9	**3**	**T**	204	0	204	0	0	0	0	-	0
10	**4**	**T**	205	0	204	0	0	0	0	-	0
11	**5**	**C**	205	0	0	0	192	0	1	T (1)	0
12	**6**	**T**	206	0	206	0	0	0	0	-	0
13	**7**	**G**	207	50	0	153	0	0	0	-	2
14	**8**	**C**	205	0	0	0	201	0	1	CA (1)	1

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.

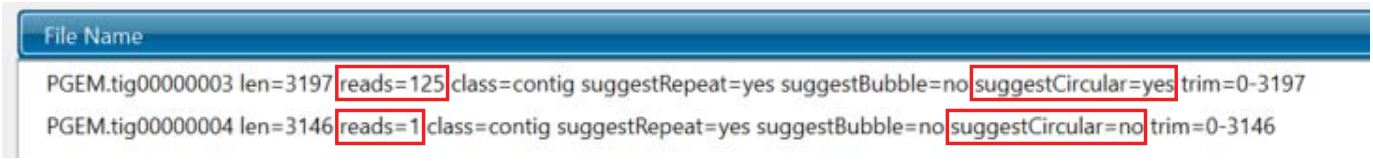
4 Variant Calling (Continued)

There is also a potential for multiple assemblies to be created resulting in multiple contigs. If this happens, an *allContigs.fasta* file will be created with all the contigs assembled.



images	5/16/2023 6:47 PM	File folder	
PGEM.fastq.gz	5/16/2023 6:42 PM	GZ File	623 KB
PGEM_allContigs.fasta	5/16/2023 6:42 PM	FASTA DNA	7 KB
PGEM_longestContig.fasta	5/16/2023 6:42 PM	FASTA DNA	4 KB
PGEM_longestContig.fastq	5/16/2023 6:42 PM	FASTQ Sequence Trace	7 KB
PGEM_longestContig_annot.csv	5/16/2023 6:42 PM	Microsoft Excel Comma Separat...	5 KB
PGEM_longestContig_annot.gbk	5/16/2023 6:42 PM	GenBank DNA	9 KB
PGEM_longestContig_annot.html	5/16/2023 6:42 PM	Chrome HTML Document	75 KB
PGEM_longestContig-readCounts-variation.csv	5/16/2023 6:42 PM	Microsoft Excel Comma Separat...	97 KB
PGEM_longestContig-readCounts-variation.xls	5/16/2023 6:42 PM	Microsoft Excel 97-2003 Worksh...	604 KB
PGEM-AssemblyReport.html	5/16/2023 6:42 PM	Chrome HTML Document	84 KB
summary_PGEM.html	5/16/2023 6:42 PM	Chrome HTML Document	3,815 KB

If you open this file in SnapGene Viewer, you will see a list of all contigs, as well as some details from the assembler.



In the case above, two contigs were assembled with one generated with 125 reads and a suggested complete circular contig, while the other is created with a single read and is not a complete circular sequence. Thus, the second contig is likely an artifact of assembly and not a true variant.

If in contrast, you see multiple circular contigs with a high number of reads used to generate the contig (e.g., > 30), then this might represent multiple plasmid variants in your sample.

If you want to see the annotation for any of the other contigs, you can simply copy and paste the sequence into Plannotate. This will generate an annotation with a GenBank file and a CSV file with the annotation parts.

If you see all contigs represented by a few reads, then this might indicate a lower quality assembly for your sample.

5. Why Did My Sample Fail to Produce an Assembly?

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.