真核生物ゲノムのde novo アセンブリ

フィルジェン株式会社
バイオインフォマティクス部(biosupport@filgen.jp)

2024.01

はじめに

- 比較的ゲノムサイズの大きい真核生物のde novo アセンブリは、 高スペックPCやサーバーの導入が必要
- コマンドライン型のツールかつパラメータ設定が重要なため操作が煩雑

OmicsBox

OmicsBoxのDNA-seq de novo アセンブリ機能

- メーカーのサーバーで高速計算　高価なPCの購入は不要
- マウス操作で簡単に解析
- 真核生物に適切なアルゴリズムと高品質なアノテーションを付与

必要なファイル

Fastq

シーケンサーから出力された生データ
もしくは受託サービスで得られたクリーンリードデータ

解析ワークフロー

QC・トリミング

- NGSより出力された生データが良好か、下流分析に影響する問題がないか確認。
  de novo アセンブリに有用な高品質なリードデータが得られる。
  FastQCとTrimmomaticツールを統合

de novo アセンブリ

- 高品質なリードデータのみで（リファレンスゲノムなしで）新規にゲノム配列を構築する解析。
  構築された長い連続的な配列（コンティグ）に関するFASTAファイルが得られる。
  ABySS、SPAdes、Flyeツールを統合

Repeat Masking

- コンティグデータ中の反復配列を見つけ塩基を「N」、「X」、などに置き換える解析。
  この変換により、以下の下流分析のツールに、これらの領域が反復配列であることを認識させることができ、
  結果として、予測結果の精度を上げることができる。
  RepeatMaskerを統合

遺伝子構造予測

- 近縁種の遺伝子情報やRNA-Seqデータなどを使用して遺伝子構造を予測する。
  ORFなどの情報をもつアノテーションファイル（GFF3/GTF）を取得できる。
  AUGUSTUSを統合

遺伝子機能情報付与

- Blastや7000以上の研究引用実績のあるBlast2GOアルゴリズムにより
  高品質な遺伝子機能情報を付与することができる。

QC・トリミング

・データが良好か、下流分析に影響する問題がないか確認

Welcome Message | FASTQ Quality Check (Dataset) | FASTQ Quality Check (ERR1948631_1.fastq) | FASTQ Quality Check (clean_ERR1948631_1.fq) | *Chart: Adapter Conten

FASTQ Quality Check

Name: Dataset

Overall Results

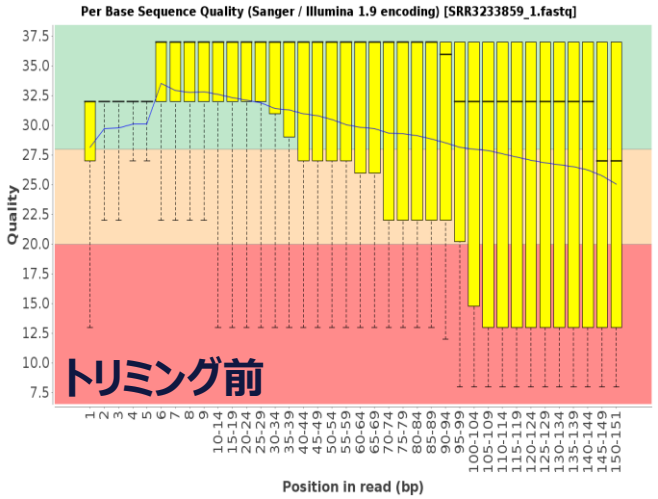
Name	Per Base Sequence Quality	Per Sequence Quality Scores	Per Base Sequence Content	Per Sequence GC Content	Per Base N Content
ERR1948631_1.fastq	PASS	PASS	FAIL	PASS	PASS
clean_ERR1948631_1.fq	PASS	PASS	FAIL	PASS	PASS

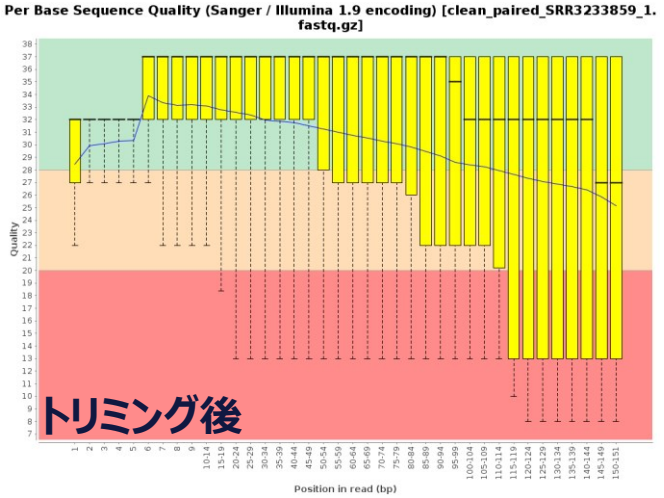
Name	Sequence Length Distribution	Adapter Content	Overrepresented Sequences	Sequence Duplication Levels	Report
ERR1948631_1.fastq	PASS	FAIL	WARNING	FAIL	
clean_ERR1948631_1.fq	WARNING	PASS	WARNING	FAIL	

The FASTQ quality check task is performed by nine analysis modules. The table above provides a quick evaluation of whether the results of each module seem entirely normal (pass), sightly abnormal (warning) or very unusual (fail). Note that these evaluations must be taken in the context of what is expected from the library. For example, some experiments may be expected to produce libraries which are biased in particular ways. Therefore, the summary evaluations should be treated as pointers that guide the preprocessing of the libraries.

✓ 解析が終了するとレポートが作成

正常（PASS）
わずかに異常（WARNING）
異常（FAIL）

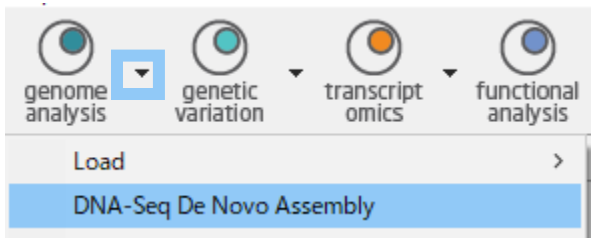




シーケンスデータの品質をすばやく評価

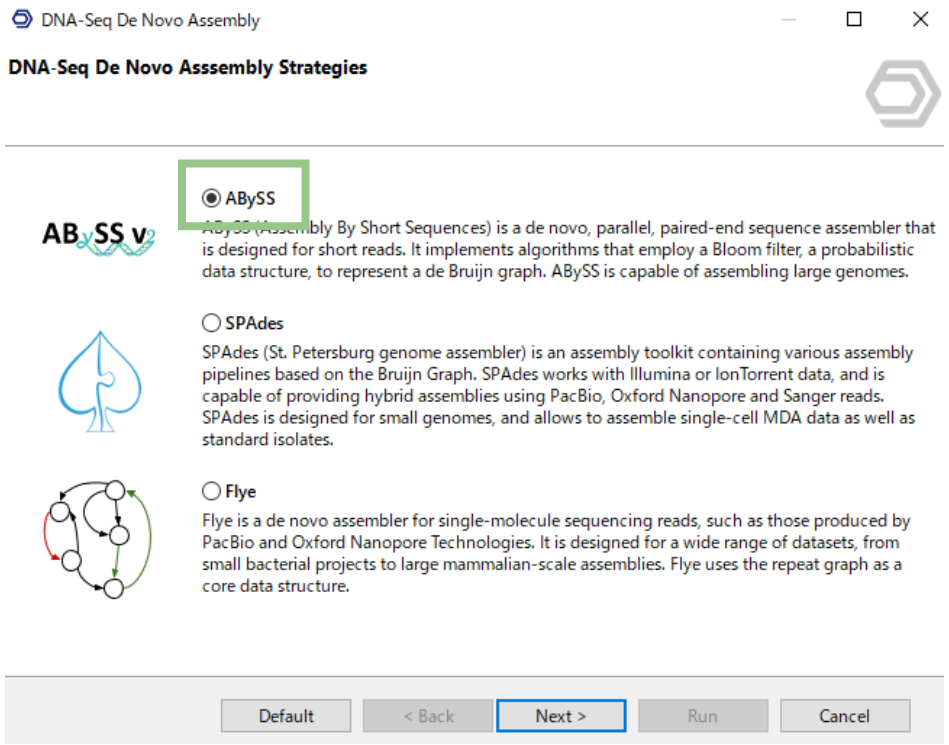
レポートのアイコンをクリック→さらに詳細な結果を見ることが可能

de novo アセンブリ

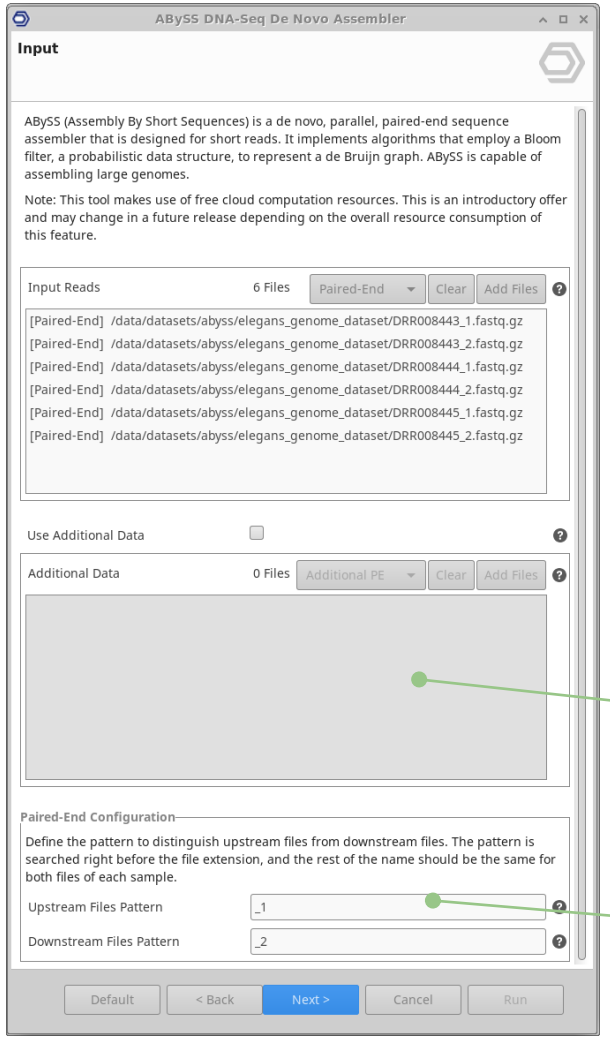


① de novo アセンブリツールをクリック

Genome Analysis Moduleを使用します。



② 使用するアセンブリアルゴリズムの中から「ABySS」を選択

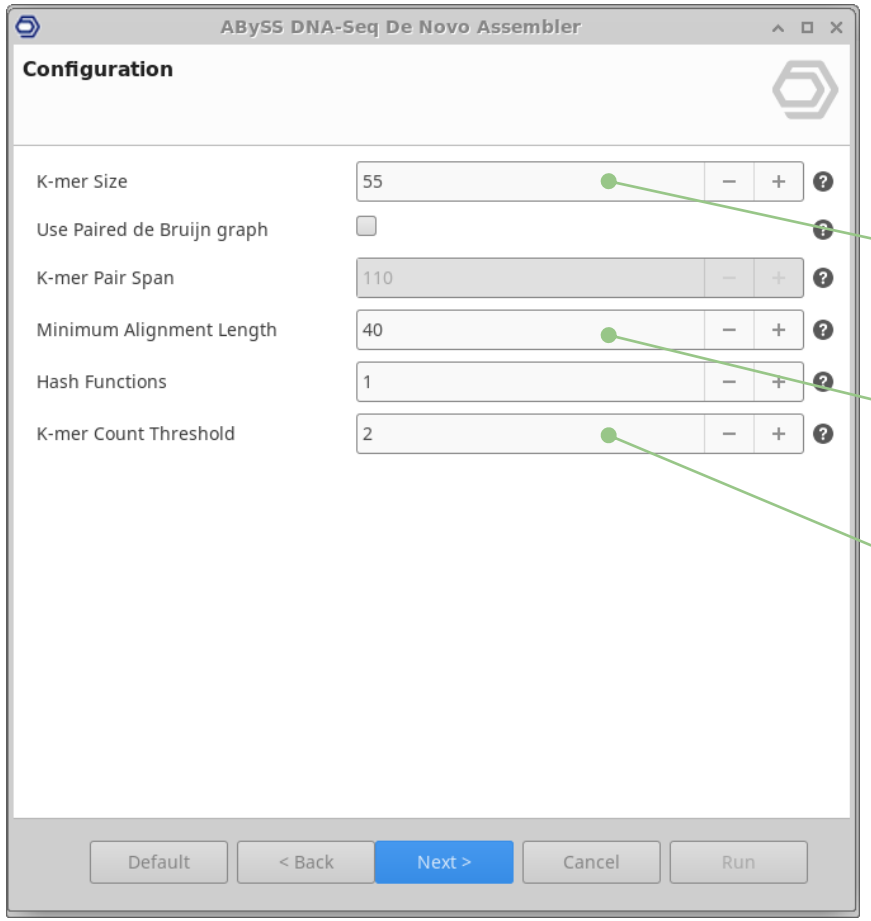


オプションでメイトペアリードがあればこちらに入力することで、コンティグの品質を改善できます。（無くても解析可能です。）

ペアエンドリードの場合リバースとフォワードを識別する部分の文字列を入力します。

③ QCを実行した高品質なリードデータを入力

de novo アセンブリ

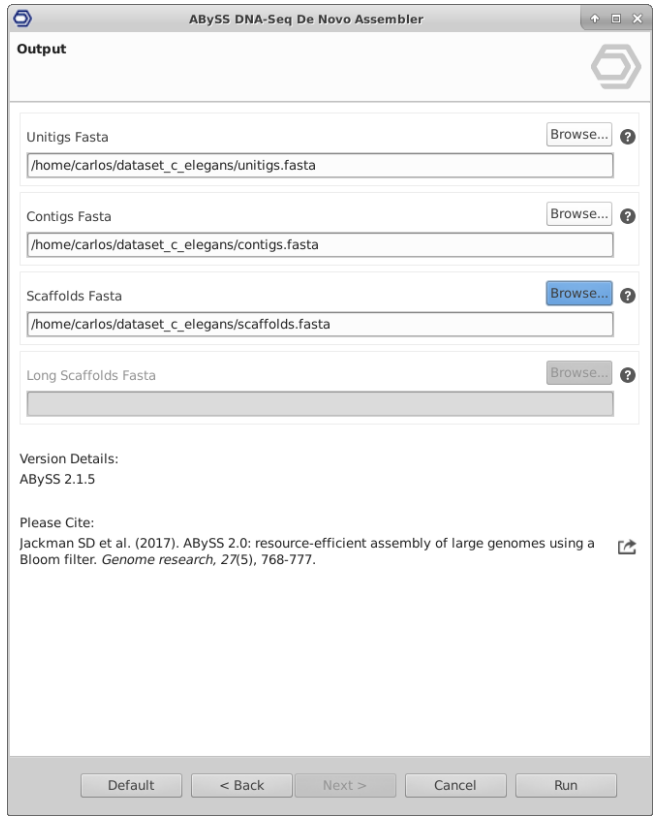


特に重要なパラメータ

k-merサイズの選択は、シーケンスアセンブリにさまざまな影響を与えます。リードの長さの少なくとも半分の奇数値を使用することをお勧めします。

リードの最小アライメント長を調整することもできます。

エラーや重要度の低いリードである可能性のある低カバレッジk-merをフィルタリングできます。カバレッジが非常に高いデータセットでは、この値を最大4まで設定できます。



③パラメータを設定
ゲノム構築のプロセスは複数の要因の影響を受けるため、
推奨される真のパラメーター設定はありません。
異なる値を試し、結果を確認して最適な値を選択することをお勧めします。

④ データの保存先を指定

de novo アセンブリ

結果のFASTAファイルに加えて、レポートとチャートが生成されます。
様々なパターンの設定を行い、設定ごとのレポートを比較します。
全ての項目について比較を行うことを推奨しますが、ここでは特に重要なものをご説明いたします。

偽陽性率を示しています。この数値を5%未満にすることを推奨します。

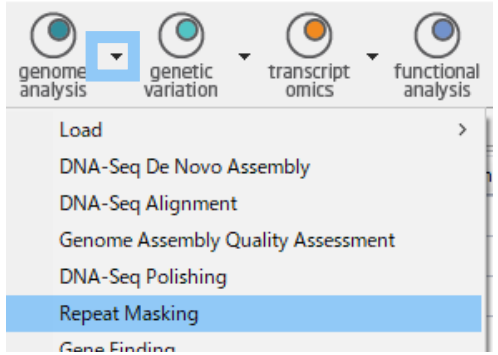
N50はアセンブルの結果の良し悪しを判断する指標です。
この値が高いほど、より良いアセンブリを示しています。

Results Overview

Bloom filter False Positive Rate (FPR): 0.056%.

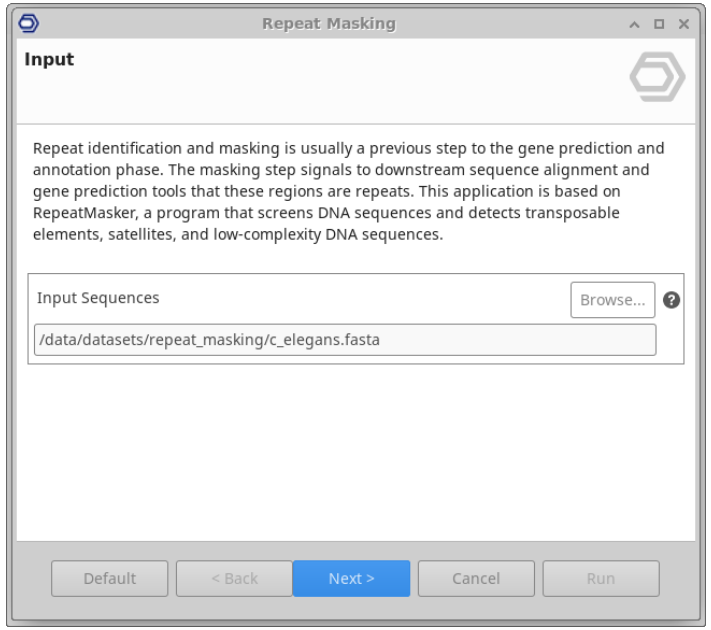
Statistic	assembly_unitigs	assembly_contigs	assembly_scaffolds
Number of Contigs (>= 0 bp)	9,036	6,704	6,328
Number of Contigs (>= 1000 bp)	1,801	494	266
Number of Contigs (>= 5000 bp)	1,360	430	202
Number of Contigs (>= 10000 bp)	1,041	398	191
Number of Contigs (>= 25000 bp)	581	328	170
Number of Contigs (>= 50000 bp)	253	245	152
Total Length (>= 0 bp)	42,852,845	42,903,496	42,890,544
Total Length (>= 1000 bp)	41,950,921	42,219,326	42,225,854
Total Length (>= 5000 bp)	40,791,399	42,054,369	42,060,897
Total Length (>= 10000 bp)	38,431,541	41,823,237	41,985,378
Total Length (>= 25000 bp)	30,767,430	40,645,628	41,639,105
Total Length (>= 50000 bp)	18,981,278	37,524,920	41,004,559
For Contigs >= 500			
Number of Contigs	1,977	544	315
Largest Contig	210,594	671,270	1,337,175
Total Length	42,079,243	42,255,468	42,261,476
GC (%)	42.01	42	42
N50	44,425	166,474	358,508
N75	23,565	98,900	204,792
L50	297	80	40
L75	614	164	78
Number of N's per 100 kbp	0	12.02	29.36

Repeat Masking

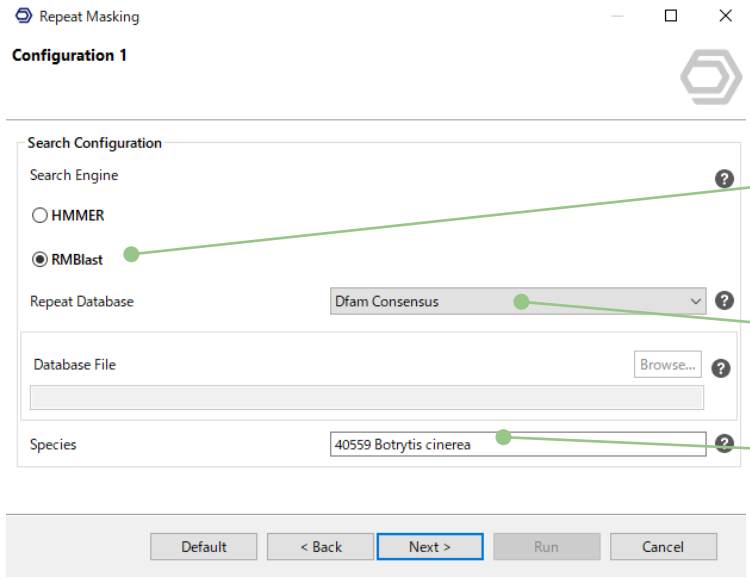


① Repeat Maskingツールをクリック

Genome Analysis Moduleを使用します。



② 前項で作成した**scaffolds.fasta**を入力します。

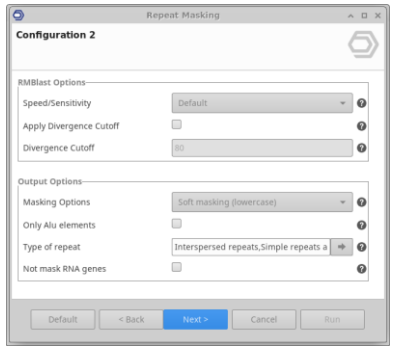


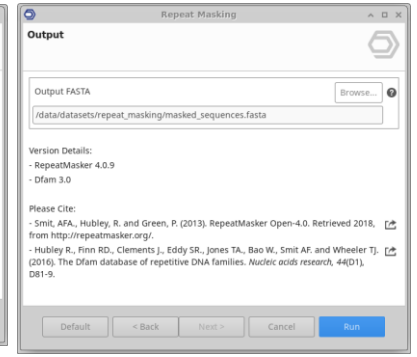
検索エンジンを選択します。
(セミナーではRMBlastを選択)

データベースを選択します。
(セミナーではDfamを選択)

種名を入力します。

③ 検索エンジン等の設定を行います。

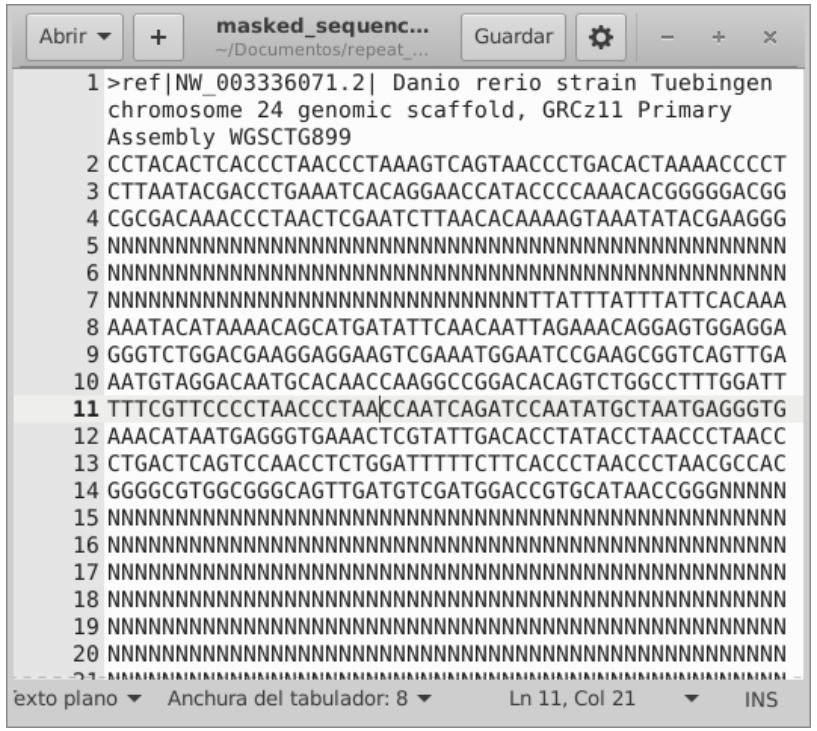




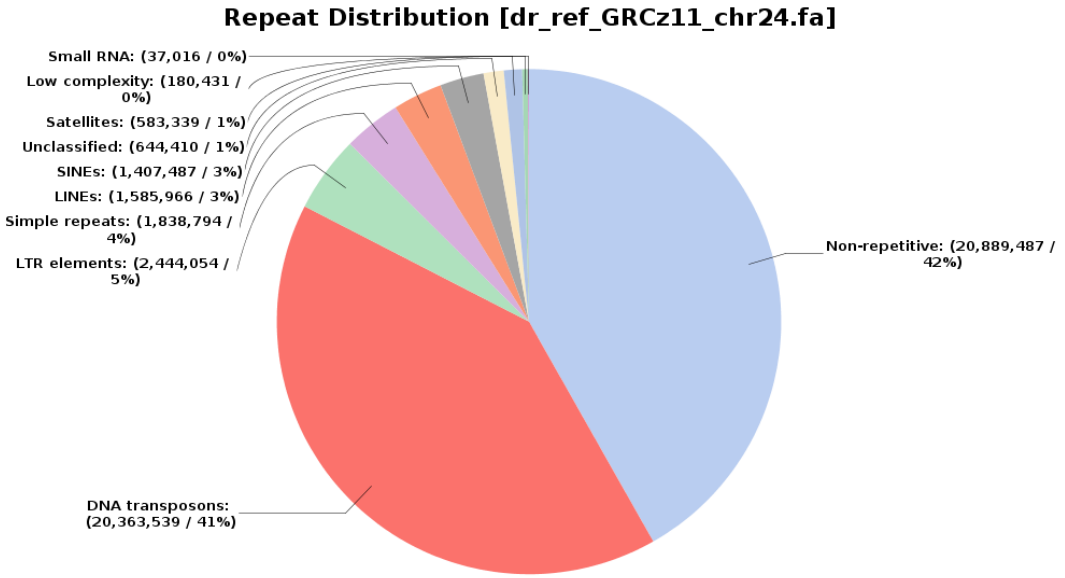
④ 任意で高度な設定を行い、最後に保存先を指定します。

Repeat Masking

・結果　Repeat Maskingプロセスは、
マスクされたシーケンス（FASTA形式）、検出されたリピートの位置（GFF形式）、レポートが保存されます。

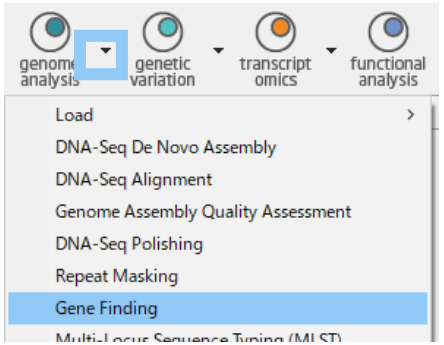


マスクされたシーケンス



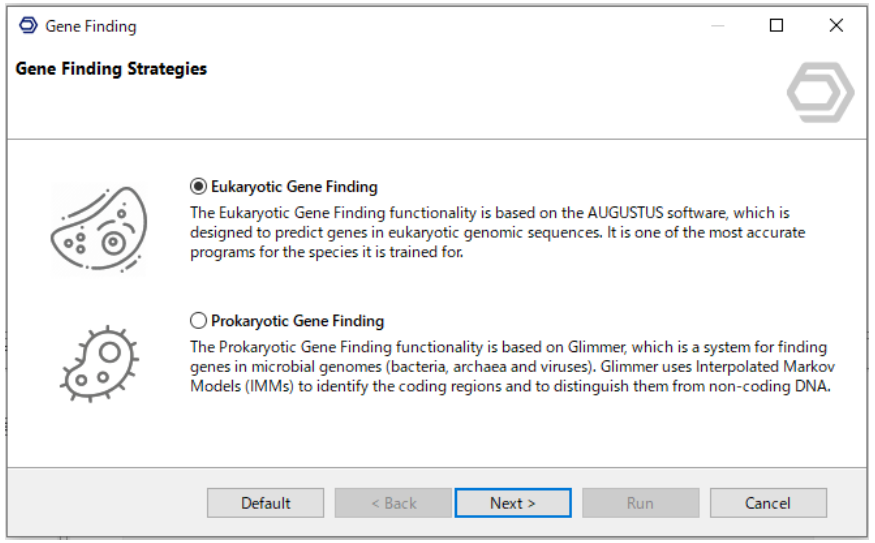
加えて、レポートとチャートが生成されます。
チャートは、各リピートクラスがカバーするシーケンスの割合を示しています。

遺伝子構造予測

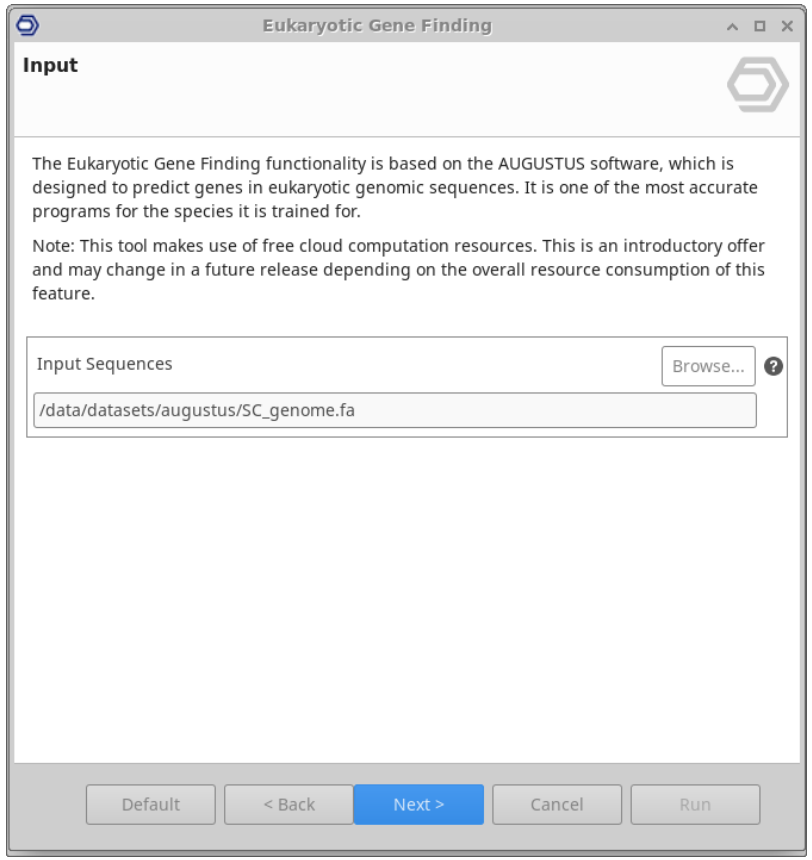


① Gene Findingツールをクリック

Genome Analysis Moduleを使用します。

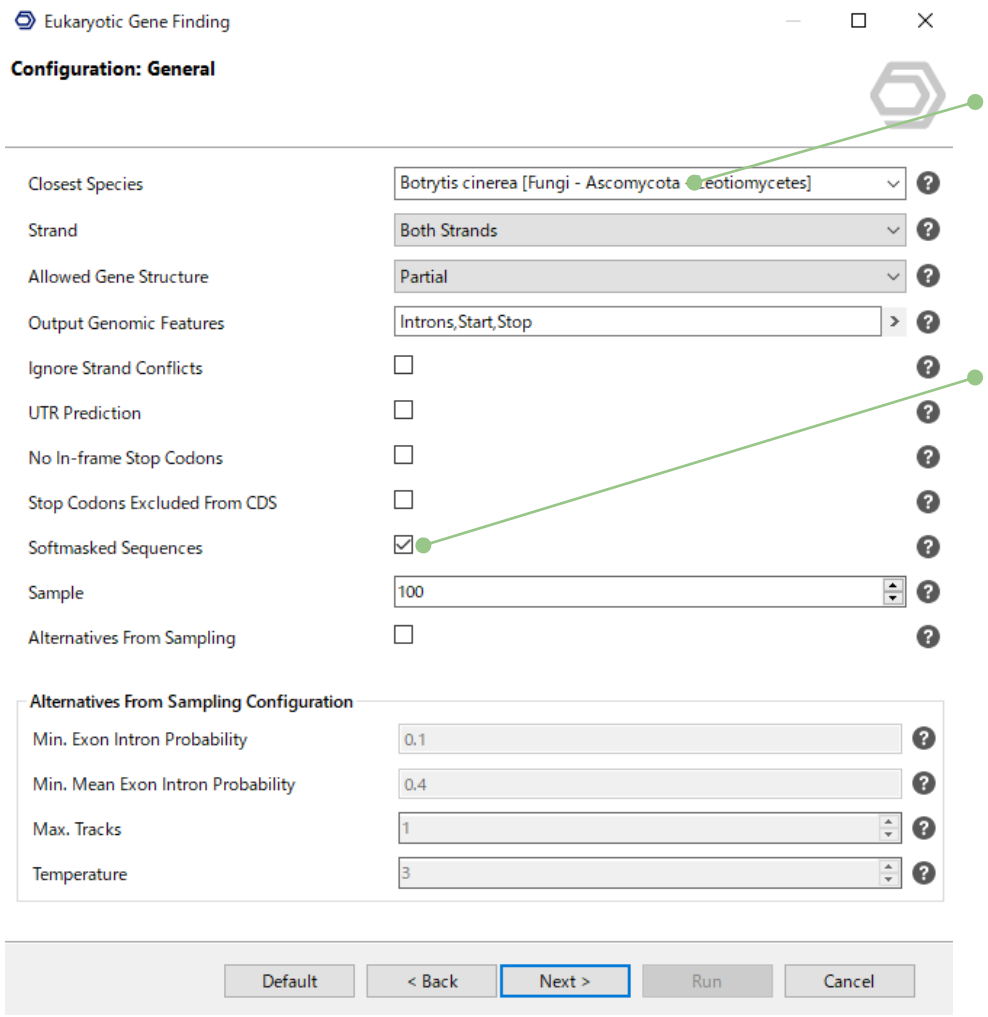


② Eukaryotic Gene Findingをクリック



③ Repeat Maskingで得られたFASTA形式のDNA入力シーケンスを入力します。

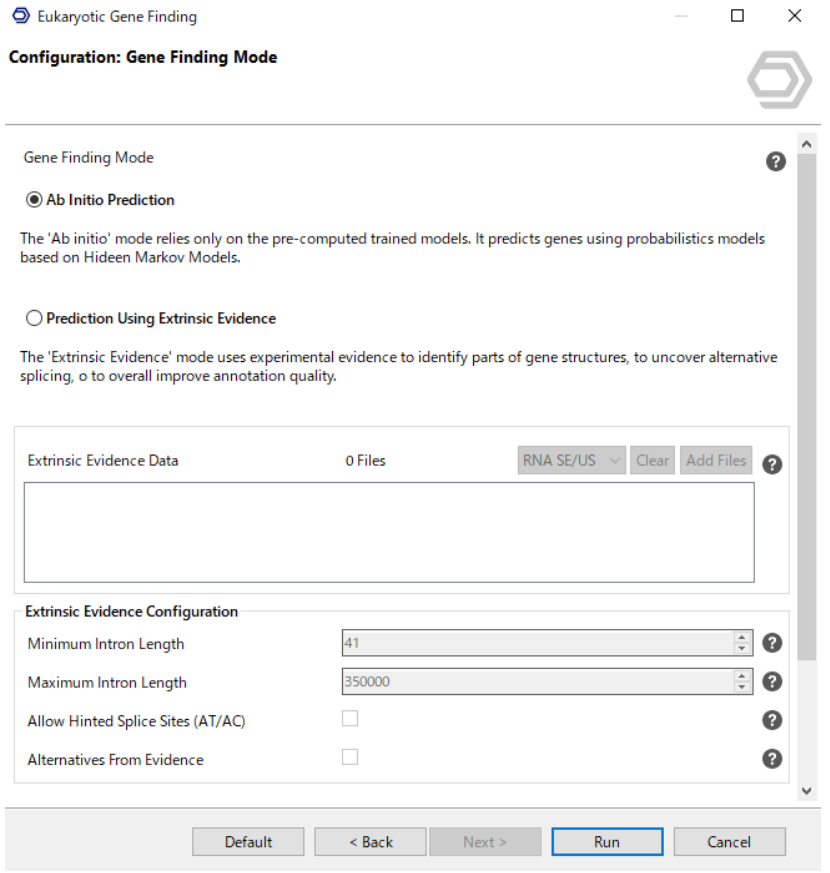
遺伝子構造予測



最も正確な予測を取得するために、
クエリに最も近い関連生物を選択します。

Repeat Masking処理後の
データを使用した場合は、
このオプションをマークします。

④パラメータを設定

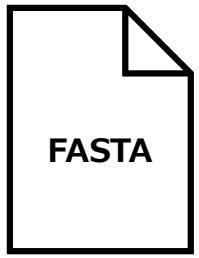


⑤「③」で選択したデータのみを使用する場合は、Ab initioモードを選択します。

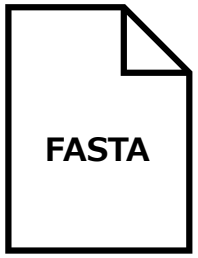
RNA-Seqデータやタンパク質、EST/cDNAデータがある場合は、Extrinsic Evidenceモードを選択することで、アノテーションの全体的な品質を向上させることができます。

遺伝子構造予測

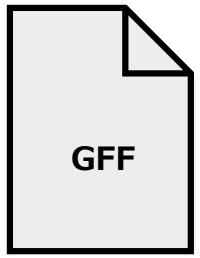
・結果



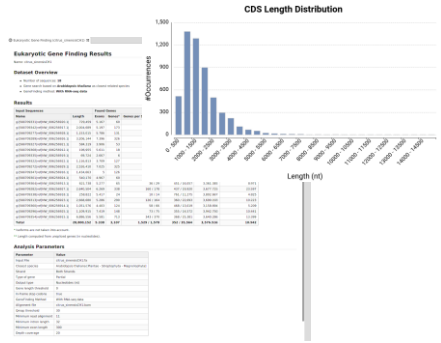
CDS配列



タンパク質配列

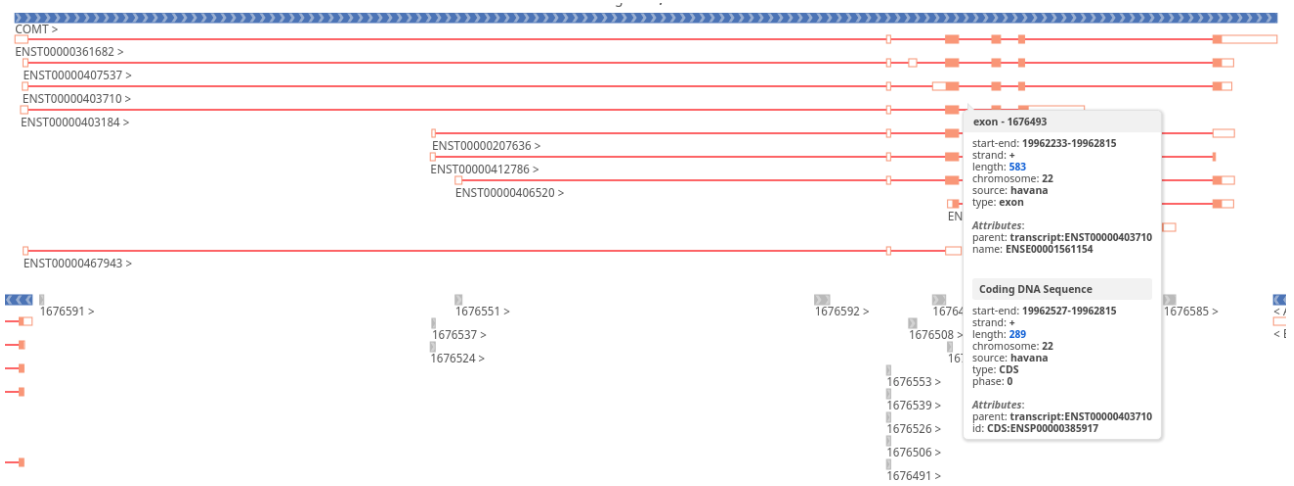


ゲノムアノテーション
ファイル



レポート・チャート

ゲノムアノテーションファイルは、
Genome Browserを使用して閲覧できます。



遺伝子機能情報付与

*Functional Analysis Moduleを使用します。

・BLASTを実行すると、遺伝子構造予測の結果のデータ(FASTAファイル)に
各配列のトップヒット配列の情報が表示されます。

・InterProScanを実行すると タンパク質のドメイン構造やモチーフなどの特徴を推定できます。

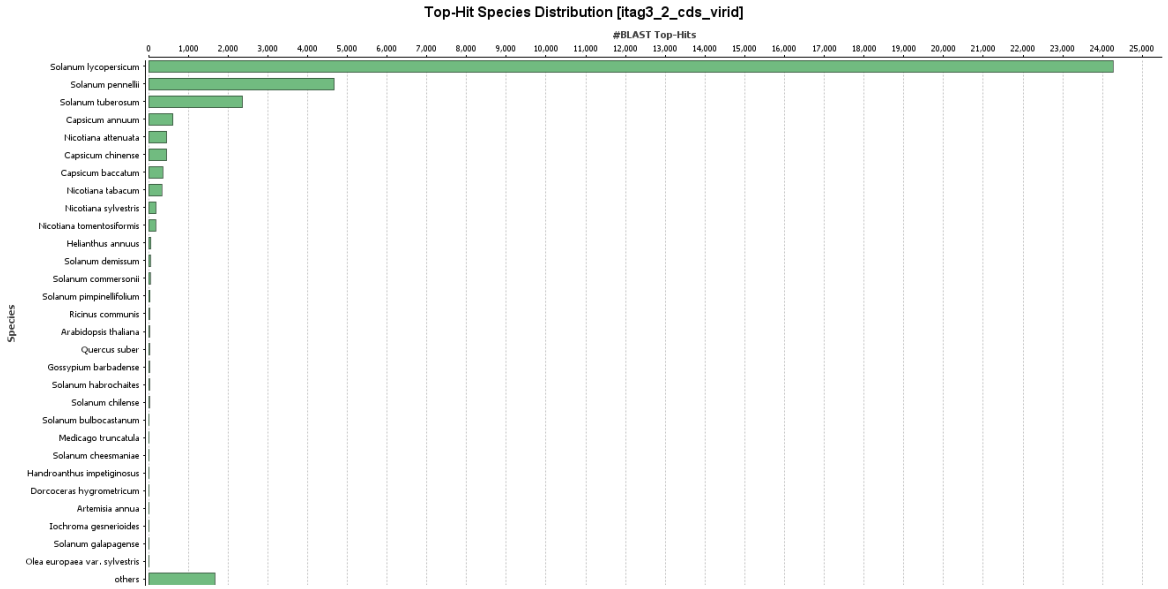
・メーカ独自のアルゴリズムBlast2GO方法論に基づき、7000件以上の研究引用の実績があります。

Nr	Tags	SeqName	Description	Length	#Hits	e-Value	sim mean	#GO	GO IDs	GO Names	Enzyme Codes	Enzyme Na...	InterPro IDs	InterPro GO IDs	InterPro GO Names
344	INTERPRO BLASTED MAPPED ANNOTATED	g344.t1	guanine deaminase protein	456	20	0E0	96.69%	4	P:GO:0006147; F:GO:0008270; F:GO:0008892; C:GO:0005829	P:guanine catabolic process; F:zinc ion binding; F:guanine deaminase activity; C:cytosol	EC:3.5.4.3	guanine deaminase	G3DSA:3.20.20.140 (GENE3D); IPR006680 (PFAM); IPR014311 (TIGRFAM); IPR011059 (G3DSA:2.30.40.GENE3D); PTHR11271 (PANTHER); IPR014311 (PTHR11271:PANTHER); IPR032466 (SUPERFAMILY)	P:GO:0006147; F:GO:0008270; F:GO:0008892; F:GO:0016787; F:GO:0016810	P:guanine catabolic process; F:zinc ion binding; F:guanine deaminase activity; F:hydrolase activity; F:hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds

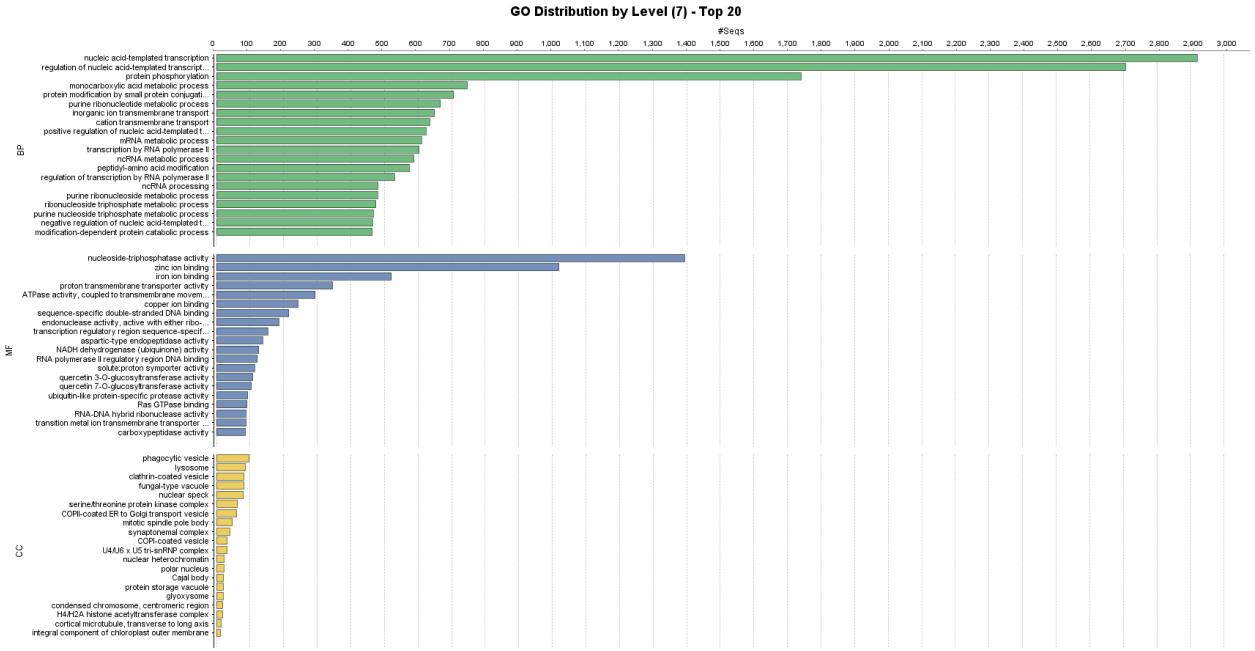
上記の例だと、
コンティグ「g344.t1」はグアニンデアミナーゼに関連し、
guanine catabolic processやguanine deaminase activityなどの関連するGO情報が
紐づけられたことが読み取れる。

遺伝子機能情報付与

アノテーション解析結果から様々なグラフを簡単に作成できます。
ここではその一部をご紹介します。



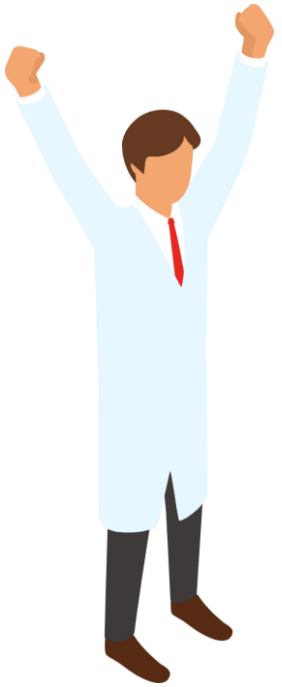
BLASTで検索された生物種のヒット数



各GO結果のランク図

OmicsBox のDNA-seq de novo アセンブリ

- 適切なアゼンブリツールを使用して解析可能
- 真核生物専用の遺伝子構造予測ツールを搭載
- 初心者でも解析できるインターフェース
- 7日間無料のデモライセンス→ 詳細(PDF)



Information

弊社HP紹介ページ

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