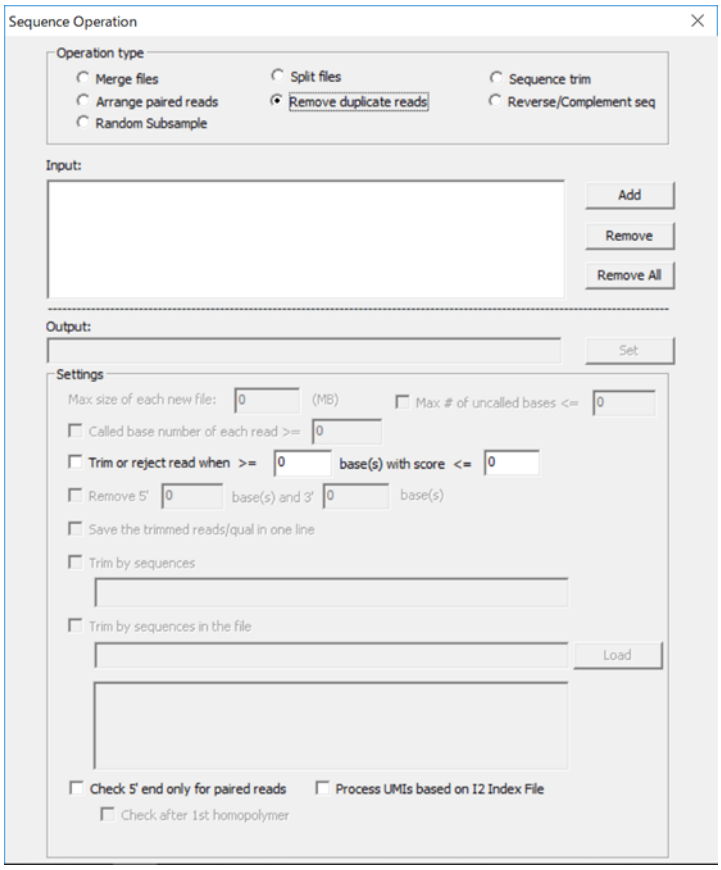
To remove duplicate reads

重複リードの除去には 2 つのアルゴリズムが使用できます；

- アルゴリズム 1 は塩基配列を使用します。次項「The Nucleotide Sequence algorithm for removing duplicate reads」を参照してください。
- アルゴリズム 2 は Unique Molecular Identifiers（UMI）と関連したインデックスファイルを使用します。「The UMI algorithm for removing duplicate reads」（374 ページ）を参照してください。

図：Sequence Operation ダイアログボックス、Remove duplicate reads 選択時



The Nucleotide Sequence algorithm for removing duplicate reads

Remove Duplicate Reads を選択した場合、Sequence Operation ツールは数値をリードの各塩基に割り当てるアルゴリズムを使用します（A=0、C=1、G=2、T=3)。

各リードについて次式によりハッシュ値が計算されます：

Sum（塩基コード×（4^塩基ポジション）

ここで、スタートの塩基ポジションは＝0 です。例えば、配列 ATTC については、ハッシュ値は次のように計算されます；

0*(4^0) + 3*(4^1) + 3*(4^2) + 1*(4^3) = (0*1) + (3*4) + (3*16) + (1*64) = 124

複数リードが同じハッシュ値をもつ場合、同一の配列と同一の配列長を示し、この配列の一つのコピーが保持されます。ペアリードについては、両方のフォワードリードが同じハッシュ値をもち、両方のリバースリードも同じハッシュ値をもつ複数のペアがあった場合、同一の配列と同一の配列長を示し、その複数リードのうち１ペアのみが保持されます。例えば、リード１F＝リード 2F とリード１R＝リード２R の場合、１ペアのリードのみが保持されますが、リード１F＝リード２F だがリード１R≠リード２R の場合は両方のペアのリードが保持されます。

1. Input ペインで Add をクリックして重複リードを除去する FASTA もしくは FASTQ ファイルを選択して下さい。

2. 重複リード除去の設定を行ってください。

設定	概要
Check 5’ end only for paired reads	このオプションを選択すると、ペアリード両方の 5’末端最初の 32bp のみをチェックし、Duplicate を判断します。
Check after 1st homopolymer	Check 5’ end only for paired reads オプションを選択しているときのみ使用できます。選択すると、1 番目のホモポリマー配列後ろの、最初の 32bp に基いて重複リードをチェックします。

3. Output フィールドをデフォルト値のままにするか、Set をクリックして出力ファイルの保存場所を指定してください。

4. OK をクリックしてください。

   処理が完了するとメッセージが開き、2 つの出力ファイルが生成されます。Duplicate.fasta は解析から除外された重複リードを含むファイルです。Unique.fasta は重複が無かった全リードと、重複リードの 1 コピーを含むファイルです。ログファイル（RemoveDuplicates_Log.txt）も生成されます。このログファイルは入力ファイルやリード（リード総数、ユニークリード数、重複リード数)、リード分布やそのカウント数の情報が含まれます。

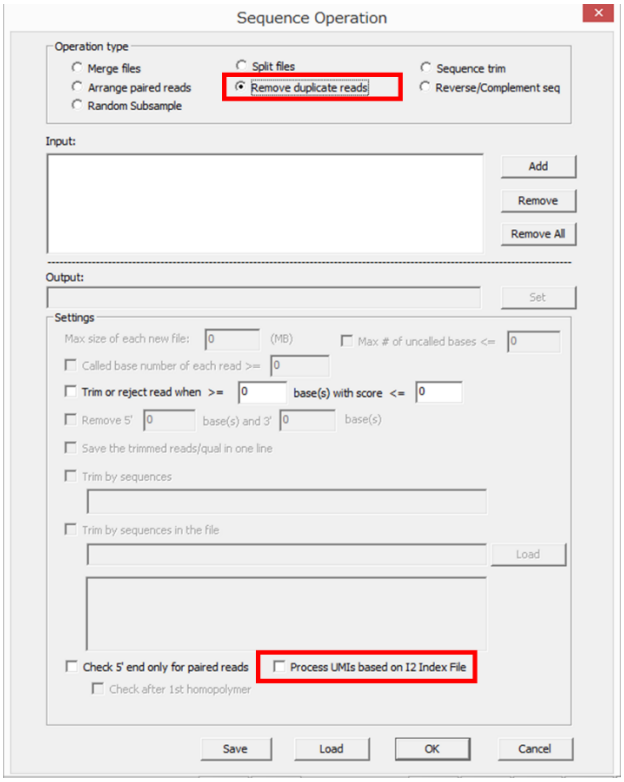
The UMI algorithm for removing duplicate reads

Unique molecular identifiers（UMI）はランダムな塩基配列で、ライブラリ増幅の前に全ての DNA 分子（断片）に付加され、タグ配列として PCR duplicate の識別に使用されます。イルミナ社の次世代シーケンサーは UMI 情報を含む I2 インデックスファイルを生成します。NextGENe はこの I2 インデックスファイルから UMI 情報を読み、同一の UMI を持つ全てのリードを識別し Total Quality Score の最も高いリードのみを残し、次の処理に回します。同じ UMI をもった他の全てのリードは PCR duplicate として分類され除去されます。

UMI アルゴリズムを使用して重複リードを除去するためには、フォーマット変換を行う前に次の手順を実行してください：

1. Tool メニュー＞Sequence Operation で Sequence Operation ダイアログボックスを開き、Operation type から「Remove duplicate reads」を、ダイアログボックス右下の「Process UMIs based on I2 Index File」を選択します。

   図：Sequence Operation ダイアログボックス、Remove duplicate reads 選択時



2. Input パネルで Add をクリックして、重複リードを除去する R1 および R2. fastq ファイルを選択します。

3. 同じく Input パネルに必要な I2 インデックスファイル（必要に応じて I1 インデックスファイル）を加えます。

4. Output フィールドで出力ファイルの保存先を選択します。

5. OK をクリックします。

   処理が完了するとメッセージが開きます。以下の 2 つのデータファイルと 1 つのログファイルが出力されます：

   _Removed.fastq：UMI 重複を含み解析から除去されるリード

   _Processed.fastq：重複を除外した単一リード

   RemoveDuplicates_Log.txt：入力ファイル、リード（総リード数、ユニークリード数、重複リード数)、リード分布の情報が含まれます。

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