Chapter 4
Sequence Condensation Tool

NextGENe の Condensation ツールでは、プロジェクトの必要に応じて、カバレッジ深度を使って同一リードを統合したり、リード数を保持したままベースコールエラーを含むシーケンスリードを補正したりリードを伸長したりすることが可能です。

このチャプターでは以下のトピックをフォローしています；

- Overview of the NextGENe Sequence Condensation Tool（97 ページ）
- Sequence Condensation Tool - General Settings（102 ページ）
- Sequence Condensation Tool Output Files（113 ページ）

Overview of the NextGENe Sequence Condensation Tool

NextGENe の Condensation ツールでは、プロジェクトの必要に応じて、カバレッジ深度を使って同一リードを統合したり、リード数を保持したままベースコールエラーを含むシーケンスリードを補正したりリードを伸長したりすることが可能です。

Condensation には、Consolidation、Elongation、Error Correction の 3 つの手法が利用できます。3 手法全て、クラスター化したリードのコンセンサス配列を生成することにより、低頻度のシーケンシングエラーを補正します。使用できる手法は解析するデータタイプに依存します。

✓ 解析に複数サンプルファイルをロードする場合は、個別のサンプルファイルとしてではなく複数サンプルファイル全体として評価されます。

Illumina, SOLiD System and Ion Torrent data

Illumina データ、SOLiD System や Ion Torrent データを解析している場合、3 つの Condensation 法が全て使用できます。この 3 つの Condensation 法は、類似しているリードのクラスタリングやコンセンサス配列生成に同じ手法を用います。複数のリードから見出させるアンカー配列（12bp の共通配列）や共通の指標によりリードは評価され、同じアンカー配列を含むリードからグループが形成されます。この配列はゲノム中でユニークな配列ではないため、グループはさらにアンカー配列に隣接する左右の塩基配列（ショルダー配列）に基づいて別々のサブグループに分類されます。少なくとも両方のショルダー配列を含むリードはブリッジリードと呼ばれ、ブリッジリードは伸長することや、両方のショルダー配列の橋渡しをする（Bridge）ことが可能です。サブグループを形成するには最小限のブリッジリードが必要です。アンカー配列のショルダー配列を評価することで、1 つのグループを、同一のアンカー配列をもち、かつ各サブグループ固有のショルダー配列をもつ複数のサブグループに分けることができます。複数のリードが同じアンカー配列を持っていても、ショルダー配列中の Variant や多型のため複数サブグループが存在する可能性はあります。またこの 12bp のアンカー配列がゲノムの別の領域にいくつも存在する可能性もあります。

各サブグループを使用してコンセンサス配列を生成できます。Illumina データ、SOLiD System や Ion Torrent データは、各リードの 5’末端のベースクオリティが Phred スコア 20 より高く、リードの残りの部分はよりクオリティが低いと仮定しており、結果として、正確性においてシーケンスの 5’末端により高いウェートをおいてベースコールを行っています。コンセンサスベースコールは、以下の規則に従って、あるポジションに出現する各塩基のスコア付けから計算されます；

- 5’シーケンスは 3’シーケンスより高いウェートが割り当てられます。
- あるポジションの塩基について、各リードの 5’にはスコア 7 が割り当てられます。
- 同じポジションの塩基について、各リードの 3’にはスコア 2 が割り当てられます。
- 同じ塩基の全リードのスコアは足し合わされ、その塩基のスコアとされます。

塩基のスコア X＝（7 × 5’リード数）＋（2 × 3’リード数）

例えば、あるポジションが T のリードと、同じそのポジションが C のリードをそれぞれ含むリードサブグループのそのポジションについて考えます。塩基 T は 2 リードの 5’末端にあり、6 リードの 3’末端にあります。塩基 C は 4 リードの 5’末端にあり、2 リードの 3’末端にあります。コンセンサスベースコールを決定するために、塩基 T と塩基 C 両方のクオリティスコアが以下の式で計算されます；

- 塩基 T のスコア＝（7×2）＋（2×6）＝26
- 塩基 C のスコア＝（7×4）＋（2×2）＝32

塩基 C のスコアの方が塩基 T のスコアより大きいため、このポジションのコンセンサス配列は塩基 C になります。

Consolidation

Condensation の Consolidation 法（Illumina データ、SOLiD System、Ion Torrent データ用）を使うと、オーバーラップしたシーケンスは統合され、サブグループの全オリジナルリードのコンセンサス配列が使われます。しかし、元のカバレッジ情報が失われないようオリジナルリードの情報は保持されます。Consolidation 法は Raw リードのカバレッジが非常に深いデータセットに推奨されます。図 4－1 は Consolidation 法選択時の Condensation ツールの出力結果例です。

図 4－1：Consolidation 法使用時の Condensation ツール結果

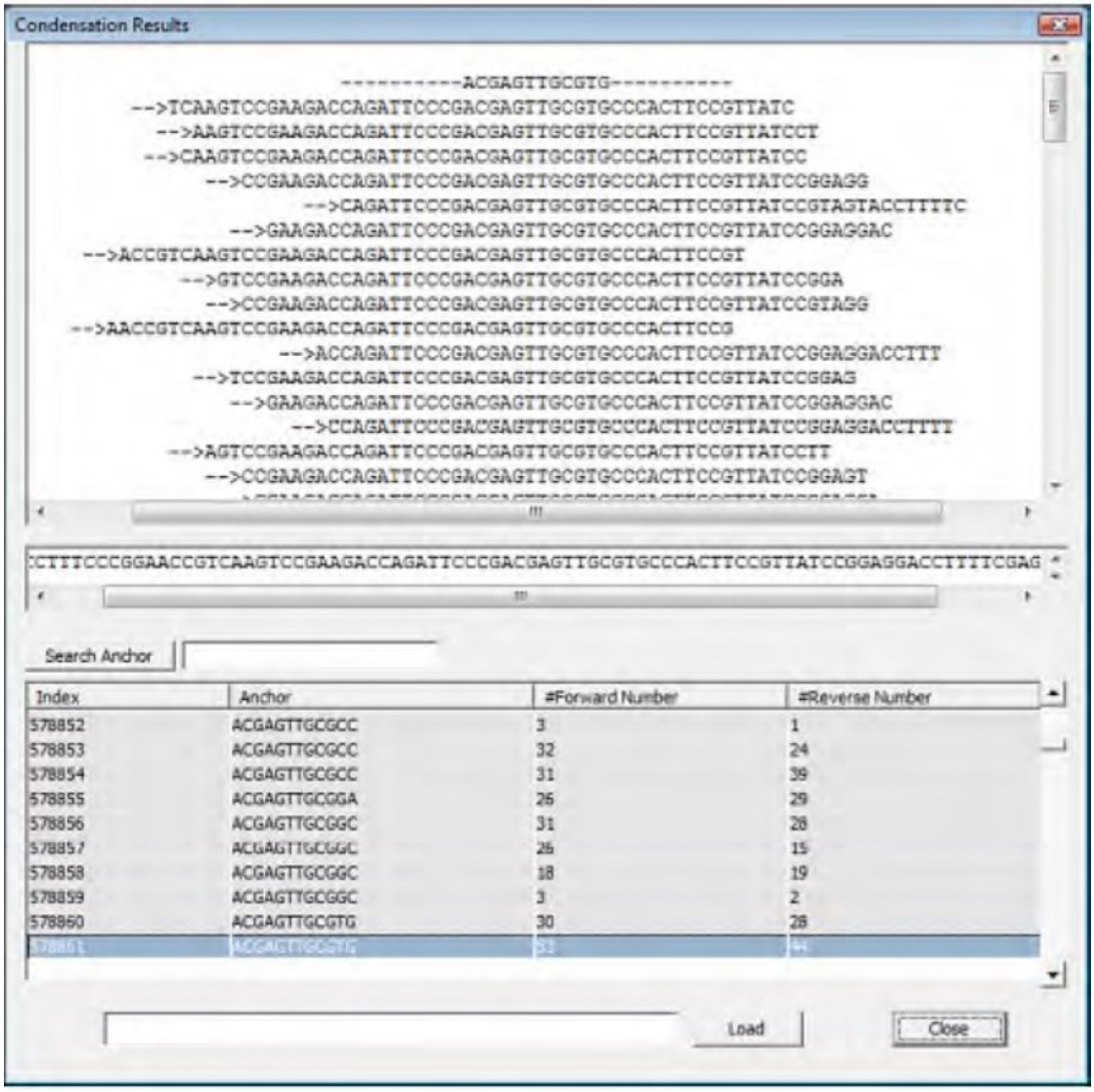


図 4－2 はコンセンサス配列出力例とリード名です。アンカー配列、ショルダー配列、使用されたフォワードリードとリバースリード数が反映されています。

図 4－2：コンセンサス配列出力例

```
>1      CTTTGGAACCTC   24      ACACATGGCT     CTCTGCCTCC     105     95
TTCAGTATTACATGACACATGGCTCTTTGGAACCTCCTCTGCCTCCACTCTGCCCAGCTG
>2      AGACCTACAAAT   24      TAGAGGAATTAA   AACAGACTGAAA   831     714
ATTATTACTAATTAGAGGAATTAAAGACCTACAAATAACAGACTGAAACAGTGGGGGAAA
```

✓ Consolidation 法選択時の Condensation ツール結果の閲覧について詳細は「The NextGENe Condensation Results Tool」（383 ページ）参照のこと。

Elongation

Condensation の Elongation 法（Illumina データ、SOLiD System、Ion Torrent データ用）を使うと、オーバーラップしたシーケンスは統合されず、代わりに、サブグループ各リードのエラー補正された伸長リードが生成されます。リードは複数のアンカー配列にマッチしている可能性が高いため、リードの全オブジェクトは複数サブグループに「そのまま」プールされます。これらの補正・伸長されたリードは、それからお互いに比較され、一つのコンセンサス配列を生成します。Consolidation 法の場合とは異なり、インデックスのどれにもマッチしないリードも除外されず、出力ファイルに出力されます。

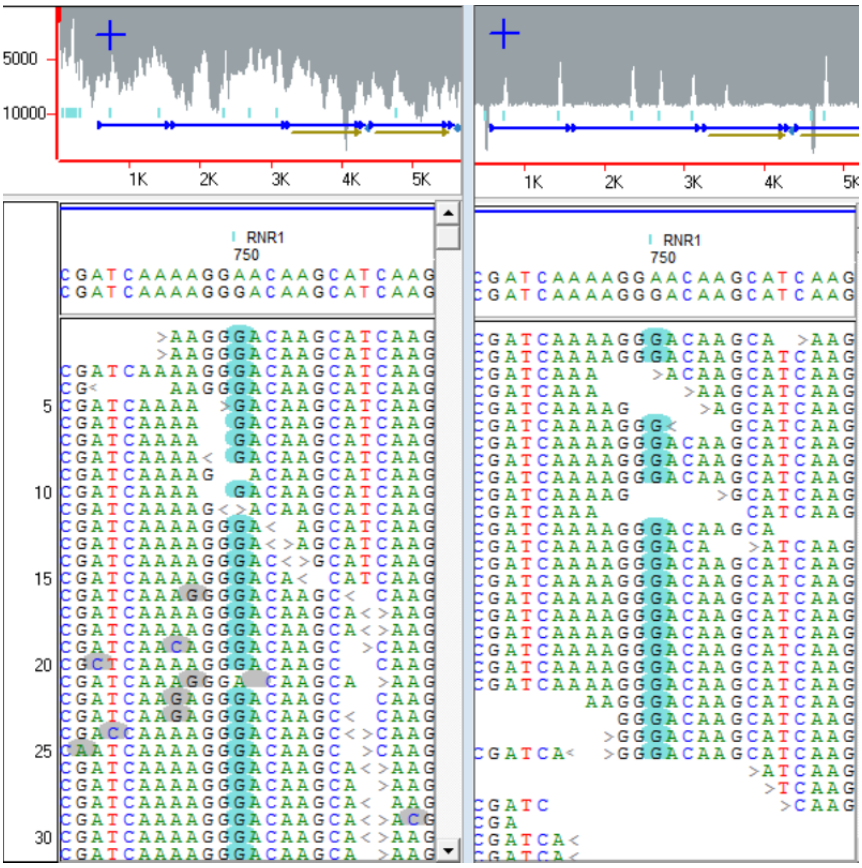
- ✓ Elongation 法は Raw リードのカバレッジが浅いデータセットやペアエンド/メイトペアに推奨されます。

Error Correction

Error Correction 法は Consolidation 法と Elongation 法に非常によく似ています。同じ方法でリードはクラスター化され、低頻度のエラーは補正されますが、リード長は伸長されず、統合もされません。代わりに各オリジナルリードは元のリード長が変わらないまま、シーケンシングエラーが補正されます。

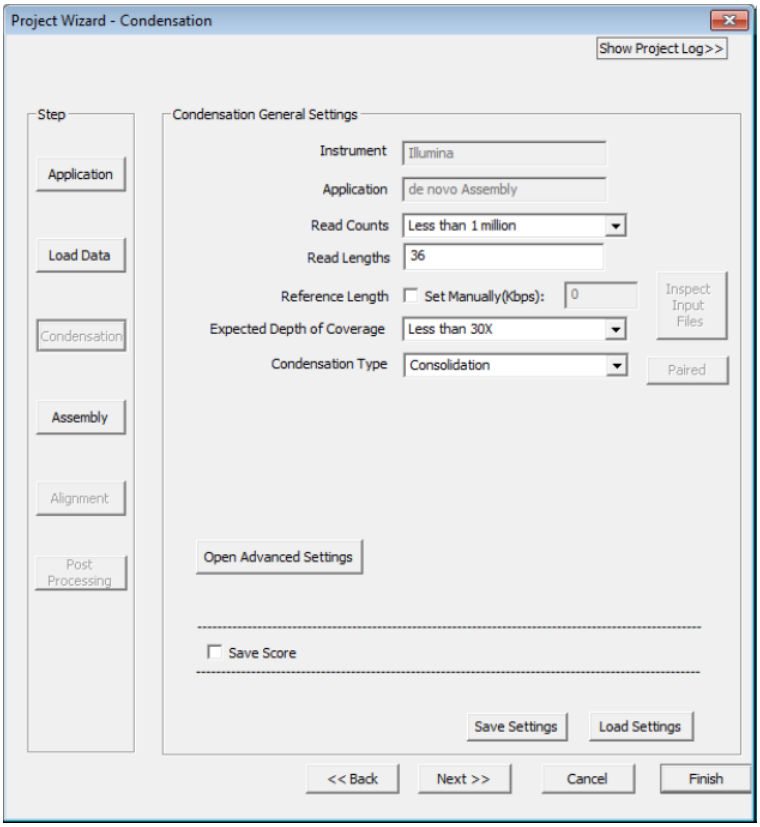
図 4－3 は Condensation ツールを用いた SNP Discovery アプリケーションの例です。この図の左側には、Raw リードがリファレンスにアライメントされています。低頻度の Variation、エラーらしき塩基はグレーでハイライトされています。一方、Variant Call は青色でハイライトされています。図の右側には、Condensation されたリードがリファレンスにアライメントされています。真の SNP が保持されている一方、エラーらしき塩基は減少しています。

図 4－3：Condensation ツールを用いた SNP Discovery



Sequence Condensation Tool
- General Settings

図 4－5：Condensation Settings ページ、General Settings



設定	概要
Inspect Input Files	Illumina、SOLiD システム、Ion Torrent データ解析時のみ使用できます。このボタンをクリックすると、Condensation ツールはデータファイルをスキャンしてこのページと Advanced Settings ページの最適設定を決定します。Inspect Input Files をクリックした後は、Illumina、SOLiD システム、Ion Torrent のデータセットについて Read Counts、Read Lengths、Reference Length（プレロードリファレンス以外）、Expected Depth of Coverage の値は自動的に設定されますが、必要に応じて変更できます。

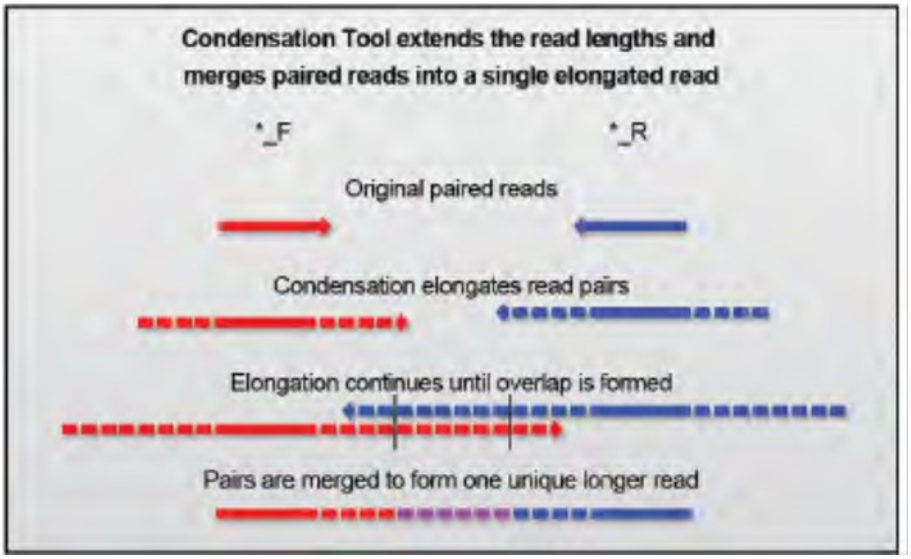
Read Counts	サンプルデータセットに含まれるリード数に最も当てはまる範囲を選択して下さい。 ✓ 複数データファイルを解析している場合は、この値には全ファイルの合計を指定してください。
Read Lengths	サンプルデータセットのリード長に最も当てはまる数を指定してください。
Reference Length	リファレンス配列長に最も当てはまる範囲を選択して下さい。Preloaded リファレンスファイルについてはこの値をマニュアルで入力してください。
Expected Depth of Coverage	サンプルデータセットの Expected Depth of Coverage に最も当てはまる範囲を選択してください。Inspect Input Files をクリックした後は、Illumina、SOLiD システム、Ion Torrent のデータセットについては、サンプルファイル内の全塩基数をリファレンスファイル内の塩基数で割った値が自動的に設定されます。 低頻度の Variation を検出するためには、Expected Depth of Coverage はマイナーアレルのカバレッジをセットすることを推奨します。
Condensation Type	Illumina、SOLiD システム、Ion Torrent データについて、以下のどれかを選択してください； ・ Consolidation（リード数を減らします。） ・ Elongation（リード数を保持します。） ・ Error Correction（リード数を減らしたりリードを伸長したりすることなくエラーを減らします。） Roche/454 データについては、Error Correction オプションのみ使用できます。
Paired	Illumina データで Elongation を選択している場合のみ使用できます。このオプションをクリックすると、Merge Overlapping Paired Reads ダイアログボックスが開きます。

	図 4－7：Merge Overlapping Paired Reads ダイアログボックス  このダイアログボックスでは、Elongation 後にオーバーラップしたペアリードを統合したいか指定できます。またオーバーラップしていないペアリードのクオリティが低い末端を無視するかも指定できます。また統合結果の長さについての設定も 2 オプションあります； ・ Merged Length [ ] bp to [1000] bp ・ Merged Length [70] bp to [130] % of the longer read length 一つもしくは両方のオプションを選択できますが、両方のオプションを選択した場合は両方の条件を満たさないと結果として残りません。 ✓ ペアリードが正確に統合されるオーバーラップの最小塩基数の推奨値は 9 です。9 より小さい値も選択できますが、ペアリード間のオーバーラップがより小さくてもよくなるため、結果の信頼性が落ちる可能性があります。9 より大きい値も選択できますが、統合するリードのオーバーラップがより長ければいけないため、統合されるペアリードが少なくなる可能性があります。「Merging Paired End Reads」（105 ページ）参照のこと。
Save Score	Consolidation の.qual ファイルを生成します。このファイルは各サブグループで使用されたリード数の情報を含みます。

Merging Paired End Reads

NextGENe のペアエンド統合機能で、ペアリード伸長によって 2 リード間のオーバーラップが存在するポイントでペアエンドリードを統合することができます。そのペアリードは連続したより長い一つのリードを形成し、以降は 1 リードとして解析の対象となります。

図 4－8：オーバーラップするペアエンドリードの統合



Elongation の必要サイクル数はリード長とライブラリーサイズに依存します。一般的に各 Condensation サイクルは、短いオリジナルリードの場合（≦36bp）は平均リード長をオリジナルリード長の 1.6 倍に、長いオリジナルリードの場合（＞36bp）はオリジナルリード長から 6bp 差し引いた 2 倍に増やします。例えば、200bp ライブラリー由来の 75bp リードについては、Elongation 1 サイクルで、ペアリードがオーバーラップするのに十分にまで伸長したリードが得られます。同じく 200bp ライブラリー由来の 35bp リードについては、Elongation 3 サイクルが必要です。ペアリードの重要な部分（Elongation したリード長のおおよそ 15%くらい）がオーバーラップすると予想されるまでリードを伸長することを推奨します。

図 4−9：各オリジナルリード長の Elongation 後の平均リード長

Original Read Length	35 bp	50 bp	75 bp
Avg Read Length After 1 Cycle of Elongation	56 bp	88bp	138 bp
Avg Read Length After 2 Cycles of Elongation	90 bp	160 bp	
Avg Read Length After 3 Cycles of Elongation	144 bp		

オーバーラップする領域がペアリード間で一致した場合のみ、そのペアリードは統合されます。シーケンシングケミストリーやベースコール、もしくは Elongation による初期アセンブリによって生じたエラーはペアリード間で一致しないため、そのペアは統合されません。

Sequence Condensation Tool Output Files

Condensation データ解析のステップが完了すると、解析結果の詳細情報の出力ファイルが作られます。Condensation 法ごとに別々の情報を表示する別々の出力ファイルがあります。

- 「Consolidation output files」
- 「Elongation output files」（114 ページ）
- 「Error Correction output files」（115 ページ）

Consolidation output files

ファイル	概要
_Condensed_Raw.fasta	このファイルには Condensation に使われたオリジナルリード全てが含まれます。
_Cycle#.fasta	このファイルは実行された Condensation の各サイクルについて作成されます。＃はサイクル番号です。このファイルはその Condensation サイクルで生成されたコンセンサスリードを含みます。
_OrgSampleID.txt	後々の解析で NextGENe がこれらを参照できるように、このファイルは元のサンプル ID を保存しています。
_Parameters.txt	このファイルはそのプロジェクトに使われた設定の情報を含みます。初期ステップとして Condensation が実行されてから同じプロジェクトの一部としてアライメントやアセンブリが実行された場合、プロジェクトの全ステップの設定を含む _Parameters.txt ファイルが作成されます。

_StatInfo.txt	このファイルには Condensation ステップの様々な Statistics が含まれます。 ・ インデックスに一致したシーケンス数 ・ 生成された Condensation リード数 ・ Condensation リードの平均長 ・ 各 Condensation リード内の平均カバレッジ ・ ユーザーマネジメントが起動している場合、解析を Run したユーザーのユーザー名
_Uncondensed_Raw.fasta	このファイルは Condensation に使われなかった全リードを含みます。
TempViewDir.giv	このファイルを使って、Condensation Results ツールで、Consolidation の結果をグラフィカルに見ることができます。「The NextGENe Condensation Results Tool」（381 ページ）参照のこと。 このファイルは View Condensation Results が選択されている場合のみ作成されます。

Condensation に Consolidation 法が選択されているとき、リード情報のキーピースを含む名称が各コンセンサスリードに割り当てられます；

- 各名称は「>」から始まり、リード名の始まりを示しています。
- そのシーケンスが一致した 12bp アンカー配列のインデックス番号
- 12bp のアンカー配列
- コンセンサス配列中のアンカー配列の開始位置を示す番号
- 左ショルダー配列
- 右ショルダー配列
- コンセンサス配列生成に使われたフォワードリード数

- コンセンサス配列生成に使われたリバースリード数

例えば、下のリード名について考えると、

>67059_TCCTGACTCCAC_19_GACGGATG_CCACACCC_42_67<

このリードはアンカー配列「TCCTGACTCCAC」を含む 67,059 番目のインデックスから生成されました。このアンカー配列はコンセンサスリードのポジション 19 から始まり、その左に配列「GACGGATG」が、右に配列「CCACACCC」があります。コンセンサス配列の生成に、42 のフォワードリードと 67 のリバースリードが使われました。

Elongation output files

ファイル	概要
_Cycle#.fasta	このファイルは実行された Condensation の各サイクルについて作成されます。＃はサイクル番号です。このファイルはその Condensation サイクルで生成されたコンセンサスリードを含みます。
_Parameters.txt	このファイルはそのプロジェクトに使われた設定の情報を含みます。Condensation が初期ステップとして実行されてから同じプロジェクトの一部としてアライメントやアセンブリが実行された場合、プロジェクトの全ステップの設定を含む _Parameters.txt ファイルが作成されます。
_StatInfo.txt	このファイルには Condensation ステップの様々な Statistics が含まれます。 ・ インデックスに一致したシーケンス数 ・ 生成された Condensation リード数 ・ Condensation リードの平均長 ・ 各 Condensation リード内の平均カバレッジ ・ ユーザーマネジメントが起動している場合、解析を Run したユーザーのユーザー名

Error Correction output files

ファイル	概要
*_ErrorCorrected.fasta	このファイルにはエラー補正されたリード全てが含まれます。このファイルはサンプルファイルとしてプロジェクトで使用できます。このファイルをサンプルファイルとして使用する時は、改めて Error Correction 法を使用しないでください。
_Parameters.txt	このファイルはそのプロジェクトに使われた設定についての情報を含みます。Condensation が予備ステップとして実行され、同じプロジェクトの一部としてその後にアライメントやアセンブリが実行された場合、プロジェクトの全ステップの設定を含むこのファイルが作成されます。
_StatInfo.txt	このファイルには Condensation ステップの様々な Statistics が含まれます。 ・ インデックスに一致したシーケンス数 ・ 生成された Condensation リード数 ・ Condensation リードの平均長さ ・ 各 Condensation リードの平均カバレッジ ・ ユーザーマネジメントが起動している場合、解析を Run したユーザーのユーザー名

お問い合わせ先

電話・E メールでのお問い合わせ

■ バイオアップロード合同会社

■ TEL：0284-22-4213

■ E-mail：info@bio-upload.com

■ 対応時間帯：平日９：００～１７：３０