

GUIDE

# Plasmid-EZ クイックスタートガイド












# 目次

1.	納品物の概要	05
2.	プロジェクトレポートの確認 プロジェクトレポートの閲覧	06
3.	サンプルレポートの評価 サンプルレポート 解析結果 アノテーションマップ アノテーションテーブル データ品質管理 (QC) アセンブリ失敗レポート	07
4.	アセンブルコンティグの確認 GeneBank Assembly FASTA Assembly FASTQ Assembly Ab1	11
5.	サンプルカバレッジの評価 変異 (Variation)	13
6.	リファレンスベース解析 カバレッジ解析 バリエーション	15
7.	Plasmid-EZアセンブリに影響する主な要因 濃度およびサンプル品質 ホモポリマー反復領域 メチル化部位	17


## Plasmid-EZ クイックスタートガイド


## 1.納品物の概要


 control_project_report.html	HTML 形式のプロジェクト概要のレポート
 raw_fastq	ZIP 圧縮された FASTQ 形式の RAW データ
 assembly_fasta	FASTA 形式のアセンブル済み配列
 assembly_fastq	FASTQ 形式のアセンブル済み配列
 genbank	GenBank 形式のアノテーションデータ
 annotation	CSV 形式の注釈付きフィーチャー一覧
 assembly_ab1	ABI 形式のアセンブル済み配列
 variation	<ul style="list-style-type: none"> <li>• RAW リードを最終コンセンサス配列にアラインメントした際の塩基カウントデータ</li> <li>• VCF 形式のバリエーションコール結果</li> <li>• VCF ファイル用インデックスファイル</li> <li>• タブ区切りテキスト形式のバリエーションコール結果</li> </ul>
 sample_reports	HTML 形式のサンプルレポート

※サンプルがアセンブリを生成できなかった場合、

サンプルフォルダーには Project Report、Sample Report、および RAW FASTQ リードのみが含まれます。

 control\_project\_report.html

 raw\_fastq

 sample\_reports

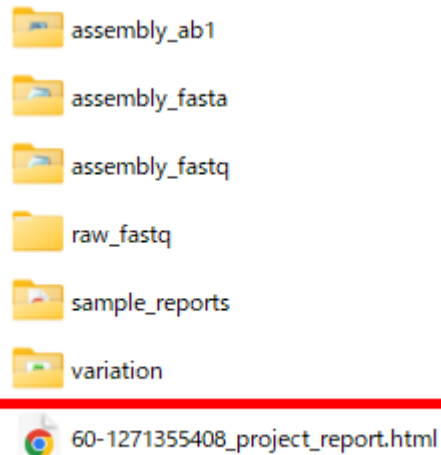
## Plasmid-EZ クイックスタートガイド

## 2. プロジェクトレポートの確認

## プロジェクトレポートの閲覧

まず60-xxxxxxx\_project\_report.html ファイルをクリックします。  
このファイルには、リード数、アセンブリのステータス、取得されたコンティグ数、マッピングされたリードの割合などをまとめた結果概要が表示されます。

バーチャルゲル画像は、各サンプルごとのリード長分布が視覚的に示されています。



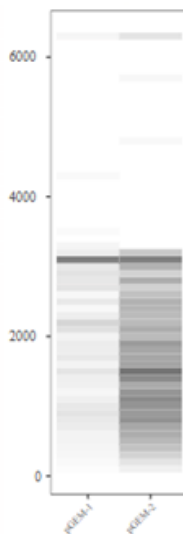
## 1 RESULTS SUMMARY

The raw data metrics and assembly results for all samples in this project are summarized in [Table 11](#).

Table 11: Summary of the assembly results

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

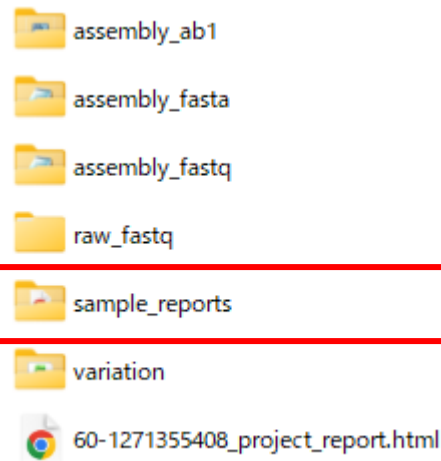
Figure below shows the read length distribution as a simulated gel image. Mouse over the bands to see the sample name and read length.



### 3. サンプルレポートの評価

#### サンプルレポート

各サンプルレポートは、sample\_reports フォルダからアクセスできます。各ファイルをクリックすると、それぞれのレポートを確認できます。



#### 解析結果 (ANALYSIS RESULTS)

サンプルレポートには、アセンブルされたコンティグ数、それぞれの長さ、トポロジー（構造）、マッピングされた塩基数、総マッピング塩基率、およびアセンブリの平均カバレッジが記載されています。

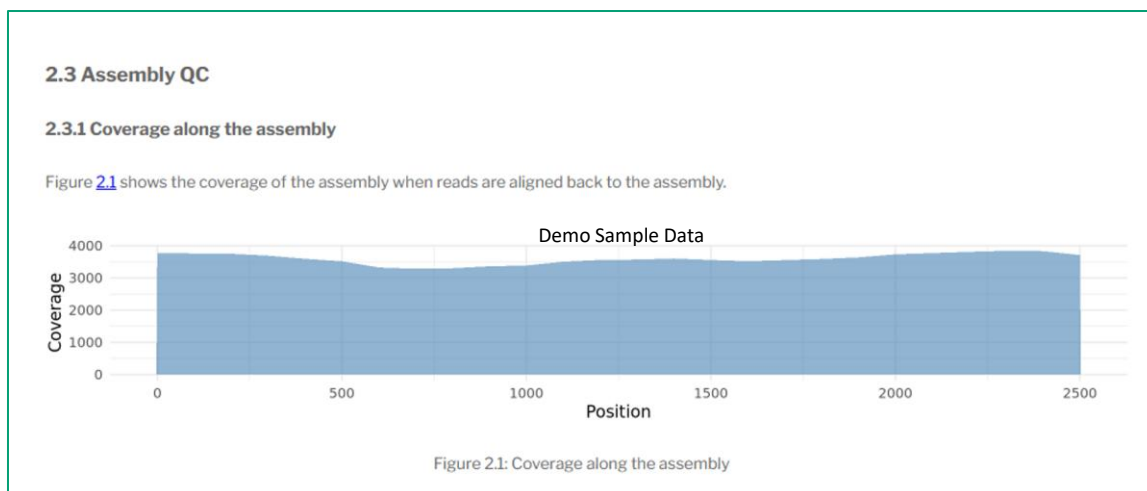
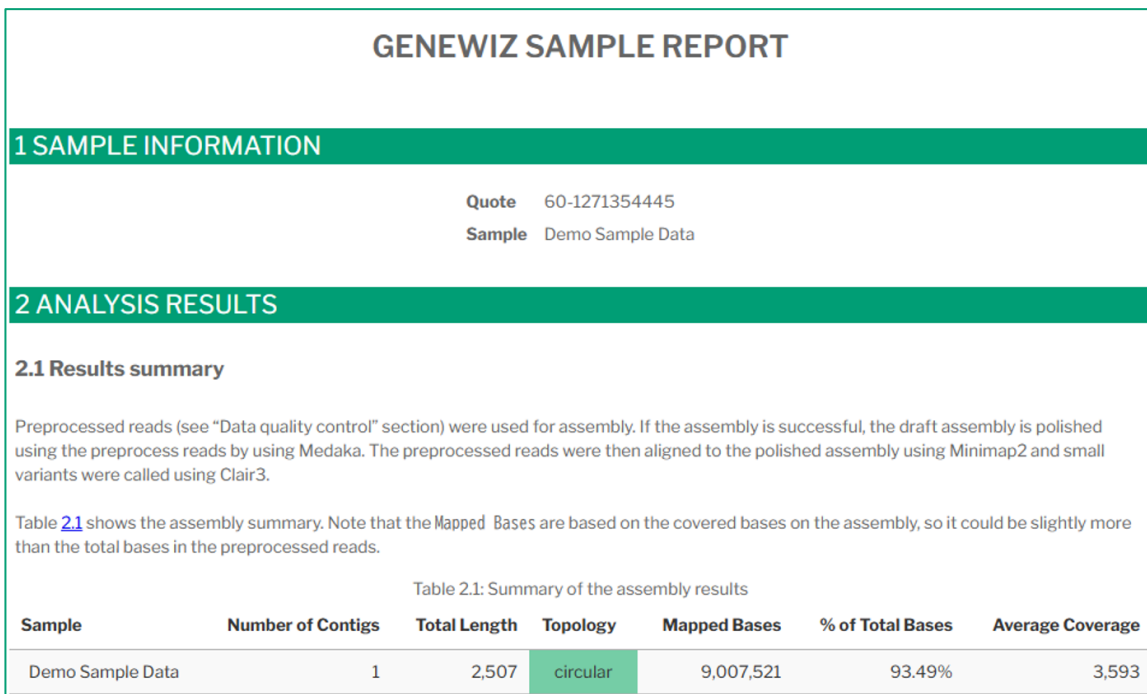
複数のコンティグが同定された場合は、各アセンブル済みコンティグごとに個別の行として、アセンブリ情報が表示されます。

代表的なトポロジーは Circular、Linear、Failed です。以下に、すべてのトポロジーの種類を示します。

- **Circular** : 閉環（クローズドループ）したコンティグがアセンブルされている
- **Linear** : 両端を持つ直線状コンティグがアセンブルされている
- **Multiple circular** : 複数の円環状コンティグがアセンブルされている
- **Mixed** : 直線状コンティグと円環状コンティグの両方がアセンブルされている
- **Multiple linear** : 複数の直線状コンティグがアセンブルされている
- **Failed** : アセンブリ未成立（アセンブルなし）

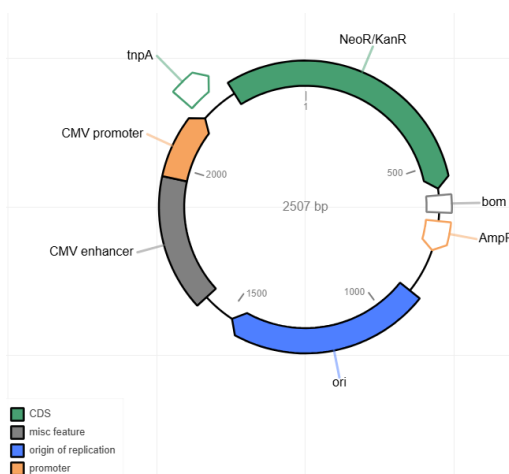
また、Assembly QC も提供されており、塩基位置ごとのカバレッジが表示されます。カバレッジが低い位置は、このグラフから容易に特定でき、ab1 ファイルや variation ファイルでの詳細確認に役立ちます。

## 解析結果（ANALYSIS RESULTS）続き



## アノテーションマップ

サンプルレポートには、インタラクティブなアノテーション付きプラスミドマップも含まれています。



## Plasmid-EZ クイックスタートガイド

## 解析結果（ANALYSIS RESULTS）続き

## アノテーションテーブル

プラスミドマップの下には、各アノテーション領域の詳細情報を示した表が表示されます。

**2.2.2 Annotation Table**

Table 2.2 shows the feature annotation details.

Show  entries

Table 2.2: Annotated features in the assembly

Search:

Contig	Feature	Type	Start	End	Strand	% Identity	% Matched Length	Description
Demo Sample Data	CMV enhancer	enhancer	1,585	1,965	1	100.0	100.0	human cytomegalovirus immediate early enhancer
Demo Sample Data	CMV promoter	promoter	1,965	2,169	1	100.0	100.0	human cytomegalovirus (CMV) immediate early promoter
Demo Sample Data	NeoR/KanR	CDS	2,284	572	1	99.7	100.0	aminoglycoside phosphotransferase from Tn5; aph(3')-II (or nptII); confers resistance to neomycin, kanamycin, and G418 (Geneticin®)

## データ品質管理（QC）

プラスミドマップの下には、各アノテーション領域の詳細情報が表示されます。

サンプルレポートのデータ品質管理（QC）では、RAW リードの長さ分布および品質分布が提供されます。

左側のグラフはリード長ごとのリード数を示しています。最適な結果では、目的のコンティグ長に対応した単一のピークが現れ、塩基レベルでの劣化（2 kb 未満のリード長）が一部に見られる程度となります。

右側のグラフは、RAW リードの品質分布を示しています。高品質なサンプルでは、グラフのピークが QS（Quality Score）20～30 付近に偏る傾向があります。

## 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 <i>E. coli</i> Sequences
Demo Sample Data	9,719,346	6,919	19.4	0.87%

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

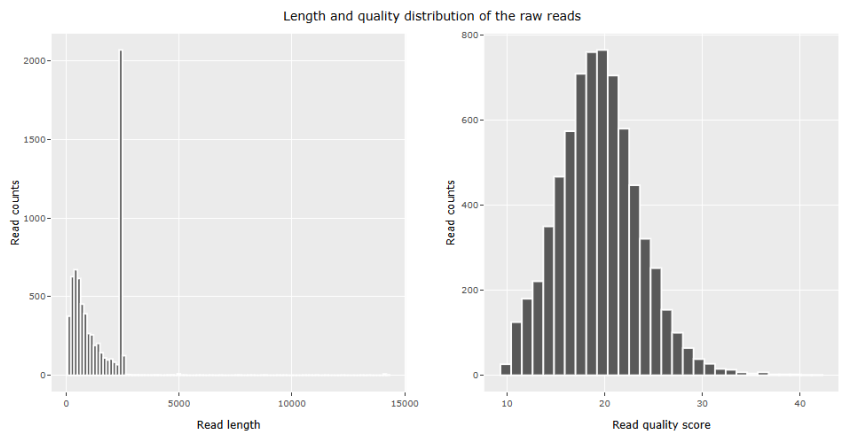


Figure 3.1: Length and quality distribution of the raw reads

## Plasmid-EZ クイックスタートガイド

## アセンブリ失敗レポート (Failed Assembly Report)

アセンブリに失敗したサンプルは、**プロジェクトレポート、サンプルレポート、および raw\_fastq フォルダーのみ**が納品されます。下の例は、アセンブリに失敗したサンプルに納品されるデータ例を示しています。

RAW リードの長さ分布および品質分布は、そのサンプルで取得された総リード数を示します。以下の失敗サンプルの QC 例では、**長さがばらついたリードがわずか 3 本しか取得されておらず、品質も低い**ことがわかります。このケースでは、**総リード数がアセンブリを行うには不十分**である事がアセンブリの失敗原因と考えられます。



## Plasmid-EZ クイックスタートガイド

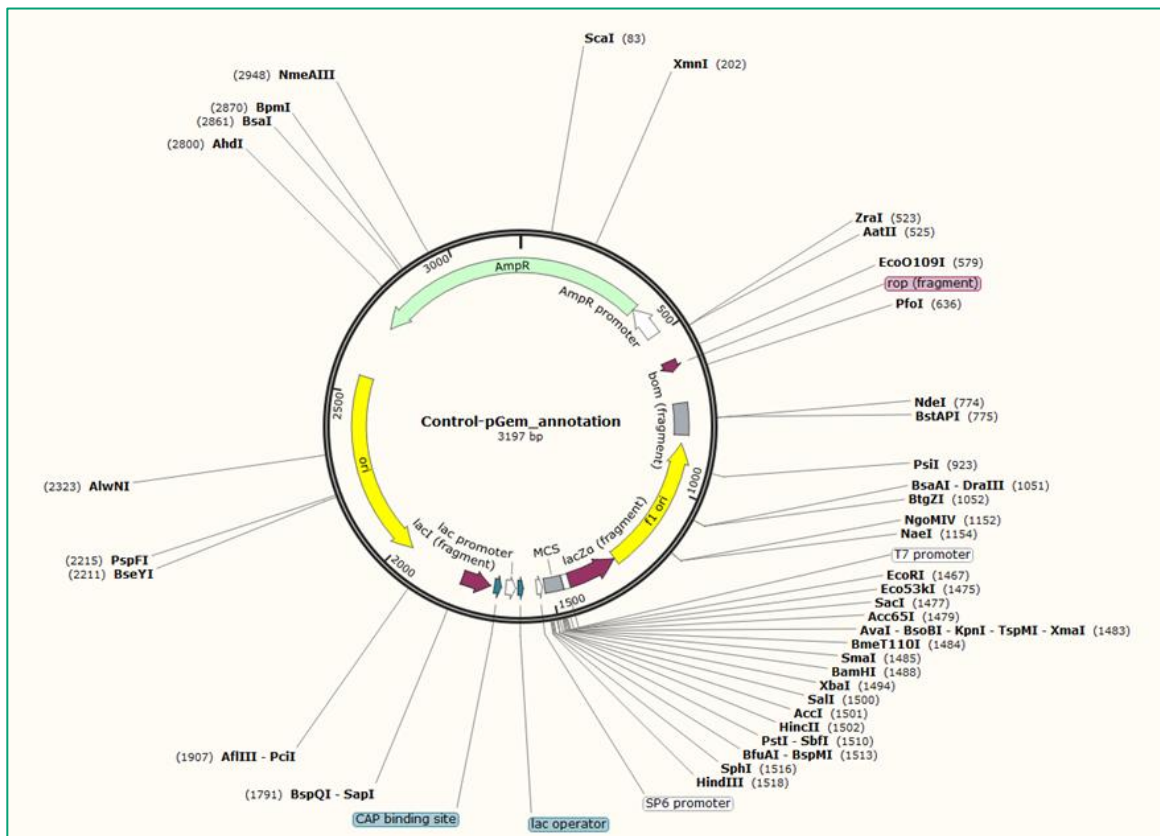
## 4.アセンブルコンティグの確認

## GeneBank

GenBank フォルダーには、アノテーション付きのアセンブル済みコンティグが含まれています。

## Assembly Fasta

assembly\_fasta フォルダーでは、アセンブルされたコンティグを確認することができます。



## Assembly Fastq

アセンブルされたコンティグの各塩基ごとの品質を迅速に確認するには、Assembly FASTQ ファイルをSnapGene Viewer、または任意の FASTA ビューアからご覧ください。各ピークにカーソルを合わせると、塩基位置および品質スコアが表示されます。



## Plasmid-EZ クイックスタートガイド

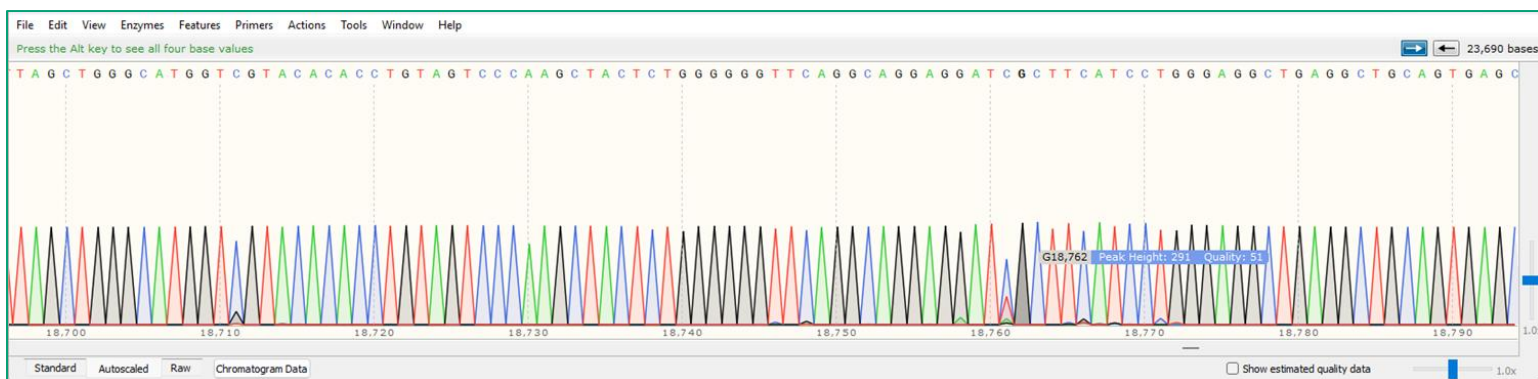
## 4.アセンブルコンティグの確認

### Assembly AB1

Sanger シーケンス形式のデータについては、確認しやすいように単一の ab1 形式で提供されています。スクロールしながら各塩基位置を確認することで、単一塩基位置に複数のピークが存在する箇所（混合塩基）を容易に特定できます。各塩基は次の色で判別可能です。**A = 緑, T = 赤, G = 黒, C = 青**

図に示すように、各位置にカーソルを合わせると、**コールされた塩基**および **品質スコア**が表示されます。

ab1 ファイルで特定された混合塩基位置については、variation の Excel ファイルで該当する位置を検索することで、混合している各塩基の**正確なリード数**をさらに詳しく確認できます。



## Plasmid-EZ クイックスタートガイド

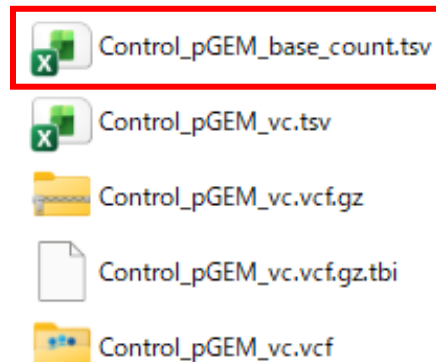
## 5. サンプルカバレッジの評価

## 変異 (variation)

variation フォルダには、**塩基コールの変動 (バリエーション)** に関する情報を複数の形式でまとめられています。

{sample\_name}\_base\_count.tsv ファイルには、**塩基位置ごとのカバレッジスコア**が記載されています。

また、このファイルでは、アセンブルされた**コンティグ内における各塩基コールの分布状況**も確認できます。



	A	B	C	D	E	F	G	H	I	J	K	L
1	Contig	Position	Reference	Coverage	A	T	G	C	N	Insertions	Top Insert	Deletions
2	Control-pGem	1 G	2573	0	0	2573	0	0	0	0 -	0	0
3	Control-pGem	2 T	2580	0	2580	0	0	0	0	0 -	0	0
4	Control-pGem	3 T	2587	0	2587	0	0	0	0	0 -	0	0
5	Control-pGem	4 A	2595	2592	0	2	1	0	0	0 -	0	0
6	Control-pGem	5 T	2596	0	2595	0	1	0	0	0 -	0	0
7	Control-pGem	6 C	2602	0	0	0	2602	0	2	AG (1)	0	0
8	Control-pGem	7 A	2607	2602	0	4	1	0	2	G (1)	0	0
9	Control-pGem	8 C	2612	0	1	0	2611	0	1	TCT (1)	0	0
10	Control-pGem	9 T	2612	0	2610	0	1	0	3	C (2)	1	1
11	Control-pGem	10 C	2617	3	4	10	2600	0	2	TTT (1)	0	0
12	Control-pGem	11 A	2625	2622	0	2	0	0	1	C (1)	1	1
13	Control-pGem	12 T	2631	0	2631	0	0	0	4	G (2)	0	0
14	Control-pGem	13 G	2636	0	0	2635	0	0	0	-	1	1
15	Control-pGem	14 G	2637	0	0	2637	0	0	1	C (1)	0	0
16	Control-pGem	15 T	2637	0	2636	0	0	0	0	-	1	1

**混合塩基、高頻度の欠失、または高頻度の挿入**が見られる位置を容易に特定するために、

{sample\_name}\_vc.tsv ファイルの確認を推奨します。このファイルでは、アセンブル済みコンティグと比較して5%を超える変動が認められる塩基位置が強調表示されます。

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

## 各略語の意味を以下に示します

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

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

## Plasmid-EZ クイックスタートガイド

## 変異 (variation) 続き

以下の例では、提供されている情報を次のように解釈できます。

Sample-1 では、5970の位置および 6556位置に変異が認められています。いずれの位置でも、主塩基は Gで、代替塩基は Aです。5970位置では、G のカバレッジが 144A のカバレッジが 46となっており、この位置での変動率は 22% です。6556位置では、G のカバレッジが 130A のカバレッジが 54であり、この位置での変動率は約 26% となります。

	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-1	6556	G	A	0/1	202	130,54	0.2673

## 各略語の意味を以下に示します

略語		略語	
CHROM	Chromosome (sample name)	GT	Genotype
POS	Position	DP	Total Depth
REF	Reference Base	AD	Allele Depth
ALT	Variant Base	AF	Allele Frequency

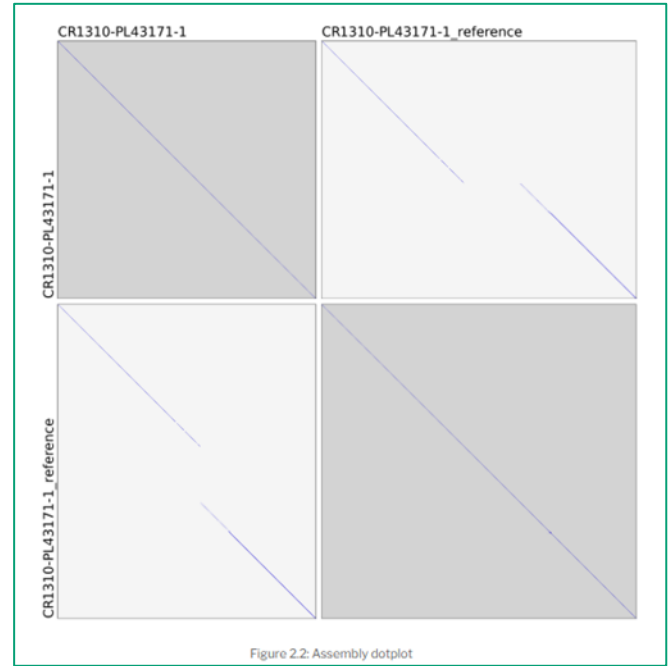
## Plasmid-EZ クイックスタートガイド

## 6. リファレンスベース解析

注文時に リファレンス配列が提供されている場合、**de novo** アセンブリ完了後にリファレンスベース解析が実行されます。  
これは、リファレンス配列によるバイアスを除去する目的で実行されます。

サンプルレポートには、アセンブル済みコンティグとリファレンス配列のアラインメントを比較したドットプロットとカバレッジプロットが表示されます。

**ドットプロット**および**カバレッジプロット**は、アセンブル済みコンティグとリファレンス配列間のアラインメント状況を迅速に確認するためのツールです。



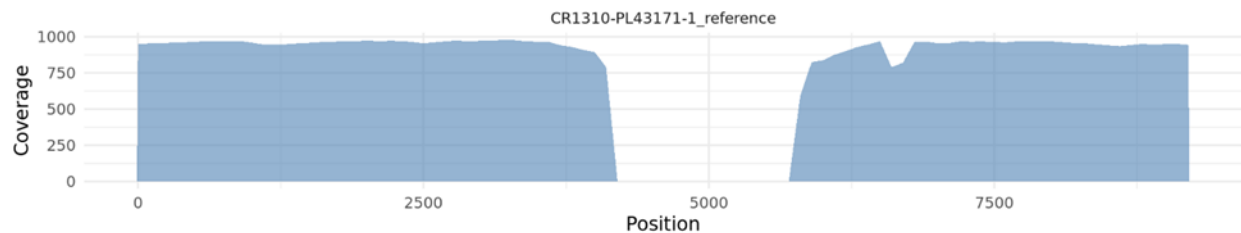
## カバレッジ解析

リファレンスベース解析で提供される **カバレッジプロット**は、アセンブル済みコンティグとリファレンス配列とのアラインメント状況を示しています。以下の例では、アセンブル済みコンティグとリファレンス配列の間で、**ほぼ全域にわたって高いカバレッジが確認できますが、アセンブル済みコンティグ側に約 1.5 kb の欠失領域が存在しています。**この結果から考えられる解釈としては、**サンプル中の挿入領域が、想定どおりに組み込まれていなかった可能性が挙げられます。**

## 2.4 Reference Based Analysis

A reference 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.








## Plasmid-EZ クイックスタートガイド

## バリエーション

reference-based analysis フォルダには、アセンブリ済みコンティグとリファレンス配列を比較するための各種 Excel ファイルが含まれています。

以下は、各サンプルで提供されるファイルの一覧です。

 Control_asm_vs_ref.var.txt	アセンブリとリファレンスファイル間の変異情報
 Control_vs_ref_base_count.tsv	コンティグとリファレンスファイル間における塩基コールおよびカウントの比較
 Control_vs_ref_vc.tsv	タブ区切りテキスト形式のバリエーションコール結果
 Control_vs_ref_vc.vcf.gz	リファレンスに基づく VCF 形式のバリエーションコール結果
 Control_vs_ref_vc.vcf.gz.tbi	VCF ファイル用インデックスファイル

annotation

assembly\_ab1

assembly\_fasta

assembly\_fastq

genebank

raw\_fastq

ref\_based\_analysis

sample\_reports

variation

30-1231967437\_project\_report.html

アセンブリで検出された変異を確認するには、vs\_ref.var.txt ファイルを参照することを推奨します。このファイルを Excel で開くには、圧縮ファイルを展開し、その後ファイルを Excel へドラッグしてください。

下の画像は、バリエーションファイルの一例で、塩基位置ごとの変異が一覧表示されています。いくつかの塩基位置の解釈は以下のとおりです。

3484-3485 位では、リファレンスの塩基が C とされていますが、アセンブル結果では G が検出されています。

4188-5829 位では、リファレンスに塩基が存在する一方で、コンティグにはこの領域が欠失していることが示されています。

	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-PL43171-1	3485	3486	+
5	V	CR1310-PL43171-1_reference	3486	3487	1	60	g	c	CR1310-PL43171-1	3486	3487	+
6	V	CR1310-PL43171-1_reference	3487	3488	1	60	t	a	CR1310-PL43171-1	3487	3488	+
7	V	CR1310-PL43171-1_reference	3489	3490	1	60	g	c	CR1310-PL43171-1	3489	3490	+
8	V	CR1310-PL43171-1_reference	3733	3734	1	60	a	g	CR1310-PL43171-1	3733	3734	+
9	V	CR1310-PL43171-1_reference	3734	3735	1	60	a	g	CR1310-PL43171-1	3734	3735	+
10	V	CR1310-PL43171-1_reference	3735	3736	1	60	g	c	CR1310-PL43171-1	3735	3736	+
11	V	CR1310-PL43171-1_reference	3736	3737	1	60	c	g	CR1310-PL43171-1	3736	3737	+
12	V	CR1310-PL43171-1_reference	4085	4086	1	60	a	g	CR1310-PL43171-1	4085	4086	+
13	V	CR1310-PL43171-1_reference	4090	4091	1	60	c	a	CR1310-PL43171-1	4090	4091	+
14	V	CR1310-PL43171-1_reference	4091	4092	1	60	g	c	CR1310-PL43171-1	4091	4092	+
15	V	CR1310-PL43171-1_reference	4188	5829	1	60	cggcgccg	-	CR1310-PL43171-1	4188	4188	+

## Plasmid-EZ クイックスタートガイド

## 7. Plasmid-EZアセンブリに影響する主な要因

## 濃度およびサンプル品質

サンプルがアセンブリを生成できなかった場合、納品データには **Project Report**、**Sample Report**、および **RAW FASTQ リードのみ**が含まれます。

raw\_fastq

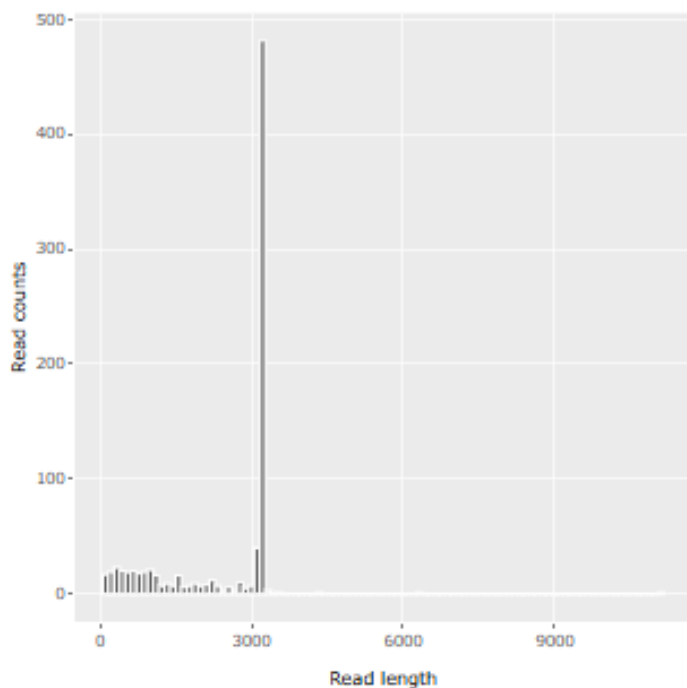
sample\_reports

30-1238252628\_project\_report.html

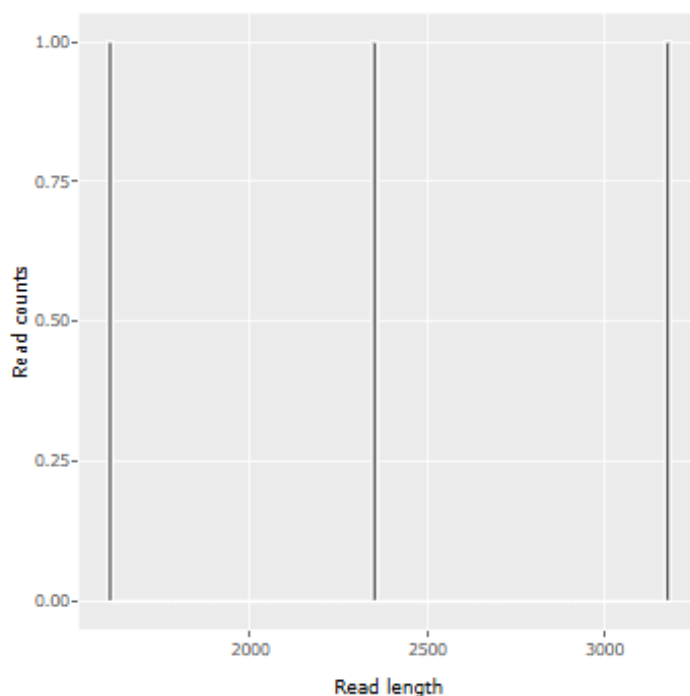
迅速な解析と低コストを実現するため、当社では**アセンブリに失敗した原因を特定するためのサンプルQCは実施していません**。しかし、最も一般的な失敗原因は、サンプル濃度が必要条件である 50 ng/μl に達していないことです。濃度が低い場合、**ライブラリー調製中に断片化が増加する**、あるいは**十分なリード数が得られない**といった問題が生じる可能性があります。失敗のリスクを低減するため、**サンプル送付前に Qubit または同等の機器で濃度を確認することを強く推奨します**。

下図の **左** は、アセンブリに成功したサンプルにおける**理想的なリード長分布**を示しています。一方、**右** はアセンブリに失敗したサンプルのリード長グラフです。

左図のように **明確なプラスミドピークが確認できるサンプル**は、アセンブリが正常に成立する傾向があります。これに対し、**リード数が少なく、明確なコンティグが形成されないサンプル**では、アセンブリに失敗することが多く見られます。



約 3 kb 付近に主要なピークが 1つ確認され、それ以外はリード数が非常に少ないことが分かります。この分布は、高品質な単一コンティグが存在することを示しています。



3種類の長さに対してそれぞれ1リードのみが検出されています。このようなリード数および分布はアセンブリには不十分であり、サンプル濃度が低いことを示唆しています。

## Plasmid-EZ クイックスタートガイド

## ホモポリマー反復領域

Plasmid-EZ では、**ホモポリマー反復領域の直後に挿入や欠失 (insertions/deletions) が検出されることがあります**。この際には、該当領域での挿入／欠失の有無を確認するため、**variation ファイルをご確認ください**。以下の例では、ホモポリマー領域の開始部分に A が 1 塩基追加で挿入されていることが示されています。さらなる確認のために、該当領域について **Sanger シーケンス解析を実施し、十分なカバレッジを確保すること**を推奨しています。

下に示す **base\_count.tsv の例**では、ホモポリマー反復領域が確認できます。

いくつかの位置で挿入の可能性が見られますが、**有意な数の挿入が確認されたのは位置 5053 のみ**です。

	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-PL43171-1	5053	T	779	84	682	1	0	0	267	AAA (50)	12
4	CR1310-PL43171-1	5054	A	903	569	3	0	0	0	0	-	331
5	CR1310-PL43171-1	5055	A	619	619	0	0	0	0	0	-	0
6	CR1310-PL43171-1	5056	A	666	666	0	0	0	0	0	-	0
7	CR1310-PL43171-1	5057	A	697	697	0	0	0	0	0	-	0
8	CR1310-PL43171-1	5058	A	736	736	0	0	0	0	1	T (1)	0
9	CR1310-PL43171-1	5059	A	762	761	1	0	0	0	3	TTTT (2)	0

同一サンプルの **vc.tsv ファイル**では、**有意な挿入が認められた位置 5053 のみ**が強調表示されています。このチャートによると、**位置 5053 の主塩基は T**ですが、**代替塩基として TAAA が約 5% の変動率で検出**されています。

	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

## メチル化部位

**GATC、CCAGG、CCTGG** は、E. coli における **メチル化部位**です。これらの部位では、塩基コールに不一致が生じることがあり、**variation ファイルでの確認が必要**となる場合があります。

不一致が **GATC** メチル化部位に起因する場合、**variation ファイル上では、A と G がほぼ半分ずつ検出**される (half A / half G) ことが挙げられます。

こうした場合には、**塩基位置をさらに確認**するために、**Sanger シーケンスによる追加解析**が必要となる場合があります。

## Plasmid-EZ の各種お問い合わせ

E-mail (DNASeq.JP@azenta.com)

TEL (03-6628-2950)