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

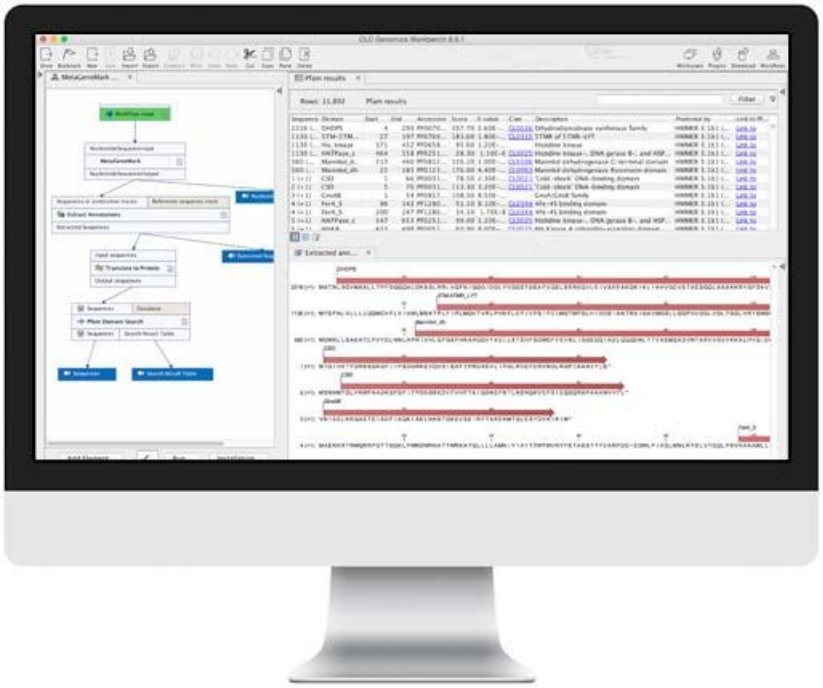
フィルジェン株式会社　バイオサイエンス部
(biosupport@filgen.jp)

はじめに

- 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。



3種類のプラグイン

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。

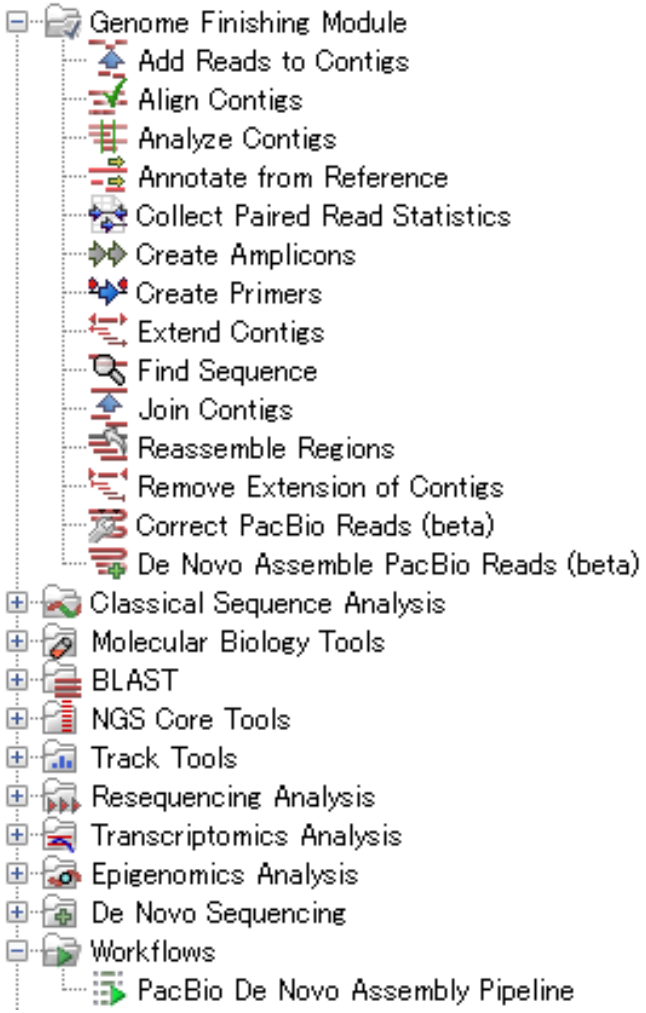
CLC Microbial Genomics Module

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。

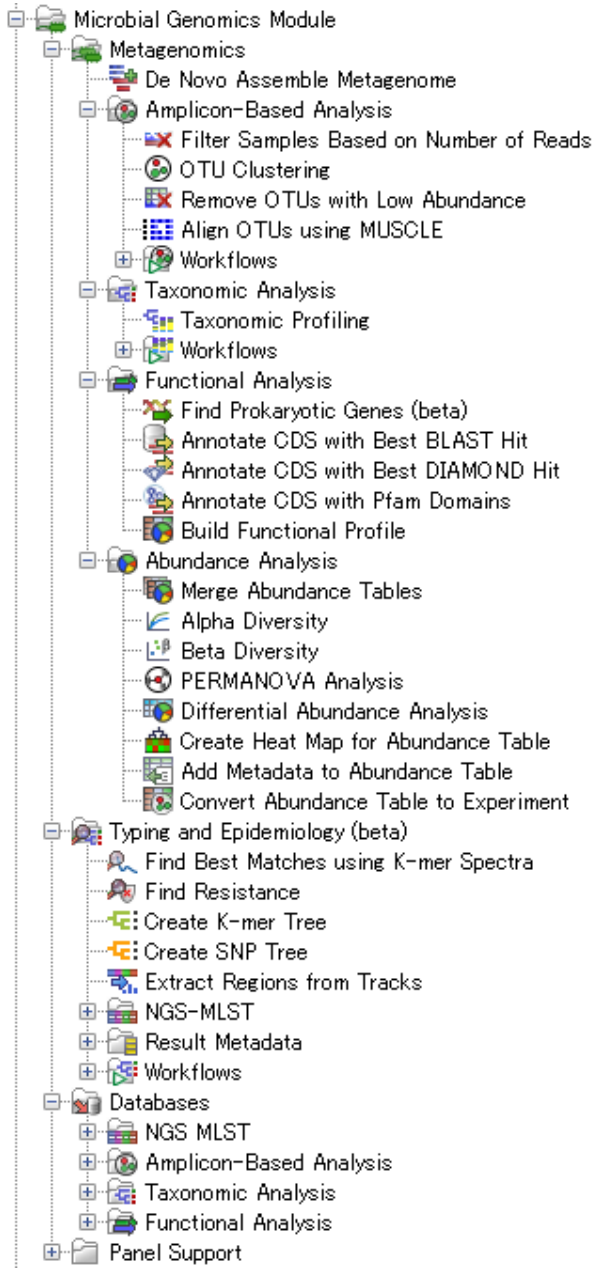
CLC Genome Finishing Module



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センブル、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センブル（ベータ版）
- PacBio De Novoアセンブル用の解析パイプライン（ベータ版）

CLC Microbial Genomics Module



使用可能になるアプリケーション

Amplicon-Based OTU Clustering

- 16S rRNAなどアンプリコンシークエンスデータの各種QCチェックおよびOTUクラスタリングによる菌種組成解析

Taxonomic Analysis

- ショットガンメタゲノムデータを用いた、宿主ゲノム配列データの除去および菌種組成解析

Functional Analysis（別途有償プラグイン「MetaGeneMark」が必要）

- メタゲノムシークエンスデータのDe Novoアセンブル
- BLAST検索、Pfamドメイン検索による遺伝子機能アノテーション付けと組成解析

Abundance Analysis

- 菌種組成データからのα多様性とβ多様性の計算
- 菌種または遺伝子機能組成データからの、サンプル間比較やヒートマップ作成

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

Databases

- NCBIからの、バクテリアやウイルスなどのゲノム配列データの一括ダウンロード
- Greengenes, SILVA, UNITEなどのOTU配列データ、MLSTスキーマや薬剤耐性遺伝子配列データのダウンロード
- カスタムデータからのデータベース作成

MetaGeneMark

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。

データ解析の流れ

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

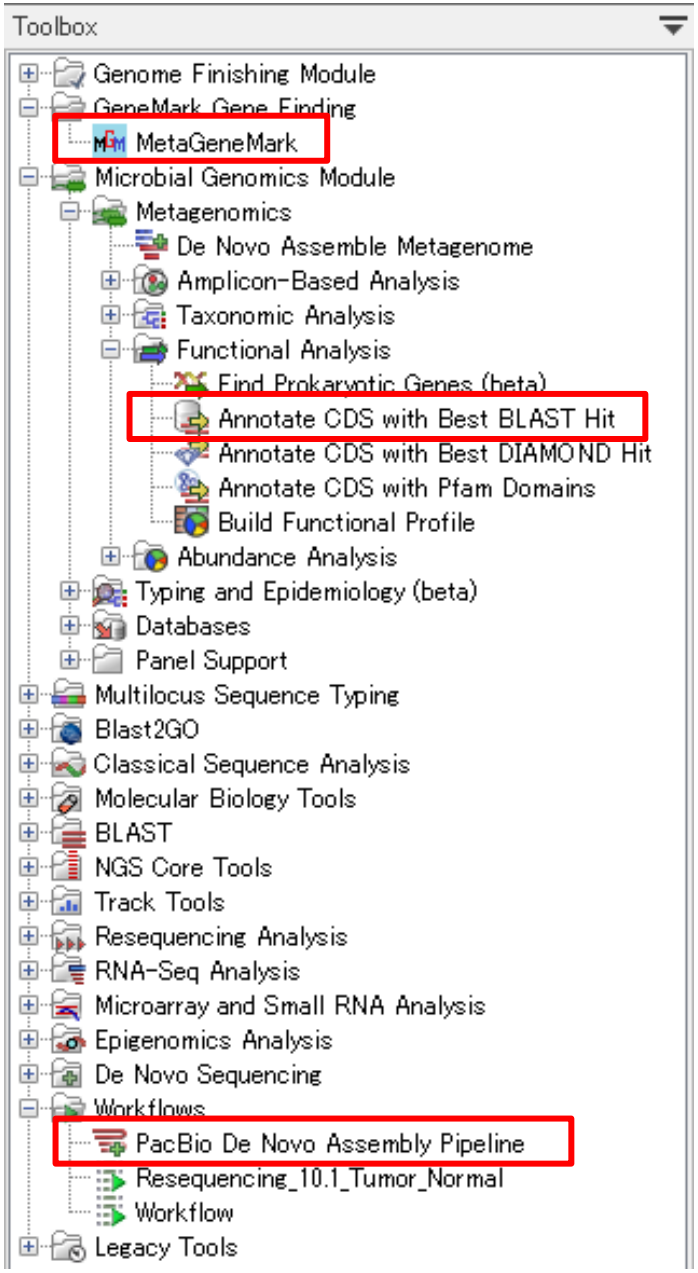
手順2： 遺伝子領域の予測

- 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

ツール使用の流れ



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)

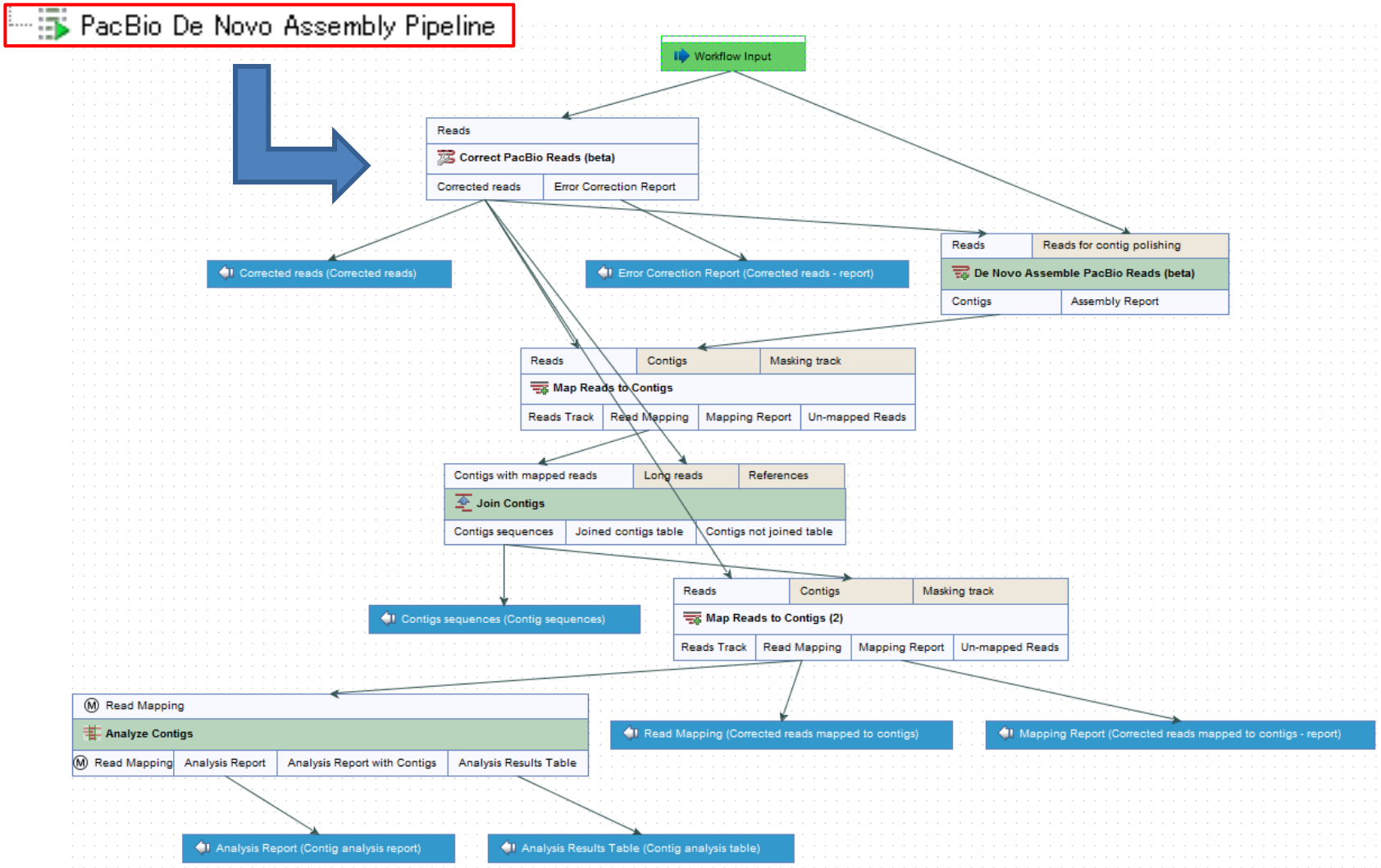
Annotate CDS with Best BLAST Hit

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

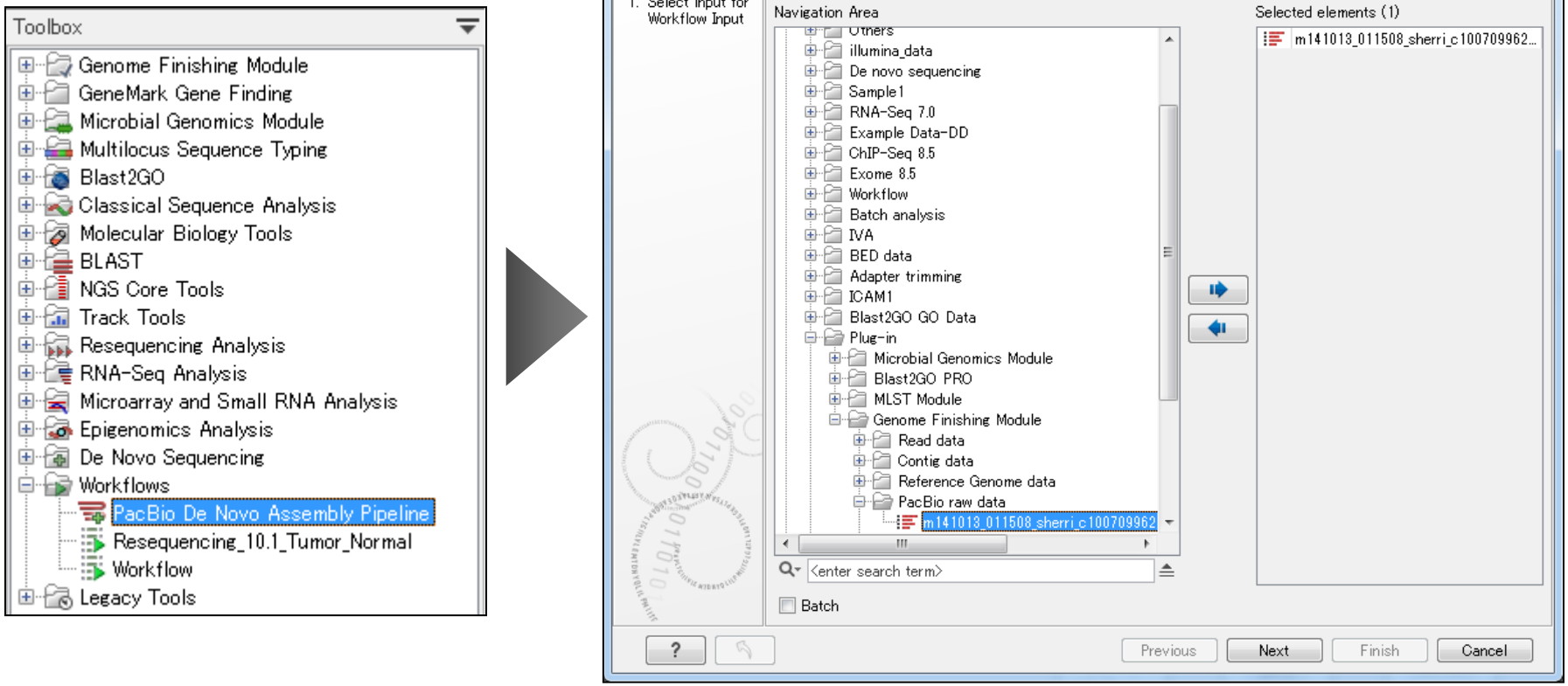
手順1：Contig配列データの作成

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、アセンブル、レポート作成などをまとめて実行することができる。

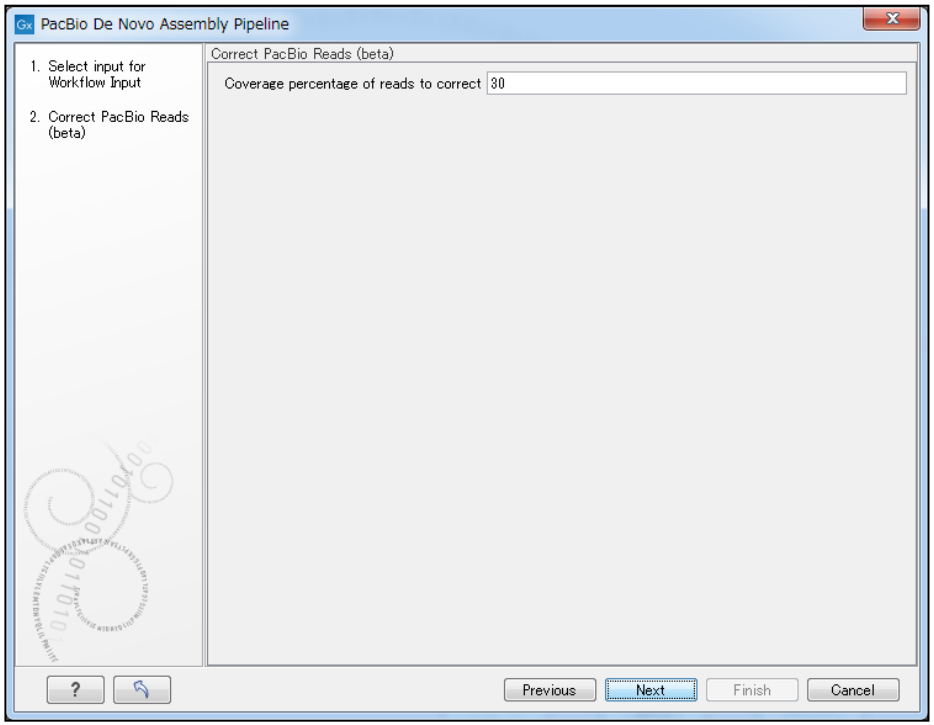


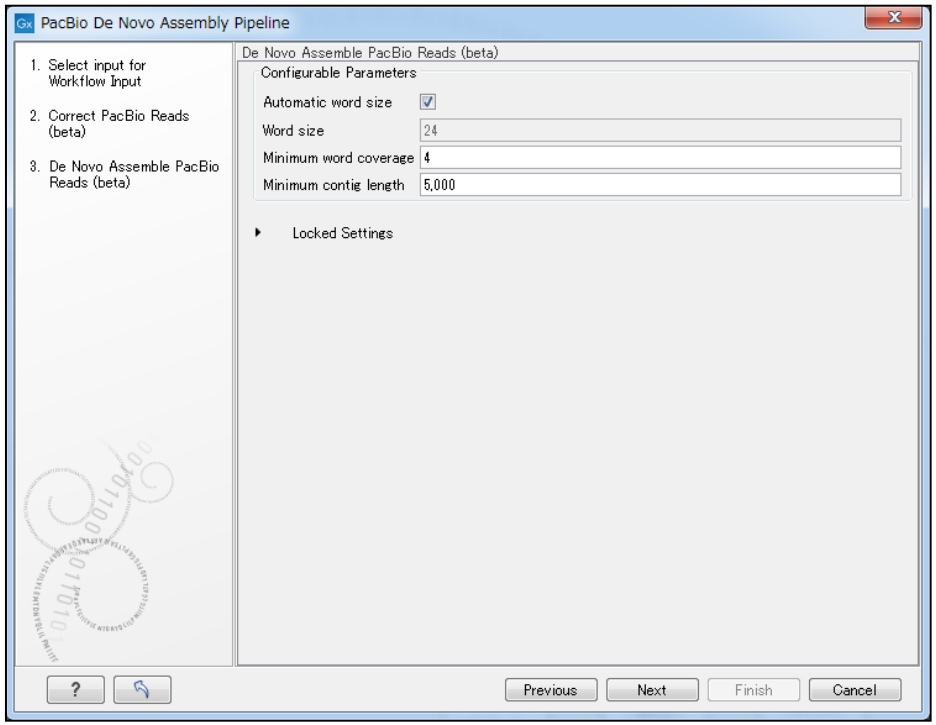
PacBio De Novo Assembly Pipeline



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

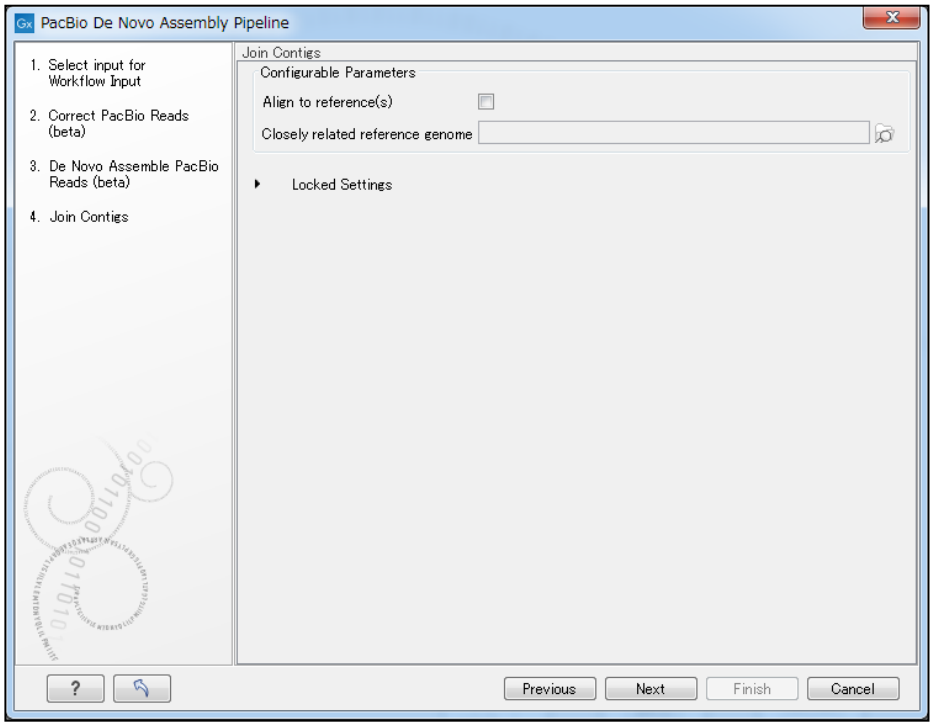


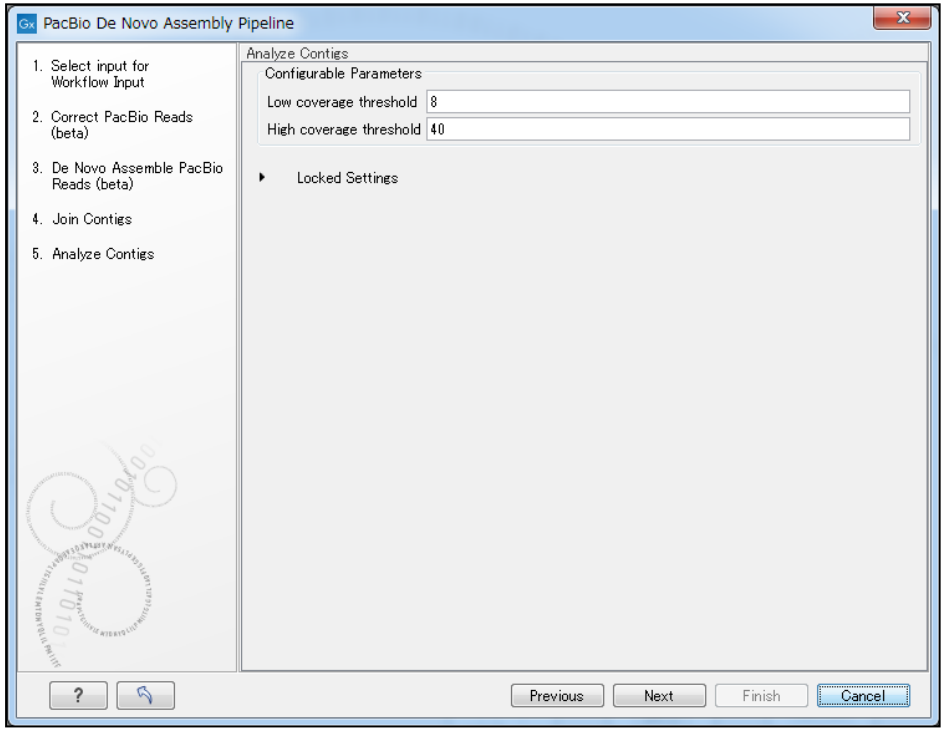


- Coverage percentage of reads to correct： 全リードのうち、配列長が大きい何%のリードを基準として、エラー補正を行うかを指定する。

- 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。

PacBio De Novo Assembly Pipeline

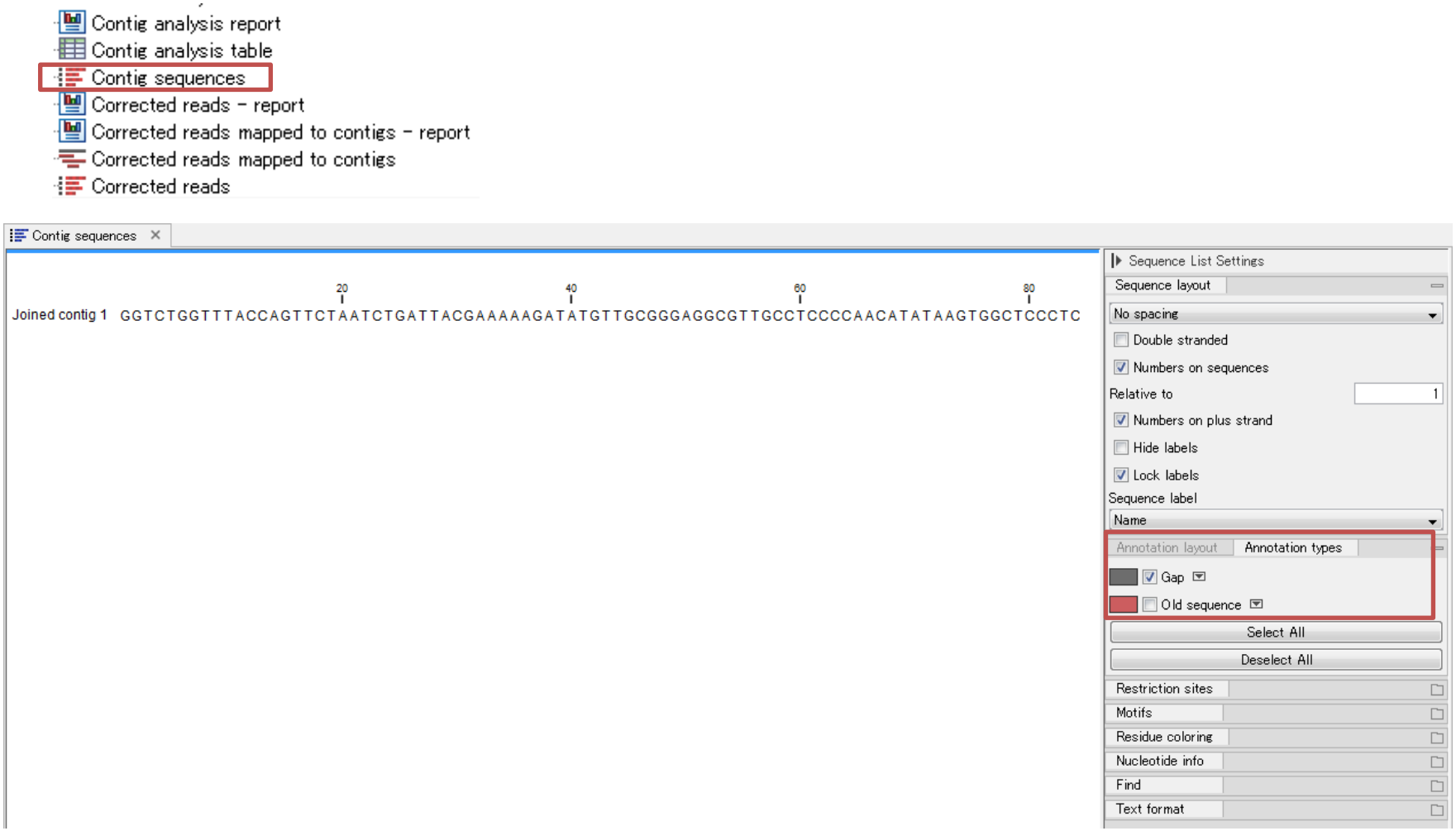




- Align to reference(s): 指定されたリファレンス配列データを使い、コンティグ同士を結合させる。
- Closely related reference genome：リファレンス配列データを指定する。

- 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。

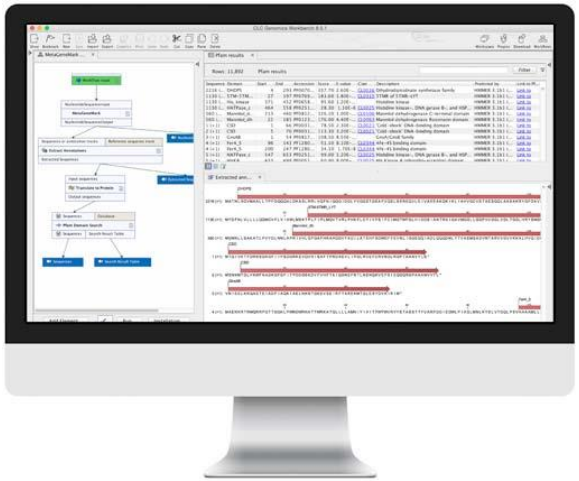
PacBio De Novo Assembly Pipeline



- 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。

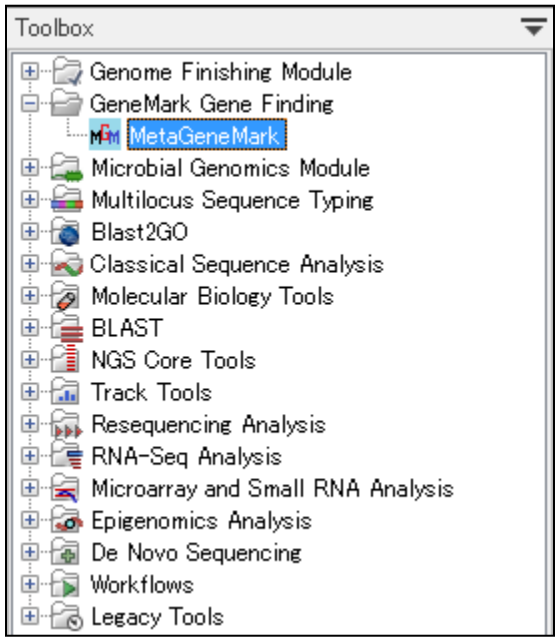
手順2：遺伝子領域の予測

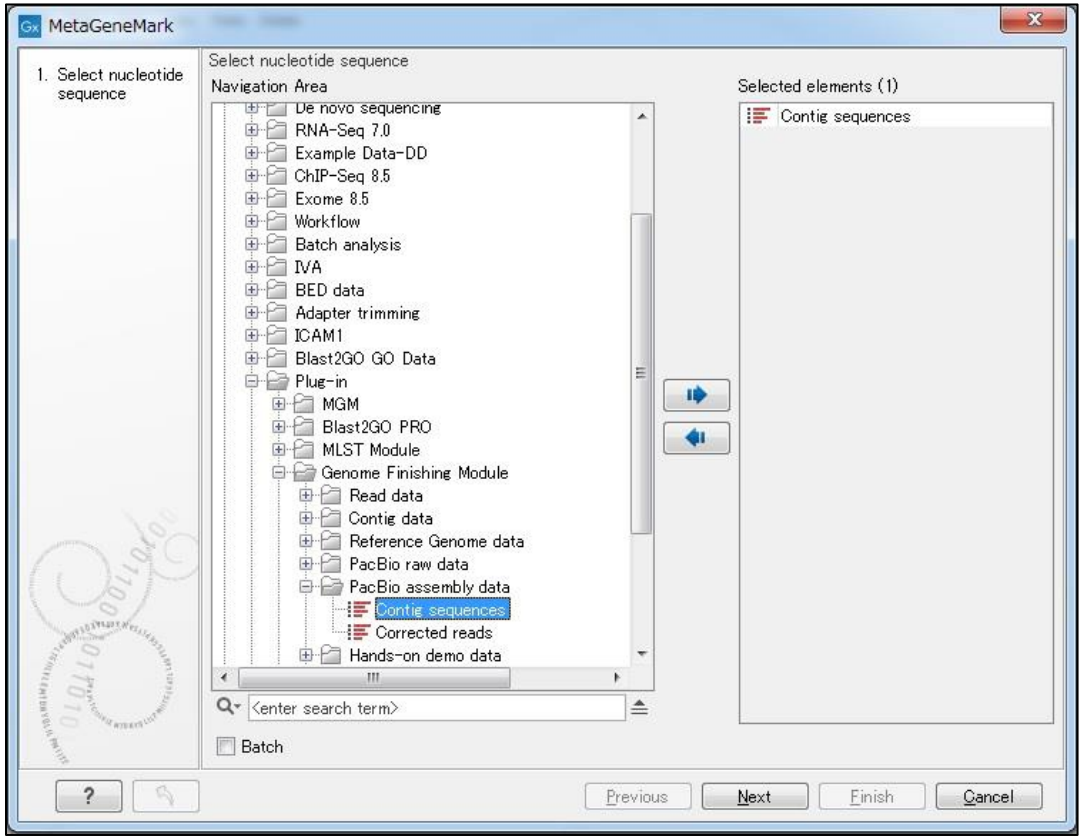
MetaGeneMark



- 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。
- 外部データリ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。

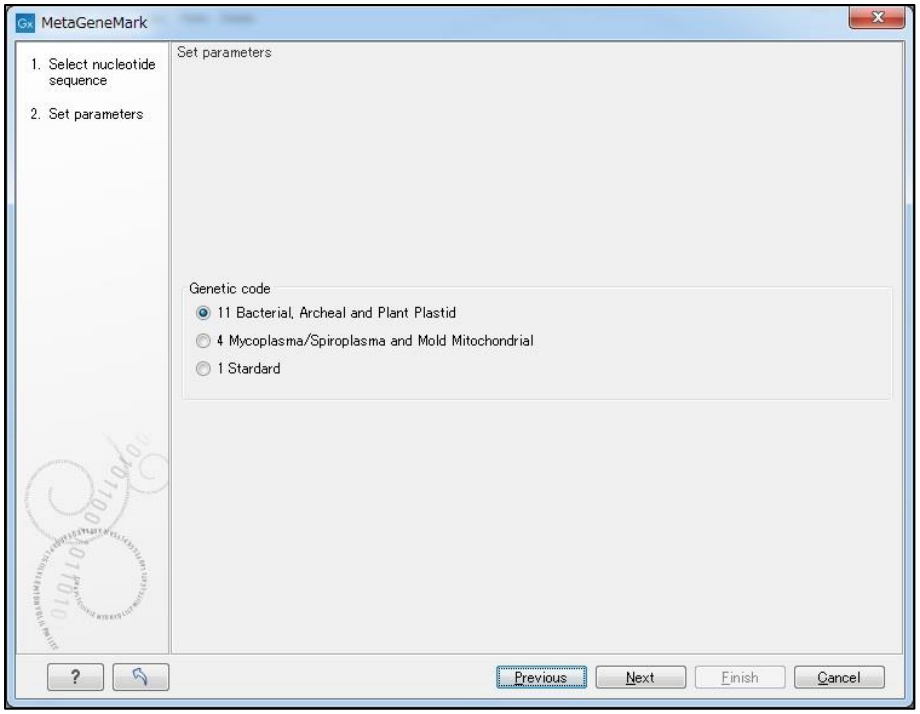
MetaGeneMark

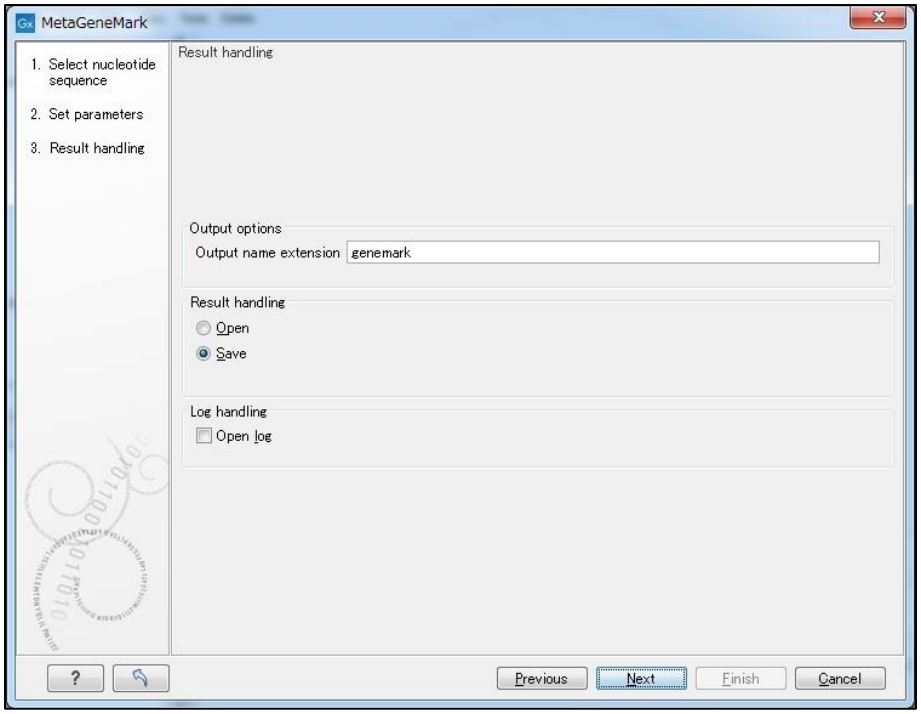




1. MetaGeneMarkを選択してダブルクリック。
2. コンティグ配列データを選択。

MetaGeneMark

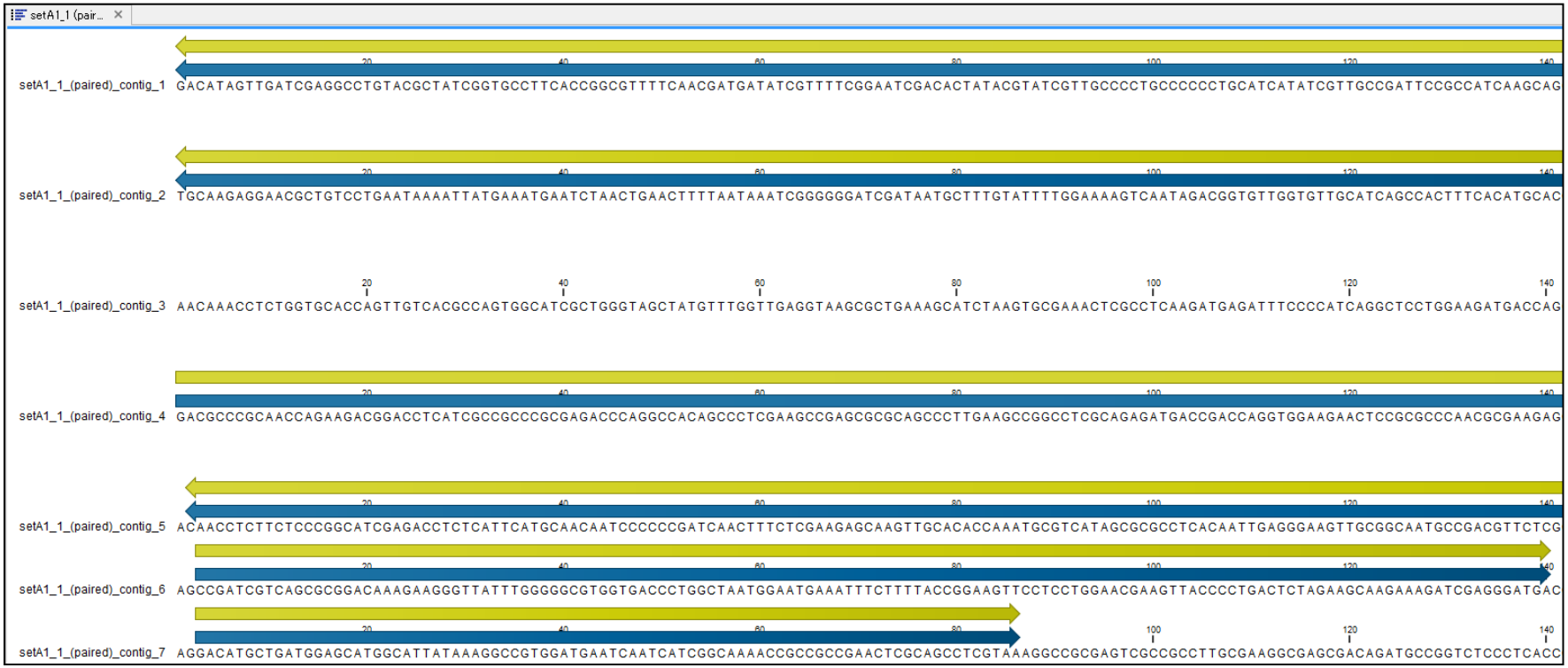




Genetic code:
- 解析データの生物種に応じて、CDS領域の塩基配列をアミノ酸に翻訳する際の翻訳テーブルを指定する。

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。

MetaGeneMark



- 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

遺伝子機能情報の付加

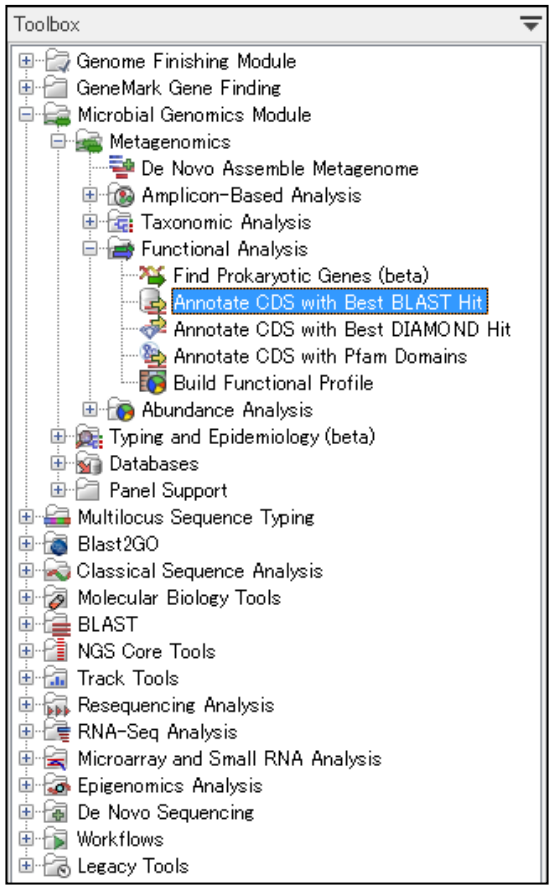
- MetaGeneMarkによって検出された各CDSデータに対して、配列の相同性比較や様々なデータソースから、遺伝子機能情報を付加する。

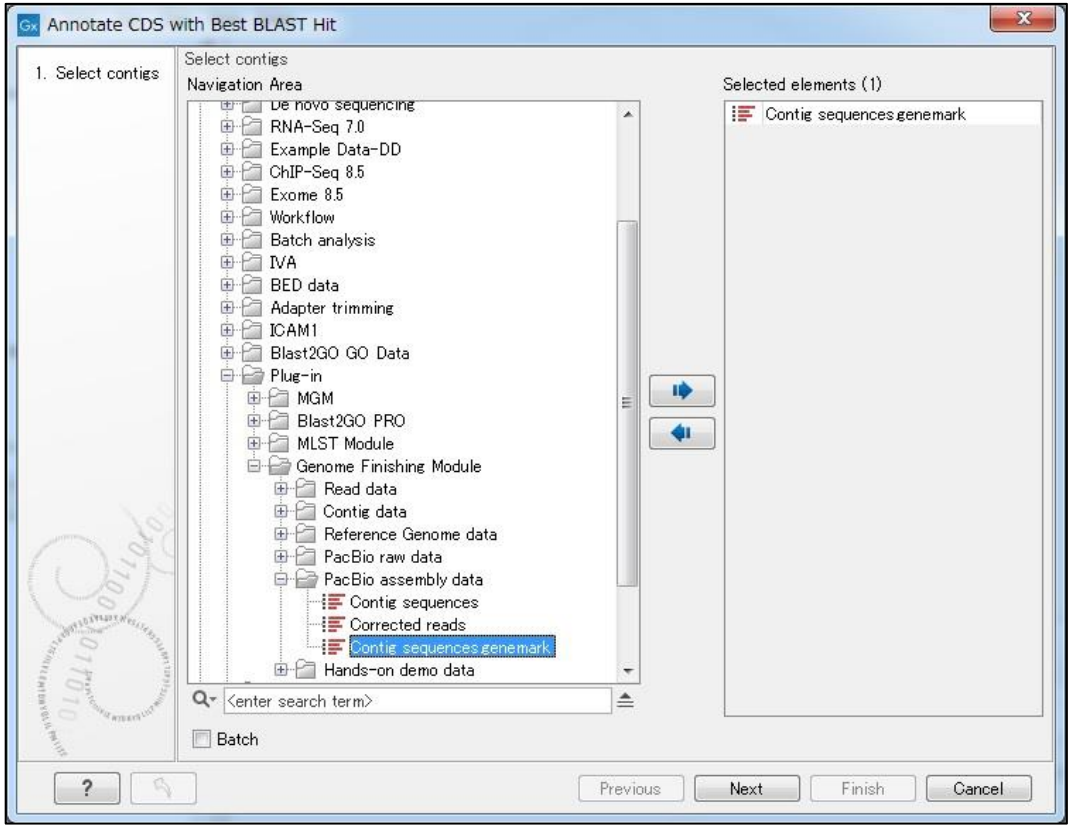
1. BLASTを利用した、配列データのターゲットデータベースとの相同性比較によるアノテーション付け
   （使用ツール：Annotate CDS with Best BLAST Hit）

2. DIAMOND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 DIAMOND Hit）

3. 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）

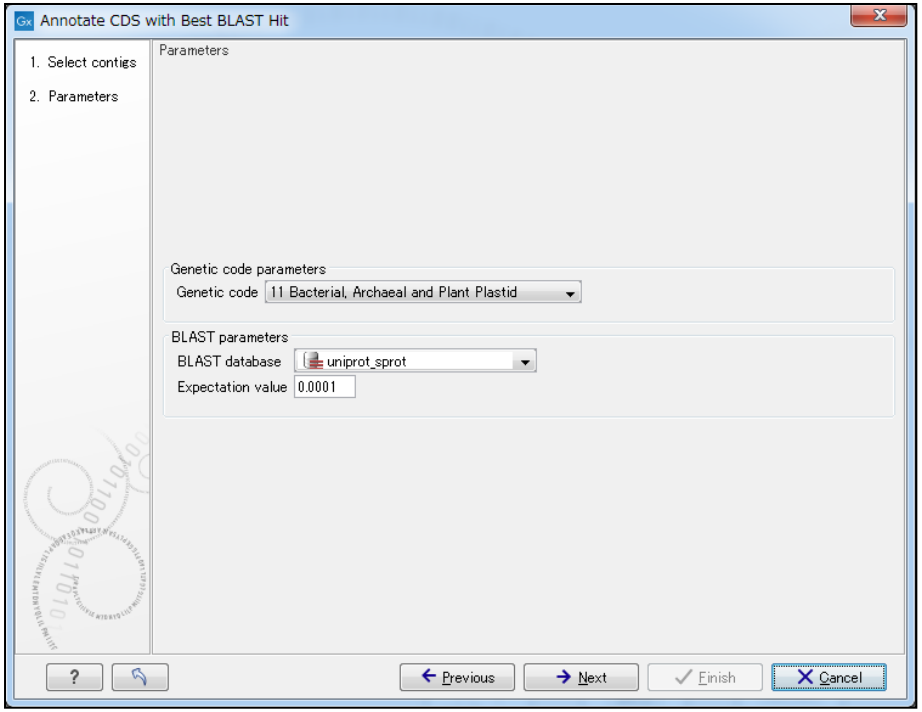
Annotate CDS with Best BLAST Hit

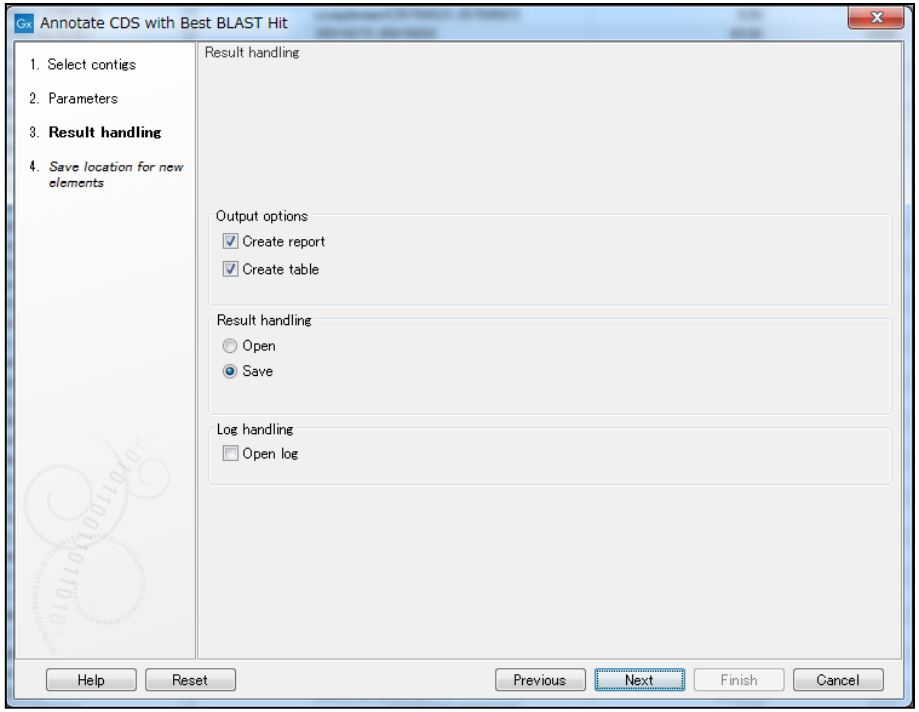




1. Annotate CDS with Best BLAST Hitを選択してダブルクリック。
2. CDSアノテーション付けされたコンティグ配列データを選択。

Annotate CDS with Best BLAST Hit



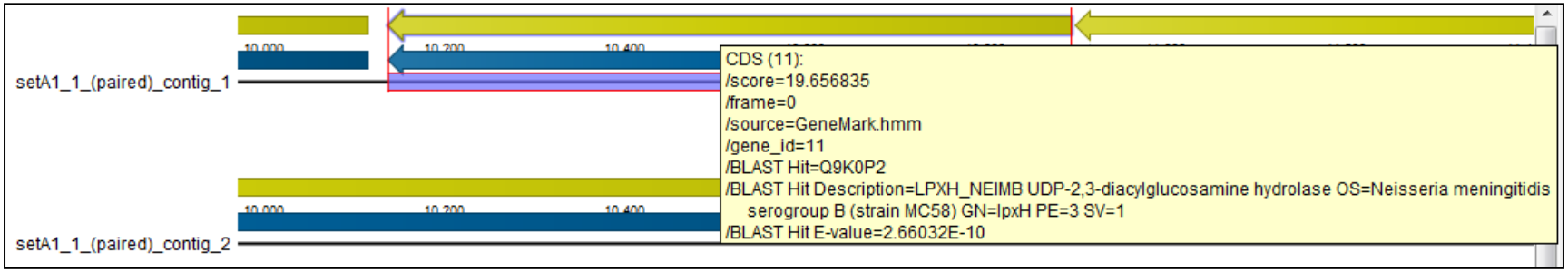


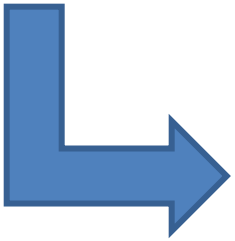
Genetic code parameters:
- Genetic code: 解析データの生物種に応じて、CDS領域の塩基配列をアミノ酸に翻訳する際の翻訳テーブルを指定する。

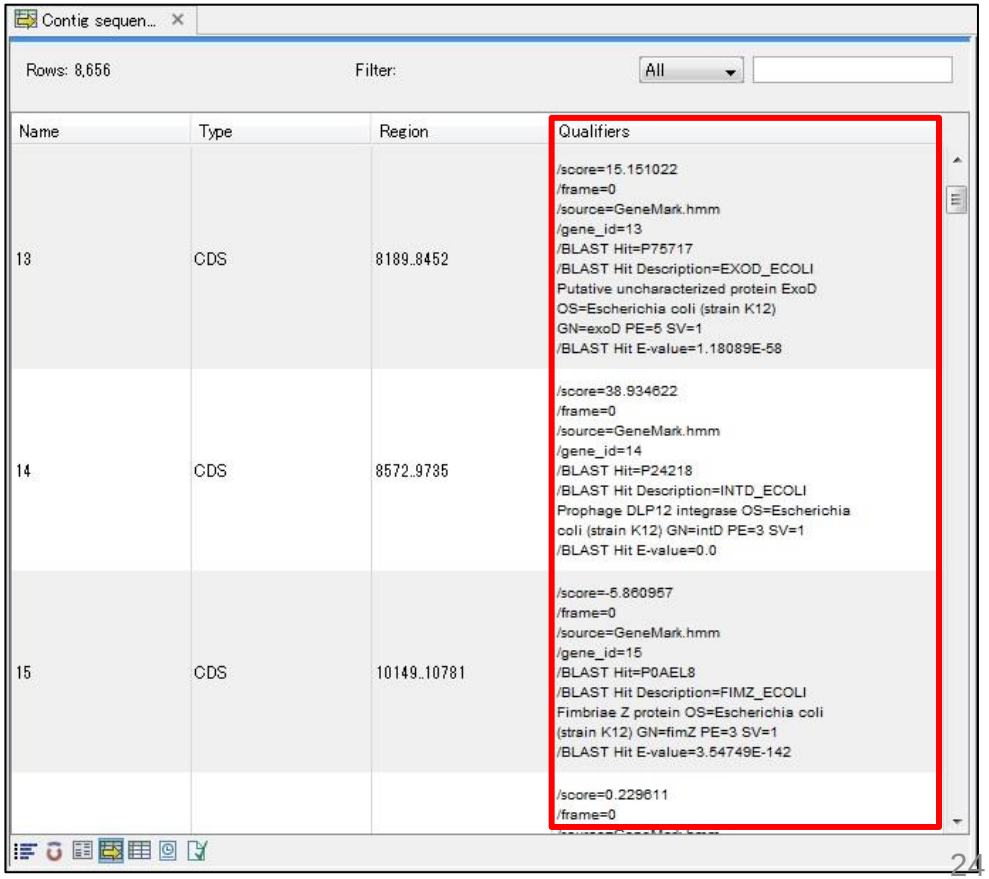
BLAST parameters:
- BLAST database: ターゲットデータベースを指定する。
（ターゲットデータベースの取得は、BLASTアプリケーションの「Download BLAST Databases」や「Create BLAST Database」から可能。）
- Expectation value: BLAST検索のE-valueの閾値。

Output options:
- Create report: 解析結果をまとめたレポートを作成する。
- Create table: 解析結果をまとめたテーブルを作成する。

Annotate CDS with Best BLAST Hit







- 遺伝子機能情報がCDSアノテーションに付加される。

Annotate CDS with Best BLAST Hit

Filgen® biosciences & nanosciences

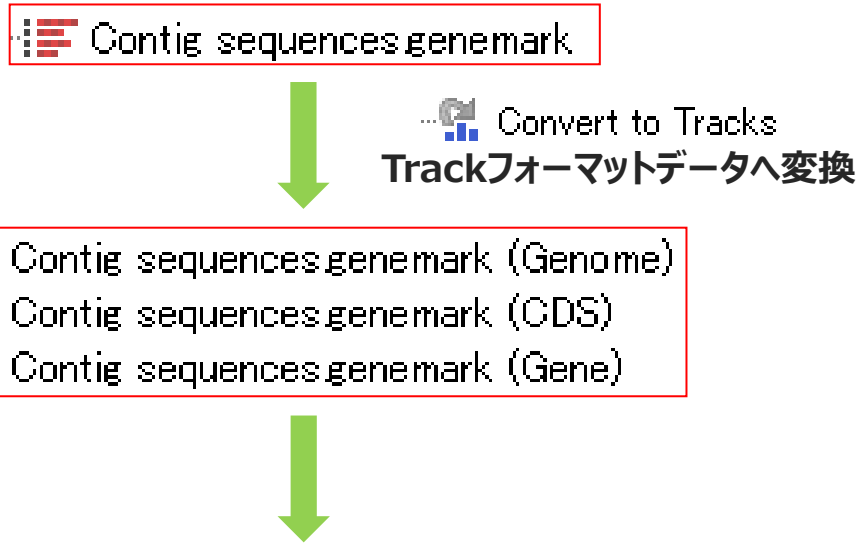
Rows: 4,246 Blast results Filter

Sequence	Accession	Query start	Query end	Hit start	Hit end	Score	E-value	Description
Joined contig 1 CDS 1071	P0AFB8	1	469	1	469	2,300.00	0.00	NTRC_ECOLI DNA-binding transcriptional regulator NtrC OS=Escherichia coli (strain K12) OX=83333 GN=glnG PE=1 SV=1
Joined contig 1 CDS 544	P0AE94	1	157	1	157	827.00	2.70E-114	CREA_SHIFL Protein CreA OS=Shigella flexneri OX=623 GN=creA PE=3 SV=1
Joined contig 1 CDS 1	Q47274	1	127	1	127	671.00	1.95E-91	REQ1_ECOLI Prophage antitermination protein Q homolog QuuD OS=Escherichia coli (strain K12) OX=83333 GN=quuD PE=3 SV=1
Joined contig 1 CDS 2	Q47272	1	46	1	46	239.00	5.57E-28	YLCG_ECOLI Uncharacterized protein YlcG OS=Escherichia coli (strain K12) OX=83333 GN=ylcG PE=1 SV=1
Joined contig 1 CDS 3	Q1RCY4	1	120	1	120	639.00	9.41E-87	RUSA_ECOUT Crossover junction endodeoxyribonuclease RusA OS=Escherichia coli (strain UTI89 / UPEC) OX=364106 GN=rusA P
Joined contig 1 CDS 4	P68661	1	96	1	96	504.00	5.88E-67	YBCO_ECOLI Putative nuclease YbcO OS=Escherichia coli (strain K12) OX=83333 GN=ybcO PE=1 SV=1
Joined contig 1 CDS 5	Q47269	1	151	1	151	816.00	1.02E-112	YBCN_ECOLI Uncharacterized protein YbcN OS=Escherichia coli (strain K12) OX=83333 GN=ybcN PE=4 SV=1
Joined contig 1 CDS 6	P77634	1	265	1	265	1,421.00	0.00	YBCM_ECOLI Uncharacterized HTH-type transcriptional regulator YbcM OS=Escherichia coli (strain K12) OX=83333 GN=ybcM PE=
Joined contig 1 CDS 7	P77368	1	183	1	183	966.00	1.36E-134	YBCL_ECOLI UPF0098 protein YbcL OS=Escherichia coli (strain K12) OX=83333 GN=ybcL PE=1 SV=1
Joined contig 1 CDS 8	P77698	1	508	1	508	2,704.00	0.00	YBCK_ECOLI Uncharacterized protein YbcK OS=Escherichia coli (strain K12) OX=83333 GN=ybcK PE=4 SV=1
Joined contig 1 CDS 9	P23895	1	110	1	110	555.00	2.97E-74	EMRE_ECOLI Multidrug transporter EmrE OS=Escherichia coli (strain K12) OX=83333 GN=emrE PE=1 SV=1
Joined contig 1 CDS 10	P0CF85	1	288	1	288	1,558.00	0.00	INSF_ECO11 Transposase InsF for insertion sequence IS3 OS=Escherichia coli O111:H- OX=168927 GN=insF PE=3 SV=1
Joined contig 1 CDS 11	P59445	1	99	1	99	498.00	6.73E-66	INSE_SHIFL Transposase InsE for insertion sequence IS3 OS=Shigella flexneri OX=623 GN=insE1 PE=3 SV=2
Joined contig 1 CDS 12	P77528	1	70	9	78	353.00	1.47E-44	PEAD_ECOLI Putative protein PeaD OS=Escherichia coli (strain K12) OX=83333 GN=peaD PE=5 SV=2
Joined contig 1 CDS 13	P75717	1	87	1	87	457.00	3.95E-60	EXOD_ECOLI Putative uncharacterized protein ExoD OS=Escherichia coli (strain K12) OX=83333 GN=exoD PE=5 SV=1
Joined contig 1 CDS 14	P24218	1	387	1	387	1,982.00	0.00	INTD_ECOLI Prophage integrase IntD OS=Escherichia coli (strain K12) OX=83333 GN=intD PE=3 SV=1
Joined contig 1 CDS 15	P0AEL8	1	210	1	210	1,034.00	5.36E-144	FIMZ_ECOLI Fimbriae Z protein OS=Escherichia coli (strain K12) OX=83333 GN=fimZ PE=3 SV=1
Joined contig 1 CDS 16	P38052	1	141	31	171	727.00	3.91E-99	SFMF_ECOLI Uncharacterized fimbrial-like protein SfmF OS=Escherichia coli (strain K12) OX=83333 GN=sfmF PE=2 SV=2
Joined contig 1 CDS 17	P75715	1	325	3	327	1,729.00	0.00	SFMH_ECOLI Uncharacterized fimbrial-like protein SfmH OS=Escherichia coli (strain K12) OX=83333 GN=sfmH PE=2 SV=2
Joined contig 1 CDS 18	P77468	1	867	1	867	4,532.00	0.00	SFMD_ECOLI Outer membrane usher protein SfmD OS=Escherichia coli (strain K12) OX=83333 GN=sfmD PE=2 SV=1
Joined contig 1 CDS 19	P77249	1	230	1	230	1,104.00	4.05E-154	SFMC_ECOLI Probable fimbrial chaperone SfmC OS=Escherichia coli (strain K12) OX=83333 GN=sfmC PE=2 SV=1
Joined contig 1 CDS 20	P0ABW5	1	180	1	180	722.00	1.43E-97	SFMA_ECOLI Uncharacterized fimbrial-like protein SfmA OS=Escherichia coli (strain K12) OX=83333 GN=sfmA PE=2 SV=1
Joined contig 1 CDS 21	P24186	1	288	1	288	1,510.00	0.00	FOLD_ECOLI Bifunctional protein FolD OS=Escherichia coli (strain K12) OX=83333 GN=folD PE=1 SV=4
Joined contig 1 CDS 22	P0AAS8	1	70	1	70	357.00	2.83E-45	YBCJ_SHIFL Uncharacterized protein YbcJ OS=Shigella flexneri OX=623 GN=ybcJ PE=3 SV=1
Joined contig 1 CDS 23	P45570	1	153	1	153	692.00	3.17E-93	YBCI_ECOLI Inner membrane protein YbcI OS=Escherichia coli (strain K12) OX=83333 GN=ybcI PE=1 SV=1
Joined contig 1 CDS 24	P21888	1	461	1	461	2,310.00	0.00	SYC_ECOLI Cysteine--tRNA ligase OS=Escherichia coli (strain K12) OX=83333 GN=cysS PE=1 SV=2
Joined contig 1 CDS 25	P23869	1	164	1	164	876.00	1.95E-121	PPIB_ECOLI Peptidyl-prolyl cis-trans isomerase B OS=Escherichia coli (strain K12) OX=83333 GN=ppiB PE=1 SV=2
Joined contig 1 CDS 26	P43341	1	240	1	240	1,275.00	7.84E-180	LPXH_ECOLI UDP-2,3-diacylglucosamine hydrolase OS=Escherichia coli (strain K12) OX=83333 GN=lpxH PE=1 SV=2
Joined contig 1 CDS 27	P0AG18	1	169	1	169	763.00	3.72E-104	PURE_ECOLI N5-carboxyaminoimidazole ribonucleotide mutase OS=Escherichia coli (strain K12) OX=83333 GN=purE PE=1 SV=2
Joined contig 1 CDS 28	P09029	1	355	1	355	1,808.00	0.00	PURK_ECOLI N5-carboxyaminoimidazole ribonucleotide synthase OS=Escherichia coli (strain K12) OX=83333 GN=purK PE=1 SV=2
Joined contig 1 CDS 29	P37306	1	297	1	297	1,473.00	0.00	ARCC_ECOLI Carbamate kinase OS=Escherichia coli (strain K12) OX=83333 GN=arcC PE=3 SV=2
Joined contig 1 CDS 30	P0AAS5	1	256	1	256	1,262.00	9.77E-177	YLBF_ECOLI Uncharacterized protein YlbF OS=Escherichia coli (strain K12) OX=83333 GN=ylbF PE=4 SV=1
Joined contig 1 CDS 31	P77129	1	419	1	419	2,150.00	0.00	YLBE_ECOLI Uncharacterized protein YlbE OS=Escherichia coli (strain K12) OX=83333 GN=ylbE PE=4 SV=1
Joined contig 1 CDS 32	Q47208	1	555	1	555	2,705.00	0.00	FDRA_ECOLI Protein FdrA OS=Escherichia coli (strain K12) OX=83333 GN=fdrA PE=1 SV=1
Joined contig 1 CDS 33	P77555	1	349	1	349	1,876.00	0.00	ALLD_ECOLI Ureidoglycolate dehydrogenase (NAD(+)) OS=Escherichia coli (strain K12) OX=83333 GN=allD PE=1 SV=1
Joined contig 1 CDS 34	P77425	1	411	1	411	2,226.00	0.00	ALLC_ECOLI Allantoate amidohydrolase OS=Escherichia coli (strain K12) OX=83333 GN=allC PE=1 SV=1
Joined contig 1 CDS 35	P75713	1	261	1	261	1,304.00	0.00	ALLE_ECOLI (S)-ureidoglycine aminohydrolase OS=Escherichia coli (strain K12) OX=83333 GN=allE PE=1 SV=1

- テーブル形式のデータも取得が可能。

リファレンスゲノムデータの作成

リファレンスゲノムデータの作成



- 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	3738:c.293_294insG	3738:p.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	3740:c.289C>G	3740:p.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 (9:00～18：00)

FAX　052-624-4389

E-mail: biosupport@filgen.jp