NGSデータ解析：ロングリードアセンブリ

～OmicsBoxを使用した解析機能のご紹介～

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)

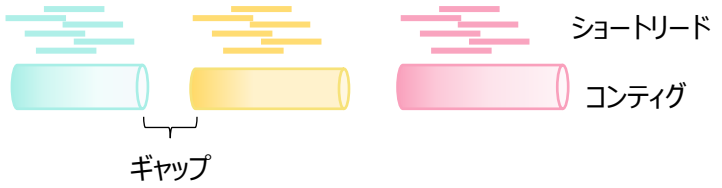
2020.06

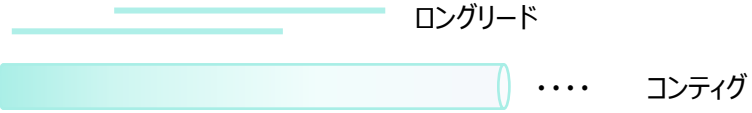
はじめに

ロングリードの利点

第3世代シーケンステクノロジーにより
長いシーケンスリードの生成が可能に。

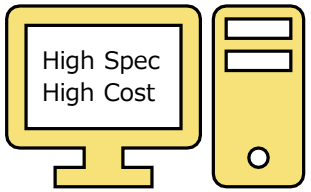
→全ゲノムシーケンスプロジェクトで使用すると、
リピート配列のシーケンスや連続したゲノムアセンブリを生成 !!





その反面・・・

エラー率が比較的高いため、
正確な最終シーケンスの生成が困難



計算量が多いため
高度なスペックのPCが必要

有名なアルゴリズムを使用した解析は、
コマンドライン操作が必要

はじめに

OmicsBox

✓非モデル生物に対応

OmicsBoxはリファレンスゲノムがないデータでも解析が実行可能です。
農学系のユーザーに適した解析ツールが搭載されています。

✓初めてでも簡単に使用することができます。

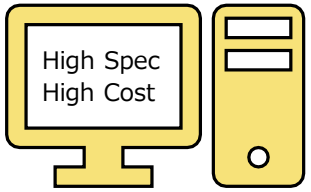
実績は高いがコマンドライン型であったりOSに制限がある
オープンソースソフトウェアを多数組み込みマウス操作で簡単に
解析できるようしたのがOmicsBoxの特徴の1つです。

✓高価で高スペックなPCは不要

解析や計算は、統合させたウェブサイトや
Biobam社のクラウドを通して行われるため、
安定したインターネット接続があれば解析が可能です。

解析する内容

エラー率が比較的高いため、
正確な最終シーケンスの生成が困難



計算量が多いため
高度なスペックのPCが必要

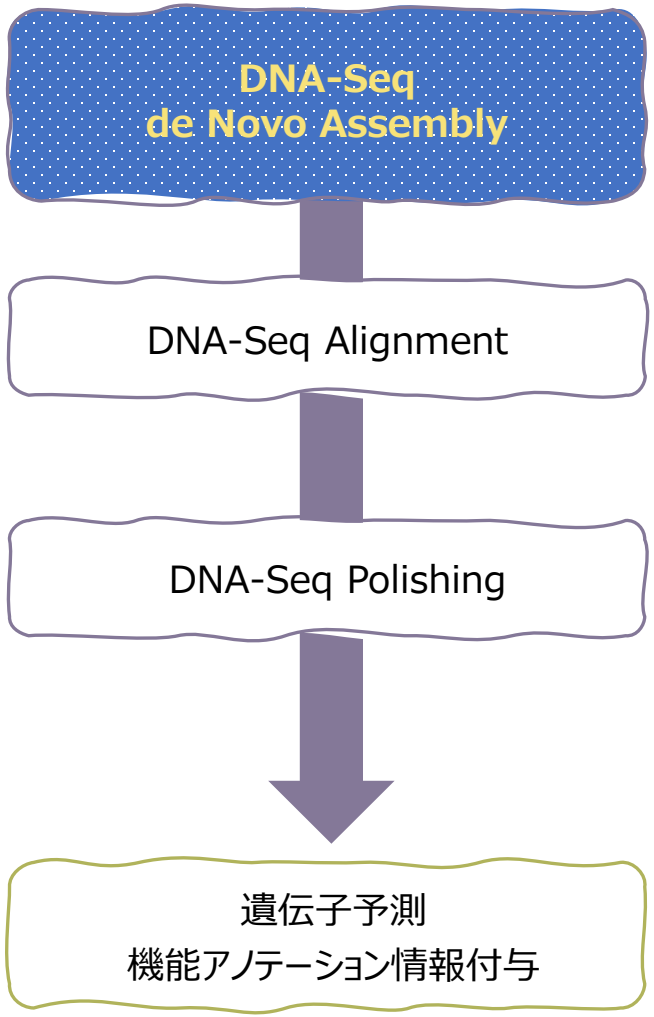
有名なアルゴリズムを使用した解析は、
コマンドライン操作が必要

De-Novo Assembly(ロングリードデータ)解析の課題

目的

・複雑な手順なしで
ロングリードのデータを用いて連続したゲノムアセンブリを生成する。

・低いエラー率のゲノムアセンブリを生成する。



DNA-Seq de Novo Assembly

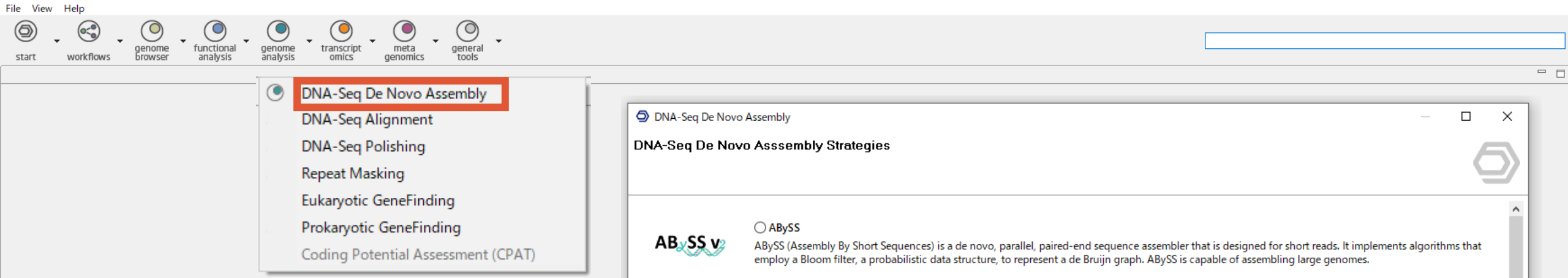
・参照ゲノム配列なしで1からゲノムを再構築します。

・シークエンサーから出力したリードから長い連続したシーケンス（コンティグ）を作成することが本解析の目的です。

・現在(2020.06)3つのアセンブリ戦略を実行できます。
本セミナーではロングリードの解析をご紹介いたします。

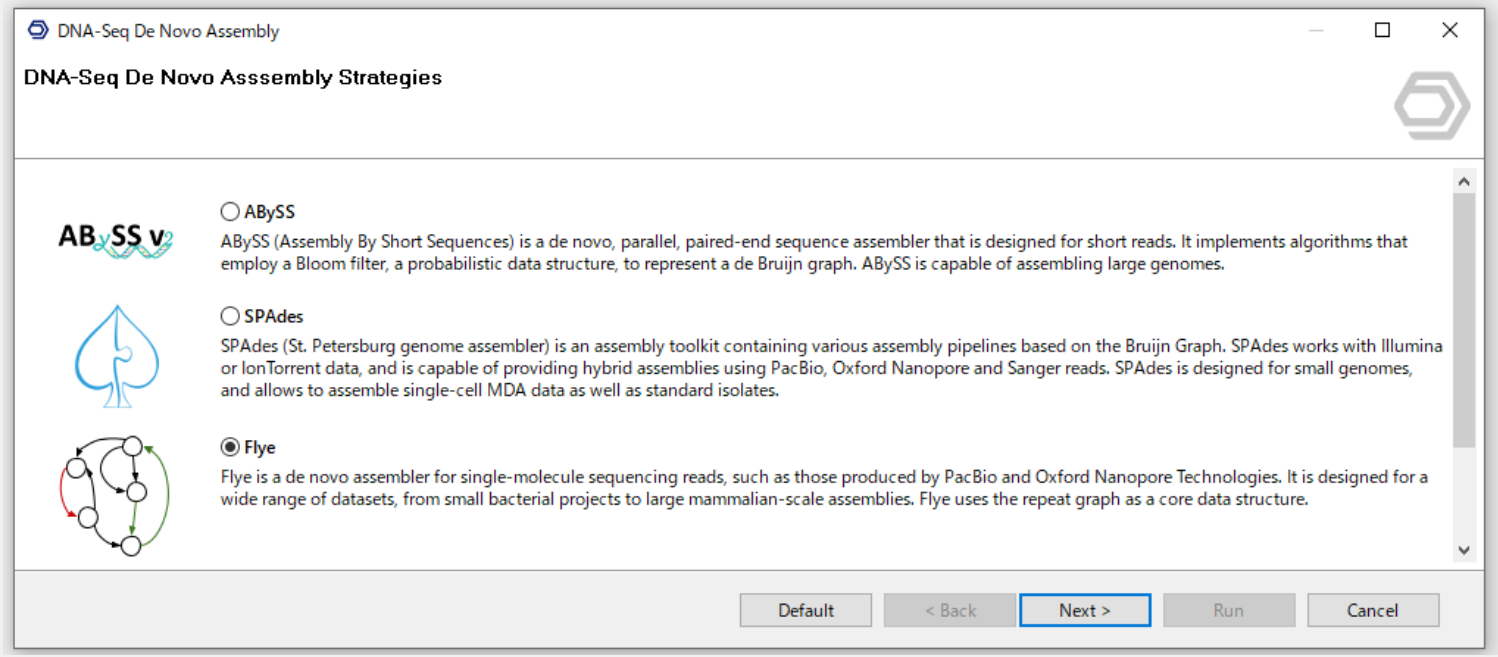
通常ハイスペックPCでの解析が必要ですが
OmicsBoxではハイスペックPCの購入なしで実行可能です。

ロングリードデータを組み立てる



- 行いたい解析は画面上部のアイコンから選択

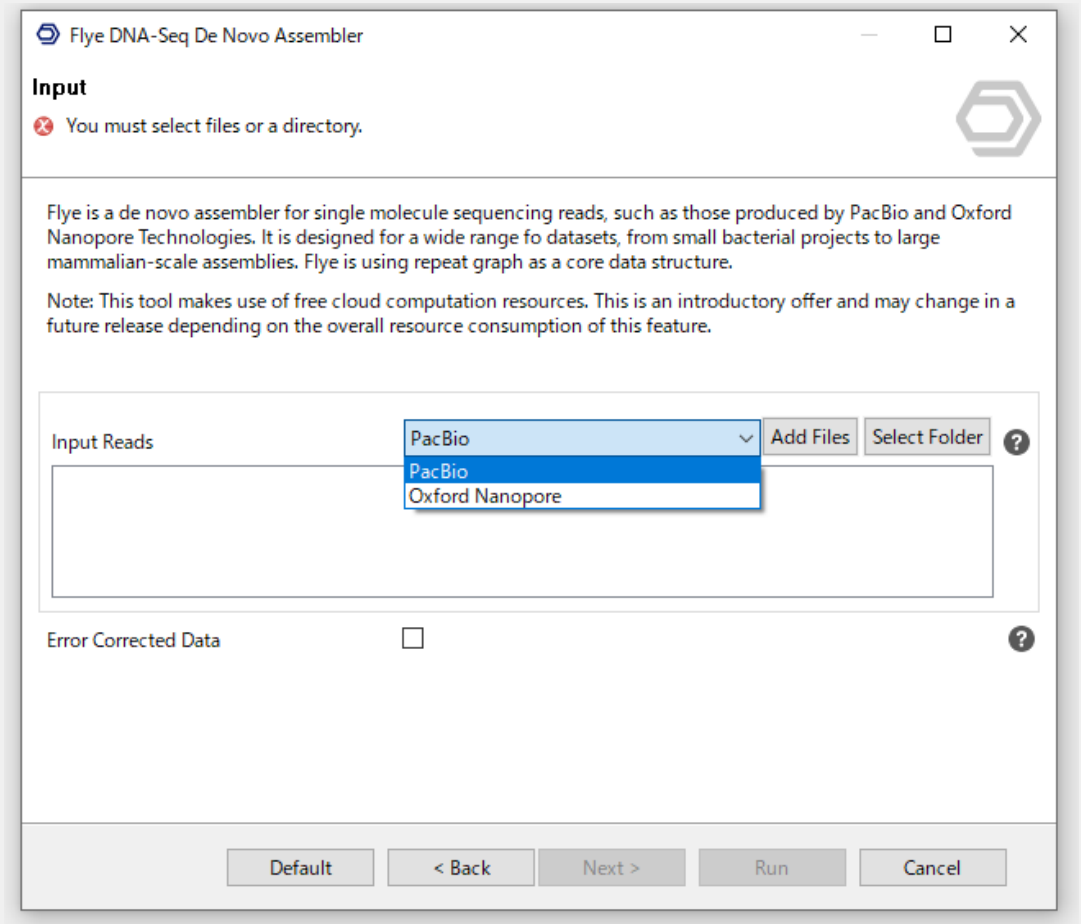
✓ DNA、RNA、メタゲノムなど各カテゴリーに
機能が集約されているので簡単に探すことができます。



- アゼンブラーのアルゴリズムを選択
- ロングリードデータの場合「Flye」を選択
Flyeは小さな細菌プロジェクトから哺乳類規模の大規模アセンブリまで、
幅広いデータセット用に設計されています。

✓ 使用するリードデータに合わせて最適なアルゴリズムで解析できます。

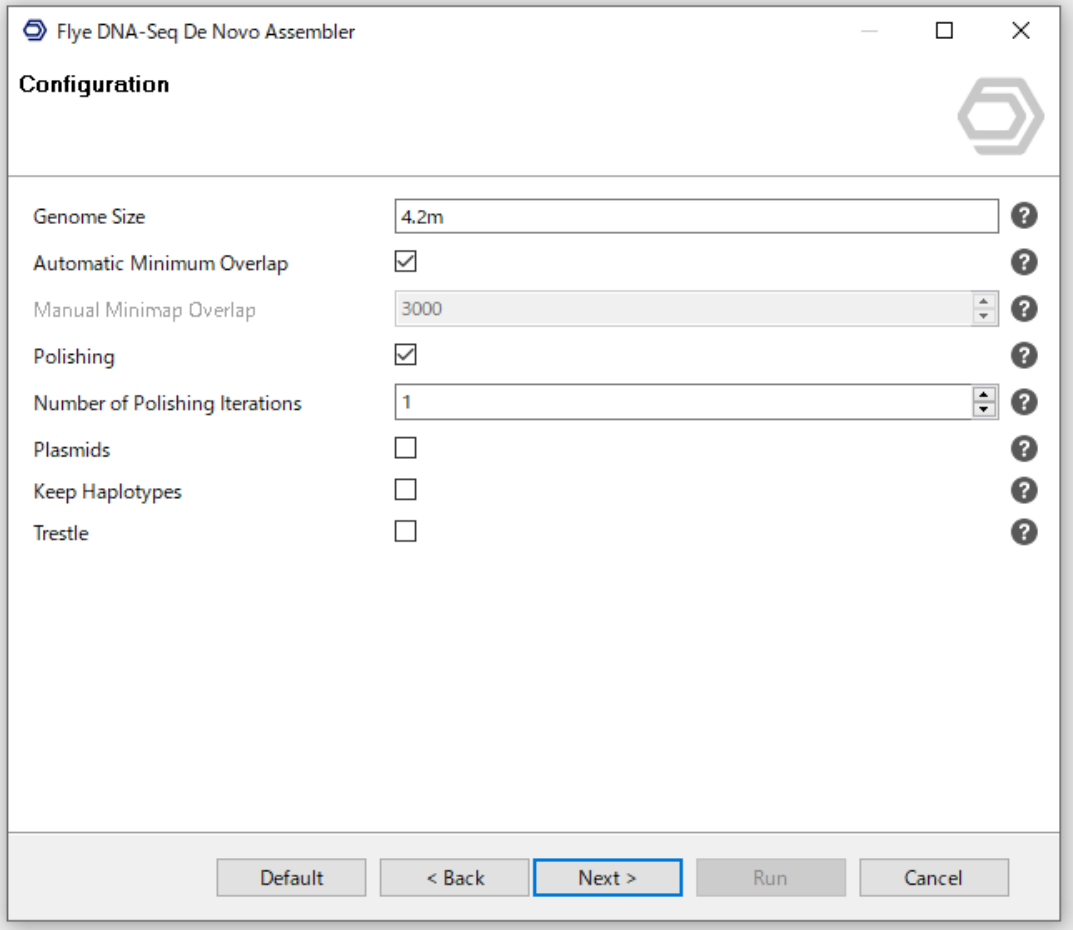
ロングリードデータを組み立てる



- 次世代シークエンサーより出力されるデータを入力Fastqファイルである必要があります。

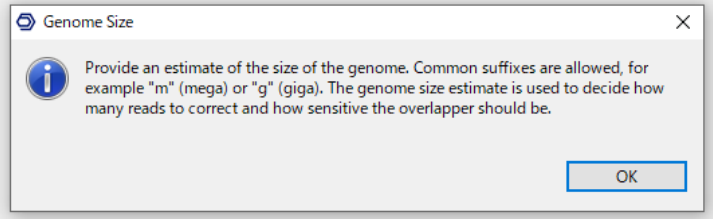
✓ PacBioおよびOxford Nanopore Technologiesがサポートされています

ロングリードデータを組み立てる



- 予想されるゲノムサイズやPolishing回数などのパラメータ設定

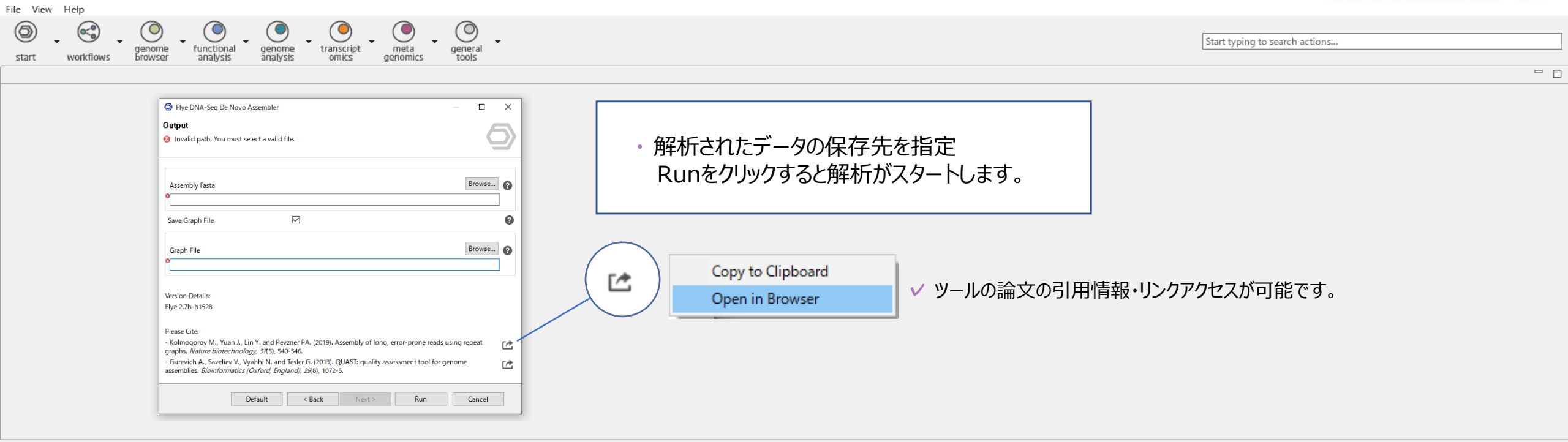
✓ ? を押すと パラメータの説明をみることができます。



Genome Sizeのパラメータ設定の ?

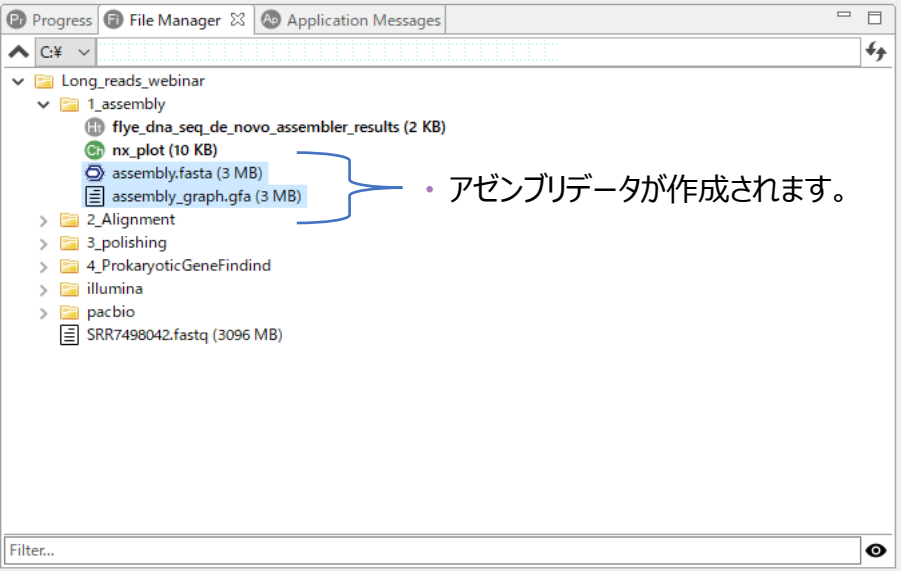
ロングリードデータを組み立てる

Filgen® biosciences & nanosciences



- 解析されたデータの保存先を指定
  Runをクリックすると解析がスタートします。

✓ ツールの論文の引用情報・リンクアクセスが可能です。



- アゼンブリデータが作成されます。

- 解析後、図・レポートが作成されます。

ロングリードデータを組み立てる

Flye DNA-Seq De Novo Assembler Results

Name: Flye output

Input: Sequencing Data

A total of 1 library has been processed.

Sample Name	Sequencing	File Name	Format
SRR7498042	PacBio	SRR7498042.fastq.gz	FASTQ

Results Overview

Statistic	assembly
Number of Contigs (>= 0 bp)	1
Number of Contigs (>= 1000 bp)	1
Number of Contigs (>= 5000 bp)	1
Number of Contigs (>= 10000 bp)	1
Number of Contigs (>= 25000 bp)	1
Number of Contigs (>= 50000 bp)	1
Total Length (>= 0 bp)	4,045,279
Total Length (>= 1000 bp)	4,045,279
Total Length (>= 5000 bp)	4,045,279
Total Length (>= 10000 bp)	4,045,279
Total Length (>= 25000 bp)	4,045,279
Total Length (>= 50000 bp)	4,045,279
Number of Contigs	1
Largest Contig	4,045,279
Total Length	4,045,279
GC (%)	43.94
N50	4,045,279
N75	4,045,279
L50	1
L75	1
Number of N's per 100 kbp	0

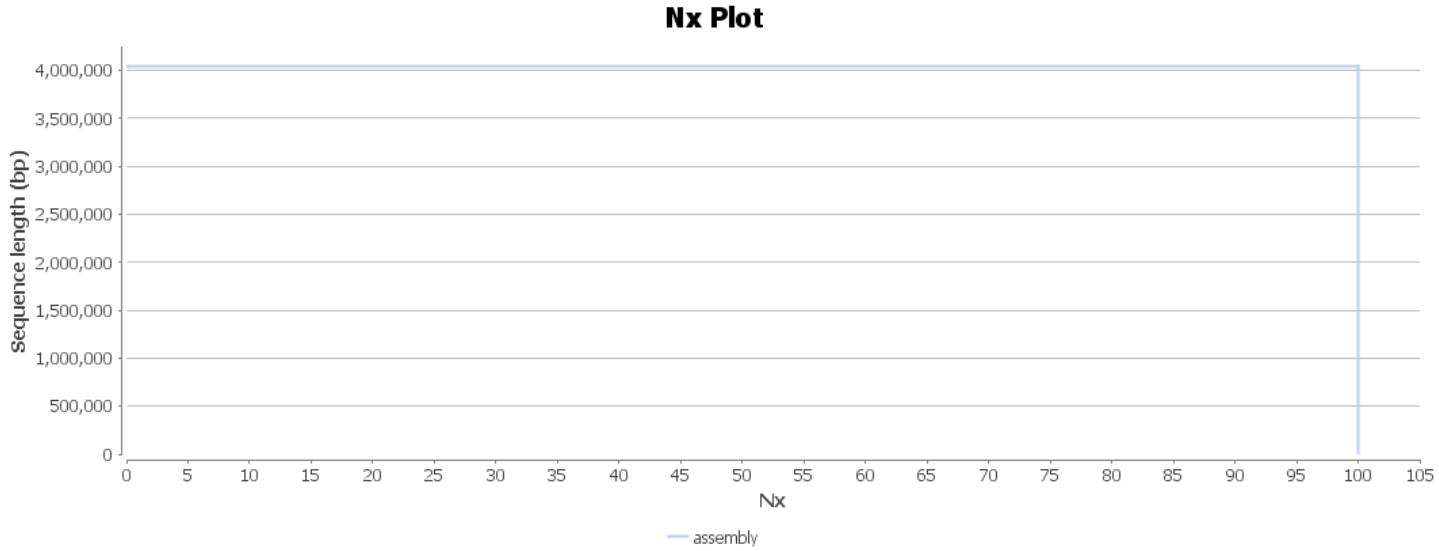
サマリーレポート

どのくらいの数のコンティグが作成できたか
短すぎるコンティグは多くないか？

予想される配列長に近いか？

アセンブリの品質の統計値

例)N50
作成されたコンティグを長い順に並べ順に足していった時
合計が全長の半分となるときのコンティグの長さ。
N50の値が高いほど、アセンブリが優れていることを示します。



Nxプロット

xが0から100%まで変化するときのNx値を示しています。
視覚的に品質を理解することができます。

✓ コンティグの品質を確かめて次の解析に進むことができます。

低いエラー率の配列を作成する

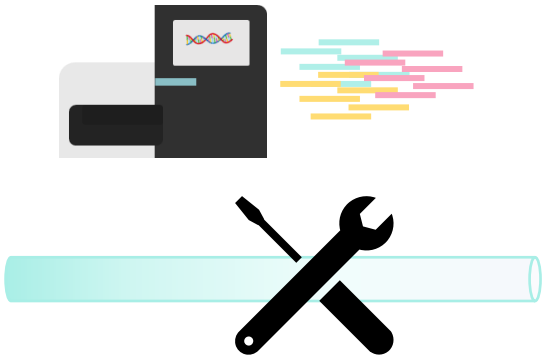
DNA-Seq
de Novo Assembly

DNA-Seq Alignment

DNA-Seq Polishing

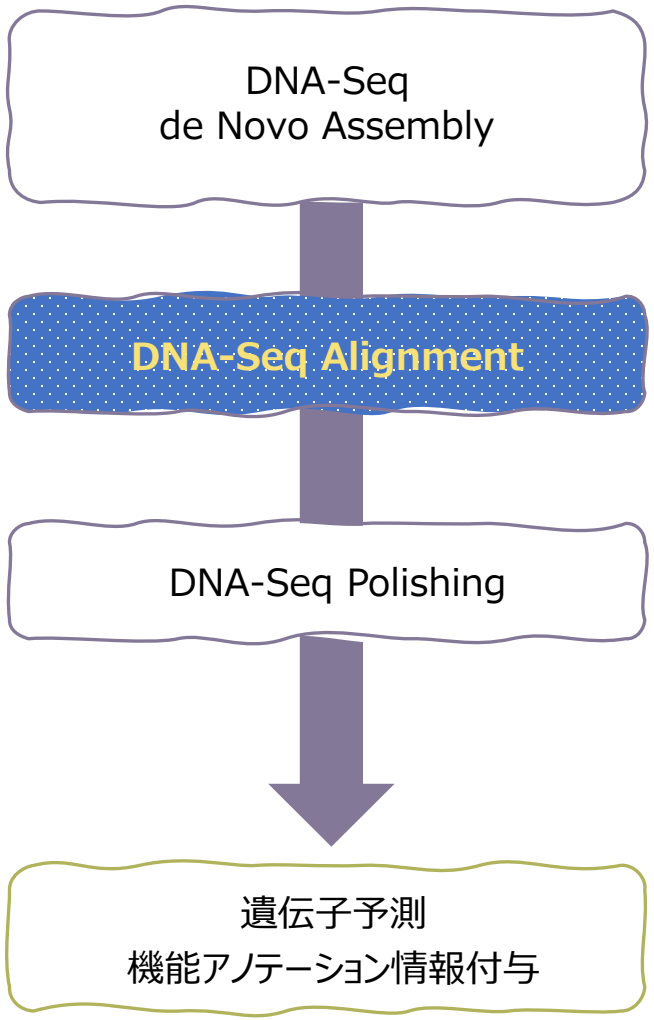
遺伝子予測
機能アノテーション情報付与

ロングリードは連続したゲノムアセンブリを作成できるが
エラー率が比較的高いため、
正確な最終シーケンスの生成が困難



ショートリードデータを使用することで、
修正とPolishingによりエラーを補正

✓ OmicsBoxにはde Novo Assemblyだけでなく
コンティグデータのエラーを補正できるツールが備わっています。



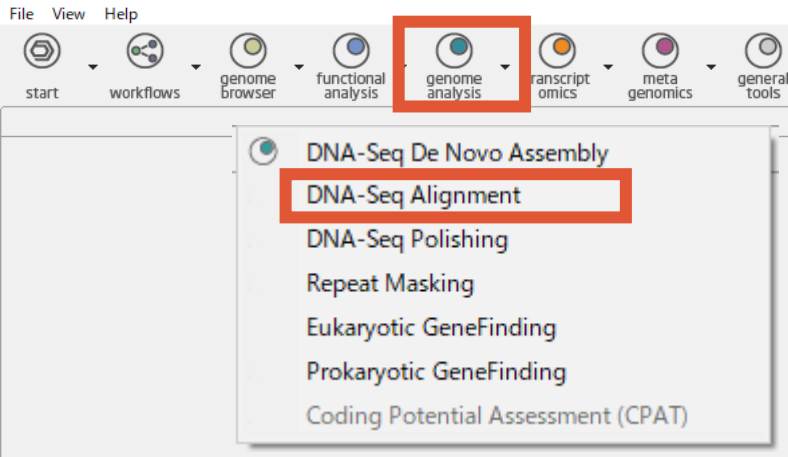
DNA-Seq Alignment

・シーケンスリードを参照ゲノムに効率的にマッピングして、
シーケンスリードのエラーを考慮しながら、リードの元となった「正しい」ゲノム
遺伝子座を特定することが本解析の目的です。

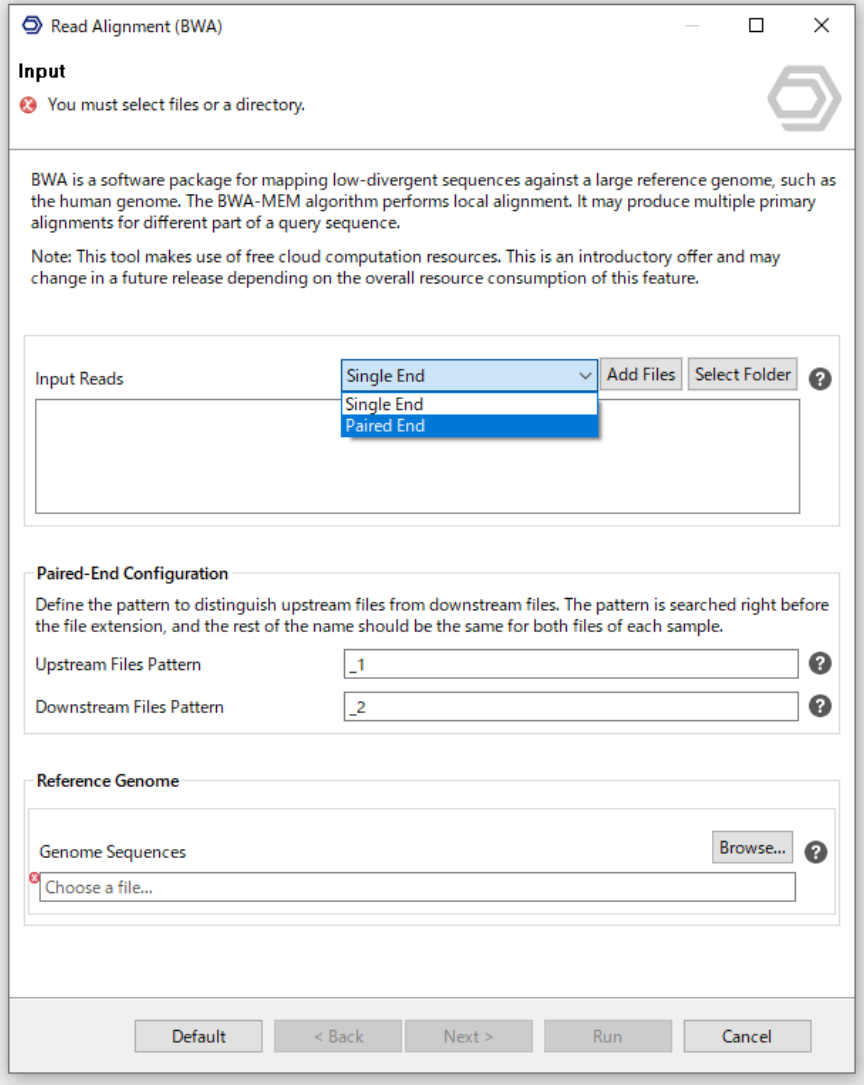
・ここでの参照ゲノムを前項で作成したコンティグデータにすることで
Polishingに必要なデータを作成することができます。

・本機能は、有名なリードアライメントパッケージ
BWA（Burrows-Wheeler Aligner）に基づいています。

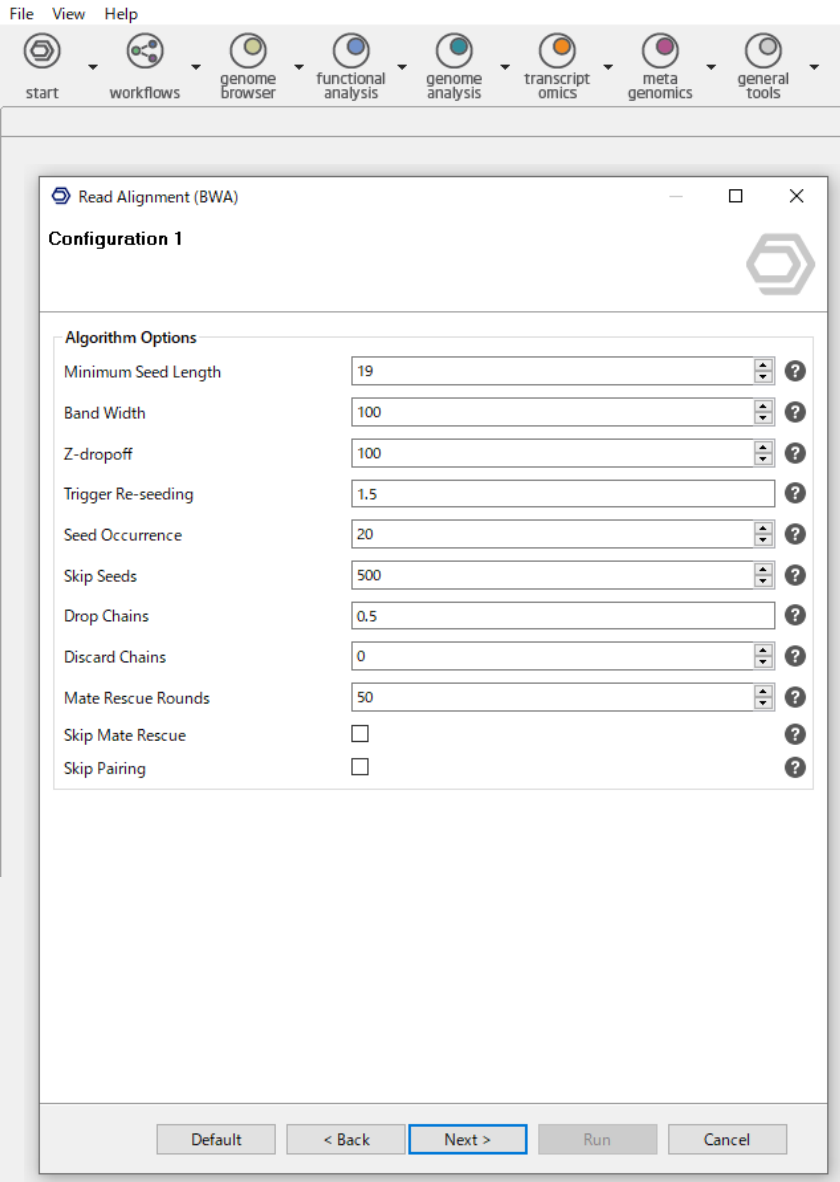
コンティグデータの修正とPolishing



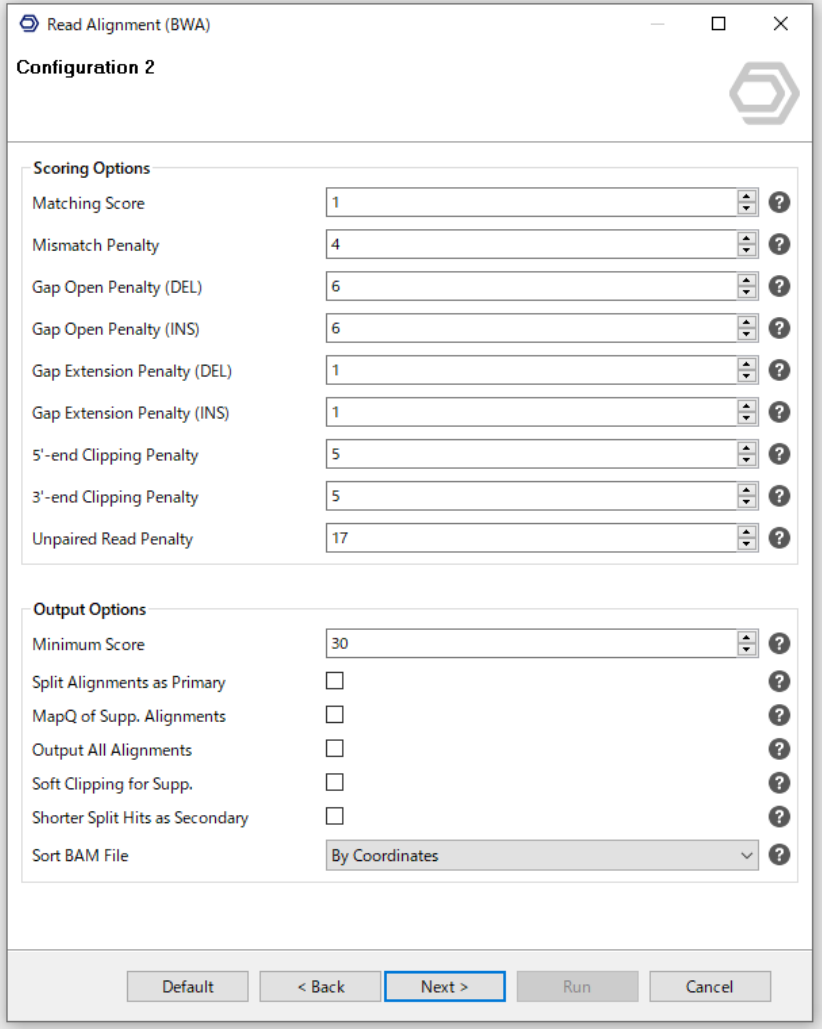
- この解析では作成したロングリードのコンティグデータに
  ショートリードをマッピングします。
- Inputにはショートリード、
  Referenceには作成したコンティグデータを指定



コンティグデータの修正とPolishing

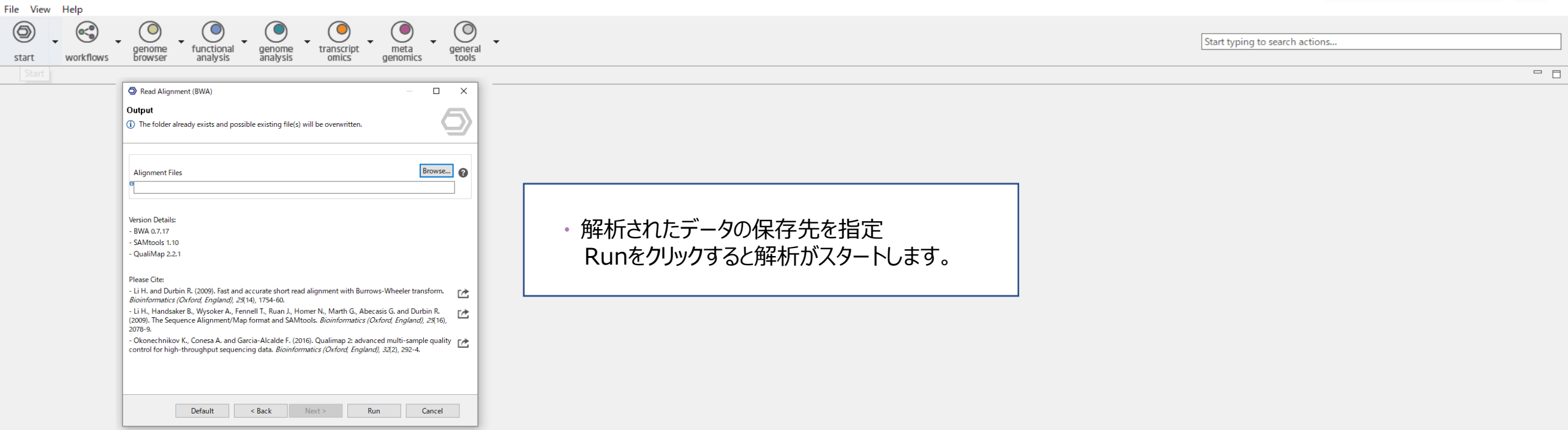


アルゴリズムオプション
（例：最小シード長、
バンド幅…）

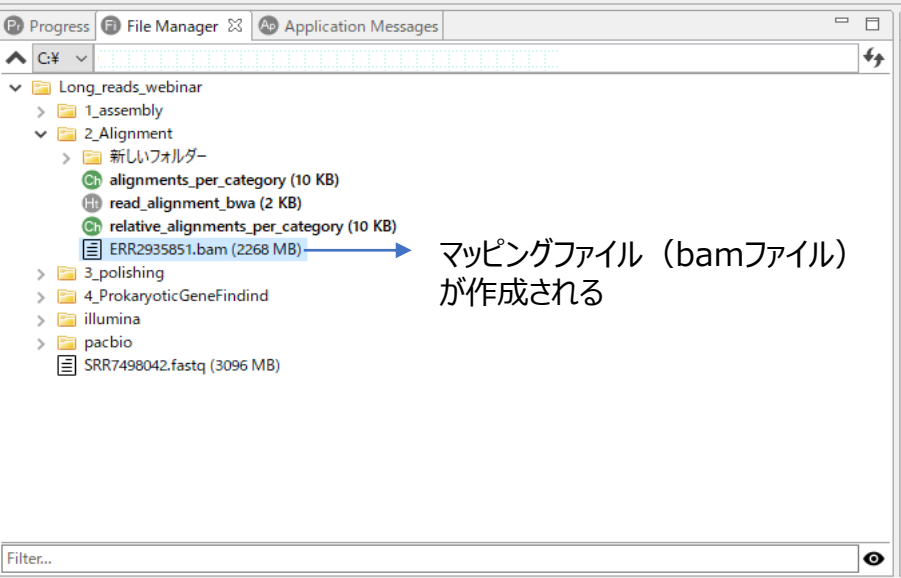


スコアオプション
（例：マッチングスコア、
ミスマッチペナルティ…）

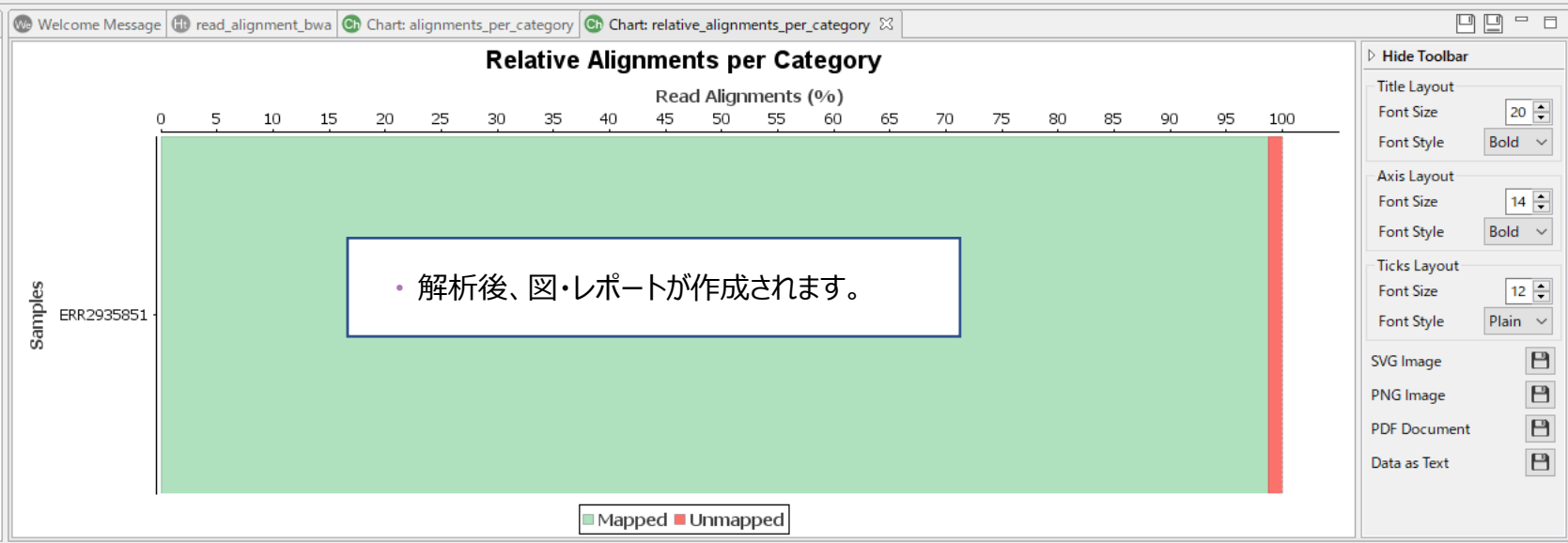
コンティグデータの修正とPolishing



- 解析されたデータの保存先を指定
  Runをクリックすると解析がスタートします。



マッピングファイル（bamファイル）が作成される



- 解析後、図・レポートが作成されます。

ロングリードデータを組み立てる

Read Alignment (BWA) Results

Input 1: Reference Genome Sequences

assembly

Sequences	Minimum Length	Maximum Length	Average Length	Total Length
1	4,045,279	4,045,279	4,045,279	4,045,279

Input 2: FASTQ Files

File Name	Sample Name	Format	Sequencing
ERR2935851_1.fastq.gz, ERR2935851_2.fastq.gz	ERR2935851	FASTQ	Paired End

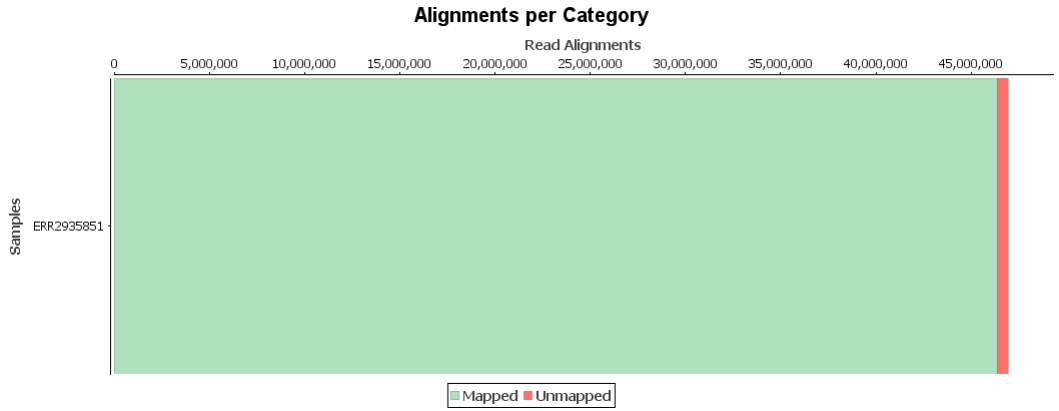
Results Overview

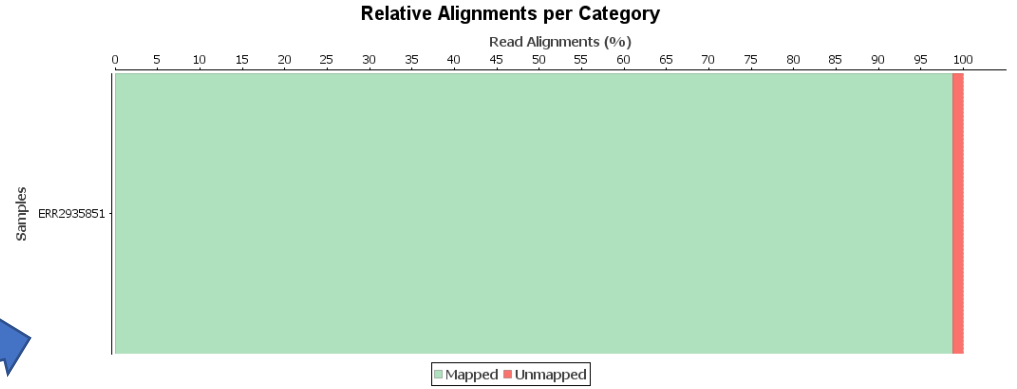
Globals

Sample	Total Alignments	Mapped	Supplementary	Unmapped	Duplicated Reads (estimated)	Duplication Rate
ERR2935851	46,952,430	46,358,042 / 98.734%	623,990 / 1.329%	594,388 / 1.266%	42,318,677 / 90.131%	99.4

どのくらいリードがマッピングされたか？

サマリーレポート

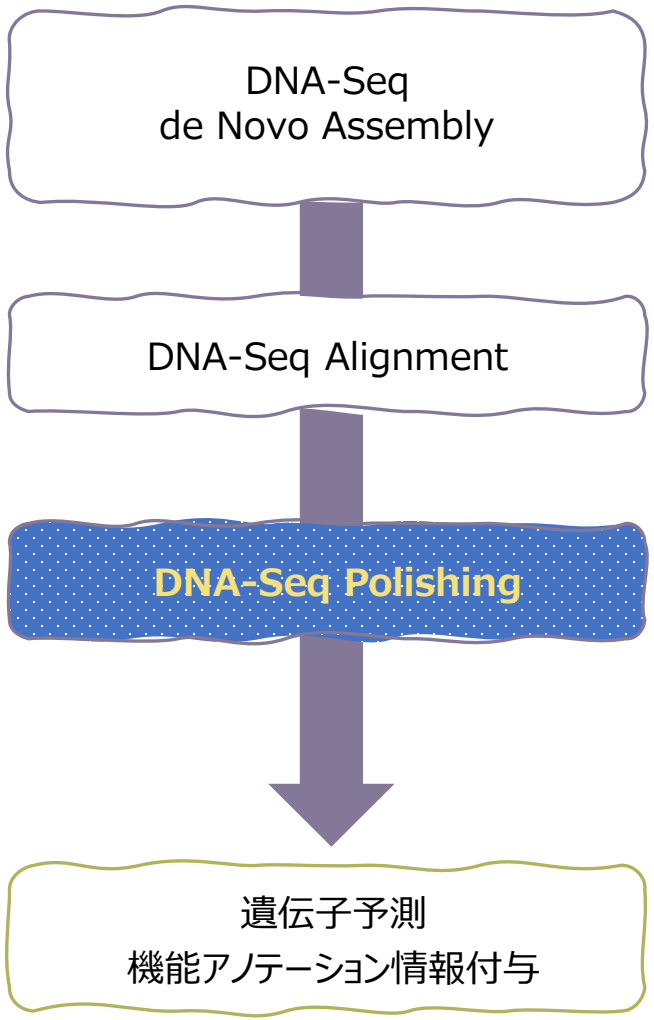




カテゴリーグラフ

上図：リード長
下図：%

✓ 結果を確かめて次の解析に進むことができます。



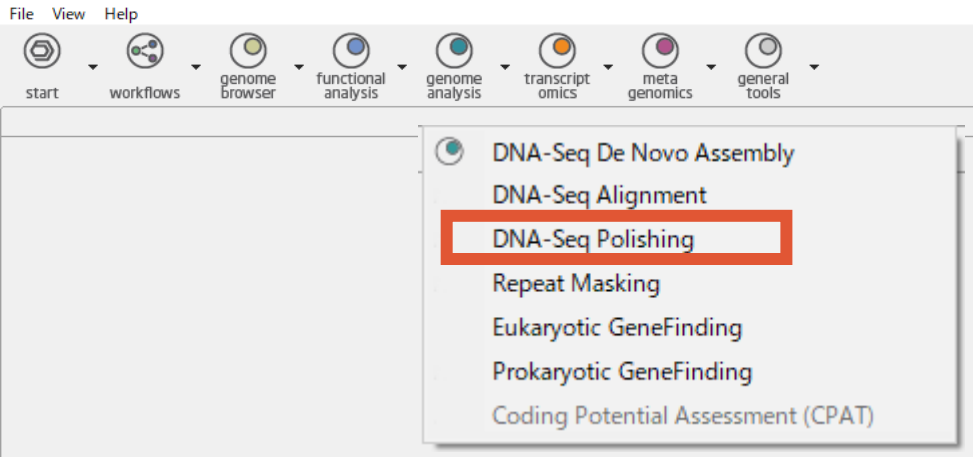
DNA-Seq Polishing

・誤ったアセンブリを修正し、ギャップを埋めることにより、
ドラフトゲノムアセンブリを大幅に改善します。

・エラーが少ない連続したゲノムを生成し、
より生物学的に関連性の高い遺伝子の同定を可能にします。

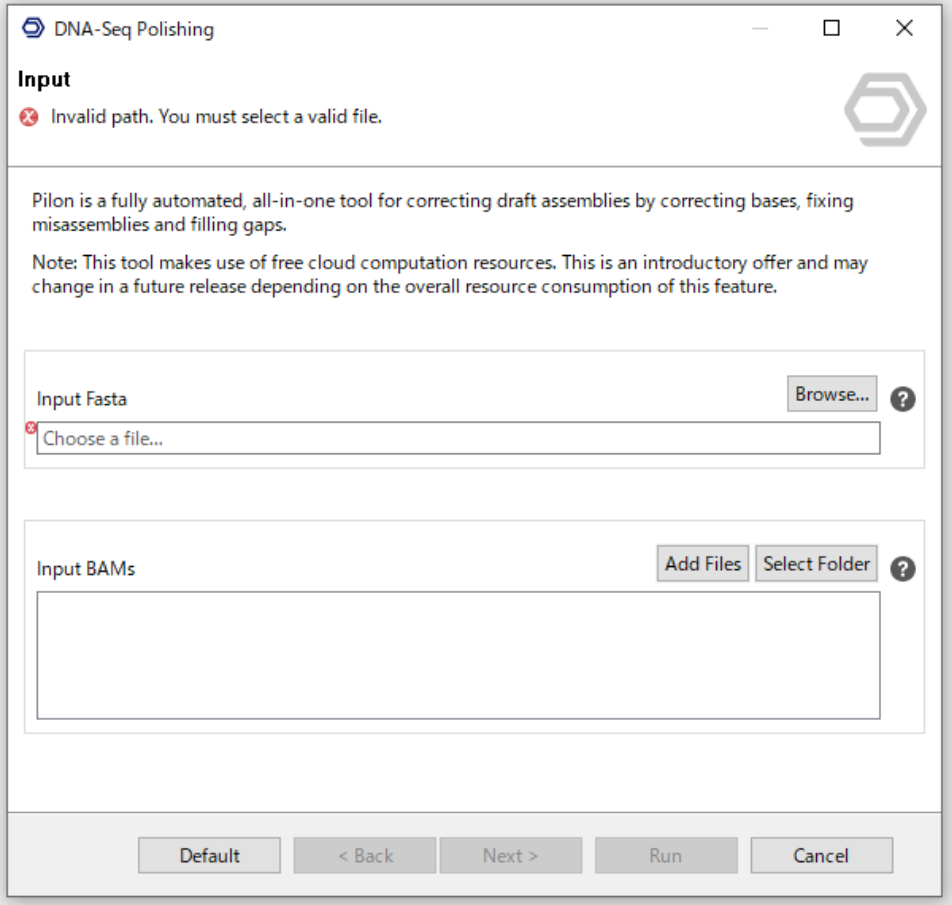
・本機能は、Pilonに基づいています。

コンティグデータの修正とPolishing

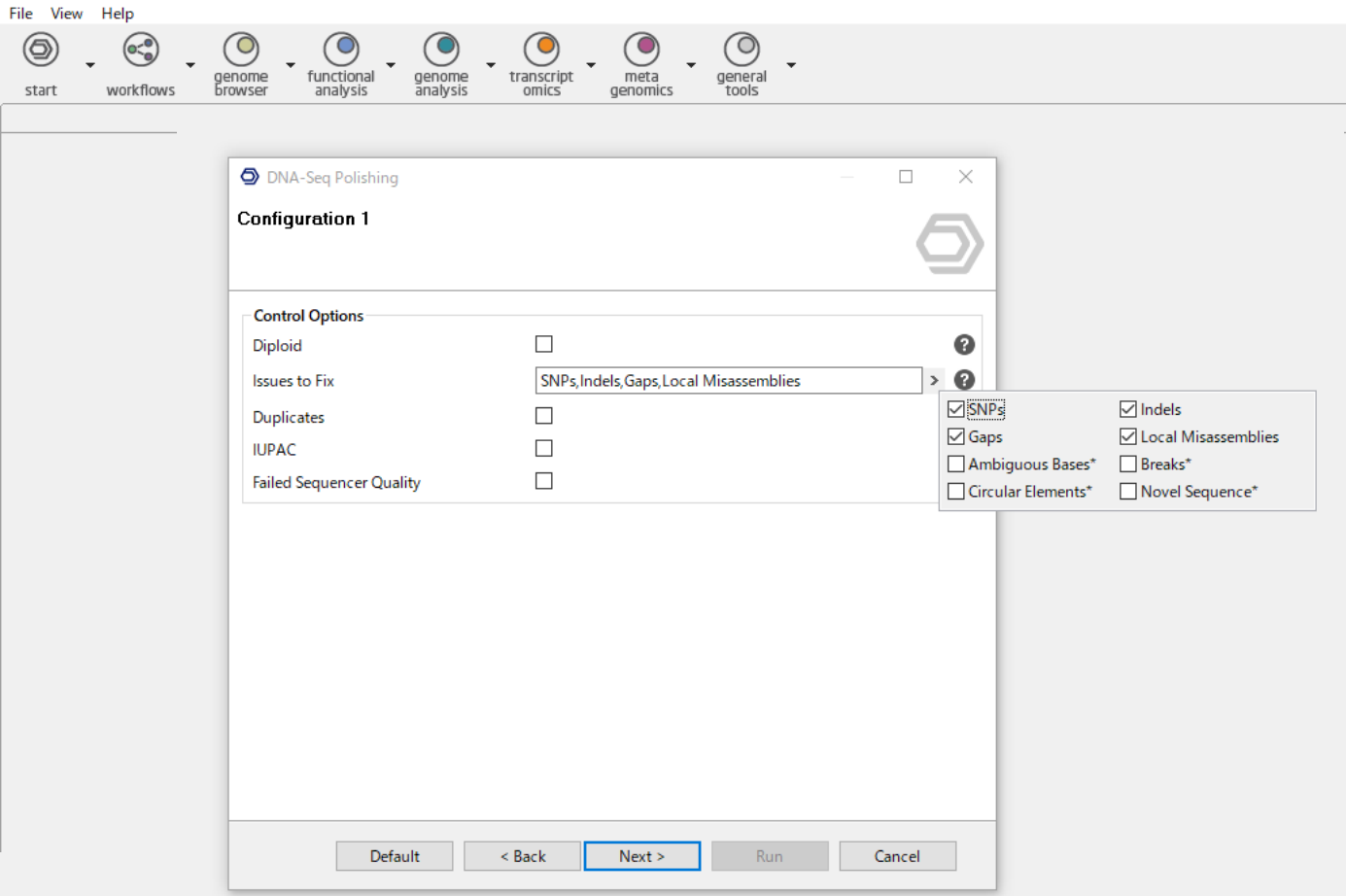


- この解析では作成したマッピングファイルを使用して
  ロングリードのコンティグデータをPolishingします。

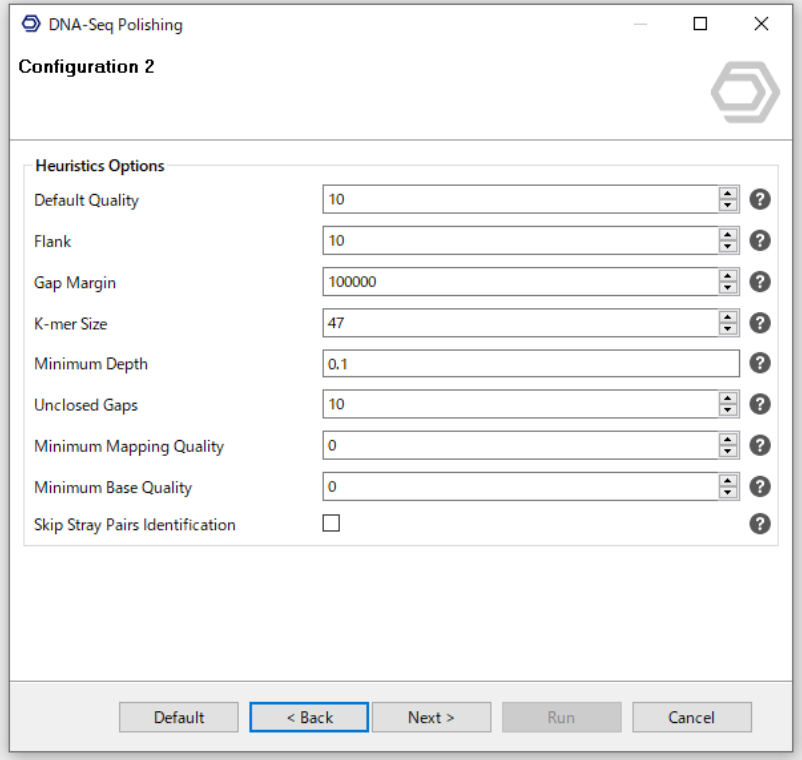
- Input Fastaはコンティグデータ（DNA-Seq de Novo Assemblyの結果）
  Input BAMsにはマッピングデータ（DNA-Seq Alignmentの結果）



コンティグデータの修正とPolishing

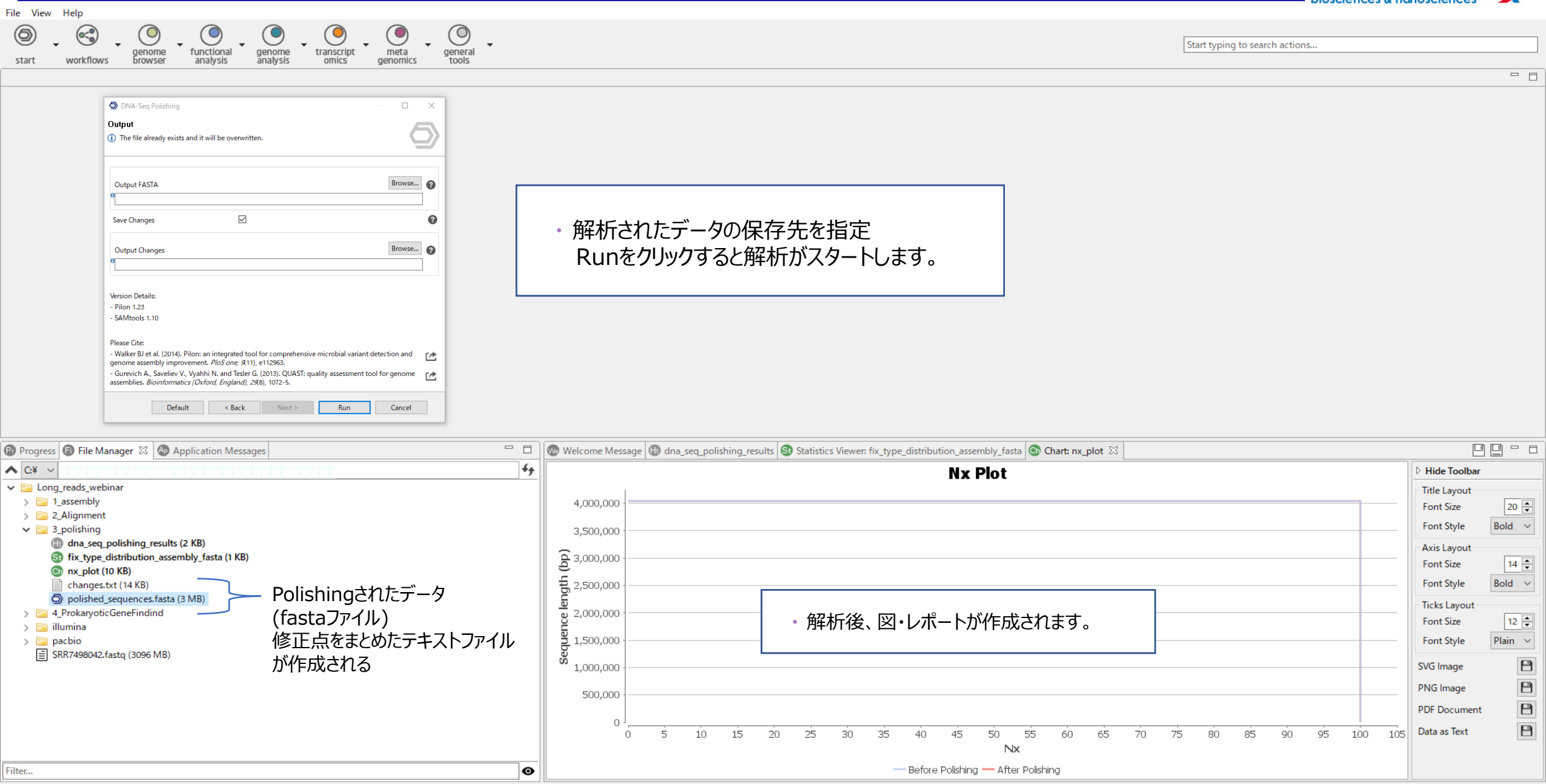


修正を行うカテゴリーについて選択します。
(e.g. SNPs, indels, gaps, local misassemblies…)



アルゴリズムの調整を行うことも可能

コンティグデータの修正とPolishing



- 解析されたデータの保存先を指定
  Runをクリックすると解析がスタートします。

Polishingされたデータ
(fastaファイル)
修正点をまとめたテキストファイル
が作成される

- 解析後、図・レポートが作成されます。

コンティグデータの修正とPolishing

Results Overview

Total number of changes: 345

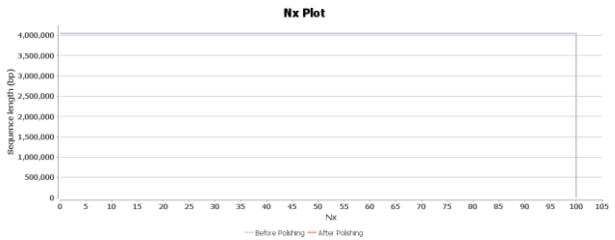
- Number of single nucleotide changes: 0
- Number of insertions: 3β9
- Number of deletions: 6
- Number of segmental changes: 0

Polishingにより改善された情報

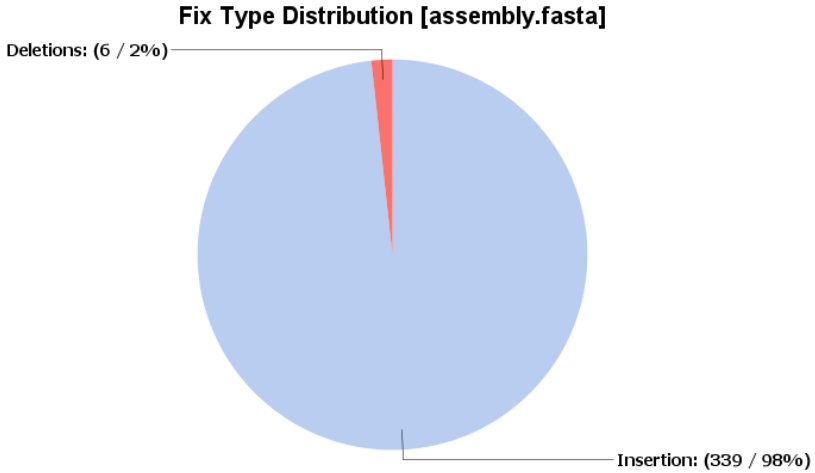
Statistic	Before Polishing	After Polishing
Number of Contigs (>= 0 bp)	1	1
Number of Contigs (>= 1000 bp)	1	1
Number of Contigs (>= 5000 bp)	1	1
Number of Contigs (>= 10000 bp)	1	1
Number of Contigs (>= 25000 bp)	1	1
Number of Contigs (>= 50000 bp)	1	1
Total Length (>= 0 bp)	4,045,279	4,045,615
Total Length (>= 1000 bp)	4,045,279	4,045,615
Total Length (>= 5000 bp)	4,045,279	4,045,615
Total Length (>= 10000 bp)	4,045,279	4,045,615
Total Length (>= 25000 bp)	4,045,279	4,045,615
Total Length (>= 50000 bp)	4,045,279	4,045,615
Number of Contigs	1	1
Largest Contig	4,045,279	4,045,615
Total Length	4,045,279	4,045,615
GC (%)	43.94	43.94
N50	4,045,279	4,045,615
N75	4,045,279	4,045,615
L50	1	1
L75	1	1
Number of N's per 100 kbp	0	0

サマリーレポート

DNA-Seq de Novo Assemblyと
見方は同じ



Nxプロットも同時に作成される



修正された総数・割合・内訳が図にて表示可能

✓ エラー率の少ない配列データを作成できます

DNA-Seq
de Novo Assembly

DNA-Seq Alignment

DNA-Seq Polishing

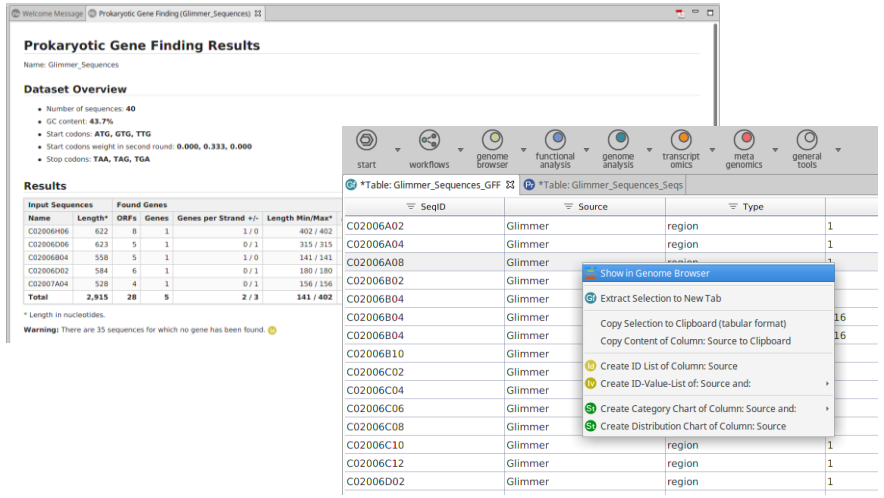
遺伝子予測
機能アノテーション情報付与

遺伝子予測

・作成したコンティグ配列を用いて、配列内の遺伝子やタンパク質コード領域を予測

・原核生物と真核生物解析用の2種類のプログラムを搭載

・遺伝子領域の詳細をまとめたGFFファイルとCDS配列リスト、
さらに解析結果をまとめたレポート



フィルジェン WEBセミナー

**DNA配列を組み立て
遺伝子の位置と構造を予測する**

～OmicsBoxを使用したNGS解析機能のご紹介～

詳しくはこちら
フィルジェン Webセミナー(2019年10月10日)

DNA-Seq
de Novo Assembly

DNA-Seq Alignment

DNA-Seq Polishing

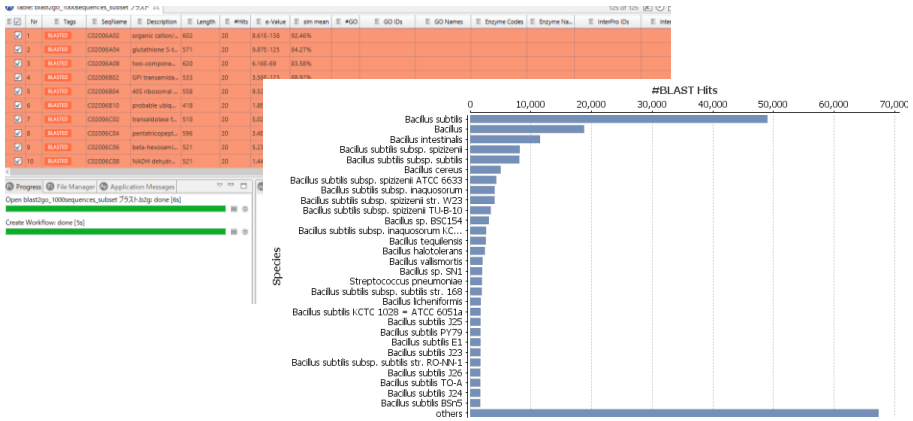
遺伝子予測
機能アノテーション情報付与

機能アノテーション情報付与

・BLASTを実行すると、CDS配列リスト(遺伝子予測の結果のデータ)に
各配列のトップヒット配列の情報が表示される

・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
推定できる

・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。



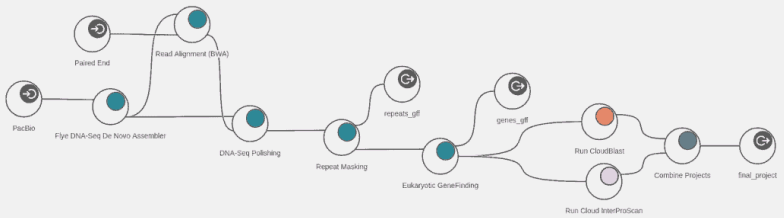
詳しくはこちら
フィルジェン Webセミナー(2019年4月18日)

まとめ

✓ OmicsBoxに完全に実装された戦略により、複雑なコマンドライン操作なしでロングリードアセンブリを実行できます。
さらにロングリードとショートリードを使用する戦略により、全体的なエラー率が非常に低い非常に連続したアセンブリが作成できます。

簡単に解析

✓実績の高いオープンソースソフトウェアを
簡単に使用することができます



OmicsBoxではワークフロー機能もあり、
今回解析した内容を1度に実行することも可能

オリジナルのWorkflow作成や既に組み立てられたWorkflow
を選択できます。（左図は既存のLong Reads Workflow）

高速計算

✓インターネット接続さえあれば高価なPCは必要ありません

1.4時間

解析時間

サンプルデータ量：約4.2M

DNA-Seq de Novo Assembly：約1時間
DNA-Seq Alignment：約30分
DNA-Seq Polishing：約10分

まとめ

Functional Analysis

機能アノテーションツール

・BlastとInterProの高速解析
・機能アノテーション情報付与
・エンリッチメント解析

Genome Analysis

新規ゲノムの配列決定

・De-Novo Assembly
・Repeat Masking
・真核生物 原核生物のORF領域の
遺伝子予測

Transcriptomics

RNA-seqデータ解析

・De-Novo Assembly
・発現値定量（モデル生物・非モデル生物対応）
・発現変動遺伝子の同定

Metagenomics

メタゲノム解析
（16S・WGS）

・Toxonomic Classification
・OTU Abundances Table
・Metagenomic Assembly/遺伝子予測

※1ライセンスの場合です。同時解析は1台となります。

2台のPCにインストールできるので
在宅勤務でも強力なデータ解析を実行できます。※

Information

弊社HPでの紹介ページ

カタログ　1週間使用可能なデモライセンスがあります。
ご希望の場合はbiosupport@filgen.jpまでお問い合わせください。

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