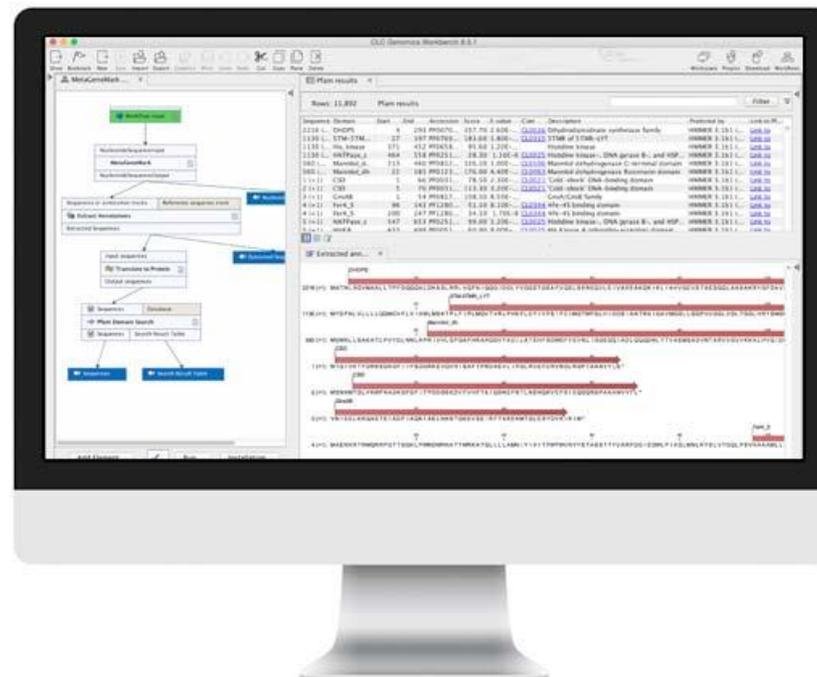


CLC bioプラグインを用いた微生物ゲノム 配列決定とアノテーション解析

フィルジエン株式会社 バイオサイエンス部
(biosupport@filgen.jp)

- 新規ゲノムの配列解析を行う場合、ゲノム全体の塩基配列決定に加えて、各遺伝子の配列やアノテーションなどの機能解析も行う必要がある。
- CLC Genomics Workbenchには、おもに微生物ゲノム解析用に開発された3種類の有償プラグインがあり、これらを活用することによって、配列決定やアノテーション解析を容易に行うことが可能になる。
- アノテーション解析まで行っておけば、今後そのデータをリファレンスゲノムとして利用して、SNPなどの変異解析や、RNA-Seq解析に応用できる。



CLC Genome Finishing Module

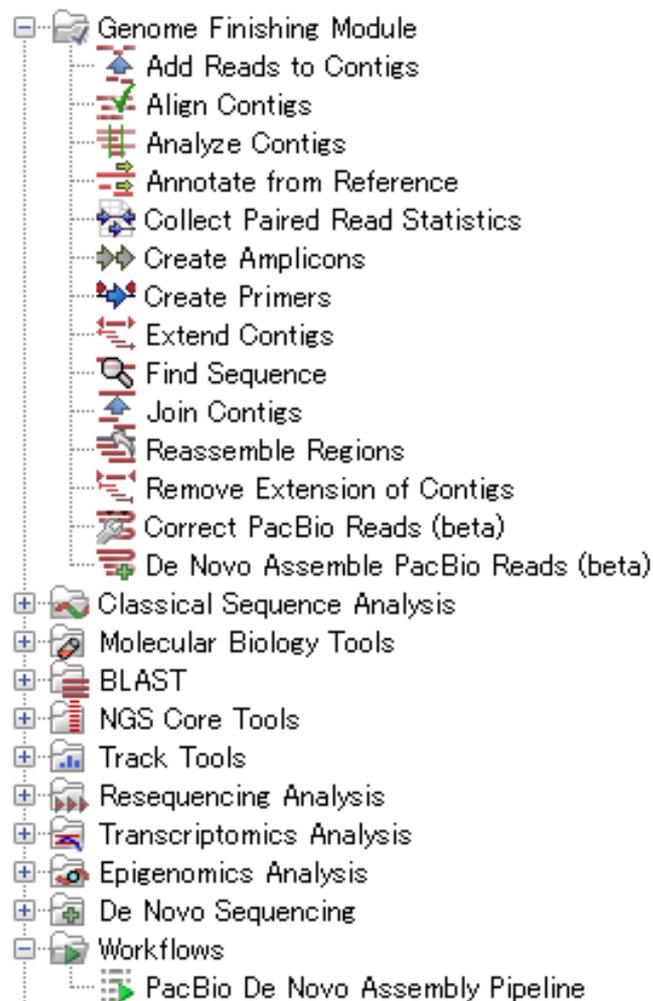
主にスモールサイズゲノムにおける、高品質な完全長ゲノム配列を得ることを目的としたGenome Finishingを行うためのプラグイン。PacBioロングリードを用いた、De Novo Assemblyが可能。

CLC Microbial Genomics Module

微生物ゲノム解析用の専用プラグイン。16S rRNA菌叢解析や病原菌タイピング・系統樹解析、全メタゲノム解析といった、多数の専用アプリケーションが使用可能になる。

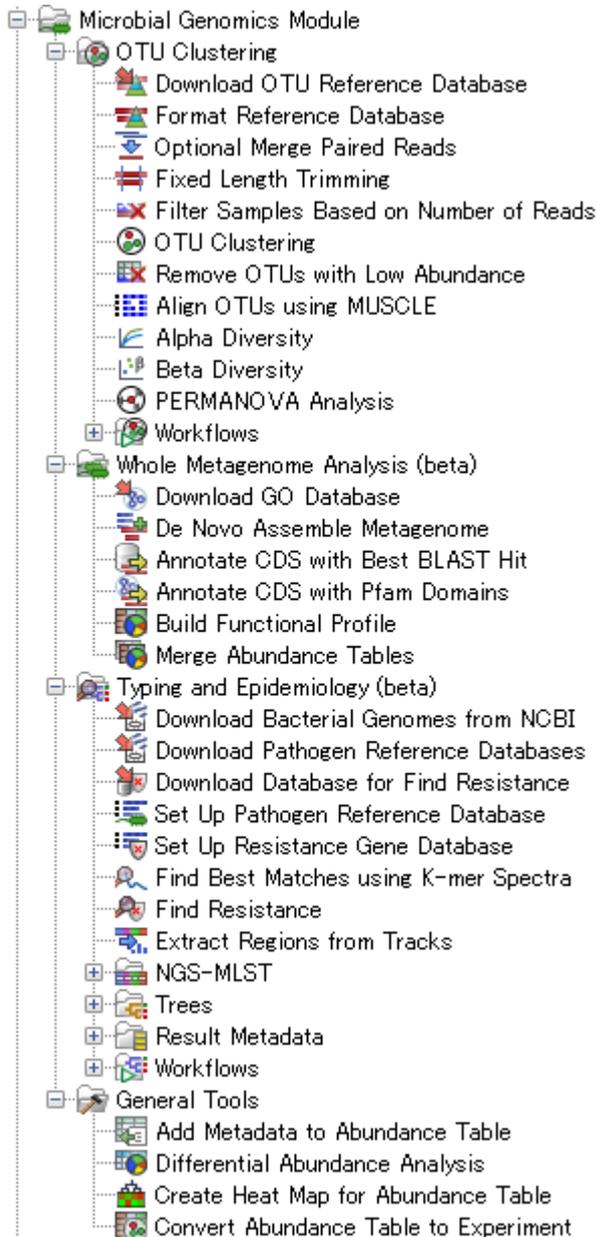
MetaGeneMark

微生物のメタゲノム配列データに対して、遺伝子とタンパク質コード領域の検出を行うプラグイン。



使用可能になるアプリケーション

- コンティグ配列における、リファレンス配列またはコンティグ配列自身へのアライメント
- コンティグ配列同士の結合
- コンティグ配列におけるミスアSEMBル、低カバレッジ領域、ペアリード配列のマッピング状況などの確認
- PCRプライマーの自動設計
- PacBioロングリード配列データのインポート
- PacBioロングリード配列データのエラー補正およびアSEMBル（ベータ版）
- PacBio De NovoアSEMBル用の解析パイプライン（ベータ版）



使用可能になるアプリケーション

OTU Clustering

- リード配列データの各種QCチェックおよびOTUクラスタリングの実行
- α 多様性と β 多様性の計算
- Greengenes, SILVA, UNITEの各種データベースからのリファレンスデータの自動ダウンロード

Typing and Epidemiology (ベータ版)

- NGS-MLST (Multi Locus Sequence Typing) 解析による病原菌のタイピングおよび薬剤耐性の確認
- K-mer Treeによる複数菌種のゲノム配列の類似度の比較
- SNP Treeによる分子系統樹の作成

Whole Metagenome Analysis (ベータ版)

(別途有償プラグイン「MetaGeneMark」が必要)

- メタゲノムシーケンスデータのDe Novoアセンブル
- BLAST検索、Pfamドメイン検索によるアノテーション付け



使用可能になるアプリケーション

微生物ゲノムまたはトランスクリプトーム配列データに対して、遺伝子領域とタンパク質コード領域を予測し、アノテーションを付加する。

手順1 : **Contig配列データの作成**

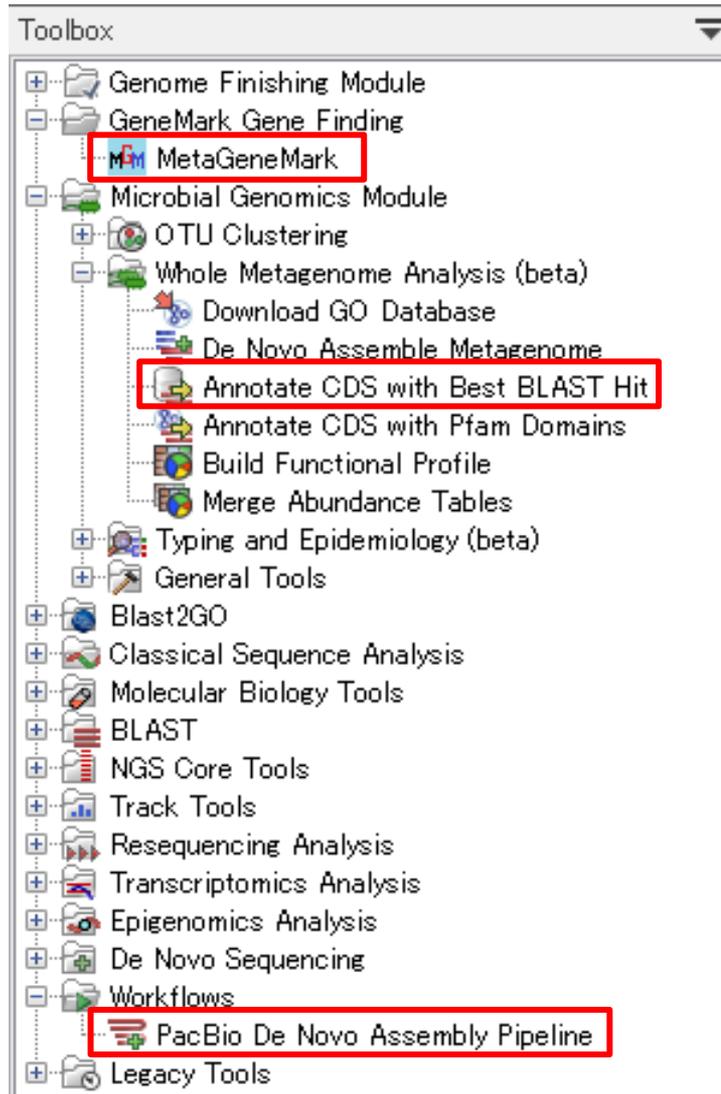
- CLC Genome Finishing Moduleを用いた、PacBioロングリード配列データのDe Novo AssemblyによるContig配列データの作成

手順2 : **遺伝子領域の予測**

- MetaGeneMarkを用いた、Contig配列内の遺伝子・タンパク質コード領域の予測

手順3 : **遺伝子機能アノテーションの付加**

- CLC Microbial Genomics Moduleを用いた、Contig配列内の各遺伝子に対する機能アノテーション情報の付加



PacBio De Novo Assembly Pipeline

- PacBioロングリード配列データを使用した De Novo Assembly (CLC Genome Finishing Module)



MetaGeneMark

- Contig配列上の遺伝子領域の予測 (MetaGeneMark)



Annotate CDS with Best BLAST Hit

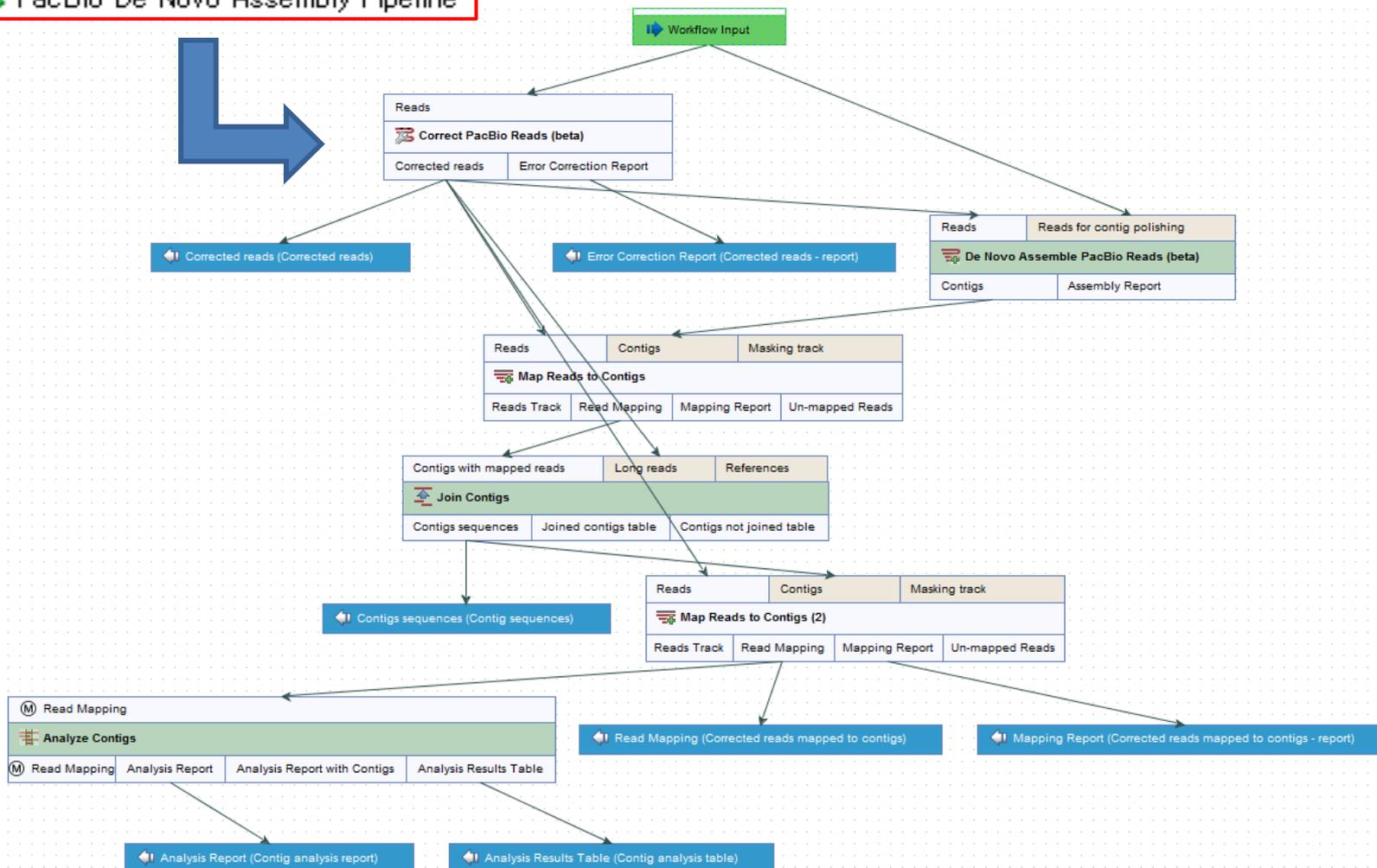
- CDS配列データへのアノテーション付け (CLC Microbial Genomics Module)

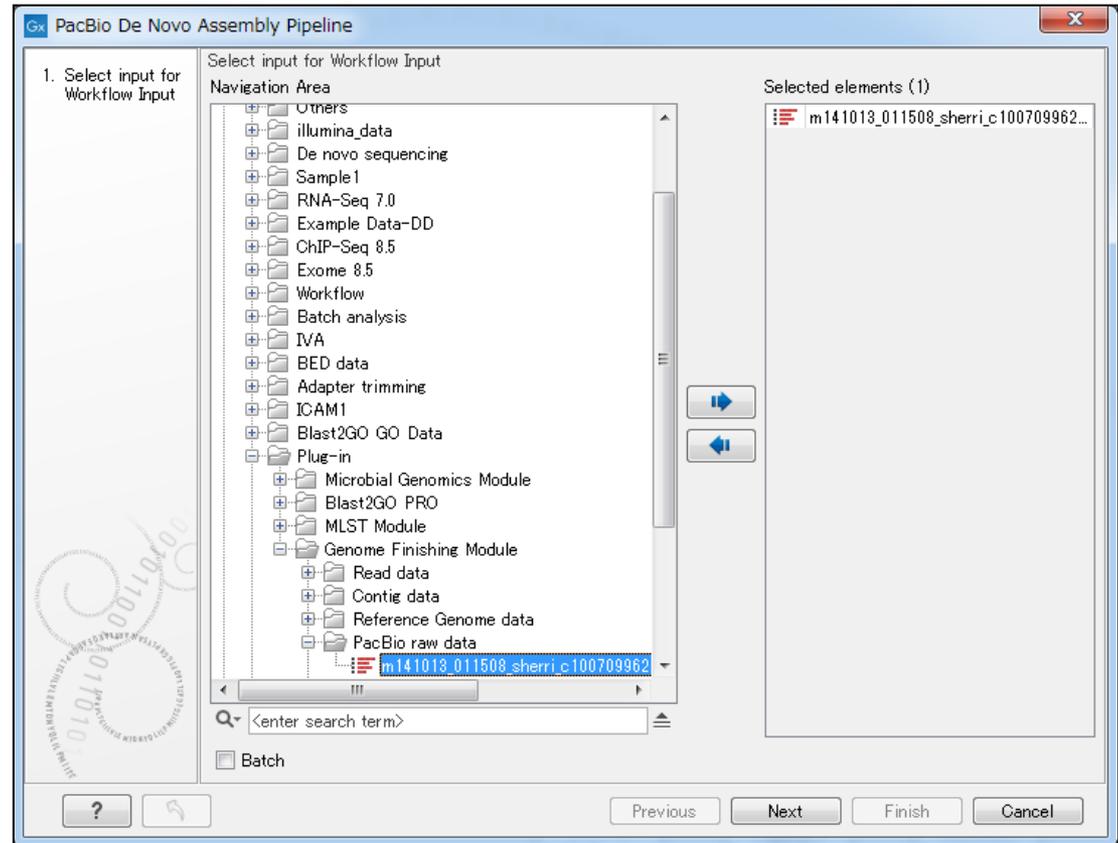
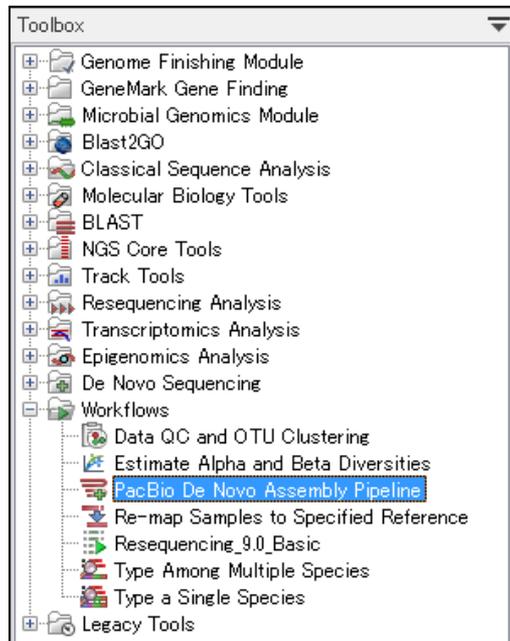
手順1 : Contig配列データの作成

PacBio De Novo Assembly Pipeline

- 標準搭載の解析パイプラインを使用することで、PacBioロングリード配列データのエラー補正、アSEMBル、レポート作成などをまとめて実行することができる。

PacBio De Novo Assembly Pipeline





1. PacBio De Novo Assembly Pipelineを選択し、ダブルクリック。
2. PacBioロングリードデータを選択。

PacBio De Novo Assembly Pipeline

1. Select input for Workflow Input

2. Correct PacBio Reads (beta)

Correct PacBio Reads (beta)

Coverage percentage of reads to correct 30

Previous Next Finish Cancel

- Coverage percentage of reads to correct : 全リードのうち、配列長が大きい何%のリードを基準として、エラー補正を行うかを指定する。

PacBio De Novo Assembly Pipeline

1. Select input for Workflow Input

2. Correct PacBio Reads (beta)

3. De Novo Assemble PacBio Reads (beta)

De Novo Assemble PacBio Reads (beta)

Configurable Parameters

Automatic word size

Word size 24

Minimum word coverage 4

Minimum contig length 5,000

Locked Settings

Previous Next Finish Cancel

- Automatic word size : 自動で設定されたword sizeを使用するか、または任意の値を入力する。
- Word size : 任意で設定する場合のword sizeを入力する。
- Minimum word coverage : リード配列中に、指定のwordが出現する最低の回数。
- Minimum contig length : コンティグの配列長の最小値。

The screenshot shows the 'Join Contigs' step of the PacBio De Novo Assembly Pipeline. The left sidebar lists the workflow steps: 1. Select input for Workflow Input, 2. Correct PacBio Reads (beta), 3. De Novo Assemble PacBio Reads (beta), and 4. Join Contigs. The main panel is titled 'Join Contigs' and contains 'Configurable Parameters' with a checkbox for 'Align to reference(s)' and a text field for 'Closely related reference genome'. Below this is a 'Locked Settings' section. At the bottom, there are navigation buttons: '?', a back arrow, 'Previous', 'Next', 'Finish', and 'Cancel'.

- Align to reference(s): 指定されたリファレンス配列データを使い、コンティグ同士を結合させる。
- Closely related reference genome : リファレンス配列データを指定する。

The screenshot shows the 'Analyze Contigs' step of the PacBio De Novo Assembly Pipeline. The left sidebar lists the workflow steps: 1. Select input for Workflow Input, 2. Correct PacBio Reads (beta), 3. De Novo Assemble PacBio Reads (beta), 4. Join Contigs, and 5. Analyze Contigs. The main panel is titled 'Analyze Contigs' and contains 'Configurable Parameters' with two input fields: 'Low coverage threshold' set to 8 and 'High coverage threshold' set to 40. Below this is a 'Locked Settings' section. At the bottom, there are navigation buttons: '?', a back arrow, 'Previous', 'Next', 'Finish', and 'Cancel'.

- Low coverage threshold : 低カバレッジ領域の閾値。
- High coverage threshold : 高カバレッジ領域の閾値。

- Contig analysis report
- Contig analysis table
- Contig sequences**
- Corrected reads - report
- Corrected reads mapped to contigs - report
- Corrected reads mapped to contigs
- Corrected reads

Contig sequences x

Joined contig 1 GGTCTGGTTTACCAGTTCTAATCTGATTACGAAAAAGATATGTTGCGGGAGGCGTTGCCTCCCAACATATAAGTGGCCTCCCTC

Sequence List Settings

Sequence layout

No spacing

Double stranded

Numbers on sequences

Relative to

Numbers on plus strand

Hide labels

Lock labels

Sequence label

Name

Annotation layout Annotation types

Gap

Old sequence

Select All

Deselect All

Restriction sites

Motifs

Residue coloring

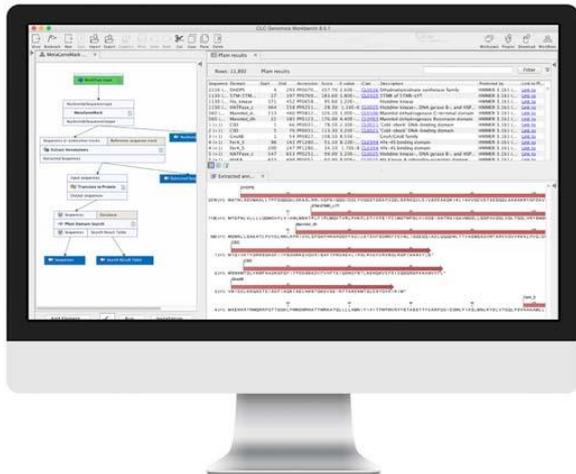
Nucleotide info

Find

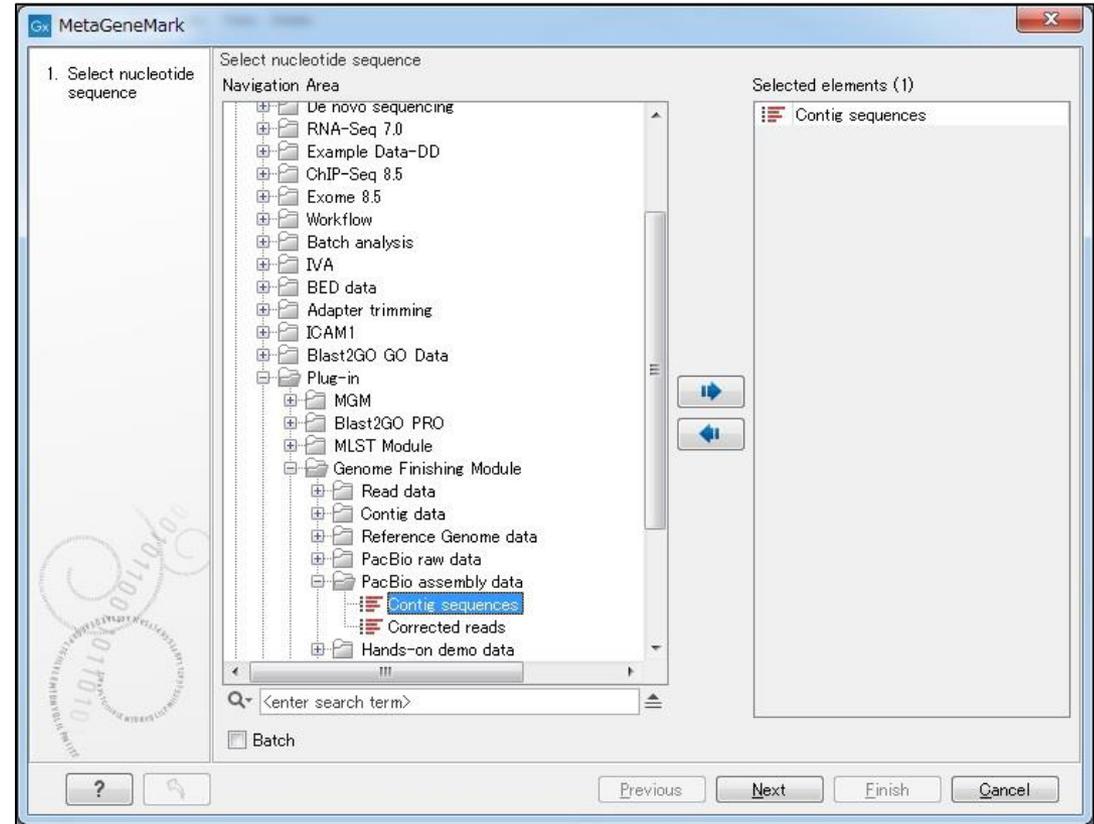
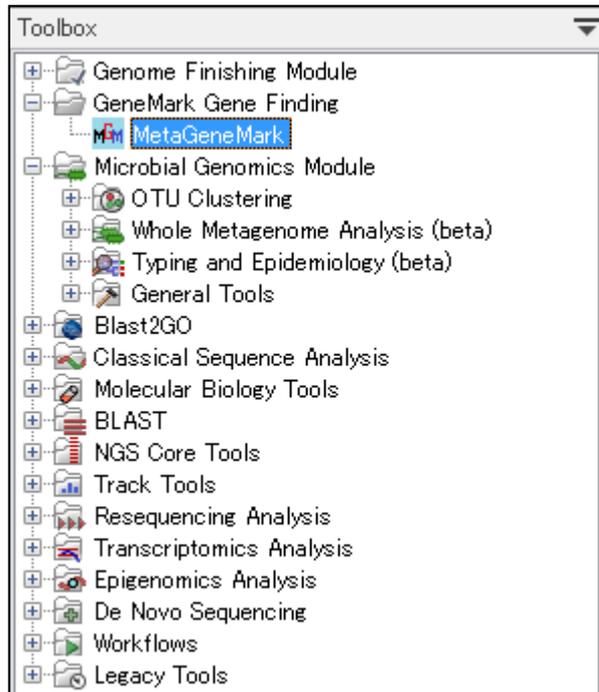
Text format

- 最終的に作成されたコンティグ配列と、アノテーションも同時に確認できる。

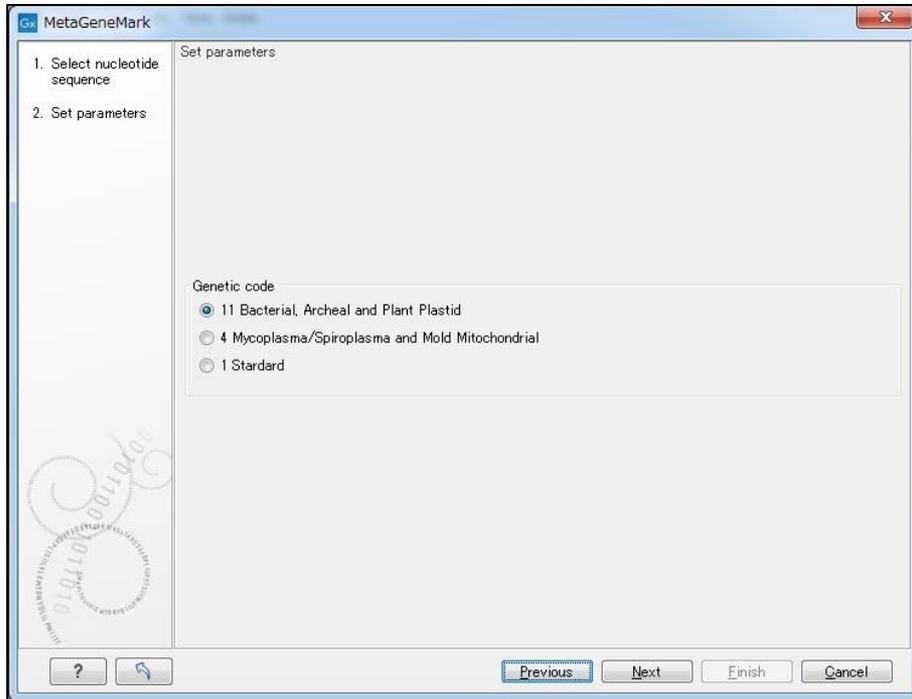
手順2：遺伝子領域の予測



- 微生物ゲノム配列データから、遺伝子CDS領域予測を行うのに必要な有償プラグイン。
- コンティグ配列データに、CDS領域のアノテーションを付加するために必要。
- 外部データソースや細かいパラメータの設定なしに、高感度な予測が可能。

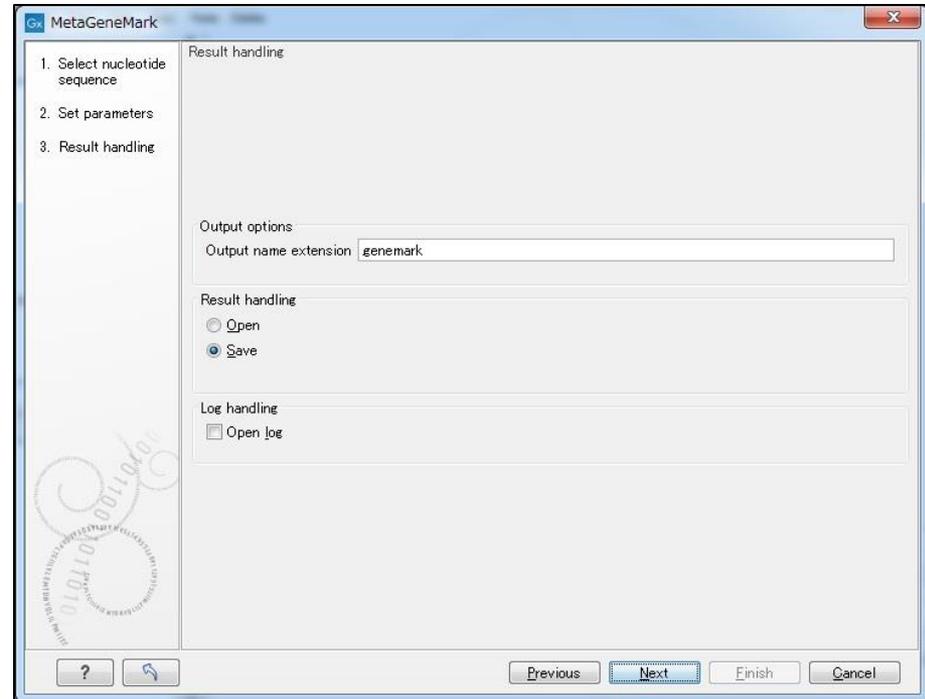


1. MetaGeneMarkを選択してダブルクリック。
2. コンティグ配列データを選択。



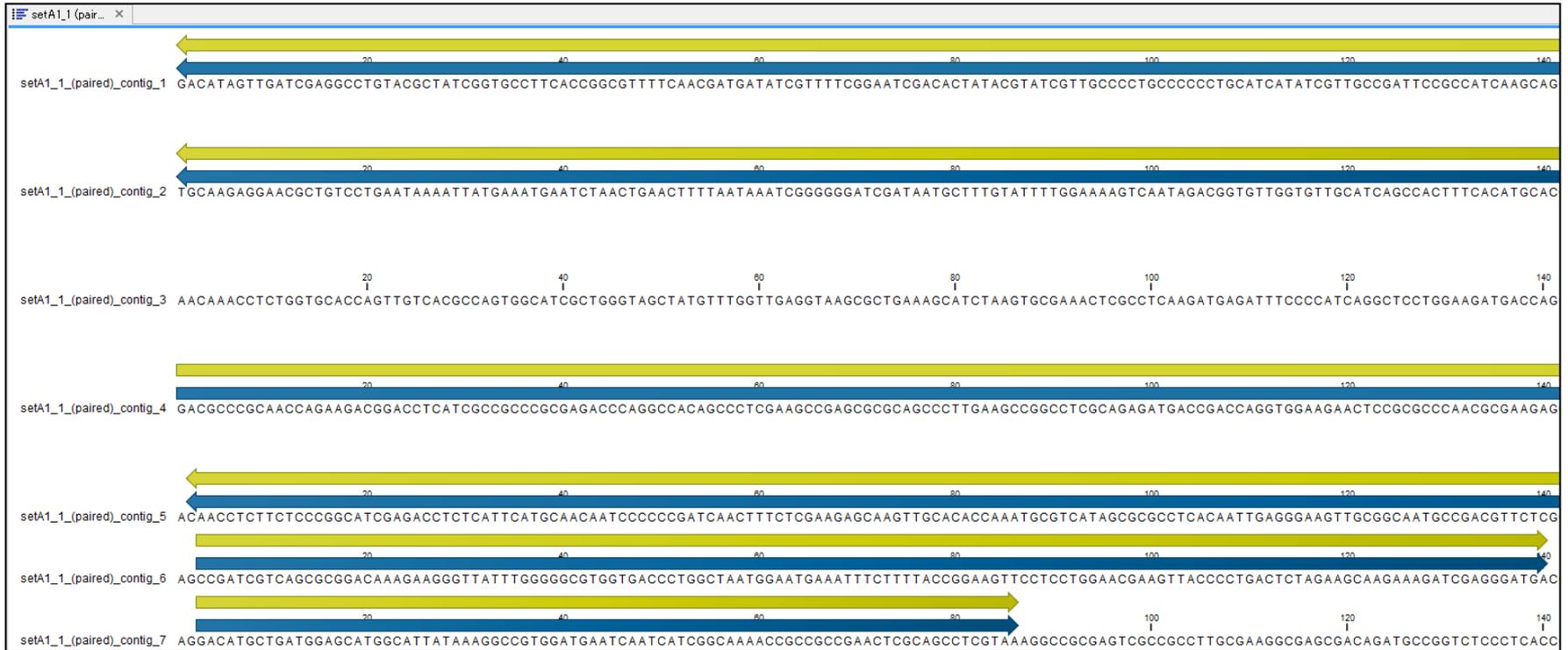
Genetic code:

- 解析データの生物種に応じて、CDS領域の塩基配列をアミノ酸に翻訳する際の翻訳テーブルを指定する。



Output options:

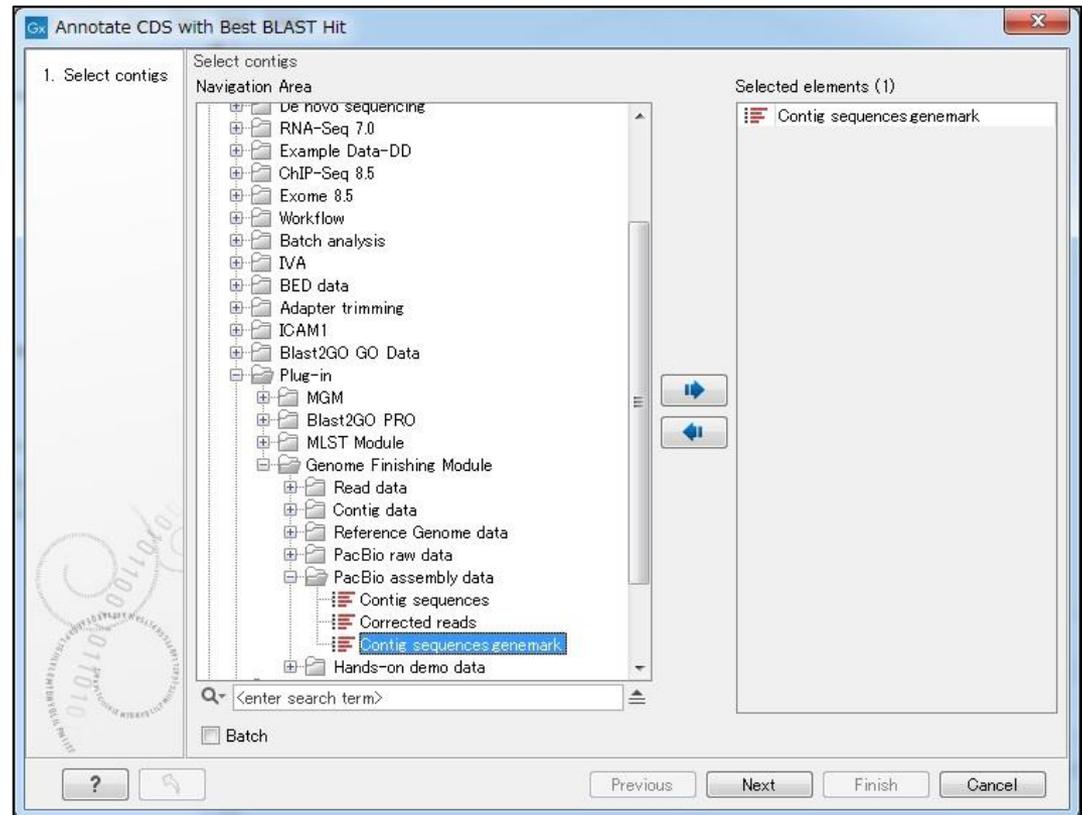
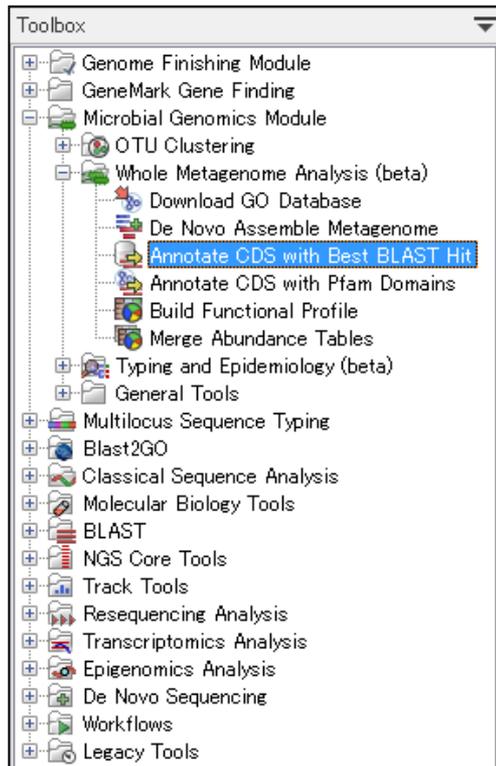
- Output name extension: 解析結果のデータ名に付加される文字列を指定する。



- 遺伝子・CDS領域を表すアノテーションが、コンティグ配列に付加される。

手順3：遺伝子機能アノテーションの付加

- MetaGeneMarkによって検出された各CDSデータに対して、配列の相同性比較や様々なデータソースから、遺伝子機能情報を付加する。
- どちらか一方のみの使用でもよい。
 1. BLASTを利用した、配列データのターゲットデータベースとの相同性比較によるアノテーション付け
(使用ツール : Annotate CDS with Best BLAST Hit)
 2. PfamとGO (Gene Ontology)データベースを利用したアノテーション付け
(使用ツール : Annotate CDS with Pfam Domains)



1. Annotate CDS with Best BLAST Hitを選択してダブルクリック。
2. CDSアノテーション付けされたコンティグ配列データを選択。

The screenshot shows the 'Parameters' tab of the 'Annotate CDS with Best BLAST Hit' application. The window title is 'Annotate CDS with Best BLAST Hit'. On the left, a sidebar lists '1. Select contigs' and '2. Parameters'. The main area is titled 'Parameters' and contains two sections: 'Genetic code parameters' and 'BLAST parameters'. Under 'Genetic code parameters', there is a dropdown menu for 'Genetic code' set to '11 Bacterial, Archaeal and Plant Plastid'. Under 'BLAST parameters', there is a dropdown menu for 'BLAST database' set to 'uniprot_sprot' and a text input field for 'Expectation value' set to '0.0001'. At the bottom, there are navigation buttons: '?', a home icon, 'Previous', 'Next', 'Finish', and 'Cancel'.

The screenshot shows the 'Result handling' tab of the 'Annotate CDS with Best BLAST Hit' application. The window title is 'Annotate CDS with Best BLAST Hit'. On the left, a sidebar lists '1. Select contigs', '2. Parameters', and '3. Result handling'. The main area is titled 'Result handling' and contains three sections: 'Output options', 'Result handling', and 'Log handling'. Under 'Output options', there is a checked checkbox for 'Create report'. Under 'Result handling', there are radio buttons for 'Open' and 'Save', with 'Save' selected. Under 'Log handling', there is a checked checkbox for 'Open log'. At the bottom, there are navigation buttons: '?', a home icon, 'Previous', 'Next', 'Finish', and 'Cancel'.

Genetic code parameters:

- Genetic code: 解析データの生物種に応じて、CDS領域の塩基配列をアミノ酸に翻訳する際の翻訳テーブルを指定する。

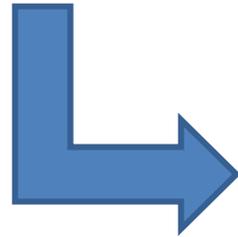
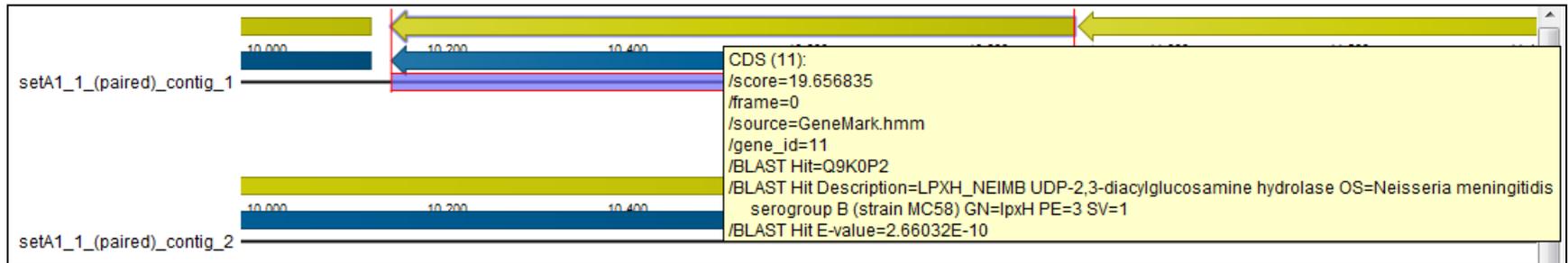
BLAST parameters:

- BLAST database: ターゲットデータベースを指定する。
(ターゲットデータベースの取得は、BLASTアプリケーションの「Download BLAST Databases」や「Create BLAST Database」から可能。)
- Expectation value: BLAST検索のE-valueの閾値。

Output options:

- Create report: 解析結果をまとめたレポートを作成する。

Annotate CDS with Best BLAST Hit



Contig sequen... x

Rows: 8,656 Filter: All

Name	Type	Region	Qualifiers
13	CDS	8189..8452	<pre>/score=15.151022 /frame=0 /source=GeneMark.hmm /gene_id=13 /BLAST Hit=P75717 /BLAST Hit Description=EXOD_ECOLI Putative uncharacterized protein ExoD OS=Escherichia coli (strain K12) GN=exoD PE=5 SV=1 /BLAST Hit E-value=1.18089E-58</pre>
14	CDS	8572..9735	<pre>/score=38.934622 /frame=0 /source=GeneMark.hmm /gene_id=14 /BLAST Hit=P24218 /BLAST Hit Description=INTD_ECOLI Prophage DLP12 integrase OS=Escherichia coli (strain K12) GN=intD PE=3 SV=1 /BLAST Hit E-value=0.0</pre>
15	CDS	10149..10781	<pre>/score=-5.860957 /frame=0 /source=GeneMark.hmm /gene_id=15 /BLAST Hit=P0AEL8 /BLAST Hit Description=FIMZ_ECOLI Fimbriae Z protein OS=Escherichia coli (strain K12) GN=fimZ PE=3 SV=1 /BLAST Hit E-value=3.54749E-142</pre>
			<pre>/score=0.229811 /frame=0</pre>

- 遺伝子機能情報がCDSアノテーションに付加される。

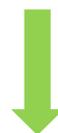
リファレンスゲノムデータの作成

☰ Contig sequences.genemark



☰ Convert to Tracks
Trackフォーマットデータへ変換

☰ Contig sequences.genemark (Genome)
☰ Contig sequences.genemark (CDS)
☰ Contig sequences.genemark (Gene)



- リシークエンス解析
- RNA-Seq解析
- ChIP-Seq解析 など

Chromosome	Region	Type	Reference	Allele	Zygosity	Count	Coverage	Frequency	Coding region change	Amino acid change	BLAST Hit ...	BLAST Hit Description (Contig sequences.genemark (CDS))
Joined contig 1	4009847~4009848	Insertion	-	G	Homozygous	16	17	94.12	3738c.293_294insG	3738p.Asp100fs	P09154	YMFS_ECOLI Uncharacterized protein YmfS OS=Escherichia
Joined contig 1	4009968	SNV	T	T	Heterozygous	13	18	72.22			P09154	YMFS_ECOLI Uncharacterized protein YmfS OS=Escherichia
Joined contig 1	4010106~4010107	Insertion	-	-	Heterozygous	19	21	90.48			P09154	YMFS_ECOLI Uncharacterized protein YmfS OS=Escherichia
Joined contig 1	4011423	SNV	G	C	Heterozygous	28	44	63.64	3740c.289C>G	3740p.Leu97Val	P33227	STFE_ECOLI Putative protein StfE (Fragment) OS=Escherich
Joined contig 1	4011423	SNV	G	G	Heterozygous	16	44	36.36			P33227	STFE_ECOLI Putative protein StfE (Fragment) OS=Escherich

コンティグ配列に付加されたアノテーション由来のデータ

お問い合わせ先: フィルジェン株式会社

TEL 052-624-4388 (9:00~17:00)

FAX 052-624-4389

E-mail: biosupport@filgen.jp