



NGSデータ解析：ロングリードアセンブリ

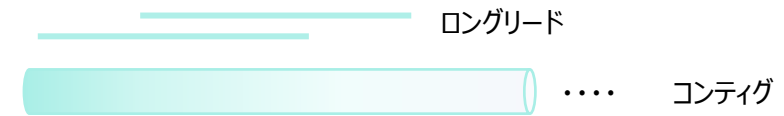
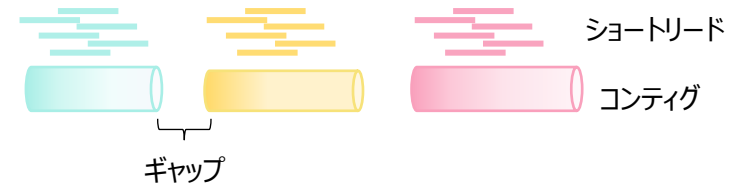
～OmicsBoxを使用した解析機能のご紹介～

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)

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。

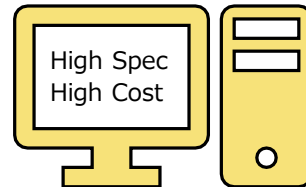
→全ゲノムシーケンスプロジェクトで使用すると、
リピート配列のシーケンスや連続したゲノムアセンブリを生成!!



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



✓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。

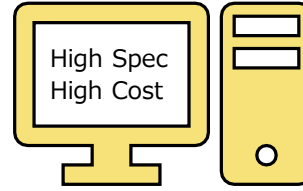
実績は高いがコマンドライン型であったりOSに制限がある
オープンソースソフトウェアを多数組み込みマウス操作で簡単に
解析できるようしたのがOmicsBoxの特徴の1つです。

✓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。



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

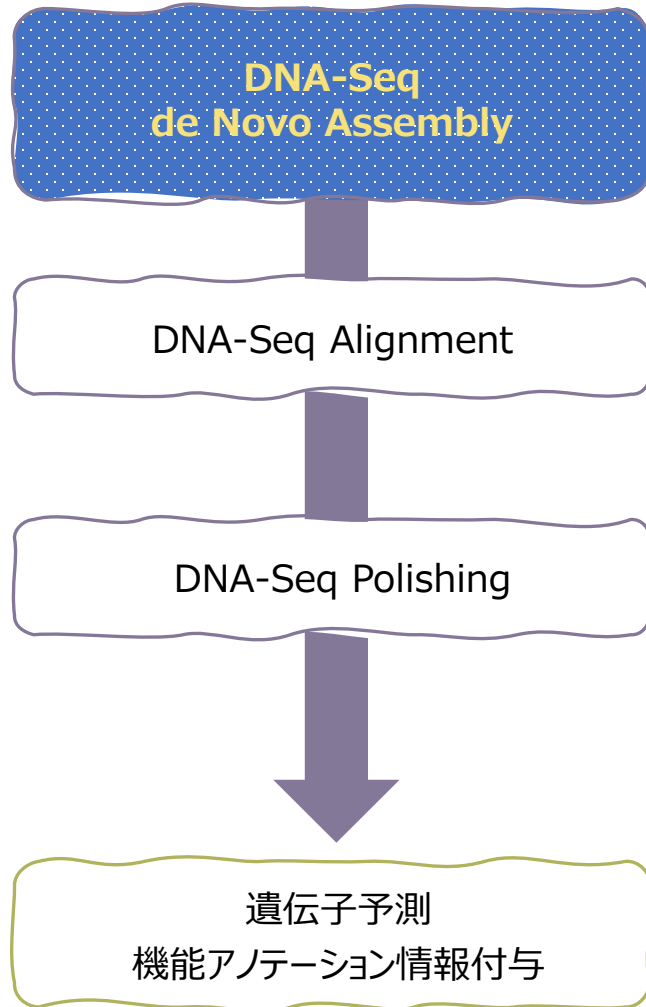


有名なアルゴリズムを使用した解析は、
コマンドライン操作が必要

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

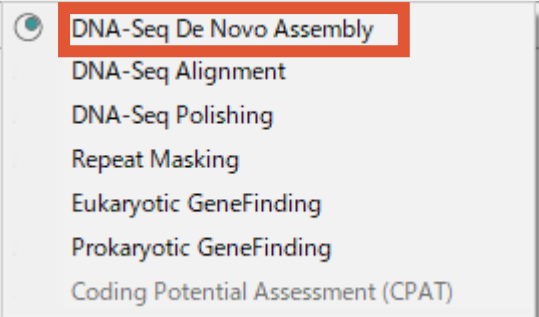
- 参照ゲノム配列なしで1からゲノムを再構築します。
- シークエンサーから出力したリードから長い連続したシーケンス（コンティグ）を作成することが本解析の目的です。
- 現在(2020.06)3つのアセンブリ戦略を実行できます。
本セミナーではロングリードの解析をご紹介します。



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。

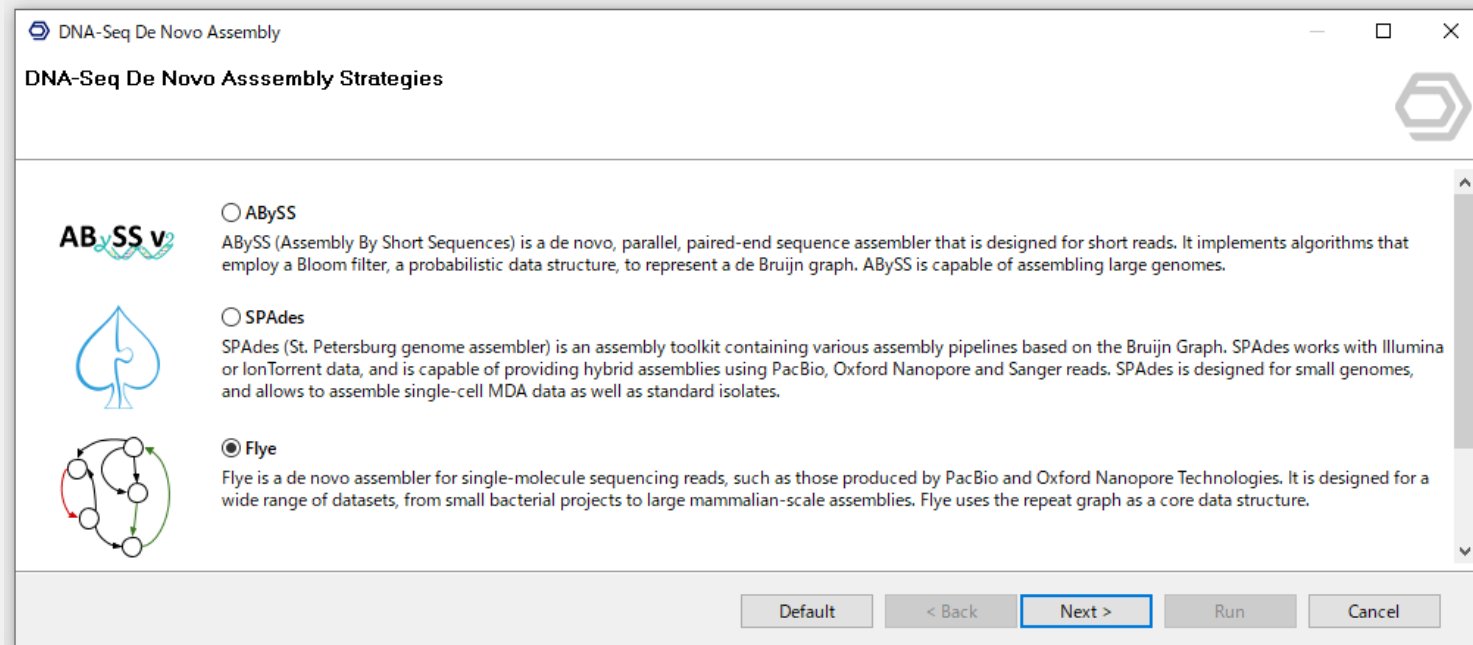
ロングリードデータを組み立てる

File View Help



• 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。



• アセ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

File View Help



Flye DNA-Seq De Novo Assembler

Input

You must select files or a directory.

Flye is a de novo assembler for single molecule sequencing reads, such as those produced by PacBio and Oxford Nanopore Technologies. It is designed for a wide range fo datasets, from small bacterial projects to large mammalian-scale assemblies. Flye is using repeat graph as a core data structure.

Note: This tool makes use of free cloud computation resources. This is an introductory offer and may change in a future release depending on the overall resource consumption of this feature.

Input Reads Add Files Select Folder ?

Error Corrected Data ?

Default < Back Next > Run Cancel

- 次世代シーケ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

File View Help



Flye DNA-Seq De Novo Assembler

Configuration

Genome Size	<input type="text" value="4.2m"/>	?
Automatic Minimum Overlap	<input checked="" type="checkbox"/>	?
Manual Minimap Overlap	<input type="text" value="3000"/>	?
Polishing	<input checked="" type="checkbox"/>	?
Number of Polishing Iterations	<input type="text" value="1"/>	?
Plasmids	<input type="checkbox"/>	?
Keep Haplotypes	<input type="checkbox"/>	?
Trestle	<input type="checkbox"/>	?

Default < Back **Next >** Run Cancel

- 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

i Provide an estimate of the size of the genome. Common suffixes are allowed, for example "m" (mega) or "g" (giga). The genome size estimate is used to decide how many reads to correct and how sensitive the overlapper should be.

OK

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

File View Help



Start typing to search actions...

Flye DNA-Seq De Novo Assembler

Output
Invalid path. You must select a valid file.

Assembly Fasta

Save Graph File

Graph File

Version Details:
Flye 2.7b-b1528

Please Cite:
- Kolmogorov M., Yuan J., Lin Y. and Pevzner PA. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature biotechnology*, 37(5), 540-546.
- Gurevich A., Saveliev V., Vyahhi N. and Tesler G. (2013). QUASt: quality assessment tool for genome assemblies. *Bioinformatics (Oxford, England)*, 29(8), 1072-5.

• 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。



Copy to Clipboard
Open in Browser

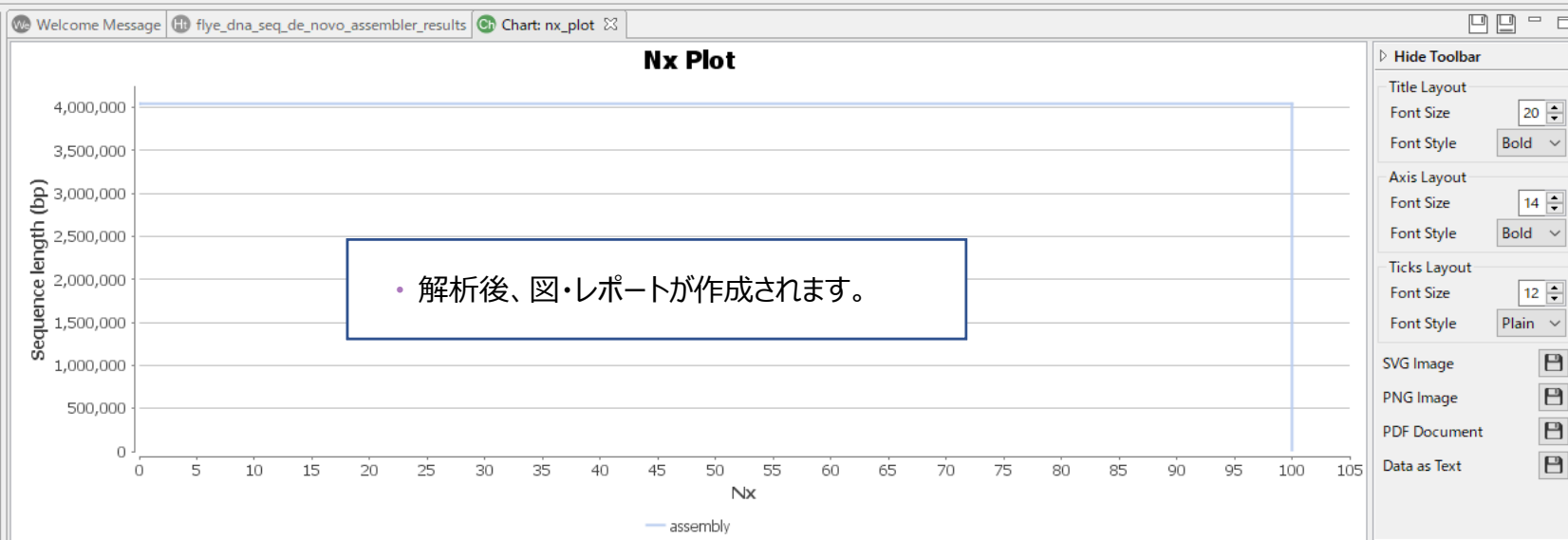
✓ ツールの論文の引用情報・リンクアクセスが可能です。

Progress File Manager Application Messages

Long_reads_webinar

- 1_assembly
 - flye_dna_seq_de_novo_assembler_results (2 KB)
 - nx_plot (10 KB)
 - assembly.fasta (3 MB)
 - assembly_graph.gfa (3 MB)
- 2_Alignment
- 3_polishing
- 4_ProkaryoticGeneFindind
- illumina
- pacbio
- SRR7498042.fastq (3096 MB)

• アセ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。

Hide Toolbar

Title Layout
Font Size 20
Font Style Bold

Axis Layout
Font Size 14
Font Style Bold

Ticks Layout
Font Size 12
Font Style Plain

SVG Image

PNG Image

PDF Document

Data as Text

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

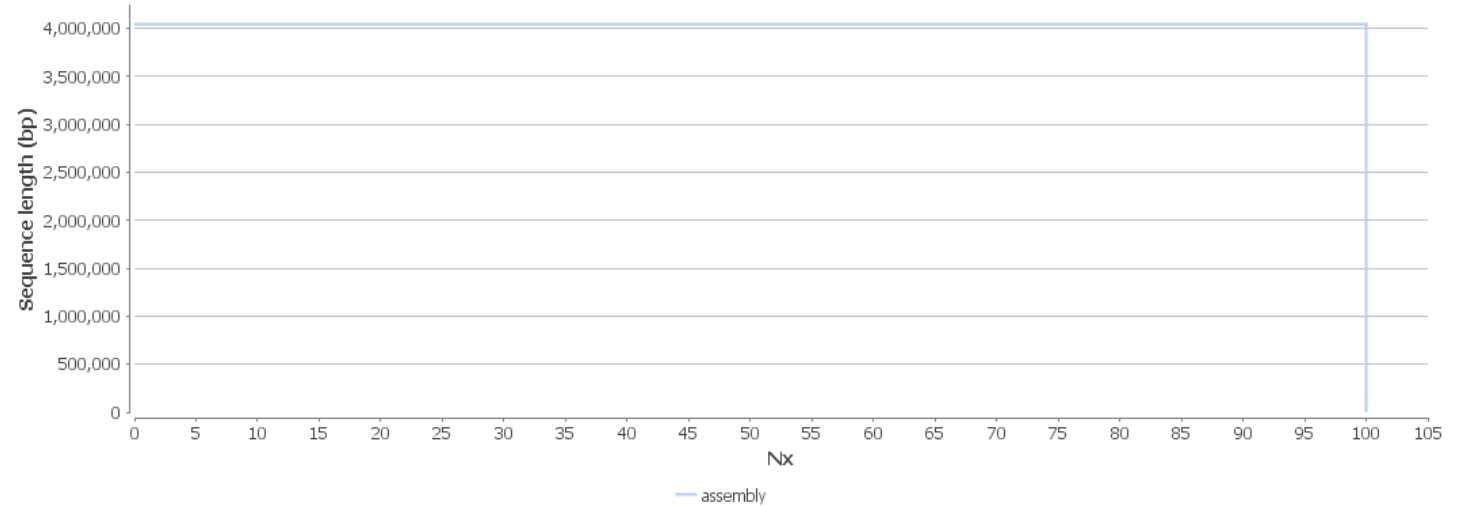
どのくらいの数のコンティグが作成できたか
短すぎるコンティグは多くないか？

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 Plot

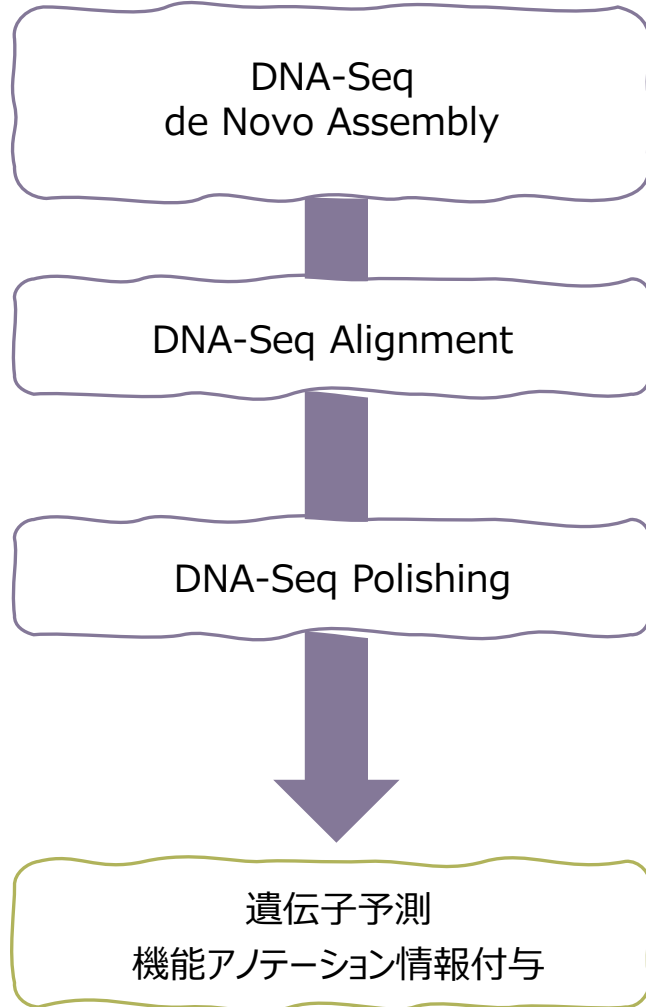


Nxプロット

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。

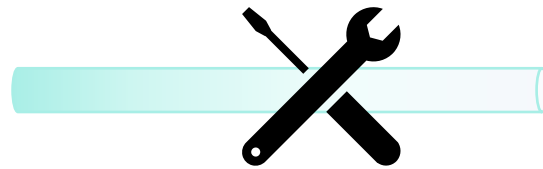
サマリーレポート



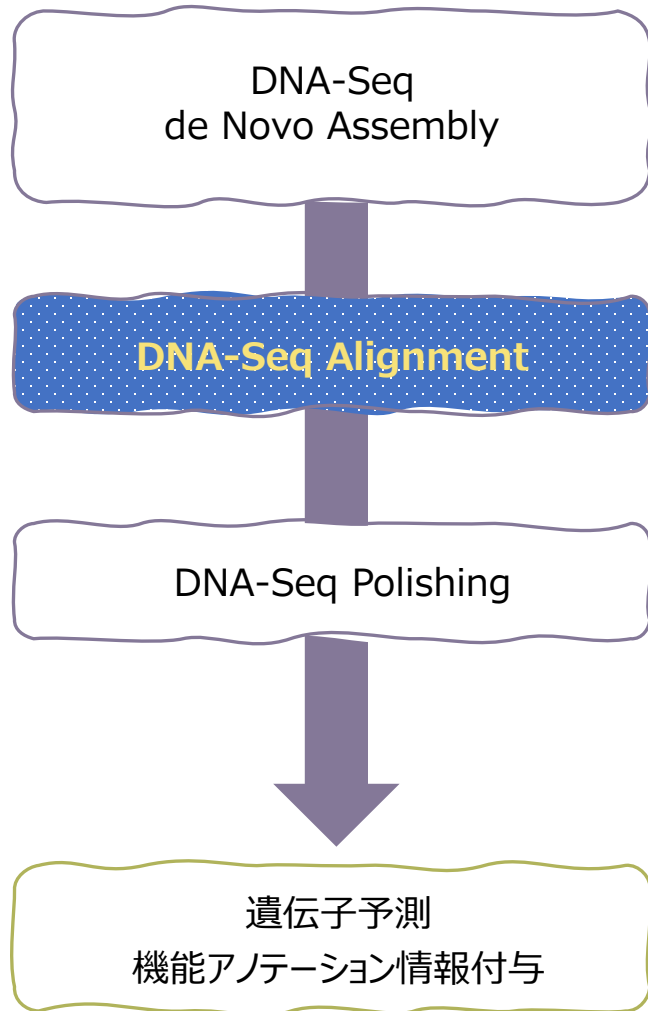
ロングリードは連続したゲノムアセンブリを作成できるがエラー率が比較的高いため、正確な最終シーケンスの生成が困難



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

File View Help



- DNA-Seq De Novo Assembly
- DNA-Seq Alignment**
- DNA-Seq Polishing
- Repeat Masking
- Eukaryotic GeneFinding
- Prokaryotic GeneFinding
- Coding Potential Assessment (CPAT)

- 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

Read Alignment (BWA)

Input

✖ You must select files or a directory.

BWA is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome. The BWA-MEM algorithm performs local alignment. It may produce multiple primary alignments for different part of a query sequence.

Note: This tool makes use of free cloud computation resources. This is an introductory offer and may change in a future release depending on the overall resource consumption of this feature.

Input Reads Single End Add Files Select Folder ?

Single End
Paired End

Paired-End Configuration

Define the pattern to distinguish upstream files from downstream files. The pattern is searched right before the file extension, and the rest of the name should be the same for both files of each sample.

Upstream Files Pattern ?

Downstream Files Pattern ?

Reference Genome

Genome Sequences Browse... ?

✖ Choose a file...

Default < Back Next > Run Cancel

コンティグデータの修正とPolishing

File View Help



Read Alignment (BWA)

Configuration 1

Algorithm Options

Minimum Seed Length	19
Band Width	100
Z-dropoff	100
Trigger Re-seeding	1.5
Seed Occurrence	20
Skip Seeds	500
Drop Chains	0.5
Discard Chains	0
Mate Rescue Rounds	50
Skip Mate Rescue	<input type="checkbox"/>
Skip Pairing	<input type="checkbox"/>

Default < Back Next > Run Cancel

アルゴリズムオプション
(例：最小シード長、
バンド幅…)

Read Alignment (BWA)

Configuration 2

Scoring Options

Matching Score	1
Mismatch Penalty	4
Gap Open Penalty (DEL)	6
Gap Open Penalty (INS)	6
Gap Extension Penalty (DEL)	1
Gap Extension Penalty (INS)	1
5'-end Clipping Penalty	5
3'-end Clipping Penalty	5
Unpaired Read Penalty	17

Output Options

Minimum Score	30
Split Alignments as Primary	<input type="checkbox"/>
MapQ of Supp. Alignments	<input type="checkbox"/>
Output All Alignments	<input type="checkbox"/>
Soft Clipping for Supp.	<input type="checkbox"/>
Shorter Split Hits as Secondary	<input type="checkbox"/>
Sort BAM File	By Coordinates

Default < Back Next > Run Cancel

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

File View Help



Start typing to search actions...

Read Alignment (BWA)

Output
① The folder already exists and possible existing file(s) will be overwritten.

Alignment Files Browse...

Version Details:
- BWA 0.7.17
- SAMtools 1.10
- QualiMap 2.2.1

Please Cite:
- Li H. and Durbin R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14), 1754-60.
- Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G. and Durbin R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), 2078-9.
- Okonechnikov K., Conesa A. and Garcia-Alcalde F. (2016). QualiMap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics (Oxford, England)*, 32(2), 292-4.

Default < Back Next > Run Cancel

- 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。

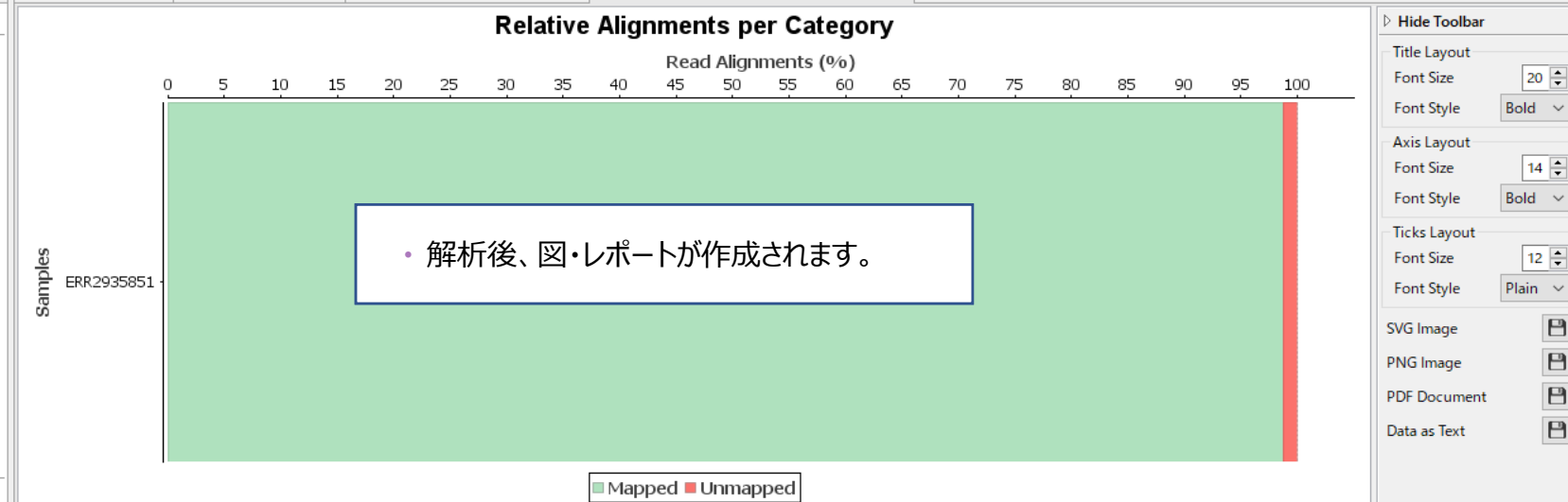
Progress File Manager Application Messages

C:\>

- Long_reads_webinar
 - 1_assembly
 - 2_Alignment
 - 新しいフォルダー
 - alignments_per_category (10 KB)
 - read_alignment_bwa (2 KB)
 - relative_alignments_per_category (10 KB)
 - ERR2935851.bam (2268 MB) → マッピングファイル (bamファイル) が作成される
 - 3_polishing
 - 4_ProkaryoticGeneFindind
 - illumina
 - pacbio
 - SRR7498042.fastq (3096 MB)

Filter...

Welcome Message read_alignment_bwa Chart: alignments_per_category Chart: relative_alignments_per_category



Hide Toolbar

Title Layout
Font Size: 20
Font Style: Bold

Axis Layout
Font Size: 14
Font Style: Bold

Ticks Layout
Font Size: 12
Font Style: Plain

SVG Image

PNG Image

PDF Document

Data as Text

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

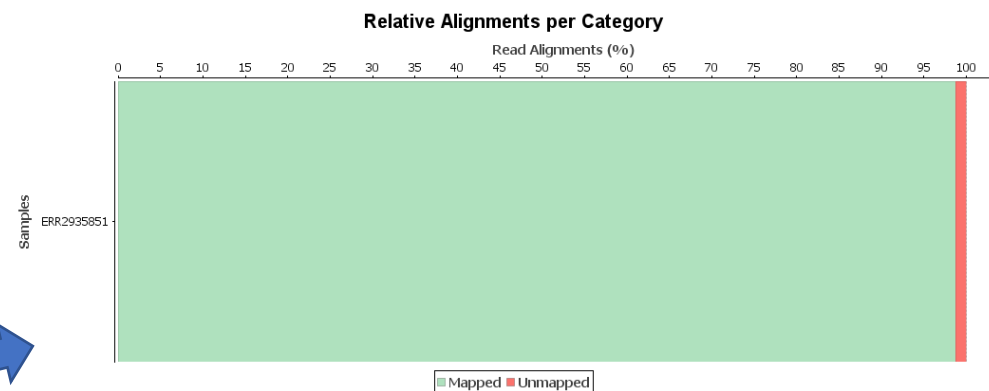
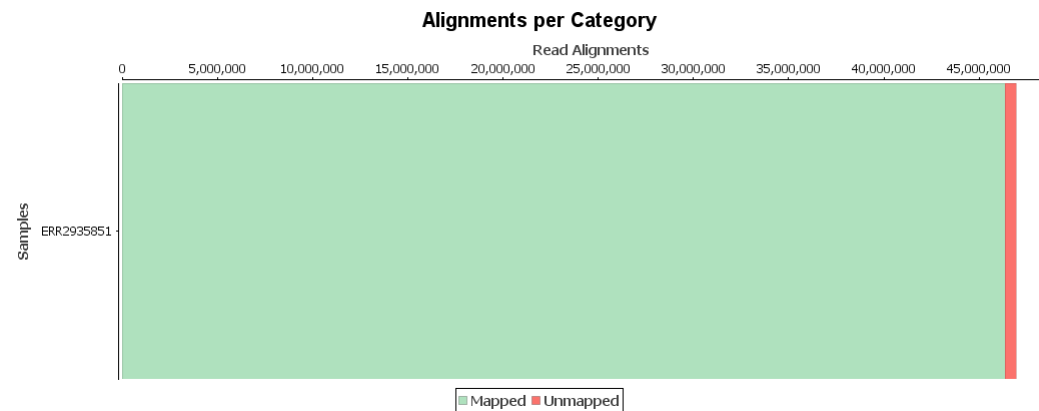
Results Overview

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？

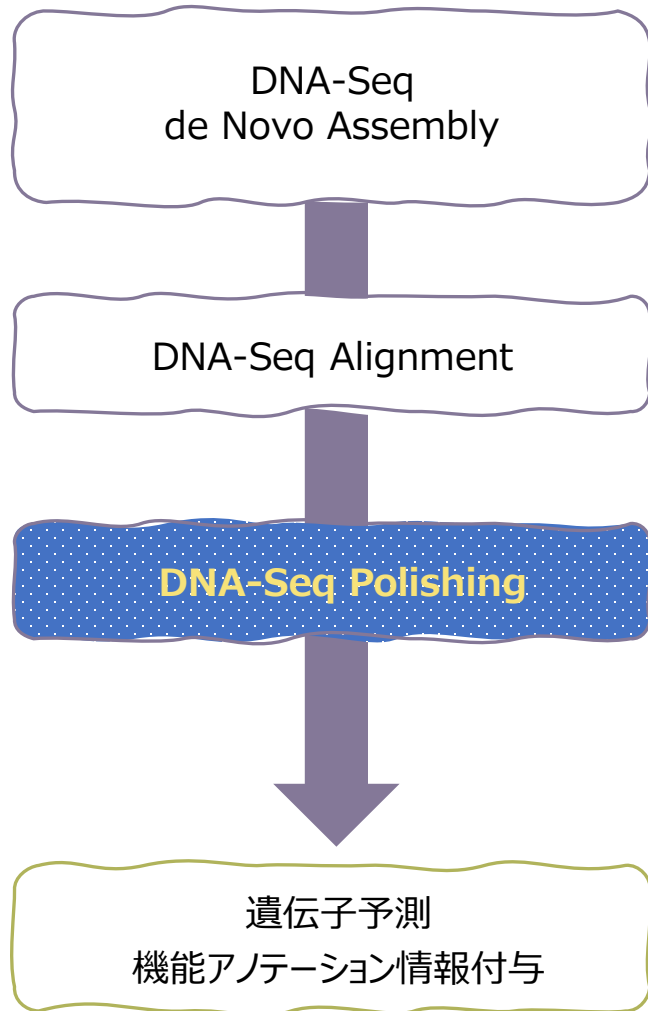
サマリーレポート



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

File View Help



- DNA-Seq De Novo Assembly
- DNA-Seq Alignment
- DNA-Seq Polishing**
- Repeat Masking
- Eukaryotic GeneFinding
- Prokaryotic GeneFinding
- Coding Potential Assessment (CPAT)

- 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)

DNA-Seq Polishing

Input

Invalid path. You must select a valid file.

Pilon is a fully automated, all-in-one tool for correcting draft assemblies by correcting bases, fixing misassemblies and filling gaps.

Note: This tool makes use of free cloud computation resources. This is an introductory offer and may change in a future release depending on the overall resource consumption of this feature.

Input Fasta Browse... ?

Input BAMs Add Files Select Folder ?

Default < Back Next > Run Cancel

コンティグデータの修正とPolishing

File View Help



DNA-Seq Polishing

Configuration 1

Control Options

Diploid

Issues to Fix

Duplicates

IUPAC

Failed Sequencer Quality

SNPs Indels
 Gaps Local Misassemblies
 Ambiguous Bases* Breaks*
 Circular Elements* Novel Sequence*

Default < Back Next > Run Cancel

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

DNA-Seq Polishing

Configuration 2

Heuristics Options

Default Quality

Flank

Gap Margin

K-mer Size

Minimum Depth

Unclosed Gaps

Minimum Mapping Quality

Minimum Base Quality

Skip Stray Pairs Identification

Default < Back Next > Run Cancel

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

File View Help



Start typing to search actions...

DNA-Seq Polishing

Output
The file already exists and it will be overwritten.

Output FASTA Browse... ?

Save Changes ?

Output Changes Browse... ?

Version Details:
- Pilon 1.23
- SAMtools 1.10

Please Cite:
- Walker BJ et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS one*, 9(11), e112963.
- Gurevich A., Saveliev V., Vyahhi N. and Tesler G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics (Oxford, England)*, 29(8), 1072-5.

Default < Back Next > Run Cancel

- 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。

Progress File Manager Application Messages

File Manager

- Long_reads_webinar
 - 1_assembly
 - 2_Alignment
 - 3_polishing
 - dna_seq_polishing_results (2 KB)
 - fix_type_distribution_assembly_fasta (1 KB)
 - nx_plot (10 KB)
 - changes.txt (14 KB)
 - polished_sequences.fasta (3 MB)
 - 4_ProkaryoticGeneFindind
 - illumina
 - pacbio
 - SRR7498042.fastq (3096 MB)

Filter...

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

Welcome Message dna_seq_polishing_results Statistics Viewer: fix_type_distribution_assembly_fasta Chart: nx_plot



Hide Toolbar

Title Layout
Font Size: 20
Font Style: Bold

Axis Layout
Font Size: 14
Font Style: Bold

Ticks Layout
Font Size: 12
Font Style: Plain

SVG Image

PNG Image

PDF Document

Data as Text

Results Overview

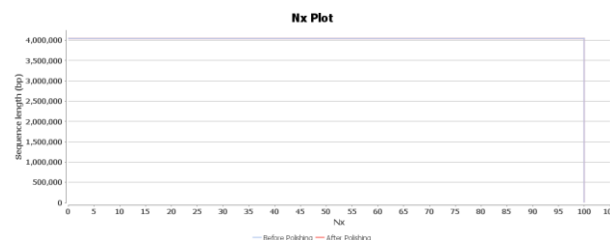
Total number of changes: 345

- Number of single nucleotide changes: 0
- Number of insertions: 339
- Number of deletions: 6
- Number of segmental changes: 0

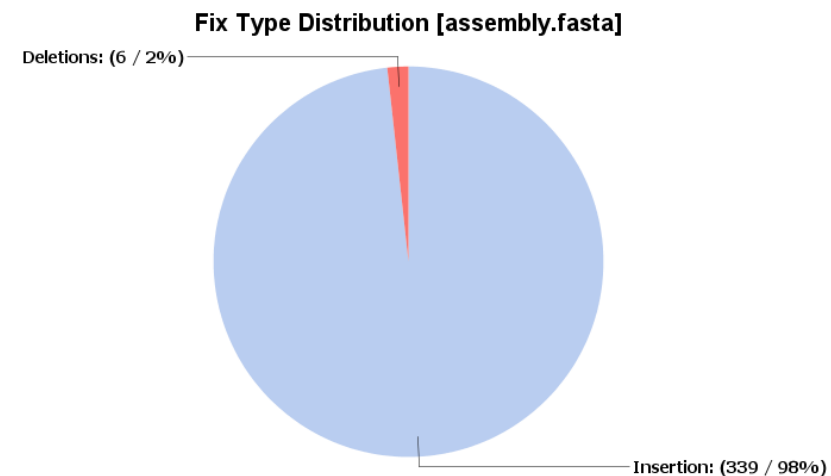
Polishingにより改善された情報

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

DNA-Seq de Novo Assemblyと見方は同じ



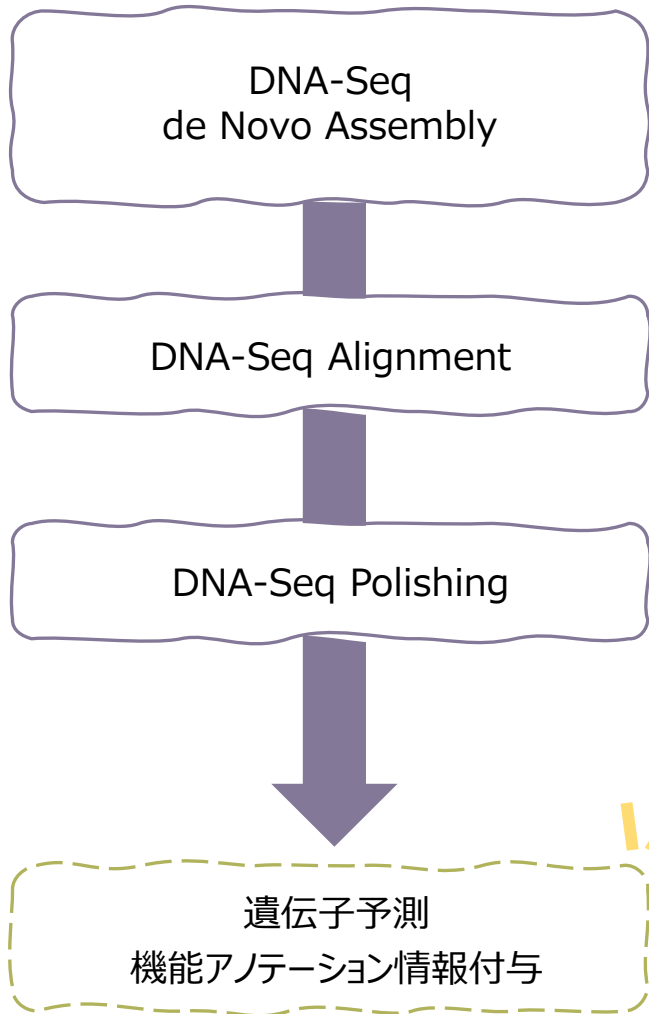
Nxプロットも同時に作成される



修正された総数・割合・内訳が図にて表示可能

サマリーレポート

✓ エラー率の少ない配列データを作成できます



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

Prokaryotic Gene Finding Results
Name: Glimmer_Sequences

Dataset Overview

- Number of sequences: 40
- GC content: 43.7%
- Start codons: ATG, GTG, TTG
- Start codons weight in second round: 0.000, 0.333, 0.000
- Stop codons: TAA, TAG, TGA

Results

Input Sequences	Found Genes					
Name	Length*	ORFs	Genes	Genes per Strand	+/-	Length Min/Max*
C02006H06	622	8	1	1/0		402 / 402
C02006D06	623	5	1	0/1		315 / 315
C02006B04	558	5	1	1/0		141 / 141
C02006D02	584	6	1	0/1		180 / 180
C02007A04	528	4	1	0/1		156 / 156
Total	2,815	28	5	2/3		141 / 402

* Length in nucleotides.
Warning: There are 35 sequences for which no gene has been found.

Genome browser view showing a table with columns: SeqID, Source, Type. A context menu is open over the table with options like 'Show in Genome Browser', 'Extract Selection to New Tab', etc.

フィルジェン WEBセミナー

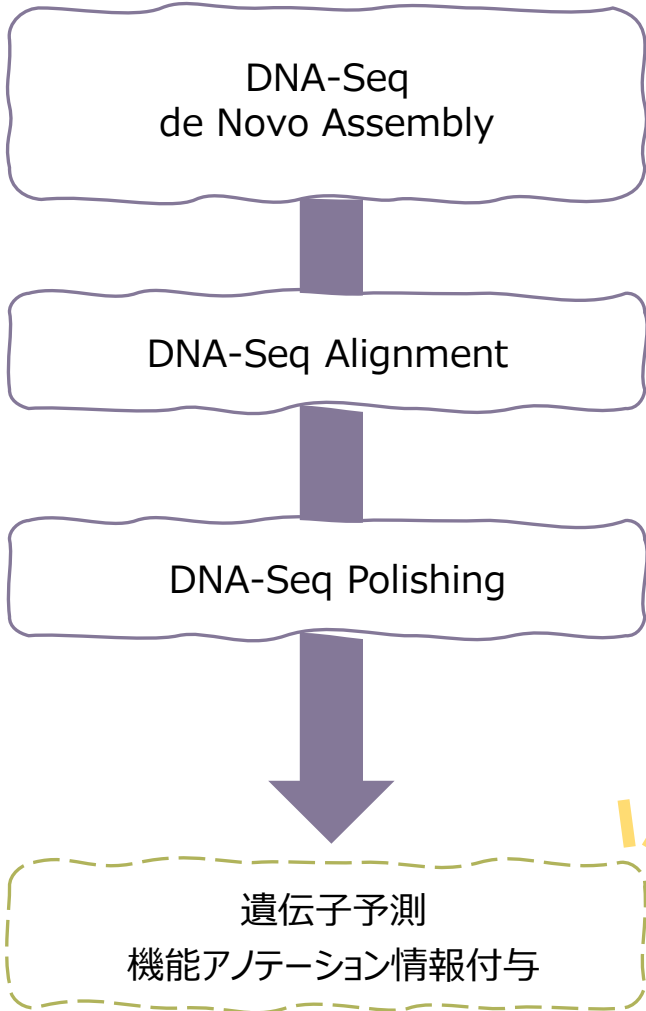
DNA配列を組み立て
遺伝子の位置と構造を予測する

～OmicsBoxを使用したNGS解析機能のご紹介～



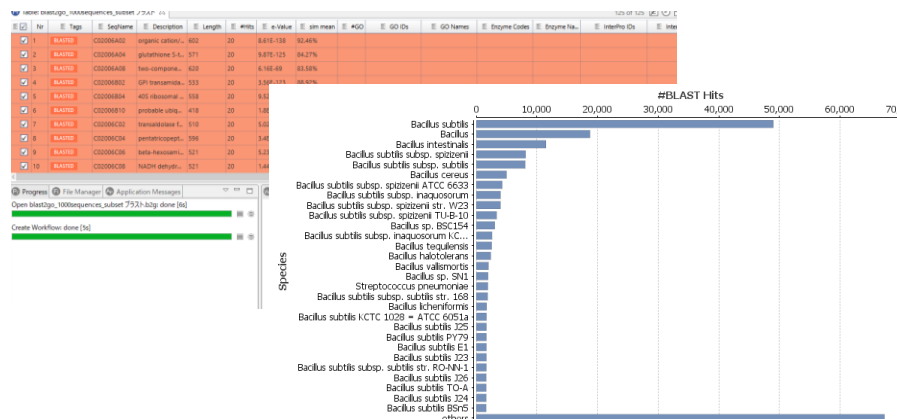
詳しくはこちら

フィルジェン Webセミナー(2019年10月10日)



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
- メーカー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。



フィルジェン WEBセミナー

高速BLAST・アノテーション解析

~OmicsBoxを使用した解析機能のご紹介~



[詳しくはこちら](#)

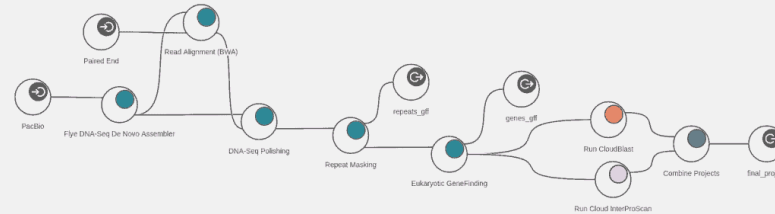
フィルジェン Webセミナー(2019年4月18日)



- ✓ 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



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)

- Taxonomic Classification
- OTU Abundances Table
- Metagenomic Assembly/遺伝子予測



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。※

※1ライセンスの場合です。同時解析は1台となります。

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