de novo トランスクリプトームの特性評価

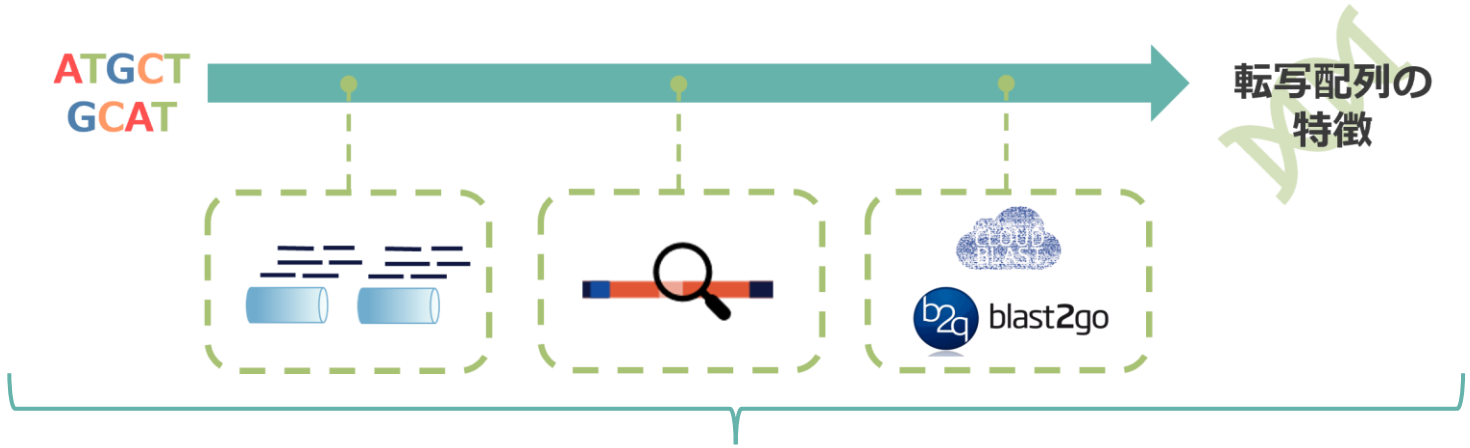
フィルジェン株式会社
バイオインフォマティクス部(biosupport@filgen.jp)

2024.03

はじめに

- de novo トランスクリプトームはリファレンスゲノムの無い生物種や
  アノテーションの無い新規転写産物の探索に有用な解析アプローチです。

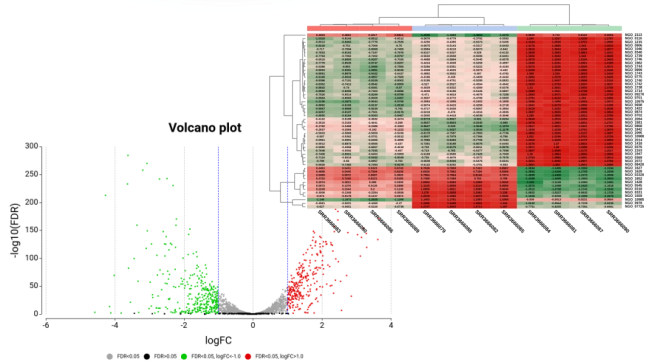
- この解析により、タンパク質の機能推定を得るだけでなく、
  リファレンスゲノムの無いあるいは不十分な生物種に対して、
  発現値定量や発現比較などのRNA-Seq解析を行うことができるようになります。



de novo トランスクリプトームの特性評価

転写配列の特徴

RNA-Seq解析



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。
  RNA-Seq 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

RNA-Seq 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。
  Trinityツールを統合

コンティグの完全性評価

- RNA-Seq de novo アセンブリで作成したトランスクリプトーム配列の品質を確認する。
  BUSCOを統合

配列クラスタリング

- 配列類似性に基づいてクラスタリングを行うことでトランスクリプトーム配列冗長性を取り除く。
  CD-HITを統合

コーディング領域予測

- トランスクリプトーム配列からコード領域を探す。
  TransDecoderを統合

アノテーション

- BlastやInterProScanにより遺伝子機能情報を付与することができる

OmicsBoxで簡単に解析

- 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

- RNA-Seq de novo アセンブリやBlast解析は、 高スペックPCやサーバーの導入が必要

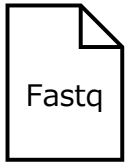
OmicsBox

OmicsBoxのde novo トランスクリプトームの特性評価

- 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＆既成のワークフローで簡単に解析
- de novo トランスクリプトームの特性評価を行える数少ないソフトウェア

使用データ

必要なファイル

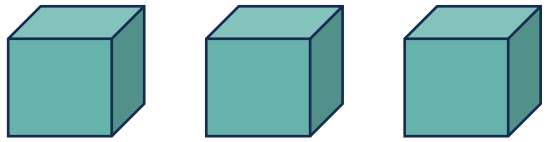


シーケンサーから出力された生データ
もしくは受託サービスで得られたクリーンリードデータ

例えば以下のような実験を行うとします：

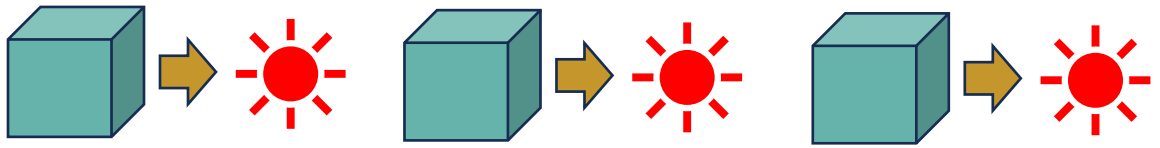
果実の褐色腐敗病の原因となる*Monilinia laxa*

暗所で 4 日間生育した菌糸体



n=3

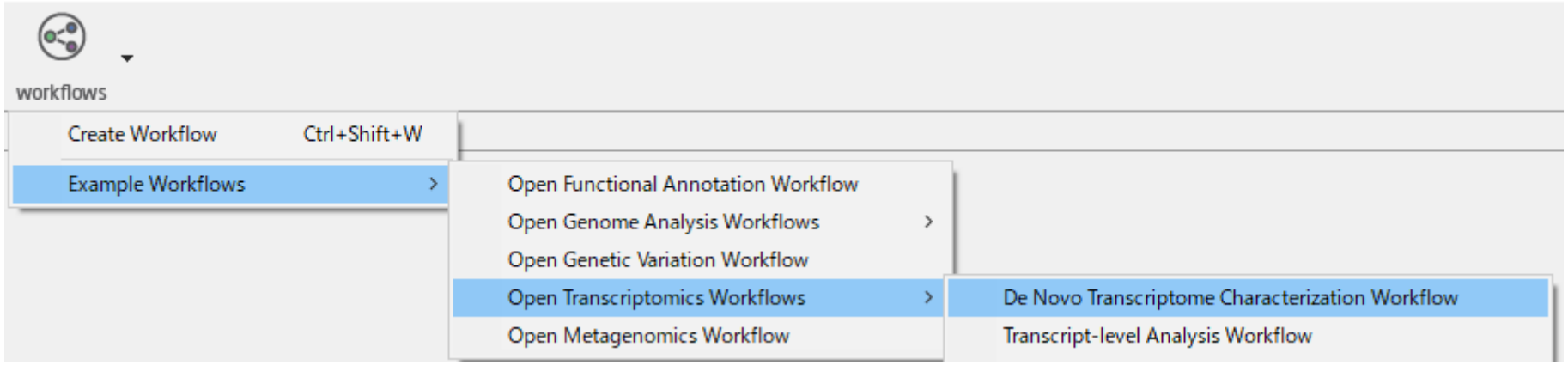
暗所で 2 日間生育し、2 日間光に曝露した菌糸体

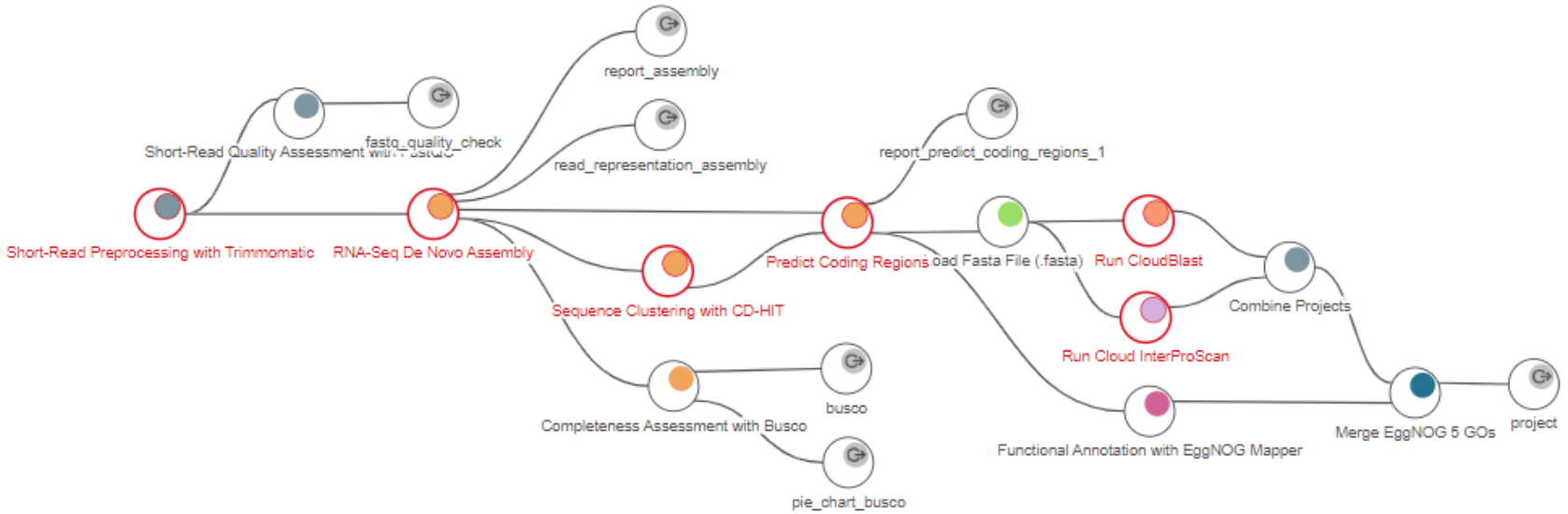


n=3

この場合、すべてのサンプル（計6サンプル）をまとめて解析し
1つのトランスクリプトーム配列を構築することになります。

OmicsBoxのワークフロー機能

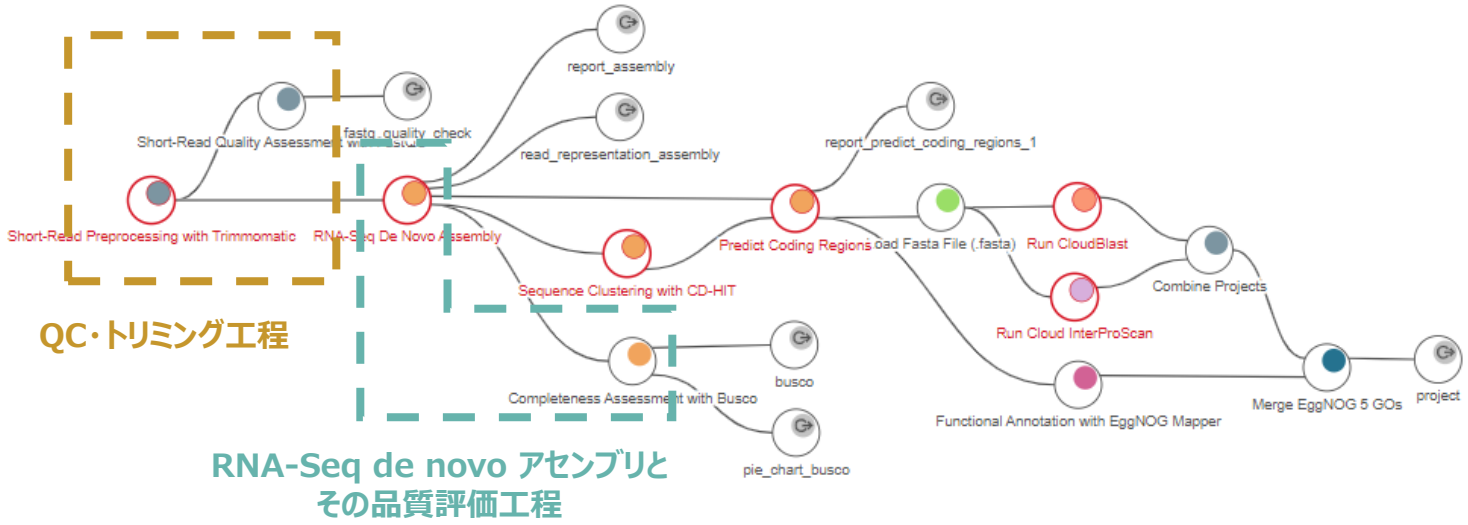




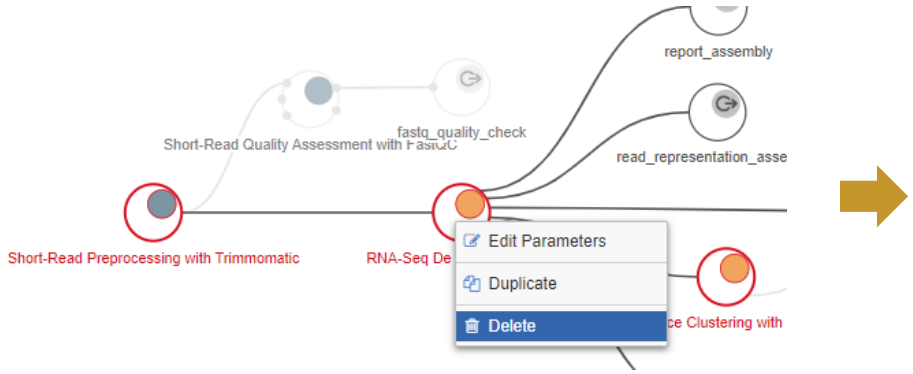
この解析のための既成のワークフローが搭載されています。

*Transcriptomics ModuleとFunctional Analysis Moduleを使用します。

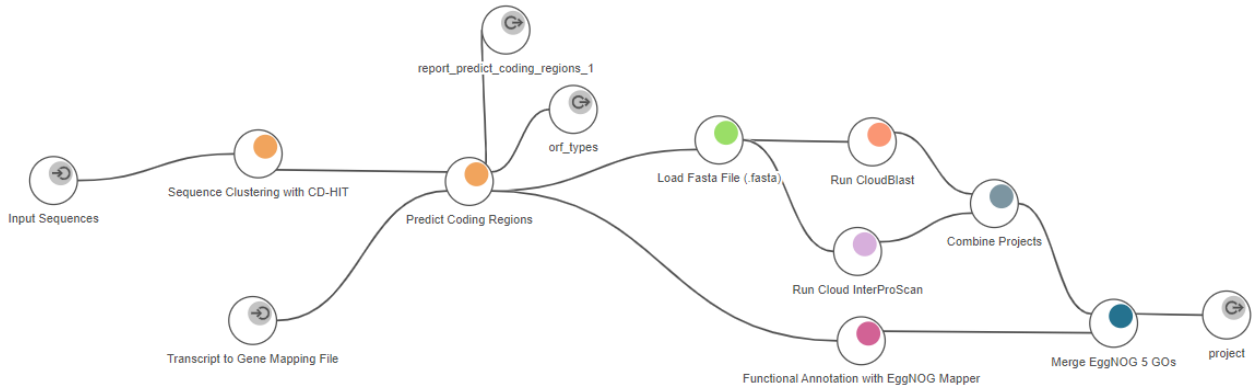
OmicsBoxのワークフロー機能



だだし、QC・トリミングとRNA-Seq de novo アセンブリとその品質評価については結果によっては、
この解析を繰り返す必要があるため、これらの工程は個別に解析を行い以下のようにワークフローを修正できます。



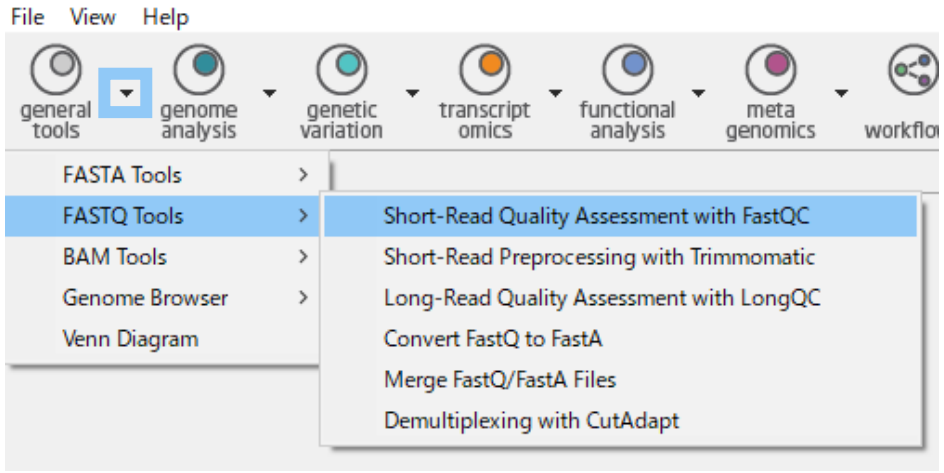
該当ステップを右クリックで削除



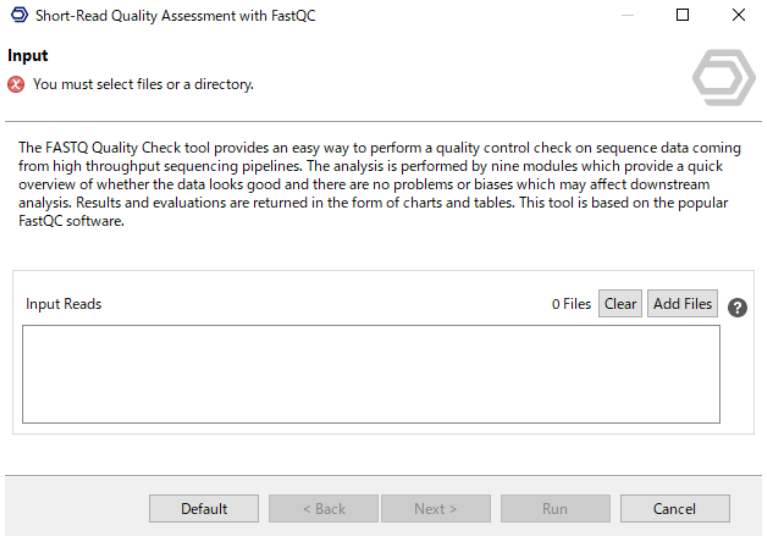
修正後のワークフロー

QC・トリミング

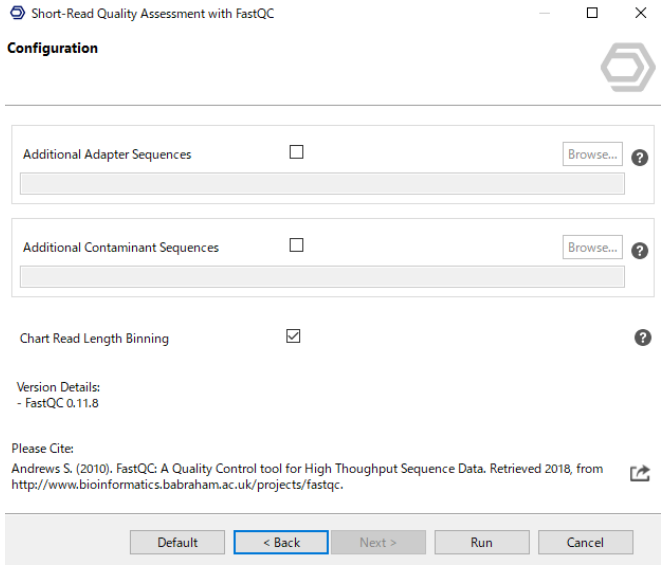
・データが良好か、下流分析に影響する問題がないか確認



画面上部のアイコンからQCツールを起動



全てのサンプルをこちらに指定



任意でアダプタリストを追加

結果

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

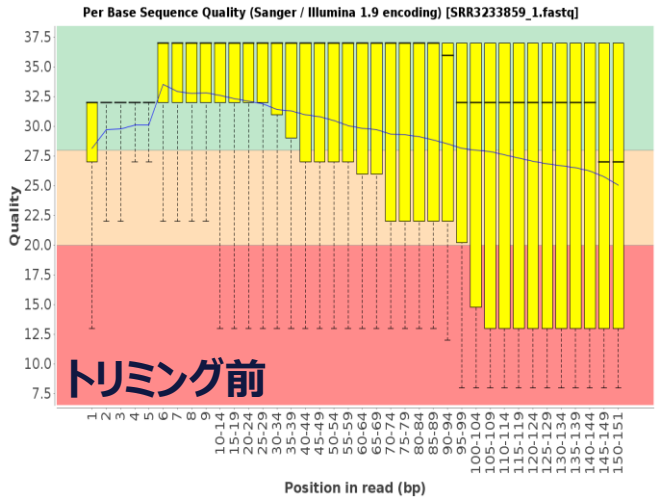
Name	Per Base Sequence Quality	Per Sequence Quality Scores	Per Base Sequence Content	Per Sequence GC Content	Per Base N Content
ERR1948631_1.fastq	PASS	PASS	FAIL	PASS	PASS
clean_ERR1948631_1.fq	PASS	PASS	FAIL	PASS	PASS

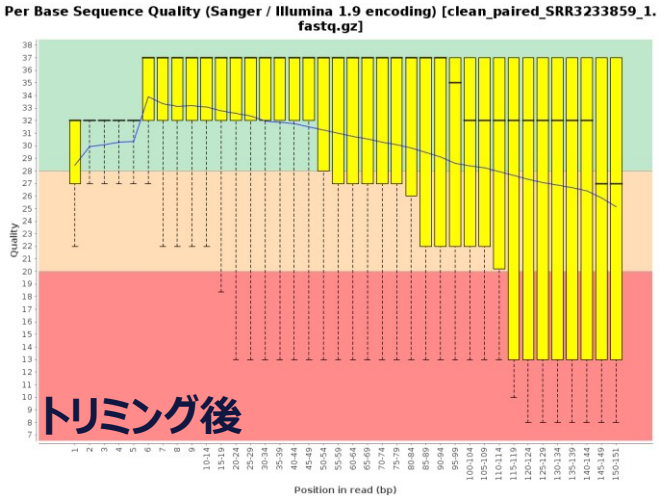
Name	Sequence Length Distribution	Adapter Content	Overrepresented Sequences	Sequence Duplication Levels	Report
ERR1948631_1.fastq	PASS	FAIL	WARNING	FAIL	
clean_ERR1948631_1.fq	WARNING	PASS	WARNING	FAIL	

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）

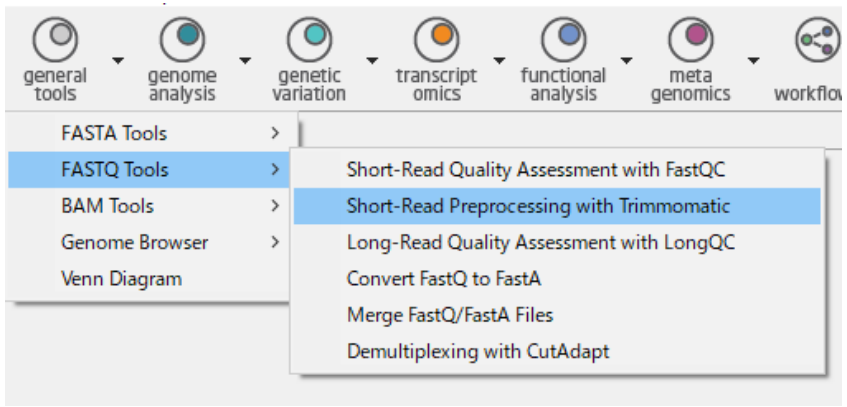




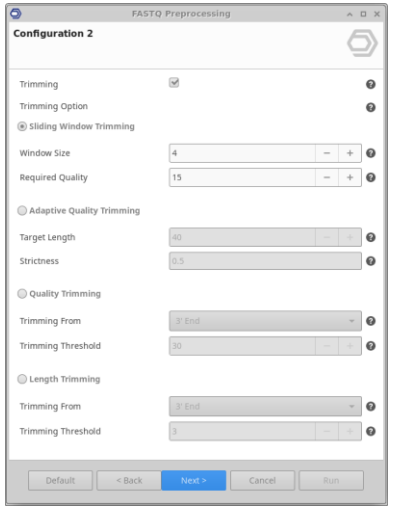
シーケンスデータの品質をすばやく評価

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

QC・トリミング



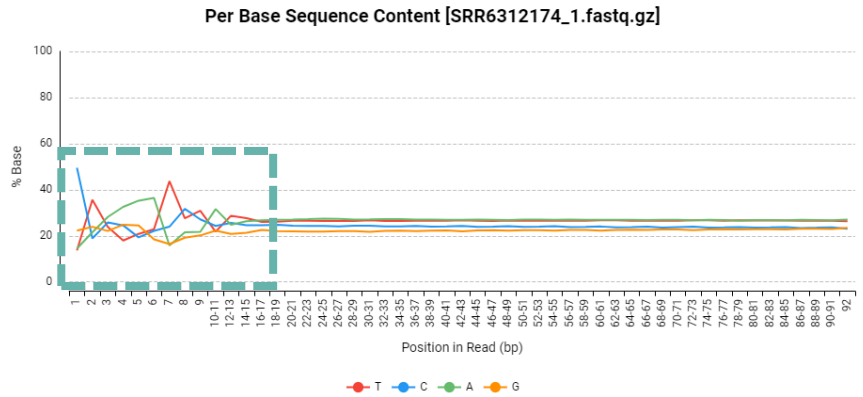
QC後必要に応じてトリミングを行う



様々なトリミング方法
トリミング後再度QCを行う

【補足】必ずしもすべてのQC項目を正常（PASS）にする必要はありません。

Name	Per Base Sequence Quality	Per Sequence Quality Scores	Per Base Sequence Content
SRR6312175_1.fastq.gz	PASS	PASS	FAIL
SRR6312174_1.fastq.gz	PASS	PASS	FAIL
SRR6312174_2.fastq.gz	PASS	PASS	FAIL



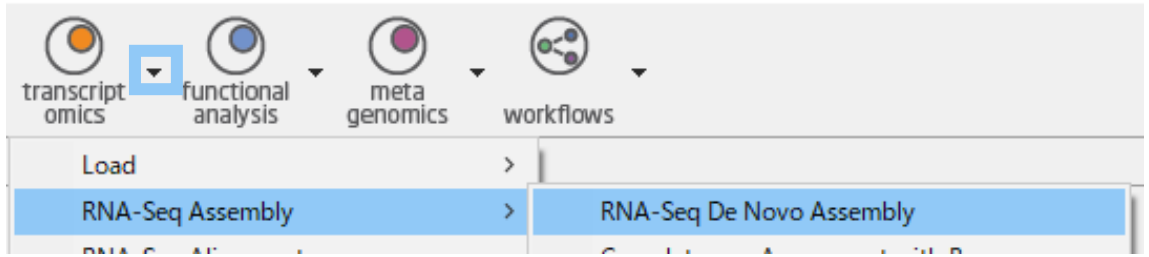
FalseやWarningと判定されていても、それを説明できる理由があれば、適切なリードであると判断することが多いです。

例）

左の結果ではPer base Sequence ContentがFalseですが、
こちらはRNA-Seqデータであれば想定される形状に見られます。

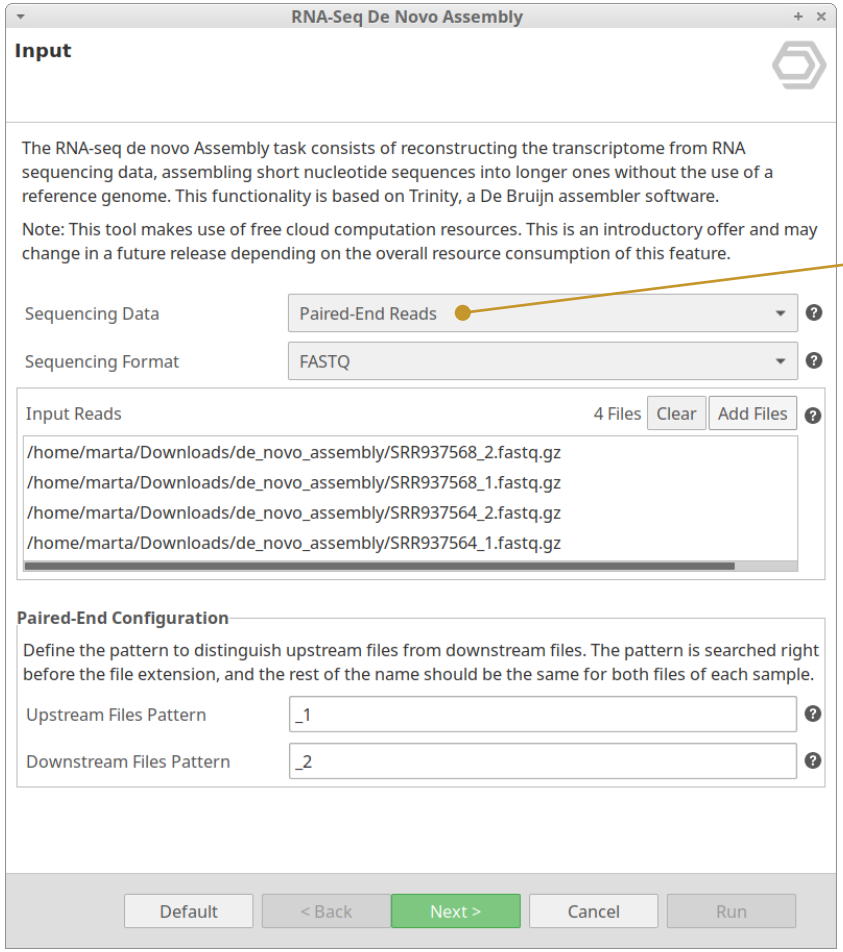
RNA-seqライブラリー調製におけるプライミングはそれほどランダムではないので、
最初の塩基の塩基含量のシフトが起こることが予想されるためです。

de novo アセンブリ



① de novo アセンブリツールをクリックします。

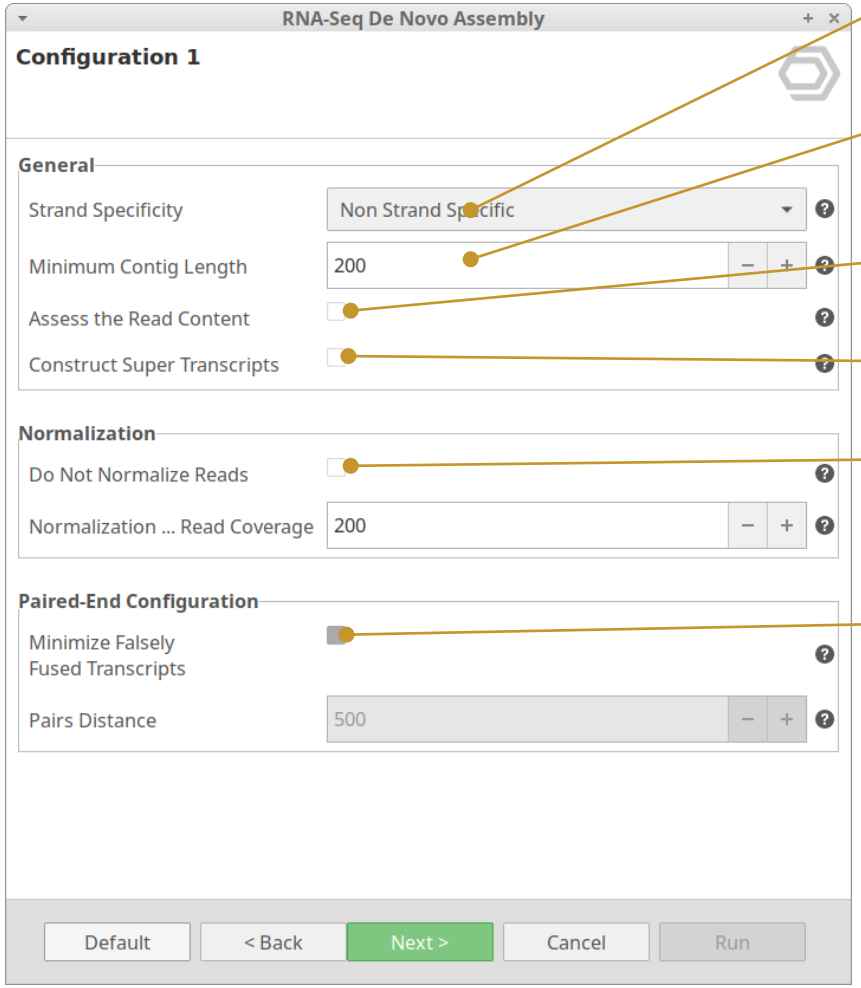
Transcript Moduleを使用します。



ペアエンドリードか
シングルエンドリードか
を指定

② 使用するデータをすべて入力します。

de novo アセンブリ



RNA-seqリードの鎖長を定義します。
ライブラリー調整の条件や受託元にご確認ください。

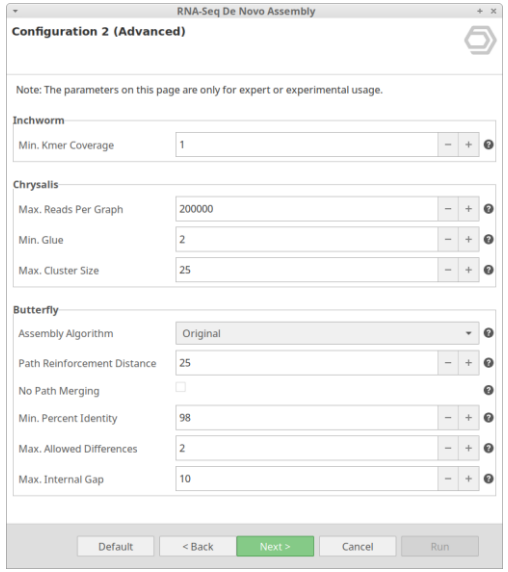
コンティグの最小長

Bowtie2を使用して入力RNA-Seqリードを
トランスクリプトームアセンブリにアラインメントします。

アセンブリ結果からSuperTranscriptsを構築します。

正規化をオフ
通常チェックを入れない

融合転写産物を最小限に抑えることができます
コンパクトな真菌ゲノムでは、強くお勧めします。
注）負荷の高い操作なので、必要がない限り
使用しないでください。



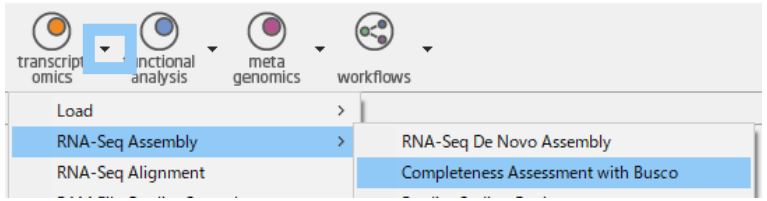
④次の画面のパラメータは、Trinityツールに詳しい
ユーザー向けの設定を目的としています。
（セミナーではデフォルトの設定で実行）

③ 必要に応じてパラメータ設定を変更します。

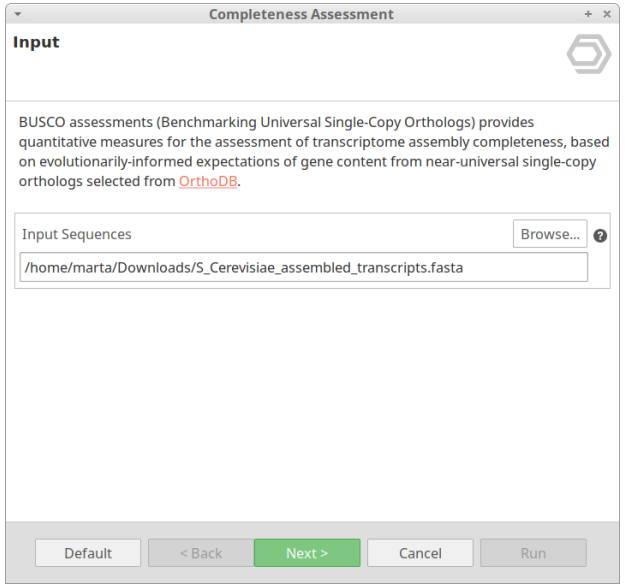
（セミナーで使用したサンプルは真菌のためPaired-End Configurationのみチェックをいれ後はデフォルトの設定で実行）

de novo アセンブリ・コンティグの完全性評価

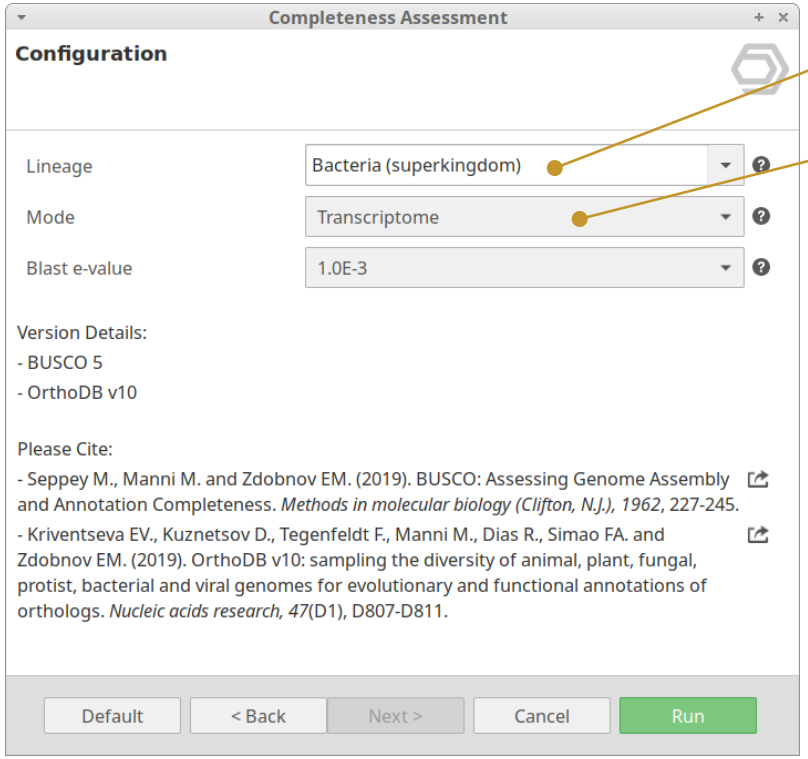
結果のFASTAファイルに加えて、レポートが生成されます。
このFASTAファイルをしようしてコンティグの完全性を評価します。



①完全性評価ツールをクリックします。



②de novo アセンブリ結果のFASTAファイルを入力します。



評価する種に応じて、適切な系統を選択します。

Transcriptomeを選択します。

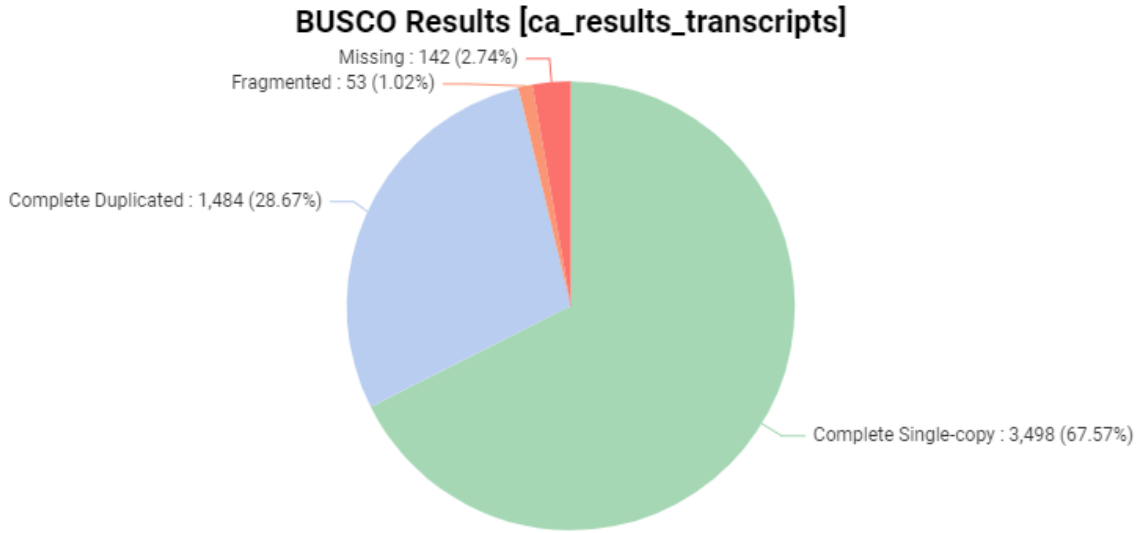
③データに合った設定を行います。

de novo アセンブリ・コンティグの完全性評価

BUSCO Results

96.23% of the BUSCO groups have complete gene representation (single-copy or duplicated), whi

BUSCO Notation	BUSCOs	BUSCO IDs	Sequences	Sequence IDs
Complete Single-copy	3,498 / 67.57%	Id	3,359 / 10.4%	Id
Complete Duplicated	1,484 / 28.67%	Id	3,405 / 10.54%	Id
Fragmented	53 / 1.02%	Id	53 / 0.16%	Id
Missing	142 / 2.74%	Id	25,630 / 79.34%	Id

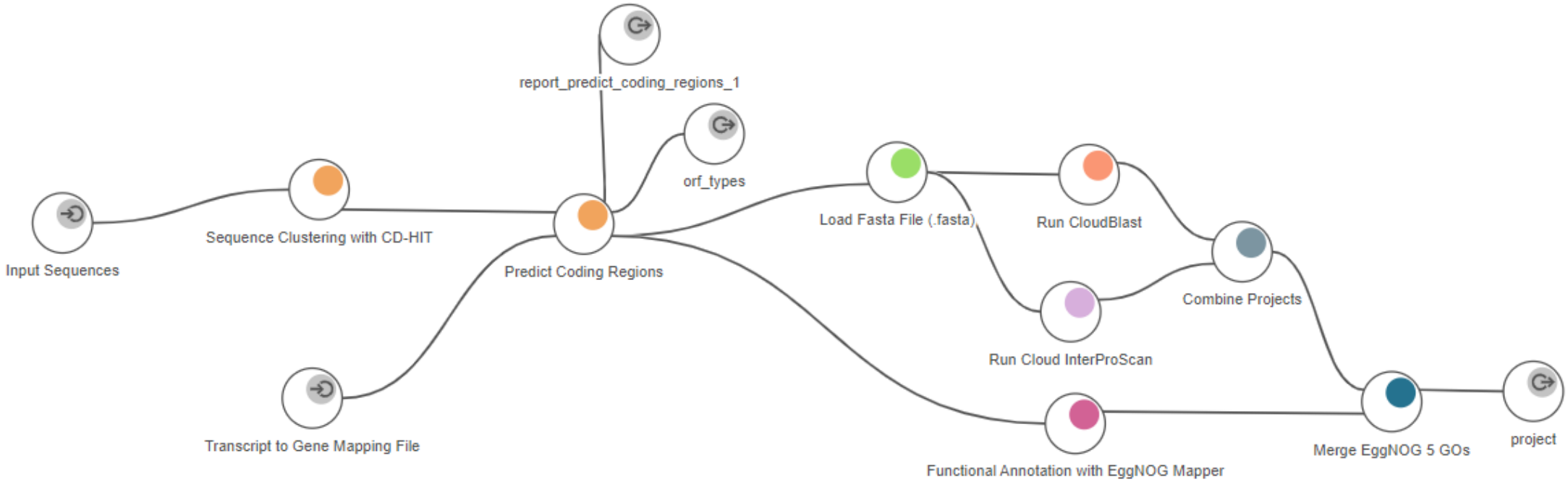


レポートやチャートからコンティグの完全性を簡単に評価できます。

上記の例ではアセンブルされた配列中のおおよそ96%がCore gene setに含まれていたので
下流分析に使用できると判断することができます。

ワークフローの設定

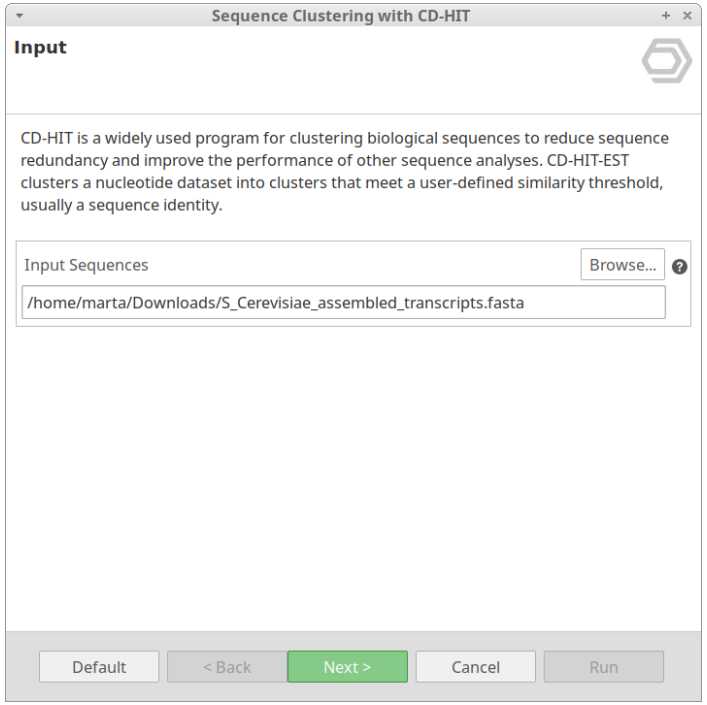
残りの解析についてはワークフローを使用します。

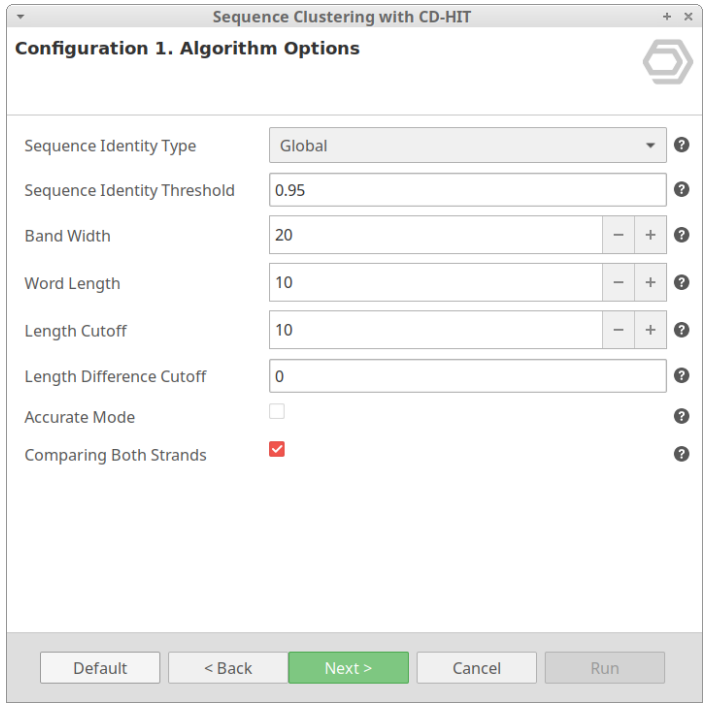


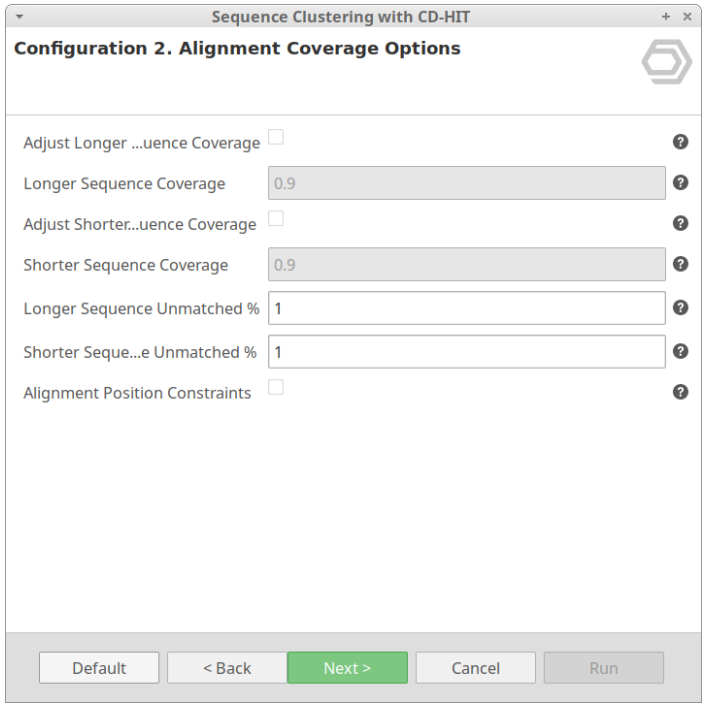
ワークフローでは各ツールがアイコンで表現されています。
赤枠のアイコンは設定が不十分でワークフローを開始できません。
またアイコンの設定はすべてデフォルトの設定になっているので、各ツールの設定を確認・修正が必要です。
設定はアイコン上で右クリック、もしくはダブルクリックで行います。

配列クラスタリングの設定

の設定画面を開くと次のような画面が現れます。







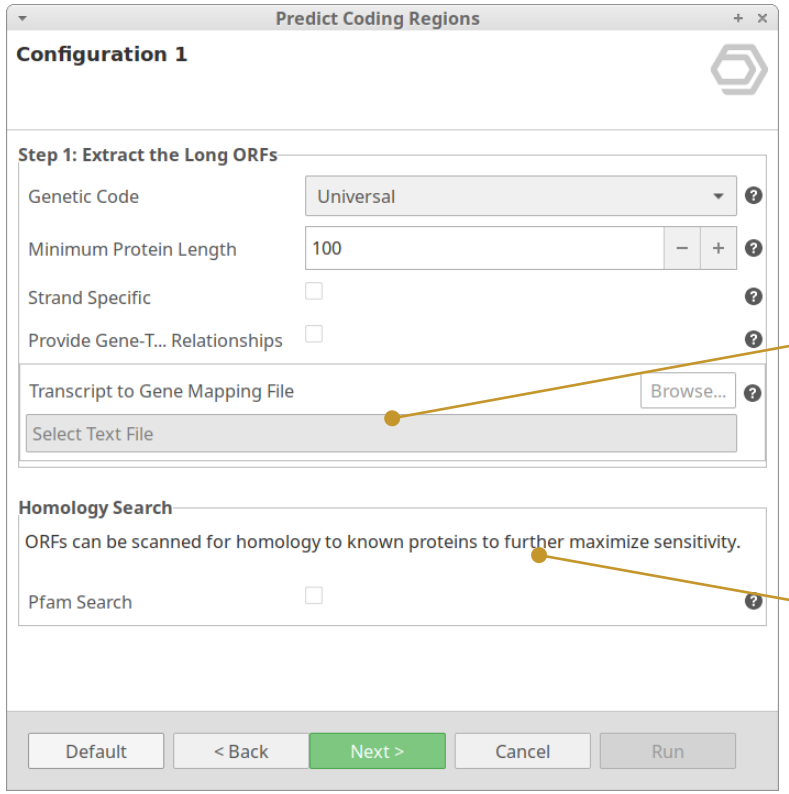
①de novo アセンブリ結果のFASTAファイルを入力します。

②任意で設定を行います。

（セミナーではデフォルトの設定で実行）

コーディング領域予測の設定

の設定画面を開くと次のような画面が現れます。

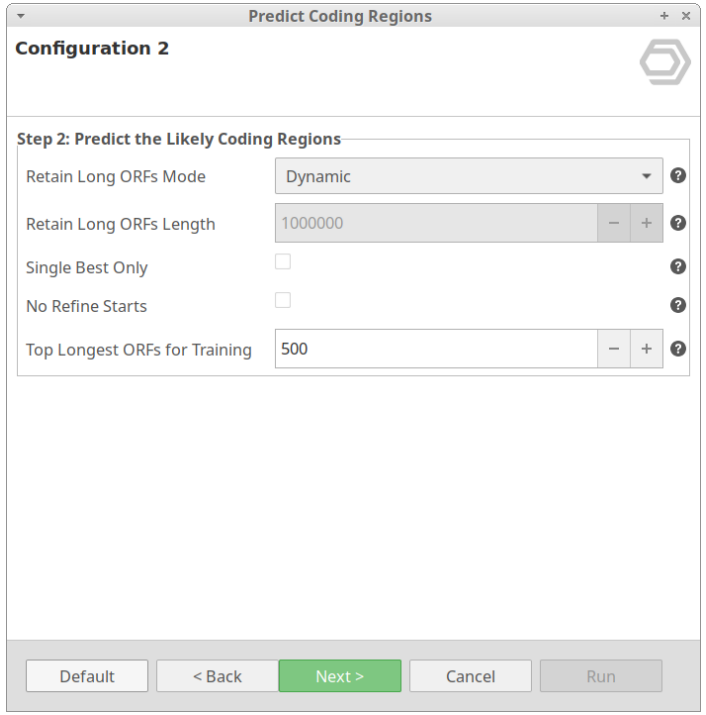


de novo アセンブリの工程で作成されたTranscript to Gene Mapping Fileテキストデータを入力します。

Pfam検索により、既知のタンパク質との相同性を持つ ORF を特定します。有効にすることをお勧めいたしますが、実行時間が大幅に増加することに注意してください。

①必要に応じてパラメータ設定を変更します。

（セミナーではTranscript to Gene Mapping File の入力とPfam検索のみデフォルト値から変更）

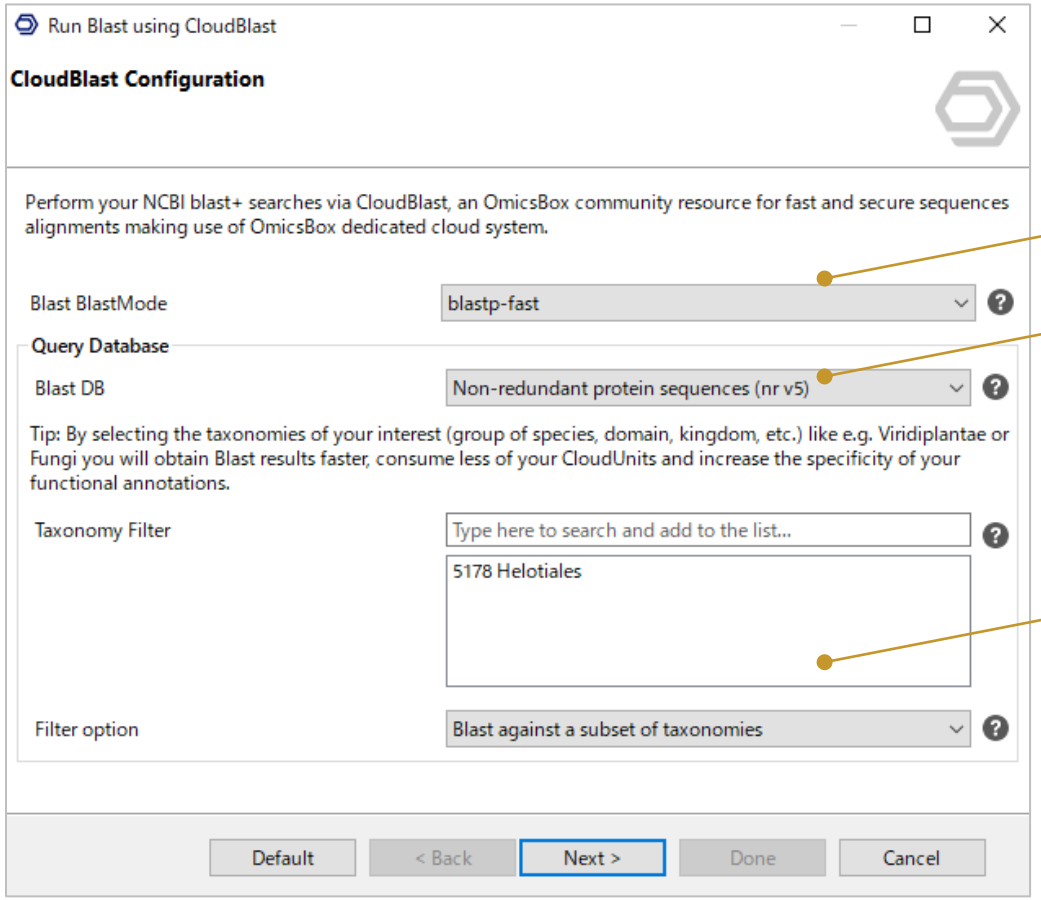


②コーディング領域の配置の可能性を予測するための設定を任意で変更します。

（セミナーではsingle Best Onlyにチェックを入れ
それ以外はデフォルトの設定で実行）

Blastの設定

の設定画面を開くと次のような画面が現れます。



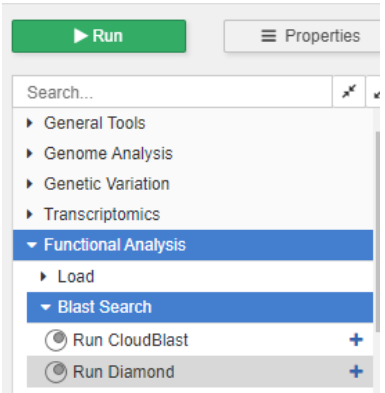
タンパク質配列のためblastp(あるいはblastp-fast)を指定します。

任意のデータベースを指定します。

分類を選択することによってより高速のBlast結果が得られ、Cloud Computation Unit*の消費量が少なくなり、機能アノテーションの具体性が高まります。

*Computation Unitsは、CloudBlast解析を行うごとに消費されます。（InterProScanのCloud解析でも消費）
すべてのUnitsを使いきってしまった場合は、追加で 6 million Computation Units を購入することができます。

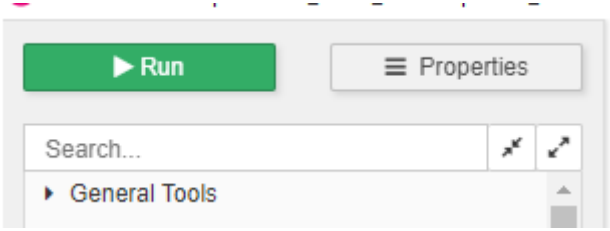
【補足】
Cloud　BlastをDiamondに変更できます。



ワークフロー左側のメニューからRun Diamondの「＋」をクリックすると解析アイコンが出現します。
後はCloud　Blastを削除しこのアイコンに置き換えます。

その他の設定とワークフローの開始

の設定も任意に変更可能　（セミナーではデフォルトの設定で実行）
*InterProScanは保存先を指定するだけで赤枠のアイコンから通常のアイコンに変わります。



全ての設定が完了したら、Runボタンをクリックすることで
一連の解析がスタートします。

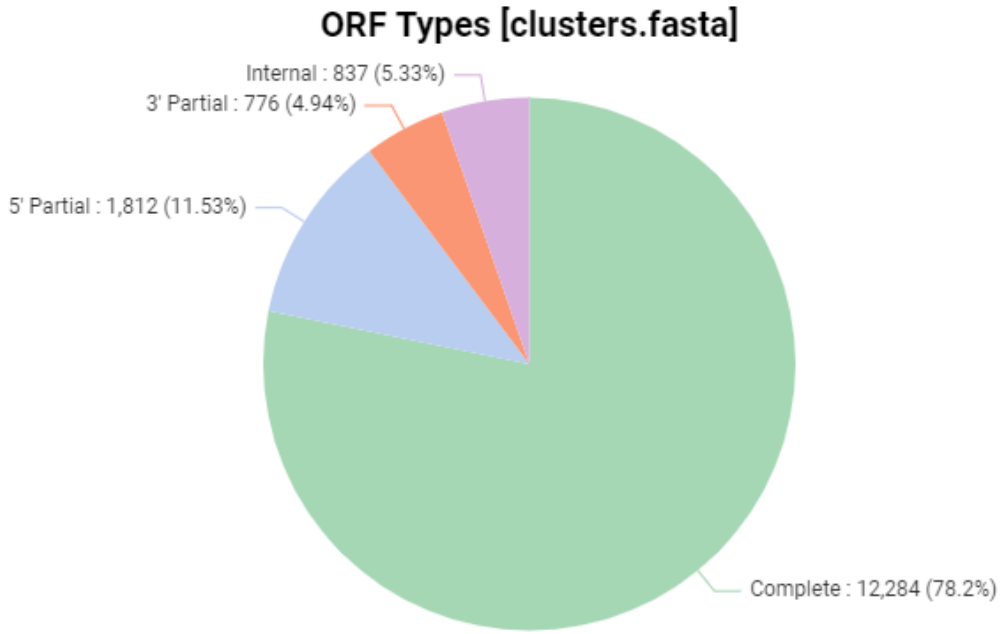
ワークフローの結果

コーディング領域予測の結果

TransDecoder Results

ORF Type	Sequences
Complete	12,284
5' Partial	1,812
3' Partial	776
Internal	837

Longest ORF per gene:

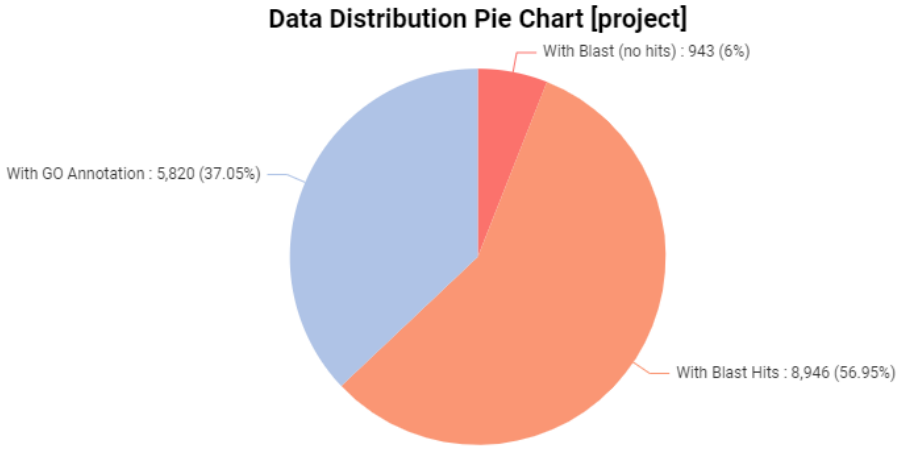


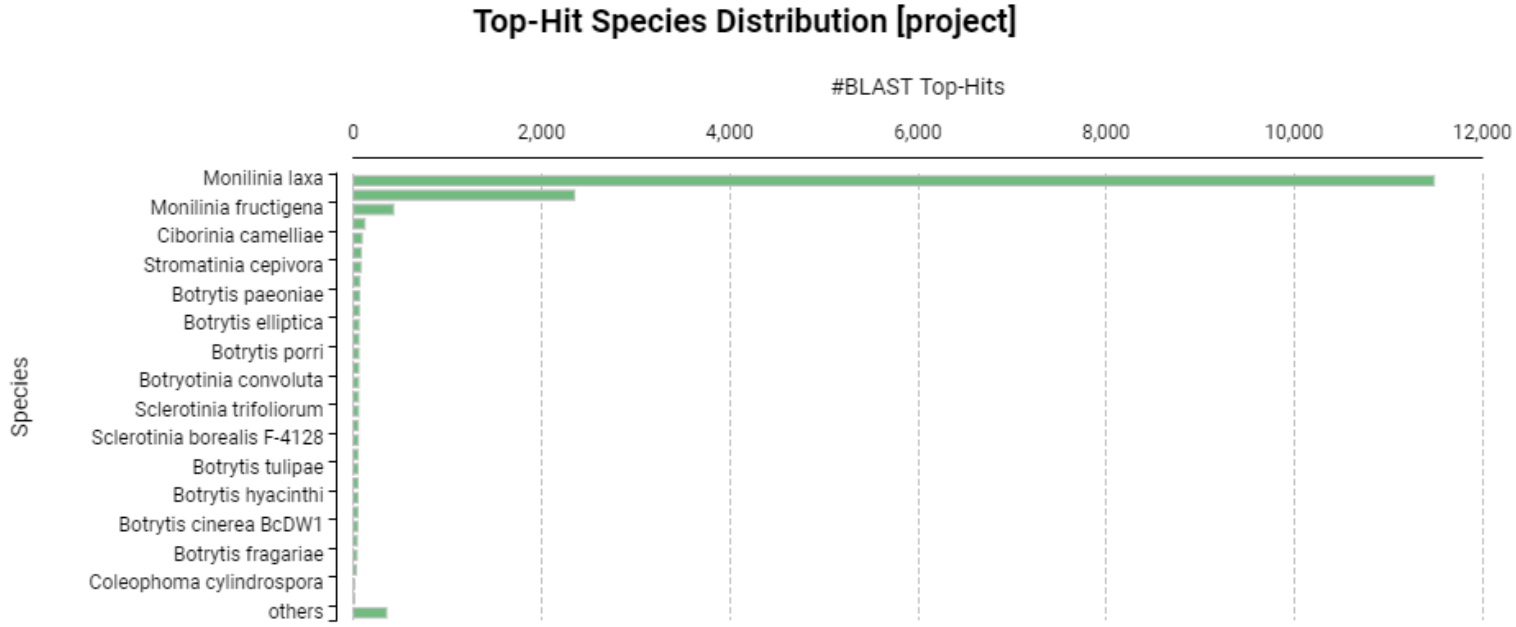
レポートやチャートから予測されたコーディング領域属性を簡単に評価できます。

Complete:開始コドンと停止コドンが含まれています。
5' partial: 開始コドンが欠落しており、おそらく N 末端の一部が欠落しています。
3' partial: 終止コドンが欠落しており、おそらく C 末端の一部が欠落しています。
Internal: 5‘ と 3’ の両方とも部分的です。

ワークフローの結果

Blat、InterProScan、EggNog Mapperの結果

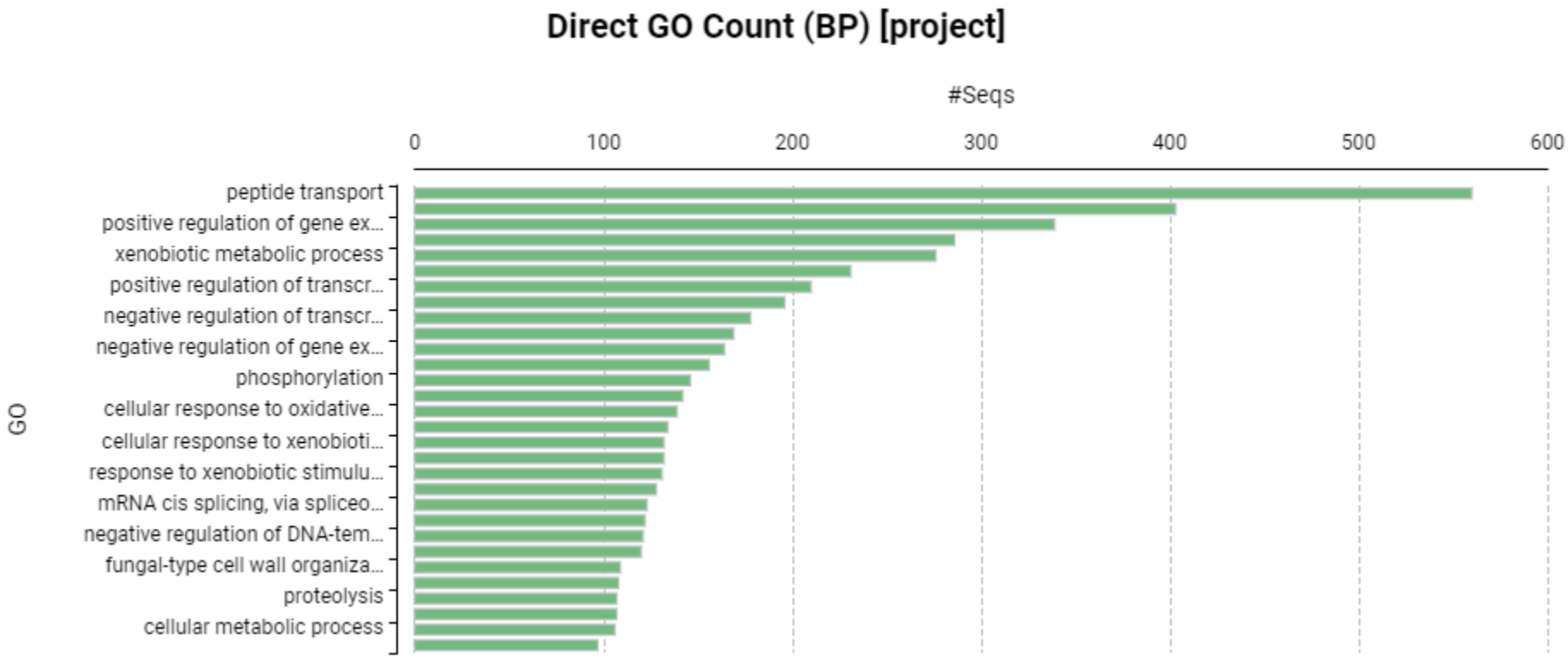




どのぐらいのトランスクリプトーム配列に特徴がつけられたか
どの生物種のアノテーションが付与されたかグラフで表現できます。

ワークフローの結果

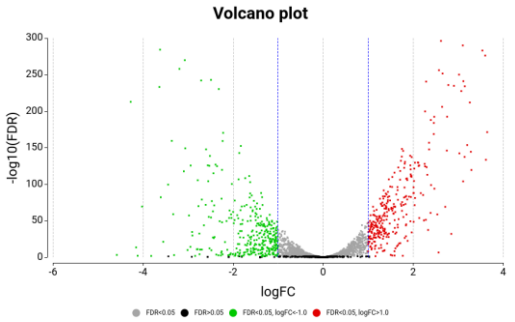
Blat、InterProScan、EggNog Mapperの結果

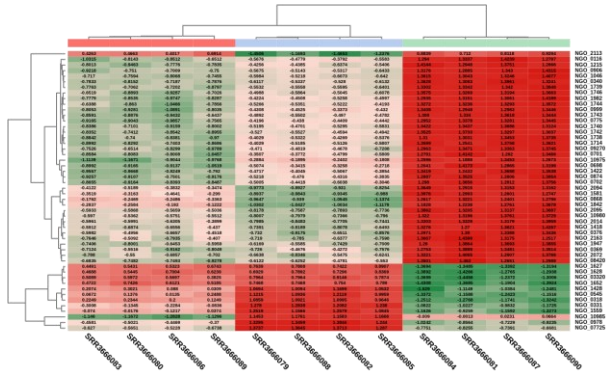


データセット内で最も頻繁に使用される GO Termを表示します(上図はbiological processカテゴリ)。

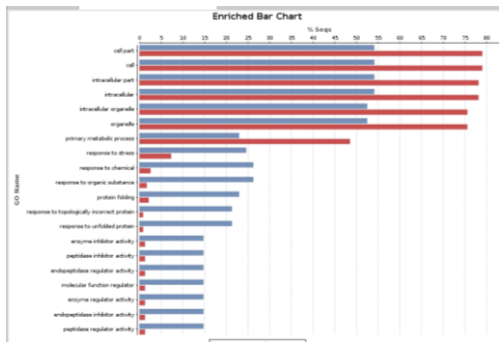
de novo トランスクリプトームの応用

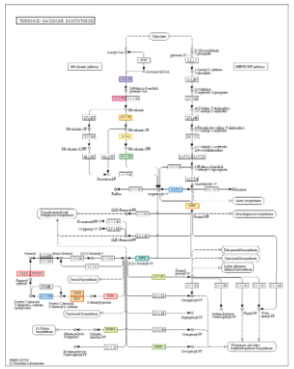
コーディング領域予測の結果あるいはCH-Hitの結果





RNA-Seq解析





機能エンリッチメント解析・パスウェイ解析

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