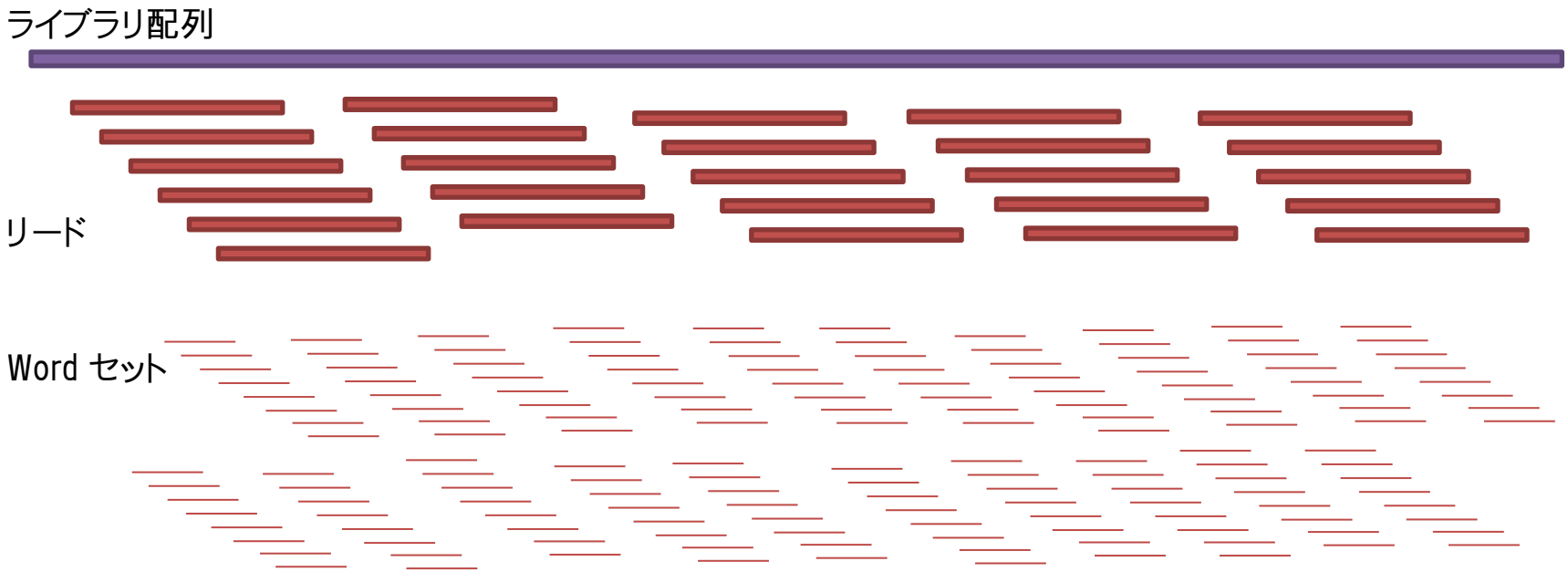
De Novo　アセンブリ

De Novo 原理

- Genomics Workbench では de brujin グラフというネットワーク理論に基づいた方法で de novo アセンブリを実行します。
- 各リードからさらに短い長さの配列のセットを作成し、グラフを作成。
- de Bruijn を利用しているオープンソースの方法ではvelvet が有名です。



De Novo 原理

Word Size

- de Bruijn グラフではリードを短い配列に分断し(word)、グラフを作成します。

(例) リード長　20, word size = 10 の場合は11個のwordができる。

リード

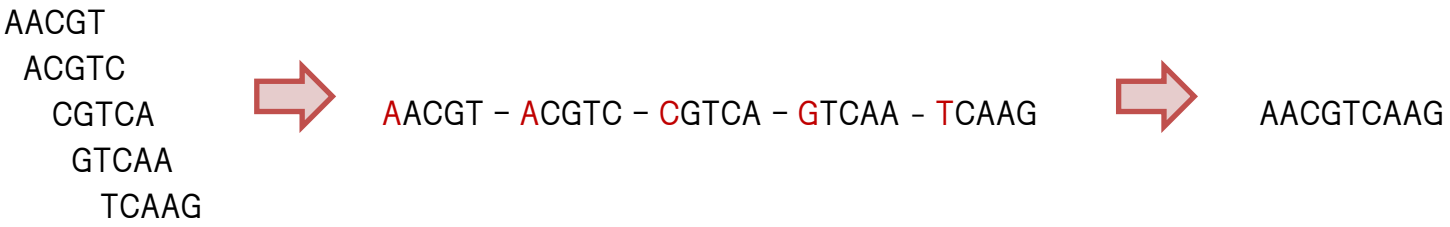
AGTTGATCTTACTAGAGGAA

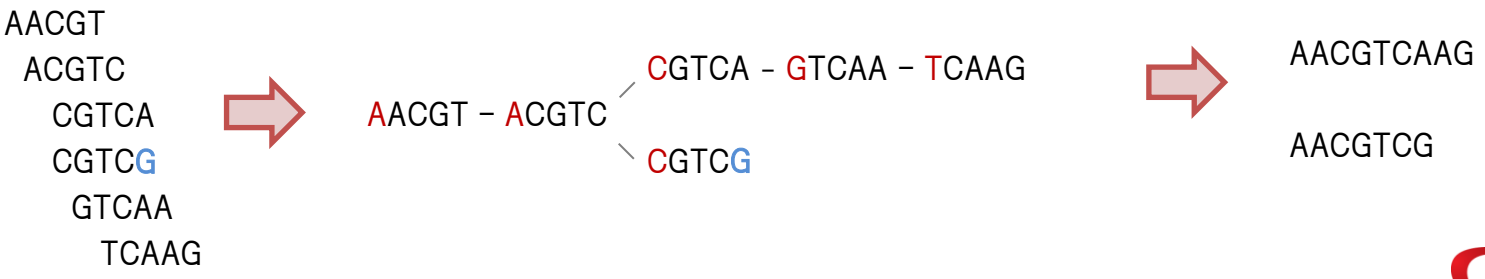
```
1  AGTTGATCTT
2   GTTGATCTTA
3    TTGATCTTAC
4     TGATCTTACT
5      GATCTTACTA
6       ATCTTACTAG
7        TCTTACTAGA
8         CTTACTAGAG
9          TTACTAGAGG
10          TACTAGAGGA
11           ACTAGAGGAA
```

すべてのリードに対して、同様にWordを作成。

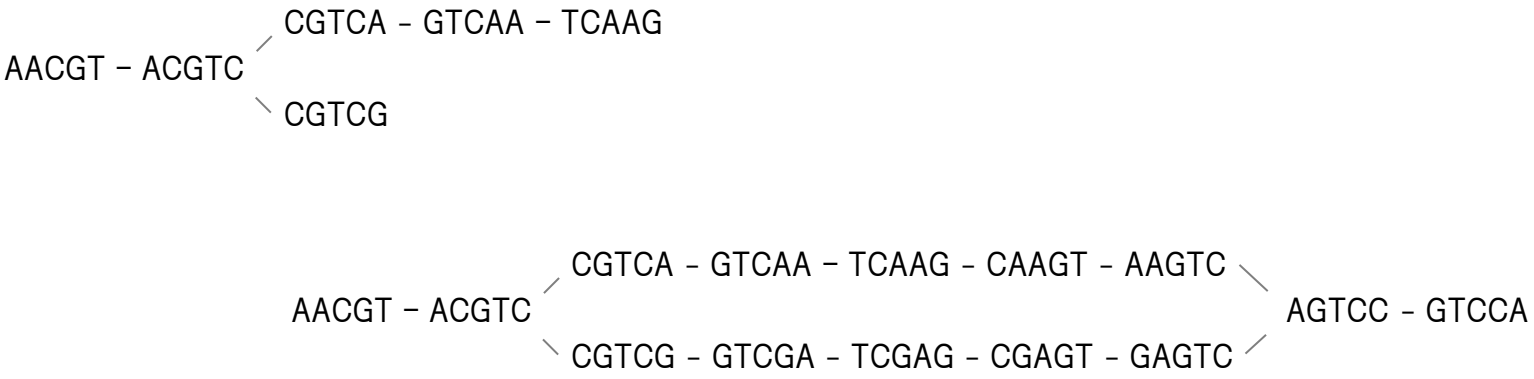
De Novo 原理

- グラフ作成 (簡単な例としてWord size = 5 で考える)



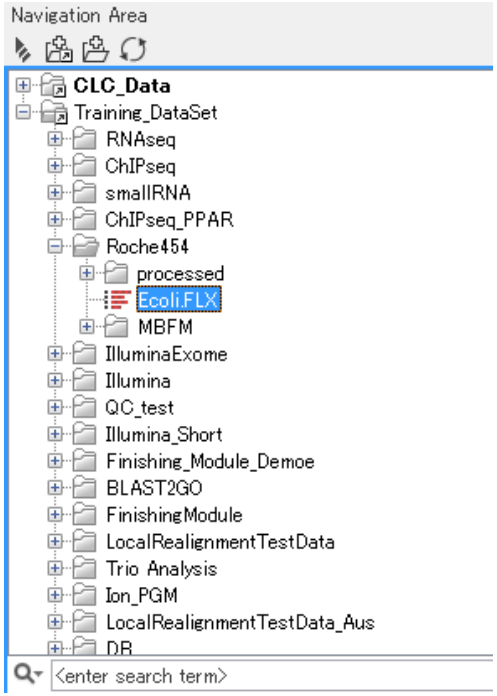


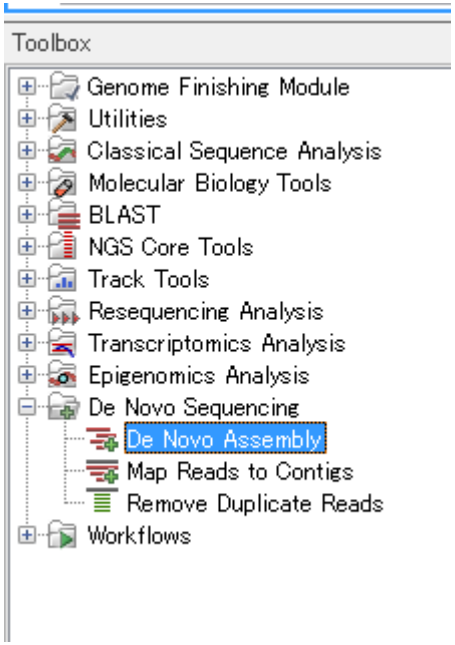
De Novo 原理

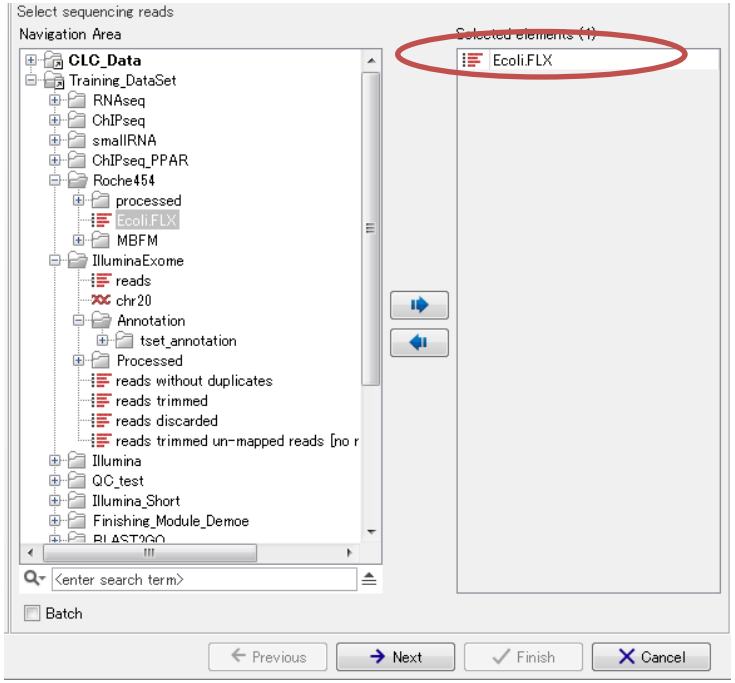


このように作成される多くのグラフから様々なステップを経て、より確からしいContigを作成していく。

De Novo アセンブリ

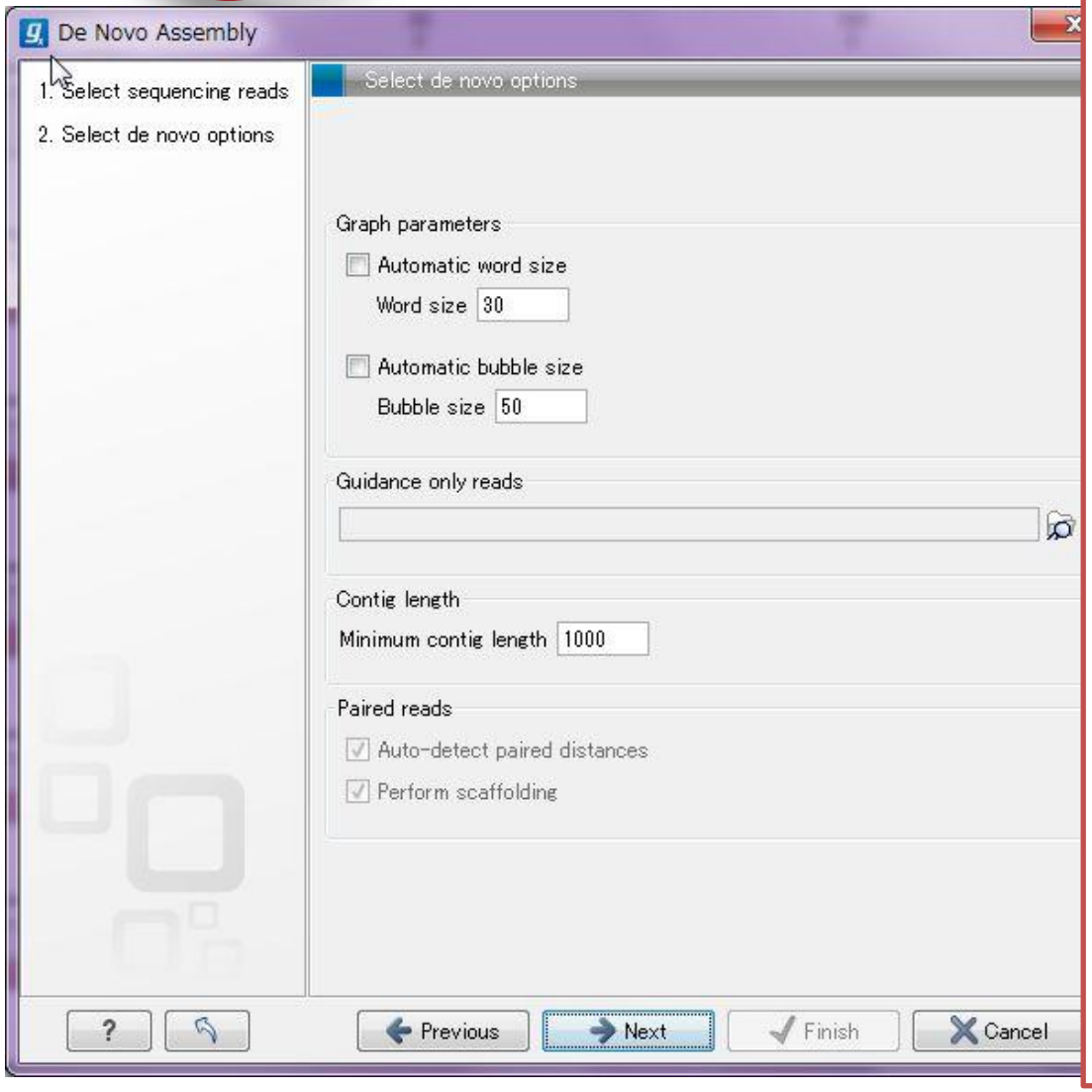






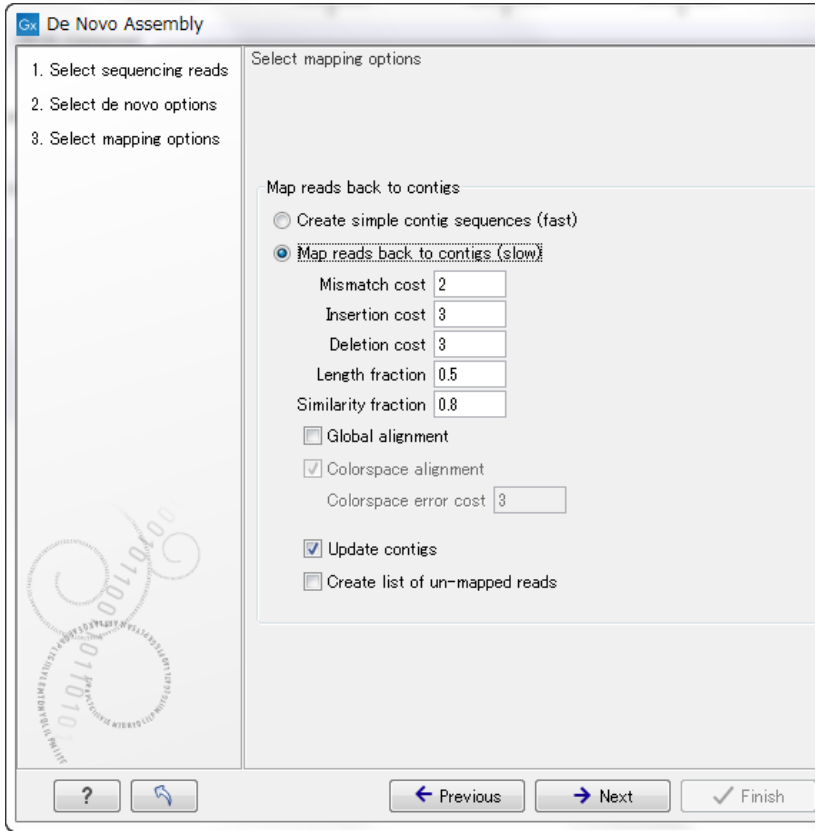
- Navigation Areaからシーケンスリストデータを選択。
- Toolboxから De Novo Sequencing > De Movo Assembly を選択、ダブルクリック。
- ウィザードが起動し、選択したデータが選ばれていることを確認。

De Novo アセンブリ

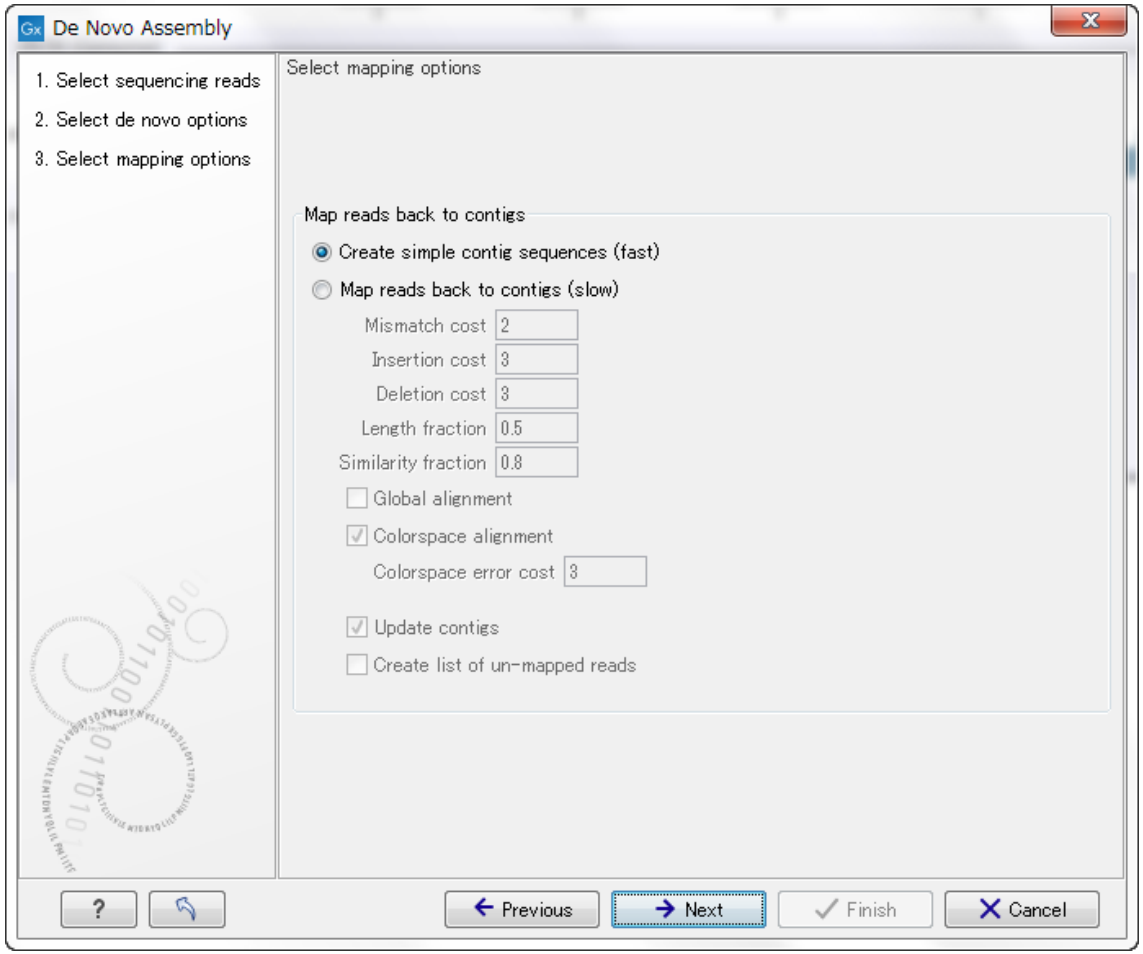


- Graph parameters
  - Automatic word size:これにチェックを入れると、Wordサイズは自動で入力されたリード数に応じて決定される。チェックをはずすと、任意で指定可能となる。
  - Automatic bubble size:これにチェックを入れると自動でbubble size が決まる。自動の場合、110bp以下のリード長では50、それ以上ではリードの平均の長さがbubble sizeとなる。チェックをはずすと任意で指定可能。
- Guidance only reads:ここで指定したリードのセットはグラフ作成には使われず、グラフにより作成されたContigの分岐やリピートを解消するために利用されます。
- Contig length:作成するContigの最小長
- Paired reads:
  - Auto-detect paired distances:ペアの距離を自動で推定する
  - Scaffoldingを行うかどうか。

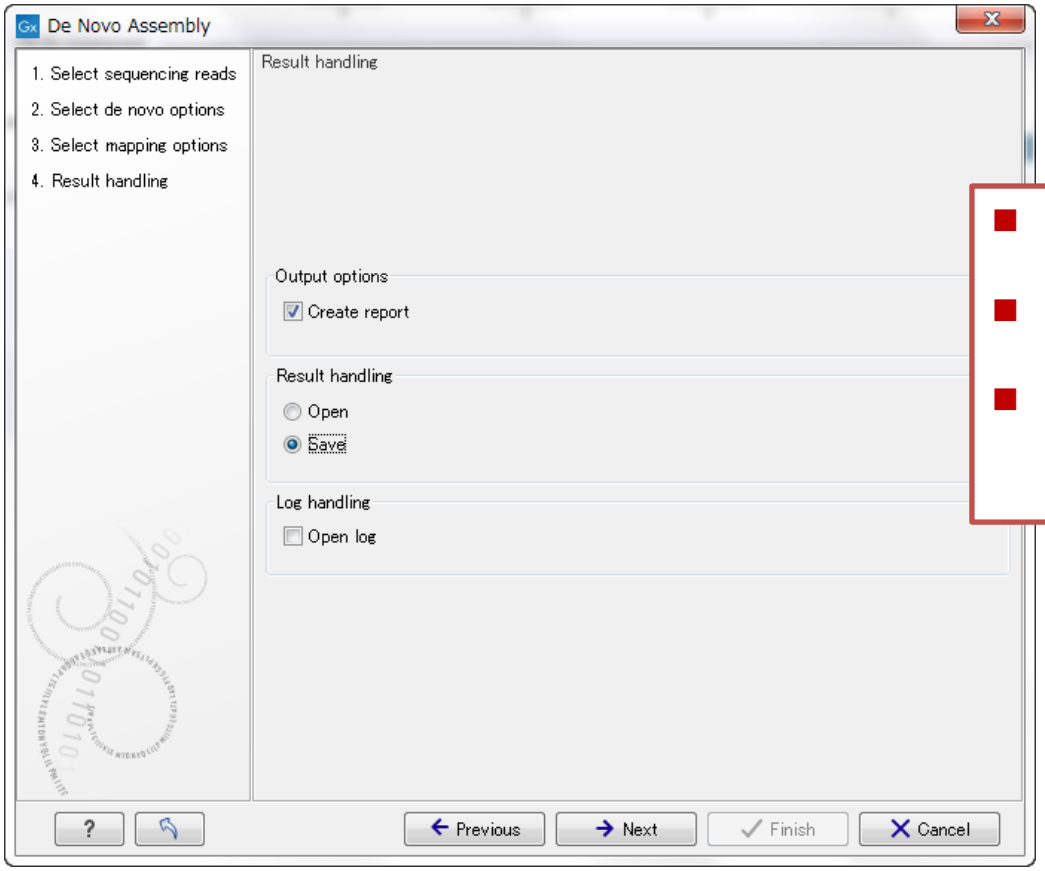
De Novo アセンブリ



- Map reads back to contigs
  - Create simple contig sequence (fast):De Novo実行時にContigの作成のみを行う。
  - Map reads back to contigs (slow):De NovoでContigを作成後、そのContigにリードをマップする。(作成されたContigの評価目的。)
    - Mismatch cost:ミスマッチコスト
    - Insertion cost:挿入のコスト
    - Deletion cost:欠失のコスト
    - Length fraction:フィルタリングで対象とする長さ
    - Similarity fraction:Length fractionのうち、どの程度の一致率以上のものを残すか。
    - Global alignment:グローバルアライメントの有無
    - Colorspace alignment, cost:カラースペースのオプション
    - Update contigs:マッピングの結果をContigに反映させるかどうか
    - Create list of un-mapped reads:マップされなかったリードのリストを作成するかどうか。


De Novo Assembly
1. Select sequencing reads
2. Select de novo options
3. Select mapping options
Select mapping options
Map reads back to contigs
Create simple contig sequences (fast)
Map reads back to contigs (slow)
Mismatch cost 2
Insertion cost 3
Deletion cost 3
Length fraction 0.5
Similarity fraction 0.8
Global alignment
Colorspace alignment
Colorspace error cost 3
Update contigs
Create list of un-mapped reads
Previous
Next
Finish
Cancel

De Novo アセンブリ



- Output option
  - Create report:レポート作成の有無。
- Result handling
  - 結果をすぐに開くか、保存するか
- Log handling
  - Make a log:ログを作成するかどうか。

De Novo 結果

Ecoli.FLX_Ass...

1 Ecoli.FLX trimmed assembly summary report

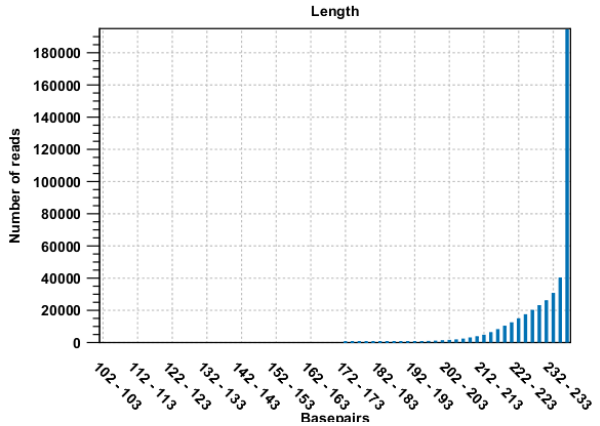
1.1 Nucleotide distribution

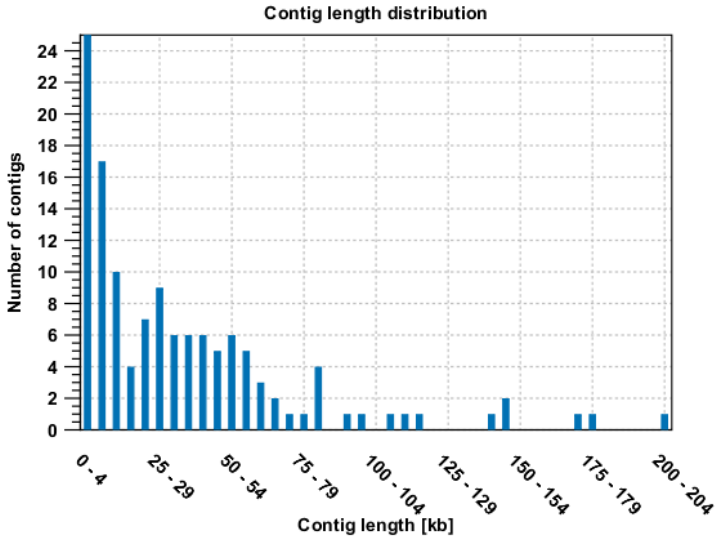
Nucleotide	Count	Frequency
Adenine (A)	1,122,585	24.6%
Cytosine (C)	1,150,465	25.3%
Guanine (G)	1,160,811	25.5%
Thymine (T)	1,121,096	24.6%

1.2 Contig measurements

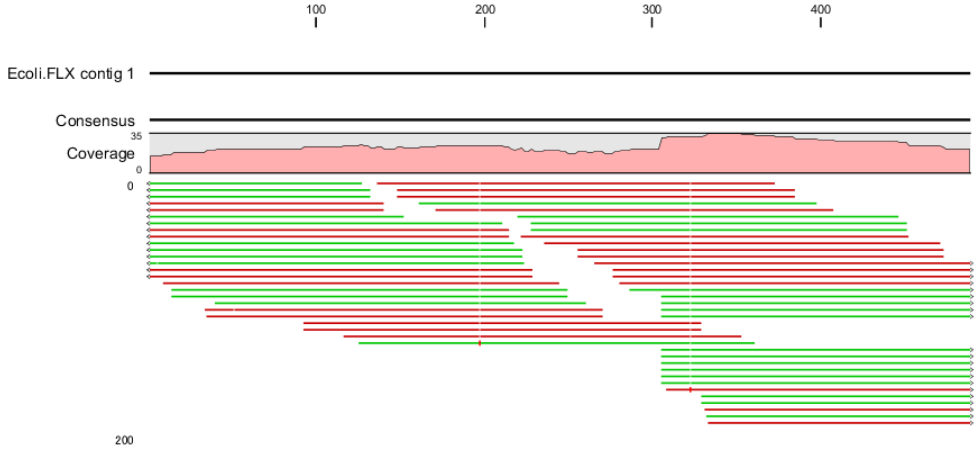
	Length
N75	38,261
N50	59,509
N25	112,279
Minimum	854
Maximum	201,817
Average	35,586
Count	128
Total	4,554,957

1.5 Distribution of read length





De Novo 結果



コンティグの結果は、配列として見ることも、コンティグにリードをマッピングさせた状態で見ることも可能。

De Novo 注意点

クオリティとトリミング

- De Novoでは、エラーが多いデータを使うと、非常に複雑なグラフが作成され、メモリが非常に多く必要となります。データのクオリティを確認しながら、必要に応じて厳しめのトリミングを実行してください。
- クオリティの向上は作成されるContig数を減らすだけでなく、必要となるメモリが少なくなり、計算速度にも影響してきます。

De Novo 注意点

Duplicate除去

- Whole Genome De Novoの場合は、Duplicateの除去は重要となります。エラーによるDuplicateが多く残っていると、作成されたグラフから正しいContigを選ぶ妨げとなります。また余分なメモリを使う原因ともなるため、エラーによるDuplicateはきちんと取り除きましょう。
- Whole Genome De Novo にてDuplicateの除去を行うために、ベータ版ですが「Remove Duplicate Reads」というプラグインが利用できます(詳細は次ページ以降)。
- Transcriptome De Novo の場合、Duplicateではなく高発現による場合もあるため、注意が必要です。そのため、そのまま実施されることも多いです。

De Novo 注意点

パラメータ設定

- 最適なWordサイズは、データ毎に異なります。何度か設定を変えて実行し、最適な値を決定するようになります。まずは自動で行ってみて、自動で決定されたWord sizeの前後10bp、または20bpなど幅をとって値を変更し、N50やContigの数が減少するかなどを見て、最適な値を決定してみてください。
- バブルサイズも同様ですが、バブルサイズを任意で変更して効果があるデータは454やIon PGMなどリード長が長く、ホモポリマーのエラーなどが含まれる場合に改善することがあります(詳細はホワイトペーパーに記載されています)。サイズの設定は、リード長の半分程度から、いくつか値を振って最適な値を検討するようになります。